READING THE ADAPTIVE RECEPTOR REPERTOIRES

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF IMMUNOLOGY
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Jacob Eli Gunn Glanville

June 2017

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Mark Davis) Principal Co-Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Scott Boyd) Principal Co-Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Andrew Fire)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Olivia Martinez)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Kari Nadeau)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Robert Tibshirani)

Approved for the University Committee on Graduate Studies.

Abstract

We humans evolve slower than our pathogens. It is the cost of multicellular bodies and long lifespans: how to survive to reproductive age while under continuous siege by viruses and bacteria that can mutate and divide in as little as 30 minutes? To keep up against this ever-changing sea of foes, our bodies produce a special class of remarkably diverse cells: our T and B lymphocytes. As each individual B and T cell matures, it uses specialized genomic shuffling machinery to generate a unique random receptor. Collectively, this results in an army of over 100 million unique B cell receptors and T cell receptors displayed across all of our lymphocytes: the adaptive immune receptor repertoire. The repertoire functions as a pattern recognition system, able to bind to and respond to nearly any pathogen or protein surface. It operates through brute force: by having hundreds of millions of unique receptors, the body will almost always have some with shape complementary to any molecular surface.

While this extraordinary diversity is central to our ability to fend off pathogens, it has also made the adaptive immune system unusually difficult to study. In the history of immunology, every step to characterize these immune receptors, from their isolation, to their genomic locus sequencing, to solving their crystal structures, have been confounded by the underlying universe of variation. Even now, although a history of every immunological battle we have experienced is pumping through our veins, it has remained tantalizingly out of reach for routine inspection. As a consequence, the underlying etiologies of many immune-mediated diseases remain unresolved, and the molecular details of adaptive immune recognition are only partially characterized in almost all adaptive immune responses.

Here we present a means of decoding the adaptive repertoire. We begin in chapter 1 by presenting the development of a series of enabling technologies for analyzing adaptive immune interfaces. Based on high-throughput sequencing, these technologies allow millions of datapoints to be collected for B cell receptors, T cell receptors, single-cell paired heterodimer sequencing, high dimensional association of phenotypes, high-throughput HLA genotyping, and peptide-MHC complexes: all of the core interacting components of the adaptive immune repertoire. In chapter 2, we introduce convergence analysis: a statistical framework for identifying receptors and cells that perform a common function with non-identical but similar bits. We demonstrate that across αβ TCRs, γδ TCRs, BCRs, pMHC pools and phenotypes, convergence analysis provides a unifying framework for interpreting the adaptive immune system, and any degenerate system where combinatorial diversity exceeds functional diversity. In chapter 3, we review the consequences of clonal analysis and convergence groups on the relationship between genetics, environment, and random chance in the repertoire's ability to form specific clones and shared specific responses to common immune stimulation. In chapter 4, a series of studies are presented that illustrate many of the findings that are being made possible by reading the repertoires. In chapter 5, we review interventional studies of selection pressures in the adaptive repertoire that are made evident through the construction of synthetic repertoires in controlled settings. In the final chapter, we comment on some of the new findings and new classes of research that can be anticipated based on these methods.

Repertoire analysis is not a panacea. However, it will improve diagnostics by being able to rapidly identify thousands of reactivities by directly reading the BCR and TCR receptor repertoires. It will improve vaccine optimization by being able to rapidly assay relationships between changes in vaccine composition, and changes in number of clones, affinity of clones, isotype and addinity maturation of clones, and focus on different epitope targeting receptor groups that emerge in response. It will facilitate the rapid de-orphaning of antigen-specificities of TCRs in diseases of unknown cause. It may provide some benefit to discovery of therapeutically relervant receptors from natural responses, and guidance in further engineering of those receptors. Companioned with good experimental designs, it can further our understanding

of the relationship between our genetics, our environment, and the random receptor generation engine inside of us.

Acknowledgement

Behind every paper is a story. In reviewing the body of my scientific work, I am humbled by the combination of those stories, and the remarkable cast of allies that have brought me here.

I'd like to begin by thanking Mark Davis. "Watch out - he insists on innovation." It was the first I'd heard of Mark, and it was a warning. It was also all I needed to hear: I immediately asked to work with him. I couldn't have found a better mentor than Mark. With a fencer's talent for finding an unexpected way immediately to the heart of people and problems, I learned from him how to tease out the essential from the chaff, how to navigate collaborations without being either a pushover or unfair, and that you don't win a race in clean shoes. It was training in a powerful philosophy for approaching science that I will carry forward with me.

I'd also like to thank Scott Boyd. Working with him and his team I was able to dive in immediately to massive new datasets of repertoire data. His hybrid group of bench scientists, reagent engineers, computer scientists and statisticians provided a think tank for reviewing and optimizing high-throughput repertoire characterization methods, and his talent for getting people to collaborate made for a very open sharing environment. I also appreciate him taking me on as an understudy to watch his process of setting up a lab, expanding grant footprints and ultimately obtaining tenure – an anthropological exercise that I hope is of use to me in my career.

I'm deeply appreciative of my Stanford collaborators. Huang Huang, my co-author and partner in crime on GLIPH convergence research really stands out. Brilliant, driven and genuine - I was a better scientist for working with him and our work is something that I will be proud of. Olivia Hatton and Olivia Martinez were early

believers in the convergence analysis, and their deep knowledge of clean successful experimental strategies early on set the snowball rolling for that successful work. The campus was a little less bright when Hatton left. I'm deeply indebted to Allison Nau for our collaborative work together. Spookily-well trained and with huge potential: I'm keeping an eye on her career. Arnold Han was a thrill to work with - he had unique style: a bullish charge on victory every time that I found very impressive, and resulted in some great studies. Holden Maecker and his awesome team at the HIMC - Ji, Weiqi, Yael, Gerlinde. I've learned so much from Holden that I wanted to add him to my committee but I thought Maureen might strangle me for committee bloat. Kari and her great team were wonderful to work with on the peanut allergy work and pretty damn inspiring in her commitment to therapies that can work today: Na lu e-govaned win. I hung on every word while Chris Garcia took the time to describe his strategies for dealing with reviewers and focusing on what matters in a project. He has a stunning ability to identify the essential revelation of an experiment and I hope a little rubbed off. Likewise working with Michael Birnbaum was a rush: such a clever devil with a plan every time. Ramona Hoh really knew her assays and always had time to teach me methods - I hope one day I can repay her for all of that. I could barely keep up with Krishna's epic hacker skills but some rubbed off there too. The whole lab of post-docs and grad students in the Davis and Boyd labs were incredible and I feel lucky to be considered part of their ranks. Rachel Hovde was a joy to work with on the peanut work and generally has been an invaluable resource to bounce ideas off once I feel I'm getting out of my depth statistically. Rishi is one of the cleverest folks I've ever met, and a joy to work with - I look forward to us getting some of our cooler projects out the door over the summer. There are too many other great collaborators to mention.

Beyond just the direct collaborators, the greater immunology and research community was a big part of what made the program so great. Maureen was absolutely wonderful, not to mention infinitely patient, when dealing with my various paperwork muckymucks and generally just always going out of her way to help make sure things came together well. LeScarf was a great source of valuable feedback on how to approach grant-writing and good scientific communication in general. Of course

a critical part of the experience are my fellow PhD Immunology students. Yekyung, Jolien, Tom, Ian, Cesar, Nick, and lil' Jake: you are incredible. Brilliant, hilarious, thoughtful: I don't know how I got so lucky with you guys, but I can't imagine my PhD without you and I know I have made lifelong friends. You kept me humble ("No, Jake, you do not have special powers"), aspirational ("No, it does matter, because if you do this experiment this way you won't be talking to some veterinary group in Kansas, you'll be talking to the goddamn Army!"), attentive ("Well, Ian, what have we here?" "Well, Cesar, looks like we have Jake in the RAGECAGE"), and just generaly epic people. Beyond my (obviously superior vintage) 2012 cohort, the larger group of the years to either direction were just awesome and incredible people. As a bit of a lone wolf all my life, the feeling of community was very touching to me about my PhD experience, something I did not expect and will not forget. I found my peoples here.

I'd also like to thank some of the people outside of Stanford that brought me here and enabled this journey. I'm deeply appreciative to Arvind Rajpal for his scientific mentorship: how to ruthlessly challenge your own ideas from every angle, and only spend your time following the strongest signals - that we are carving our legacy in the choices we make in science, and to build that legacy only atop the strongest foundations. In particular it was Sawsan and Achim Moesta who provided me guidance on how to get into the Stanford Immunology program and thrive on arrival. They also provided me a broadened foundation in immunology beyond my starting points in MHC biology. Rinat was a powerful training that prepared me incredibly well for Stanford because I was surrounded by incredible scientists and a powerful organic culture of solid science: Jaume, Javier, Shelton, Yasmina. Sawsan: you made me a scientist and a bioengineer. Thank you.

Back at UC Berkeley, I owe thanks to Glenys Thompson and Kimmen Sjolander. With Glenys as a freshman I got pulled into the thrilling world of scientific research. Her brilliance at the principles of population genetics, and my realization that I could use programming for something more that computer security pranks, has left deep roots in my life and the choices that followed. It was Kimmen who taught me how to be an algorithm developer, and how to understand and analyze selection pressures

of protein variant populations. A ruthlessly brilliant perfectionist, echoes of Kimmen can be find in every one of my studies and I owe so much of everything that followed to that early training.

Many of these studies were made possible by collaborators across the country and world. Wayne Marasco and Corey Watson - in putting together this thesis I realize just how much our collaborations have enriched my thinking about these problems. It's pretty remarkable what we've accomplished over long distance and the occasional tequila shot. Likewise Yoram Louson, it has been a pleasure working with you and your team back since almost a decade now. Andew Bradbury - congrats again on your new ventures and thanks again for the works we have done together. Jamie Scott and Felix Brendon - you have given me advice throughput my PhD experience that I have found so helpful, and for which I am grateful. Over at USF, I'd like to thank Jen Dever, Cary Lai, Christina, and Patricia for all the work we have done together in training those graduate students. Some of them have helped me win a Gates grant (thanks Nikhil and Chris!), and others have contributed to analysis that I have added to some of my publications (thanks Christina Pettus!). It's been a pleasure working with you and an opportunity for me to ask myself whether I enjoy mentoring, and whether I was any good at it. I think the answer is yes, but maybe I've just been spoiled by the great folks in your program.

On a related note must give thanks to my partners in industry at Distributed Bio. Giles and Chris – you guys have been immensely supportive as I struck out on the audacious task of PhD and building a company simultaneously. Thanks for your patience and believing in me. I'd also like to thank my team - both for being so incredibly competent that I was able to not be around all the time, but also in being so incredibly interesting that I would want to, and could always trust an awesome conversation about heavy-duty bioengineering problems if I needed a mental break from my PhD work. Maurer - it's a joy working with you, and being able to jump into the headcandy of cell optimization. You are unnaturally creative and brilliant in a way that I always find so refreshing. Sarah, the universal epitope focusing project was a massive set of responsibilities with so many moving pieces, collaborators and countries to contend with, and thanks to you, one of my favorite dreams has come

to life. You've also just saved me from my own nonsense: I can't even keep count of the number of projects you've saved from disaster by coming in with a classic Sarah "Jake, I've been thinking, and this whole project is wrong unless we do X,Y,Z." Lauren, we've achieved remarkable success since your arrival, and the timing was perfect for a period where I needed to go the final mile with the PhD work. Thank you for taking the pressure off and being such a remarkably talented collaborator and rock for the whole group. We work well together. Devanshi, Christina, Ray - you are awesome: keep killin' it.

Out side of these communities of science, my close friends have carried me through this process: taking me out for boxing practice, or playing Munchkins, or dungeons and dragons. There are too many of you to list individually, but thank you all.

Finally, I would like to thank my family. Erin Flynn, my partner in life, adventure, science, and love - we have been there for each other for these many years of our PhDs. You put up with me and comforted me during the haze of my crazy dream of growing a startup and finishing a PhD at the same time. Your epic scientific adventures to gather fish eggs from the Antarctic ice sheet, swallows from the windy arctic, and seals along the beaches of Marine have constantly fascinated and inspired me (and sound like 2/3s of a damn fine omelet!). I love being able to come home and talk science with you. Your patience and social graces have always balanced out my Chaos Muppet tendencies, and I am greater with you. Thank you for everything that you are. My Alaskan family - Cathy, Rod, Laura, Colin... thank you for your love and support all these years, and for Erin. Finally, mother, father, and brother. No matter how chaotic our lives, you always poured more than you had to make sure I had an education. From Calvert school at home during the civil war, to driving us across a choppy lake at night to make Panajachel classes, to finding me more advanced math tutors when I outgrew what dear shirtless Jack could teach me in 6th grade, to helping support me in college - you never gave up on me and always were there to take less yourself so that I would push further. This work is for you.

Contents

List of Tables

List of Figures

Chapter 1

Establishing Repertoire Analysis Technologies

1.1 Introduction

The adaptive B and T lymphocytes perform molecular recognition of antigen with extremely diverse receptor repertoires. Since 2008, the emergence of high throughput DNA sequencing technologies have provided a mechanism for deep characterization of millions of receptors, HLA genotypes, and cognate antigens. Combined with appropriate reagent engineering and computational platforms to productively process and interpret the output data, this new class of repertoire analysis technologies provides a means of analyzing the adaptive immune repertoires at sampling scales proximal to their diversities. In chapter 1 we review the development of repertoire analysis technologies for antibody repertoires, αβ TCR repertoires, g/d TCR repertoires, and pMHC antigen repertoires: the central interacting components of the adaptive immune receptor system.

In the first study, "Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire," by Glanville et al in PNAS from 2009, we provide the first published example of high-throughput sequencing of human antibody repertoires. The study developed Hidden Markov Models as an instrument of analysis of adaptive repertoires that are resilient to read

error, PCR error, affinity maturation, and genotype polymorphism. By evaluating the repertoire before and after selection against antigen when displayed on phage, it provided a controlled environment for quantifying antigen-driven repertoire bias. In the second study, "Diversity and clonal selection in the human T-cell repertoire," by Qi et al in PNAS from 2013, we present a primer set for T-cell receptor bulk sequencing on Illumina MiSeq that has been adopted in multiple subsequent studies, and established multiple valuable metrics for estimating the diversity of a natural repertoire. In the third study, "Linking T-cell receptor sequence to functional phenotype at the single-cell level" by Han, Glanville Hansmann and Davis, we present an easily adoptable method for multi-locus single-cell sequencing. Demonstrated on the α-chain, β-chain, and a panel of phenotypic markers of tumor-inflitrating lymphocytes, it provided a mechanism for extracting the complete αβ heterodimer receptor as well as phenotypic encoding of phenotypic markers on over 2000 single cells in a single sequencing experiment. In the fourth study, "Deconstructing the peptide-MHC specificity of T cell recognition," the Garcia laboratory developes a yeast-display library of peptide-MHC complexes, where a library of variable peptides are presented on an engineered MHC. When panning this library with a TCR of interest, we repeatedly demonstrate that a swarm of thousands of related peptides can be selected for that share amino acid features: an early demonstration of the properties of convergence that will be greatly expounded on in chapter 2. We end chapter 1 with a review of the technological, bioengineering, bioinformatic, and theoretical considerations and best practices required for successful repertoire analysis.

1.2 Antibody repertoire sequencing

Antibody repertoire diversity, potentially as high as 10^11 unique molecules in a single individual, has historically confounded characterization by conventional sequence analyses. In this study, we present a general method for assessing human antibody sequence diversity displayed on phage using massively parallel pyrosequencing, a novel application of Kabat column-labeled profile Hidden Markov Models, and translated complementarity determining region (CDR) capture-recapture analysis.

Pyrosequencing of domain amplicon and RCA PCR products generated 1.5 x 10^6 reads, including more than 1.9 x 10^5 high quality, full-length sequences of antibody variable fragment (Fv) variable domains. Novel methods for germline and CDR classification and fine characterization of sequence diversity in the 6 CDRs are presented. Diverse germline contributions to the repertoire with random heavy and light chain pairing are observed. All germline families were found to be represented in 1.7 x 10^4 sequences obtained from repeated panning of the library. While the most variable CDR (CDR-H3) presents significant length and sequence variability, we find a substantial contribution to total diversity from somatically mutated germline encoded CDRs 1 and 2. Using a capture-recapture method, the total diversity of the antibody library obtained from a human donor Immunoglobulin M (IgM) pool was determined to be at least 3.5 x10^10. The results provide insights into the role of IgM diversification, display library construction, and productive germline usages in antibody libraries and the humoral repertoire.

1.2.1 Introduction

The humoral immune response recognizes novel molecular surfaces by exposure to a vast repertoire of potential binding partners (1). Antibody paratopes, the agents of humoral molecular recognition, mediate specific binding through a protein-antigen interface that varies dramatically between molecules. When confronted with a novel antigen, the chance that any given antibody in the pool will bind is low. Therefore, it is primarily the diversity of the antibody repertoire that determines whether a specific complementary paratope will be recovered (2).

Under such selective pressures, a number of mechanisms to maximize the recognition potential of the antibody repertoire have evolved. Antibody paratopes are found at the hypervariable region of a light and heavy chain heterodimer. Each chain contributes 3 loops to a spatial cluster of complementarity determining regions (CDRs). CDRs 1 and 2 are encoded in germline V-segment loci: 51 VH and 70 Vk/l loci, each with unique amino acid encodings, exist in a typical human haplotype (3–5). Diversity in each chain is determined by combinatorial VH-(DH)-JH (for the heavy) or

Vk/lJk/l (for the light) rearrangements, P and N-addition, junctional flexibility, and somatic hypermutation of variable domain nucleotides, with a concentration on CDR encoding regions (6, 7). The combinatorial association of such stochastically generated light and heavy chains has the potential to generate many orders of magnitude more diversity than can be uniquely displayed on the 10^11 B-cells in a single individual's lymphocyte population (2, 8). With each antibody variable fragment (Fv) encoded by at least 650 base pairs, the presented repertoire is potentially 4 orders of magnitude larger than the entire human diploid genome (6.4 x 10^9 bp).

Such extreme sequence diversity poses multiple challenges to repertoire characterization efforts. Achieving sufficient sampling depth to determine total diversity is impractical with Sanger-based sequencing (6, 7). High-throughput sequencing methods, while able to address sampling depth, have until recently produced read lengths under 200 bp; too short to span 3 CDRs in a single read (9). While in other settings these technologies can rely on assembly to overcome short read lengths (10), the diverse yet repetitive char acter of antibody Fv reads cause assembly to either fail or return erroneous chimeric contigs that do not represent individual population members (10, 11). Once sequences are obtained, somatic hypermutation and junctional diversity pose a challenge to reliable CDR boundary identification (12–14). The problem is significant: a recent study reports that over 10% of the variable domain sequences in the Kabat antibody database have been misnumbered by existing methods (15). Given these limitations, past diversity assessment efforts have focused on low-resolution length-based CDR3 spectratyping (16) and local nucleotide-level V-(D)-J assessment of more limited TCR b-chain (10) and zebrafish repertoires (17). While these approaches provide valuable insights into specific features of binding site diversification, it has not yet been feasible to characterize the combined effects of diversification on the complete translated paratope at molecular resolution.

Long-read high-throughput sequencing chemistry, Bayesian fold recognition and a single chain variable fragment (scFv) architecture have created an opportunity for complete paratope repertoire analysis. Recent advances in high-throughput pyrosequencing chemistry have allowed 10^6 400 bp sequences to be generated in a single run: deep enough for capture-recapture diversity assessments and long enough to span

all 3 CDRs of a chain in a single read. Once read, a novel application of Kabat-labeled profile Hidden Markov Models (HMM) borrows from advances in remote homology fold recognition to provide an O(n) fast, highly accurate unified Bayesian framework for domain recognition, CDR boundary identification, and multiple sequence alignment (18–21). With reliable access to the entire CDR contribution of a variable domain in a single read combined with shotgun reads spanning the heavy-light chain pair-ing, it becomes possible to directly estimate and characterize the number of unique binding surfaces presented by an antibody library repertoire.

Accurate diversity assessment is of particular interest during the construction of combinatorial antibody repertoires (22, 23). Phage display libraries allow an antibody repertoire to be queried with a candidate antigen directly, without the need to proceed through in vivo immunization (24, 25). A number of strategies for introducing repertoire diversity during library construction have been proposed (26–28) but existing methods to assess the final functional library diversity are based on estimates of transformation efficiency and limited sequence sampling. Here we present the design and assessment of a scFv library built directly from the complete germline diversity of 654 human donor Immunoglobulin M (IgM) repertoires; a lymphocyte reservoir that includes naïve, memory, plasma, and preimmune somatically altered paratopes (29–31). The available diversity of the entire library was assessed directly using high-throughput pyrosequencing to generate datasets large enough to perform capture-recapture diversity (17) and chain assortment estimates. We also compare the repertoire and diversity of the input library to functional binders derived from panning the library against 16 diverse antigens. The results provide a powerful method for monitoring diversity during future library construction efforts and fundamental insights into the strategy of functional paratope diversification elected by evolutionary forces.

1.2.2 Results

Antibody Library Generation. Heavy and light chain V-genes from 654 healthy human donors were separately amplified by PCR using equimolar mixture of degenerate

family primers (32, 33) individually validated at a common reaction condition of 25 PCR cycles at 94°C for 45 sec, 58°C for 45 sec and 72°C for 60 sec. The heavy and light domain products were randomly associated in a scFv VH-(G4S1)3linker-VL architecture. A total of 120 ug of scFv antibody repertoire from 72 ug of VH-Vk and 48 ug of VH-Vl was obtained and ligated into a display vector. Three hundred and ten transformations yielded 302 ml total library volume. Plating of serial dilutions for colony counting resulted in an estimated 3.1 x 10^10 (SD 0.7 x 10^10) successful transformants containing scFv antibodies in the total remaining 301 ml library pool.

Antibody Library Selections. Sequences for more than 1.7 x 10^4 nonredundant antibodies were obtained from output generated by panning against 16 human and nonhuman targets. These sequences were CDR clustered using the profile HMM CDR and germline classification methods used on the library. For each antigen, an average of 30 representative sequences from distinct CDR clusters were selected for further characterization: affinities obtained ranged from less than 100 pM to over 1 uM.

454 Sequencing. Each of the 2 samples, rolling circle amplified (RCA) shotgun and variable domain PCR amplicon, were sequenced using the GS FLX Titanium large PicoTiter plate in 2 separate sequencing runs. The 2 sequencing runs combined yielded 1,452,529 and 1,602,399 raw well reads for the shotgun and amplicon library, respectively. After the signal processing step of the 454 data analysis pipeline, where reads may be rejected by multiple signal and quality filters, we obtained 923,876 (shotgun) and 554,310 (amplicon) quality filter-passing reads.

Accuracy of Profile HMM CDR Classification and Kabat Numbering. In 779 benchmark cases, 99.8% of CDR-H3 loops were classified correctly, and all other CDRs received perfect classification (supporting information (SI) Fig. S1A). The single error was due to assignment of stem residue, H102, to a neighboring insert state at the C terminus of the H3 (Fig. S1B). With a CDR boundary insert correction, all cases classified correctly. Using the scFv HMM (Smith/Waterman local alignment, expectation value <1e-10, >70% match state occupancy in all FW regions in single reading frame), 96,303 heavy and 98,946 light chain reads spanning entire variable domains in a single reading frame were identified in the pyrosequencing results, aligned, and CDR labeled.

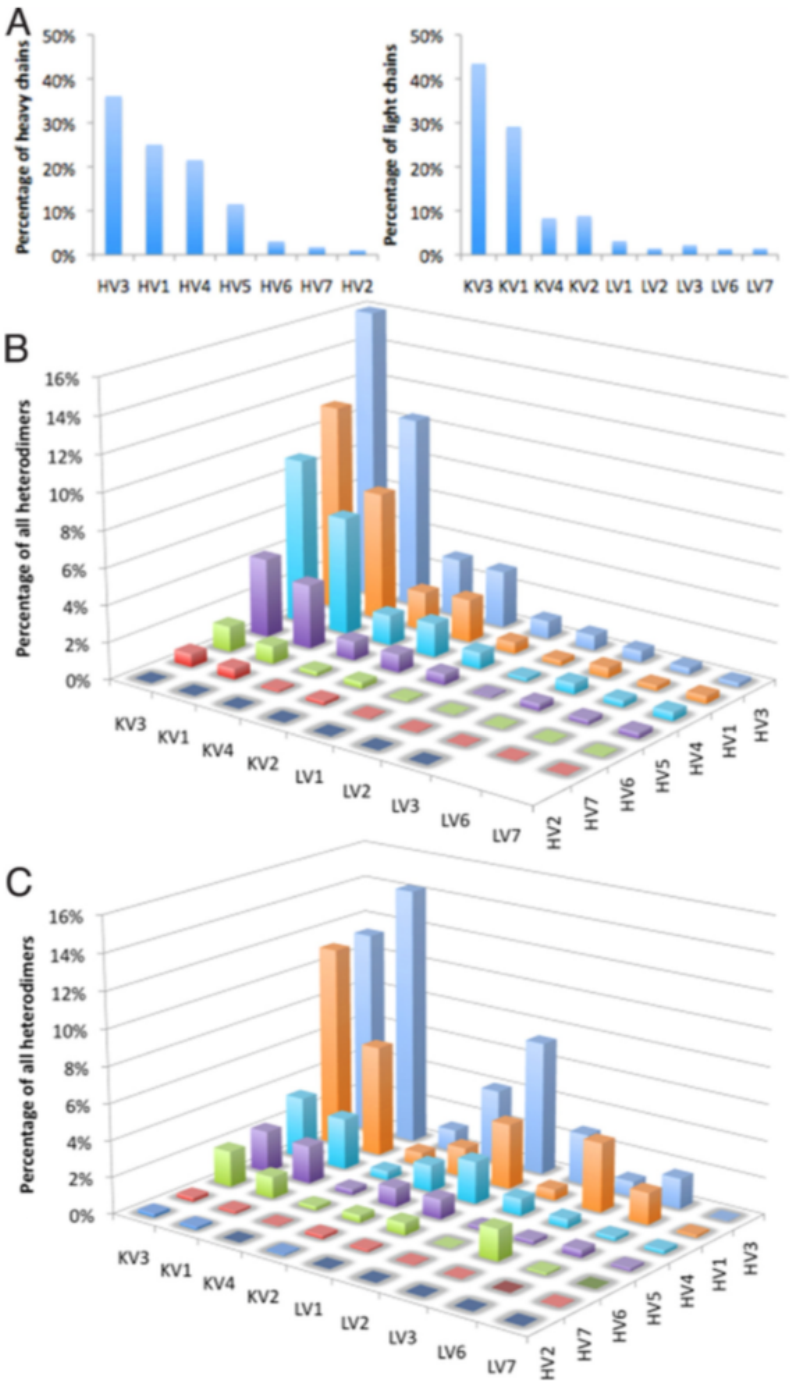


Figure 1.1: Heavy and light family frequencies and pairing observed in 18,158 RCA reads. (A) Heavy (Left) and light chain (Right) family frequencies observed in the library. Only families at 1% chain frequency are shown. (B) Heavy and light chain family pairings occur in proportion to the abundance of their partner, indistinguishable from a null model of random assortment by Chi-squared (P value: 0.9512). (C) Heavy and light chain families pairings in second and third round Panning show pairing preferences that cannot be explained by random assortment (P value: 0.00107).

HV	Ampl.	RCA	Pann.	Kabat	KV	Ampl.	RCA	Pann.	Kabat
5–51	17.21	14.17	7.94	6.14	3–20	27.90	22.98	14.55	7.32
1–69	9.97	10.00	12.51	5.70	1–39	11.69	11.12	11.89	7.32
1–2	8.34	7.35	5.85	2.78	3–15	9.61	9.48	2.70	3.38
4–59/61	7.66	8.22	4.67	6.38	4–1	8.29	8.57	5.71	3.78
3–30/33	7.65	7.74	6.75	10.10	1–5	6.34	7.00	7.40	2.63
3–23	6.09	6.39	8.39	10.13	2–28	4.87	5.28	5.43	3.72
1–46	5.19	5.43	3.50	1.51	3–11	3.57	3.68	2.42	6.24
1–18	4.54	4.59	5.12	2.45	1–12	3.36	2.81	2.84	1.03
6–1	3.80	3.21	5.85	5.42	1–33	2.24	2.68	1.75	2.00
1–8	3.45	2.89	2.92	1.02	1–16	1.43	1.49	0.60	0.69
3–7	2.99	2.97	3.55	4.40	3-NL5	1.37	0.99	0.70	0.06
4–4	2.24	2.46	2.26	1.51	1–27	1.31	1.11	0.78	1.09
3–48	1.95	2.11	7.27	2.81	2–30	1.28	2.13	1.58	1.26
3–9	1.87	2.04	6.66	1.29	1–9	1.10	0.97	4.80	1.66
7–4–1	1.84	2.43	0.84	0.66	1–17	0.93	1.09	0.41	1.09
4–30–4/31	1.72	1.48	0.73	2.42	1–6	0.68	0.77	0.96	0.46
3–74	1.57	2.11	1.72	1.21	2–24	0.58	0.87	0.35	0.51
4–39	1.44	1.69	0.88	3.22	2–40	0.51	0.80	0.32	0.23
3–53/66	1.28	1.13	2.44	1.43	1–13	0.51	0.71	0.25	0.17
3–21	1.28	1.24	1.16	2.06	1D-16	0.42	0.23	0.09	0.23
3–11	1.10	1.21	0.72	1.95	6–21	0.39	0.49	0.34	0.11
3–15	1.02	1.53	1.35	2.86	2D-29	0.37	0.36	0.17	0.51
4-b	0.84	1.39	3.05	0.83	3-NL1	0.28	0.23	0.03	0.00
1–3	0.76	1.07	1.01	1.95	1D-17	0.26	0.24	0.09	0.06
5-a	0.57	0.51	0.26	1.87	1–8	0.25	0.38	0.18	0.57
2–70	0.54	0.83	0.22	0.44	3D-20	0.23	0.21	0.04	0.23
3–64	0.45	0.43	0.20	0.41	2–29	0.22	0.27	0.25	0.00
3–13	0.40	0.54	0.23	0.52	**LV**	**Ampl.**	**RCA**	**Pann.**	**Kabat**
3–49	0.38	0.46	0.26	0.66	6–57	1.62	1.77	6.02	0.11
1–58	0.31	0.35	0.08	0.22	3–21	1.36	2.05	6.38	5.26
3–72	0.29	0.37	0.15	0.33	1–40	1.00	1.19	4.42	4.92
3–73	0.25	0.32	0.11	1.43	2–14	1.00	1.13	3.74	9.21
1–24	0.21	0.31	0.12	0.30	1–44	0.82	1.03	4.74	3.78
2–5	0.18	0.29	0.29	1.27	3–19	0.66	0.62	0.53	2.80
3–43	0.16	0.17	0.17	0.08	3–1	0.57	0.69	0.63	4.23
3-h	0.12	0.10	0.14	0.03	1–47	0.54	0.70	2.26	1.72
3–20	0.11	0.12	0.33	0.14	8–61	0.51	1.19	0.79	2.17
1–45	0.06	0.07	0.04	0.17	7–46	0.44	0.55	0.61	0.63
4–30–2	0.05	0.05	0.03	0.50	1–51	0.43	0.58	1.83	3.49
4–34	0.05	0.09	0.05	10.81	7–43	0.30	0.56	0.26	1.20
4–28	0.04	0.06	0.00	0.17	10–54	0.18	0.20	0.33	0.69
1-f	0.03	0.05	0.10	0.17	2–23	0.13	0.13	0.07	1.26
2–26	0.01	0.03	0.06	0.25	2–8	0.08	0.11	0.18	2.52

Table 1.1: Partial list of V-segment germline distribution in 100% nonredundant Kabat sequences (Kabat), scFv domain-specific amplicons (ampl.), scFv RCA, and Sanger sequencing from 1.7 x 10^4 nonredundant round 2 and 3 binders against 16 unique antigen targets (Pann.). Forty-eight of 51 IMGT functional HV germlines were recovered, 53 of 70 K/L germlines were recovered. (see Table S1 and S2 for complete list)

GS-Linker Assessment. Of the subset of RCA shotgun library reads that spanned the GS-linker and framework regions of VH and Vk/l domains to either side, 95.6% appeared as expected by design. The remaining 4.4% had predominantly single errors that could be genuine linker errors or pyrosequencing read errors.

Germline Classification of Library Clones. In >250,000 simulations, sequences with less than 30 mutations in the V-segment were never misclassified. Even with up to 50 simulated mutations, 99.97% of test sequences were correctly classified, with only 5.8% receiving reduced family-level classification (Fig. S2 A and B). This limited rate of resolution reduction corrects 84% of sequences misclassified by a naïve

approach. By comparison, 95% of all antibodies recovered from the library differed by less than 30 mutations from the closest germline allele (Fig. S2C).

Using the classification method, all Ig-bearing library sequence reads were classified to germlines (Table 1). Forty-eight heavy and 53 light known functional V-segment germline loci were encountered at least 10 times in the full-domain sequence reads (Tables S1 and S2). In addition, 2 germlines listed as pseudogenes by International ImMunoGeneTics Information System (IMGT) (HV3-h and HV3–71) were recovered in rearranged form (127 and 240 times, respectively). Germlines sampled were consistent between variable-domain amplicon and RCA-amplified shotgun library samples and resemble the distribution found in Kabat database, although some differences are observed (Kolmogorov-Smirnov assessments: Amplicon vs. RCA D = 0.0930; P = 0.989, Amplicon vs. panned D = 0.1163, P = 0.917, Amplicon vs. Kabat D = 0.1395, P = 0.765).

Heavy and Light Chain Pairing. Heavy and light chain family pairings in the library were found to occur in proportion to the abundance of the respective families: indistinguishable from a null model of random assortment (X2 observed vs. expected: H/L: P value: 0.9512) (Fig. 1 A and B). While most of the dominant germline families are represented in the chain pairings of leads generated after panning, in panned sequences we observe nonrandom assortment of families (X2 panned vs. library H/L: P value: 0.00107), illustrated by the deemphasis of KV4 and the increased lambda contribution (Fig. 1C).

CDR-H3 Diversity. The CDR-H3 length distribution was consistent across site-directed and RCA-amplified shotgun library preparation approaches. A Poisson distribution with mean 11.5 (Kabat 95–102), as observed by others, was consistent with results found here (7), although an increase in H3 of length 5 was observed. (Fig. 2A).

Somatic Mutation Distribution. In V-segment encoded CDRs (1 and 2), 17% of sequences were unaltered from germline, while 78% of sequences had between 1 and 6 aa mutations (Fig. 2C; see Tables S1 and S2 for details). The definition of somatic mutation used counts distance from closest germline allele and could therefore be inflated by novel allelic and copy number loci variations not found in IMGT. Position

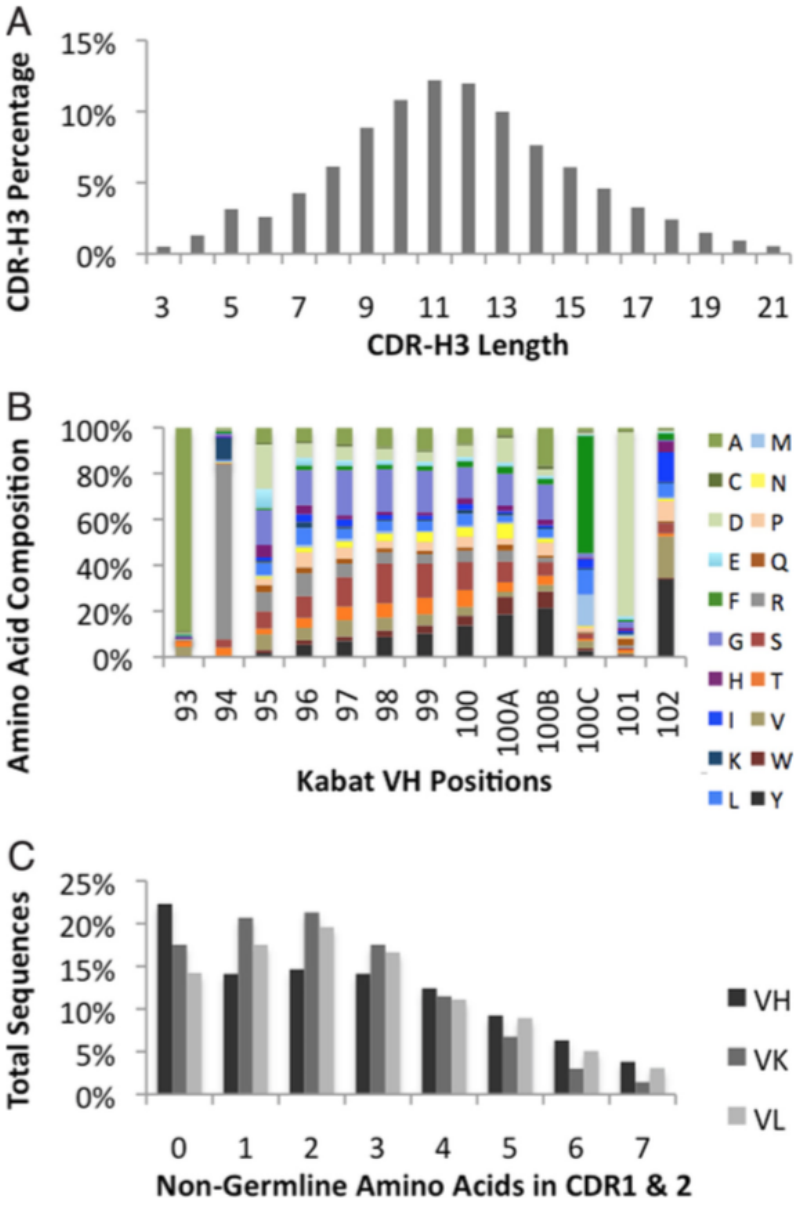


Figure 1.2: CDR-H3 length and amino acid composition for the most common length bin. (A) Observed CDR-H3 length diversity from 65,240 amplicon and 22,769 RCA reads. (B) Position specific amino acid frequencies for 10,281 length 11 (determined by positions 95–102) CDR-H3s. (C) Number of nongermline encoded amino acids found in CDRs 1 and 2, by domain type.

specific scoring matrices of all sequences for each germline show a pattern of somatic hypermutation consistent with that previously reported (34).

Total Diversity Estimate. The observed diversity in the heavy chain is dominated by contributions from the CDR-H3 (Fig. 3A), while that observed in the light chain is more evenly contributed by all 3 CDRs (Fig. 3B). In the total paratope contribution from the heavy chain, diversity contributions from H1 and H2 more than double the diversity found in H3 alone. In the light chain, combined CDR diversity is more than 6-fold higher than diversity in any single CDR. Figure 3 C and D, showing percent recapture at sampling depths for the heavy and light chain CDRs, respectively, recapitulates patterns observed in the diversity estimates (Fig. 3 A and B) with CDR-H1 and CDR-H2 approaching saturation at lower sampling depths than CDR-H3. In the light chain, CDR-L2 saturates at lower sampling depths than both CDR-L1 and CDR-L3. In general, the approach to saturation for the light chain is more rapid than the heavy chain counterpart. A lower bound estimate of 2.2 x 10^5 (SD 2.2 x 10^3) diversity for heavy domain, and 1.6 x 10^5 (SD 0.8 x 10^3) for the light domain is obtained by nonredundant capture-recapture at M = 33,000 as rarefaction trends toward asymptote in Figs. 3 A and B.

1.2.3 Discussion

Direct analysis of phage displayed paratope diversity was made possible by recent advances in long-read pyrosequencing, a novel application of profile HMM-based sequence analysis, and syntenic placement of Fv chains in the scFv construct. Long read lengths allowed the entire CDR contributions from variable chains to be assessed in concert without assembly. A combination of variabledomain amplicon and RCA of plasmid library sample preparations used for sequencing allowed for the depth required for effective capture-recapture based diversity assessments, heavy and light chain paring assignment, and scFv construct integrity (including the GS linker) to be evaluated directly. HMM-based residue labeling provided the flexibility required for

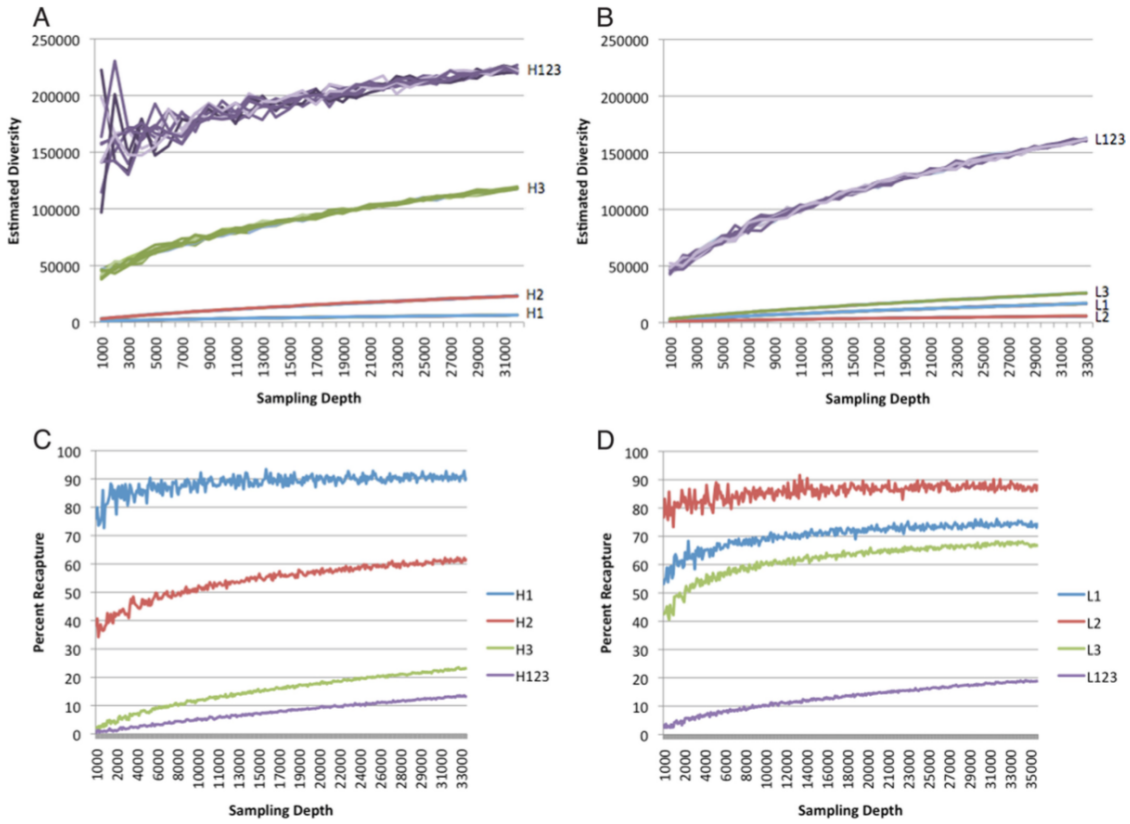


Figure 1.3: Diversity estimates for heavy and light chain CDRs in the antibody library. (A and B) 10 capture-recapture rarefaction results for CDR1, 2, 3 and concatenated CDRs for heavy and light chain, respectively. (C and D) Percent recapture during rarefaction analysis.

accurate CDR identification and residue-specific somatic hypermutation rate assessment in a diverse sequence space. Germline analysis provided a mechanism to assess the dispersion of potential paratope diversity in an antibody repertoire.

Considering only CDRs and using a stringent definition of diversity that requires at least 2 amino acid mutations from any other sequence for each chain to be considered unique, a lower bound diversity estimate of 2.2 x 10^5 was determined for the heavy chain CDRs, and 1.6 x 10^5 for the light chain CDRs. While it is possible that this strict definition of diversity underestimates the total number of unique molecules available in the library, doing so minimizes the chance that read errors in assembly-free sequence could contribute to artificially inflating the recapture diversity estimates. Given the observed random pairing of heavy and light chain variable domain families in the library (Fig. 1), we estimate the combined library diversity of unique nonredundant paratopes to be near 3.5 x 10^10. This estimate is quite similar to the predicted number of transformants recovered during library construction (3.1 x 10^10, SD 0.7 x 10^10). While rarefaction studies do not show complete saturation for the concatenated CDRs at sampling depths displayed (Fig. 3 A and B), the projected asymptotes suggest that the diversity of a well-constructed phage displayed antibody library is limited by transformation efficiency.

The single largest source of paratope diversity in antibodies is derived from variation in the heavy chain CDR3 loop. In the library, CDR-H3 lengths ranged from 1 to 31 aa, following an apparent Poisson distribution (Fig. 2A). While every amino acid could be found at most positions in CDR-H3, the distribution of abundances varied depending on sequence length (Fig. S3) in a manner characteristic of these loops previously observed in human antibodies (27, 34).

Diversity in V-segment encoded CDR 1 and 2 is dispersed by germline origin and diversified by somatic hypermutation. The 48 heavy and 53 light germlines found in the library provide 2,544 distinct potential heterodimer pairings. Library antibodies pair randomly at the family level in the scFv construct. If this trend can be assumed to hold true at the level of germlines, then based on frequencies observed (Table 1; see Tables S1 and S2 for complete list) and a library size of 10^10, one can expect to find between 105–107 unique combinations of most heterodimer germline pairs

in the library. In each chain, germline encoded CDRs 1 and 2 vary by average of 4.4 aa from the next closest germline. This gap in CDR sequence space between germlines is thoroughly explored by somatic hypermutation, with 78% of antibody chains displaying between 1–6 aa mutations in their V-segment encoded CDRs. The total contribution of germline origins and somatic mutation is substantial: estimates that consider all 3 CDRs on each chain result in a 12-fold increase in total diversity compared an estimate derived from H3 and L3 alone (Fig. 3).

When panning 16 diverse protein test antigens with the phage displayed antibody library, multiple binders were always successfully recovered. In sequencing over 1.7 x 10^4 leads recovered during the second and third round of panning, antibodies were found to be derived from almost every germline available in the library. For each antigen panned, at least 30 unique antibody clusters were identified that explored multiple distinct epitopes across their respective target's surface. The diversity of binders recovered is a strong indication that the deliberate inclusion of diverse frameworks into the library contributed directly to accessible diversity. While some variation in germline frequency (Table 1) and chain pairings (Fig. 1C) were observed, the overall similarity in distributions between the phage display library and panned binder products suggests that germlines were recruited in rough proportion to their availability. This indicates that the display of antibody on phage in a scFv format does not significantly restrict the available diversity accessible to in vitro panning.

Even with 302 transformations, diversity came close to saturating transformants. While heavy and light chain shuffling and pooling of cDNA from multiple donors during library construction makes it difficult to speculate on whether this is a direct reflection of the diversity of individual antibody diversity or of the degree of humoral overlap between individuals, it does suggest that the maximum possible repertoire diversity supported by human germlines may yet be fully explored by phage display. While this library design was successful, biases during construction can always impact the final functional diversity and not be readily noticeable using traditional diversity estimation techniques. In future library development projects, deep sequencing in conjunction with germline and CDR analysis could provide a powerful quality assurance technique to identify and correct biases that may be introduced during each

stage of library construction. This general methodology has immediate applications in quality assurance during library construction as well as potential applications in antibody repertoire assessment in disease states.

1.2.4 Methods

Construction of Human Naïve scFv Library. Total RNA and/or mRNA were obtained from 637 healthy human peripheral blood leukocyte donors (BioChain Institute and Clontech) and 17 human spleens (BioChain Institute, Clontech, and OriGene). First strand cDNA was synthesized by using

human heavy chain constant region primer

HuIgM (5'-TGGAAGAGGCACGTTCTTTTCTTT-3'),

human k constant region primer

HuCk (5'-AGACTCTCCCCTGTTGAAGCTCTT-3')

and human l constant region primer

HuCl (5'-TGAAGATTCTGTAGGGGCCACTGTCTT-3')

according to vendor specifications (Invitrogen). Human heavy and light chain V-genes were separately amplified by PCR using equimolar mixture of degenerate family primers (32, 33): 9 VH and 4JH for VH genes, 7 Vk and 5 Jk for Vk genes, and 9 Vl and 3 Jl for Vl genes at a final concentration of 10pmol/ul of each primer. The amplified products were assembled as VH-(G4S1)3linker-VL scFv antibodies according to Marks and Bradbury (32). The assembled scFv antibody repertoire were purified, cut with SfiI, and cloned into a vector. TG1 competent cells (Stratagene) were transformed in with the scFv vector by electraporation using BTX 1 mm gap cuvettes in 310 parallel reactions. Transformation efficiency was estimated by colony counting of plated serial dilutions drawn from 1 ml of the 302 ml posttransformant pool before any incubation.

Selection of Human Antibodies from scFv Phage Display Library. Phage antibodies were prepared from scFv library glycerol stocks according to published protocols (33). The specific antibodies to 16 diverse antigens were selected and screened by

ELISA and BIAcore assay after 3–4 rounds of biopanning through either immobilized antigen at 5–10 ug/ml for solid phase or biotinylated antigen at 20 –200 nM for solution phase depending on the antigens according to standard protocols (33). The obtained antibodies were sequenced and grouped by CDR clustering for further analysis. 454 Titanium Sequencing. Two sample preparation strategies were used for sequencing the scFv library. High-depth bidirectional variable domain coverage was provided by PCR amplification of the VH and Vk/Vl scFv insert regions using amplicon specific primers complementary to the constant regions of the vector and the GS-linker and harboring the 454 Titanium adaptor sequences to generate Ig amplicon libraries. Vector composition and read error-rate were assessed by sequencing single-stranded RCA-shotgun libraries generated by RCA of the whole vector, followed by random shearing and ligation of 454 Titanium adaptors. The sequencing runs using the Roche/454 Genome Sequencer FLX were set up according to the 454 Titanium Sequencing protocol. Please see SI Text for RCA and shotgun Library preparation, PCR amplicon library preparation (including primer sequences used, Table S3), and sample library titration for 454 Titanium sequencing. Information on controls performance, loading density, and signal intensities in addition to signal processing and run yield are listed as well. Sequence Analysis. Translation, multiple sequence alignment, Kabat numbering and identification of structurally conserved CDR boundary positions were performed with HMMER profile hidden Markov models (18) designed to represent the scFv architecture. The HMM was trained with normalized concatenations of 95% nonredundant IMGT (35) germline V and J segment amino acid multiple sequence alignments, a direct GS-linker encoding, and a permissive insert D segment. Direct mapping of Kabat numbering system to columns in the HMM allowed specific Kabat positions and ranges to be identified. CDRs were bounded by conserved Kabat positions H1 31–35, H2 51– 61, H3 93–102; L1 24 –34, L2 50 –56, and L3 89 –97 that could be identified with high accuracy (details in SI Text). The approach was evaluated with a benchmark of 779 superposed nonredundant antibody structures from multiple species. Reads with frames bearing 10^-10 or better expectation values to the model were aligned to and annotated by the profile. Seventy percent match state

$$S_i = 1 - \sum_{g=1}^{G} \frac{P((\lambda - \delta_{ig}),(\mu_g - \delta_{ig}))}{P(\lambda,\mu_g)} \Big| g \neq i$$

Figure 1.4: Probabilistic Germline Classification Formula 1, where Si is the confidence that primary hypothesis i is correct, l is the common alignable query sequence length, u is the observed number of mutations between query and germline i, d is the distance difference between the primary hypothesis i and the alternate g, and G is the total number of germlines. In cases where the sum of alternate hypotheses was greater than the cutoff 10^-3, family classification was attempted.

coverage in neighboring framework regions was a read-through requirement for CDR and GS-linker analysis.

Germline Classification. Classification of V segment germlines was performed by nucleotide comparison to IMGT database of allelic variants using BLAST (36). Classification was made to germline, and a subset of very similar germlines were pre-grouped during classification (Indicated by solidus in Table 1). To address increased risk of misclassification in a mutation-rich sequence space, confidence in classification was assigned using a benchmarked probabilistic framework. Confidence that the top hit was the correct germline was determined by Probabilistic Germline Classification Formula 1, where Si is the confidence that primary hypothesis i is correct, l is the common alignable query sequence length, u is the observed number of mutations between query and germline i, d is the distance difference between the primary hypothesis i and the alternate g, and G is the total number of germlines. In cases where the sum of alternate hypotheses was greater than the cutoff 10^-3, family classification was attempted. The method's ability to classify sequences with somatic mutations and read errors was benchmarked by simulation: >250,000 sequences derived from human frameworks and bearing progressive simulated somatic mutation loads.

Diversity Assessment. Nonredundant functional binder diversity was assessed per domain, using only translated CDR sequences from single reading frames spanning the entire variable domain. By ignoring silent mutations and any mutations that occurred in the framework, sequencing error effects were minimized and only variation most likely to impact binding specificity was evaluated. Diversity was determined

with capture-recapture rarefaction as previously described (17) with 2 modifications: 1) the entity compared when assessing recapture was the translated amino acid content of CDRs, and 2) a rigorous nonredundant diversity assessment, counting any CDR concatenation with less than 2 amino acid differences as being a functionally equivalent recapture, was performed for each estimate.

454 Sequencing: Rolling Circle Amplification and Shotgun Preparation. The TempliPhi DNA Amplification kit (GE Healthcare) was used to exponentially amplify the double-stranded circular plasmid DNA templates containing the scFv insert by RCA. The RCA reactions were set up for 18 h using the manufacturer's protocol. Three micrograms of amplified plasmid DNA sample was sheared per microTube on the Covaris E210 instrument (Covaris Inc.). The high molecular weight RCA amplified plasmid DNA was fragmented for 87 sec using 5% duty cycle, 200 cycles per burst and intensity of 3. Fragmented DNA was purified using AMPure SPRI beads (Agencourt) to remove fragments smaller than 400 bp. One ul of purified DNA was analyzed using the DNA 7500 chip assay on the 2100 Bioanalyzer (Agilent Technologies) to make sure that >80% of the fragments were between 350-1000 bp. The purified fragmented DNA was further processed according to the 454 FLX Titanium Library construction kit and protocol (Roche Applied Science) to ligate adaptors specific to the Titanium sequencing chemistry. The resulting single-stranded DNA library was assessed for size distribution using the RNA 6000 Pico chip on the 2100 Bioanalyzer (Agilent Technologies) and quantified using the Ribogreen RNA Quantitation Kit (Invitrogen) on a Synergy 2 (BioTek Instruments Inc.). Emulsion PCR (emPCR) and titration by enrichment at 0.5, 1, 2, 4 copies per bead (cpb) was carried out according to the 454 Titanium emPCR protocol to determine the optimal ratio of the library DNA fragments to emPCR beads. Large volume emulsions were set up at the optimal ratio of 1.5 DNA copy per bead.

454 Sequencing: PCR Amplicon Preparation. Two pairs of PCR primers were designed per amplicon to span the VH and Vk/l regions of the scFv insert. One PCR primer pair contained the Titanium adaptor sequence with the sequencing primer binding site incorporated as part of the PCR forward primer. The second PCR primer pair contained the Titanium adaptor sequence with the sequencing primer binding

site incorporated as part of the PCR reverse primer (See Table S1 for oligonucleotide sequences). This design allowed for the bidirectional sequencing of the amplicon libraries using the Titanium chemistry and the currently available Titanium emPCR kit from Roche. The Titanium library adaptor sequences were kindly provided by Roche. Four independent PCR reactions were set up with each of the 2 PCR primer pairs for the VH and Vk/l amplicons using Platinum TaqDNA Polymerase (Invitrogen). The doublestranded DNA libraries were assessed for quality on the Agilent Bioanalyzer 2100 DNA 1000 chip and quantified using the Picogreen assay (Invitrogen). Small volume emPCR and titration by enrichment at 0.075, 0.15, 1, 2, 4 cpb was carried out according to the 454 Titanium emPCR protocol to determine the optimal cpb for each of the 4 amplicon libraries. Final small volume emulsions were set up at the optimal cpb of 0.10. The DNA beads from the 4 amplicon libraries were pooled before sequencing to enable bidirectional reads for the VH and Vk/l regions.

454 Sequencing: Sample Titration. Before emulsion PCR and sequencing, sample DNA size and concentration was determined by f luorometr y and capillar y electrophoresis to estimate optimum ratio of DNA molecules and beads for the emulsion PCR process. Since emPCR performance and yield of clonally amplified fragments depends on this ratio, the Roche/454 sequencing protocols recommend an empirical optimization using smallscale emPCR reactions and measuring yields of microparticles with positively amplified DNA by solid-phase capture of DNA beads by oligonucleotide hybridization (process called “enrichment”) (1). It has been empirically determined that enrichment yields around 10% of the starting number of microparticles used in an emPCR reaction indicate that majority of DNA beads will carry clonally amplified sequences (2). A 3-point titration was used for the shotgun library and an optimum ratio has been determined to be between 1–2 DNA copies per bead, resulting in 7–10% yield of enriched beads (R2 = 0.99). The amplicon libraries titrations were less linear (R2 = 0.65) and the titration range had to be extended over 6 points (0.075, 0.15, 0.5, 12, 4 cpb) to achieve enrichment yields around 10%. At 0.1 cpb, the final enrichment yields for amplicon libraries still varied between 10 and 20%.

454 Sequencing: Titanium Sequencing. The sequencing runs using the Roche/454 Genome Sequencer FLX were set up according to the 454 Titanium Sequencing protocol (3). One 4-region sequencing run was initially set up to evaluate the performance of the Titanium RCA shotgun and amplicon library sample preparations. Greater depth of coverage on these libraries were consequently obtained using a 2-region Titanium sequencing run.

454 Sequencing: Controls Performance, Loading Density, and Signal Intensities. The quality of 454 sequencing depends on multiple parameters optimized for each particular application (reviewed in 1). Among these parameters, the PicoTiter plate loading density and intensity of light signals produced in wells are among the most important, since these parameters affect well to well signal cross-talk. Since optimized protocols for sequencing amplicon libraries using the GS Titanium kits were not available at the time of these experiments, amplicon library samples were loaded onto the PicoTiter plate and sequenced in the same way as the shotgun library. Sequencing run performance was checked based on metrics generated for control DNA sequences, provided as a part of the Roche sequencing reagents. Sequencing performance of controls was about 20% lower in regions containing amplicon libraries compared to the shotgun library regions, measured as percentage Control wells sequencing at 98% accuracy over 200 bases. PicoTiter plates were loaded with equal amounts of DNA beads for both shotgun and amplicon libraries. Due to bead deposition variability, the region containing amplicon library had 14% more wells occupied by DNA beads. In addition, the average light signal intensity per nucleotide incorporation in the amplicon library sample was about 1.8 times stronger than the signals produced by the shotgun librar y. It is thus reasonable to conclude that the combination of loading density and signal intensity reduced quality of sequencing both for the control DNA templates as well as the amplicon library sample, resulting in more amplicon library reads being rejected by the 454 software read quality filters as described below.

454 Sequencing: Titanium Signal Processing and Run Yield. The 2 sequencing runs combined yielded 1,452,529 and 1,602,399 raw well reads for the shotgun and amplicon library, respectively. For both libraries, more than 99% (99.53% and 99.23%) of raw reads contained the 4-base key sequence. After the signal processing step of

the 454 data analysis pipeline, where reads may be rejected by multiple signal and quality filters, we obtained 923,876 (shotgun) and 554,310 (amplicon) filter-passing reads. The shotgun library sample had a ratio of filter-pass to key-pass reads of 63%, whereas the amplicon library sample produced only 35% filter-pass sequence reads of the key-pass reads. When examining which of the signal processing filters rejected most of the reads, we have observed that for the shotgun library sample, rejected reads were equally split between the "Dots and Mixed" filter and the "Quality Score Trimming" filter (18.18% and 18.48%), both metrics being below the recommended threshold of 20%. For the amplicon library sample, only 13.21% of reads were rejected by the "Dots and Mixed" filter, but 51.7% of reads failed the "Quality Score Trimming" filter. Despite this limitation of the GS data analysis software for our application, we have obtained 117Mb of high quality sequence for the amplicon library and 375Mb for the shotgun library.

Sequence Analysis: Germline Classification Benchmark. The reliability of germline assignment was evaluated by simulation of somatic hypermutation on germline v-segments. Somatic mutations were randomly applied to IMGT germline sequences, after which an attempt was made to then classify the derived sequence to nodes in an NJ reference tree built from the IMGT reference alignment. The distribution of somatic mutations in the actual pyrosequencing reads showed 95% of sequences with less than 30 apparent somatic mutations and >99.5% of sequences with less than 40 somatic mutations in the library as a whole (Fig. S2C). To thoroughly evaluate the space, the number of somatic mutations evaluated was 1 to 50, with each germline sequence retested 10 times at each level of somatic mutation (every functional human germline was tested 500 times in total). The simulation was run in a distributed cluster that included a 64 CPU XSERVE cluster and multiple XGRID-linked desktops using analysis libraries written in Perl. The assessment resulted in an extremely low rate of misclassification even when 50 somatic mutations were applied: in the combined datasets, 4 out of 152,500 VH simulations were misclassified and 14 out of 103,000 Vk/l simulations were misclassified(Fig. S2 A and B). This represents 84% correction over a naïve assessment, where the best BLAST hit is automatically accepted (33 Vh and 83 Vk/l sequences are misclassified in the naive

case). A reduction in resolution did occur: when above 30 mutations, heavy and light chains began to show an increased rate of lower resolution calls (assigning a sequence to an internal node instead of an explicit germline leaf node of the reference tree). As expected, this was disproportionately taking place for germlines that shared high sequence identity with another germline (less than 5 distinguishing polymorphisms). About 95% of germlines were still classified to the fullest resolution when even 50 mutations had been applied. It was concluded that all sequences along the observed distribution of somatic mutations in the actual pyrosequencing reads will be reliably classified with this approach. (Fig. S2C).

Sequence Analysis: HMM Validation. Direct mapping of Kabat numbering system to columns in the scFv HMM allow specific Kabat positions and ranges to be identified with high fidelity. The CDR boundary positions used were CDR-H1 Kabat 31–35, CDR-H2 Kabat 51–61, CDR-H3 Kabat 93–102; CDR-L1 Kabat 24 –34, CDR-L2 Kabat 50 –56, CDR-L3 Kabat 89 –97. Boundar y positions were used to identify CDRs for comparison. The ability of the HMM to automatically annotate CDR boundaries was assessed with 779 nonredundant PDB antibody structures from multiple species. The structures were superposed to a reference 1ck0 without information from the HMM. The HMM was then used to identify CDR stem boundary positions. Carbon-a backbone distance at those positions and the reference was obtained (Fig. S1A), and compared to the closest carbon-a from the chain. In all but 2 cases, the boundary positions were successfully labeled in every CDR in every reference. The 2 error cases only failed in one position: problems in setting the C-terminal CDR-H3 H102 boundary position. One case occurred with an insert placement in H3 offsetting residue by one position. The HMM correctly numbered the residues from position H103 onward. The second case was evaluated manually (Fig. S1B) and determined to be an unusual H3 conformation causing a proximity-based evaluation to fail, although the HMM alignment had annotated the H102 residue correctly. Ignoring this odd CDR, the error rate for the method was therefore 1 out of 1558 attempts. A smoothing function, restacking any CDR border inserts into the CDR, was applied during CDR identification.

1.2.5 Acknowledgements

I'm very greatful to my co-authors Wenwu Zhai, Jan Berka, Dilduz Telman, Gabriella Huerta, Gautam R. Mehta, Irene Ni, Li Mei, Purnima D. Sundar, Giles M. R. Day, David Cox, Arvind Rajpal, and Jaume Pons for making this work possible. We would like to acknowledge Lin T. Guey, Peter Henstock, Tenshang Joh, and Albert Seymour for their assistance in statistical analyses. We are appreciative of the helpful correspondence with Joshua Weinstein regarding his application of capture-recapture for antibody diversity assessment. We are also grateful to Andrea Rossi for reading the manuscript and sharing his insights on structural biology considerations.

1.2.6 References

1. Kindt TJ, Capra JD (1984) The Antibody Enigma (Plenum Press, New York).

2. Perelson AS, Oster GF (1979) Theoretical studies of clonal selection: Minimal antibody repertoire size and reliability of self-non-self discrimination. J Theor Biol 81:645–670.

3. Huber C, et al. (1993) The V kappa genes of the L regions and the repertoire of V kappa gene sequences in the human germ line. Eur J Immunol 23:2868–2875.

4. KawasakiK,etal.(1995)Theorganizationofthehumanimmunoglobulinlambda-gene locus. Genome Res 5:125–135.

5. Matsuda F, et al. (1998) The complete nucleotide sequence of the human immuno- globulin heavy chain variable region locus. J Exp Med 188:2151–2162.

6. Tonegawa S (1983) Somatic generation of antibody diversity. Nature 302:575–581.

7. Wu TT, Kabat EA (1970) An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody comple- mentarity. J Exp Med 132:211–250.

8. Trepel F (1974) Number and distribution of lymphocytes in man. A critical analysis. Klin Wochenschrift 52:511–515.

9. Margulies M, et al. (2005) Genome sequencing in microfabricated high-density pico- litre reactors. Nature 437:376–380.

10. Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA (2009) Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. Genome Res 19:1817– 1824.

11. Mavromatis K, et al. (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat Methods 4:495–500.

12. Gelfand I, Kister A, Kulikowski C, Stoyanov O (1998) Algorithmic determination of core positions in the VL and VH domains of immunoglobulin molecules. J Comput Biol 5:467–477.

13. Honegger A, Pluckthun A (2001) Yet another numbering scheme for immunoglobulin variable domains: An automatic modeling and analysis tool. J Mol Biol 309:657–670.

14. Lefranc MP, et al. (2003) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev Comp Immunol 27:55–77.

15. Abhinandan KR, Martin AC (2008) Analysis and improvements to Kabat and structur- ally correct numbering of antibody variable domains. Mol Immunol 45:3832–3839.

16. Gorski J, et al. (1994) Circulating T cell repertoire complexity in normal individuals and bone marrow recipients analyzed by CDR3 size spectratyping. Correlation with im- mune status. J Immunol 152:5109–5119.

17. Weinstein JA, Jiang N, White RA, III, Fisher DS, Quake SR (2009) High-throughput sequencing of the zebrafish antibody repertoire. Science 324:807–810.

18. Eddy SR (2000) Profile hidden markov models for biological sequence analysis. HMMER . Available at http://hmmer.janelia.org.

19. Eddy SR, Mitchison G, Durbin R (1995) Maximum discrimination hidden Markov models of sequence consensus. J Comput Biol 2:9–23

20. Karplus K, et al. (1997) Predicting protein structure using hidden Markov models. Proteins Suppl 1:134–139.

21. Park J, et al. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol 284:1201–1210.

22. Huse WD, et al. (1989). Generation of a large combinatorial library of the immuno- globulin repertoire in phage lambda. Science 246:1275–1281.

23. Barbas CF, III, Lerner RA (1991) Combinatorial immunoglobulin libraries on the surface of phage (phabs): Rapid selection of antigen-specific Fabs. Methods: Companion Methods Enzymol 2:119–124.

24. Griffiths AD, et al. (1993) Human anti-self antibodies with high specificity from phage display libraries. EMBO J 12:725–734.

25. Marks JD, et al. (1991) By-passing immunization. Human antibodies from V-gene libraries displayed on phage. J Mol Biol 222:581–597.

26. Hoet RM, et al. (2005) Generation of high-affinity human antibodies by combining donor-derived and synthetic complementarity-determining-region diversity. Nat Bio- technol 23:344–348.

27. Knappik A, et al. (2000) Fully synthetic human combinatorial antibody libraries (Hu- CAL) based on modular consensus frameworks and CDRs randomized with trinucle- otides. J Mol Biol 296:57–86.

28. Vaughan TJ, et al. (1996) Human antibodies with sub-nanomolar affinities isolated from a large non-immunized phage display library. Nat Biotechnol 14:309–314.

29. Klein U, Kuppers R, Rajewsky K (1997) Evidence for a large compartment of IgM- expressing memory B cells in humans. Blood 89:1288–1298.

30. Weller S, et al. (2004) Human blood IgM "memory" B cells are circulating splenic marginal zone B cells harboring a prediversified immunoglobulin repertoire. Blood 104:3647–3654.

31. Weller S, et al. (2008) Somatic diversification in the absence of antigen-driven re- sponses is the hallmark of the IgM IgD CD27 B cell repertoire in infants. J Exp Med 205:1331–1342.

32. Marks JD, Bradbury A (2004) PCR cloning of human immunoglobulin genes. Methods Mol Biol 248:117–134.

33. Marks JD, Bradbury A (2004) Selection of human antibodies from phage display libraries. Methods Mol Biol 248:161–176.

34. Zemlin M, et al. (2003) Expressed murine and human CDR-H3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. J Mol Biol 334:733–749.

35. LefrancMP,etal.(2009)IMGT,theinternationalImMunoGeneTicsinformationsystem. Nucleic Acids Res 37:D1006–D1012.

36. Altschul SF, et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25:3389.

1.2.7 Copyright

1.3 T-cell receptor repertoire sequencing

This study saw the developement of standardized set of optimized TCR sequencing primers that were DNA multiplex compatible for sequencing up to 184 samples in a MiSeq sequencing run, and could be adapted for greater depth.

T-cell receptor (TCR) diversity, a prerequisite for immune system recognition of the universe of foreign antigens, is generated in the first two decades of life in the thymus and then persists to an unknown extent through life via homeostatic proliferation of naïve T cells. We have used next-generation sequencing and nonparametric statistical analysis to estimate a lower bound for the total number of different TCR beta (TCRB) sequences in human repertoires. We arrived at surprisingly high minimal estimates of 100 million unique TCRB sequences in naïve CD4 and CD8 T-cell repertoires of young adults. Naïve repertoire richness modestly declined twoto fivefold in healthy elderly. Repertoire richness contraction with age was even less pronounced for

memory CD4 and CD8 T cells. In contrast, age had a major impact on the inequality of clonal sizes, as estimated by a modified Gini–Simpson index clonality score. In particular, large naïve T-cell clones that were distinct from memory clones were found in the repertoires of elderly individuals, indicating uneven homeostatic proliferation without development of a memory cell phenotype. Our results suggest that a highly diverse repertoire is maintained despite thymic involution; however, peripheral fitness selection of T cells leads to repertoire perturbations that can influence the immune response in the elderly.

A decline in the diversity of the T-cell receptor repertoire owing to thymic involution has been implicated as causing defective immune responses in the elderly. By applying next-generation sequencing of replicate TCRB libraries from highly purified T-cell subsets, and using nonparametric statistical analysis, we obtain estimates of repertoire richness in the young adult that are higher than previously reported. Although contracting with age, the repertoire remains highly diverse. These data challenge the paradigm that thymic rejuvenation is needed to maintain diversity and prevent immune incompetence in the elderly. However, we observe an increasing inequality of clonal sizes with age even among naïve T cells. This clonal selection could result in biased and possibly autoreactive immune responses.

1.3.1 Introduction

The ability of the adaptive immune system to respond to a wide variety of pathogens depends on a large repertoire of unique T-cell receptors (TCRs). TCR diversity is generated by the random and imprecise rearrangements of the V and J segments of the TCR alpha (TCRA) and V, D, and J segments of the TCR beta (TCRB) genes in the thymus. Thymic production of T cells is the sole mechanism to generate TCR diversity. With the involution of the thymus with age, the generation of new naïve T cells with new TCRs dwindles (1, 2). In the mouse, residual thymic activity provides an ongoing source of naïve T cells, although not sufficient to maintain the compartment size. In contrast, homeostatic proliferation of peripheral T cells is the predominant mechanism of T-cell generation in the human adult (3). Given the dramatic decline

in thymic T-cell production, thymic involution has been implicated in defective T-cell responses with aging (4). Elderly individuals are more prone to develop complications from infections; in particular, they are susceptible to newly arising infectious organisms such as West Nile fever or severe acute respiratory syndrome (5, 6). Moreover, immune responses to vaccination are dramatically reduced (7). Shrinkage in the size of the naïve T-cell compartment and holes in the repertoire have been discussed as potential causative factors in age-related impaired adaptive immunity (8). Indeed, holes have been identified in the murine repertoire to contribute to the increased morbidity from influenza infection (9, 10).

Pioneering early work on estimates of diversity of the human TCR repertoire was based on sequencing a few hundred sequences and extrapolating to the scale of the entire repertoire, yielding an estimated lower limit estimate of fewer than 1 million different TCRB genes (11). Studies in more recent years have used highthroughput sequencing to base estimates on greater sequencing depth. Warren et al. (12) measured 1 million different TCRB sequences in a peripheral blood sample. Robins et al. (13) inferred the number of unseen TCRB genes from deep sequencing data by applying a Poisson process model and estimated a diversity of 3–4 million TCRB sequences in the total T-cell populations of two healthy donors.

Although high-throughput DNA sequencing provides the tool to gather extensive datasets, estimates of TCR richness remain a challenge because clinical samples represent a small fraction of the 10^11 cells in the overall T-cell compartment. Richness, the actual number of unique TCR sequences in a T-cell population, is dominated by the infrequent species that are observed rarely or not at all. The major factors affecting accurate identification of rare species in TCR repertoires are the number of cells examined, the variable efficiency of PCR amplification, the expression levels of TCR transcripts in different T cells, and sequencing error rates. In addition, peripheral blood T cells consist of a mixed population of naïve, memory, and effector CD4 and CD8 cells that differ markedly in diversity, with memory populations being less complex than naïve populations.

The current study was designed explicitly to improve estimation of the TCR repertoire richness of stringently defined naïve and memory T cells. By sequencing multiple

replicate TCRB libraries of cells from each T-cell subset and applying nonparametric statistical analysis, we find that human TCR repertoires are an order of magnitude more diverse than previously estimated. Despite significant age-related decreases in richness, humans maintain high diversity during healthy aging. Strikingly, we find age-associated changes in the distribution of clonal sizes, in particular in the naïve compartment, which may reflect unevenness of homeostatic expansion with clonal expansion of some naïve T cells that equal or exceed clonal sizes of memory T cells. The inequalities in clonal sizes could compromise the immune response to the majority of antigenic epitopes while causing an increased responsiveness to selected few epitopes.

1.3.2 Results

Gene Segment Use and CDR3 Features in Young and Elderly TCRB Rearrangements. In initially evaluating TCRB repertoires in young and elderly subjects we compared the composition of TCRB gene rearrangements at the level of TCRBV and TCRBJ gene segment use and the features of the CDR3-encoding junctional nucleotides. Apheresis lymphocytes were obtained from four 20to 35and five 70to 85-y-old healthy adults who were regular platelet donors. Naïve CD4 and CD8 T cells were purified by cell sorting. We used a stringent definition of naïve cells (CD3+CD4+ or CD8+CCR7+CD45RAhighCD28+) and very restrictive gating to ensure purity. Approximately 1.5–3 million sequence reads were obtained for each T-cell subset of each donor (Table S1). TCRBV and TCRBJ gene segments were used at comparable frequencies in the repertoires of naïve CD4 and CD8 T cells in young and old individuals (Fig. S1A), whereas the memory repertoires of CD4 and, particularly, CD8 T cells show variable and individual-specific gene segment use frequencies (Fig. S1B). CDR3 sequences in the young and elderly were comparable in length and did not show any definitive age-related features (Fig. S1 C and D).

High Richness of Naïve CD4 and CD8 TCRB Repertoires in Young and Elderly Adults. To obtain sequence data to estimate global TCRB repertoire richness we used the experimental design of analyzing a series of replicate libraries from independent cell aliquots from each T-cell subset in each individual. These replicates allowed us

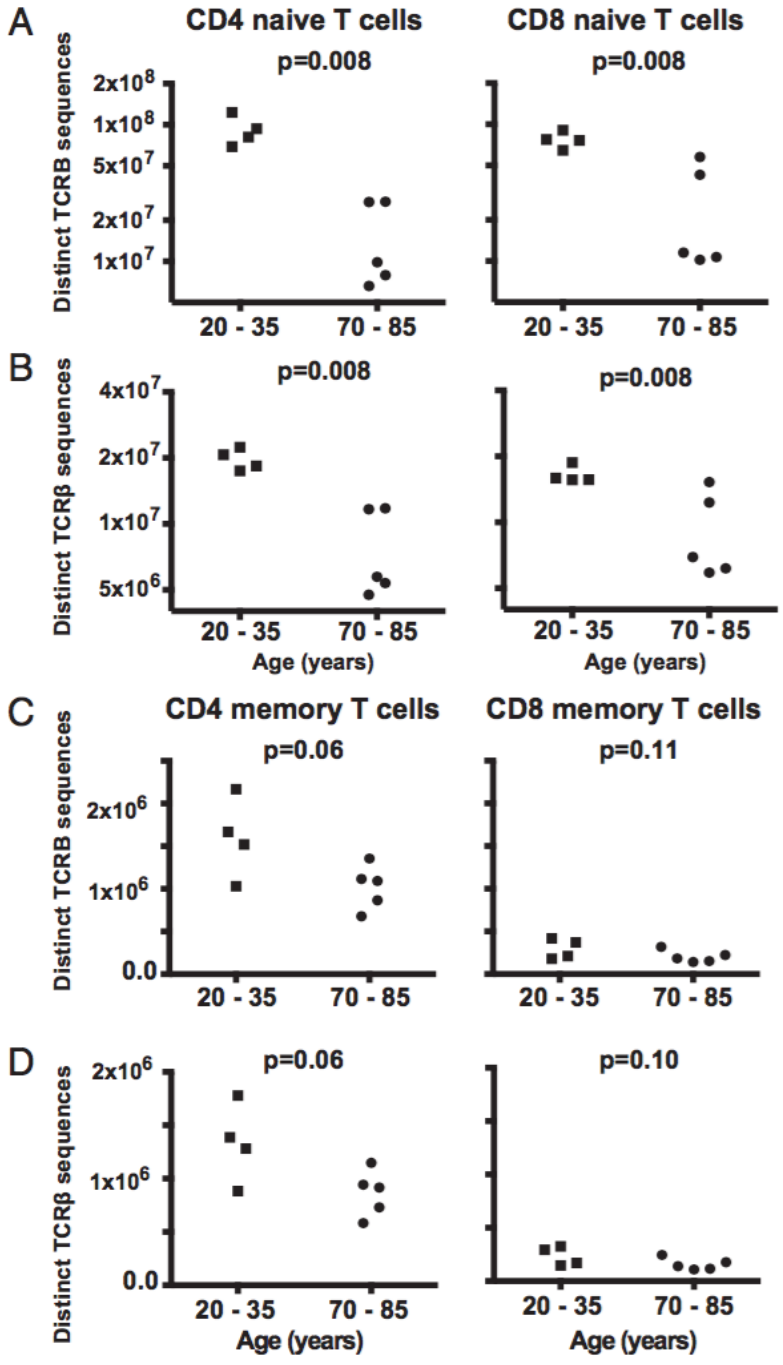


Figure 1.5: Age is associated with a modest decrease in diversity of the TCRB repertoire. TCRB sequences were obtained from replicate samples of naïve (A and B) and memory (C and D) CD4 and CD8 T cells. A lower bound of TCRB richness was estimated by applying nonparametric statistics using the Chao2 estimator. Results are shown for nucleotide (A and C) and derived amino acid sequences (B and D). Estimates were compared by Wilcoxon–Mann–Whitney test. Increase in age is associated with a decline in richness of naïve CD4 and CD8 T cells; however, the repertoire in the elderly remains highly diverse. Richness in CD4 and CD8 memory T cells markedly differed, whereas the impact of age was negligibly small.

to calculate repertoire richness by applying the "Chao2" estimator, a nonparametric estimator of unseen species (14). The approach allows estimation of the extent to which the full repertoire is covered and use of this information to determine a lower bound of the total number of species in the repertoire. Because the Chao2 estimator requires only a binary characterization of presence or absence of each clone in each replicate library, it circumvents the challenges that arise in experimental designs using only a single library and avoids confusing the effects of PCR amplification with the presence of expanded T-cell clones. To not count possible sequencing errors as independent sequences, we rejected single sequences as erroneous if a highly similar clone of greater frequency was identified in the same library (see Materials and Methods for the definition of similarity).

The lower bounds on TCRB gene richness obtained with this approach yielded higher estimates than previous studies (Fig. 1). Young adults carried an estimated 60–120 million different TCRB genes, both in the CD4 and CD8 naïve T-cell repertoires. This high diversity in nucleotide sequences was reflected in a large functional repertoire of TCR β chains with a lower boundary of ∼20 million different amino acid sequences. To determine the robustness of our estimates, we used two approaches to estimate confidence intervals. We applied the BCa variant of bootstrapping that is designed for obtaining confidence intervals when the underlying bootstrap distribution is not symmetric about its center (15). Second, we estimated the confidence intervals using the approach originally developed by Chao (16). The 95% confidence intervals with both methods were very narrow (Table S2).

Naïve TCRB repertoire richness declined significantly in the 70to 85-y-old adults to a lower bound richness of 8–57 million different nucleotide sequences encoding ∼5–15 million TCR β-chain amino acid sequences ($P = 0.008$, Fig. 1 A and B). Interestingly, the estimates in elderly CD4 and CD8 naïve T cells were similar despite the greater decline in CD8 compared with CD4 naïve T-cell numbers with aging (17, 18).

CD4 and CD8 Memory T Cells Differ in TCRB Richness Independent of Age. Memory cells have been selected from the naïve repertoire and clonally expanded in

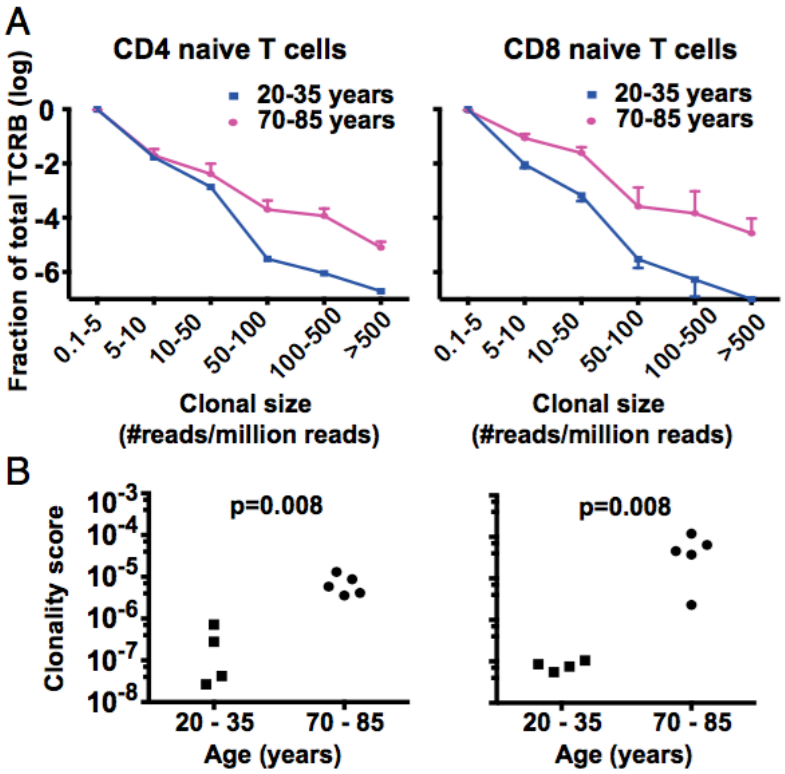


Figure 1.6: Increase in clonal expansions within naïve CD4 and CD8 T-cell compartments with age. (A) The mean number of replicate TCRB sequences was used as an estimate of approximate clonal sizes. The frequency distribution of clonal size bins is shown as (log mean) ± SD of the four young and the five elderly adults. Nondetectable clonal sizes were set at a frequency of 1 in 107. (B) Replicate samples of naïve CD4 or naïve CD8 T cells were analyzed for the shared occurrence of TCRB sequences to estimate a clonality score, defined as the probability of two independently identified sequences originating from the same clone. Clonality scores were compared by Wilcoxon–Mann–Whitney test. The observed increase indicate inequality in clonal sizes in the elderly naïve T-cell repertoire with age with increasing number of large clones.

response to antigen and are therefore expected to have a lower richness. With advancing age, the number of memory T cells increases and end-differentiated effector CD8 T cells responsive to latent viruses, in particular CMV, accumulate (19, 20). In initial experiments, we compared richness in CD8 central, effector memory, and end-differentiated effector T cells in elderly individuals who did or did not have antibodies to CMV. Richness of effector cell populations was lower than that of central memory T cells and was further compromised in CMV-positive individuals (Fig. S2). To examine the effect of age independent of CMV-induced repertoire changes, we excluded individuals who had positive CMV serology and excluded terminally differentiated effector cells that are known to include clonally expanded effector T cells to latent viruses. Richness in the CD4 memory compartment was about 50-fold lower than in the naïve repertoire, with a lower bound of about 1 million different TCRB genes in each individual. Surprisingly, richness showed negligible contraction with age (Fig. 1 C and D). CD8 memory T cells were 5to 10-fold less diverse than CD4 memory T cells, irrespective of age.

Age-Dependent Clonal Expansion Within the Naïve T-Cell Compartment. In addition to richness, distributions of clonal size could affect the functionality of the T-cell repertoire. Owing to limitations in sampling ($5 \times 10\hat{}6$ naïve T cells out of a total repertoire of up to $10\hat{}11$ cells per individual), estimates of clonal size distributions are only reliable for clones above a clonal size threshold. In the young as well as the old, most TCR sequences are derived from relatively small clones. In Fig. 2A, the frequencies of clones are plotted against clonal sizes (defined as the number of occurrences of identical sequences in the five combined samples of each subset). The frequency distributions for clones above the detection threshold in the repertoires of young and elderly individuals clearly differed, with a small number of naïve T-cell clones showing a striking expansion with age. To compare the contributions of expanded clones to the repertoires, we used a "clonality score" summary metric that is independent of sequencing depth. This metric can be thought of as the probability that two sequence reads selected at random from different replicate library pools will be members of the same clone. Clonality scores in the elderly are approximately

>100-fold higher for naïve CD8 T cells and >10-fold higher in CD4 T cells compared with younger individuals (Fig. 2B).

Recent studies have identified public TCR sequences that are shared by unrelated individuals (21, 22). Typically, public TCR sequences are those with simple junctions, such as those that lack N-region nontemplated nucleotides (22). To exclude that sequences detected more than once represented coincidentally identical TCR rearrangements rather than clonally expanded T cells, we analyzed sequences present at frequencies of greater than 0.01% in elderly individuals. A TCR sequence was called as public when found in more than one of the nine individuals studied. Twenty-three of the 483 naïve CD4 and 81 of the 878 naïve CD8 frequent TCRB sequences in old individuals were found in more than one individual (Table 1). In contrast, a higher number of 92 of the 483 most frequent CD4 ($P < 0.001$) and 129 of 878 most frequent CD8 TCRB sequences ($P < 0.05$) in the naïve repertoire of young individuals are shared in different individuals. The data indicate that the observed increased clonality in the elderly individuals represent true clonal expansions, whereas many of the apparently clonally expanded sequences in the young repertoires may reflect the presence of simple and public TCR rearrangements.

Effect of Age on Clonality in the Memory T-Cell Compartment. In contrast to the age-associated increases in clonality that we observed in naïve CD4 and CD8 cell compartments, there was little effect of age on the clonality of memory CD4 or CD8 T-cell populations. By virtue of memory T cells being present in larger clonal sizes than naïve T cells, clonal size distributions could be examined for a larger fraction of the repertoire. Distributions of clonal sizes were mostly identical for young and elderly individuals and an age-related increase in clonal sizes was seen only for the largest clones (Fig. 3A). Accordingly, clonality scores only slightly increased with age (Fig. 3B). Notably, CD8 exceeded CD4 memory T cells in clonal sizes.

Clonally Expanded Naïve T Cells Express TCRB Sequences Distinct from Memory T Cells. Effector T cells can revert to a phenotype sharing cell surface markers with naïve cells (23). To examine the possibility that the clonal expansions within the naïve compartment represented a contamination with effector T cells masquerading as naïve T cells, we analyzed all TCRB sequences that were found at more than 0.01%

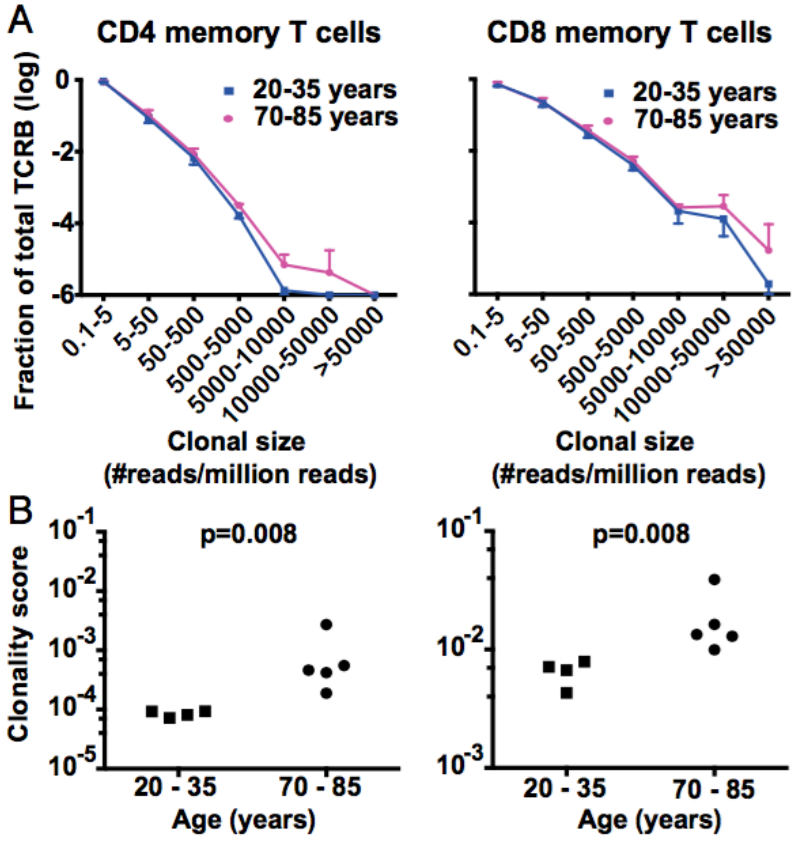


Figure 1.7: Greater clonality in the repertoire of CD8 than CD4 memory T cells. Frequency distributions of categorized clonal size bins (A) and clonality scores (B) for memory CD4 and CD8 T cells were determined as described in Fig. 2 for naïve T cells. Nondetectable clonal sizes were set at the minimal frequency of 1 in 106 (A). Clonality scores were compared by Wilcoxon–Mann– Whitney test. Large clonal expansions were more frequent in CD8 than in CD4 T cells with only minor influence of age on clonality in both compartments.

frequency in the naïve T-cell population for their representation in the memory population. The reverse analysis was done for clonally expanded sequences in the memory compartments. Results suggested that clonal expansions in the naïve compartment of elderly individuals were not contaminating memory cells. For example, of 121 TCRB sequences clonally expanded in the CD4 naïve T-cell compartment of one individual, 34 were unique for naïve T cells. Conversely, of 920 CD4 memory TCRB sequences, 449 were unique for memory cells. A similar distribution was seen for CD8 T cells: 215 of 379 clones originally identified in the CD8 naïve compartment were uniquely found in that compartment, whereas 349 of 846 memory sequences were not seen at all in naïve CD8 T cells. To further examine whether even the shared clonally expanded populations can be predominantly assigned to one compartment, the frequency of each of these shared sequences in each of the five naïve and five memory CD4 and CD8 T-cell replicate samples was determined. The heat plots in Fig. 4 show the frequency distributions of these clonally expanded CD4 and CD8 TCRB sequences for the different naïve and memory cell samples from two elderly adults. Each row represents one sequence and each lane an independent library of naïve or memory T cells. Although sequences were selected to be present in naïve and memory T cells, very few sequences were found at similar frequencies in both compartments. Importantly, TCRB sequences originally identified in the naïve compartment were virtually absent in memory cell pools, suggesting that these clonal expansions originated from naïve cells and maintained their naïve phenotype.

Mechanisms of Clonal Expansion Within the Naïve T-Cell Compartment. A likely interpretation of the increasing inequality of clonal sizes with age is that the homeostatic proliferation of naïve T cells is associated with fitness selection. Clonally expanded T cells might have a competitive growth advantage that in part could be conferred by increased responsiveness to homeostatic cytokines. To test this hypothesis, naïve CD8 T cells labeled with carboxyfluorescein diacetate succinimidyl ester were cultured in the presence of IL-7 and IL-15. Cells that had undergone one or more or two or more divisions after 7 d were purified and sequenced. TCRB sequences from cultured T cells were analyzed for their presence in the replicate libraries of unstimulated naïve T cells. T cells that divided in culture in response to cytokine stimulation

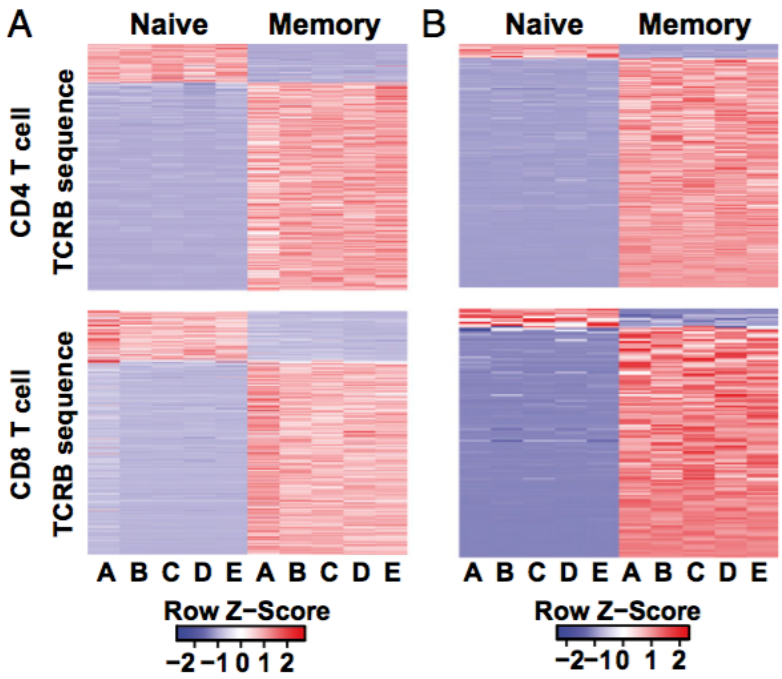


Figure 1.8: Characterization of clonally expanded naive T cells. TCRB sequences that had frequencies of $\geq$0.01% and were found in the naïve as well as the memory compartment were analyzed for their frequencies in five independent naïve and memory replicate samples (samples A–E). Representative results of two elderly individuals (one in A and the other in B) are shown as heat maps using the color scheme from infrequent (blue) to frequent (red). The data indicate that T cells expanded in the naïve and memory compartments are different.

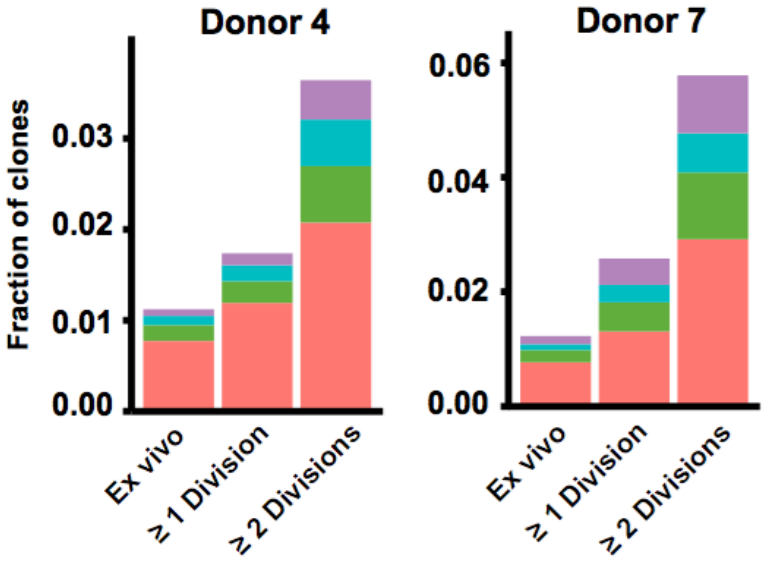


Figure 1.9: Increased responsiveness of in vivo expanded naïve CD8 T cells to cytokineinduced proliferation. Naïve CD8 T cells were cultured with IL-7 and IL-15. TCRB sequences from naïve CD8 T cells that had divided equal to or more than once or twice were compared with sequences present in the peripheral blood repertoire of the individual. Results from two individuals are shown. The proportions of sequences from the cultured cells that were members of clones detected in four (purple), three (blue), two (green), or one (red) replicates of the original noncultured T-cell libraries from the peripheral blood are represented as cumulative bar graphs. The fastest-proliferating cultured T cells (right column) show enrichment of large clones found in four replicate libraries from the blood ($P < 0.001$).

show an increased proportion of clones that were expanded in the peripheral blood (Fig. 5). The enrichment was more pronounced for T cells that divided more rapidly. In the population that had divided two or more times, large clones (defined as being found in four replicates of the original blood sample) are enriched compared with small clones (defined as present in one replicate) ($P < 0.001$). In contrast, enrichment was not obvious when T cells purified for high expression of the IL-7 and IL-15 cytokine receptors CD127 and CD215, respectively, were analyzed.

1.3.3 Discussion

In this study we combined next-generation sequencing with a nonparametric statistical approach using the Chao2 estimator to estimate a lower bound for TCR richness. We found a higher richness in CD4 and CD8 naïve T cells than previous studies. Even though diversity contracts with age, we find that elderly individuals still possess a diverse T-cell repertoire. However, we observe robust clonal expansion with age in the naïve compartments, suggesting that homeostatic proliferation is associated with fitness selection. Finally, we found lower richness in CD8 than in CD4 memory cells, a difference that was preserved during aging. Thymic involution is the most dramatic age-related change in the human immune system. Understanding whether the T-cell repertoire can be maintained in the absence of thymic activity and whether repertoire contraction contributes to the immune defects in the elderly is critical for designing possible interventions. Conclusions for the human repertoire from animal models are unreliable because the size of the T-cell compartment and mechanisms and kinetics of T-cell homeostasis are fundamentally different in humans and mice (3).

Whether thymic T-cell generation in humans is of any quantitative importance for the steady state of T-cell populations after the end of the growth period has been controversial. Some residual thymic tissue persists in elderly humans (24); however, even in the lymphopenic host after chemotherapy or bone marrow transplantation or in HIV patients after initiation of highly active antiretroviral therapy, resurgence of thymic activity does not occur in the majority of individuals older than 40–50 y (25). In the healthy adult, TCR excision circle (TREC) byproducts of TCR rearrangement that do not replicate are the best surrogate marker for thymic activity. Mathematical modeling has suggested that the age-related decline in TRECs is best explained by cell loss compensated by homeostatic naïve T-cell turnover rather than by declining thymic efflux. Thymic activity may not play a significant role in the maintenance of TCR diversity in adult life (26).

High-throughput DNA sequencing now enables extensive measurement of TCR populations, but estimation of TCR repertoire richness from sequencing of blood samples has remained a challenge. Here, we combined next-generation sequencing with a nonparametric statistical approach to estimate richness. The use of multiple

replicate sequencing libraries generated from independent samples of T cells for each subset enabled us to identify the rare sequences that are critical for estimates of the number of unobserved species and overall TCR richness in the repertoire. We corrected for sequencing errors by excluding sequences that were related to detected peaks in sequence space. In contrast to previously used corrections, our approach does not eliminate essentially all rare clones, but it does still give a conservative appraisal of diversity.

Based on our estimate of 100 million TCRB gene rearrangements and considering that each TCR β-chain in the naïve T-cell repertoire combines with about 25 TCR α-chains (11), the average clonal size of a naïve cell would be about 100–200 cells. This approximation is of the same magnitude as predicted from the dilution of TREC-bearing T cells in the naïve repertoire. In the neonate, naïve T cells proliferate to fill the compartment (27). Recent thymic emigrants in the newborn undergo three to four divisions in the periphery after TCRA gene rearrangement, thereby establishing a minimal clonal size of 10 cells (28). TREC frequencies decline further in subsequent years, suggesting that naïve T cells undergo approximately three more divisions during the growth period in the presence of thymic activity, to give a final average naïve T-cell clonal size of about 100 cells.

Our analysis does not address directly whether a decline from about 20 million to 10 million TCR β-chain sequences leads to immune defects or holes in the repertoire in the elderly. Tetramer studies in noninfected individuals have suggested that the frequency of T cells specific for many exogenous viral peptides, or self-peptides, is on the order of 1 in 1 million (29), suggesting that the estimated elderly repertoires of more than 100 million TCR α-β dimers could be diverse enough to recognize most peptides bound to a given MHC molecule and avoid frank "holes" in the repertoire.

A striking age-related change in our studies was an increase in T-cell clonality in the naïve repertoire. The contribution of clonally expanded T cells to the observed repertoire increased by a factor of >100 for naïve CD8 and of >10 for naïve CD4 T cells compared with young adults. Analysis of these expanded clones showed that the clones in the naïve subsets were distinct from those in the memory compartments, although many of them were present in clonal sizes that were comparable to those

of memory cells. If these naïve clones derive from uneven homeostatic proliferation, they maintain a naïve phenotype, in contrast to findings in mice, in which naïve T cells undergoing homeostatic proliferation tend to acquire a memory-like phenotype (30, 31).

Based on the limited number of cells that can be sampled in humans, our clonality index only allows conclusions on the impact of clones that are above a certain size. However, for naïve T cells age-associated increases were observed for the entire range of clonal sizes that could be assessed. This observation is consistent with the interpretation that homeostatic proliferation during aging is associated with increasing unevenness in clonal size distributions of the entire repertoire owing to peripheral selection. In support of this interpretation, we have recently performed an agent-based stochastic in silico simulation of repertoire complexity under homeostatic proliferation and have found only a minimal contraction in diversity over a human lifetime, as long as peripheral fitness selection during homeostatic proliferation can be ignored; even with no thymic activity after age 20 y and 90% shrinkage of the naïve T-cell compartment size (32). Depending on the initial clonal size, increasing variance in clonal abundances will not necessarily lead to clonal extinction. In our data, naïve CD8 T cells had a much higher clonality score than naïve CD4 T cells; however, richness in both compartments was very similar, suggesting a higher variance in clonal sizes in CD8 T cells (Figs. 1 and 2). A large number of homeostatic T-cell divisions or, as explored in our in silico simulation, a rather dramatic co-occurrence of selective forces is required before homeostatic proliferation leads to contraction in richness. However, the inequalities in clonal sizes with some large and many very small clones may generate a less complex T-cell response to peptide antigens, which may be more important than the fewer TCRs available in the repertoires of older compared with younger individuals. Such an unevenness cannot be improved by restoring thymic T-cell generation (32).

Age-associated skewing in clonal size distributions explains why some prior studies have observed a higher decline in diversity with aging. Because sequencing depths in most previous studies were not sufficient and/or the analytical design did not take into account increasing inequality in clonal size distributions, estimates of richness were

not reliable. Based on this model, our previous report of a stable TCR repertoire up to the seventh decade of life with a sudden change thereafter indicates an occurrence of increasing T-cell clonal expansions late in life (33). This interpretation is also consistent with observations in a nonhuman primate model (34).

Homeostatic proliferation and survival of naïve T cells are thought to require the recognition of self-antigen in the presence of homeostatic cytokines (30, 35). Inequalities in clonal sizes could therefore reflect a peripheral selection based on cytokine responsiveness and/or self-recognition. In support of this interpretation, in vitro culture of naïve T cells in the presence of homeostatic cytokines, but absence of foreign antigens, selected for TCR sequences clonally expanded in vivo. Homeostatic proliferation associated with a peripheral selection may therefore result in a more autoreactive repertoire (36). This may explain why lymphopenia-induced homeostatic proliferation confers an increased risk for autoimmunity (37, 38) and supports the model that increased homeostatic proliferation and associated changes in the repertoire found in patients with rheumatoid arthritis may predispose for the disease (39–41).

Surprisingly, and in contrast to naïve T cells, the impact of age on the memory T-cell repertoire was minimal. Although the size of the memory compartment is known to increase with age, richness did not change significantly, and clonality increased only for the very largest clones. This observation is particularly surprising for CD8 T cells where the terminally differentiated effector T-cell population tends to increase at the expense of central and effector memory T cells (17). It should be noted, however, that we have excluded individuals who were CMVpositive to analyze age effects without the confounding factor of CMV-induced repertoire changes.

Our data highlight striking differences in the repertoires of CD4 and CD8 T memory cells, independent of age. Richness was higher in CD4 memory than in CD8 T cells, whereas clonality was higher in CD8 memory T cells. The prominent clonality in CD8 memory T cells is consistent with previous spectratyping studies demonstrating clonal peaks in the CD8 memory repertoire (42). Our studies now show that the difference between human CD4 and CD8 memory T cells is not limited to clonally expanded CD8 T cells, but also includes a globally decreased richness of the entire CD8 memory repertoire. Given that richnesses in naïve CD4 and CD8 T cells are

approximately equal, our data suggest that formation or maintenance of the memory repertoire is more constrained for CD8 than for CD4 T cells. Future studies of the repertoires of antigen-specific T cells after vaccination or infection in humans will be required to further explore the consequences of T-cell subset-specific repertoire contractions and the contribution of clonal expansions to the increased vulnerability of the elderly to common pathogens.

1.3.4 Methods

TCRB cDNA libraries were generated from five replicate samples of FACSsorted naïve and memory CD4 and CD8 T cells from apheresis samples of young and elderly healthy adults and sequenced with an Illumina Miseq sequencer. TCRB repertoire richness was determined by applying the Chao2 nonparametric estimator of the lower bound of species richness in a population after correcting for possible sequencing errors and eliminating TCRB sequences that are close to peaks in sequence space. A clonality score, adapted from the Gini–Simpson index, was determined using the lymphclon inference algorithm. A detailed description of the experimental design and procedures and the statistical analysis is given in SI Materials and Methods. Primers are described in Table S3.

Study Subjects. Platelet donor apheresis lymphocytes were obtained from young (aged 20–35 y) and elderly (aged 70–85 y) adults from the Stanford Blood Center. All individuals were healthy, regular platelet donors without a history of autoimmune disease, diabetes mellitus, or chemotherapy and were analyzed for their CMV serology. Informed consent for research use of blood samples was provided by all participants.

Cell Purification and RNA Extraction. T cells were enriched by negative selection using the human T-cell RosetteSep enrichment kit (StemCell Technologies) and then stained with FITC-CD3(BioLegend), V450-CD4-(Affymetrix eBioscience), PECy7-CD8-, PerCP/Cy5.5-CCR7(BioLegend), PE-CD28and APC-CD45RA(BD Biosciences) antibodies as well as Live/Dead Fixable Aqua Dead cell stain dye (Life Technologies), followed by cell sorting using a BD Aria3 cell sorter to obtain naïve CD4 T cells (CD3+CD4+ CCR7+CD45RAhighCD28+), naïve CD8 T

cells (CD3+CD8+ CCR7+CD45RAhighCD28+), memory CD4 T cells (CD3+CD4+ CD45RA–), and memory CD8 T cells (CD3+CD8+ CD45RA–). For young individuals, five replicates with 1 × 106 cells per aliquot of naïve CD4, naïve CD8, and memory CD4 T cells and five replicates with 0.25 × 106 per aliquot of memory CD8 T cells were collected. For naïve CD8 T cells in elderly individuals, five replicates of 0.25 × 106 cells were collected; cell numbers for all other populations were the same as in young individuals. In additional experiments, central memory (CD3+ CD8+ CCR7+ CD45RA–), effector memory (CD3+CD8+CCR7–CD45RA–), and terminal effector memory (CD3+CD8+CCR7–CD45RAhighCD28–) CD8 T cells from CMV-positive and CMV-negative elderly individuals were sorted and five replicates with 0.2 × 106 cells per aliquot of each CD8 T-cell population were collected. In two elderly individuals, CD4+CD25highCD127low (Treg) and CD127+ and CD215+ naïve CCR7+CD45RAhighCD28+ CD8+ were purified for analysis. RNA was extracted using AllPrep DNA/RNA mini kit (Qiagen). cDNA was generated using SuperScript III reverse transcription kit (Invitrogen) with random hexamer oligonucleotides.

Cytokine-Mediated in Vitro T-Cell Expansion. T cells were stained with V450-CD4-, PECy7-CD8-, APC-CD45RA-, and PerCP/ Cy5.5-CCR7antibodies, followed by cell sorting to obtain naïve CD8 T cells (CD3+ CD8+ CCR7+ CD45RAhigh). Cells were labeled with carboxyfluorescein diacetate succinimidyl ester (CFSE) and cultured in vitro in the presence of 10 ng/mL IL-15 and 10 ng/mL IL-7 (Peprotech). After 7 d of culture, T cells were stained with V450-CD4, APC-CD3, and PECy7-CD8, followed by cell sorting to obtain T cells (CFSElowCD8+ CD3+) that had divided in response to cytokine stimulation. RNA was extracted using RNA micro kit (Qiagen). cDNA was generated using SuperScript VILO MasterMix (Invitrogen).

Generation of T-Cell Receptor Beta Gene Libraries. To amplify rearranged T-cell receptor beta (TCRB) genes from cDNA, 39 forward primers, each specific to one or two functional TCRBV segments, and one reverse primer specific to the CB segments were designed based on the repertoire of TCR germ-line segments present in the ImMunoGeneTics (IMGT) database (www. imgt.org) (Table S3). A 12-bp nucleotide unique "barcode" sequence was inserted in the reverse primer between the CB sequence and Illumina universal reverse sequence. First PCR amplification was

TRBV10-1-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaAAGTCTCAGATGGCTACAGTGTCTCTAG
TRBV10-2-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaTCCAGATCCAAGACAGAGAATTTCCCC
TRBV10-3-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaGAAGTCTCAGATGGCTATAGTGTCTCTAG
TRBV12-5-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaGGGATGCCGAAGGATCGATTC
TRBV13-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaAGCTCAACAGTTCAGTGACTATCATTCTG
TRBV14-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaTTAGCTGAAAGGACTGGAGGGACG
TRBV16-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaATTTTCAGCTAAGTGCCTCCCAAATTCAC
TRBV18-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaTGCTGAATTTCCCAAAGAGGGCC
TRBV19-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaAGCGTCTCTCGGGAGAAGAAGG
TRBV2-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaAATTCTCAGTTGAAAGGCCTGATGGATC
TRBV20-1-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaATGCAAGCCTGACCTTGTCCAC
TRBV24-1-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaAGTGTCTCTCGACAGGCACAGG
TRBV25-1-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaACAGTCTCCAGAATAAGGACGGAGC
TRBV27-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaTGAAGGGTACAAAGTCTCTCGAAAAGAG
TRBV28-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaGATATTCCTGAGGGGTACAGTGTCTC
TRBV6-1-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaAATGTCTCCAGATTAAACAAACGGGAGTTC
TRBV6-3-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaTCCCTGATGGCTACAATGTCTCCAG
TRBV6-4-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaTATAGTGTCTCCAGAGCAAACACAGATG
TRBV6-8-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaTAGATTAAACACAGAGGATTTCCCACTCAG
TRBV6-9-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaACAATGTATCCAGATCAAACACAGAGG
TRBV7-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaAGAGAGGCCTGAGAGATCCGTCTC
TRBV9-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaATTCTTGAACGATTCTCCGCACAACAG
TRVB11-1.11-2.11-3-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaCAGTTGCCTAAGGATCGATTTTCTGC
TRVB12-3.12-4-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaGGGATGCCCGAGGATCGATTC
TRVB15-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaACCCCTGATAACTTCCAATCCAGGAG
TRVB29-1-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaATCAGCCGCCCAAACCTAACATTC
TRVB3-1-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaTTCTCACCTAAATCTCCAGACAAAGC
TRVB30-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaTGCCCCAGAATCTCTCAGCCTC
TRVB4-1.4-2.4-3-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaAGTGTGCCAAGTCGCTTCTC
TRVB5-1-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaAACTTCCYTGGTCGATTCTCAGG
TRVB5-4-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaTAGATTCTCAGGTCTCCAGTTCCC
TRVB5-5-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaTCTCAGCTCGCCAGTTCCCTAAC
TRVB5-6-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaATTCTCAGGTCACCAGTTCCCTAACTATAG
TRVB5-8-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaTAGATTTTCAGGTCGCCAGTTCCC
TRVB6-5.6-6-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaTCTCCAGATCAACCACAGAGGATTTCC
TRVB7-2-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaGGGCTGCCCAGTGATCGC
TRVB7-6.7-7-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaGGGCTGCCCARTGATCGG
TRVB7-8-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaTGATCGCTTCTTTGCAGAAAGGCC
TRVB7-9-F	GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNactaggtaRGGCTGCTCAGTGATCGG
TCRBconst	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNacgtcgtaCGACCTCGGGTGGGAACAC

Lowercase nucleotides represent barcode sequences.

Table 1.2: TCRb-specific Illumina MiSeq NGS sequencing primers with poly-N dephasing, DNA multiplex identifiers, and MiSeq partial Sequencing Adaptors. Note that complete sample prep requires secondary amplification with PE1/PE2 extension primers.

performed using a Multiplex PCR kit (Qiagen) with 3 μL of template cDNA, 50 pmol of CB primer, and 1.28 pmol of each V primer per 25 μL of reaction for 15 min at 95 °C 15, followed by 30 cycles of 94 °C for 30 s, 59 °C for 90 s, and 72 °C for 60 s, then 72 °C for 10 min. Full-length Illumina sequencing adaptors were added in a second 15-cycle PCR with 0.5 μL of the first PCR product and 10 pmol of forward and reverse Illumina adaptor primers. The final PCR products were pooled in equal amounts and separated on 2% agarose gel. PCR products of 350–450 bp were excised and extracted using a Qiaquick Gel Extraction Kit (Qiagen).

Illumina Sequencing and Sequence Data Analysis. TCRB libraries were sequenced with an Illumina Miseq sequencer (2 × 250-bp paired-end reads). Each paired-end read was first trimmed to 150 bp on each side and then assembled into a single sequences using COPE (1). The reads were then mapped to reference sequences in the IMGT information system (www.imgt.org) using modified IgBLAST. The TCRBV and TCRBJ regions for each sequence were annotated to best-matching sequences. CDR3 sequences were identified as nucleotide sequences between the second conserved cysteine at the 3´ end of the VB gene segment and the conserved phenylalanine at the 5´ portion of the JB segment. Sequences successfully mapped to known VB and JB segments and containing an in-frame CDR3 sequence and JB segment were accepted for further analysis.

TCRB Repertoire Richness Statistics. We used multiple replicate libraries generated from distinct aliquots of T cells for each subset and applied the Chao2 nonparametric estimator of the lower bound of species richness in a population (2, 3). The Chao2 approach is a standard tool in ecology and intends to make use of the information provided by all species in a dataset, rare or not. Instead of excluding all singleton sequences, as was done in previous studies, we corrected for sequencing errors by examining each putative clone S that appeared only in one replicate and decided whether it could be a sequencing error with respect to another clone using the following rule set: S is considered a sequencing error if another clone in the data from that T-cell subset (i) shares the same V and J annotations, (ii) has the same CDR3 length, (iii) has a Hamming distance of two or one to S in the CDR3 nucleotides, and (iv) has higher abundance. If we find another clone satisfying all four criteria,

then the putative clone S is rejected as a clone. Computing Chao2 entails examining the replicates for clones that are found in multiple replicates and contrasting this quantity to the number of clones found in only one in a mathematically informed manner. Chao2 essentially contrasts these two quantities to estimate the extent to which we have covered the full repertoire and uses that to inform how many species there are in total. We used two strategies to estimate confidence intervals of the Chao2 estimator. First, we have produced an empirical confidence interval based on bootstrapping. We have used the BCa bootstrap from DiCiccio and Efron (4) that is a variant of bootstrapping designed in part for confidence intervals when the underlying bootstrap distribution is not symmetric about its center. Following standard technique, we converted the replicated data into sufficient statistics necessary to compute the Chao2, in particular, the per-clone incidence frequencies across the replicates (i.e., the number of replicates that contain the clone). We performed sampling with replacement on this sufficient statistic over all clones. In our bootstrap estimates, we performed 1,000 bootstrapping iterations for each estimate. Second, we estimated the confidence intervals using the formula developed by Chao (2). Fewer cells were available for analysis in our CD8 memory T-cell samples compared with the CD4 memory cells (5 replicates of 250,000 each vs. 5 replicates of 1,000,000 each); therefore, we ran simulations to evaluate the robustness of our estimates of richness, given the differences in sampling. We generated an underlying repertoire with an uneven Zipf distribution (power = -0.5), spread across 1 million clones. From this repertoire, we sampled 5 replicate libraries, for 250,000 virtual cells per replicate, equivalent to the number of CD8 memory cells studied in our experiments. To simulate PCR amplification, each virtual cell was subject to a standard lognormal scaling. Across each of the 100 random iterations of these noisy samplings, we calculated a Chao2 repertoire estimate of richness. The resulting estimates of a mean of 924,400 with a SD of 1756 were very close to the true value of 1 million different TCRs. For comparison, we also performed simulations using sampling depths equivalent to those obtained for CD4 memory T-cell experiments (5 replicates of 1 million cells each). This resulted in a mean of 945,800 and a SD of 800, demonstrating the robustness of the approach to variation in sample sizes.

Clonality Statistics. The "clonality score" was recently developed as a measure to estimate contribution of clonally expanded sequences within a repertoire, as an adaptation of the Gini–Simpson index modified for application to multiple replicate sequencing datasets and corrected for covariance between replicate sequence libraries (5, 6). To perform inference of the clonality score, we used the lymphclon package available on R CRAN to perform accurate estimation of this quantity on the repertoires. The lymphclon inference algorithm explicitly takes advantage of multiple replicated sequencing experiments by modeling the experimental covariances and the per-replicate amplification differences. A description of this methodology is available in ref. 7.

1.3.5 Acknowledgements

This work was made possible by my co-authors, including first author Qian Qia, as well as Yi Liu, Yong Cheng, David Zhang, Ji-Yeun Lee, Richard A. Olshen, Cornelia M. Weyand, Scott D. Boyd, and Jörg J. Goronzy. We thank Dr. A. Chao, National Tsing Hua University, for statistical advice and helpful discussions. This work was supported by National Institutes of Health (NIH) Grants U19 AI090019, R01 AG015043, U19 AI057266, and R01 AI108891 (to J.J.G.); NIH Grant U19 AI090019 and a New Scholar in Aging grant from the Ellison Medical Foundation (to S.D.B.); NIH Grants R01 AR042527, R01 AI044142, and P01 HL058000 (to C.M.W.); and an American Federation for Aging Research Postdoctoral Fellowship in Aging Research (to Q.Q.).

1.3.6 References

1. Palmer DB (2013) The effect of age on thymic function. Front Immunol 4:316.

2. Taub DD, Longo DL (2005) Insights into thymic aging and regeneration. Immunol Rev 205:72 – 93.

3. den Braber I, et al. (2012) Maintenance of peripheral naive T cells is sustained by thymus output in mice but not humans. Immunity 36(2):288 – 297.

4. Nikolich-Zugich J (2008) Ageing and life-long maintenance of T-cell subsets in the face of latent persistent infections. Nat Rev Immunol 8(7):512 – 522.

5. Jean CM, Honarmand S, Louie JK, Glaser CA (2007) Risk factors for West Nile virus neuroinvasive disease, California, 2005. Emerg Infect Dis 13(12):1918 – 1920.

6. Peiris JS, et al.; HKU/UCH SARS Study Group (2003) Clinical progression and viral load in a community outbreak of coronavirus-associated SARS pneumonia: A prospective study. Lancet 361(9371):1767 – 1772.

7. Goodwin K, Viboud C, Simonsen L (2006) Antibody response to influenza vaccination in the elderly: A quantitative review. Vaccine 24(8):1159 – 1169.

8. Cicin-Sain L, et al. (2010) Loss of naive T cells and repertoire constriction predict poor response to vaccination in old primates. J Immunol 184(12):6739 – 6745.

9. Ahmed M, et al. (2009) Clonal expansions and loss of receptor diversity in the naive CD8 T cell repertoire of aged mice. J Immunol 182(2):784 – 792.

10. Yager EJ, et al. (2008) Age-associated decline in T cell repertoire diversity leads to holes in the repertoire and impaired immunity to influenza virus. J Exp Med 205(3):711 – 723.

11. Arstila TP, et al. (1999) A direct estimate of the human alphabeta T cell receptor di- versity. Science 286(5441):958 – 961.

12. Warren RL, et al. (2011) Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured rep- ertoire size of at least 1 million clonotypes. Genome Res 21(5):790 – 797.

13. Robins HS, et al. (2009) Comprehensive assessment of T-cell receptor beta-chain di- versity in alphabeta T cells. Blood 114(19):4099 – 4107.

14. Chao A, Bunge J (2002) Estimating the number of species in a stochastic abundance model. Biometrics 58(3):531 – 539.

15. DiCiccio TJ, Efron B (1996) Bootstrap confidence intervals. Stat Sci 11(3):189 – 228.

16. Chao A (1987) Estimating the population size for capture-recapture data with un- equal catchability. Biometrics 43(4):783 – 791.

17. Czesnikiewicz-Guzik M, et al. (2008) T cell subset-specific susceptibility to aging. Clin Immunol 127(1):107 – 118.

18. Wertheimer AM, et al. (2014) Aging and cytomegalovirus infection differentially and jointly affect distinct circulating T cell subsets in humans. J Immunol 192(5):2143 – 2155.

19. Weekes MP, Carmichael AJ, Wills MR, Mynard K, Sissons JG (1999) Human CD28-CD8 + T cells contain greatly expanded function al virus-specific memory CTL clones. J Immunol 162(12):7569 – 7577.

20. Newell EW, Sigal N, Bendall SC, Nolan GP, Davis MM (2012) Cytometry by time-of- flight shows combinatorial cytokine expression and virus-specific cell niches within a continuum of CD8 + T cell phenotypes. Immunity 36(1):142 – 152.

21. Robins HS, et al. (2010) Overlap and effective size of the human CD8 + T cell receptor repertoire. Sci Transl Med 2(47):47ra64

22. Venturi V, et al. (2011) A mechanism for TCR sharing between T cell subsets and in- dividuals revealed by pyrosequencing. J Immunol 186(7):4285 – 4294.

23. Berard M, Tough DF (2002) Qualitative differences between naïve and memory T cells. Immunology 106(2):127 – 138.

24. Poulin JF, et al. (1999) Direct evidence for thymic function in adult humans. J Exp Med 190(4):479 – 486.

25. Hakim FT, et al. (2005) Age-dependent incidence, time course, and consequences of thymic renewal in adults. J Clin Invest 115(4):930 – 939.

26. Hazenberg MD, Borghans JA, de Boer RJ, Miedema F (2003) Thymic output: A bad TREC record. Nat Immunol 4(2):97 – 99.

27. Min B, et al. (2003) Neonates support lymphopenia-induced proliferation. Immunity 18(1):131 – 140.

28. Schönland SO, et al. (2003) Homeostatic control of T-cell generation in neonates. Blood 102(4):1428 – 1434.

29. Su LF, Kidd BA, Han A, Kotzin JJ, Davis MM (2013) Virus-specific CD4(+) memory- phenotype T cells are abundant in unexposed adults. Immunity 38(2):373 – 383.

30. Boyman O, Krieg C, Homann D, Sprent J (2012) Homeostatic maintenance of T cells and natural killer cells. Cell Mol Life Sci 69(10):1597 – 1608.

31. Akue AD, Lee JY, Jameson SC (2012) Derivation and maintenance of virtual memory CD8 T cells. J Immunol 188(6):2516 – 2523.

32. Johnson PL, Yates AJ, Goronzy JJ, Antia R (2012) Peripheral selection rather than thymic involution explains sudden contraction in naive CD4 T-cell diversity with age. Proc Natl Acad Sci USA 109(52):21432 – 21437.

33. Naylor K, et al. (2005) The influence of age on T cell generation and TCR diversity. J Immunol 174(11):7446 – 7452.

34. Cicin-Sain L, et al. (2007) Dramatic increase in naive T cell turnover is linked to loss of naive T cells from old primates. Proc Natl Acad Sci USA 104(50):19960 – 19965.

35. Rudd BD, et al. (2011) Nonrandom attrition of the naive CD8 + T-cell pool with aging governed by T-cell receptor:pMHC interactions. Proc Natl Acad Sci USA 108(33): 13694 – 13699.

36. Goronzy JJ, Weyand CM (2012) Immune aging and autoimmunity. Cell Mol Life Sci 69(10):1615 – 1623.

37. Datta S, Sarvetnick N (2009) Lymphocyte proliferation in immune-mediated diseases. Trends Immunol 30(9):430 – 438.

38. Jones JL, et al. (2013) Human autoimmunity after lymphocyte depletion is caused by homeostatic T-cell proliferation. Proc Natl Acad Sci USA 110(50):20200 – 20205.

39. Wagner UG, Koetz K, Weyand CM, Goronzy JJ (1998) Perturbation of the T cell repertoire in rheumatoid arthritis. Proc Natl Acad Sci USA 95(24):14447 – 14452.

40. Goronzy JJ, Weyand CM (2001) Thymic function and peripheral T-cell homeostasis in rheumatoid arthritis. Trends Immunol 22(5):251 – 255.

41. Koetz K, et al. (2000) T cell homeostasis in patients with rheumatoid arthritis. Proc Natl Acad Sci USA 97(16):9203 – 9208.

42. Nikolich-Zugich J, Slifka MK, Messaoudi I (2004) The many important facets of T-cell repertoire diversity. Nat Rev Immunol 4(2):123 – 132.

1.3.7 Copyright

This work was published in the Proceedings of the National Academy of Science with the following reference: Qi, Q., Liu, Y., Cheng, Y., Glanville, J., Zhang, D., Lee, J.Y., Olshen, R.A., Weyand, C.M., Boyd, S.D. and Goronzy, J.J., 2014. Diversity and clonal selection in the human T-cell repertoire. Proceedings of the National Academy of Sciences, 111(36), pp.13139-13144.

1.4 Single cell receptor and phenotype sequencing

Although each T lymphocyte expresses a T-cell receptor (TCR) that recognizes cognate antigen and controls T-cell activation, different T cells bearing the same TCR can be functionally distinct. Each TCR is a heterodimer, and both a and b-chains contribute to determining TCR antigen specificity. Here we present a methodology enabling integration of information about TCR specificity with information about T cell function. This method involves sequencing of TCRa and TCRβ genes, and amplifying functional genes characteristic of different T cell subsets, in single T cells. Because this approach retains information about individual TCRa-TCRβ pairs, TCRs of interest can be expressed and used in functional studies, for antigen discovery, or in therapeutic applications. We apply this approach to study the clonal ancestry and differentiation of T lymphocytes infiltrating a human colorectal carcinoma.

1.4.1 Introduction

Single-cell analysis can reveal important functional insights that are masked in bulk analysis of cell populations(1–3). Recent technological advances have improved our ability to query expression of multiple genes in single cells simultaneously, thereby helping to resolve the complexity inherent in heterogeneous populations of cells including T lymphocytes. These technologies include time-of-flight mass cytometry (CyTOF), RNA sequencing (RNA-seq) and quantitative RT-PCR(4–7). However, these technologies have not thus far been applied in a high-throughput manner to include the most distinctive genes a T cell expresses: the genes that encode the TCR.

The TCR, which determines which complexes of antigenic peptide–major histocompatibility complex (MHC) the T cell responds to, plays a major role in controlling the selection, function and activation of T cells(8). Because the TCR expressed in each T cell is composed of αand β-chain genes that are derived by somatic V(D)J recombination, the TCR repertoire in any given individual is tremendously diverse(9). Therefore, the TCR also serves as a unique identifier of a T-cell's ancestry, because it is likely that any two T cells expressing the same TCRαβ pair arose from a common T-cell clone.

There is great potential synergy in pairing TCR sequences (which can reveal information about T-cell ancestry and antigen specificity) with information about expression of genes characteristic of particular T-cell functions. Integrating these two types of information can allow one to comprehensively describe a given T cell. For example, it is becoming clear that T cells responding to different antigens can have very different phenotypic and functional properties, even if these antigens are derived from the same pathogen(10). The ability to link T-cell function and TCR specificity will enable one to determine which functional subsets of T cells have undergone clonal expansion and which clones exhibit plasticity, ultimately give rise to progeny that express the same TCRαβ heterodimers, but exhibit diverse functional phenotypes. It will also allow identification of TCRαβ heterodimers expressed in individual T cells of interest without expansion of the T-cell population in vitro, which can result in loss of functional integrity. These heterodimers can be invaluable in studies designed to discover antigens11 or in therapeutic applications(12).

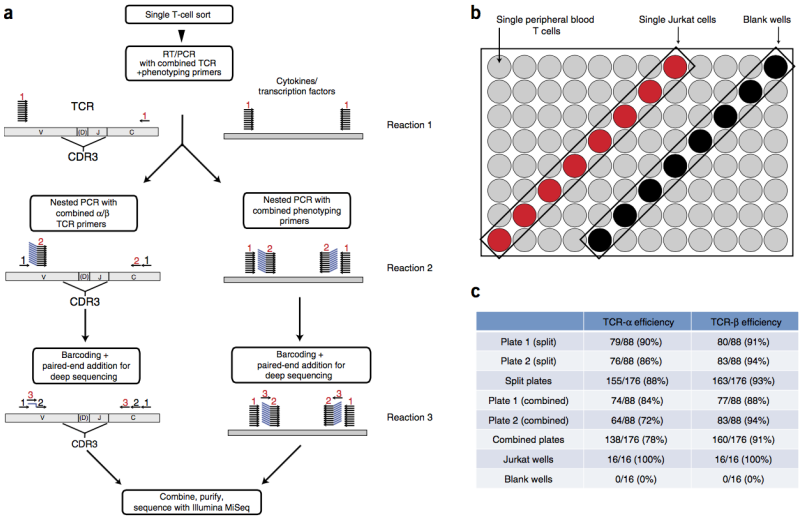
Here we present an approach enabling the simultaneous sequencing of TCRα and TCRβ genes and amplification of transcripts of functional interest in single T cells. This approach enables both TCR sequencing and extensive phenotypic analysis in single T cells, linking TCR specificity with information about T-cell function.

1.4.2 Results

Strategy We and others have successfully sequenced TCR genes from single, sorted T cells using a nested PCR approach followed by Sanger sequencing(13–15). Here we

devise a strategy enabling simultaneous sequencing of rearranged TCR genes and multiple functional genes in single, sorted T cells through deep sequencing. In addition to enabling the analysis of multiple functional genes in parallel with TCR sequencing, this approach has several advantages over previous TCR sequencing methods that utilize Sanger sequencing(13–15). First, it is efficient (5,000–10,000 cells can be sequenced in one sequencing run) and less labor intensive as individual PCR products do not need to be purified and sequenced separately. Second, it is also very accurate as consensus sequences are determined from a high number of independent sequencing reads (often exceeding 1,000) per TCR gene, essentially eliminating the effect of sequencing error. Third, it is well-established that individual T cells can express two TCRα genes(16,17). Our approach uniquely enables sequencing of multiple TCRα genes from most single T cells and determination of which of these are functional.

In our method, single T cells are sorted into 96-well PCR plates (Fig. 1a). An RT-PCR reaction is done using 76 TCR primers and 34 phenotyping primers (Supplementary Fig. 1 and Supplementary Tables 1–3). The products are then used in a second PCR reaction—either one that uses nested primers for TCR genes or one that uses nested primers for phenotypic markers, including cytokines and transcription factors. A third reaction is then performed that incorporates individual barcodes into each well (Supplementary Fig. 2)(18). The products are combined, purified and sequenced using the Illumina MiSeq platform. The resulting paired-end sequencing reads are assembled and deconvoluted using barcode identifiers at both ends of each sequence by a custom software pipeline to separate reads from every well in every plate (Supplementary Note). The resulting sequences are then analyzed using a program called VDJFasta19, which we have adapted to resolve barcodes and analyze sequences with a customized gene-segment database that includes relevant transcription factors and cytokine genes. The population of annotated sequences above background levels in each well is then measured (see Online Methods for details on data processing). For TCR sequences, the CDR3 nucleotide sequences are then extracted and translated. For phenotypic parameters, the presence or absence of a transcript in a particular well is scored.



	TCR-α efficiency	TCR-β efficiency
Plate 1 (split)	79/88 (90%)	80/88 (91%)
Plate 2 (split)	76/88 (86%)	83/88 (94%)
Split plates	155/176 (88%)	163/176 (93%)
Plate 1 (combined)	74/88 (84%)	77/88 (88%)
Plate 2 (combined)	64/88 (72%)	83/88 (94%)
Combined plates	138/176 (78%)	160/176 (91%)
Jurkat wells	16/16 (100%)	16/16 (100%)
Blank wells	0/16 (0%)	0/16 (0%)

Figure 1.10: Strategy for single-cell TCR sequencing and phenotyping, and determination of TCR-sequencing efficiency. (a) Strategy for simultaneous TCR-sequence determination and phenotyping of single, sorted T cells. Single T cells were sorted into 96-well plates. The initial RT-PCR reaction (reaction 1) uses 76 TCR primers and 34 phenotyping primers. An aliquot of the product of reaction 1 is used for two separate second nested PCR reactions (reaction 2), one for TCR sequencing and one for phenotyping. Using an aliquot of reaction 2 product as a template, a third PCR reaction is performed that incorporates individual barcodes into each well and enables subsequent sequencing using the Illumina MiSeq platform. For TCR sequencing, the third reaction can be split into separate TCRα and TCRβ reactions (for optimal efficiency), or the two TCR chains can be included in a single reaction. The products of reaction 3 are then combined and sequenced using the Illumina MiSeq platform. (b,c) Accuracy and efficiency of TCR sequencing using this method. (b) Strategy used to validate TCR sequencing. Into each 96-well test plate, individual human peripheral blood T cells were sorted into 80 wells (gray). Single Jurkat T cells were sorted into eight other wells (red), and the remaining eight wells (black) were left empty (blank). For sequencing of these test plates, reaction 3 was initially performed separately for TCRα and TCRβ (split). It was also repeated with TCRα and TCRβ amplified together in the same reaction (combined). (c) Efficiency of TCRα and TCRβ sequencing in split or combined formats. Plate 1 contained 80 single CD45RA+CD4+TCRαβ+ T cells, and plate 2 contained 80 single CD4+ or CD8+ TCRαβ+ T cells sorted from peripheral blood of the same healthy human donor. Identical Jurkat sequences were obtained from all Jurkat wells. No sequences were obtained from any empty wells.

TCR sequencing validation To validate our TCR sequencing methodology, 80 single CD45RA+CD4+TCRαβ+ T cells were sorted from peripheral blood of a healthy human donor into one 96-well plate, and 80 single CD4+ or CD8+TCRαβ+ T cells were sorted from the same sample into a second plate. CD45RA marks naive CD4+ T cells that are not expected to have undergone much clonal expansion(20). The Jurkat human T-leukemic cell line was used as a positive control(21). Into both plates, individual Jurkat T cells were sorted into eight wells, and eight wells were left blank (Fig. 1b). These plates were initially subjected to the first reaction containing 74 TCR variable (V)-region primers, 2 constant (C)-region primers and 34 phenotyping primers. Phenotyping primers were included in the first reaction to demonstrate that the inclusion of these primers does not interfere with TCR sequencing. The subsequent nested PCR and barcoding reactions were then performed according to the protocol shown in Figure 1a; reaction products were subsequently sequenced and analyzed.

Out of 160 wells into which single, peripheral blood αβ T cells were randomly sorted, productive TCRβ sequences were successfully obtained in 147 wells (92%), and at least one productive TCRα sequence was found in 139 wells (87%) wells (Fig. 1c and Supplementary Table 4). Productive TCR genes have been joined in the proper reading frame by V(D)J recombination without premature stop codons, enabling expression of a complete TCRα or β chain. Paired, productive TCRαβ sequences were found in 131 (82%) wells. Identical Jurkat TCRαβ sequences were found in 16/16 wells into which Jurkat cells were sorted but in no other wells on the plates (Fig. 1c and Supplementary Table 4). There were no sequences above background found in the wells into which no cell was sorted (see Supplementary Fig. 3 for details regarding background). The absence of sequences from wells with no cells and the presence of Jurkat sequences only in the 16 designated Jurkat wells indicates that cross-contamination of wells was minimal. Optimal TCR-sequencing efficiency was obtained when the third PCR reaction (barcoding) was done in two separate plates—one for TCRα and one for TCRβ; there was a only a marginal loss of efficiency when these reactions were done together in one plate (Fig. 1c).

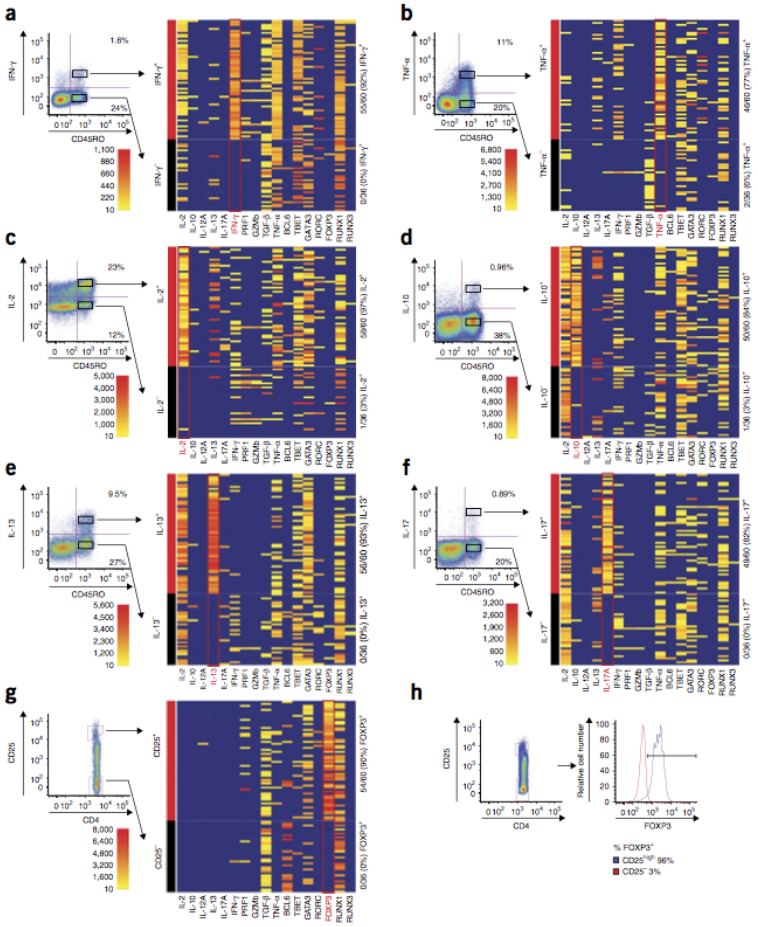


Figure 1.11: Accuracy of phenotypic analysis. (a–f) Peripheral blood T cells were stimulated for 3 h with PMA and ionomycin, and expression of the indicated cytokines was analyzed by cytokine secretion assays that do not require cell fixation: IFN-γ (a), TNF-α (b), IL-2 (c), IL-10 (d), IL-13 (e), IL-17 (f). Sixty single CD45RO+CD4+ T cells that were clearly positive for the indicated cytokine protein and 36 single CD45RO+CD4+ T cells that were clearly negative for the indicated cytokine protein were sorted, and expression of the same cytokine genes were measured by the method depicted in Figure 1a. Seventeen independent phenotypic parameters were assayed in single, sorted cells, and the phenotypic parameter on which cells were sorted is indicated in red. Heatmaps indicate read count of each parameter (x axis) within each particular well (y axis). Scale indicates number of reads obtained from a given well for the indicated parameter. Wells indicated in blue did not display any reads that reached threshold. (g) Unstimulated CD4+ T cells were sorted based upon CD25 expression to validate phenotypic analysis for FOXP3. Sixty single CD4+ T cells with high CD25 expression and 36 single CD4+ T cells that were negative for CD25 expression by flow cytometry were sorted and assayed as in a–f. (h) Expression of CD25 and FOXP3 protein was measured by flow cytometry. Cells from the same donor were fixed and stained with anti-CD25 and anti-FOXP3 antibodies. Histograms on right depict FOXP3 expression in gated CD25high and CD25− populations.

T cells often express two recombined TCRα genes(16,17). Sanger sequencing cannot be performed on heterogeneous products, therefore, methods that rely on Sanger sequencing cannot easily identify multiple TCRα chains from a single cell(15). Furthermore, the presence of multiple TCRα chains can hinder the efficiency and accuracy of sequencing in methods based on Sanger sequencing. Because our strategy employs deep sequencing, wherein each template is amplified and sequenced independently, we can readily derive multiple TCRα sequences from individual cells. On average (assuming 20 96-well plates on a single sequencing run), we obtain ~5,000 total TCRα or TCRβ sequences with the same set of barcodes, specifying they are derived from the same well. To distinguish between TCR sequences that differ owing to a sequencing and/or PCR error and those that are probably derived from different TCR genes, our software determines a cutoff value in similarity based upon the assumed rate of sequencing and/or PCR error(22). All sequences exceeding this value (e.g., they are very similar to one another) are assumed to derive from the same TCR gene and a consensus sequence is determined. We can therefore accurately identify multiple TCR gene sequences among a heterogeneous group of sequences tagged with the same barcode.

We detected multiple TCRα chains in 80/139 (58%) of wells containing at least one productive TCRα chain sequence (Supplementary Table 4). In contrast, we did not detect multiple TCRβ chains or multiple nonproductive TCRα chains in any wells. This indicates that cross-contamination of wells or the erroneous sorting of two cells into wells is minimal. With the exception of designated Jurkat wells, there were no repeated TCRs present in the first plate containing 80 naive CD45RA+CD4+ T cells. This is consistent with the expectation that naive T cells have undergone minimal clonal expansion and therefore it is unlikely that two identical clones would be in the population of 80 cells sorted into a single plate. In our second plate, which contained 80 total (naive and non-naive) TCRαβ+ T cells, we detected four repeated TCR sequences in 11 different wells (Supplementary Table 5). All these repeated T cells were scattered across the plates and not within close proximity to each other, suggesting that the repetition did not arise as a result of cross-well contamination. For one TCRβ sequence that was repeated across four wells (CAWTLGGNEQFF),

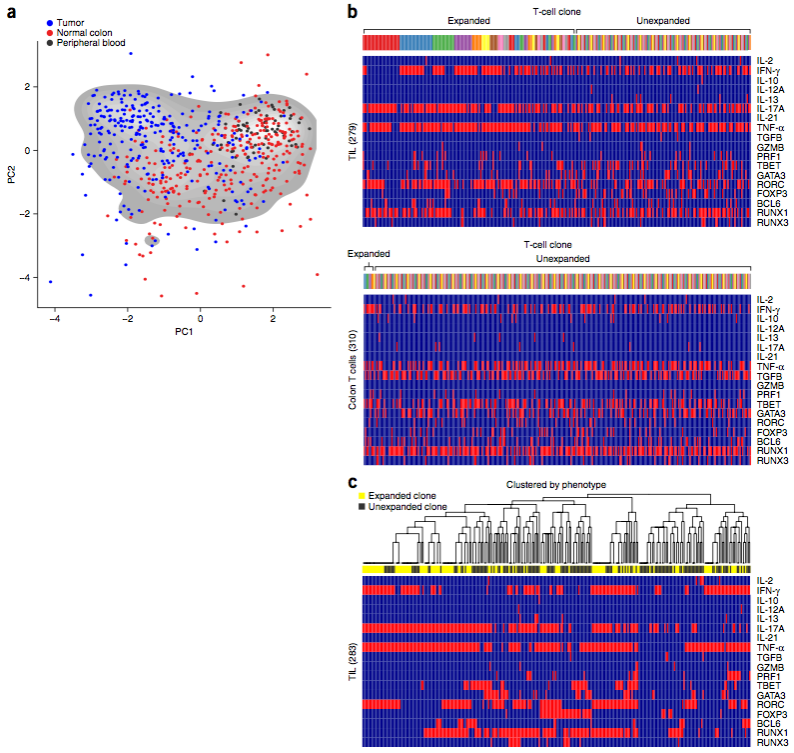


Figure 1.12: TCR sequencing and phenotypic analysis of single human TILs. (a) T cells were sorted and analyzed using the procedure from Figure 1a. PCA to depict phenotypic + T cells from tumor (blue) and adjacent colon (red) of a single patient, and from peripheral blood of another healthy diversity of PMAplus ionomycin-stimulated CD4 donor (black). PCA parameter loadings are shown in Supplementary Figure 7. Each dot represents a single T cell. (b) Top two panels: 17-parameter (parameters listed on x axis) phenotypic analysis of stimulated CD4+ T cells from tumor (top) and colon (bottom) of a single patient. Individual T cells are grouped by TCR sequence; each color on the bar above the heat maps represents a distinct TCR sequence. (c) Hierarchical clustering of different cells by phenotype, with expanded (yellow) and unexpanded (black) T-cell clones (read out by TCR sequence) indicated in the horizontal bar above the heat map.

each well contained sequences of the same two productive TCRα genes. For a TCRβ (CASSYGDPGGLDGELFF) that was repeated across three wells, the same productive TCRα gene was detected in all three wells. Additionally, within two of these three wells, an identical nonproductive TCRα gene was detected. These findings confirm that detection of two TCRα rearrangements in a particular cell is repeatable and reliable, and not likely to be the result of contamination or error.

We detected two productive TCRα genes in 19/139 (14%) wells containing at least one productive TCRα chain (Supplementary Table 4). This reinforces the importance of single-molecule sequencing methods like ours to determine true TCRαβ heterodimers. Methods that can detect only one TCRα gene per cell may falsely identify a TCRαβ heterodimer because in cells expressing two productive TCRα genes, only one TCRα gene product is thought to be expressed at the T-cell surface(16). Further, in cases where only one TCRα chain is detected in a particular cell, there is a possibility that another productive TCRα chain is present but not detected. This possibility is unlikely with our methodology given its efficiency and the fact that we detected all variable (V)-regions in our TCRα data even in the presence of other TCRα chains within the same T cell (Supplementary Fig. 4). However, due to this possibility, all TCRs derived through this method that are reconstituted for use in functional studies should be validated.

Phenotyping validation In addition to TCR sequencing, this method enables one to simultaneously query multiple phenotypic parameters from single T cells. In our phenotyping panel, we included multiple cytokines and transcription factors that influence T-cell function and define certain T-cell subsets (Supplementary Table 2). Our method is uniquely enabling in this regard, as other methods for measuring expression of cytokines and transcription factors (e.g., flow cytometry) generally require cellular fixation, which compromises the integrity of nucleic acids and makes it difficult to perform TCR sequencing. Furthermore, cellular fixation–based methods for detecting transcription factors are arduous and unreliable, even compared to methods for detecting intracellular cytokine expression(23).

For example, the functional diversity of CD4+ T cells is dependent upon expression of various transcription factors(24). Some of these transcription factors are

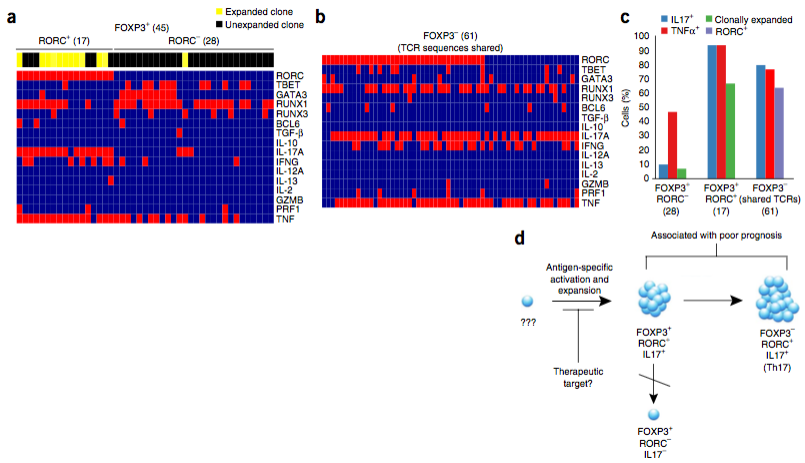


Figure 1.13: TCR sequencing and phenotypic analysis of FOXP3+ TILs. (a) Clonal expansion (indicated in horizontal bar above heatmaps) and phenotype (parameters indicated on right column) of FOXP3+RORC− T cells and FOXP3+RORC+ T cells. (b) Phenotypic analysis of FOXP3− T cells that share TCR sequences with FOXP3+ populations (total numbers of cells analyzed are indicated in parentheses). (c) Frequency of cells within each indicated population of T cells (x axis) expressing phenotypic markers indicated (right). (d) Model suggested by analysis of TILs. A single T cell is activated by antigen and expands to give rise to FOXP3+RORC+ IL-17–producing T cells, which eventually lose expression of FOXP3 to give rise to FOXP3−RORC+ IL-17–producing T cells. FOXP3+RORC− T cells do not show evidence of clonal expansion, and are thus unlikely to arise from FOXP3+RORC+ cells that lose expression of RORC.

termed "master regulators" and their expression has been used to specify particular T-cell lineages. We included T-bet, GATA3, RoRγT (RAR-related orphan receptor gamma T, which is encoded by RORC), BCL-6 and FOXP3 (Forkhead box P3), which have been used to specify helper T type 1 (Th1) cells, Th2, Th17, follicular helper (TfH) and regulatory T (Treg) cells, respectively25,26, in our phenotyping analysis. We also included the runt-related transcription factors Runx1 and Runx3, which influence T-cell differentiation(27). Lastly, we also included both pro-inflammatory and inhibitory cytokines that mediate T-cell effector function and define the various T-cell subsets; these include interferon (IFN)-γ (Th1), interleukin (IL)-13 (Th2), IL-17 (Th17), IL-10 and TGFβ (Treg).

To validate this part of our methodology, we used flow cytometry– based cytokine capture assays (Miltenyi), which enable the determination of cytokine expression without the need for cell fixation(28). We tested expression of the following cytokines for which cytokine secretion assays are commercially available: TNF-α, IFN-γ, IL-2, IL-10, IL-13 and IL-17. Into each plate we sorted 60 single CD4+CD45RO+ memory phenotype T cells from healthy human peripheral blood that were positive for protein expression of a particular cytokine and 36 single CD4+CD45RO+ T cells that were negative for cytokine protein expression (Fig. 2 and Supplementary Table 6). These plates were initially amplified with the first reaction containing 74 TCR Vregion primers, 2 C-region primers and 34 phenotyping primers. TCR primers were included in the first reaction to demonstrate that their presence does not interfere with subsequent phenotyping reactions.

Nested PCR, barcoding and sequencing analysis was performed for phenotypic parameters. We detected transcripts in single, cytokinepositive T cells with 77–97% sensitivity (Fig. 2 and Supplementary Table 7). Our false-positive rate was very low; the specificity of our assay was 94–100% when compared to the relevant cytokine capture assays (Fig. 2 and Supplementary Table 6).

For some of the cytokines genes we validated, there does appear to be a low false-positive rate compared to cytokine secretion assays. Because these wells clearly exceed background levels (Methods and Supplementary Fig. 3), this suggests that these rare cells do indeed express the particular mRNA, although its protein product

is not detected. This is not surprising given that cytokine genes are subject to particularly tight regulation, including translational repression that might prevent protein expression even in the presence of mRNA29.

We also readily detected expression of all the transcription factors in our panel in single T cells. For most of these transcription factors, there are no available surface markers that reliably predict expression. An exception is FOXP3, whose expression correlates well with high expression of the surface marker CD25 in CD4+ T cells30. To validate our methodology for FOXP3 expression, we sorted 60 single CD25highCD4+ T cells and 36 single CD25−CD4+ T cells into a single plate. We detected FOXP3 in 54/60 (90%) of CD25high cells and 0/36 (0%) of CD25− cells (Fig. 2g). We also fixed and stained T cells from the same donor with both CD25 and FOXP3 to confirm the correlation between the high expression of CD25 and FOXP3 (Fig. 2h).

We could detect as little as one molecule of template in a given cell, although sensitivity improves with increased template abundance (Supplementary Fig. 5). Whereas frequency of detection improves with template abundance, read number of a given transcript in the cells called as positive does not (Supplementary Fig. 5). This demonstrates that our methodology is binary, in that it indicates presence or absence of a given transcript. Therefore, read number per well should not be considered a quantitative indicator of the abundance of that transcript.

It is very possible that a particular mRNA might be expressed but not detected in a particular cell, especially at lower copy number (Supplementary Fig. 5). Therefore, we expect that false negatives will occur with this method. However, our data show that false positives do not occur at a significant rate (Supplementary Fig. 3). Thus, for practical purposes, the positive predictive value of our assay exceeds its negative predictive value for any given parameter. One should consider this when analyzing data using this methodology.

Despite the many factors that might contribute to discordance between mRNA and protein detection, our data correlate remarkably well with data from cytokine capture assays and with CD25 protein expression in the case of FOXP3 (Fig. 2 and Supplementary Table 6). However, we considered either the cytokine secretion assays or CD25 staining as the gold standard and did not take into account the possibility

TRAV1, Reaction 1	CTGCACGTACCAGACATCTGGGTT	TRAV1, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**AGGTCGTTTTTCTTCATTCCTTAGTC
TRAV2, Reaction 1	GGCTCAAAGCCTTCTCAGCAGG	TRAV2, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**ACGATACAACATGACCTATGAACGG
TRAV3, Reaction 1	GGATAACCTGGTTAAAGGCAGCTA	TRAV3.1, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**CTTTGAAGCTGAATTTAACAAGAGCC
TRAV4, Reaction 1	GGATACAAGACAAAAGTTACAAACGA	TRAV4.1, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**CTCCCTGTTTATCCCTGCCGAC
TRAV5, Reaction 1	GCTGACGTATATTTTTTCAAATATGGA	TRAV5.1, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**AAACAAGACCAAAGACTCACTGTTC
TRAV6, Reaction 1	GGAAGAGGCCCTGTTTTCTTGCT	TRAV6, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**AAGACTGAAGGTCACCTTTGATACC
TRAV7, Reaction 1	GCTGGATATGAGAAGCAGAAAGGA	TRAV7, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**ACTAAATGCTACATTACTGAAGAATGG
TRAV8, Reaction 1	AGGACTCCAGCTTCTCCTGAAGTA	TRAV8, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GCATCAACGGTTTTGAGGCTGAATTTAA
TRAV9, Reaction 1	GTATGTCCAATATCCTGGAGAAGGT	TRAV9, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GAAACCACTTCTTTCCACTTGGAGAA
TRAV10, Reaction 1	CAGTGAGAACACAAAGTCGAACGG	TRAV10, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**TACAGCAACTCTGGATGCAGACAC
TRAV12.1, Reaction 1	CCTAAGTTGCTGATGTCCGTATAC	TRAV12, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GAAGATGGAAGGTTTACAGCACA
TRAV12.2, Reaction 1	GGGAAAAGCCCTGAGTTGATAATGT	TRAV13.1, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GACATTCGTTCAAATGTGGGCGAA
TRAV12.3, Reaction 1	GCTGATGTACACATACTCCAGTGG	TRAV13.2, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GGCAAGGCCAAAGAGTCACCGT
TRAV13.1, Reaction 1	CCCTTGGTATAAGCAAGAACTTGG	TRAV14, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**TCCAGAAGGCAAGAAAATCCGCCA
TRAV13.2, Reaction 1	CCTCAATTCATTATAGACATTCGTTC	TRAV16, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GCTGACCTTAACAAAGGCGAGACA
TRAV14, Reaction 1	GCAAAATGCAACAGAAGGTCGCTA	TRAV17, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**TTAAGAGTCACGCTTGACACTTCCA
TRAV16, Reaction 1	TAGAGAGAGCATCAAAGGCTTCAC	TRAV18, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GCAGAGGTTTTCAGGCCAGTCCT
TRAV17, Reaction 1	CGTTCAAATGAAAGAGAGAAACACAG	TRAV19, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**TCCACCAGTTCCTTCAACTTCACC
TRAV18, Reaction 1	CCTGAAAAGTTCAGAAAACCAGGAG	TRAV20, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GCCACATTAACAAAGAAGGAAAGCT
TRAV19, Reaction 1	GGTCGGTATTCTTGGAACTTCCAG	TRAV21, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GCCTCGCTGGATAAATCATCAGGA
TRAV20, Reaction 1	GCTGGGGAAGAAAAGGAGAAAGAAA	TRAV22, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**ACGACTGTCGCTACGGAACGCTA
TRAV21, Reaction 1	GTCAGAGAGAGCAAACAAGTGGAA	TRAV23, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**CACAATCTCCTTCAATAAAAGTGCCA
TRAV22, Reaction 1	GGACAAAACAGAATGGAAGATTAAGC	TRAV24, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**ACGAATAAGTGCCACTCTTAATACCA
TRAV23, Reaction 1	CCAGATGTGAGTGAAAAGAAAGAAG	TRAV25, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GTTTGGAGAAGCAAAAAAGAACAGCT
TRAV24, Reaction 1	GACTTTAAATGGGGATGAAAAGAAGA	TRAV26.1, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**CAGAAGACAGAAAGTCCAGCACCT
TRAV25, Reaction 1	GGAGAAGTGAAGAAGCAGAAAAGAC	TRAV26.2, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**ATCGCTGAAGACAGAAAGTCCAGT
TRAV26.1, Reaction 1	CCAATGAAATGGCCTCTCTGATCA	TRAV27, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**ACTAACCTTTCAGTTTGGTGATGCAA
TRAV26.2, Reaction 1	GCAATGTGAACAACAGAATGGCCT	TRAV29, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**CTTAAACAAAAGTGCCAAGCACCTC
TRAV27, Reaction 1	GGTGGAGAAGTGAAGAAGCTGAAG	TRAV30, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**AATATCTGCTTCATTTAATGAAAAAAAGC
TRAV29, Reaction 1	GGATAAAAATGAAGATGGAAGATTCAC	TRAV34, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**CCAAGTTGGATGAGAAAAAGCAGCA
TRAV30, Reaction 1	CCTGATGATATTACTGAAGGGTGGA	TRAV35, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**CTCAGTTTGGTATAACCAGAAAGGA
TRAV34, Reaction 1	GGTGGGGAAGAGAAAAGTCATGAA	TRAV36, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GGAAGACTAAGTAGCATATTAGATAAG
TRAV35, Reaction 1	GGTGAATTGACCTCAAATGGAAGAC	TRAV38, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**CTGTGAACTTCCAGAAAGCAGCCA
TRAV36, Reaction 1	GCTAACTTCAAGTGGAATTGAAAAGA	TRAV39, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**CCTCACTTGATACCAAAGCCCGT
TRAV38, Reaction 1	GAAGCTTATAAGCAACAGAATGCAAC	TRAV40, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**AGGCGGAAATATTAAAGACAAAAACTC
TRAV39, Reaction 1	GGAGCAGTGAAGCAGGAGGGAC	TRAV41, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GATTAATTGCCACAATAAACATACAGG
TRAV40, Reaction 1	GAGAGACAATGGAAAACAGCAAAAAC	TRBV2, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GCCTGATGGATCAAATTTCACTCTG
TRAV41, Reaction 1	GCTGAGCTCAGGGAAGAAGAAGC	TRBV3-1, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**TCTCACCTAAATCTCCAGACAAAGCT
TRBV2, Reaction 1	CTGAAATATTCGATGATCAATTCTCAG	TRBV4, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**CCTGAATGCCCCAACAGCTCTC
TRBV3-1, Reaction 1	TCATTATAAATGAAACAGTTCCAAATCG	TRBV5-4,8, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**CTCTGAGCTGAATGTGAACGCCT
TRBV4, Reaction 1	AGTGTGCCAAGTCGCTTCTCAC	TRBV5-1, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**CGATTCTCAGGGCGCCAGTTCTCT
TRBV5-4,8, Reaction 1	CAGAGGAAACTYCCCTCCTAGATT	TRBV6-1, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**TGGCTACAATGTCTCCAGATTAAACAA
TRBV5-1, Reaction 1	GAGACACAGAGAAACAAAGGAAACTTC	TRBV6-2,3, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**CCCCTGATGGCTACAATGTCTCCAGA
TRBV6-1, Reaction 1	GGTACCACTGACAAAGGAGAAGTCC	TRBV6-4, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GTGTCTCCAGAGCAAACACAGATGATT
TRBV6-2,3, Reaction 1	GAGGGTACAACTGCCAAAGGAGAGGT	TRBV6-5,6, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GTCTCCAGATCAACCACAGAGGAT
TRBV6-4, Reaction 1	GGCAAAGGAGAAGTCCCTGATGGTT	TRBV6-8, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GTCTCTAGATTAAACACAGAGGATTTC
TRBV6-5,6, Reaction 1	AAGGAGAAGTCCCSAATGGCTACAA	TRBV6-9, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GGCTACAATGTATCCAGATCAAACA
TRBV6-8, Reaction 1	CTGACAAAGAAGTCCCCAATGGCTAC	TRBV7-2, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**TCGCTTCTCTGCAGAGAGGACTGG
TRBV6-9, Reaction 1	CACTGACAAAGGAGAAGTCCCCGAT	TRBV7-3, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**CGGTTCTTTGCAGTCAGGCCTGA
TRBV7-2, Reaction 1	AGACAAATCAGGGCTGCCCAGTGA	TRBV7-8, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**CCAGTGATCGCTTCTTTGCAGAAA
TRBV7-3, Reaction 1	GACTCAGGGCTGCCCAACGAT	TRBV7-4,6, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**TCTCCACTCTGAMGATCCAGCGCA
TRBV7-8, Reaction 1	CCAGAATGAAGCTCAACTAGACAA	TRBV7-7, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GCAGAGAGGCCTGAGGGATCCAT
TRBV7-4,6, Reaction 1	GGTTCTCTGCAGAGAGGCCTGAG	TRBV7-9, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**CTGCAGAGAGGCCTAAGGGATCT
TRBV7-7, Reaction 1	GGCTGCCCAGTGATCGGTTCTC	TRBV9, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**CTCCGCACAACAGTTCCCTGACTT
TRBV7-9, Reaction 1	GACTTACTTCCAGAATGAAGCTCAACT	TRBV10-1,3, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**CAGATGGCTAYAGTGTCTCTAGATCAAA
TRBV9, Reaction 1	GAGCAAAAGGAAACATTCTTGAACGATT	TRBV10-2, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GTTGTCTCCAGATCCAAGACAGAGAA
TRBV10-1,3, Reaction 1	GGCTRATCCATTACTCATATGGTGTT	TRBV11, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GCAGAGAGGCTCAAAGGAGTAGACT
TRBV10-2, Reaction 1	GATAAAGGAGAAGTCCCCGATGGCT	TRBV12-3,4, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GCTAAGATGCCTAATGCATCATTCTC
TRBV11, Reaction 1	GATTCACAGTTGCCTAAGGATCGAT	TRBV12-5, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**CTCAGCAGAGATGCCTGATGCAACT
TRBV12-3,4, Reaction 1	GATTCAGGGATGCCCGAGGATCG	TRBV13, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**TCTCAGCTCAACAGTTCAGTGACTA
TRBV12-5, Reaction 1	GATTCGGGGATGCCGAAGGATCG	TRBV14, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GCTGAAAGGACTGGAGGGACGTAT
TRBV13, Reaction 1	GCAGAGCGATAAAGGAAGCATCCCT	TRBV15, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GATAACTTCCAATCCAGGAGGCCG
TRBV14, Reaction 1	TCCGGTATGCCCAACAATCGATTCT	TRBV16, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GCTAAGTGCCTCCCAAATTCACCC
TRBV15, Reaction 1	GATTTTAACAATGAAGCAGACACCCCT	TRBV18, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GGAACGATTTTCTGCTGAATTTCCCA
TRBV16, Reaction 1	GATGAAACAGGTATGCCCAAGGAAAG	TRBV19, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GGTACAGCGTCTCTCGGGAGAAGA
TRBV18, Reaction 1	TATCATAGATGAGTCAGGAATGCCAAAG	TRBV20-1, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GGACAAGTTTCTCATCAACCATGCAA
TRBV19, Reaction 1	GACTTTCAGAAAGGAGATATAGCTGAA	TRBV24-1, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**TGGATACAGTGTCTCTCGACAGGC
TRBV20-1, Reaction 1	CAAGGCCACATACGAGCAAGGCGTC	TRBV25-1, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**CAACAGTCTCCAGAATAAGGACGGA
TRBV24-1, Reaction 1	CAAAGATATAAACAAAGGAGAGATCTCT	TRBV27-1, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**TACAAAGTCTCTCGAAAAGAGAAGAGGA
TRBV25-1, Reaction 1	AGAGAAGGGAGATCTTTCCTCTGAGT	TRBV28, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GGGGTACAGTGTCTCTAGAGAGA
TRBV27-1, Reaction 1	GACTGATAAGGGAGATGTTCCTGAAG	TRBV29, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**GTTTCCCATCAGCCGCCCAAACCTA
TRBV28, Reaction 1	GGCTGATCTATTTCTCATATGATGTTAA	TRBV30, Reaction 2	**CCAGGGTTTTCCCAGTCACGAC**CAGACCCCAGGACCGGCAGTTCAT
TRBV29, Reaction 1	GCCACATATGAGAGTGGATTTGTCATT	TRAC, Reaction 2	CAGACAGACTTGTCACTGGATTTAG
TRBV30, Reaction 1	GGTGCCCCAGAATCTCTCAGCCT	TRBC, Reaction 2	CTTTTGGGTGTGGGAGATCTCTG
TRAC, Reaction 1	CGGTGAATAGGCAGACAGACTTGT		
TRBC, Reaction 1	ACCAGTGTGGCCTTTTGGGTGTG		

Table 1.3: TCR sequencing primers for the first two PCR reactions. Common sequences are indicated in bold.

AlphaBC1	CTGCTGAACCGCTCTTCCGATCTNN**GTTCA**GTCACTGGATTTAGAGTCTCTCAG
AlphaBC2	CTGCTGAACCGCTCTTCCGATCTNN**CAGGA**GTCACTGGATTTAGAGTCTCTCAG
AlphaBC3	CTGCTGAACCGCTCTTCCGATCTNN**TTATA**GTCACTGGATTTAGAGTCTCTCAG
AlphaBC4	CTGCTGAACCGCTCTTCCGATCTNN**CCTGT**GTCACTGGATTTAGAGTCTCTCAG
AlphaBC5	CTGCTGAACCGCTCTTCCGATCTNN**ACCGC**GTCACTGGATTTAGAGTCTCTCAG
AlphaBC6	CTGCTGAACCGCTCTTCCGATCTNN**ACTTA**GTCACTGGATTTAGAGTCTCTCAG
AlphaBC7	CTGCTGAACCGCTCTTCCGATCTNN**GCTAG**GTCACTGGATTTAGAGTCTCTCAG
AlphaBC8	CTGCTGAACCGCTCTTCCGATCTNN**GACGT**GTCACTGGATTTAGAGTCTCTCAG
AlphaBC9	CTGCTGAACCGCTCTTCCGATCTNN**GGCTA**GTCACTGGATTTAGAGTCTCTCAG
AlphaBC10	CTGCTGAACCGCTCTTCCGATCTNN**GAATG**GTCACTGGATTTAGAGTCTCTCAG
AlphaBC11	CTGCTGAACCGCTCTTCCGATCTNN**CCAAC**GTCACTGGATTTAGAGTCTCTCAG
AlphaBC12	CTGCTGAACCGCTCTTCCGATCTNN**GAGAC**GTCACTGGATTTAGAGTCTCTCAG
BetaBC1	CTGCTGAACCGCTCTTCCGATCTNN**GTTCA**GAGATCTCTGCTTCTGATGGCTC
BetaBC2	CTGCTGAACCGCTCTTCCGATCTNN**CAGGA**GAGATCTCTGCTTCTGATGGCTC
BetaBC3	CTGCTGAACCGCTCTTCCGATCTNN**TTATA**GAGATCTCTGCTTCTGATGGCTC
BetaBC4	CTGCTGAACCGCTCTTCCGATCTNN**CCTGT**GAGATCTCTGCTTCTGATGGCTC
BetaBC5	CTGCTGAACCGCTCTTCCGATCTNN**ACCGC**GAGATCTCTGCTTCTGATGGCTC
BetaBC6	CTGCTGAACCGCTCTTCCGATCTNN**ACTTA**GAGATCTCTGCTTCTGATGGCTC
BetaBC7	CTGCTGAACCGCTCTTCCGATCTNN**GCTAG**GAGATCTCTGCTTCTGATGGCTC
BetaBC8	CTGCTGAACCGCTCTTCCGATCTNN**GACGT**GAGATCTCTGCTTCTGATGGCTC
BetaBC9	CTGCTGAACCGCTCTTCCGATCTNN**GGCTA**GAGATCTCTGCTTCTGATGGCTC
BetaBC10	CTGCTGAACCGCTCTTCCGATCTNN**GAATG**GAGATCTCTGCTTCTGATGGCTC
BetaBC11	CTGCTGAACCGCTCTTCCGATCTNN**CCAAC**GAGATCTCTGCTTCTGATGGCTC
BetaBC12	CTGCTGAACCGCTCTTCCGATCTNN**GAGAC**GAGATCTCTGCTTCTGATGGCTC
PhenotypeBC1	CTGCTGAACCGCTCTTCCGATCTNN**GTTCA**AGCGGATAACAATTTCACACAGGA
PhenotypeBC2	CTGCTGAACCGCTCTTCCGATCTNN**CAGGA**AGCGGATAACAATTTCACACAGGA
PhenotypeBC3	CTGCTGAACCGCTCTTCCGATCTNN**TTATA**AGCGGATAACAATTTCACACAGGA
PhenotypeBC4	CTGCTGAACCGCTCTTCCGATCTNN**CCTGT**AGCGGATAACAATTTCACACAGGA
PhenotypeBC5	CTGCTGAACCGCTCTTCCGATCTNN**ACCGC**AGCGGATAACAATTTCACACAGGA
PhenotypeBC6	CTGCTGAACCGCTCTTCCGATCTNN**ACTTA**AGCGGATAACAATTTCACACAGGA
PhenotypeBC7	CTGCTGAACCGCTCTTCCGATCTNN**GCTAG**AGCGGATAACAATTTCACACAGGA
PhenotypeBC8	CTGCTGAACCGCTCTTCCGATCTNN**GACGT**AGCGGATAACAATTTCACACAGGA
PhenotypeBC9	CTGCTGAACCGCTCTTCCGATCTNN**GGCTA**AGCGGATAACAATTTCACACAGGA
PhenotypeBC10	CTGCTGAACCGCTCTTCCGATCTNN**GAATG**AGCGGATAACAATTTCACACAGGA
PhenotypeBC11	CTGCTGAACCGCTCTTCCGATCTNN**CCAAC**AGCGGATAACAATTTCACACAGGA
PhenotypeBC12	CTGCTGAACCGCTCTTCCGATCTNN**GAGAC**AGCGGATAACAATTTCACACAGGA
PEprimer1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
PEprimer2	AAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

Table 1.4: Column barcoding primers used for the third PCR reaction in Illumina paired-end primers.

of true discordance between mRNA and protein expression. Clearly, mRNA expression does not always correlate with protein expression as many genes are subject to post-transcriptional regulation. Cytokine gene expression is subject to particularly complex regulation, including mechanisms affecting translation and/or mRNA stability(29). Because there is likely discordance between mRNA and protein expression within cells, the data we show on sensitivity and specificity should be used only as a guide (Supplementary Table 7).

Our strategy can also be customized or expanded. For example, the phenotyping panel can be customized to include additional genes of functional interest. In addition, because we can obtain the sequence of any given parameter, we can also design assays to measure genetic polymorphism, somatic mutation or splice variation of genes in single cells. Of course, because it is difficult to predict the cumulative affect of additional primers in a multiplexed PCR reaction, addition of parameters would likely require additional optimization and validation. But because the presence of additional transcripts does not affect the sensitivity of detection of a given transcript in our current panel of 17 different phenotypic parameters, (Supplementary Fig. 5), substantial expansion of this panel may be possible even with current sequencing technology.

Analysis of tumor-infiltrating lymphocytes To demonstrate one application of this method, we analyzed tumor-infiltrating lymphocytes (TILs) from a human colorectal cancer sample. Therapies designed to incite anti-tumor T-cell responses have recently shown great promise in the treatment of human cancer(31,32), and in colorectal cancer, the presence of TILs correlates strongly with a positive prognosis(33,34). To date, however, phenotypic characteristics and TCR sequences of TILs have generally been studied at the bulk population rather than at the single-cell level(33–36). Thus, there is some controversy over their function and clinical significance in different tumors37 and no consensus view on their specificity or functional properties.

We applied our methodology to 736 sorted, single CD4+ TILs from one patient volunteer who underwent a colectomy for stage T3N1 rectal adenocarcinoma. For comparison, we also analyzed 372 CD4+ T cells derived from resected adjacent colon tissue from the same patient, as well as peripheral blood T cells from a different

healthy donor. TCRβ sequences were successfully obtained from 597 of the 736 CD4+ T cells (81%), and we were able to assign productive, paired TCRαβ sequences to 503 of these (68% of total). In this particular tumor, we detected marked T-cell clonal expansion; the most frequent TCRβ was detected in 52/597 cells, and ten TCRβ sequences were detected in at least 8 cells (Supplementary Table 8). Out of 299 unique TCRβ sequences, the ten most frequent sequences accounted for 215/597 (36%) of the cells where sequences were recovered; 236 sequences (40%) were detected in only one cell (Supplementary Table 8).

Among the 372 CD4+ T lymphocytes derived from resected adjacent colon tissue from the same patient, we obtained TCRβ sequences from 309 cells (83%), and we were able to assign productive, paired TCRαβ sequences to 217 of these (58% of total). In contrast to the TCR repertoire from intratumoral T cells, clonal expansion was minimal, with only four TCR clones detected twice within the population (Supplementary Table 9). Also, not a single TCRαβ heterodimer sequence was shared between T cells in the tumor and adjacent colon tissue. This suggests that expanded T-cell clones present within tumors may be reacting to tumor antigens.

Next we searched for homology between TCR sequences in the intratumoral T-cell population to determine whether T-cell expansion was due to antigen-specific responses. The most highly expanded TCRβ (CASSLASMGVGELFF) sequence within our sample set varied by only two amino acids from another expanded TCRβ (CASSSASGGVGELFF) sequence. These TCRβ sequences comprised, respectively, 52 and 8 of 597 total T cells. These two expanded clones also used the same TCRα chain (CAYRPNYGGATNKLIF), although the TCRα chains used different nucleotide sequences in the two clones and were not present elsewhere within the sample set, indicating that this finding was not a result of cross-contamination (Supplementary Fig. 6 and Supplementary Table 8). Each T-cell clone also expressed a different non-productive TCRα chain. This finding further confirms that the common TCRα chain was indeed the TCRα chain that was used by these T-cell clones, because both clones expressed only one productive TCRα chain gene. Junctional diversity describes the process of random nucleotide addition (Nnucleotide addition) or subtraction at the junctions of V(D)J rearrangements, which markedly add to TCR diversity9. In both

T-cell clones, common sequence motifs within TCRα and TCRβ genes were generated as a result of significant N-nucleotide addition, indicating that these sequences would not be very common by chance (Supplementary Fig. 6). These findings strongly suggest that these two T-cell clones, which comprised over 10% (60/597) of the total CD4+ T cells we analyzed from this tumor, have been selected and activated by the same peptide-MHC ligand.

In addition to TCR sequencing, we phenotyped these cells using the 17-parameter panel described above. We stimulated half of the T cells for 3 h with PMA (phorbol 12-myristate 13-acetate) and ionomycin. Consistent with previous findings(38–41), the stimulated CD4+ TILs displayed a phenotype distinct from stimulated CD4+ T cells obtained from adjacent colon or peripheral blood (Fig. 3a and Supplementary Tables 10 and 11). A higher percentage of the stimulated TIL cells expressed RORC (146/279, 52%), IL-17 (184/279, 66%), TNF-α (217/279, 78%) and IFN-γ (148/279, 74%; Fig. 3b). To visualize the data, we used principal component analysis, which concentrates the most important sources of variation in larger data sets(2). This allows us to readily visualize the phenotypic diversity of CD4+ T cells (Fig. 3a and Supplementary Fig. 7). Although there is substantial overlap between the phenotypes of CD4+ T cells derived from tumor, colon and blood, these three populations of cells cluster discretely (Fig. 3a). Such phenotypic diversity is not as apparent in the absence of stimulation (Supplementary Fig. 8).

Although CD4+ TILs were largely distinguished from the other populations by their co-expression of IL-17, RORC, TNFα and IFN-γ, there was also notable heterogeneity within each T-cell population (Fig. 3b,c). Also, individual cells frequently co-expressed multiple, different, master-regulator transcription factors, showing that the categorization of CD4+ T cells into specific subsets is not always straightforward (Fig. 3b,c).

A major advantage of our methodology is that it enables us to compare the phenotypic and functional range of T cells that can arise from a single TCR clone. For instance, we observed that compared to unexpanded T cells, a significantly higher percentage of highly expanded ($\geq$10) T-cell clones expressed IL-17 (70/80 versus 65/126, $P < 0.005$) or RORC (50/80 versus 43/126, $P < 0.005$). Conversely, FOXP3

was less likely to be expressed in highly expanded versus unexpanded cells (5/80 versus 32/126, $P < 0.005$, Fig. 3b). When clustering analysis is applied, certain phenotypic clusters are preferentially occupied by unexpanded versus expanded cells or vice versa (Fig. 3c).

We also looked more closely at the FOXP3+ TILs (Fig. 4). The function of Treg cells in cancer has been the subject of much debate and FOXP3+ T-cell infiltration in tumors has been correlated with both favorable and poor prognoses38–41. Within this particular tumor, we found two distinct subsets of FOXP3+CD4+ T cells, differentiated by their expression of RORC (Fig. 4a). Within FOXP3+RORC+ cells, the overwhelming majority of cells expressed IL-17 (16/17, 94%), whereas IL-17 expression was rare within FOXP3+RORC− cells (3/28, 11%) (Fig. 4a). These two subsets also varied greatly with respect to the degree of clonal expansion. The FOXP3+RORC+ population consisted largely of clones that were expanded within our data set (12/17, 71%), whereas clonal expansion was rare in the FOXP3+RORC− population (1/28, 4%). Incidentally, the only FOXP3+RORC− T cell that was clonally expanded did express IL-17.

FOXP3+RORC+ IL-17–expressing T cells, described in both human colorectal cancer and in mouse models of polyposis, have been shown to have potent T-suppressive activity while being pro-inflammatory in their expression of IL-17 (refs. 39,40). Although the consequences of FOXP3+ T cell infiltration into tumors are unclear, the presence of IL-17 has been associated with tumorigenesis and poor prognosis(41–43). Based on this, IL-17 has been proposed as a therapeutic target. Both FOXP3+RORC+ T cells and FOXP3−RORC+ Th17-phenotype T cells may produce IL-17 within tumors, however, the relationship between these two populations of T cells is unclear. It has been proposed that they are unrelated given the discordance between their numbers within tumors(42).

To address this question, we searched for T cells that shared TCR$\alpha\beta$ sequences with FOXP3+RORC+ T cells within our data set. We found 61 instances of FOXP3− T cells sharing TCR sequences with FOXP3+RORC+ T cells (Fig. 4b). The majority of these FOXP3− T cells also expressed IL-17 (49/61, 80%) and/or RORC (39/61, 64%). These findings indicate that these two populations of IL-17–expressing T cells

share a common ancestry and are consistent with the idea that FOXP3+RORC+ T cells within tumors lose FOXP3 expression to become Th17 cells(44,45). The relationship between FOXP3+RORC− T cells and FOXP3+RORC+ T cells is not as clear. We cannot say whether the FOXP3+RORC+ T cells we observed originated as FOXP3+RORC− T cells which underwent clonal expansion. However, we did not detect TCR sequences shared between these two populations, and this suggests that FOXP3+RORC− T cells did not originate from clonally expanded FOXP3+RORC+ T cells.

Interestingly, both of the expanded TIL clones that express highly similar TCRs (Supplementary Fig. 6) contained cells expressing IL-17 and RORC. Among the 27 stimulated cells in the first TCRβ clone (CASSLASMGVGELFF), 24 expressed IL-17 and 16 expressed RORC. One cell co-expressed both FOXP3 and RORC. For the second TCRβ clone (CASSSASGGVGELFF), only two of eight sequences were present in the stimulated sample. Both of these T cells expressed IL-17 and one expressed RORC. Taken together, our data show clear heterogeneity between FOXP3+ T cells within tumors, which might help explain the why the data regarding the function of Treg cells in tumors have been controversial.

1.4.3 Discussion

The approach we describe here enables highly efficient TCR determination and multiparametric phenotypic analysis in single T cells. It requires no proprietary reagents or materials, and can be performed in any standardly equipped laboratory with access to flow cytometry and deep sequencing. To our knowledge, we describe the highest reported efficiency in sequencing TCRs from single T cells. Furthermore, our method is uniquely suited to identifying multiple TCRα chains from single T cells.

We demonstrate the utility of this technology by analyzing TILs, and show how TCR sequences can add an invaluable dimension to multiparametric phenotypic analysis by marking the ancestry of particular T cells, especially when the antigen is not known. For example, we show that FOXP3+RORC+ T cells and FOXP3−RORC+

Th17-phenotype cells can share a common ancestry (indicated by identical TCR sequence), indicating that the activation of a single T cell can lead to subsequent clonal expansion, loss of FOXP3 expression and differentiation to Th17-phenotype cells (Fig. 4d). Furthermore, we show an example of two expanded T-cell clones with highly homologous TCR sequences; among the two clonal populations are members expressing IL-17 and FOXP3. Because these two highly expanded T-cell clones appear to be responding to the same peptide-MHC, antigen-specificity is likely important to the selection of these T cells (Fig. 3d). More work is needed to understand the signals and antigens that lead to activation, loss of FOXP3 expression and clonal expansion of these IL-17–producing T cells within tumors. Also, TILs from colorectal cancer have been shown to be heterogeneous with respect to IL-17 secretion, so these results need to be validated on additional tumor samples40. But given the association of IL-17 with tumorigenesis and poor clinical outcomes, the event(s) responsible for initially activating these cells might represent an attractive therapeutic target.

This technology is also complementary to recently developed methods to determine ligands for TCRs using random peptide-MHC libraries(11). These complemetary approaches enable the identification of potentially important disease-associated TCRs and subsequent discovery of their antigens. Many human diseases including cancer, autoimmune disease and infectious diseases are characterized by T-cell activation and proliferation; what antigens are driving these T-cell responses may not be known. Identification of these antigens could be invaluable in understanding the origin of both beneficial and potentially detrimental T-cell responses and provide targets for therapeutic intervention.

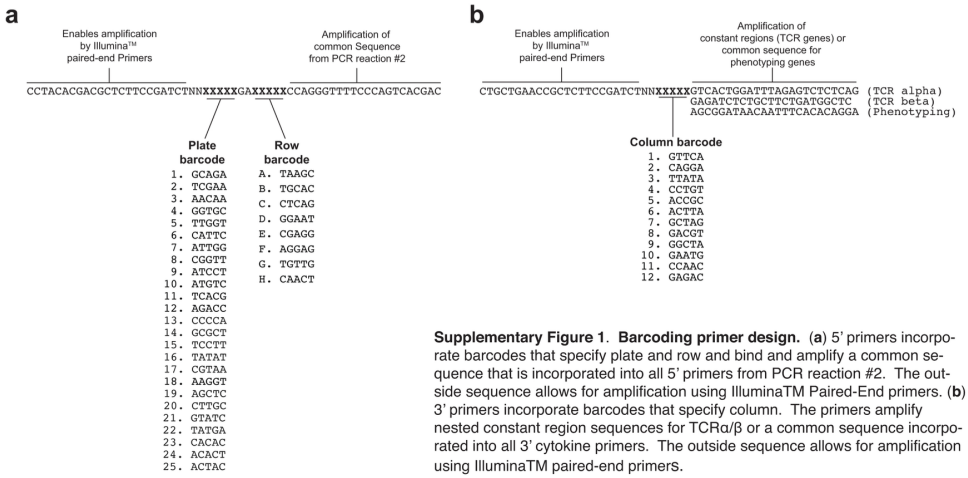
1.4.4 Methods

Single-cell sorting and flow cytometry. All fluorescence-activated cell sorting (FACS) experiments were performed on ARIA II instruments (Becton Dickinson) in the Stanford Shared FACS Facility. Cytokine capture assays (Miltenyi Biotec) were performed per manufacturer's instructions on freshly isolated human peripheral blood mononuclear cells (PBMCs). PBMCs were collected from a healthy donor who gave informed

consent. The Jurkat T-cell leukemia cell line (Clone E6-1) was obtained from ATCC. The following antibody clones were used for flow cytometry: anti-CD3 (SK7, BioLegend), anti-CD4 (RPA-T4, BioLegend), anti-CD8 (OKT8, eBiosciences), anti-αβTCR (IP26, BioLegend), anti-CD25 (2A3, Becton Dickinson) and anti-FOXP3 (PCH101, eBiosciences). Dead cells were excluded using a LIVE/DEAD Fixable Dead Cell Stain kit (Invitrogen).

TIL preparation. The Stanford University Institutional Review Board approved all protocols for collection of human tissue and blood. Tissue was collected with informed consent from a patient undergoing colon resection for colon cancer at Stanford University Hospital after initially being processed by the Department of Pathology. Tumor tissue was cut into small pieces and incubated in 10 mM EDTA (EDTA) in PBS for 30 min. Cells in suspension were passed through a 10-μM nylon cell strainer (Becton Dickinson). Tissue was then incubated in RPMI with 5% FCS containing 0.5 mg/ml of Type 4 collagenase for 30 min (Worthington Biochemical). Tissue was periodically disrupted during incubation by passing through a syringe topped with a blunt-ended 16-gauge needle. Lymphocytes were enriched through Percoll (GE Healthcare) gradient centrifugation. Cells were frozen in complete RPMI containing 10% DMSO (dimethylsulfoxide) and 40% FCS (FCS) for later use. Prior to use, cryopreserved lymphocytes were thawed and washed with complete RPMI before overnight recovery at 37 °C. Cells were transferred to tubes, washed and resuspended in cytometry buffer (PBS + 0.05% sodium azide + 2 mM EDTA + 2% FCS) for staining. For stimulation, cells were cultured for 3 h at approximately 15 × 106/ml in complete RPMI (10% FCS) and 150 ng/ml PMA + 1 μM ionomycin. At the end of the 3 h stimulation, cells were pipetted vigorously to remove adherent cells from the plate and transferred to tubes, washed and resuspended in cytometry buffer (PBS + 0.05% sodium azide + 2 mM EDTA + 2% FCS).

TCR sequencing and phenotyping. Single-cell sorting was performed using an ARIA II cell sorter (Becton Dickinson). TCR sequence and gene expression analysis from single cells were obtained by a series of three nested PCR reactions as described13. Cells were sorted directly into RT-PCR buffer. For the first reaction, reverse transcription and preamplification were performed with a One-Step RT-PCR



Supplementary Figure 1. **Barcoding primer design.** (**a**) 5' primers incorporate barcodes that specify plate and row and bind and amplify a common sequence that is incorporated into all 5' primers from PCR reaction #2. The outside sequence allows for amplification using IlluminaTM Paired-End primers. (**b**) 3' primers incorporate barcodes that specify column. The primers amplify nested constant region sequences for TCRα/β or a common sequence incorporated into all 3' cytokine primers. The outside sequence allows for amplification using IlluminaTM paired-end primerS.

Figure 1.14: Barcode architecture.

kit (Qiagen) using multiplex PCR with multiple Vα and Vβ region primers, Cα and Cβ region primers, and phenotyping primers in a 20-μl reaction (Supplementary Tables 1 and 2). For the PCR reaction #1, the final concentration of each TCR V-region primer is 0.6 μM, each C-region primer is 0.3 μM, each phenotyping primer is 0.1 μM. A 25-cycle first RT-PCR reaction was done per manufacturer's instructions using the following cycling conditions:50°C30min; 95°C 15 min; 94°C 30s, 62°C 1 min, 72°C 1 min × 25 cycles; 72 °C 5 min; 4 °C. Next, a 1-μl aliquot of the first reaction was used as a template for second 20-μl PCR using HotStarTaq DNA polymerase (Qiagen) for either TCR sequencing or phenotyping. The following cycling conditionswere:95°C15min;94°C30s, 64°C 1 min, 72°C 1 min × 25 cycles (TCR) or 35 cycles (phenotyping); 72 °C 5 min; 4 °C. For the TCR sequencing reaction, multiple internally nested TCRVα, TCRVβ, TCRCα and Cβ primers (Supplementary Table 1) were used (V primers 0.6 μM, C primers 0.3 μM). For the phenotyping reaction, multiple internally nested phenotyping primers (Supplementary Table 2) were used (0.1 μM). The second set of TCR V-region primers and 5´ phenotyping primers contained a common 23-base sequence at the 5´ end to enable further amplification (during the third reaction) with a common 23-base primer. The second set of 3´ phenotyping primers contained a common 24-base sequence to enable further amplification (during the third reaction). 1-μl aliquot of the second PCR product was used as a

template for the third 20-μl PCR reaction, which incorporates barcodes and enables sequencing on the Illumina MiSeq platform. For the third and final PCR reaction for TCR sequencing, amplification was performed with HotStarTaq DNA polymerase for 36 cycles using a 5′ barcoding primer (0.05 μM) containing the common 23-base sequence and a 3′ barcoding primer (0.05 μM) containing sequence of a third internally nested Cα and/or Cβ primer, and Illumina Paired-End primers (0.5 μM each, Supplementary Table 3). For tumor-infiltrating and colonic T-cell analysis, the final barcoding PCR reactions for TCR alpha and TCR beta were combined. When the third reaction was performed together, the 3′ Cα barcoding primer was used in threefold excess to the 3′ Cβ barcoding primer (0.15 μM and 0.5 μM, respectively). In addition to the common 23-base sequence at the 3′ end (that enables amplification of products from the second reaction) and a common 23-base sequence at the 5′ end (that enables amplification with Illumina Paired-End primers), each 5′ barcoding primer contains a unique 5-base barcode that specifies plate and a unique 5-base barcode that specifies row within the plate (Supplementary Fig. 1). These 5′ barcoding primers were added with a multichannel pipette to each of 12 wells within a particular row within a particular plate. In addition to the internally nested TCR C-region sequence and a common 23-base sequence at the 3′ end (that enables amplification with Illumina Paired-End primers), each 3′ barcoding primer contains a unique 5-nucleotide barcode that specifies column. These 3′ barcoding primers were added with a multichannel pipette to each of eight wells within a column within all plates. The third reaction for phenotyping was performed in a similar manner to the TCR sequencing, except that the 3′ primer contains the common 24-base sequence contained in all 3′ primers from the second reaction rather than the internally nested TCR C-region primer. The same 5′ barcoding primers were used for the third phenotyping reaction as the TCR sequencing reaction. After the third and final PCR reaction, each PCR product should have a unique set of barcodes incorporated that specifies plate, row and column and have Illumina Paired-End sequences that enable sequencing on the Illumina MiSeq platform. The PCR products were combined at equal proportion by volume, run on a 1.2% agarose gel, and a band around 350 to 380

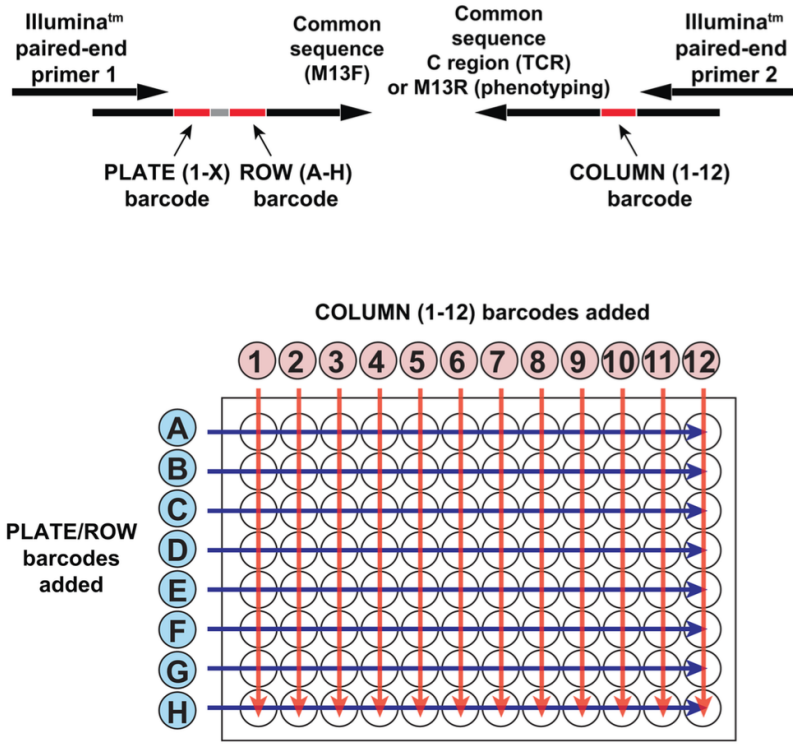


Figure 1.15: Schematic for barcoding (third) PCR reaction. An aliquot from the second PCR reaction product is used as a template for this reaction. To each well within a particular row within a given plate, a distinct 5' primer is added by multichannel pipette that specifies row. To each well within a column, a distinct 3' primer is added by multichannel pipette that specifies column. The reaction is performed with Illumina paired-end primers in all wells, which enable sequencing on the Illumina MiSeq platform.

bp was excised and gel purified using a Qiaquick gel extraction kit (Qiagen). This purified product was then sequenced.

PCR primer design. All primer sequences are provided in Supplementary Figure 1 and Supplementary Tables 1–3. All primers were designed to have a Tm of 70–72 °C (Tm = 4×[GC]+2[AT]). For TCR primers, base degeneracy was incorporated into the primers when necessary to account for TCR polymorphism and ensure amplification of all known functional Vα, Vβ, Cα and Cβ regions identified in the IMGT database (http://www.imgt.org/). V-region primers were designed to be at least 50 bases from the distal end to ensure inclusion of the entire CDR3 region. All

TCR and phenotyping primers for the second reaction contain the common sequence CCAGGGTTTTCCCAGTCACGAC at the 5´ end, which enables amplification with barcoding primers during the third reaction. All phenotyping primers for the second reaction contain the common sequence AGCGGATAACAATTTCACACAGGA at the 5´ end, which enables amplification with barcoding primers during the third reaction. After all reactions are performed, TCR primers amplify a segment of the TCR of approximately 250 bp. The final product for sequencing is approximately 380 bp. Phenotyping PCR primers were designed to span introns and amplify all major variants of the genes present in the NCBI database (http://www.ncbi. nlm.nih.gov). After the second reaction is performed, phenotyping primers amplify a gene segment of approximately 200 bp, and the final sequencing product is approximately 350 bp.

Sequencing data analysis. Raw sequencing data were processed and demultiplexed using a custom software pipeline to separate reads from every well in every plate according to specified barcodes. All paired ends are assembled by finding a consensus of at least 100 bases in the middle of the read. The resulting paired-end reads are then assigned to wells according to barcode. Primer dimers are filtered out by establishing minimum length of 100 bases for each amplicon. See Supplementary Note for further details on software. For example, in a recent sequencing run consisting of 2,164 cells, we obtained 2.01 $\times$ 10^7 raw reads, 1.95 $\times$ 10^7 pass-filtered reads (Illumina.com), forward/reverse consensus sequences were obtained and barcodes assigned to 1.66 $\times$ 10^7 reads, with 1.60 $\times$ 10^7 reads above 100 bases. The average read number per well was 7,382 $\pm$ 5,366. A consensus sequence is obtained for each TCR gene. Because multiple TCR genes might be present in a given well, our software establishes a cutoff of >95% sequence identity within a given well. All sequences exceeding 95% sequence identity are assumed to derive from the same TCR gene and a consensus sequence is determined. The 95% cutoff conservatively ensures all sequences derived from the same transcript would be properly assigned, even given a PCR rate of 1/9,000 bases, and sequencing error rate up to 0.4%22. TCR V, D and J segments were assigned by VDJFasta19. For phenotyping transcripts, the number of reads containing a 95% match to the customized database of transcription factor and cytokine genes are scored.

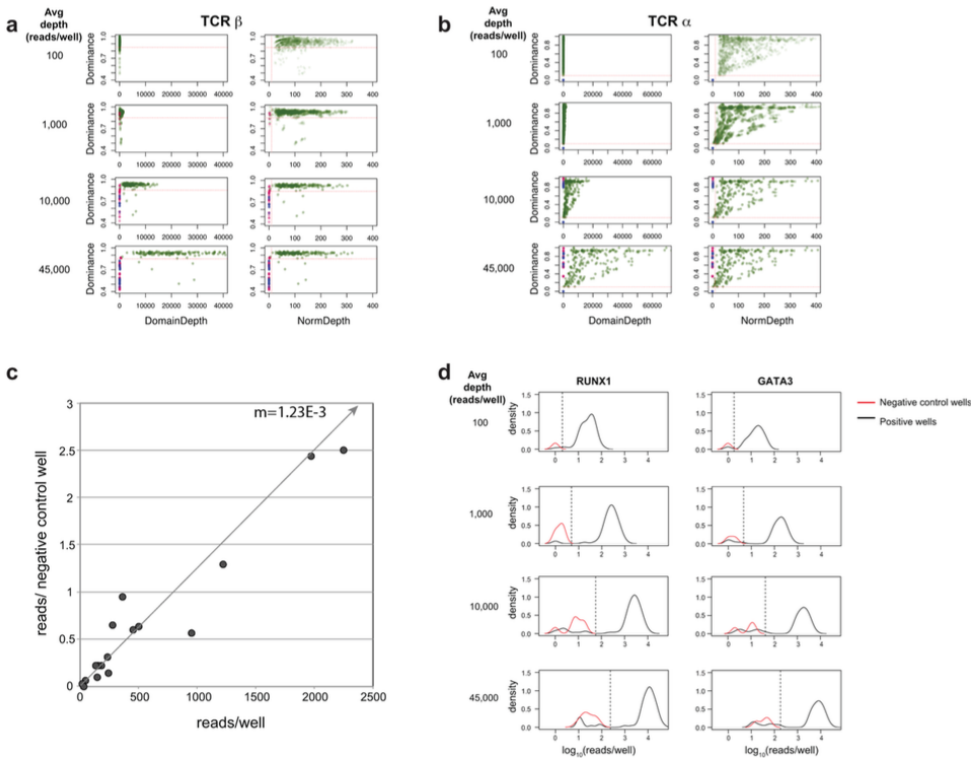


Figure 1.16: Validation of true-positive cutoff criteria by very deep sequencing (See online SI methods).

Single-well depth and dominance cutoff parameter validation. For both TCR and phenotypic parameters, there is a low background of unrelated sequences (Supplementary Fig. 3). We quantified potential background through highdepth sequencing and set thresholds accordingly. To quantify this background, sequencing was done at high depth. Single, PMA and ionomycin–stimulated T cells were sorted into two plates and processed for TCR and phenotypic analysis. Into these two plates, no cells were sorted into 16 wells, scattered through all columns and rows. Eight of these wells were processed normally with all reagents added. Eight of these wells were left completely blank throughout the analysis with no reagents added. These two plates (as opposed to the usual 20 to 25 plates) were run on a single sequencing run to give a sequencing depth more than tenfold higher than usual.

There was no significant difference in TCR background reads in negativecontrol wells without sorted T cells, regardless of whether wells were processed with reagents (Supplementary Fig. 3a,b). These data indicate that background is primarily due to error in PCR, sequencing or oligonucleotide synthesis within the barcodes and not due to cross-contamination.

For TCR reads within the two test plates, we validated cutoff criteria by simulated subsampling (Supplementary Fig. 3a,b). Plates were sequenced to an average depth of >45,000 reads per well, and subsampled to depths ranging from 100 to 45,000 average reads per well. By quantifying background signal (negative-control wells), we provide justification for thresholds set in the analysis. For TCR analysis, we establish a threshold normalized depth (based on average number of reads per well in the plate) of 10%. Using normalized depth independently, there is a clear separation between wells containing cells and background signal in negative-control wells at all depths down to 100 reads/well. For TCR analysis, establishing thresholds for clone dominance within the well further excludes the majority of negative-control wells and wells potentially containing more than one cell. For beta chains, a domain dominance cutoff is set at >85%. Domain dominance is determined based on 100% identity in sequence. Thus, this threshold of 85% is considerably lower than 100% because it accounts for the presence of PCR mutation or sequencing error. Because multiple

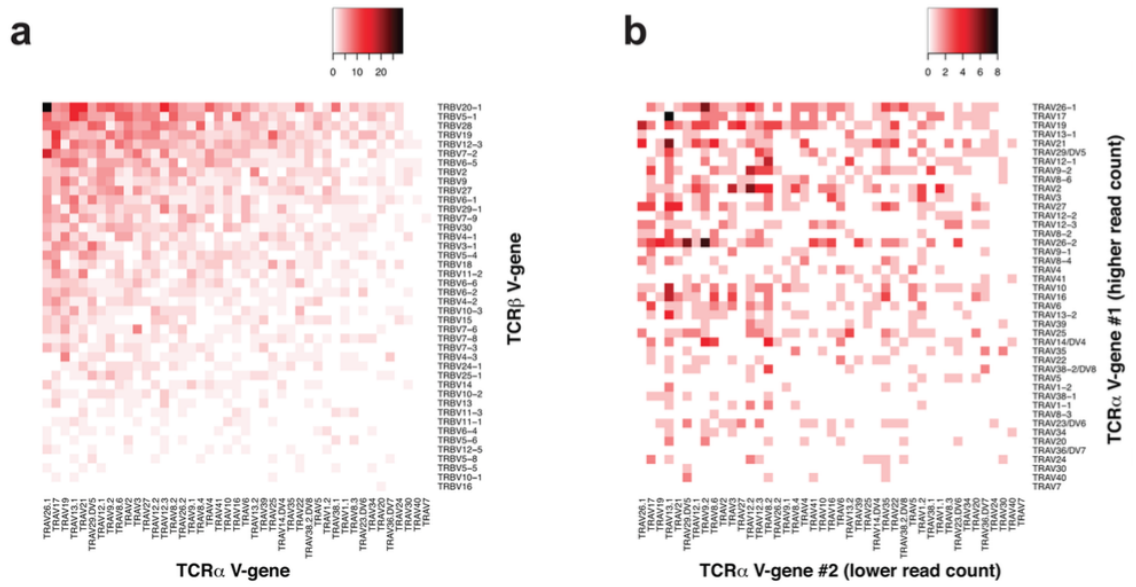


Figure 1.17: Human TCR V-gene usage in single T cells.

TCRα chains can exist within a given cell, the threshold for domain dominance is more permissive and set to 10%.

For phenotypic parameters, unlike for TCR genes, not all cells express a given parameter. Therefore, background is expected to depend upon number of cells expressing a given parameter as well as read depth. To investigate background for phenotypic parameters, we performed analysis on two plates; we sorted PMA plus ionomycin–stimulated IL-17+ single T cells into 40 wells, PMA plus ionomycin–stimulated IL-17− single T cells into 40 wells, and left 16 wells as negative-control wells. Eight of these negative-control wells were processed normally with all reagents added. Eight of these wells were left completely blank throughout analysis with no reagents added. IL-17+ and IL-17− T cells were sorted because this population gives a variable range of cells expressing all phenotypic parameters within the plate. We assessed background levels of each phenotypic parameter signal in negative-control wells. As was the case with TCR, there was no marked difference in background between negative wells processed with (0.54 background reads/well) or without reagents (0.72 background reads/well), suggesting that background is primarily due to error in

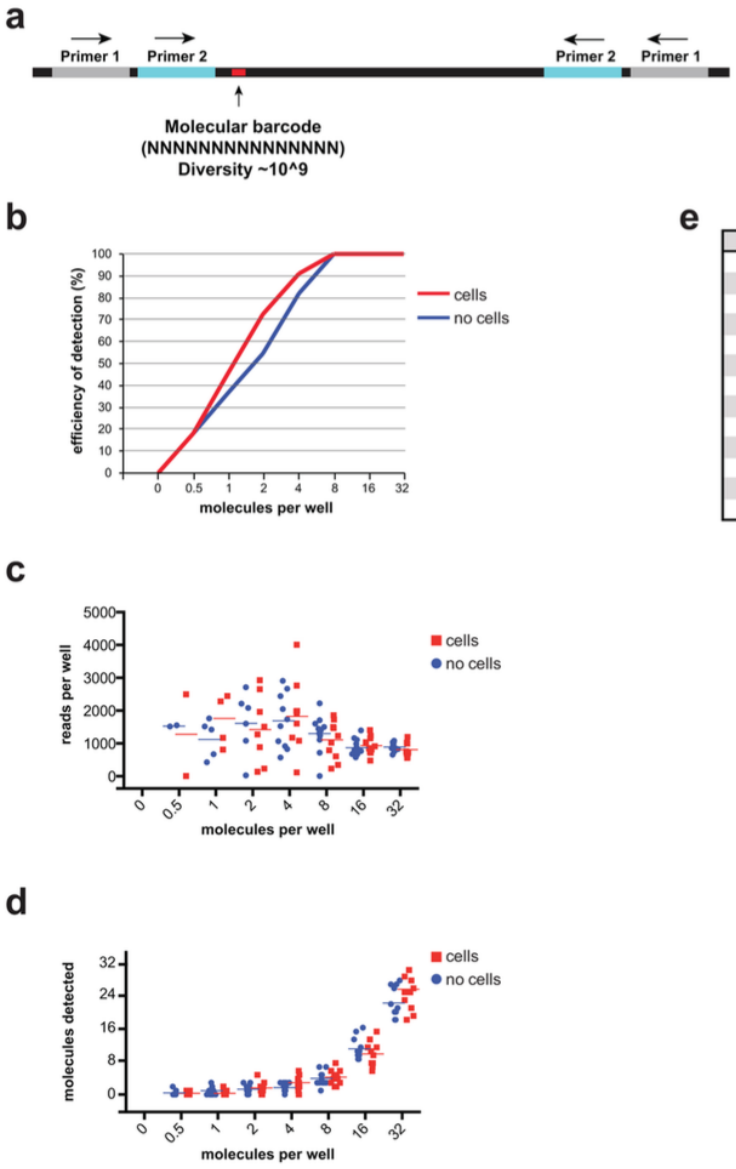
PCR, sequencing or oligonucleotide synthesis within the barcodes and not due to cross-contamination. Background was directly proportional to number of reads for each particular parameter on a plate and number of cells expressing a given parameter (Supplementary Fig. 3c). The ratio of reads/ negative-control well versus total reads/well for each phenotypic parameter in a given plate is approximately 1.23 $\times$ 1e$-$3. This ratio is constant, independent of the frequency of cells expressing a given parameter.

We then performed high-depth analysis on one plate containing 80 wells with single T cells and 16 negative-control wells to further investigate background per well. The plate was sequenced to an average depth of >45,000 reads per well, and subsampled to depths ranging from 100 to 45,000 average reads per well. We individually assessed the two phenotypic parameters with the highest level of background on this plate, RUNX1 and GATA3. For RUNX1 and GATA3, respectively, the ratio of reads/negative-control well versus total reads/well was 1.30 $\times$ 10$-$3 and 1.71 $\times$ 10$-$3 consistent with levels established in the analysis of the prior plate (Supplementary Fig. 3c). This indicates that relative background does not vary significantly, even at high-read depth. We assessed RUNX1 and GATA3 signal in 80 wells containing T cells and 16 negative-control wells (Supplementary Fig. 3d). Setting a threshold to 1 s.d. below the mean of log read counts per well (in all wells within a sequencing run expressing a given parameter) provides a scale-free means of conservatively excluding all background signals for phenotypic parameters. The accuracy of this threshold does not vary as a function of frequency of cells expressing the parameter, as only wells expressing a given parameter are included.

Sensitivity of detection. We further investigated the sensitivity of our method for detection of a particular transcript. A synthetic double-stranded (ds)DNA was constructed that contains binding sites for our IL-17 primers (Supplementary Fig. 5a). The construct is identical to the exogenous IL-17 amplicon except 15 nucleotides of endogenous IL-17 sequence is replaced with a 15-nucleotide random molecular barcode giving a theoretical diversity of >109 (415). This molecular barcode was incorporated to enable tracking of individual molecules during the analysis and further validate quantification (we should not be able to detect more molecular barcodes in a given

well than molecules added if quantification was accurate). This synthetic construct was made by PCR using a 124-base 5´ primer incorporating the primer sequences and the molecular barcode (5´ GCG TAA TAC GAC TCA CTA TAG GGA GAC AGA CAA GAA CTT CCC CCG GAC TGT GAT GGT CAA CCT GAA CAT CCA TAA CCG GAA CAT NNN NNN NNN NNN NNN CAA AAG GTC CTC AGA TTA CTA CAA C). To ensure that unique barcodes were not amplified, the template was first amplified by 60-cycle reaction using only the 5´ primer, and then 1 cycle was performed after addition of the 3´ primer. The PCR product was purified and quantified. The product was quantified by Nanodrop 2000 (Thermo Scientific) and Bioanalyzer 2100 (Agilent). Based upon these calculations, serial dilutions were performed and quantities were further verified by performing 50-cycle PCRs using primers within the template sequence. We spiked this synthetic construct into wells at different serial dilutions indicated and performed reactions and analysis on two plates. These two plates were processed identically, except a single, stimulated T cell was added to one of the plates. Into both plates, eight negative-control wells were processed without spiked template or cells.

We demonstrate that the method can detect as little as one molecule of dsDNA template (equivalent to two molecules of mRNA) (Supplementary Fig. 5b). Sensitivity improves with increased copy number and 100% sensitivity is achieved when 8 molecules of dsDNA (equivalent of 16 molecules of mRNA) are spiked into the initial reaction (Supplementary Fig. 5b). Although sensitivity does improve with increased copy number, read count per well does not significantly change (Supplementary Fig. 5c). This indicates that our readout is binary and read depth will not significantly affect sensitivity; in other words, sequencing at a higher depth will not improve identification of low-abundance transcripts in cells. Furthermore, the sensitivity of detection for one particular phenotypic parameter is not affected by the presence of other transcripts, as the sensitivity of detection for this template does not differ when stimulated T cells are added to the reaction and other amplified transcripts are present (Supplementary Fig. 5c). The number of molecular barcodes detected in a well is directly proportional to number of molecules added (Supplementary Fig. 5d). No molecular barcodes were repeated in different wells in our data set after accounting



GENE	READ/WELL
FOXP3	550
GATA3	478
GZMB	353
IFNG	605
IL17A	442
PRF1	283
RORC	974
TBET	389
TGFB1	156
TNF	524
BCL6	584
RUNX1	517
RUNX3	493

Figure 1.18: Analysis of the effect of transcript abundance on assay sensitivity and read count.

for background and the presence of PCR or sequencing error (Supplementary Fig. 5d).

Mean read counts per well for each phenotypic parameter did not vary markedly for phenotypic parameters present in at least 50 cells in our tumor and colon T-cell data set, which were sequenced to similar read depth (Supplementary Fig. 5e).

Single-cell sequencing accuracy. PCR error occurs at a rate of 1/9,000 bases and sequencing error has been reported to occur at a rate up to 0.4%22. Our method relies on generation of a consensus sequence from 10–10,000 reads, thus establishing single-cell transcript coverage far superior to that provided by genomic sequencing, mitigating the role of PCR error and largely eliminating sequencing error. To determine the accuracy of sequencing, we looked at the incidence of error within phenotyping transcripts that are entirely germline encoded, unlike TCR genes. When consensus sequence was obtained for all phenotyping transcripts within individual wells, the sequences were always identical. This indicates that despite sequencing or PCR error, the consensus sequence derived from a given well from >10 reads is 100% accurate within our data set.

Statistical analysis. To determine sensitivity (SN), specificity (SP), positive predictive value (PPV) and negative predictive value (NPV) (Supplementary Table 7), the following formulas were used: SN = TP/(TP + FN), SP = TN/(FP + TN),PPV=SN×PRV/[(SN×PRV+(1−SP)×(1−PRV)],NPV=SP×(1−PRV)/[(1−SN)×PRV+SP×(1−PRV)].TP=TruePositives,TN= True Negative, FP = False Positives, FN = False Negatives. PRV = Prevalence, determined by percentage of cells analyzed that were positive for the given parameter by flow cytometry. Chi-squared test was used to determine the statistical significance of skewing of phenotypic parameters within expanded versus unexpanded T-cell clones.

1.4.5 Acknowledgements

This work was made possible by my co-authors, including first author Arnold Han, as well as author Leo Hansmann and Mark M Davis. We thank members of the Davis laboratory and the Y.-H. Chien laboratory for helpful discussions. We thank

E. Newell for critical reading of the manuscript and helpful suggestions. We thank C. Bolen for assistance with data analysis. We thank X. Ji for deep sequencing. Tissue was obtained through the Stanford University Tissue Bank. The Stanford Shared FACS Facility provided use of equipment and the Stanford Functional Genomics Facility provided deep sequencing services. A.H. is supported by a grant from the Simons Foundation. L.H. is supported by a fellowship from the German Research Foundation (D.F.G.). M.M.D. is funded by the US National Institutes of Health and is an investigator of the Howard Hughes Medical Institute.

1.4.6 References

1. Wills QF, et al. Single-cell gene expression analysis reveals genetic associations masked in whole- tissue experiments. Nat Biotechnol. 2013; 31:748–752. [PubMed: 23873083]

2. Newell EW, Sigal N, Bendall SC, Nolan GP, Davis MM. Cytometry by time-of-flight shows combinatorial cytokine expression and virus-specific cell niches within a continuum of CD8 + T cell phenotypes. Immunity. 2012; 36:142–152. [PubMed: 22265676]

3. Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet. 2013; 14:618–630. [PubMed: 23897237]

4. Bendall SC, Nolan GP, Roederer M, Chattopadhyay PK. A deep profiler's guide to cytometry. Trends Immunol. 2012; 33:323–332. [PubMed: 22476049]

5. Spurgeon SL, Jones RC, Ramakrishnan R. High throughput gene expression measurement with real time PCR in a microfluidic dynamic array. PLoS ONE. 2008; 3:e1662. [PubMed: 18301740]

6. Wu AR, et al. Quantitative assessment of single-cell RNA-sequencing methods. Nat Methods. 2014; 11:41–46. [PubMed: 24141493]

7. Newell EW, Davis MM. Beyond model antigens: high-dimensional methods for the analysis of antigen-specific T cells. Nat Biotechnol. 2014; 32:149–157. [PubMed: 24441473]

8. Krogsgaard M, Davis MM. How T cells 'see' antigen. Nat Immunol. 2005; 6:239–245. [PubMed: 15716973]

9. Murphy, K.; Travers, P.; Walport, M.; Janeway, C. Janeway's Immunobiology. 8th. Garland Science; 2012.

10. Newell EW, et al. Combinatorial tetramer staining and mass cytometry analysis facilitate T-cell epitope mapping and characterization. Nat Biotechnol. 2013; 31:623–629. [PubMed: 23748502]

11. Birnbaum ME, et al. Deconstructing the peptide-MHC specificity of T cell recognition. Cell. 2014; 157:1073–1087. [PubMed: 24855945]

12. Hinrichs CS, Restifo NP. Reassessing target antigens for adoptive T-cell therapy. Nat Biotechnol. 2013; 31:999–1008. [PubMed: 24142051]

13. Han A, et al. Dietary gluten triggers concomitant activation of CD4 + and CD8 + alphabeta T cells and gammadelta T cells in celiac disease. Proc Natl Acad Sci USA. 2013; 110:13073–13078. [PubMed: 23878218]

14. Kim SM, et al. Analysis of the paired TCR alpha- and beta-chains of single human T cells. PLoS ONE. 2012; 7:e37338. [PubMed: 22649519]

15. Dash P, et al. Paired analysis of TCRalpha and TCRbeta chains at the single-cell level in mice. J Clin Invest. 2011; 121:288–295. [PubMed: 21135507]

16. Gascoigne NR, Alam SM. Allelic exclusion of the T cell receptor alpha-chain: developmental regulation of a post-translational event. Semin Immunol. 1999; 11:337–347. [PubMed: 10497088]

17. Malissen M, et al. Regulation of TCR alpha and beta gene allelic exclusion during T-cell development. Immunol Today. 1992; 13:315–322. [PubMed: 1324691]

18. Bentley DR, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008; 456:53–59. [PubMed: 18987734]

19. Glanville J, et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proc Natl Acad Sci USA. 2009; 106:20216– 20221. [PubMed: 19875695]

20. De Rosa SC, Herzenberg LA, Herzenberg LA, Roederer M. 11-color, 13-parameter flow cytometry: identification of human naive T cells by phenotype, function, and T-cell receptor diversity. Nat Med. 2001; 7:245–248. [PubMed: 11175858]

21. Yanagi Y, Chan A, Chin B, Minden M, Mak TW. Analysis of cDNA clones specific for human T cells and the alpha and beta chains of the T-cell receptor heterodimer from a human T-cell line. Proc Natl Acad Sci USA. 1985; 82:3430–3434. [PubMed: 3873654]

22. Nakamura K, et al. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res. 2011; 39:e90. [PubMed: 21576222]

23. Law JP, et al. The importance of Foxp3 antibody and fixation/permeabilization buffer combinations in identifying CD4 + CD25 + Foxp3 + regulatory T cells. Cytometry A. 2009; 75:1040– 1050. [PubMed: 19845018]

24. Vahedi G, Kanno Y, Sartorelli V, O'Shea JJ. Transcription factors and CD4 T cells seeking identity: masters, minions, setters and spikers. Immunology. 2013; 139:294–298. [PubMed: 23586907]

25. Oestreich KJ, Weinmann AS. Master regulators or lineage-specifying? Changing views on CD4 + T cell transcription factors. Nat Rev Immunol. 2012; 12:799–804. [PubMed: 23059426]

26. Wilson CB, Rowell E, Sekimata M. Epigenetic control of T-helper-cell differentiation. Nat Rev Immunol. 2009; 9:91–105. [PubMed: 19151746]

27. Collins A, Littman DR, Taniuchi I. RUNX proteins in transcription factor networks that regulate T-cell lineage choice. Nat Rev Immunol. 2009; 9:106–115. [PubMed: 19165227]

28. Assenmacher M, Lohning M, Radbruch A. Detection and isolation of cytokine secreting cells using the cytometric cytokine secretion assay. Curr Protoc Immunol. 2002; 4627:6.

29. Anderson P. Post-transcriptional control of cytokine production. Nat Immunol. 2008; 9:353–359. [PubMed: 18349815]

30. Fontenot JD, Gavin MA, Rudensky AY. Foxp3 programs the development and function of CD4 + CD25 + regulatory T cells. Nat Immunol. 2003; 4:330–336. [PubMed: 12612578]

31. Ribas A. Tumor immunotherapy directed at PD-1. N Engl J Med. 2012; 366:2517–2519. [PubMed: 22658126]

32. Sliwkowski MX, Mellman I. Antibody therapeutics in cancer. Science. 2013; 341:1192–1198. [PubMed: 24031011]

33. Pagès F, et al. Effector memory T cells, early metastasis, and survival in colorectal cancer. N Engl J Med. 2005; 353:2654–2666. [PubMed: 16371631]

34. Galon J, et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. Science. 2006; 313:1960–1964. [PubMed: 17008531]

35. Gerlinger M, et al. Ultra-deep T-cell receptor sequencing reveals the complexity and intratumour heterogeneity of T-cell clones in renal cell carcinomas. J Pathol. 2013; 231:424–432. [PubMed: 24122851]

36. Sherwood AM, et al. Tumor-infiltrating lymphocytes in colorectal tumors display a diversity of T cell receptor sequences that differ from the T cells in adjacent mucosal tissue. Cancer immunol immunother. 2013; 62:1453–1461. [PubMed: 23771160]

37. Sasada T, Suekane S. Variation of tumor-infiltrating lymphocytes in human cancers: controversy on clinical significance. Immunotherapy. 2011; 3:1235–1251. [PubMed: 21995574]

38. deLeeuw RJ, Kost SE, Kakal JA, Nelson BH. The prognostic value of FoxP3 + tumor-infiltrating lymphocytes in cancer: a critical review of the literature. Clin Cancer Res. 2012; 18:3022–3029. [PubMed: 22510350]

39. Scurr M, Gallimore A, Godkin A. T cell subsets and colorectal cancer: discerning the good from the bad. Cell Immunol. 2012; 279:21–24. [PubMed: 23041206]

40. Tosolini M, et al. Clinical impact of different classes of infiltrating T cytotoxic and helper cells (Th1, th2, treg, th17) in patients with colorectal cancer. Cancer Res. 2011; 71:1263–1271. [PubMed: 21303976]

41. Ladoire S, Martin F, Ghiringhelli F. Prognostic role of FOXP3 + regulatory T cells infiltrating human carcinomas: the paradox of colorectal cancer. Cancer immunol immunother. 2011; 60:909– 918. [PubMed: 21644034]

42. Blatner NR, et al. Expression of RORgammat marks a pathogenic regulatory T cell subset in human colon cancer. Sci Transl Med. 2012; 4:164ra159.

43. Gounaris E, et al. T-regulatory cells shift from a protective anti-inflammatory to a cancer- promoting proinflammatory phenotype in polyposis. Cancer Res. 2009; 69:5490–5497. [PubMed: 19570783]

44. Miyara M, et al. Functional delineation and differentiation dynamics of human CD4 + T cells expressing the FoxP3 transcription factor. Immunity. 2009; 30:899–911. [PubMed: 19464196]

45. Zhou L, Chong MM, Littman DR. Plasticity of CD4 + T cell lineage differentiation. Immunity. 2009; 30:646–655. [PubMed: 19464987]

1.4.7 Copyright

1.5 pMHC repertoire sequencing

In the following study, Birnbaum et al developed a yeast-displayed pMHC library for characterizing TCR-pMHC interfaces and de-orphaning TCR specificity. Due to the constraint of the single variable peptide and a sandwiched molecular interaction, the system provided powerful early datasets to validate some of the mathematical methods relevant to the convergence analysis framework, greatly expanded in Chapter 2.

In order to survey a universe of major histocompatibility complex (MHC)-presented peptide antigens whose numbers greatly exceed the diversity of the T cell repertoire, T cell receptors (TCRs) are thought to be cross-reactive. However, the nature and extent of TCR cross-reactivity has not been conclusively measured experimentally. We developed a system to identify MHC-presented peptide ligands by combining TCR selection of highly diverse yeast-displayed peptide-MHC libraries with deep sequencing. Although we identified hundreds of peptides reactive with each

of five different mouse and human TCRs, the selected peptides possessed TCR recognition motifs that bore a close resemblance to their known antigens. This structural conservation of the TCR interaction surface allowed us to exploit deepsequencing information to computationally identify activating microbial and self-ligands for human autoimmune TCRs. The mechanistic basis of TCR cross-reactivity described here enables effective surveillance of diverse self and foreign antigens without necessitating degenerate recognition of nonhomologous peptides.

1.5.1 Introduction

T cells are central to many aspects of adaptive immunity. Each mature αβ T cell expresses a unique ab T cell receptor (TCR) that has been selected for its ability to bind to peptides presented by major histocompatibility complex (MHC) molecules. Unlike antibodies, TCRs generally have low affinity for ligands (dissociation constant [K D] ~ 1–100 m M), which has been speculated to facilitate rapid scanning of peptide-MHC (pMHC) (Matsui et al., 1991; Rudolph et al., 2006; Wu et al., 2002). Structural studies of TCR-pMHC complexes have revealed a binding orientation where, generally, the TCR CDR1 and CDR2 loops make the majority of contacts with the tops of the MHC helices, whereas the CDR3 loops, which are conformationally malleable, primarily engage the peptide presented in the MHC groove (Davis and Bjorkman, 1988; Garcia and Adams, 2005; Rudolph et al., 2006). The low affinity and fast kinetics of TCR-pMHC binding, combined with conformational plasticity in the CDR3 loops, would seem to facilitate cross-reactivity with structurally distinct peptides presented by MHC (Mazza et al., 2007; Reiser et al., 2003; Yin and Mariuzza, 2009). Indeed, given that the calculated diversity of potential peptide antigens is much larger than TCR repertoire diversity, TCR cross-reactivity appears to be a biological imperative (Mason, 1998; Sewell, 2012). Cross-reactive TCRs have been implicated in both pathogenic and protective roles for a number of diseases (Benoist and Mathis, 2001; De la Herran-Arita et al., 2013; Shann et al., 2010; Welsh et al., 2010; Wucherpfennig and Strominger, 1995).

Nevertheless, the true extent of TCR cross-reactivity, and its role in T cell immunity, remains a speculative issue, largely due to the absence of quantitative experimental approaches that could definitively address this question (Mason, 1998; Morris and Allen, 2012; Shih and Allen, 2004; Wilson et al., 2004; Wucherpfennig et al., 2007). Although many examples exist of TCRs recognizing substituted or homologous peptides related to the antigen (Borbulevych et al., 2009, 2011; Krogsgaard et al., 2003), such as altered peptide ligands (Kersh and Allen, 1996), most of these peptides retain similarities to the wildtype (WT) peptides and are recognized in a highly similar fashion. Only a handful of defined examples exist of a single TCR recognizing nonhomologous sequences (Adams et al., 2011; Basu et al., 2000; Colf et al., 2007; Ebert et al., 2009; Evavold et al., 1995; Lo et al., 2009; Macdonald et al., 2009; Nanda et al., 1995; Reiser et al., 2003; Zhao et al., 1999).

One approach that has been used to estimate cross-reactivity utilizes pooled, chemically synthesized peptide libraries (Hemmer et al., 1998b; Wilson et al., 2004; Wooldridge et al., 2012). Usingcalculations basedupon thistechnique, it hasbeenextrapolated that ~10 6 different peptides in mixtures containing ~10 12 different peptides were agonists (Wilson et al., 2004; Wooldridge et al., 2012). Synthetic peptide libraries have been used to isolate diverse peptide sequences (Hemmer et al., 1998a), including microbialandself-ligandsforTCRsofinterest(Hemmeretal.,1997). However, most studies find only close homologs to known peptides (Krogsgaard et al., 2003; Maynard et al., 2005; Wilson et al., 1999, 2004). Furthermore, these cross-reactivity estimates were derived from the bulk stimulatory ability of libraries possessing femtomolar concentrations of any given peptide and no knowledge of peptide loading in the MHC or pMHC binding to the TCR. A more accurate estimate of cross-reactivity requires the isolation of individual sequences from a library of MHC-presented peptides based upon binding to a TCR.

Recently, we and others have created libraries of peptides linked to MHC via yeast and baculovirus display as a method to discover TCR ligands through affinity-based selections that rely on a physical interaction between the pMHC and the TCR (Adams et al., 2011; Birnbaum et al., 2012; Crawford et al., 2004, 2006; Macdonald et al., 2009; Wang et al., 2005). However, these methods have so far not been used

to address the broader question of TCR cross-reactivity, given that the requirement of manually validating and sequencing individual library "hits" has restricted the approach to discovering small numbers of peptides.

Here, we combined affinity-based selections of pMHC yeast libraries and deep sequencing to discover hundreds of unique peptide sequences recognized by multiple murine and human TCRs. Strikingly, all peptide sequences bear TCR epitopes with close similarity to their previously known agonist antigens. With an understanding of this property, we created a computational algorithm to predict naturally occurring TCR ligands using data from our deep-sequencing results. We tested a diverse set of the putative TCR-reactive peptides and found that 94% are able to elicit a T cell response. In general, TCR cross-reactivity does not appear to be characterized by broad degeneracy but rather is largely constrained to a small number of TCR contact residue "hot spots" on a peptide, while tolerating extensive diversity at other positions. This more granular understanding of the properties of TCR cross-reactivity has broad implications for ligand identification, vaccine design, and immunotherapy.

1.5.2 Results

Development and Selection of a Murine MHC Platform for Yeast Display We developed a system for the rapid and sensitive detection of TCR-binding peptides presented by the murine class II MHC I-E k . This represents an advance over previous reports of class II pMHC molecules displayed on the surface of yeast that either did not show or were not tested for the ability to bind soluble TCR (Birnbaum et al., 2012; Boder et al., 2005; Esteban and Zhao, 2004; Jiang and Boder, 2010; Starwalt et al., 2003; Wen et al., 2008, 2011). We designed our construct as a "mini" single-chain MHC Aga2 fusion, with the truncated peptide-binding b 1 a 1 domains fused and the WT peptide Moth Cytochrome C (MCC, residues 92–103) fused to the N terminus of the MHC b 1 domain via Gly-Ser linkers (Figure 1 A) (Adams et al., 2011). The initial construct was correctly routed to the yeast surface but did not have the ability to bind to TCR, indicating that the pMHC was not correctly folded (Figure 1 B).

In order to rescue folding of the pMHC, we subjected the mini I-E k to error-prone mutagenesis combined with introduction of solubility-enhancing mutations. We selected this mutagenized mini scaffold for binding to the 2B4 TCR, which recognizes MCC-I-E k with moderate affinity and slow kinetics (Newell et al., 2011). Our final construct contained solubilizing mutations in what was previously the a 1 b 1a 2 b 2 domain interface and one mutation between the MHC helix and the beta sheets (Figure 1 B). None of the mutated MHC residues contacted either the peptide or the TCR. The evolved construct retained specific binding to severalMCC-I-E k -recognizingTCRsandshowedcomparableaffinity to the WT pMHC (Figure 1 B and Figure S1 available online). We then created a peptide library tethered to the MHC construct. Based upon the recently solved 2B4-MCC-I-E k structure (Newell et al., 2011), we mutagenized the peptide from P(-2) to P10 (Figure 1 C). Limited diversity was introduced at the two most distal residues and the primary MHC-binding anchor residues at P1 and P9 to maximize the number of peptides capable of being correctly displayed by the MHC (Figure 1 C).

Our first attempts at screening involved "manual curation" of selections conducted with multivalent TCR. The library showed enrichment after three rounds of selection using highly avid TCR-coated streptavidin beads followed by a higher stringency "polishing" round of selection using TCR tetramers. The three peptides that were recovered via sequencing of 12 individual, hand-picked clones after selection were related to the WT MCC peptide—the P2, P5, and P8 TCR contacts were all conserved, whereas P3 showed a conservative Tyr-to-Phe mutation (Figure 1 D). We surmised that these enriched WT-like sequences present in the later rounds dominated the selections, preventing alternative, potentially nonhomologous sequences from being recovered. For this reason, we turned to deep sequencing at each step of the selection process to recover all enriched peptides.

Deep Sequencing of Selections for TCR-Binding Peptides Analysis of the pooled yeast library DNA after each successive round of selection with 2B4 via deep sequencing showed enrichment from an essentially random distribution of amino acids to a highly WT-like TCR recognition motif (Figures 2 A and S2 A). After the third round, there were nonhomologous amino acids at P5 and P8 selected above background (Met

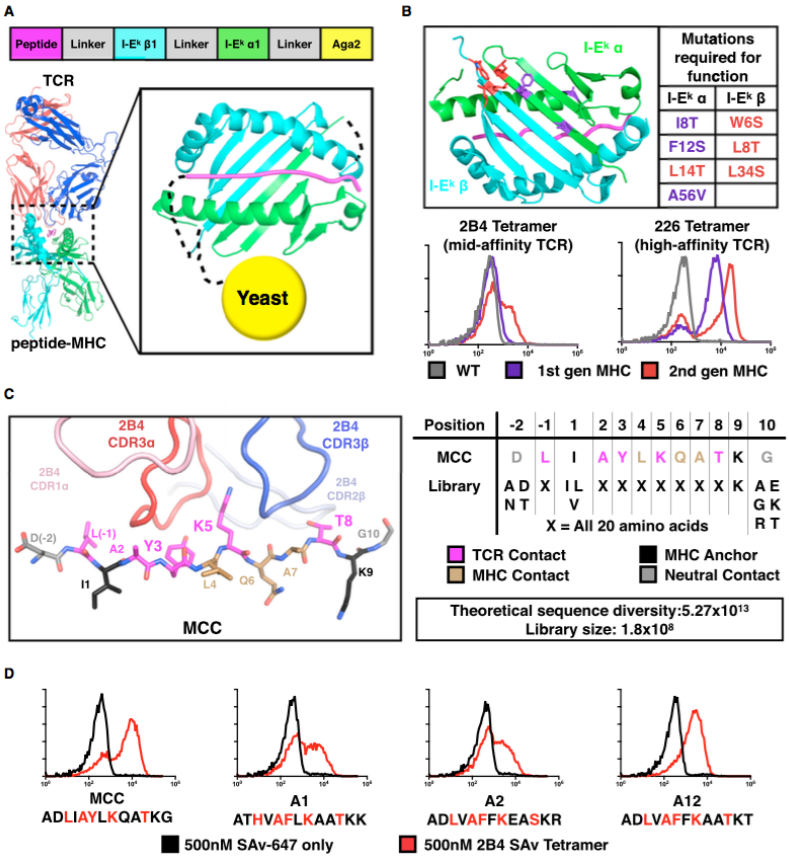


Figure 1.19: Library Design and Selection of I-Ek, a Murine Class II MHC Molecule (A) Schematic of the murine class II MHC I-Ek displayed on yeast, as b1a1 mini MHC with peptide covalently linked to MHC N terminus. (B) Mutations required for correct folding of the b1a1 mini I-Ek (top). Mutations were derived from error-prone mutagenesis and selection (purple) and rational design (red). Staining with 2B4 and 226 tetramers demonstrates function of error-prone-only construct (1st gen MHC) as well as error-prone + designed mutant construct (2nd gen MHC) (bottom). (C) Design of the peptide library displayed by I-Ek. Design is based upon the structure of 2B4 bound to MCC-I-Ek (left). Residues from P(2) to P10 are randomized, with limited diversity at P(2), P10, and the P1/P9 anchors (right). (D) 500 nM TCR tetramer staining of three clones selected for binding to 2B4 TCR compared to MCC (WT). TCR contact residues are colored red. See also Figure S1.

and Ser for P5, Ile and Leu for P8) that were outcompeted by the WT-like motif by the final round of selection. Overall, the number of unique peptides observed via deep sequencing progressed from 132,000 unique in-frame peptides observed in the sequenced portion of preselection library to only 207 unique peptides after the 3 rd round of selection (Figures 2 B, 2C, S2 A, and S2B). By the final round of selection, the library was dominated by a handful of sequences, matching the result obtained by manual curation (Figures 1 D, 2 B, and 2C).

We repeated the selections with two other TCRs reactive to MCC-I-E k : 226 and 5cc7. We analyzed enrichment for each TCR after the third round of selection, where there is enrichment for a binding motif but before complete convergence to a small number of sequences (Figures 2 A, 3 A, S2 B, and S3 A). Although all three TCRs retain a WT-like TCR recognition motif (indicated by the outlined boxes in the heatmaps), each TCR also shows some variation in positional preferences (Figure 3 A). For example, where 2B4 can recognize P5 Met and Ser, 5cc7 can accommodate P5 Leu, Val, and Arg. The P3 TCR contact position showed the least variance across all three TCRs, with either Phe or Tyr being required for 2B4 and 5cc7, and Phe, Tyr, or Trp being required for 226 (Figure 3 A). 226, as previously reported, showed a greater degree of cross-reactivity, able to recognize 897 unique peptide sequences. The larger number of peptides recognized was largely a function of a higher tolerance for substitutions on TCR-neutral and MHC-contacting residues, such as at positions P(-1) and P4 (Figures 3 A and S3 A) (Ehrich et al., 1993; Newell et al., 2011).

The large collection of peptides recovered via deep sequencing enabled us to apply covariation analysis to discover intrapeptide structure-activity relationships that were not previously accessible with traditional single-residue substitution analysis (Figure 3 B) (Ehrich et al., 1993; Newell et al., 2011; Reay et al., 1994; Wilson et al., 1999). By using covariation analysis of the central P5 residue and the C-terminal P8 residue, a pattern emerged: the native, MCC-like "up-facing" TCR-contact motifs for each TCR (P5 Lys, P8 Ser/Thr) were strongly correlated, whereas the altered residues (P5 Ser/P8Leu for 2B4, P5 Leu or Arg/P8 Phe for 5cc7) independently segregated (Figure 3 B). These results highlight a degree of cooperativity in the composition of residues comprising a "TCR epitope" that is clearly revealed with deep sequencing.

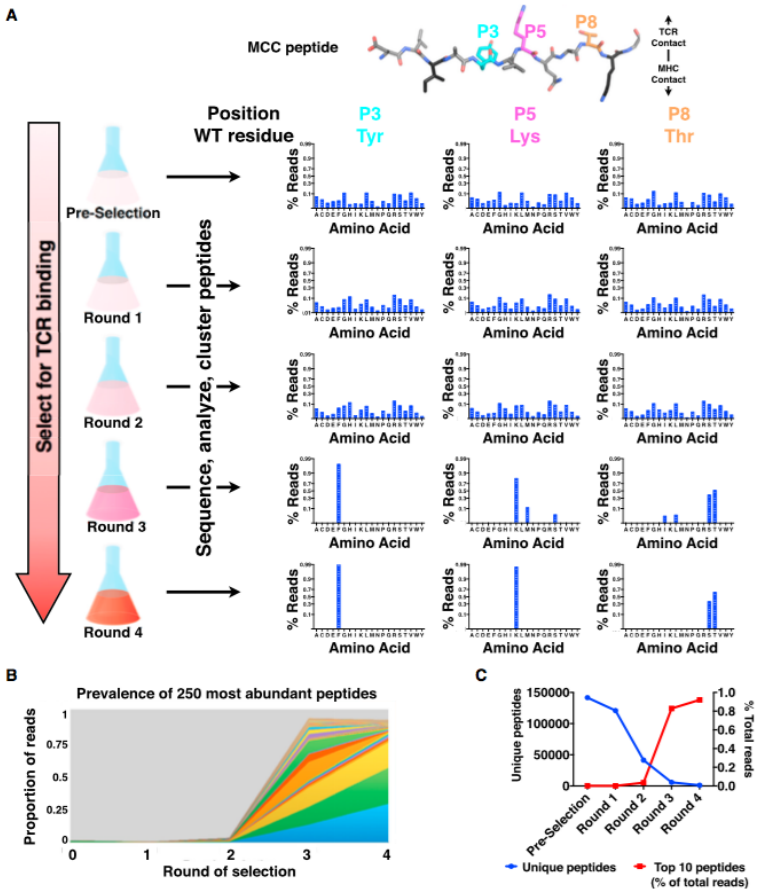


Figure 1.20: Deep Sequencing of Peptide Selections on I-Ek Converges on One Dominant Epitope for 2B4 TCR Recognition (A) Plots for amino acid prevalence at the three primary TCR contact positions (P3 [cyan], P5 [magenta], and P8 [orange]) show that the peptide library enriches from even representation of all amino acids in the preselection library to a WT-like motif at each position. A secondary preference can be seen at P5 and P8 in round 3 but is outcompeted by round 4. (B) Sequence enrichment of 250 most abundant peptides shows a convergence from a broad array of sequences to a few clones. Area in gray represents all clones other than the most prevalent 250. (C) Comparison of total number of peptides and prevalence of 10 most abundant peptides for each round of selection. See also Figure S2.

Furthermore, such intrapeptide residue coupling reveals how cross-reactivity can occur through mutually compensatory substitutions to the parent peptide.

Although the selected ligands for all three TCRs possessed shared features, each TCR also selected for a subset of sequences that were not selected by the other two. We applied distance clustering to the peptides selected by all three TCRs to determine whether all selected sequences were part of the larger MCC-like peptide family or were distinct families (Figure 3 C). We found that although sequences recognized by individual TCRs clustered more closely to each other, essentially all of the selected sequences formed one large cluster of peptides no more than three amino acids different than at least one other peptide in the cluster (Figures 3 C and S3 B). Therefore, the selected peptides for all three TCRs are related via a common specificity domain and, importantly, to the parent MCC ligand. Even though we conducted unbiased selections of random libraries, the only ligands that were recovered were remarkably similar to the WT ligand at the TCR interface.

Functional Characterization of I-E k Library Hits We synthesized 44 of the library peptides selected for binding to the TCRs and examined their ability to stimulate T cell blasts from 2B4 and 5cc7 transgenic mice as assayed by CD69 upregulation and IL-2 production. The majority of the peptides predicted to bind 2B4 (19/19) and 5cc7 (17/21) expressing T cells induced CD69 upregulation (Figures 4 A, 4B, and S4 A–S4D). The peptides had a wide range of potencies, including 50-fold more potent than the WT peptide MCC (colored red). When we compared the presence of the MCC-like TCR recognition epitope with TCR signaling, we found that in general, sequences that shared the MCC-like epitope at all three major TCR contacts (colored blue) were more potent in inducing signaling than those peptides that were more distantly related (colored black) (Figures 4 A and 4B). We also tested the peptides selected for binding to one TCR for their ability to cross-react with the other MCC-reactive T cells. Surprisingly, a large proportion of these peptides potently activated TCR signaling (Figures 4 A, 4B, and S4 A–S4D). In general, thesequencesthat showed themostrobustactivationwere again the ones that most closely shared the MCC TCR-binding epitope.

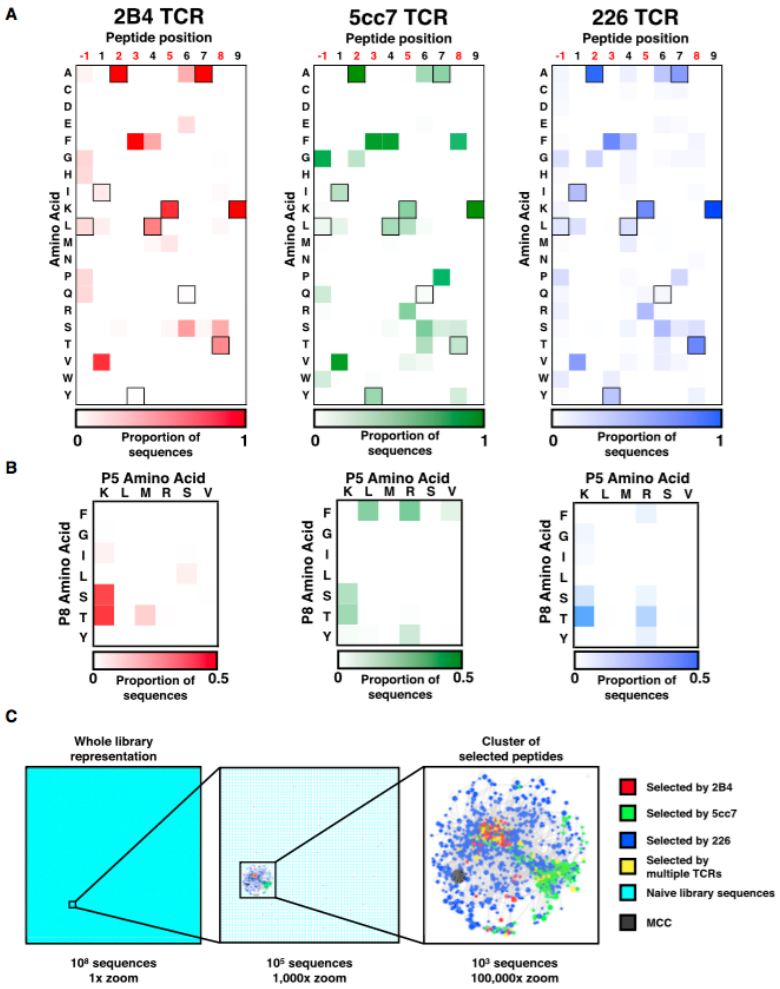


Figure 1.21: Three Different MCC/I-Ek-Reactive TCRs Require a WT-like Recognition Motif in the Peptide Antigens (A) Heatmaps of amino acid preference by position for 2B4 (left, red) 5cc7 (center, green), and 226 (right, blue) TCRs after three rounds of selection. The sequence for MCC is represented via outlined boxes. TCR contact residues are labeled red on the x axis. (B) Covariation analysis of TCR contact positions P5 (x axis) and P8 (y axis) show distinct coupling of amino acid preferences. (C) Minimum distance clustering of all TCR sequences selected above background shows that sequences for all TCRs form one large cluster with MCC (black circle, not represented in library but added for reference). Sequence cluster is placed in a representation of whole-library sequence space (left: 13 magnification, center: 10003 magnification) for reference.

We additionally chose nine peptides from our initial set of 44 and exchanged them into soluble I-E k MHC for TCR affinity measurements via surface plasmon resonance (SPR). For 2B4 and 5cc7, TCR bound the pMHC of interest with affinities ranging from K D of ˜1 m M (over 10-fold better than that of MCC) to those with binding only barely detectable at 100 m M TCR (Figures S4 E and S4F). When we compared the activity and affinity of our selected peptides, there is a loose but positive correlation between strength of TCR-pMHC binding and potency of activation (Figure 4 C).

The Structural Basis of TCR Recognition of Cross-Reactive Peptides Todeterminethemolecularbasisofthe TCRs'abilitytorecognize the most diverse peptides selected from our I-E k libraries, we determined the crystal structures of 2B4 in complex with a peptide termed 2A bound to I-E k as well as 5cc7 in complex with two peptides bound to I-E k , termed 5c1 and 5c2 (Table S1 ; Figures 5 A and 5B). When these complexes were aligned with the previously solved TCR-pMHC complex structures (2B4 and 226 binding MCC-I-E k), very little deviation in overall TCR-pMHC complex geometry was observed (Figures 5 A and 5B) (Newell et al., 2011). Because the 5cc7-MCC-I-E k complex is not solved, 5c1 and 5c2 were compared to 226-MCC-I-E k , which shares the TCR b chain with 5cc7 and therefore likely retains a close footprint. The contacts between TCR germline-derived CDR1/2 loops and MHC helices, which make up roughly 50% of the binding interface between TCR and pMHC, were essentially unchanged in the new peptide complexes versus MCC (Figure 5 C). When we examined the chemistry of MCC versus 2A and MCC versus 5c1 peptide recognition by their respective TCRs, we saw that the interactions between the TCR a CDR loops and the N-terminal halves of the peptides are essentially invariant (Figures 5 A and 5B, lower panels). Each peptide backbone makes a hydrogen bond at the P3 carbonyl with Arg29 a in the TCR CDR1 a loop (Figure S5 A). The contacts of 2B4 CDR3 a with P2 and P3 in MCC and 2A are essentially identical (Figure 5 A, lower panels). Although an exact analogy cannot be made between 5cc7 recognizing 5c1 and 226 recognizing MCC due to sequence differences in their CDR3 loops, 5cc7 and 226 CDR3 a loop conformations and peptide contacts are extremely similar (Figure 5 B, lower panels).

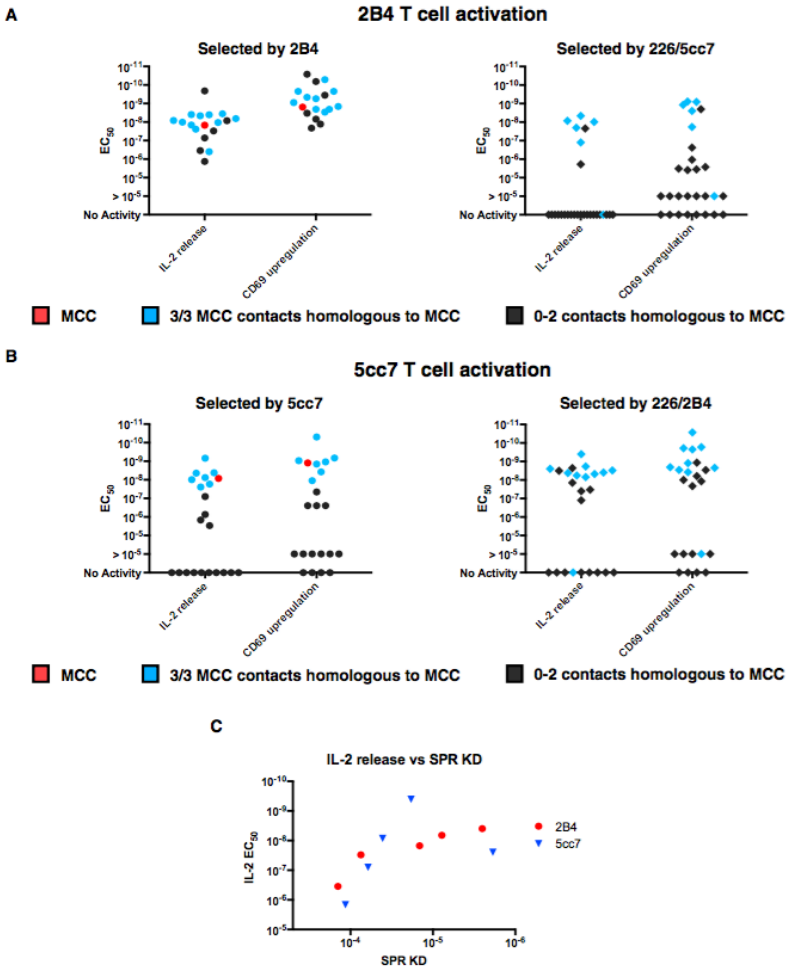


Figure 1.22: Relationships between Affinity and Activity of Peptides Selected for Binding to I-Ek-Reactive TCRs (A) EC50s of IL-2 release and CD69 upregulation for 2B4 T cells stimulated with peptides selected with 2B4 TCR, plus MCC (red) (left), or peptides selected with 226 or 5cc7 TCRs (right). Sequences with close homology to MCC at P3, P5, and P8 are represented in blue. Sequences that do not share 3/3 TCR contacts with MCC are in black. (B) EC50s as in (A) for 5cc7 T cells with peptides selected with 5cc7 (left) or 226/2B4 (right) TCRs. (C) Correlation between TCR-pMHC affinity and peptide signaling potency. Each data point represents one peptide.

In contrast, 2B4 and 5cc7 b chain CDR loop interactions with the C-terminal halves of the peptides show marked changes to accommodate the non-MCC sequences. For 2B4, the CDR3 b loop conformation completely rearranges to engage the alternate P5 and P8 residues on the 2A peptide (Figure 5 A, lower panels). Gln100 b , a residue that makes no contact with the peptide in the 2B4-MCC-I-E k complex, flips its side chain by 180 degrees to form hydrogen bonds with the peptide backbone carbonyl oxygens at P5 and P6 (Figure 5 A, lower panels). The side chains of Trp98 b and Ser99 b form hydrogen bonds with the P5 Ser hydroxyl moiety (Figure 5 A). Asp101 b , one of the main contacts with P5 Lys in MCC, forms a hydrogen bond with Ser95 b on the other end of the CDR3 b loop, significantly altering the overall topology of the loop (Figure S5 B).

In the 5cc7-5c1-I-E k complex, there are fewer hydrogen bonds formed between the peptide and TCR due to the replacement of P5 Lys with Leu in the 5c1 peptide (Figure 5 B, lower panels). Asn98 b changes its hydrogen-bonding network from engaging only the carbonyl of P6 on the MCC peptide backbone to simultaneously interacting with the carbonyl oxygen of P6 and the amide nitrogen of P8 of the 5c1 peptide (Figure 5 B). The second 5cc7-reactive peptide, 5c2, is recognized essentially identically by 5cc7 as 5c1 despite the substitution of P5 to Arg (Figure S5 C). The substitution of a bulkier side chain at P8 (Phe instead of Thr) results in a rocking of 5cc7 such that the TCR C b FG loop is translated by 15 A $^{\circ}$ relative to the 226-MCC structure (Figures S5 D and S5E). It is interesting to note that all tested peptides with P8 Phe signal less efficiently than MCClike peptides, even when affinities are closely matched (Figures S4 E and S4F). These structures raise the question of whether a minor tilt of the TCR relative to the MHC can have consequences for signaling.

Upon closer inspection, we find that homologies between what appear to be unrelated peptide sequences emerge from sequence clustering and structural analysis. For example, close structural relationships between the interaction modes of the 2B4-reactive peptides MCC and 2A are apparent even though the peptides show little homology at 4/5 TCR contact positions (Figure 5 A). We set out to determine whether we could identify intermediate sequences that would “evolutionarily” link these two peptide sequences, given that both reside in the same sequence cluster (

Figure 3 C). Using our data set of peptide sequences selected for 2B4 binding, we were able to populate a family of peptides that would incrementally link MCC and 2A, with each peptide differing by only one TCR contact from the peptide before and after it (Figure 5 D). Thus, connectivity can be established between MCC and 2A through stepwise single amino acid drifts from their parent sequences.

Collectively, despite differences in peptide sequences, all MCC and library-peptide-derived complexes share many common features with regards to docking geometry and interaction chemistry. Changes in up-facing peptide residue sequence (e.g., P5, P8) are accommodated "locally" in a structurally parsimonious fashion that preserves most of the parent MCC peptide complex features, as opposed to accommodation through largescale repositioning of the CDR loops on the pMHC surface.

Development and Selection of a Human MHC Platform for Yeast Display To exploit our technology to find ligands for TCRs relevant to humandisease, wealso engineered the humanMHC HLA-DR15, an allele with genetic linkage to multiple sclerosis (Hafler et al., 2007; Patsopoulos et al., 2013). For yeast surface display, HLA-DR15 was constructed comparably to the murine I-E k b 1 a 1 mini MHC (Figure 6 A). We chose to examine two closely related TCRs, Ob.1A12 and Ob.2F3, that were cloned from a patient with relapsing-remitting multiple sclerosis and recognize HLA-DR15 bound to an immunodominant epitope of myelin basic protein (MBP, residues 85–99) (Wucherpfennig et al., 1994b). These two TCRs utilize the same V a -J a and V b -J b gene segments and differ at one position in the CDR3 a loop and two positions in CDR3 b . Ob.1A12 is sufficient to cause disease in a humanized TCR transgenic mouse model (Harkiolaki et al., 2009; Hausmann et al., 1999; Madsen et al., 1999). A structure of Ob.1A12 complexed with MBP-HLA-DR15 revealed an atypical docking mode, with the TCR shifted toward the N terminus of the peptide and primarily interacting with a P2-His/P3-Phe TCR contact motif (Figure 6 A) (Hahn et al., 2005; Wucherpfennig et al., 1994a).

Because the initial WT MBP-HLA-DR15 yeast display construct was not stained by Ob.1A12 TCR tetramers, we subjected the construct to error-prone mutagenesis and selected for binding with Ob.1A12. Our final construct combined the most heavily selected mutation (Pro11Ser on HLA-DR15 b) with two solubility-enhancing

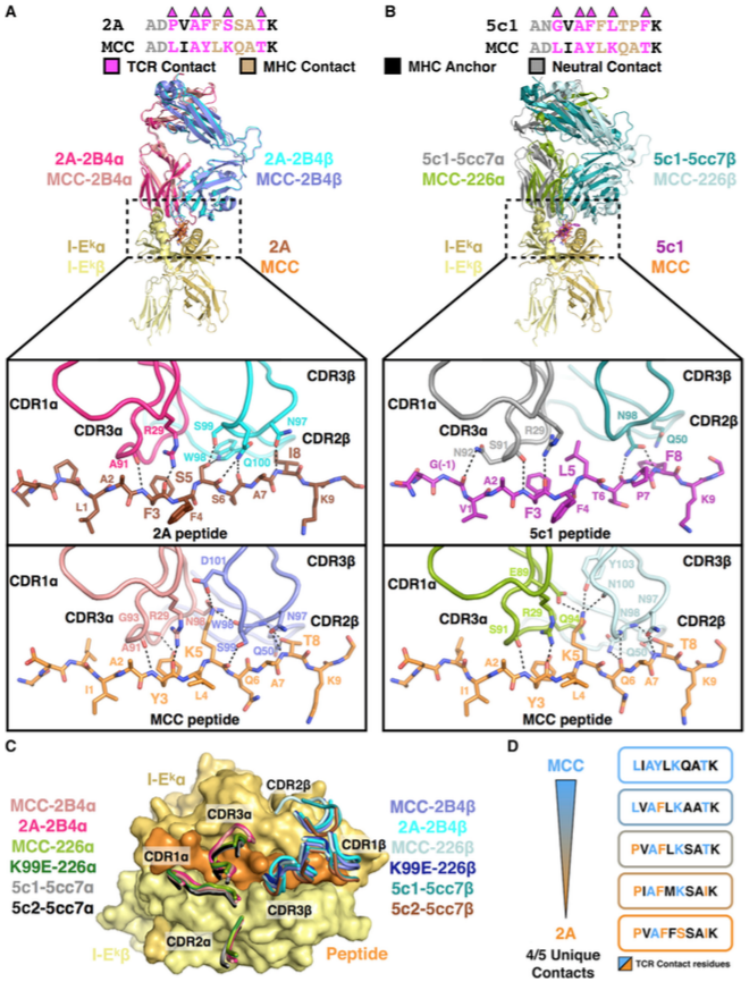


Figure 1.23: Peptides Distantly Related to MCC Show Highly Similar Mechanism of Recognition and Linkages to the Cognate Antigen (A and B) Comparison of crystal structures of TCR-pMHC complexes for 2B4-2A-I-Ek and 2B4-MCC-I-Ek (PDB ID: 3QIB) (A) and 5cc7-5c1-I-Ek and 226-MCC-I-Ek (PDB ID: 3QIU) (B). TCR contacts are shown in magenta (top, noted with triangles). There is very little change in overall binding geometry despite significant variation of peptide sequence. The TCRs accommodate differences in peptide sequence primarily through differences in CDR3b (bottom). (C) TCR CDR loop footprints for 2B4 recognizing MCC and 2A peptides, 226 recognizing MCC and MCC K99E peptides, and 5cc7 recognizing 5c1 and 5c2 peptides show very little deviation. (D) Relationship between MCC and 2A peptides revealed through intermediate selected peptide sequences.

mutations on the bottom of the MHC platform (Figure S6 A). The final construct stained robustly with Ob.1A12 and Ob.2F3 TCR tetramers (Figure S6 B).

We designed a peptide library within the HLA-DR15 mini MHC scaffold to find novel Ob.1A12and Ob.2F3-reactive peptides (Figure 6 A). Given that Ob.1A12 binds its cognate pMHC shifted toward the N terminus of the peptide, we extended the library, randomizing from P(

4) to P10 (Hahn et al., 2005). The P1 and P4 peptide anchors for HLA-DR15 were afforded limited diversity. When we selected with Ob.1A12 and Ob.2F3 TCRs, we observed a strong convergence to a WT MBP-like TCR recognition motif for the primary TCR contacts (P2 His, P3 Phe, and P5 Lys) (termed the "HF" motif) (Figures 6 B, S6 C, and S6D).

Given the dominance of the HF motif in the selection results, we sought to determine whether alternative cross-reactive TCR epitopes would emerge if the motif were suppressed. We made a library that allowed every amino acid except for His at P2, Phe at P3, and Lys at P5 (Figure 6 C). After selection, the TCR-binding clones still converged to a central HF motif by register shifting toward the C terminus of the peptide by one amino acid, allowing the previous P4 Phe anchor to be repurposed as the P3 TCR contact and the P3 position of the library to become the new P2 His TCR contact (Figure 6 C). Furthermore, when we subsequently prevented both His and Phe at P2 and P3 in a new library to suppress potential register shifting, we did not isolate any Ob.1A12or Ob.2F3-binding peptides (data not shown). These results show that the HF motif is required for TCR recognition and its enrichment is a function of TCR preference, not any inherent biases caused by the library or MHC anchor positions of the peptide.

Clustering analysis of the selected peptides for both Ob.1A12 and Ob.2F3 showed distinct clusters consisting of peptides no more than four amino acids different from each other (Figure 6 D). When the stringency of clustering is increased to allow no more than three amino acid differences, matching the analysis done for I-E k , there were several more sparse clusters (Figure S6 E). Because Ob1.A12 and Ob.2F3 are so focused on the HF motif, there are fewer total hot-spot residues distributed on the peptide compared to the MCC-reactive TCRs we studied.

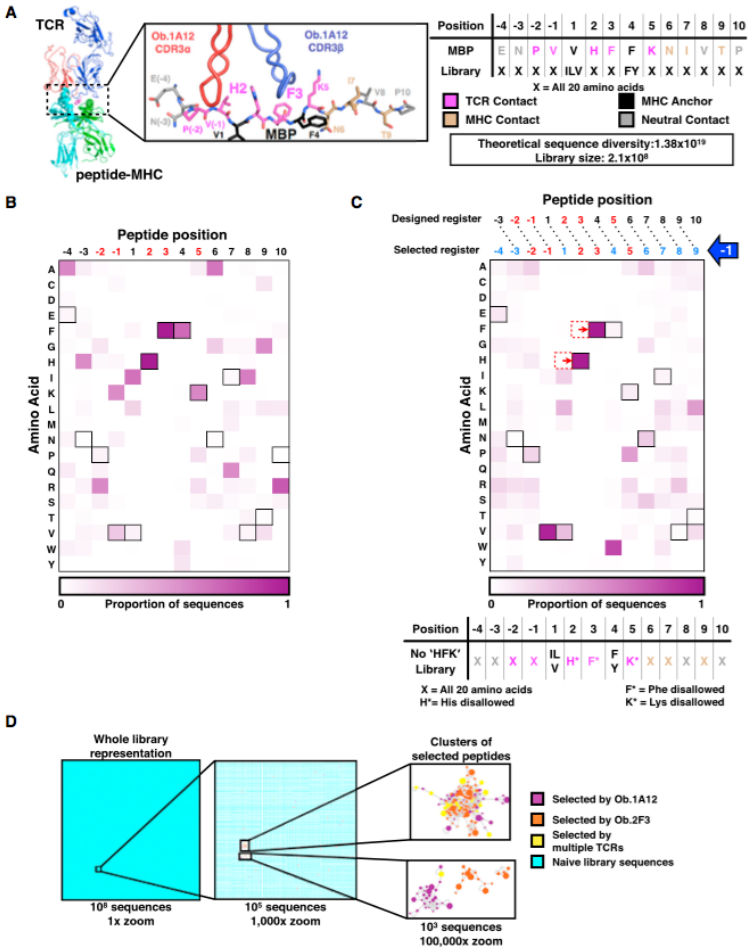


Figure 1.24: Design and Selection of HLA-DR15-Based Libraries for MBP-Reactive Human TCRs (A) HLA-DR15 library design based upon structure of Ob.1A12-MBP-HLA-DR15 complex. Residues P(4)–P10 are fully randomized, except for the P1 and P4 anchors (in black). TCR contacts are colored magenta. (B) Heatmap of amino acid preference by position for Ob.1A12 TCR. The sequence for MBP is represented via outlined boxes. TCR contacts are labeled red on the x axis. (C) Design and selection results of library that suppresses central HF TCR recognition motif at P2–P3 of peptide. Resulting register shift is shown in blue on x axis. (D) Sequence clustering shows distinct, related clusters of selected peptides. Sequence cluster is placed in a representation of whole-library sequence space (left: 13 magnification, center: 10003 magnification) for reference.

High-Confidence Prediction of Naturally Occurring TCR-Reactive Peptides The surprisingly limited degeneracy of TCR recognition suggests that it may be feasible to identify naturally occurring TCR ligands with a random peptide library. However, library selections and deep sequencing alone are not sufficient to identify naturally occurring ligands for two reasons. First, the size of yeast libraries relative to all possible MHC-displayed peptides makes it unlikely that any given peptide sequence exists in the library. Second, the amino acid substitutions that are permitted at each position along the peptide represent a complex, and as our covariation analysis indicated, cooperative interplay between the peptide, MHC, and TCR that may not be well described by common substitution matrices such as BLOSUM (Henikoff and Henikoff, 1992). For example, even though manual inspection of Ob.1A12-binding sequences readily shows the WT-like HF motif, the sequences do not find MBP as a match in blastp searches (data not shown).

We therefore set out to develop an algorithm to use the aggregate data from our selection results to inform searches for candidate TCR antigens. First, we created a substitution matrix that used the positional frequency information derived from our Ob.1A12 and Ob.2F3 deep-sequencing data (Zhao et al., 2001).Because the limited coverage of our libraries could lead to appearance of residue biases at noncritical (i.e., neutral) peptide positions that do not reflect actual selective pressure, we created a new HLA-DR15-based library where we fixed the dominant Ob.1A12-binding motif (P2 His, P3 Phe, and P5 Lys/ Arg) along with the P1 and P4 MHC-binding anchors, while randomizing the remaining residues. When the selected libraries were sequenced, we found that whereas some proximal positions such as P(-1) and P(-2) still showed distinct residue preferences, other positions such as P7 and P8 showed less convergence relative to the original HLA-DR15 library (Figure S7 A). The more granular substitution data for peptide positions distal to the TCR-binding hot spot allowed us to construct a more reliable algorithm.

We compiled two 14 3 20 substitution matrices consisting of the observed frequencies of the 20 amino acids at each of the 14 positions of the library peptides from the focused DR15 pMHC libraries selected by Ob.1A12 and Ob.2F3 (Figures 7 A and S7 A; Table S2)(Zhao et al., 2001). Given that minimal residue covariation was

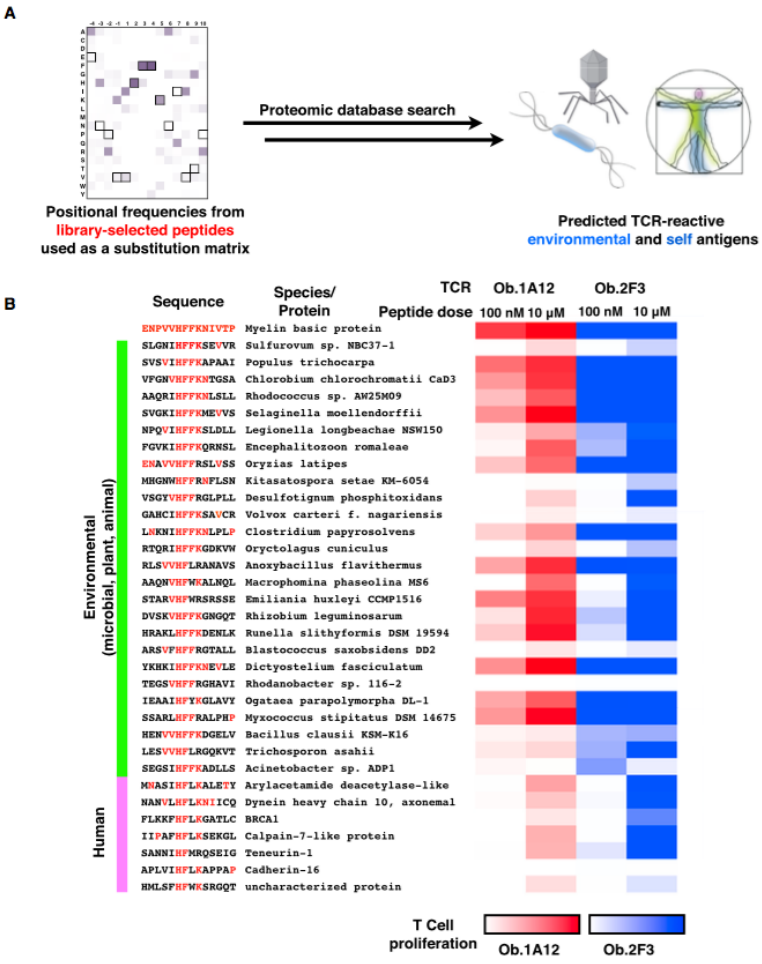


Figure 1.25: Discovery of Naturally Occurring TCR Ligands through Deep Sequencing and Substitution-Matrix-Based Homology Search (A) Schematic for ligand search strategy, in which a positional substitution matrix is generated from deep-sequencing data and then used to find naturally occurring peptides that are represented within the matrix. (B) Functional characterization of a selection of naturally occurring peptides with predicted activity. Activity is tested via proliferation of T cells when exposed to peptide. Heatmaps are normalized to 10 mM dose of MBP peptide for each T cell clone.

observed for Ob.1A12 and Ob.2F3 selections, each position was treated independently (Figure S7 B). Our peptide database search using the Ob.1A12-based matrix yielded 2,330 unique hits, including MBP. For the search based on the Ob.2F3 matrix, we had 4,824 unique hits, again including MBP. The peptide hits shared the central P(-1)-P5 motif of MBP, but the flanking residues showed very little sequence homology to either MBP or each other (Figure 7 B; Table S3). The predicted peptides are from diverse microbial sources, such as bacteria; from environmental sources, such as antigens expressed by plants; and from proteins in the human proteome.

To test our computationally predicted ligands for Ob.1A12 and Ob.2F3, we synthesized a diverse set comprising 26 of the potentialenvironmentalantigensaswellas7novelhumanpeptidespredicted to cross-react with Ob.1A12 and Ob.2F3. When we tested the 33 putative ligands for activity, 25/26 of the environmental antigens and 6/7 of the human peptides induced proliferation for Ob.1A12 and/or Ob.2F3, a success rate of 94% (Figure 7 B).

1.5.3 Discussion

The concept of TCR cross-reactivity is important because key aspects of T cell biology, including thymic development, pathogen surveillance, autoimmunity, and transplant rejection, seemingly require recognition of diverse ligands. In this study, we aimed to define the mechanisms underlying TCR specificity and cross-reactivity using a combinatorial, biochemical approach that yielded massive data sets based on direct selection. This has given us insight into the structural basis of TCR cross-reactivity and also provides a robust way to discover peptide ligands for a TCR of interest.

Our results clarify previous controversies on whether TCRs are highly cross-reactive or highly specific by leveraging large amounts of experimental data found via direct binding of pMHC to TCR. We find that structural principles allow for the TCR to engage large numbers of unique pMHC without requiring degeneracy in pMHC recognition. If the criterion of cross-reactivity is simply the number of unique peptide sequences that can be recognized by any given TCR, then TCRs do exhibit a high degree of cross-reactivity. Given that the libraries greatly undersample all

possible sequence combinations, it is likely that our hundreds of discovered peptides are emblematic of thousands of different peptides that can be recognized by the studied TCRs. However, when cross-reactive peptides are examined en masse, we find conserved TCR-binding (i.e., up-facing) motifs. TCR cross-reactivity is not achieved by each receptor recognizing a large number of unrelated peptide epitopes but rather through greater tolerance for substitutions to peptide residues outside of the TCR interface, differences in residues that contact the MHC, and relatively conservative changes to the residues that contact the TCR CDR loops. The segregation of TCR recognition and MHC binding allows for TCRs to simultaneously accommodate needs for specificity and cross-reactivity.

Although we believe this mechanism will be general for ab TCRs, recognition of nonhomologous antigens certainly occurs to varying degrees in the TCR repertoire, although molecularly defined examples are surprisingly rare. The ability for one TCR to bind to multiple MHCs (e.g., alloreactivity), for one TCR to bind in multiple orientations on one MHC, for a peptide to noncanonically bind MHC (e.g., partially filled MHC grooves, registershifted peptides), or for a TCR to have TCR-peptide contacts as a disproportionately large or small part of the overall interface (e.g., "super-bulged" peptides) will grant some receptors a greater degree of epitope promiscuity (Adams et al., 2011; Colf et al., 2007; Maynard et al., 2005; Morris and Allen, 2012; Morris et al., 2011; Tynan et al., 2005). It is also possible that class I versus class II MHC-specific TCRs could exhibit different degrees of cross-reactivity as a consequence of the "low-lying" peptides in the class II groove, versus the elevated or "higherprofile" peptides presented by class I. Indeed, in a prior study, multiple peptides reactive with a class I-specific (H-2L d) murine TCR were identified through manual curation, and the structures indicated a diverse recognition chemistry by the TCR CDR3 loops (Adams et al., 2011). In retrospect, a close inspection reveals striking commonalities in the peptide-binding chemistry by the TCR, in particular a requirement for a hydrophobic contact at the apex of the P7 "bulge" that forms the principal site of contact with the TCR CDR3 b . In contrast, a second class I TCR, 2C, was not found to exhibit peptide degeneracy, instead exhibiting specificity for its endogenous

antigen, QL9, in a manner similar to that of the class II-specific TCRs studied here (unpublished data).

An important implication of our findings, which is consistent with previous studies (Macdonald et al., 2009), is that identification of endogenous antigens of TCRs is feasible using pMHC libraries. In our previous view of cross-reactivity, we assumed that a given TCR would cross-react with so many peptides in a library that elucidation of "natural" leads from a background of degenerately binding sequences would be extremely difficult. Additionally, the sparse coverage of possible sequences renders it unlikely that any given sequence of interest will be represented with 100% identity in our library. However, limited TCR epitope cross-reactivity allows us to use selection results to constrain computational searches of protein databases, which proves to be a highly successful strategy for finding naturally occurring TCR ligands. Thus, this approach now opens up the possibility of peptide ligand discovery for "orphan" TCRs such as those from regulatory T cells and tumor-infiltrating lymphocytes (TILs).

Although the naturally occurring peptides in this study were found as a proof of principle for our methodology, they further support the hypothesis that autoimmune T cells have the ability to be activated by immunogens encountered in the environment, which may serve as the trigger for the initiation of autoimmunity (De la Herran-Arita et al., 2013; Harkiolaki et al., 2009; Hausmann et al., 1999; Wucherpfennig and Strominger, 1995). Additionally, the potential for other human peptides to cross-react with autoimmune TCRs with previously "known" antigens presents the intriguing possibility that individual TCRs can recognize multiple self-peptides, potentially contributing to T cell pathologies in autoimmune disease. This notion is supported by the finding that a murine TCR specific for myelin-oligodendrocyte glycoprotein cross-reacts with a second CNS antigen, neurofilament M. Due to this unexpected cross-reactivity, these T cells remained pathogenic even in MOG-deficient mice (Krishnamoorthy et al., 2009). Our approach for systematic discovery of peptides recognized by human TCRs thus has the potential to advance our understanding of complex pathogenesis of immune-mediated diseases.

1.5.4 Methods

Creation and Selection of pMHC Libraries Peptide libraries were created through use of mutagenic primers allowing all 20 amino acids via NNK codons. The libraries allowed limited diversity at the known MHC anchor residues to maximize the number of correctly folded and displayed pMHC clones in the library. Yeast libraries were created by electroporation of competent EBY-100 cells via homologous recombination of linearized pYAL vector and mutagenized pMHC construct essentially as described previously (Adams et al., 2011; Chao et al., 2006). Final libraries contained approximately 2 3 10 8 yeast transformants. Yeast libraries were selected for binding to the TCR of interest coupled to streptavidin-coated magnetic beads (Miltenyi) through magnetic-activated cell sorting. After libraries enriched above the baseline of streptavidin beads alone (typically after three rounds of selection), a final round of selection was conducted with fluorescently labeled streptavidin tetramers.

Deep Sequencing of pMHC Libraries Pooled plasmids from 5 3 10 7 yeast from each round of selection were isolated via yeast miniprep (Zymoprep II kit, Zymo Research) and used as PCR template to prepare sequencing samples. The adaptor and barcode sequences were appended via nested 25-round cycles of PCR of the purified plasmids using Phusion polymerase (NEB). Deep sequencing was conducted on an Illumina MiSeq sequencer at the Stanford Stem Cell Institute Genome Center.

Profile-Based Searches for Naturally Occurring Peptide Ligands The positional frequencies from round 3 of the fixed HF library were used to generate a 14 3 20 substitution matrix. Each protein in the NR (NCBI) or human protein (Uniprot) databases was scanned using a 14 position sliding window and scored as a product of the positional substitution matrix (Cockcroft and Osguthorpe, 1991). In this way, a candidate peptide containing even a single disallowed substitution would be excluded as a possible hit.

Structural Determination of pMHC-TCR Complexes All crystallographic data were collected at the Stanford Synchroton Radiation Lightsource (Stanford, CA, USA) beamlines 11-1 and 12-2. Data were indexed, integrated, and scaled using either the XDS or the HKL-2000 program suites (Kabsch, 2010; Otwinowski et

al., 1997). All structures were solved via molecular replacement using the program Phaser (McCoy, 2007) and refined with Phenix (Adams et al., 2010).

Extended Experimental Procedures Further details for the design, selection, and sequencing of yeast display libraries; methods for production, characterization, and crystallization of proteins; and computational discovery and functional validation of peptide hits can be found online in the Extended Experimental Procedures.

ACCESSION NUMBERS

The coordinates and structure factors for the reported crystal structures are deposited in the Protein Data Bank (PDB) under PDB IDs 4P2O, 4P2Q, and 4P2R. Deep-sequencing data can be accessed via the Sequence Read Archive (SRA) under project code SRP040021.

Supplemental Information includes Extended Experimental Procedures, seven figures, and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2014.03.047 .

1.5.5 Acknowledgements

I feel privileged to be invited to collaborate by this excellent group of researchers, including first author Michael E. Birnbaum, as well as Juan L. Mendoza, Dhruv K. Sethi, Shen Dong, Jessica Dobbins, Engin O¨ zkan, Mark M. Davis, Kai W. Wucherpfennig, and K. Christopher Garcia. We thank Suzanne Fischer, Gary Mantalas, and Nelida Prado for technical assistance and Jarrett Adams, Lauren Ely, and Evan Newell for helpful discussions. M.E.B. was supported by a Regina Casper Stanford Graduate Fellowship, a Gerald J. Lieberman Fellowship, and a National Science Foundation Graduate Fellowship. J.L.M. was supported by NCI-K01CA175127 and a Helen Hay Whitney Foundation postdoctoral fellowship. This work was supported by the NIH (PO1 AI045757 to K.W.W., R01 AI03867 to K.C.G., R01 AI022511 to M.M.D., and U19 4100041120 to M.M.D. and K.C.G.) and HHMI (to M.M.D. and K.C.G.). Use of the Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, is supported by the U.S. Department of Energy under Contract No. DE-AC02-76SF00515. The SSRL Structural Molecular Biology Program is

supported by the DOE Office of Biological and Environmental Research and by the National Institutes of Health (including P41GM103393).

1.5.6 References

Adams, J.J., Narayanan, S., Liu, B., Birnbaum, M.E., Kruse, A.C., Bowerman, N.A., Chen, W., Levin, A.M., Connolly, J.M., Zhu, C., et al. (2011). T cell recep- tor signaling is limited by docking geometry to peptide-major histocompatibil- ity complex. Immunity 35 , 681–693.

Adams, P.D., Afonine, P.V., Bunko´ czi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.W., Kapral, G.J., Grosse-Kunstleve, R.W., et al. (2010). PHENIX: a comprehensive Python-based system for macromolecular struc- ture solution. Acta Crystallogr. D Biol. Crystallogr. 66 , 213–221.

Basu, D., Horvath, S., Matsumoto, I., Fremont, D.H., and Allen, P.M. (2000). Molecular basis for recognition of an arthritic peptide and a foreign epitope on distinct MHC molecules by a single TCR. J. Immunol. 164 , 5788–5796.

Benoist, C., and Mathis, D. (2001). Autoimmunity provoked by infection: how good is the case for T cell epitope mimicry? Nat. Immunol. 2 , 797–801.

Birnbaum, M.E., Dong, S., and Garcia, K.C. (2012). Diversity-oriented approaches for interrogating T-cell receptor repertoire, ligand recognition, and function. Immunol. Rev. 250 , 82–101.

Boder, E.T., Bill, J.R., Nields, A.W., Marrack, P.C., and Kappler, J.W. (2005). Yeast surface display of a noncovalent MHC class II heterodimer complexed with antigenic peptide. Biotechnol. Bioeng. 92 , 485–491.

Borbulevych, O.Y., Piepenbrink, K.H., Gloor, B.E., Scott, D.R., Sommese, R.F., Cole, D.K., Sewell, A.K., and Baker, B.M. (2009). T cell receptor cross- reactivity directed by antigen-dependent tuning of peptide-MHC molecular flexibility. Immunity 31 , 885–896.

Borbulevych, O.Y., Piepenbrink, K.H., and Baker, B.M. (2011). Conformational melding permits a conserved binding geometry in TCR recognition of foreign and self molecular mimics. J. Immunol. 186 , 2950–2958.

Chao, G., Lau, W.L., Hackel, B.J., Sazinsky, S.L., Lippow, S.M., and Wittrup, K.D. (2006). Isolating and engineering human antibodies using yeast surface display. Nat. Protoc. 1 , 755–768.

Cockcroft, V.B., and Osguthorpe, D.J. (1991). Relative-residue surface-accessibility patterns reveal myoglobin and catalase similarity. FEBS Lett. 293 , 149–152.

Colf, L.A., Bankovich, A.J., Hanick, N.A., Bowerman, N.A., Jones, L.L., Kranz, D.M., and Garcia, K.C. (2007). How a single T cell receptor recognizes both self and foreign MHC. Cell 129 , 135–146.

Crawford, F., Huseby, E., White, J., Marrack, P., and Kappler, J.W. (2004). Mimotopes for alloreactive and conventional T cells in a peptide-MHC display library. PLoS Biol. 2 , E90.

Crawford, F., Jordan, K.R., Stadinski, B., Wang, Y., Huseby, E., Marrack, P., Slansky, J.E., and Kappler, J.W. (2006). Use of baculovirus MHC/peptide display libraries to characterize T-cell receptor ligands. Immunol. Rev. 210 , 156–170.

Davis, M.M., and Bjorkman, P.J. (1988). T-cell antigen receptor genes and T-cell recognition. Nature 334 , 395–402.

De la Herran-Arita, A.K., Kornum, B.R., Mahlios, J., Jiang, W., Lin, L., Hou, T., Macaubas, C., Einen, M., Plazzi, G., Crowe, C., et al. (2013). CD4+ T cell auto- immunity to hypocretin/orexin and cross-reactivity to a 2009 H1N1 influenza A epitope in narcolepsy. Sci. Transl. Med. 5 , 216ra176.

Ebert, P.J., Jiang, S., Xie, J., Li, Q.J., and Davis, M.M. (2009). An endogenous positively selecting peptide enhances mature T cell responses and becomes an autoantigen in the absence of microRNA miR-181a. Nat. Immunol. 10 , 1162–1169.

Ehrich, E.W., Devaux, B., Rock, E.P., Jorgensen, J.L., Davis, M.M., and Chien, Y.H. (1993). T cell receptor interaction with peptide/major histocompatibility complex (MHC) and superantigen/MHC ligands is dominated by antigen. J. Exp. Med. 178 , 713–722.

Esteban, O., and Zhao, H. (2004). Directed evolution of soluble single-chain human class II MHC molecules. J. Mol. Biol. 340 , 81–95.

Evavold, B.D., Sloan-Lancaster, J., Wilson, K.J., Rothbard, J.B., and Allen, P.M. (1995). Specific T cell recognition of minimally homologous peptides: evidence for multiple endogenous ligands. Immunity 2 , 655–663.

Garcia, K.C., and Adams, E.J. (2005). How the T cell receptor sees antigen—a structural view. Cell 122 , 333–336.

Hafler, D.A., Compston, A., Sawcer, S., Lander, E.S., Daly, M.J., De Jager, P.L., de Bakker, P.I., Gabriel, S.B., Mirel, D.B., Ivinson, A.J., et al.; International Multiple Sclerosis Genetics Consortium (2007). Risk alleles for multiple scle- rosis identified by a genomewide study. N. Engl. J. Med. 357 , 851–862.

Hahn, M., Nicholson, M.J., Pyrdol, J., and Wucherpfennig, K.W. (2005). Unconventional topology of self peptide-major histocompatibility complex binding by a human autoimmune T cell receptor. Nat. Immunol. 6 , 490–496.

Harkiolaki, M., Holmes, S.L., Svendsen, P., Gregersen, J.W., Jensen, L.T., McMahon, R., Friese, M.A., van Boxel, G., Etzensperger, R., Tzartos, J.S., et al. (2009). T cell-mediated autoimmune disease due to low-affinity cross- reactivity to common microbial peptides. Immunity 30 , 348–357.

Hausmann, S., Martin, M., Gauthier, L., and Wucherpfennig, K.W. (1999). Structural features of autoreactive TCR that determine the degree of degener- acy in peptide recognition. J. Immunol. 162 , 338–344.

Hemmer, B., Fleckenstein, B.T., Vergelli, M., Jung, G., McFarland, H., Martin, R., and Wiesmu¨ ller, K.H. (1997). Identification of high potency microbial and self ligands for a human autoreactive class II-restricted T cell clone. J. Exp. Med. 185 , 1651–1659.

Hemmer, B., Vergelli, M., Gran, B., Ling, N., Conlon, P., Pinilla, C., Houghten, R., McFarland, H.F., and Martin, R. (1998a). Predictable TCR antigen recogni- tion based on peptide scans leads to the identification of agonist ligands with no sequence homology. J. Immunol. 160 , 3631–3636.

Hemmer, B., Vergelli, M., Pinilla, C., Houghten, R., and Martin, R. (1998b). Probing degeneracy in T-cell recognition using peptide combinatorial libraries. Immunol. Today 19 , 163–168.

Henikoff, S., and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA 89 , 10915–10919.

Jiang, W., and Boder, E.T. (2010). High-throughput engineering and analysis of peptide binding to class II MHC. Proc. Natl. Acad. Sci. USA 107 , 13258–13263.

Kabsch, W. (2010). Xds. Acta Crystallogr. D Biol. Crystallogr. 66 , 125–132. Kersh, G.J., and Allen, P.M. (1996). Essential flexibility in the T-cell recognition of antigen. Nature 380 , 495–498.

Krishnamoorthy, G., Saxena, A., Mars, L.T., Domingues, H.S., Mentele, R., Ben-Nun, A., Lassmann, H., Dornmair, K., Kurschus, F.C., Liblau, R.S., and Wekerle, H. (2009). Myelin-specific T cells also recognize neuronal autoantigen in a transgenic mouse model of multiple sclerosis. Nat. Med. 15 , 626–632.

Krogsgaard, M., Prado, N., Adams, E.J., He, X.L., Chow, D.C., Wilson, D.B., Garcia, K.C., and Davis, M.M. (2003). Evidence that structural rearrangements and/or flexibility during TCR binding can contribute to T cell activation. Mol. Cell 12 , 1367–1378.

Lo, W.L., Felix, N.J., Walters, J.J., Rohrs, H., Gross, M.L., and Allen, P.M. (2009). An endogenous peptide positively selects and augments the activation and survival of peripheral CD4+ T cells. Nat. Immunol. 10 , 1155–1161.

Macdonald, W.A., Chen, Z., Gras, S., Archbold, J.K., Tynan, F.E., Clements, C.S., Bharadwaj, M., Kjer-Nielsen, L., Saunders, P.M., Wilce, M.C., et al. (2009). T cell allorecognition via molecular mimicry. Immunity 31 , 897–908.

Madsen, L.S., Andersson, E.C., Jansson, L., krogsgaard, M., Andersen, C.B., Engberg, J., Strominger, J.L., Svejgaard, A., Hjorth, J.P., Holmdahl, R., et al. (1999). A humanized model for multiple sclerosis using HLA-DR2 and a human T-cell receptor. Nat. Genet. 23 , 343–347.

Mason, D. (1998). A very high level of crossreactivity is an essential feature of the T-cell receptor. Immunol. Today 19 , 395–404.

Matsui, K., Boniface, J.J., Reay, P.A., Schild, H., Fazekas de St Groth, B., and Davis, M.M. (1991). Low affinity interaction of peptide-MHC complexes with T cell receptors. Science 254 , 1788–1791.

Maynard, J., Petersson, K., Wilson, D.H., Adams, E.J., Blondelle, S.E., Boulanger, M.J., Wilson, D.B., and Garcia, K.C. (2005). Structure of an autoimmune T cell receptor complexed with class II peptide-MHC: insights into MHC bias and antigen specificity. Immunity 22 , 81–92.

Mazza, C., Auphan-Anezin, N., Gregoire, C., Guimezanes, A., Kellenberger, C., Roussel, A., Kearney, A., van der Merwe, P.A., Schmitt-Verhulst, A.M., and Malissen, B. (2007). How much can a T-cell antigen receptor adapt to structur- ally distinct antigenic peptides? EMBO J. 26 , 1972–1983.

McCoy, A.J. (2007). Solving structures of protein complexes by molecular replacement with Phaser. Acta Crystallogr. D Biol. Crystallogr. 63 , 32–41.

Morris, G.P., and Allen, P.M. (2012). How the TCR balances sensitivity and specificity for the recognition of self and pathogens. Nat. Immunol. 13 , 121–128

Morris, G.P., Ni, P.P., and Allen, P.M. (2011). Alloreactivity is limited by the endogenous peptide repertoire. Proc. Natl. Acad. Sci. USA 108 , 3695–3700.

Nanda, N.K., Arzoo, K.K., Geysen, H.M., Sette, A., and Sercarz, E.E. (1995). Recognition of multiple peptide cores by a single T cell receptor. J. Exp. Med. 182 , 531–539.

Newell, E.W., Ely, L.K., Kruse, A.C., Reay, P.A., Rodriguez, S.N., Lin, A.E., Kuhns, M.S., Garcia, K.C., and Davis, M.M. (2011). Structural basis of speci- ficity and cross-reactivity in T cell receptors specific for cytochrome c-I-E k . J. Immunol. 186 , 5823–5832.

Otwinowski, Z., Minor, W., and Carter, C.W., Jr. (1997). Processing of X-ray diffraction data collected in oscillation mode. In Methods in Enzymology (New York: Academic Press), pp. 307–326.

Patsopoulos, N.A., Barcellos, L.F., Hintzen, R.Q., Schaefer, C., van Duijn, C.M., Noble, J.A., Raj, T., Gourraud, P.A., Stranger, B.E., Oksenberg, J., et al.; IMSGC; ANZgene (2013). Fine-mapping the genetic association of the major histocompatibility complex in multiple sclerosis: HLA and non-HLA effects. PLoS Genet. 9 , e1003926.

Reay, P.A., Kantor, R.M., and Davis, M.M. (1994). Use of global amino acid replacements to define the requirements for MHC binding and T cell recogni- tion of moth cytochrome c (93-103). J. Immunol. 152 , 3946–3957.

Reiser, J.B., Darnault, C., Gre´ goire, C., Mosser, T., Mazza, G., Kearney, A., van der Merwe, P.A., Fontecilla-Camps, J.C., Housset, D., and Malissen, B. (2003). CDR3 loop flexibility contributes to the degeneracy of TCR recognition. Nat. Immunol. 4 , 241–247.

Rudolph, M.G., Stanfield, R.L., and Wilson, I.A. (2006). How TCRs bind MHCs, peptides, and coreceptors. Annu. Rev. Immunol. 24 , 419–466.

Sewell, A.K. (2012). Why must T cells be cross-reactive? Nat. Rev. Immunol. 12 , 669–677.

Shann, F., Nohynek, H., Scott, J.A., Hesseling, A., and Flanagan, K.L.; Working Group on Nonspecific Effects of Vaccines (2010). Randomized trials to study the nonspecific effects of vaccines in children in low-income countries. Pediatr. Infect. Dis. J. 29 , 457–461.

Shih, F.F., and Allen, P.M. (2004). T cells are not as degenerate as you think, once you get to know them. Mol. Immunol. 40 , 1041–1046.

Starwalt, S.E., Masteller, E.L., Bluestone, J.A., and Kranz, D.M. (2003). Directed evolution of a single-chain class II MHC product by yeast display. Protein Eng. 16 , 147–156.

Tynan, F.E., Burrows, S.R., Buckle, A.M., Clements, C.S., Borg, N.A., Miles, J.J., Beddoe, T., Whisstock, J.C., Wilce, M.C., Silins, S.L., et al. (2005). T cell receptor recognition of a 'super-bulged' major histocompatibility complex class I-bound peptide. Nat. Immunol. 6 , 1114–1122.

Wang, Y., Rubtsov, A., Heiser, R., White, J., Crawford, F., Marrack, P., and Kappler, J.W. (2005). Using a baculovirus display library to identify MHC class I mimotopes. Proc. Natl. Acad. Sci. USA 102 , 2476–2481.

Welsh, R.M., Che, J.W., Brehm, M.A., and Selin, L.K. (2010). Heterologous immunity between viruses. Immunol. Rev. 235 , 244–266

Wen, F., Esteban, O., and Zhao, H. (2008). Rapid identification of CD4+ T-cell epitopes using yeast displaying pathogen-derived peptide library. J. Immunol. Methods 336 , 37–44.

Wen, F., Sethi, D.K., Wucherpfennig, K.W., and Zhao, H. (2011). Cell surface display of functional human MHC class II proteins: yeast display versus insect cell display. Protein Eng. Des. Sel. 24 , 701–709.

Wilson, D.B., Pinilla, C., Wilson, D.H., Schroder, K., Boggiano, C., Judkowski, V., Kaye, J., Hemmer, B., Martin, R., and Houghten, R.A. (1999). Immunoge- nicity. I. Use of peptide libraries to identify epitopes that activate clonotypic CD4+ T cells and induce T cell responses to native peptide ligands. J. Immunol. 163 , 6424–6434.

Wilson, D.B., Wilson, D.H., Schroder, K., Pinilla, C., Blondelle, S., Houghten, R.A., and Garcia, K.C. (2004). Specificity and degeneracy of T cells. Mol. Immunol. 40 , 1047–1055.

Wooldridge, L., Ekeruche-Makinde, J., van den Berg, H.A., Skowera, A., Miles, J.J., Tan, M.P., Dolton, G., Clement, M., Llewellyn-Lacey, S., Price, D.A., et al. (2012). A single autoimmune T cell receptor recognizes more than a million different peptides. J. Biol. Chem. 287 , 1168–1177.

Wu, L.C., Tuot, D.S., Lyons, D.S., Garcia, K.C., and Davis, M.M. (2002). Two-step binding mechanism for T-cell receptor recognition of peptide MHC. Nature 418 , 552–556.

Wucherpfennig, K.W., Sette, A., Southwood, S., Oseroff, C., Matsui, M., Strominger, J.L., and Hafler, D.A. (1994a). Structural requirements for binding of an immunodominant myelin basic protein peptide to DR2 isotypes and for its recognition by human T cell clones. J. Exp. Med. 179 , 279–290.

Wucherpfennig, K.W., Zhang, J., Witek, C., Matsui, M., Modabber, Y., Ota, K., and Hafler, D.A. (1994b). Clonal expansion and persistence of human T cells specific for an immunodominant myelin basic protein peptide. J. Immunol. 152 , 5581–5592. Wucherpfennig, K.W., and Strominger, J.L. (1995). Molecular mimicry in T cell- mediated autoimmunity: viral peptides activate human T cell clones specific for myelin basic protein. Cell 80 , 695–705. Wucherpfennig, K.W., Allen, P.M., Celada, F., Cohen,

I.R., De Boer, R., Garcia, K.C., Goldstein, B., Greenspan, R., Hafler, D., Hodgkin, P., et al. (2007). Poly- specificity of T cell and B cell receptor recognition. Semin. Immunol. 19 , 216–224.

Yin, Y., and Mariuzza, R.A. (2009). The multiple mechanisms of T cell receptor cross-reactivity. Immunity 31 , 849–851.

Zhao, R., Loftus, D.J., Appella, E., and Collins, E.J. (1999). Structural evidence of T cell xeno-reactivity in the absence of molecular mimicry. J. Exp. Med. 189 , 359–370.

Zhao, Y., Gran, B., Pinilla, C., Markovic-Plese, S., Hemmer, B., Tzou, A., Whitney, L.W., Biddison, W.E., Martin, R., and Simon, R. (2001). Combinatorial peptide libraries and biometric score matrices permit the quantitative analysis of specific and degenerate interactions between clonotypic TCR and MHC peptide ligands. J. Immunol. 167 , 2130–2141.

1.5.7 Copyright

1.6 Reagent, sequencing & computational considerations

High throughput sequencing is poised to change all aspects of the way antibodies, T-cell receptors, and other binders are discovered and engineered. In synthetic immunology, millions of available sequence reads provide an unprecedented sampling depth able to guide the design and construction of effective, high quality naïve libraries containing tens of billions of unique molecules. Furthermore, during selections, high throughput sequencing enables quantitative tracing of enriched clones and position-specific

guidance to amino acid variation under positive selection during antibody engineering. Successful application of the technologies relies on specific PCR reagent design, correct sequencing platform selection, and effective use of computational tools and statistical measures to remove error, identify antibodies, estimate diversity, and extract signatures of selection from the clone down to individual structural positions. Here we review these considerations and discuss some of the remaining challenges to the widespread adoption of the technology. We place particular emphasis on the controlled setting of phage display libraries: in-vitro fixed repertoires that are of considerable value in generating binding reagents, but also powerful control systems for optimizing repertoire sequencing technologies without having to contend with the intra-individual differences, tissue distributions, and dynamic change of repertoire composition over time that is observed in in-vivo repertoires.

1.6.1 Introduction

Next generation sequencing (NGS) has transformed genomics. Its impact in antibody and TCR receptor research has been slower, but is likely to be equally disruptive. In many ways, the display technologies and deep sequencing are approaching a perfect match as sequencing technologies improve. For synthetic antibody library analysis, total numbers of bases sequenced is less important than the number of reads and their length. Present sequencing technology is able to generate up to 40 million reads from a single MiSeq run (figure 1). A an antibody repertoire, T-cell repertoire, or displayed antibody library could potentially have a diversity at least 100 fold greater ($\geq$10^10), the true diversity of which can be estimated using the methods described below. However, once these libraries are subject to selection by phage or yeast display, diversity is reduced to ~10^6 after a single round, allowing comprehensive analysis of the complete diversity of dozens of different selections in a single MiSeq run. After two or more rounds of selection, diversity is reduced still further, and the percentage of positive clones increases significantly; making analysis of $\geq$100 selections in a single run relatively straightforward. Read lengths vary, depending upon the technology (figure 1). Although 454 and PacBio provide the longest reads, the higher read

number and low cost have made paired end MiSeq (2x300bp) or Ion Torrent (400bp) sequencing the most commonly used for library analysis. While MiSeq will completely cover variable domains, encompassed by ≤600 bp (e.g. single Ig-like domain – VH domain of a scFv, camelid VHH's or fibronectin domains, smaller DARPINs, affibodies), it is presently insufficient to completely cover both the VH and VL chains found in an scFv in a single read. We expect this problem to be overcome as read lengths increase with further technology development.

The convergence of these technologies is important in structural biology for the increased use of antibody fragments [1] and other binders [2–4], as crystallization chaperones. While such chaperones were originally derived from immunized animals, recombinant display techniques using immunized or naïve binder sources as starting materials has broadened the nature of molecules used to include synthetic recombinant Fabs [5,6], designed ankyrin repeat proteins (DARPINs) [7–9], fibronectin domains [10] and nanobodies [11]. Any method that simplifies the generation of suitable crystallization chaperones is to be welcomed, and it is anticipated that the combination of NGS with display technologies will facilitate the development of effective chaperones, particularly if selection strategies can be specifically designed to select such molecules directly.

Here we review the technology and the informatic analyses required before describing the insights that can be gained from the use of next generation sequencing in library selection projects.

1.6.2 Sequencing technologies

The ability to assess the entire diversity of an antigen-specific sub-library allows the identification of all unique species in a sub-library, independently of their relative enrichment during the selection process. In fact, the wide span of relative abundances within a selected population is a known bias in the random screening process [12,13]. NGS technologies can successfully interrogate, at the deepest levels, theoretically every individual molecule, hence their increasing use in the screening of selected sub-libraries.

Several NGS platforms, each with specific advantages and, usually, preferred applications, are available. As a general consideration, read length and depth of sequencing are inversely proportional: technologies that provide the longest read lengths have the lowest throughputs, and vice versa for platforms that favor depth over length. The choice of sequencing usually lies in the nature of the selected library to be investigated: single scaffold synthetic libraries are generally easy to analyze because the sequencing can be focused on that limited region of the scaffold molecule that encodes the diversity. NGS platforms generating short reads are preferred in this case. Antibody-derived molecules, such as scFvs (single chain Fragment variable), represent a more complex scenario, where diversity is spread along a ~800bp-long gene. In this specific case, full-length gene sequencing would be ideal. Sanger sequencing provides sufficient read length, to cover the entire gene, but the low-throughput and high expense only allows a very limited snapshot of the true diversity of the selection process. At present among the different NGS platforms, PacBio is the only one able to provide sufficiently long read lengths to cover the entire scFv, but to the detriment of throughput [14] and read quality, where the ability to discriminate minimal sequence differences with certainty[15] has had limited adoption.

For selection projects, where depth of sequencing is preferred, platforms that provide shorter reads become the obvious alternative. However, this imposes a choice on the region of the gene that is to be analyzed. Roche's 454 and Illumina's MiSeq paired-end sequencing allow the coverage of the entire VL or VH domain [14,16]. The cheapest and fastest sequencing runs are provided by IonTorrent, and MiSeq single or paired-end reads; here, the main drawback is represented by the read length, which in the case of the single reads is sufficient to only partially cover one of the domains. In this case, the general consensus is to analyze the heavy chain VH domain, as it contains the complementarity determining region 3 (CDR3) as the primary signature of clonality, as well as amino acid variation in H1 and H2, framework mutations, biochemical liabilities in the variable domain and the identity of the V-gene and J-gene scaffold elements. As shown in Figure 1, CDR3s have the highest variability of all CDRs in both variable light (VL) and heavy (VH) domains, with HCDR3 being considered the principal determinant of specificity in antigen binding and, consequently,

a surrogate for scFv identity in a naive library[17,18] due to its diversity in length and aa composition[19,20]. The light chains are often sequenced as well, but given their relatively low diversity, are usually insufficiently diverse to reliably and uniquely indicate clonality. Efforts to utilize paired-end reads in the VH and the VL could provide a means of tracking VL diversity for each HCDR3. Table 1 summarizes the features of some of the most popular NGS platforms in selection projects.

The genetic material used to perform the NGS analysis on scFv-based libraries is usually a plasmid preparation of the selected sublibrary, from which the relevant immunoglobulin coding regions can be extrapolated by PCR amplification; the relevant coding regions, can consist of an entire scFv, an entire domain or just a portion of it, according to the chosen platform.

The amplification step requires the design of primers complementary to the target regions of interest; when the entire scFv gene or domain are sequenced (by PacBio, 454 or MiSeq paired ends), the use of external primers (mapping on the plasmid or linker flanking regions) is to be preferred, as these anneal to constant sequence elements. This provides the least biased means of amplification and makes the entire variable domain accessible. Shorter amplicons have also been designed around specific regions of interest by using multiple primers mapping upstream of the desired region (i.e. HCDR3), but these primers need to be carefully designed, in order to avoid amplification bias: due to the diversity of the antibody variable region frameworks, the primers are usually designed for families of antibody genes (consensus sequences) able to detect the highest number of gene segments[21]. As physical amplicon tolerances and read lengths increase, invariant vector primers have emerged as a standard.

As a general consideration, the primers are designed to allow the amplification of the target sequence and to carry adapters. These are platform-specific sequences that allow: i) annealing of sequencing primers; ii) anchoring of the amplicon to beads or other solid substrate during sequencing; iii) amplification of the single DNA molecule on the solid substrate. The primers can be further modified to allow multiplexing: the ability to sequence multiple selection outputs in a single run (Figure 2). Different selections are distinguished from one another with short sequences, unique for each selection inserted into the primers. This allows for significant reductions in costs as

(except for naïve library analysis) platform throughputs vastly exceed diversity in most selection outputs, providing sufficient depth to allow comprehensive analysis.

Multiplexing is achieved by adding unique DNA sequences (usually 6–8 bp) at the 5´ end of the gene-specific region on the primer (Figure 2). The sequencing of the barcode, along with the gene sequence, allows for the association of a read to a specific sub-library within the sequenced sample. Over 100 samples can easily be barcoded using such schemes, and by extending the barcode length multiplexing can be extended still further, and arbitrary numbers of samples could theoretically be generated, with primer costs becoming the primary limitation. For high number of multiplexed samples, the most efficient method is combinatorial barcoding: different barcodes are added to each end of the sub-library-specific amplicon, thus allowing for hundreds of different sub-libraries to be pooled and sequenced in a single run. The NGS platforms most suitable for multiplexing are IonTorrent and Illumina, due to their higher sequencing depths: a 10-fold coverage of the estimated sublibrary diversity is a desirable feature that allows the minimization of the effects of PCR amplification and sequencing errors. Alternatively, when evaluating selections, a simple rule of thumb can be applied: every sample should be performed in replicate to a depth of 100k reads. The 100k read depth will allow any sequence occurring at a frequency of 1e-4 (one in 10k reads) to be observed 10 times in each replicate on average, irrespective of how diverse the background library may be. As a consequence, treating 1e-4 as a threshold of meaningful enrichment, this simple rule allows all samples from all panning rounds to be processed identically, sequenced to equal depth, and analyzed in a comparable manner.

For antibody research a key aspect is to obtain the entire sequence of the highly diverse antibody variable regions, which allows precise definition of the antibody [22]. At present, none of the existing NGS platforms can provide sufficient accuracy and read length to characterize full-length scFv genes in a large sub-library. To overcome these limitations, some methods have been proposed and successfully applied to the characterization of naïve and immune repertoires: in one instance [14,23], two independent MiSeq paired-end sequencings have been used to sequence the entire VL and VH domains, while a third sequencing, coupled with appropriate bioinformatics tools

(discussed in the next paragraph), aims to bridge the VL and VH pair. Alternatively, the sequencing of a full-length scFv could be achieved in a single run by using the same "bridging" approach, with 2 primers sequencing sequentially from the 2 ends, and a third primer (or set of primers) bridging the gap starting from the framework region 3 of the VL domain. The method is yet to be tested for feasibility.

Roche's 454 is able to generate long reads, currently around 700 bp, making it well suited for V region analysis, with the limitations of a limited number of reads (Table 1) and significantly higher cost per run. MiSeq is cheaper, has much higher throughput and a faster turnaround compared to 454. However, read lengths are shorter, making it more appropriate for analyses of single variable domains. The reads obtainable by Ion Torrent range from 35 to 400 bp, enough to cover the CDR3 region as well as a single V domain (which will provide sequences of the other CDRs and help identify the antibody family). The lower quality of the sequences and current read length are the major drawbacks, while the main advantages are speed and low price per run.

Pacific Bio is a single DNA molecule sequencing platform that gives very long reads, but with error rates >10%. While this is not a problem for genome assembly, where it is now usually combined with other platforms [24–26], it has not been used successfully for antibody analysis in a single pass mode. More recently, accuracy has been significantly increased by circularizing DNA and sequencing it multiple times [15]. However, throughput remains relatively low.

In a recent paper [21] we compare the use of 454, MiSeq and Ion Torrent to sequence the same antibody library samples, and find that each method has its advantages, as outlined above.

1.6.3 Bioinformatics toolkits

The high-throughput sequence analysis of both naïve libraries and antibody library selections follows a well-established series of common steps. First, any paired-end reads are assembled. Next, all reads are screened to distinguish reads bearing antibody-like content from off-target content [20]. Antibody analysis typically begins with identification of the V, D and J region segments found in each antibody using a known

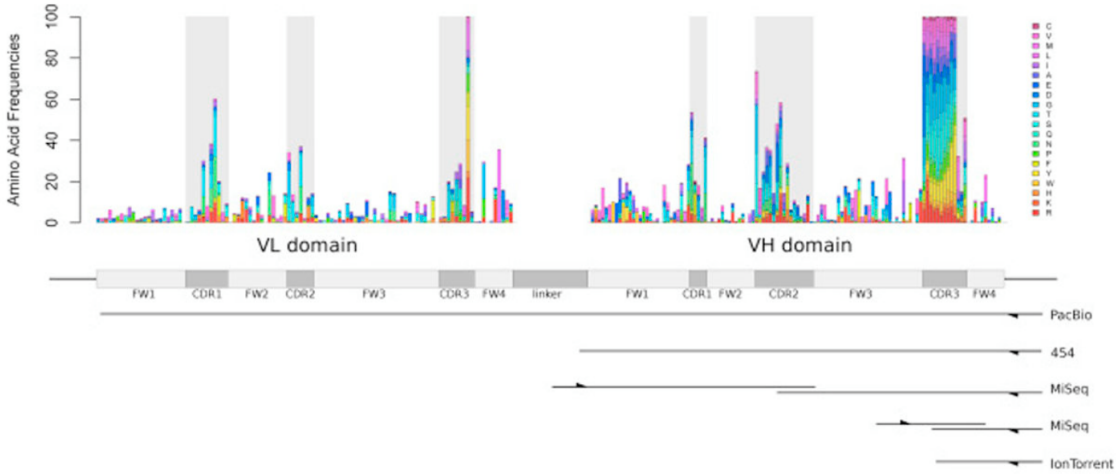


Figure 1.26: NGS sequencing on scFv genes. Variability plots for representative VL and VH genes are shown, with the CDRs shaded in grey. Length coverage for the most popular NGS platforms and scFv-based libraries targeted regions are shown. For each platform, single or double directional arrows indicate single or paired-end sequencing, respectively.

germline reference set (all methods). This reference set can be from any source, but in practice is most often obtained from IMGT [27]. Identification of reference segments enables somatic hypermutation analysis and statistical analysis of selective forces favoring individual segments [28]. Segment analysis is insufficient for selection interpretation, as antigen-driven selection acts not on the genetic elements directly, but rather on the translated CDRs. Consequently, it is critical to identify the correct frame of translation and obtain the translated CDR sequences.

The analysis of selection outputs is more straightforward than naïve library analysis. Once individual reads have been analyzed, the relative abundance of all clones in a selection is of great interest. To ensure accuracy, multiple sources of error, including PCR error, read error, bioinformatic classification ambiguity and variable read lengths, need to be accounted for. Once error processing is complete, clonal clustering can be used to gather de-facto identical clones, trace affinity maturation lineages of related clones, and even identify convergent paratopes emerging after a selection. The analysis of the clonal enrichment and clonal diversity of this final data can be

used to trace individual clones across different panning rounds or selection pressures, estimate the total diversity of responding clones after a selection, and even estimate the total diversity of the selection output and the functional fitness of every amino acid at every homologous position [18,29].

In the identification of the best lead antibodies to pursue, the translation of frameworks and CDRs additionally enables annotation of biochemical and immunological liabilities – these include non-synonymous framework mutations, N-linked glycosylation sites, deamination sites, acid hydrolysis sites, free cysteines, known aggregation and destabilizing variation, domain truncations and other gross-defects caused by library assembly.

The analysis of naïve unselected libraries, and assessments of total diversity, is far more challenging, since total potential diversity (i.e. all VH+VL combinations) almost always exceeds deep sequencing capacity ($<10^8$) (figure 3). Accumulation analysis (i.e. counting number of unique clones observed) provides a lower bound estimate of how many sequences exist in a library, and is necessarily incomplete when the entire scFv or Fab fragment cannot be observed in a single sequence. While accumulation analysis of diversity on individual CDRs is effective for H1, H2, L1, L2 and L3, the H3 diversity alone can easily be greater than that of sequencing depth, and in our experience is probably increased ten fold when diversity in the remaining CDRs and frameworks are accounted for [20]. Furthermore, extrapolation of total library diversity from individual CDR observations requires either strong assumptions of positional independence, or sophisticated mathematical models of positional relationships. More effective measures for library diversity estimation can be borrowed from field ecology – the Fisher's capture recapture [30] and the Chao statistic [31] can both be used to estimate the number of unseen species on the basis of the number and diversity of observed species, although both will likely return lower bound estimates. To complement lower bound estimates, a higher-bound estimate can be obtained by saturation analysis: subtracting the fraction of the repertoire taken up by observed high frequency clones. However, these species richness estimators are hindered by the presence of errors that inflate the number of rare species in the dataset (see next section). Used together, the methods provide a low-bound and high-bound

of diversity, allowing for a sensitive detection of library defects that reduce effective library size below 109. (Figure 3).

1.6.4 Annotating receptor sequences

The diversity of repertoires poses a number of unique bioinformatics challenges, compared with most other high-throughput sequence analysis applications (genomic sequencing, transcriptomics, chip-SEQ, microbiome analysis, virome analysis, etc). These involve mapping tens of millions of reads to a relatively finite reference set of segments that is on the order of thousands to hundred of thousands of segments (genes, exons, cDNAs, bacterial and viral genomes). In contrast, an antibody library can easily contain a billion antibodies, drawn from a VDJ rearrangement capacity exceeding 100 trillion possible combinations in most known organisms [32], and a nearly infinite molecular diversity (10^50) when considering somatic hypermutation, or synthetic libraries created by highly diverse oligonucleotides. As a consequence, some mapping shortcuts cannot be performed, and each read must be analyzed individually at early stages of repertoire analysis. This additional computational burden is addressed either through distributed computing, typically on commercial cloud computing environments, or research into novel parsimonious algorithms (regex-based methods [21], the VDJ challenge [33]).

Segment identification is confounded by somatic hypermutation, codon-reoptimized frameworks, read error, incomplete reads, and incomplete allele coverage. Incorrect segment assignment is best avoided, as it can lead to artificial separation of variants of the same clonal family in downstream analysis. For natural encodings using segments from a wellcharacterized reference species, the majority of reads can be reliably identified by best-blast based methods [34]. Read error will have little effect on segment classification, as the majority of segments are easily distinguishable even given the burden of read error typical of high throughput sequencing technologies. To improve accurate identification of segments in clones with high degrees of somatic hypermutation or incomplete reads, a probabilistic classifier can be used to assign confidence to the top hit, and reduce the resolution of assignment

when necessary [20,28]. When working with codon reoptimized frameworks, a novel reference segment database or segment assignment at the amino acid level can be attempted. The D-segment, given its short length and vulnerability to trim back, can often not be reliably classified in even the best circumstances.

Multiple analysis toolkits exist for the analysis of antibody library selections. The LANL Antibody Mining toolbox [21] operated through nucleotide-level pattern recognition of HCDR3 boundary elements. It is limited to frequency analysis of CDR3 sequences from naturally encoded human libraries, but is the fastest of all of the above algorithms, able to parse millions of reads in less than a minute, and thus well-suited for analyzing enrichment of CDR-H3 clones from a naïve library within minutes from a single computer. Most of the other toolkits provide analysis of segment identities and VDJ junctional boundaries and alignments, but at an additional computational cost that requires distributed computing capabilities to process millions of sequences efficiently. iHMMunalign uses a nucleotidelevel Bayesian Hidden Markov Model to assign probabilities to segment identities and VDJ junctional boundaries [35]. The NIH/NIAID/CIT Joinsolver operates through a combination of nucleotide-level CDR3 boundary motif recognition and parsed segment alignments [36]. IMGT's V-Quest benefits from wide breadth, performing analysis on BCRs, TCRs and multiple species including human, mouse, rat rabbit and pig, as well as a powerful reference database, classifying input sequences to their definitive reference IMGT set [37]. IMGT has also expanded their analysis support by offering High V-Quest, a web-based NGS compatible version currently able to handle up to 500k batches of sequencings, providing full annotation of VDJ segments, CDR regions, and somatic hypermutations [38]. The NIH Igblast performs blast-blast segment classification and boundary recognition [34]. The VDJFasta algorithm is the most generalizable tool, adapted for continued utility when analyzing very mutated antibodies, engineered antibodies and novel species [20,28,29]. VDJFasta utilizes amino acid profile Hidden Markov Models to identify CDRs and performs alignments by amino acid homology, then assigns segments by a probabilistic classifier. It is able to operate on any species without requiring a segment reference, as well as codon reoptimized frameworks and other heavily engineered monoclonal libraries. It can be run

either in a very fast CDR3 discovery mode, or in a more complete analytical mode that recovers alignments, segments, CDR boundaries and biochemical liabilities. It provides affinity maturation tree construction as an embedded feature of the toolkit, unique among the other tools. ImmunediveRsity, is another stand-alone pipeline for the analysis of antibody repertoire data, providing quality filtering, noise correction and repertoire reconstruction based on VDJ assignment, clonal origin and unique VH identification. Finally, the very recently published open-access software MiXCR uses an advanced alignment algorithm that enables rapid annotation of germline segments, CDRs, SHMs, and error correction (see next section for more), processing 10^7 sequence reads in minutes [39].

In addition to academically available command line resources, a set of industrial platforms, including Adaptive Analytics and the Distributed Bio AbGenesis platform, have also emerged as solutions for non-technical users. Open source community portals, such as receptormarker.com, have emerged as free academic user interfaces for specific applications [40].

The amino acid diversity of immunoglobulins presents a challenge for accurate CDR identification: even 10% of the Kabat database is estimated to be mis-numbered by their own classification system [41]. Motif-based CDR boundary recognition methods can often be used to recognize CDRs, but they will fail in more heavily mutated antibodies, engineered antibodies, and antibodies from novel species. Profile Hidden Markov Model (HMM) based Bayesian methods have emerged as powerful tools for CDR recognition, given their ability to recognize homology signatures of the frameworks to aid in contextualizing CDR diversity [20]. Such tools can operate at the nucleotide [35] or amino acid [20] level. However, they tend to be slower than motif based CDR recognition, and require longer CDR flanking sequences to function, when considering naturally encoded human antibody libraries [21]. HMMs provide a substantial advantage when analyzing codon optimized libraries, novel species, or highly affinity matured antibodies such as broadly neutralizing HIV repertoires, as the amino acid homology signatures are more robust to mutation and do not require nucleotide motif re-definition with each new species.

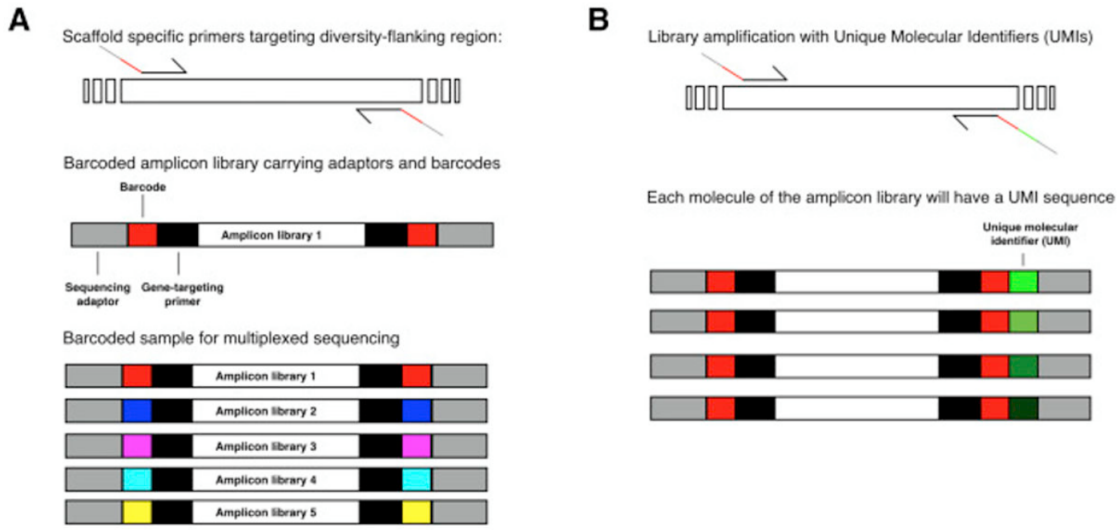


Figure 1.27: A) Schematic representation of NGS barcoded primers. Primers on conserved regions flanking the diversity carry barcode sequencing and NGS-specific adapters. The PCR product contains a unique identifier (barcode) for a specific library. Multiple amplicon libraries can be pooled in a unique sample and sequenced together. Each single sequenced DNA fragment is associated to a specific library based on its barcode. B) An amplicon library can also be generated with the inclusion of a Unique Molecular Identifier (UMI) sequence, which are added at an initial step (e.g., first-strand cDNA synthesis) resulting in each molecule of a library being tagged with a UMI. Similar to A), amplicons can be sequenced in a multiplex fashion but following NGS, sequences with identical UMIs are grouped together for consensus building-based error correction.

1.6.5 Error correction

Another challenge in repertoire analysis, particularly the assessment of naïve library diversity, is the presence of errors, as all sequencing technologies are susceptible to read error [42]. Compounding errors introduced by the sequencing platform is the fact that antibody library preparation requires PCR amplification, where DNA polymerase can also produce additional errors. Furthermore, unlike genome or transcriptome sequencing where errors can often be simply corrected by read-based consensus building or alignment with a reference genome/transcriptome, antibodies undergo somatic hypermutation making it difficult to distinguish between technical errors versus true biological mutations without an informatics-based error correction method.

Fortunately, the analysis of antibody repertoires, and in particular selections, is somewhat unique in that many types of read error have minimal impact on quantifiable features. Read error has little effect on segment identification, as it introduces proportionally less variation than somatic hypermutation into the underlying sequences. In analysis of positional amino acid frequencies in a total library, the read error rate will typically introduce less than 1% noise to observed frequencies. In selections, the reads of greatest interest will have the greatest depth of coverage, having expanded in the pool and thus receiving greater proportional read depth.

The greatest challenges lie in analysis of non-expanded clones or accurate total repertoire diversity. Errors in the CDR3 region could alter clonal diversity measurements if 100% CDR3 identity is used as the definition for clonality. The fact that CDR3 regions themselves are considered hotspots for somatic hypermutation adds further complexity to this problem. Recently the impact of errors was comprehensively evaluated where high-throughput sequencing (Illumina HiSeq) was performed on a "model repertoire" consisting of seven monoclonal cell lines expressing antibodies or TCRs [43]. Following, sequencing and annotation of CDR3 regions, there was a large number of false positive clones detected, which would have resulted in a drastic overestimation of clonal diversity. Indeed our results corroborate this. Following duplicate (using two different barcodes) Ion Torrent sequencing of the HCDR3 of a single VH region >99.5% of the ~80,000 sequences in each were correct. The remaining 0.5% comprised 166 unique HCDR3 false positives. ~40% were found in both barcodes,

and the remaining sequences were unique to each barcode. Although the majority of these were 1, 2 or 3 amino acid mutations away from the original HCDR3, there were also a number of unrelated HCDR3s, thought to be the result of contamination (unpublished data Bradbury group).

Despite the presence of errors, simple methods of informatics processing and filtering can be utilized to achieve partial correction. One example is CDR-based clustering at the amino acid level (used in the example above), a method that minimizes the footprint of read error to non-synonymous mutations in the paratope and accounts for the majority of read error, which will be single nucleotide mutations away from a true clone [20]. Another is frequency-based consensus building. At the depth of sequencing now available, higher frequency clones will often result in hundreds (or thousands) of reads, while the majority of their read error variants will typically appear as singletons that are often only a single nucleotide away from the correct higher-frequency read. Thus these singletons can be dropped from analysis or corrected by consensus alignment to the higher frequency clone [16]. Another method to alleviate overestimation of clonal diversity is to apply clonotyping, which is the grouping of similar CDR3s (e.g., 80%, 90% identity) [44], although, measuring intraclonal diversity or the number of somatic variants would still not be possible. Treebased single-linkage clustering methods are useful in libraries derived from natural repertoires where in vivo affinity maturation will generate complex SHM trees as well as read errors that may not resolve accurately by other clustering approaches [28]. In addition, another approach to partially overcome errors would be to perform replicate sequencing, in such a case only clones present in both replicates would be considered reliable [45–47]. However, this would not correct for reproducible sequencing or PCR hotspot errors; for example in our aforementioned sequencing experiment that resulted in 166 false positive HCDR3s, common false HCDR3s were found in both datasets. Others have also observed this phenomenon of reproducible systematic errors in NGS [43,48]. So while these methods are easy to implement and do improve repertoire accuracy, they still fail to provide fully accurate measurements of somatic hypermutation or clonal diversity, thus making more advanced methods necessary.

In order to correct for errors in NGS, several variations of an advanced method have been developed which rely on library preparation with unique molecular identifiers (UMIs, which are also known as unique identifiers, barcodes, molecular identifiers groups, primer IDs); UMIs are a stretch of degenerate nucleotides (e.g., NNNNNNNN) that are typically added to mRNA or cDNA molecules via reverse transcription or ligation [49,50]. Thus when PCR or sequencing introduces errors, these can be corrected by grouping sequences that share a common UMI and correcting variant sequences to the group's consensus sequence [51] (Figure 2B). The consensus is typically the correct sequence since all sequences with a common UMI are assumed to be derived from the same original template molecule. Recently UMI addition has also been applied for antibody repertoire sequencing. In one example, UMIs were incorporated into forward and reverse primers and added during first and second strand cDNA synthesis, which was combined with replicate sequencing to improve the accuracy of human B cell repertoires obtained from vaccinated individuals [52]. While UMI-based consensus building enables correction of sequencing errors, it does not address all PCR errors. For example, a polymerase-introduced error in an early PCR cycle that ends up becoming the majority positional nucleotide for that UMI group would result in a false consensus built sequence variant. While this might be considered a rare event, several reports have identified that this is more common than once believed [48,53]. To date the only method that has been developed to correct for PCR errors is based on UMI-labeling of cDNA followed by a read-gain/loss secondary correction (filtering) step [43]. Here, the original sequences or clone read counts are compared to those after consensus building. Since these early PCR errors tend to be systematic and reproducible, they will often appear in multiple UMI groups in later rounds of amplification. This phenomena leads to an overall greater number of erroneous variants being corrected, resulting in a net loss of erroneous sequence variants. This method when applied to a control repertoire was able to achieve nearly absolute error correction (removal of all false positive CDR3 variants) [43]. All UMIbased error correction methods require oversampling of UMIs (each UMI ≥ 3 reads). However, this has been challenging to accomplish, as it either requires a very high read depth or precise sample preparation and quantification methods to achieve adequate, but not excessive

oversampling. This sample preparation precision has yet to be fully standardized for antibody sequencing. Finally, it has also emerged that errors introduced to the UMIs themselves are substantially present and will thus need to be addressed in the future [53,54].

1.6.6 Repertoire size estimation

The size of naïve antibody libraries has been generally assessed by counting the number of bacterial colonies on a dilution plate after transformation, and multiplying accordingly, making the assumption that each bacterial colony represents a unique antibody. While this assumption may be reasonable in synthetic libraries, in which potential diversity usually exceeds actual diversity by orders of magnitude, this may be less true for libraries prepared from natural sources in which clonal dominance may occur if the number of donors is limited. The ability to sequence millions of antibodies in a naïve library allows a far more accurate assessment of diversity. The heavy chain variable region, and in particular, the heavy chain CDR3, are considered to be the most important determinants of recognition [17,18,55], exemplified by experiments in which HCDR3 sequences from an anti-lysozyme VHH antibody (VHH) [56] were transplanted into neocarzinostatin [57] and sfGFP [58], conferring lysozyme binding activity. HCDR3's have even been harvested as diversity elements [59,60] and binders have been selected from libraries in which they provided the only diversity [18,58]. As a result, deep sequencing initially concentrated on HCDR3, expanding to VH as read lengths increased. One somewhat surprising result from naïve natural library sequencing [20,21], is that lower bound estimates of heavy chain diversity, 3x106 unique clones (40 donors, ref [21]) and as little as 2x105 (˜654 donors, ref [20]) paratopes differing by at least 2 amino acids to any other, assessed using different methods, are significantly less than the almost limitless potential diversity of human VH rearrangement [32], and far closer to the VH diversity found in any single person (106–7) [32]. Given that most VH sequences are unique to an individual [14,28], unlike VL sequences, which are more commonly public [14], one would expect the number of unique VH sequences in a natural library to increase with the number of donors. That

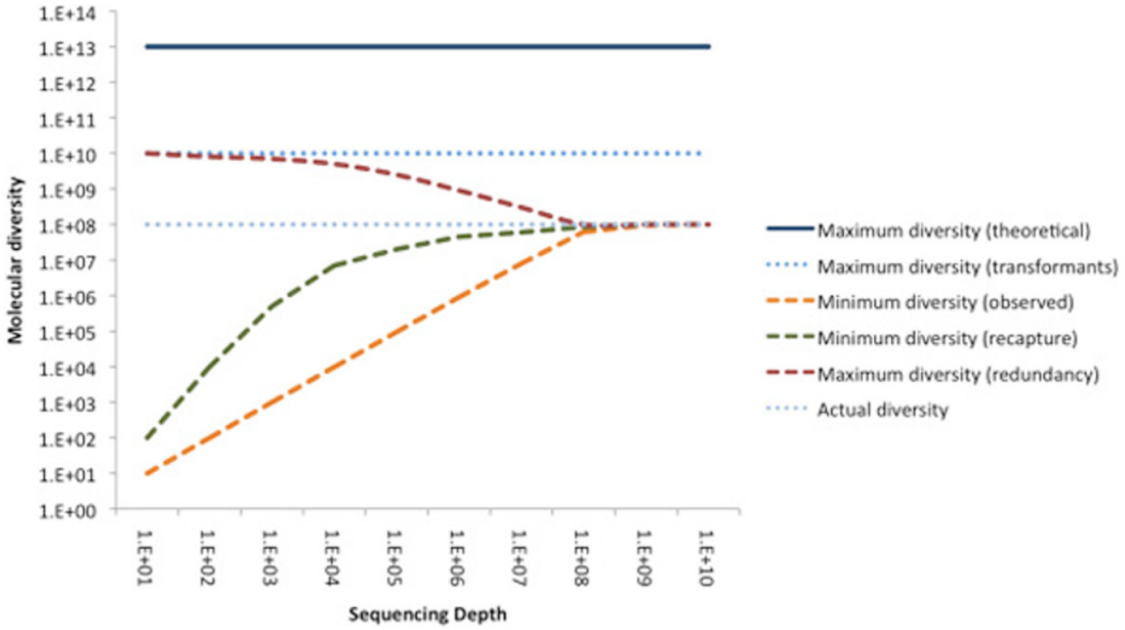


Figure 1.28: Estimating upper and lower diversity bounds as a function of sequencing depth. Maximum theoretical diversity is the total number of unique molecules that could exist in a library of this design if the number of transformants were infinite. Maximum transformant diversity is the maximum library size if every molecule in the library was non-redundant. Minimum observed diversity is the accumulated diversity observed from sequencing: the number of different clones actually seen. Minimum diversity estimated by capture-recapture methods more rapidly approaches the true diversity of the library, by anticipating library diversity from subsample overlap. Maximum diversity can be calculated by extracting known library redundancy from the transformation size. Actual diversity is the number of unique clones in the library. All measures convergence on true diversity with increasing sampling depth, although libraries with "long tails" of rare clones will converge slowly.

this does not appear to be the case is likely to be a consequence of library construction methods. In both the described cases of naïve natural library sequencing [20,21], V region amplifications were carried out on pools of B cell cDNA from different donors using pools of specific primers [61]. The extent of multiplex primer bias in repertoire sequencing has recently been carefully evaluated using 56 synthetic templates of all human TCR V-alpha genes, which revealed substantial bias, as in some cases entire V genes were not amplified at all [62]. Therefore multiplex PCR with pools of primers may be expected to bias amplification towards templates with more favorable primer-specific regions, as well as VH genes that are more abundant. One way to overcome bias would be to use both optimized primer concentrations and informatics correction, however this requires rather in-depth and sophisticated characterization studies [62]. Another possibility would be to use individual primer pairs on each individual, or small pools of individuals, rather than pooled, B cell cDNA. Although one very large library has been created using this approach [63], it has not been analyzed by deep sequencing.

For synthetic libraries, theoretical diversity in the heavy chain (but not usually the light chain) can vastly exceed the diversity achievable by bacterial transformation, depending on the design. This is confirmed by NGS: in an appropriately designed library, the vast majority of clones occurs only once, even when applying strict paratope distance measures to ensure that read error isn't artificially inflating the result [12,29,64,65].

For both natural and synthetic libraries, if VL and VH chains are assorted independently, library diversity increases enormously. This makes most sequences unique, and validates the colony counting approach to estimate library size. However, if construction methods (e.g. assembly PCR) are used where opportunities for clonal dominance exist, diversity may be overestimated by colony counting, and can only be assessed by NGS.

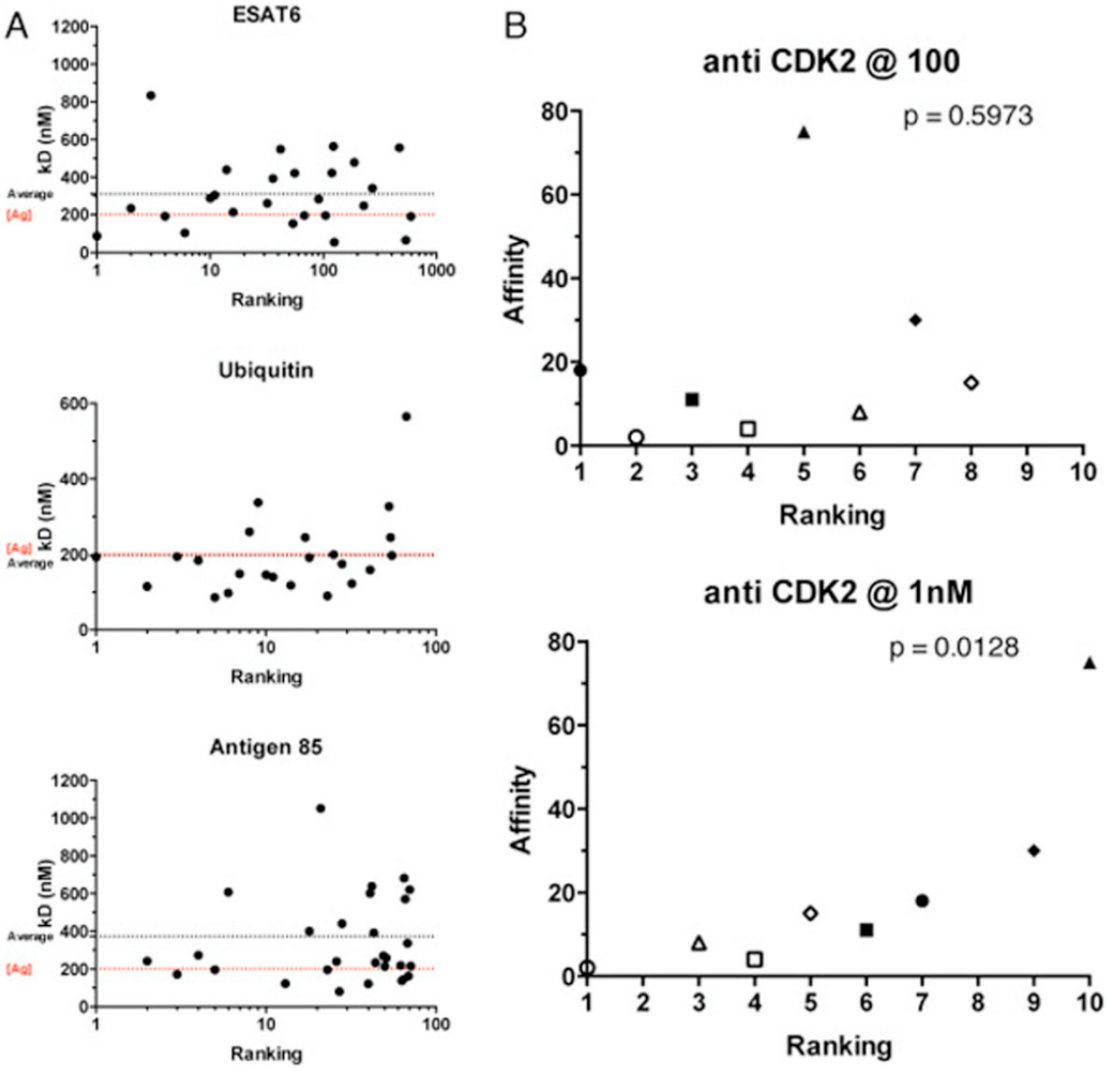


Figure 1.29: Relative abundance and affinity. Panel A: the experimentally measured kDs (nM) of selected clones are plotted in relation to their ranking position in the sequenced selection output (i.e. the clone in ranking position 1 has the highest relative abundance in the selection output). Average affinity of the clones (black) and antigen concentration used in the selection process (red) are shown as dotted lines. Data collected for three different antigen selections (ESAT6, Antigen85, and Ubiquitin) are reported. Panel B: ranking and affinity plots are shown for anti CDK2 selection at different antigen concentrations. The affinities of identical clones (identified by the same symbol in the 2 plots) found in sequenced populations selected at different antigen concentrations are shown in relation to their ranking position. P-values for significant correlation are reported.

1.6.7 Quality control and repertoire design

In addition to assessing diversity, NGS has been proposed as a straightforward method to quality control libraries after they have been created [64]. Sequencing provides accurate information on the percentages of clones that are non-functional due to stop codons or frame shifts; assesses how close actual diversity of synthetic libraries is to designed diversity; assesses the randomness of VH/VL linkage; and can assess library redundancy due to contaminants or clonal dominance.

Even when clones appear correct on the basis of sequence, they may be nonfunctional. For synthetic libraries, diversity may be functionally reduced by sequences that prevent correct folding, or are polyreactive, due to inappropriate amino acid choices in complementarity determining regions (CDRs) [66]. In the case of natural libraries, reduced functionality may be caused by excessively mutated V regions which display poorly [67], or by the restricted recognition properties of libraries with reduced VH diversity, given that VH is the major determinant of antibody recognition. The deep sequencing and analysis of well (and poorly) folding, or non-aggregating [68], antibody variable regions will result in the gradual accumulation of data on amino acid preferences at different positions (e.g. see refs [69,70]), which in turn will feed back into library design and more sophisticated functional quality control analyses that go beyond the mere identification of open reading frames.

For both library classes incompatible VH/VL pairs are also likely to reduce functional diversity. While there is likely to be significant individual variation, this may be mitigated by choosing known functional VH/VL pairs[65]. In addition to analyzing final library diversity, NGS can be used to monitor fidelity of components during construction [29]. It is expected that its continued use after each step of library construction will allow the direct analysis of the roles of different construction strategies in the generation, or loss, of diversity in the future, allowing more efficient library construction. The unexpected relatively low final VH diversity described in the two libraries above, could be better understood if NGS was applied to intermediate steps in the construction process, and indicate the insight NGS can bring to library creation methods.

Platform	Type of sequencing	Max read length (bp)	Throughput	Cost (lowest)	Accuracy	Time	Type of error
MiSeq (Illumina) v2/v3	2 x 300	600	$25x10^6$/lane	$1750/lane	>70% reads at 99.9%	55h	Substitution
	2 x 150	300	$16x10^6$/lane	$1100/lane	>80% reads at 99.9%	24h	
IonTorrent (LifeTech)-316	1 x 400	400	$2x10^6$/chip	$900/chip	> 99%	5 h	InDel
	1 x 200	200				3 h	
PacBio-RSII	1 x 8500	8500	47,000	$1050	11–15%	≤4 h	InDel
		e.g. 10 passes of 850 bp			99.999% depends on no. passes		
454 (Roche)-GS-FLWX+	1 x 700	700	50,000 in 1/8 plate	$2400/1/8 plate	99.997%	23 h	InDel
	1 x 450	450		$1900/1/8 plate	99.995%	10 h	

Table 1.5: Comparative features of NGS platforms

1.6.8 Selections

Target specific—One of the paradoxes in the early days of selection from antibody libraries was the inconsistency between the number of identified unique positive clones selected from large libraries, and the number of clones expected from theoretical calculations [71], or the scale up of the selection result from small libraries [72,73]. While the number of unique positive clones will depend upon the complexity of the antigen, the threshold affinity and the number of clones screened, practical experiments [72,73] indicate that it should be possible to select 1–5 positive clones from libraries with a diversity of 107, suggesting that libraries 100 fold greater in size (˜109) should yield 100–500 unique binders. In general, this has not been the case, unless extreme efforts have been taken to carry out selection under many different conditions [74]. Deep sequencing of selection outputs reveals that this is a combination of libraries not being as diverse as anticipated, and also that the recovery of unique clones poses a sampling problem: when only 96–384 clones are tested in different selection experiments, sometimes not even the ten most abundant clones can be identified [12,75]. Furthermore, when 96/384 clones are randomly picked, most are duplicates of abundant clones, while others represent single copies of far less abundant clones [76]. In our experience, such rare clones may individually comprise less than 0.001% of the selection output, and yet still be positive for the target, indicating that the only way to identify the full spectrum of binding antibodies after selection is to sequence and rank the complete output. This of course makes it even more paramount to apply

error correction methods, as otherwise true rare clones would not be able to be distinguished from errors. However, sequencing needs to be sufficiently deep that such clones are seen multiple times as read error correction cannot be carried out on single sequences. Sequence identification, however, is not the same as clone isolation. Once identified, clones can be isolated with inverse PCR [77], using the HCDR3 as a barcode for outward facing primers [78,79]. In order to reduce the numbers of primers required, arbitrary screening can be used initially. This usually provides many of the commonest clones, as well as a random selection of rarer ones, which, after individual sequencing, can be mapped back to the ranked list of antibodies. Inverse PCR can then be used to isolate missed clones.

NGS has also shown that the pattern of diversity found in selection outputs against different targets can be very variable. In some cases selections are dominated by single HCDR3s (or VHs), while in other cases, diversity is far broader. However, even when responses are relatively monoclonal, less abundant clones isolated by inverse PCR, are positive. Given their low abundance, NGS is the only way that these rare clones can be identified, as they cannot be found by standard screening methods. In these cases NGS is able to rescue selections that would otherwise have been considered failures due to their apparent limited diversities.

Identification of clones with desirable properties—Initial experiments in which antibodies were ranked for abundance after phage/yeast display selection and deep sequencing surprisingly revealed no correlation between affinity and abundance for all targets we have tested (fig 4a). In these experiments target concentration was kept relatively high (˜200nM), in order to preserve binding diversity, but as can be seen, the antibodies with the highest affinities (lowest Kd) are usually the less abundant ones. We believe this is because all antibodies with Kds lower than the target concentration used for sorting will bind approximately similar amounts of antigen, providing them with no selective advantage over antibodies with better affinities. As target concentration is reduced, only those yeast displaying antibodies with lower Kds are able to capture target. Further analysis revealed that at the lowest target concentration at which positive yeast can be identified by flow cytometry, there is a far better correlation between abundance rank and affinity (fig 4b). Consequently, we

believe NGS can be used to identify antibodies with the best affinities in a binding population, by sorting with diminishing target concentrations, and sequencing the output of the lowest target concentration that yields a positive population. Under these conditions, the most abundant antibodies will tend to be those with the lowest Kds. It is likely that similar approaches can be taken to similarly optimize individual steps in the selection process, including washing, temperature, incubation times and elution methods for phage display as well.

In addition to sequencing the outputs of target-specific selections, it will also be possible to apply deep sequencing to the analysis of common desirable antibody traits, such as thermostability [80], binding to protein A [81] or high display/expression levels [82], and other developability traits. We anticipate this can be carried out on individual target specific selections, or by sequencing complete libraries subjected to particular selection gates. In the latter case it may be possible to identify common sequence features correlated with desired properties, which could then be used to build and improve subsequent libraries, as described above. Such approaches could be powerfully combined with structural modeling and prediction using in silico prediction tools (e.g., Rosetta, MOE, Discovery studio) [83–86].

Identification of common clones—When display methods are used to generate antibodies, the selection targets are more complex than assumed. Antigens are usually biotinylated [87], which introduces additional complications: the presence of the biotin, the chemical moiety linking the biotin to the protein and the streptavidin (which itself may or may not be modified). Further, many targets are expressed recombinantly, and include common domains, such as peptide tags recognized by antibodies, fusion protein or His tags (see [88] for a review). All these additional common components can themselves, become targets for selection, potentially leading to antibodies that do not recognize the specific target but the common feature. Although appropriate controls and negative selections usually allow the elimination of such apparently cross-reactive antibodies, NGS can also be used to identify them after selection. In a recent paper [89], polyclonal antibodies selected from a large naïve library [90] created by recombination [91] using phage/yeast display [76,92] against a series of in vitro

biotinylated proteins were found to be strongly crossreactive with other targets. Careful analysis of the cross-reactivity revealed the polyclonal antibody pool recognized proteins biotinylated using a particular kit (Lightening Link), but not if biotinylated with other kits or in vivo. Deep sequencing of the antibody populations showing this cross-reactivity identified one common antibody in all the selections, which when tested, was found to recognize the Lightening Link biotinylation site [89]. A similar approach could be adapted to the identification of antibodies recognizing epitopes in common between different targets and enable informatics based library removal: e.g. human and murine versions of the same protein, or related therapeutic targets.

1.6.9 Acknowledgements

This work was made possible by my co-authors D'Angelo, Kan, Reddy, Naranjo, Ferrara, and Andrew Bradbury.

1.6.10 References

1 Ostermeier C, Iwata S, Ludwig B, Michel H: Fv fragment- mediated crystallization of the membrane protein bacterial cytochrome c oxidase. Nat Struct Biol 1995, 2:842-846.

2. Rasmussen SG, Choi HJ, Fung JJ, Pardon E, Casarosa P, Chae PS, Devree BT, Rosenbaum DM, Thian FS, Kobilka TS et al.: Structure of a nanobody-stabilized active state of the beta(2) adrenoceptor. Nature 2011, 469:175-180.

3. Lam AY, Pardon E, Korotkov KV, Hol WG, Steyaert J: Nanobody- aided structure determination of the EpsI:EpsJ pseudopilin heterodimer from Vibrio vulnificus. J Struct Biol 2009, 166:8-15.

4. Korotkov KV, Pardon E, Steyaert J, Hol WG: Crystal structure of the N-terminal domain of the secretin GspD from ETEC determined with the assistance of a nanobody. Structure 2009, 17:255-265.

5. Uysal S, Vasquez V, Tereshko V, Esaki K, Fellouse FA, Sidhu SS, Koide S, Perozo E, Kossiakoff A: Crystal structure of full-length KcsA in its closed conformation. Proc Natl Acad Sci U S A 2009, 106:6644-6649.

6. Ye JD, Tereshko V, Frederiksen JK, Koide A, Fellouse FA, Sidhu SS, Koide S, Kossiakoff AA, Piccirilli JA: Synthetic antibodies for specific recognition and crystallization of structured RNA. Proc Natl Acad Sci U S A 2008, 105:82-87.

7. Schweizer A, Roschitzki-Voser H, Amstutz P, Briand C, Gulotti- Georgieva M, Prenosil E, Binz HK, Capitani G, Baici A, Pluckthun A et al.: Inhibition of caspase-2 by a designed ankyrin repeat protein: specificity, structure, and inhibition mechanism. Structure 2007, 15:625-636.

8. Huber T, Steiner D, Rothlisberger D, Pluckthun A: In vitro selection and characterization of DARPins and Fab fragments for the co-crystallization of membrane proteins: the Na(+)- citrate symporter CitS as an example. J Struct Biol 2007, 159:206-221.

9. Kohl A, Amstutz P, Parizek P, Binz HK, Briand C, Capitani G, Forrer P, Pluckthun A, Grutter MG: Allosteric inhibition of aminoglycoside phosphotransferase by a designed ankyrin repeat protein. Structure (Camb) 2005, 13:1131-1141.

10. Koide A, Gilbreth RN, Esaki K, Tereshko V, Koide S: High-affinity single-domain binding proteins with a binary-code interface. Proc Natl Acad Sci U S A 2007, 104:6632-6637.

11. Low C, Yau YH, Pardon E, Jegerschold C, Wahlin L, Quistgaard EM, Moberg P, Geifman-Shochat S, Steyaert J, Nordlund P: Nanobody mediated crystallization of an archeal mechanosensitive channel. PLoS One 2013, 8:e77984.

12. Ravn U, Gueneau F, Baerlocher L, Osteras M, Desmurs M, Malinge P, Magistrelli G, Farinelli L, Kosco-Vilbois MH, Fischer N: By-passing in vitro screening – next generation sequencing technologies applied to antibody display and in silico candidate selection. Nucleic Acids Res 2010, 38:e193.

13. Di Niro R, Ziller F, Florian F, Crovella S, Stebel M, Bestagno M, Burrone O, Bradbury AR, Secco P, Marzari R et al.: Construction of miniantibodies for the in vivo study of human autoimmune diseases in animal models. BMC Biotechnol 2007, 7:46-55.

14. DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD, Georgiou G: In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. Nat Med 2014 http://dx.doi.org/10.1038/nm.3743.

15. Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW: A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res 2010, 38:e159.

16. Reddy ST, Ge X, Miklos AE, Hughes RA, Kang SH, Hoi KH, Chrysostomou C, Hunicke-Smith SP, Iverson BL, Tucker PW et al.: Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. Nat Biotechnol 2010, 28:965-969.

17. Xu JL, Davis MM: Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. Immunity 2000, 13:37-45.

18. Mahon CM, Lambert MA, Glanville J, Wade JM, Fennell BJ, Krebs MR, Armellino D, Yang S, Liu X, O'Sullivan CM et al.: Comprehensive interrogation of a minimalist synthetic CDR- H3 library and its ability to generate antibodies with therapeutic potential. J Mol Biol 2013, 425:1712-1730.

19. Zemlin M, Klinger M, Link J, Zemlin C, Bauer K, Engler JA, Schroeder HW Jr, Kirkham PM: Expressed murine and human CDR-H3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. J Mol Biol 2003, 334:733-749.

20. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, Ni I, Mei L, Sundar PD, Day GM et al.: Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proc Natl Acad Sci U S A 2009, 106:20216-20221.

21. D'Angelo S, Glanville J, Ferrara F, Naranjo L, Gleasner CD, Shen X, Bradbury AR, Kiss C: The antibody mining toolbox: an open source tool for the rapid analysis of antibody repertoires. MAbs 2014, 6:160-172.

22. Bradbury A, Pluckthun A: Reproducibility: standardize antibodies used in research. Nature 2015, 518:27-29.

23. Dekosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, Varadarajan N, Giesecke C, Dorner T, Andrews SF et al.: High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. Nat Biotechnol 2013, 31:166-169.

24. Kamada M, Hase S, Sato K, Toyoda A, Fujiyama A, Sakakibara Y: Whole genome complete resequencing of Bacillus subtilis natto by combining long reads with high-quality short reads. PLoS One 2014, 9:e109999.

25. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC et al.: Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One 2012, 7:e47768.

26. Ribeiro FJ, Przybylski D, Yin S, Sharpe T, Gnerre S, Abouelleil A, Berlin AM, Montmayeur A, Shea TP, Walker BJ et al.: Finished bacterial genomes from shotgun sequence data. Genome Res 2012, 22:2270-2277.

27. Ehrenmann F, Lefranc MP: IMGT/DomainGapAlign: the IMGT(R) tool for the analysis of IG, TR, MH, IgSF, and MhSF domain amino acid polymorphism. Methods Mol Biol 2012, 882:605-633.

28. Glanville J, Kuo TC, von Budingen HC, Guey L, Berka J, Sundar PD, Huerta G, Mehta GR, Oksenberg JR, Hauser SL et al.: Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. Proc Natl Acad Sci U S A 2011, 108:20066-20071.

29. Zhai W, Glanville J, Fuhrmann M, Mei L, Ni I, Sundar PD, Van Blarcom T, Abdiche Y, Lindquist K, Strohner R et al.: Synthetic antibodies designed on natural sequence landscapes. J Mol Biol 2011, 412:55-71.

30. Weinstein JA, Jiang N, White RA III, Fisher DS, Quake SR: High- throughput sequencing of the zebrafish antibody repertoire. Science 2009, 324:807-810.

31. Qi Q, Liu Y, Cheng Y, Glanville J, Zhang D, Lee JY, Olshen RA, Weyand CM, Boyd SD, Goronzy JJ: Diversity and clonal selection in the human T-cell repertoire. Proc Natl Acad Sci U S A 2014, 111:13139-13144.

32. Arnaout R, Lee W, Cahill P, Honan T, Sparrow T, Weiand M, Nusbaum C, Rajewsky K, Koralov SB: High-resolution description of antibody heavy-chain repertoires in humans. PLoS One 2011, 6:e22365.

33. Lakhani KR, Boudreau KJ, Loh PR, Backstrom L, Baldwin C, Lonstein E, Lydon M, MacCormack A, Arnaout RA, Guinan EC: Prize-based contests can provide solutions to computational biology problems. Nat Biotechnol 2013, 31:108-111.

34. Ye J, Ma N, Madden TL, Ostell JM: IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res 2013, 41:W34-W40.

35. Gaeta BA, Malming HR, Jackson KJ, Bain ME, Wilson P, Collins AM: iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. Bioinformatics 2007, 23:1580-1587.

36. Souto-Carneiro MM, Longo NS, Russ DE, Sun HW, Lipsky PE: Characterization of the human Ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOLVER. J Immunol 2004, 172:6790-6802.

37. Brochet X, Lefranc MP, Giudicelli V: IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V–J and V–D–J sequence analysis. Nucleic Acids Res 2008, 36:W503-W508.

38. Alamyar E, Giudicelli V, Li S, Duroux P, Lefranc MP: IMGT/ HighVQUEST: the IMGTs web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. Immunome Res 2012, 8:26.

39. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, Chudakov DM: MiXCR: software for comprehensive adaptive immunity profiling. Nat Methods 2015, 12:380-381.

40. Han A, Glanville J, Hansmann L, Davis MM: Linking T-cell receptor sequence to functional phenotype at the single-cell level. Nat Biotechnol 2014, 32:684-692.

41. Abhinandan KR, Martin AC: Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. Mol Immunol 2008, 45:3832-3839.

42. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ: Performance comparison of benchtop high- throughput sequencing platforms. Nat Biotechnol 2012 http:// dx.doi.org/10.1038/nbt.2198.

43. Shugay M, Britanova OV, Merzlyak EM, Turchaninova MA, Mamedov IZ, Tuganbaev TR, Bolotin DA, Staroverov DB, Putintseva EV, Plevova K et al.: Towards error-free profiling of immune repertoires. Nat Methods 2014, 11:653-655.

44. Wine Y, Boutz DR, Lavinder JJ, Miklos AE, Hughes RA, Hoi KH, Jung ST, Horton AP, Murrin EM, Ellington AD et al.: Molecular deconvolution of the monoclonal antibodies that comprise the polyclonal serum response. Proc Natl Acad Sci U S A 2013, 110:2993-2998.

45. Greiff V, Menzel U, Haessler U, Cook SC, Friedensohn S, Khan TA, Pogson M, Hellmann I, Reddy ST: Quantitative assessment of the robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. BMC Immunol 2014, 15:40.

46. Menzel U, Greiff V, Khan TA, Haessler U, Hellmann I, Friedensohn S, Cook SC, Pogson M, Reddy ST: Comprehensive evaluation and optimization of amplicon library preparation methods for high-throughput antibody sequencing. PLoS One 2014, 9:e96727.

47. Robasky K, Lewis NE, Church GM: The role of replicates for error mitigation in next-generation sequencing. Nat Rev Genet 2014, 15:56-62.

48. Brodin J, Mild M, Hedskog C, Sherwood E, Leitner T, Andersson B, Albert J: PCR-induced transitions are the major source of error in cleaned ultra-deep pyrosequencing data. PLoS One 2013, 8:e70388.

49. Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R: Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proc Natl Acad Sci U S A 2011, 108:20166-20171.

50. Lundberg DS, Yourstone S, Mieczkowski P, Jones CD, Dangl JL: Practical innovations for high-throughput amplicon sequencing. Nat Methods 2013, 10:999-1002.

51. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B: Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A 2011, 108:9530-9535.

52. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR: Genetic measurement of memory B-cell recall using antibody repertoire sequencing. Proc Natl Acad Sci U S A 2013, 110:13463-13468.

53. Brodin J, Hedskog C, Heddini A, Benard E, Neher RA, Mild M, Albert J: Challenges with using primer IDs to improve accuracy of next generation sequencing. PLoS One 2015, 10:e0119123.

54. Deakin CT, Deakin JJ, Ginn SL, Young P, Humphreys D, Suter CM, Alexander IE, Hallwirth CV: Impact of next-generation sequencing error on analysis of barcoded plasmid libraries of known complexity and sequence. Nucleic Acids Res 2014, 42:e129.

55. Kabat EA, Wu TT: Identical V region amino acid sequences and segments of sequences in antibodies of different specificities, Relative contributions of VH and VL genes, minigenes, and complementarity-determining regions to binding of antibody-combining sites. J Immunol 1991, 147:1709-1719.

56. Desmyter A, Transue TR, Ghahroudi MA, Thi MH, Poortmans F, Hamers R, Muyldermans S, Wyns L: Crystal structure of a camel single-domain VH antibody fragment in complex with lysozyme. Nat Struct Biol 1996, 3:803-811.

57. Nicaise M, Valerio-Lepiniec M, Minard P, Desmadril M: Affinity transfer by CDR grafting on a nonimmunoglobulin scaffold. Protein Sci 2004, 13:1882-1891.

58. Dai M, Temirov J, Pesavento E, Kiss C, Velappan N, Pavlik P, Werner JH, Bradbury AR: Using T7 phage display to select GFP- based binders. Protein Eng Des Sel 2008, 21:413-424.

59. Kiss C, Fisher H, Pesavento E, Dai M, Valero R, Ovecka M, Nolan R, Phipps ML, Velappan N, Chasteen L et al.: Antibody binding loop insertions as diversity elements. Nucleic Acids Res 2006, 34:e132.

60. Venet S, Kosco-Vilbois M, Fischer N: Comparing CDRH3 diversity captured from secondary lymphoid organs for the generation of recombinant human antibodies. MAbs 2013, 5:690-698.

61. Sblattero D, Bradbury A: A definitive set of oligonucleotide primers for amplifying human V regions. Immunotechnology 1998, 3:271-278.

62. Carlson CS, Emerson RO, Sherwood AM, Desmarais C, Chung MW, Parsons JM, Steen MS, LaMadrid- Herrmannsfeldt MA, Williamson DW, Livingston RJ et al.: Using synthetic templates to design an unbiased multiplex PCR assay. Nat Commun 2013, 4:2680.

63. Schofield DJ, Pope AR, Clementel V, Buckell J, Chapple S, Clarke KF, Conquer JS, Crofts AM, Crowther SR, Dyson MR et al.: Application of phage display to high throughput antibody generation and characterization. Genome Biol 2007, 8:R254. Demonstrates the utility of in vitro antibody selection at high throughput.

64. Ravn U, Didelot G, Venet S, Ng KT, Gueneau F, Rousseau F, Calloud S, Kosco-Vilbois M, Fischer N: Deep sequencing of phage display libraries to support antibody discovery. Methods 2013, 60:99-110.

65. Tiller T, Schuster I, Deppe D, Siegers K, Strohner R, Herrmann T, Berenguer M, Poujol D, Stehle J, Stark Y et al.: A fully synthetic human Fab antibody library based on fixed VH/VL framework pairings with favorable biophysical properties. MAbs 2013, 5:445-470.

66. Birtalan S, Zhang Y, Fellouse FA, Shao L, Schaefer G, Sidhu SS: The intrinsic contributions of tyrosine, serine, glycine and arginine to the affinity and specificity of antibodies. J Mol Biol 2008, 377:1518-1528.

67. Saggy I, Wine Y, Shefet-Carasso L, Nahary L, Georgiou G, Benhar I: Antibody isolation from immunized animals: comparison of phage display and antibody discovery via V gene repertoire mining. Protein Eng Des Select 2012, 25:539-549.

68. Dudgeon K, Rouet R, Kokmeijer I, Schofield P, Stolp J, Langley D, Stock D, Christ D: General strategy for the generation of human antibody variable domains with increased aggregation resistance. Proc Natl Acad Sci U S A 2012, 109:10879-10884.

69. Jung S, Spinelli S, Schimmele B, Honegger A, Pugliese L, Cambillau C, Pluckthun A: The importance of framework residues H6, H7 and H10 in antibody heavy chains: experimental evidence for a new structural subclassification of antibody V(H) domains. J Mol Biol 2001, 309:701-716.

70. Wang N, Smith WF, Miller BR, Aivazian D, Lugovskoy AA, Reff ME, Glaser SM, Croner LJ, Demarest SJ: Conserved amino acid networks involved in antibody variable domain interactions. Proteins 2009, 76:99-114.

71. Perelson AS, Oster GF: Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. J Theor Biol 1979, 81:645-670.

72. Marks JD, Hoogenboom HR, Bonnert TP, McCafferty J, Griffiths AD, Winter G: By-passing immunization. Human antibodies from V-gene libraries displayed on phage. J Mol Biol 1991, 222:581-597.

73. Griffiths AD, Williams SC, Hartley O, Tomlinson IM, Waterhouse P, Crosby WL, Kontermann RE, Jones PT, Low NM, Allison TJ et al.: Isolation of high affinity human antibodies directly from large synthetic repertoires. EMBO J 1994, 13:3245-3260.

74. Edwards BM, Barash SC, Main SH, Choi GH, Minter R, Ullrich S, Williams E, Du Fou L, Wilton J, Albert VR et al.: The remarkable flexibility of the human antibody repertoire; isolation of over one thousand different antibodies to a single protein, BLyS. J Mol Biol 2003, 334:103-118.

75. Di Niro R, Ferrara F, Not T, Bradbury AR, Chirdo F, Marzari R, Sblattero D: Characterizing monoclonal antibody epitopes by filtered gene fragment phage display. Biochem J 2005, 388:889-894.

76. Ferrara F, D'Angelo S, Gaiotto T, Naranjo L, Tian H, Graslund S, Dobrovetsky E, Hraber P, Lund-Johansen F, Saragozza S et al.: Recombinant renewable polyclonal antibodies. MAbs 2015, recombination. 7:32-41.

77. Hoskins RA, Stapleton M, George RA, Yu C, Wan KH, Carlson JW, Celniker SE: Rapid and efficient cDNA library screening by self- ligation of inverse PCR products (SLIP). Nucleic Acids Res 2005, 33:e185.

78. D'Angelo S, Kumar S, Naranjo L, Ferrara F, Kiss C, Bradbury AR: From deep sequencing to actual clones. Protein Eng Des Sel 2014, 27:301-307.

79. Spiliotopoulos A, Owen JP, Maddison BC, Dreveny I, Rees HC, Gough KC: Sensitive recovery of recombinant antibody clones after their in silico identification within NGS datasets. J Immunol Methods 2015 http://dx.doi.org/10.1016/j.jim.2015.03.005.

80. Orr BA, Carr LM, Wittrup KD, Roy EJ, Kranz DM: Rapid method for measuring ScFv thermal stability by yeast surface display. Biotechnol Prog 2003, 19:631-638.

81. Hillson JL, Karr NS, Oppliger IR, Mannik M, Sasso EH: The structural basis of germline-encoded VH3 immunoglobulin binding to staphylococcal protein A. J Exp Med 1993, 178:331-336.

82. Shusta EV, Kieke MC, Parke E, Kranz DM, Wittrup KD: Yeast polypeptide fusion surface display levels predict thermal stability and soluble secretion efficiency. J Mol Biol 1999, 292:949-956.

83. Sircar A, Kim ET, Gray JJ: RosettaAntibody: antibody variable region homology modeling server. Nucleic Acids Res 2009, 37:W474-W479.

84. Weitzner BD, Kuroda D, Marze N, Xu J, Gray JJ: Blind prediction performance of RosettaAntibody 3.0: grafting, relaxation, kinematic loop modeling, and full CDR optimization. Proteins 2014, 82:1611-1623.

85. Li L, Kumar S, Buck PM, Burns C, Lavoie J, Singh SK, Warne NW, Nichols P, Luksha N, Boardman D: Concentration dependent viscosity of monoclonal antibody solutions: explaining experimental behavior in terms of molecular properties. Pharm Res 2014, 31:3161-3178.

86. Sydow JF, Lipsmeier F, Larraillet V, Hilger M, Mautz B, Molhoj M, Kuentzer J, Klostermann S, Schoch J, Voelger HR et al.: Structure-based prediction of asparagine and aspartate degradation sites in antibody variable regions. PLoS One 2014, 9:e100736.

87. Hawkins RE, Russell SJ, Winter G: Selection of phage antibodies by binding affinity. Mimicking affinity maturation. J Mol Biol 1992, 226:889-896.

88. Malhotra A: Tagging for protein expression. Methods Enzymol 2009, 463:239-258.

89. Ferrara F, Naranjo LA, D'Angelo S, Kiss C, Bradbury AR: Specific binder for lightning-link(R) biotinylated proteins from an antibody phage library. J Immunol Methods 2013, 395:83-87.

90. Sblattero D, Bradbury A: Exploiting recombination in single bacteria to make large phage antibody libraries. Nat Biotechnol 2000, 18:75-80.

91. Sblattero D, Lou J, Marzari R, Bradbury A: In vivo recombination as a tool to generate molecular diversity in phage antibody libraries. Rev Mol Biotech 2001, 74:303-315.

92. Ferrara F, Naranjo LA, Kumar S, Gaiotto T, Mukundan H, Swanson B, Bradbury AR: Using phage and yeast display to select hundreds of monoclonal antibodies: application to antigen 85, a tuberculosis biomarker. PLoS One 2012, 7:e49535.

1.6.11 Copyright

Chapter 2

Reading Specificity with Convergence Analysis

2.1 Introduction

A memory of all our past immunological battles is stored in our blood. Unfortunately, extracting and interpreting immunological histories from the complex B-cell and T-cell receptor repertoires and cellular phenotypes has remained an unsolved problem. The problem has been one of diversity: while the human genome contains approximately 25 thousand genes, a single human's T-cells present approximately 100 million unique T-cell receptors. These receptors recognize a limited set of peptide antigens out of hundreds of trillions of possible variants, each displayed by a few out of 10 thousand possible HLA alleles strewn across human populations. As a consequence, common B-cell receptors and T-cell receptors are rarely encountered across individuals, and although sequencing technologies presented in Chapter 1 enable millions of receptors to be read in a single experiment, it has not been possible to recognize nearly any of them from previous studies and therefore it has not been possible to read specificity from the adaptive repertoire. What is needed is a statistical framework for interpreting systems that are more diverse than can be carried in a body.

In chapture 2, we present convergence analysis: a statistical method for decoding specificity from the complexity of the interacting molecules of the adaptive immune

system. In the first study, "Identifying specificity groups in the T-cell receptor repertoire," by Glanville and Huang in Nature 2017, we present the GLIPH, *Grouping of Lymphocyte Interations by Paratope Hotspots*, a clustering algorithm that is able to identify T-cell receptors that recognize the same antigen, predict the MHC restriction of those TCRs, and aid in the identification of their target antigen. In the second study, "A highly focused antigen receptor repertoire characterizes γδT cells that are poised to make IL-17 rapidly in naive animals," we expand the analysis of convergence TCRs to include γδ T cells, illustrating that a convergence of receptor is correlated to a convergence of phenotype. In the third study, "IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity," we demonstrate convergence analysis in B-cell receptors and antibodies, identifying a broadly neutralizing influenza HA-stem directed convergence group that can be identified from primary sequence, is predicated on allelic variation in the naive repertoire, and is predictive of seroconversion. In the fourth chapter, "Successful immunotherapy induces previously unidentified allergen-specific CD4+ T-cell subsets," we generalize the method of convergence analysis beyond amino acid variation at homologous molecular positions to any system where selection acts on only a subset of dimensionality, demonstrating that convergence of single-cell phenotypes can grant clarity to the analysis of oral immunotherapy for peanut allergy.

In the course of each study, multiple applications of convergence analysis are establish. First, it provides a means of clustering receptors and cells of common function and specificity. Second, it provides a means of "reading the repertoire:" predicting new members of the previously identified clusters. When the HLA genotype is available, the approach provides a method for predicting the MHC constriction, and rapidly de-orphaning TCRs. Finally, in both the TCR and BCR cases, it has demonstrated that the methods are sufficiently robust as to predict new de novo binders against targets, and in the case of TCRs, generate de novo binders with superior activation properties to the naturally observed receptors. This result opens up convergence analysis as a tool for optimization of receptors beyond their behavior encountered in nature.

2.2 Reading specificity in the αβ T-cell receptor repertoire

T Cell Receptor (TCR) sequences are very diverse, with many more possible sequence combinations than T cells in any one individual(1-4). To better understand how TCR antigen recognition is shared between individuals, here we define the minimal requirements for TCR antigen specificity, through a comprehensive analysis of 2068 unique TCRs of known specificity from 33 donors and a panel of pMHC tetramer sorted cells, as well as public TCR-pMHC complex structures. With this data we developed an algorithm - GLIPH, *Grouping of Lymphocyte Interactions by Paratope Hotspots,* to cluster TCRs that have a high probability of sharing specificity due to both conserved motifs and global similarity of CDR3s. After training on this data, we demonstrate that GLIPH could reliably group TCRs of common specificity from different donors, with the conserved CDR3 motifs that help define the TCR clusters often being contact points with the antigenic peptides. As an independent validation test set, we analyzed 5711 TCR β chain sequences from Mycobacterium tuberculosis (Mtb) reactive CD4+ T cells from 22 latent Mtb-infected subjects. We find 141 TCR specificity groups, including 16 distinct groups containing TCRs from multiple subjects. These TCR groups typically shared HLA alleles making it possible to successfully predict the HLA restriction, and a large collection of Mtb T cell epitopes enabled us to identify peptide-MHC ligands for all five of the groups tested. Mutagenesis and de novo TCR design confirmed the GLIPH identified motifs were critical and sufficient for shared antigen recognition. Thus the GLIPH algorithm can analyze large numbers of TCR sequences and define TCR specificity groups shared by TCRs and individuals, which by itself could greatly accelerate our ability to analyze T cell responses and expedite the identification of specific peptide-MHC ligands.

2.2.1 Introduction

The adaptive immune system uses a highly diverse population of T lymphocytes to selectively recognize and respond to antigenic proteins. Individual T lymphocytes

generate and express single recombinants from a massive repertoire of combinatorially generated T-cell receptors (TCRs) with diversification mechanisms that could theoretically produce in excess of (˜10ˆ16) unique heterodimers(4). Because of this extreme diversity, individuals typically share only ˜1% of their TCR seqences, and even monozygotic twins-with identical HLA haplotypes, only share ˜2%(1,3,24,25). The major αβ form of the TCR heterodimer typically recognizes peptide fragments of larger proteins, bound to major histocompatibility complex (MHC) or related molecules and displayed on the cell surface(26). The unique TCR recombinantly produced by each T lymphocyte imbues it with a unique TCR specificity – the ability to selectively recognize and respond to a given peptide-MHC (pMHC) complex or set of complexes. While αβ TCRs are highly specific for a given peptide-MHC ligand, such that a small change in an amino acid side chain of the peptide or TCR is sufficient to eliminate binding, it is also very common for there to be cross-reactivity to similar and even non-homologous peptides bound to the same or a different MHC. This appears to be due to a flexibility in the TCR binding site, allowing multiple stable conformations27. Although advances in high-throughput sequencing technologies now enable the routine analysis of millions of T-cell receptors in a single experiment, there has been no systematic way to organize groups of TCR sequences according to their likely antigen specificities.

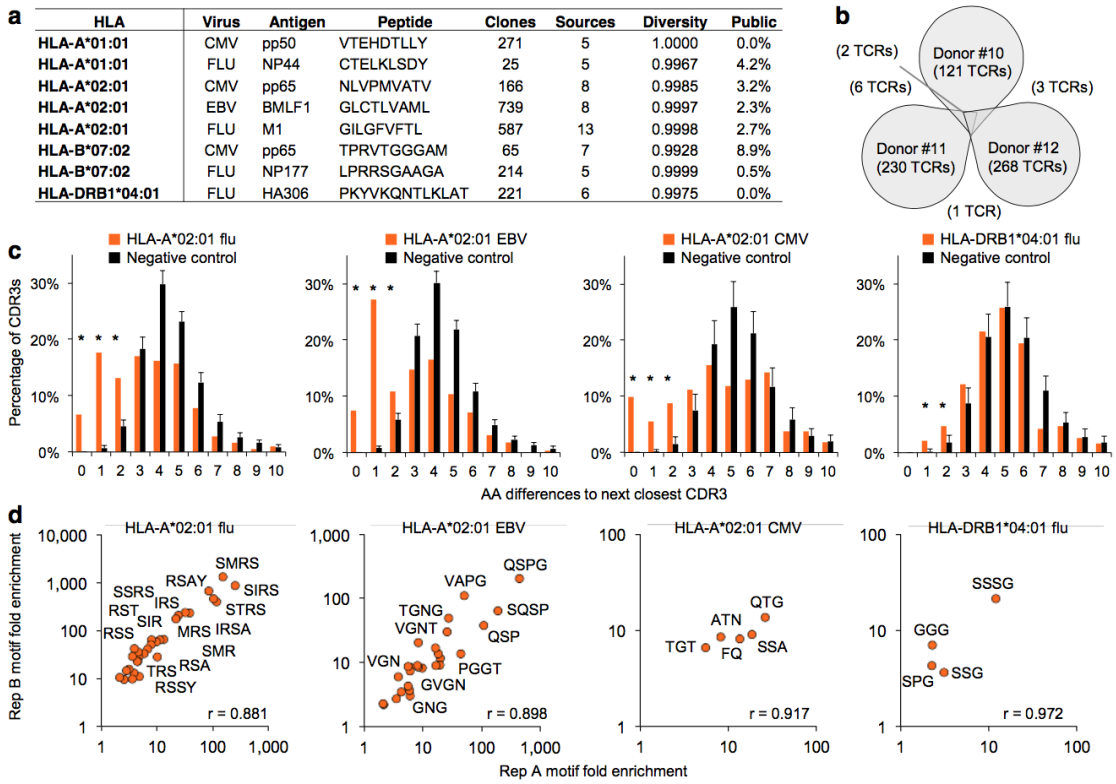
2.2.2 Results

To address this problem, we performed a comprehensive analysis of all published structural data of TCR-pMHC complexes. We aligned the TCR amino acid sequences from all 52 of the structures of TCR-pMHC complexes that have been reported, and for each alignment position calculated the proportion of all complexes where the residue at that position was within 5 Angstroms from the peptide antigen (Extended Data Fig. 1, Supplementary Table 2). This provided an a-priori probability of contact for every position in a TCR. The results showed that the majority of antigen peptide contacts were in TCR CDR3s, and only short, typically linear stretches of amino

acids in particular regions of CDR3s make contact with antigen(IMGT positions 107-116), while the majority of CDR positions did not contribute to ligand contact and in particular the stem positions of CDR3 (IMGT positions 104, 105, 106, 117, and 118) are never within 5 Angstroms of antigen(9). We also note that whereas there is always at least one CDR3β contact, there are multiple cases where no CDR3α contact is made, suggesting that the former is required, although typically both are involved (Extended Data Fig. 1). Collectively, the result suggested that sequence analysis focused entirely on the high probability contact sites in CDR3 may provide a means of clustering TCRs by shared specificity.

To evaluate whether specificity was principly mediated by these limited contact sites, we assembled a panel of eight peptide-MHC tetramers, and used them to isolate specific T cells from 4-13 blood bank donors for each HLA specificity, plus one tonsil sample for the class II specificity. These were immunodominant peptides from Epstein-Barr virus (EBV), Cytomegalovirus (CMV), and influenza in the context of HLA backgrounds HLA-A*01:01, HLA-A*02:01, HLA-B*07:02 or the class II molecule HLA-DRB1*04:01 (Fig. 1a). Antigen-specific T cells were isolated using peptide-MHC-tetramers, and characterized using either single cell TCR αβ sequencing or bulk TCR β sequencing. In addition, 229 published TCR sequences of known specificity were obtained from the literature and from crystal structures in the Protein Data Bank(10). In total, the training set consisted of 2068 unique TCRs of known specificity (Supplementary Table 1). While most specificities were recognized by hundreds of unique TCR sequences in each subject, a few were more monomorphic. For all specificities, few if any TCRs were found shared across subjects, with the majority being unique to an individual (Fig. 1b).

The specificity of some TCRs could be predicted by global similarity (Fig. 1c). Searching just within CDR3, the CDR3 sequences selected against a single specificity would often differ by only one amino acid, while this was not observed in set of unselected TCRs. It was also noted that antigen-specific pools of TCRs were enriched for more similar CDR3s on average (differing by 2-4 amino acids), although those could not individually be asserted to be antigen specific as naïve TCR populations also occasionally produced TCRs with that degree of similarity.



HLA	Virus	Antigen	Peptide	Clones	Sources	Diversity	Public
HLA-A*01:01	CMV	pp50	VTEHDTLLY	271	5	1.0000	0.0%
HLA-A*01:01	FLU	NP44	CTELKLSDY	25	5	0.9967	4.2%
HLA-A*02:01	CMV	pp65	NLVPMVATV	166	8	0.9985	3.2%
HLA-A*02:01	EBV	BMLF1	GLCTLVAML	739	8	0.9997	2.3%
HLA-A*02:01	FLU	M1	GILGFVFTL	587	13	0.9998	2.7%
HLA-B*07:02	CMV	pp65	TPRVTGGGAM	65	7	0.9928	8.9%
HLA-B*07:02	FLU	NP177	LPRRSGAAGA	214	5	0.9999	0.5%
HLA-DRB1*04:01	FLU	HA306	PKYVKQNTLKLAT	221	6	0.9975	0.0%

Figure 2.1: Characteristics of TCRs reactive to common antigens across individuals. a, MHC-tetramer sorted antigen specific TCR repertoires of common pathogen epitopes as well as public sources. Diversity is calculated as the Shannon entropy of observed clones, where clone counts are the number of individuals expressing each clone. Percentage of all clones that were found in more than one individual reported as Public. b, Representative Venn Diagram of GLC-specific clonal overlap in three HLA-A*02:01+/EBV+ donors. c, Minimum Hamming distance of CDR3βs in MHC-tetramer sorted antigen-specific pools, rendered non-redundant within each subject, compared with equal sized randomly sampled naïve control pools. s.d. of 100 repeat random samples of control TCRs reported on bars (*P<0.01 Chi-square test). d, CDR3s in MHC-tetramer sorted antigen-specific pools are enriched for a subset of motifs. Replicates A and B consisting of TCRs from different sets of donors (Supplementary Table 7), reproduce the same motifs with correlated enrichment assessed by Pearson correlation coefficient.

To further distinguish TCRs recognizing the same antigen from unrelated TCRs, we searched for enrichment of amino acid motifs of length 2, 3, and 4 in the high contact probability region of CDR3 spanning IMGT positions 107-116. As the repertoire is created through a complex V(D)J recombination process that is not known to model neatly to simple functions, we instead developed a non-parametric resampling method for detecting significant enrichment of local motifs in antigen-specific CDR3s. In this method, the similarity of receptors amongst antigen-specific repertoires were compared to repeat random subsampling from CDR3-length distribution matched unselected repertoires of 266,346 unique naïve unselected CD4 and CD8 TCRs from thirteen healthy individuals to establish a "fold enrichment" and "probability of enrichment" of any motif above its expected frequency in the naive repertoire(11,12). To avoid false positive conclusions drawn from PCR or read error, we separated biological replicates containing TCRs against the same specificity from different individuals and searched for enrichment of common motifs. Using this method, we reproducibly detected enriched motifs in CDR3s that were found within TCRs specific to a given pMHC from multiple individuals, but not in TCRs recognizing unrelated antigens (Fig. 1d).

When clustering TCRs by both global similarity (CDR3 differing by up to one amino acid) or local similarity (shared enriched CDR3 amino acid motifs: $>$10x fold-enrichment, Prob$<$0.001), we found that most of the TCR sequences selected by a particular peptide-MHC tetramer typically fell into one or a few related TCR groups (Fig. 2a). Furthermore, in four cases where there was a high resolution crystal structure involving one of these dominant homology group TCRs complexed to its peptide-MHC ligand, the significantly enriched CDR3 motifs we detected in the TCR sequences corresponded to the contact residues with the antigenic peptide (Fig. 2a, b, Extended Data Fig. 2a). These were typically three to four amino acids in length and usually contiguous. Positions outside of this central contact motif tended to tolerate more amino acid diversity. A positional weight matrix of amino acid diversity in the sequence group gives a high score to current group members and could be used to selectively recognize new group members (Extended Data Fig. 2b). In the case of the flu GIL antigen bound to HLA-A*02:01, where we had $\alpha\beta$ pairing from single cell TCR

sequencing, we would observe convergence of both α and β sequences (Fig. 2b). In other cases, including the DRB1*04:01 restricted flu peptide specificity, the conserved sequence features in the TCR alpha CDR3 were more discontinuous and could only be easily detected in the context of a solved structure, where the coordination of negative charges on the TCR with positive charges on the peptide appears to be accomplished in multiple ways by different TCRs (Extended Data Fig. 2a). Thus we have largely relied on TCR β sequences for our analysis. This also has the serendipitous effect of enabling analysis of the growing databases of those sequences in different studies, the majority of which use technologies that only sequence TCR β. An analysis of positional motif enrichment and positional amino acid enrichment relative to the unselected repertoire highlighted the specific residues and their motif relationships that the structures indicate contribute the primary contacts of antigen recognition (Extended Data Fig. 2b). The results suggest that varying degrees of sequence convergence in each chain of the TCR heterodimer may provide some information regarding the relative importance of each chain for specificity.

Collectively, these results on our tetramer-sorted TCR "training" dataset formulated the parameters of an algorithm—GLIPH (Grouping Lymphocyte Interactions by Paratope Hotspots) to search for and automatically cluster TCR sequences into distinct groups according to their likely specificity (Extended Data Fig. 3, Supplementary Discussion). We then subjected GLIPH to three major validation tests: First, we ran multiple benchmarks on our training set of the aggregate sequence data from the eight tetramers and 2068 unique sequences. We individually unit tested the global CDR3 and local motif components of GLIPH separately and in combination (Extended Data Fig. 4a). We found that local motifs clustered about 10% of the TCR database into clusters of at least 3 members, global CDR3 similarity clustered about 12% of the TCRs into clusters, and the complete GLIPH combination placed 14% of TCRs in clusters, with a trend that more sequences would result in more clusters recognized and a higher percentage clustered. In comparison, GLIPH did not generate any appreciable clusters when run on naïve TCRs at a range of depths, providing confidence in a low false positive rate. Although GLIPH clustering was

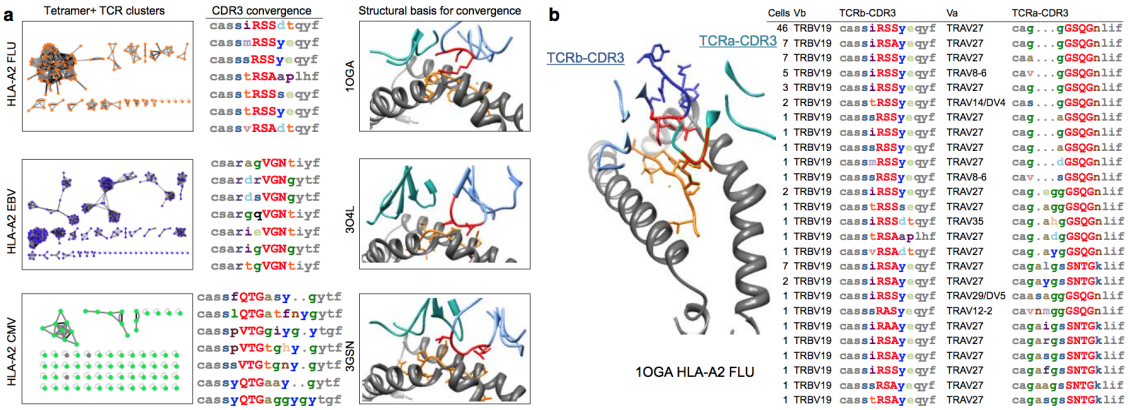


Figure 2.2: Crystal structure representatives of TCR specificity groups reveal the structural basis for antigen-specific paratope convergence. a, Network analysis of tetramer+ CDR3 clusters indicates relationships between TCRs (nodes) sharing global CDR3 similarity (black edges) or local CDR3 motifs (grey edges: motifs >10 fold enriched, 0.001>probability of enrichment by chance). Grey arrows indicate representative specificity group, accompanied with representative CDR3 alignment and crystal structure. Significant motif residues are highlighted in red in both CDR3 alignments and structure. In alignments: low contact probability: grey. In structures MHC: grey; peptide: orange; TCRβ: light blue; TCRα: cyan. b, single cell paired α/β sequencing with crystal structure representative reveals coordinated motifs in both TCRβ and TCRα CDR3 that define paratope specificity.

run on the aggregate TCR database, when evaluating the composition of the clusters, we found almost all TCRs in an individual cluster to be of a common specificity (Extended Data Fig. 4b). We further benchmarked GLIPH against a global CDR3 only approach, a local motif only approach, a local motif only approach that did not include our structural information, and an independent clustering algorithm CD-HIT (Extended Data Fig. 4c)(17). For global similarity, we found that while a clustering by CDR3 distance 1 resulted in effective grouping of TCRs of similar specificity, a distance of 2 resulted in predominantly mixed clusters. For local motif clustering, we found that good clustering of TCRs of common specificity could only be obtained when our structural masks were applied to the data. Similarily, although CD-HIT was not effective at clustering TCRs by common specificity when provided the entire TCR sequences, when offered only the high contact probability CDR3s, it was able to perform effective clusters provided an appropriate clustering threshold. GLIPH produced more clusters with higher accuracy than the other methods. Finally, as a test of whether GLIPH could recognize new members of existing specificity groups, we ran GLIPH on replicates containing only half of the subjects (Supplementary Table 7), and then used those specificity groups to score TCRs in the other half of the subjects. GLIPH was able to successfully recognize new TCRs of known specificity groups in the circulating T cells of new donors – providing a basis for reading the TCR repertoire (Extended Data Fig. 4d). The excess of specificity groups over the number of peptide-MHCs shows that there are multiple distinguishable TCR sequence solutions to a given ligand, with multiple unique clusters observed for all pMHC antigens evaluated (Fig. 2a, Extended Data Fig. 4b).

Our second validation test was to evaluate the performance of GLIPH in an independent test set: TCR sequences from Mycobacterium Tuberculosis (Mtb) specific CD4+ T cells from 22 subjects who were QuantiFERON positive, which indicates a latent infection (Supplementary Table 3). Briefly, PBMCs were stimulated with a large collection of Mtb peptides (300) or an Mtb lystate for 4 and 12 hours respectively, and Mtb-specific CD4+ T cells were selected based on the upregulation of the CD154 activation marker or cytokin secretion (Extended Data Fig. 5b, c)(18-20). Single cells were sorted into 96-well plates and amplified and sequenced for TCR $\alpha\beta$ sequences, as

	TCR ID	Donor ID	CDR3b	Freq	CDR3a	Freq	DQA1		DQB1		DRB1		DRB3/4/5	
Group I	TCR001	01/0873	CASSFEETQYF	2/168	CIVKTNSGGSNYKLTF	2/158	*05:02	*01:02	*03:19	*06:02	*11:01	*15:03	DRB3*02:02	DRB3*02:02
	TCR008	09/0018	CASSLEETQYF	2/400			*05:01	*01:02	*02:01	*06:03	*03:01	*15:03	DRB3*02:02	DRB5*01:01
	TCR010	03/0492	CASSPEETQYF	1/112			*01:02	*01:02	*06:09	*06:02	*13:02	*15:03	DRB5*01:01	DRB3*03:01
	TCR012	09/0217	CASSPEETQYF	49/166	CIVHTNSGGSNYKLTF	47/135	*01:03	*01:02	*06:04	*06:02	*13:01	*13:02	DRB3*03:01	DRB3*02:02
	TCR003	01/0430	CASSLEETQYF	1/82	CGMSGNTGKLIF	1/70	*03:03	*01:05	*02:02	*05:01	*10:01	*09:01	DRB4*01:01	DRB4*01:01
	TCR004	01/0873	CASSLEETQYF	21/168	CIEHTNSGGSNYKLTF	21/158	*05:02	*01:02	*03:19	*06:02	*11:01	*15:03	DRB3*02:02	DRB3*02:02
	TCR009	01/0873	CASSPEETQYF	2/304			*05:02	*01:02	*03:19	*06:02	*11:01	*15:03	DRB3*02:02	DRB3*02:02
	TCR011	09/0018	CASSPEETQYF	31/400	CAVPSGGANSKLTF	1/267	*05:01	*01:02	*02:01	*06:03	*03:01	*15:03	DRB3*02:02	DRB5*01:01
Group II	TCR022	01/0873	CASSVALAGAEYF	1/69	CAVGGLSGANSKLTF	1/67	*05:02	*01:02	*03:19	*06:02	*11:01	*15:03	DRB3*02:02	DRB3*02:02
	TCR023	02/0152	CASSVALASGANVLTF	2/41	CAGAGGGGFKTIF	2/28	*05:01	*01:02	*02:01	*06:01	*03:01	*15:01	DRB5*01:01	DRB3*01:01
	TCR024	03/0492	CASSVALQGVHTQYF	2/112	CAGTNTGNQFYF	2/90	*01:02	*01:02	*06:09	*06:02	*13:02	*15:03	DRB5*01:01	DRB3*03:01
	TCR026	09/0018	CASSVALYANEQFF	1/151	CAGPTTGYALNF	1/125	*05:01	*01:02	*02:01	*06:03	*03:01	*15:03	DRB3*02:02	DRB5*01:01
	TCR036	09/0772	CASSVALLGETQYF	1/107	CAGAPTGNQFYF	1/98	*05:05	*01:02	*03:01	*06:02	*03:01	*15:03	DRB3*02:02	DRB5*01:01
	TCR029	09/0328	CASSVALLGGEQYF	1/107	CAGLVGTSYGKLTF	1/73	*06:01	*04:01	*03:01	*04:02	*12:02	*03:02	DRB3*03:01	DRB3*01:01
	TCR025	03/0492	CASSVALATGEQYF	1/112	CAGPTGGSYIPTF	1/90	*01:02	*01:02	*06:09	*06:02	*13:02	*15:03	DRB5*01:01	DRB3*03:01
Group III	TCR051	02/0152	CASSLIEGGTEAFF	1/41	CVVSAITNDYKLSF	1/28	*05:01	*01:02	*02:01	*06:01	*03:01	*15:01	DRB5*01:01	DRB3*01:01
	TCR052	09/0772	CASSLIEGLEQYF	1/107	CAVQPGAGGFKTIF	1/98	*05:05	*01:02	*03:01	*06:02	*03:01	*15:03	DRB5*01:01	DRB3*02:02
	TCR053	09/0018	CASSLIENTEAFF	1/151	CAVTIGATQGGSEKLVF	1/125	*05:01	*01:02	*02:01	*06:03	*03:01	*15:03	DRB5*01:01	DRB3*02:02
	TCR054	02/0152	CASSLIEQQPQHF	1/41	CASQSNTGNQFYF	1/28	*05:01	*01:02	*02:01	*06:01	*03:01	*15:01	DRB5*01:01	DRB3*01:01
Group IV		03/0492	CASSSGQGHYNEQFF	1/162			*01:02	*01:02	*06:02	*06:09	*15:03	*13:02	DRB3*03:01	DRB5*01:01
		09/0328	CASSVGQGHYNEQFF	1/107	CAVISGGSNYKLTF	1/73	*06:01	*04:01	*03:01	*04:02	*12:02	*03:02	DRB3*03:01	DRB3*01:01
	TCR098	03/0492	CASSLGQGHYNEQFF	3/162	CAVNGGGSNYKLTF	3/134	*01:02	*01:02	*06:02	*06:09	*15:03	*13:02	DRB3*03:01	DRB5*01:01
	TCR099	09/0125	CASSPGQGHYNEQFF	4/56	CAVNSGGSNYKLTF	4/39	*06:01	*01:02	*03:01	*05:02	*12:02	*16:02	DRB3*03:01	DRB5*01:01
Group V		01/0906	CSARSSGGEAKNIQYF	2/118			*02:01	*01:02	*02:02	*06:02	*07:01	*15:01	DRB4*01:03	DRB5*01:01
		09/0018	CSARKGGGEAKNIQYF	1/182			*05:01	*01:02	*02:01	*06:03	*03:01	*15:03	DRB3*02:02	DRB5*01:01
	TCR087	03/0492	CSARAGGGEAKNIQYF	3/112	CAVSRAGAGSYQLTF	3/90	*01:02	*01:02	*06:09	*06:02	*13:02	*15:03	DRB3*03:01	DRB5*01:01
	TCR088	01/0906	CSARASGGEAKNIQYF	1/106	CAVRDPGNTDKLIF	1/72	*02:01	*01:02	*02:02	*06:02	*07:01	*15:01	DRB4*01:03	DRB5*01:01

Figure 2.3: TCR specificity groups and predicted HLA-restriction among Mtb-infected subjects. CDR3 α/β amino acid sequences from five GLIPH TCR specificity groups. Yellow colored boxes highlight the predicted common HLA class II alleles for each specificity group (combinatorial sampling probability Prob$<$0.013 DRB1*15 for group II, Prob$<$0.007 DRB1*03 for group III, Prob$<$0.03 DRB3*03 for group IV, Prob$<$0.02 DRB1*15/DRB5*01 for group V). Green colored boxes highlight the TCRs that have been validated in vitro. Red outlines indicate actual HLA as determined by reporter assay.

well as scored for a panel of 18 cytokine genes using multiplex primers as previously described (Extended Data Fig. 6)(21). The majority of cells showed a Th1*-like phenotype including IFNγ and IL-2 production, no IL-17 production, and T-bet and RORC expression consistent with previous reports(22,23). The TCRs from the samples were enriched for clonally expanded sequences compared with PBMC controls (Extended Data Fig. 7). We obtained 4464 independent TCR α and 5711 β sequences from 22 individuals and analyzed them with the GLIPH algorithm. GLIPH clustered 14% of all TCRs into 141 clusters of which 43 contained at least three unique TCRs. We focused on clusters that contained TCRs from at least 3 individuals: 16 distinct TCR βspecificity groups that were shared between three or more individuals and contained at least four uniquely derived TCR β clones. Among that set, there were 6 specificity groups that exhibited significant V-gene bias ($p<0.05$), exhibited significant CDR3 length bias ($p<0.05$), and were overrepresentated in clonally expanded T cells (Fig. 3, Supplementary Table 5).

As an initial validation of the GLIPH-predicted specificity groups in the test set, these 22 individuals were comprehensively HLA typed by sequencing in order to determine whether the GLIPH-derived TCR clusters also correlated with shared HLA alleles (Supplementary Table 6). This is expected if these TCR specificity groups are discrete specificities, as αβ T cells recognize both peptides and the HLA molecules they are bound to. We observed 69 unique HLA class II alleles across the genotypes of our 22 subjects, but observed only one or two enriched candidate HLAs within the contributors to each GLIPH sequence group. To determine whether this predicted HLA restriction was correct, we chose three or four TCR heterodimers from different individuals from five different representative TCR specificity groups (I, II, III, IV, V) that all scored well for the GLIPH parameters (Fig. 3). Using a reporter assay(24), we found that, as predicted, Group I responded to the class II allele DQA1*01:02/DQB1*06:02, Group II responded to DRB1*15:03, Group III responded to DRB1*03:01, Group IV responded to DRB3*03:01, and Group V responded to DRB5*01:01 (Fig. 4a-c, Extended Data Fig. 8). This provided initial validation that GLIPH was successfully grouping TCRs of common specificity, and

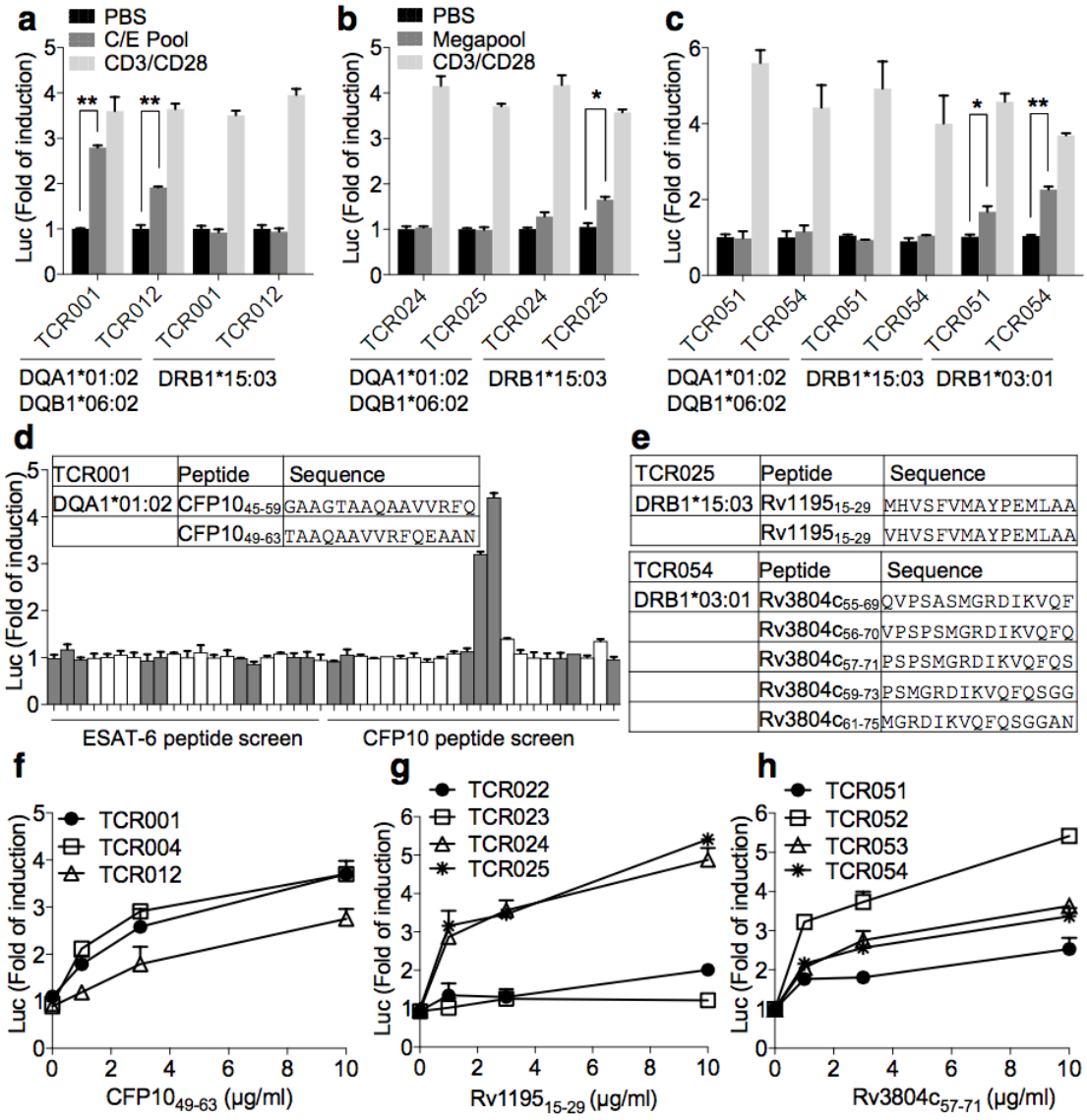


Figure 2.4: Identification of common antigen recognition by TCR specificity groups. a, Group I TCRs were tested against candidate HLA alleles using CFP10/ESAT-6 pool (C/E Pool). Group II (b) and Group III (c) TCRs were tested using Megapool. Negative control: PBS, positive control: CD3/CD28 stimulation. Mean ± s.d. (n=3) shown. *$P < 0.05$ and **$P < 0.005$ two-tailed Student's t-tests. d, Individual peptides from C/E Pool tested against TCR001. Top 15th percentile of NetMHC-predicted DQA1*01:02 binding indicated by grey bars. Insert table shows identified peptide antigen. e, Restricted HLA types and responding peptides for Group II and III TCRs. f-h, Dose-dependent response of Group I, II and III TCRs to their corresponding epitopes. Mean ± s.d. (n=3) shown.

also enabled rapid identification of an appropriate HLA-matched APC for antigen discovery, without having to test for all 69 HLAs.

To determine the Mtb peptide specificity, we used the IEDB HLA-II binding prediction algorithms to rank likely antigen candidates known to be in the Megapool(25), and then performed individual peptide stimulation assays on the HLA-matched APC cell line to quickly find the target peptide for all TCR specificity groups (I, II, III, IV, V) (Fig. 4d, e, Extended Data Fig. 8). In each case, all or most TCRs in a given group recognized the same Mtb peptide (Fig. 4f-h). We performed a sensitive Glycine mutagenesis scan of TCR025, which confirmed that GLIPH's predicted contact motif was at the center of specificity determination, with even conservative single amino acid changes (A->G and L->G) in the motif being sufficient to abolish specificity (Fig. 5a). This provided a definitive validation of GLIPH's ability to group TCRs of common specificity in a CD4 setting, while also providing a reproducible method for rapid TCR specificity analysis and antigen discovery in T cell responses generally.

As our final validation test of GLIPH's ability to identify specificity regions of a TCR, and predict the specificity of new TCRs using this information, we chose to generate de novo TCRs against Mtb's DRB1*15:03 restricted Rv119515-29. From subject-derived TCR CDR3 sequences identified by GLIPH as being convergent against this antigen (Fig. 5b), we calculated a CDR3 PWM matrix (Fig. 5c), in order to de novo design TCR β sequences (paired with the TCR α from known binder TCR025) as having the same specificity. From the GLIPH TCR PWM, we emitted the top 1000 predicted CDR3b TCRs specific to Mtb's DRB1*15:03 restricted Rv119515-29. Some of the emitted CDR3s were identical to those of observed binders in the CDR3, although when in the context of TCR025 Vβ, Jβ, Vα, Jα and CDR3α, differed by at least 45 amino acids in the total TCR (Extended Data Fig. 9). We found that many predicted binders, none of which was found in our study, had better scores than the naturally derived TCRs obtained from subjects (Fig. 5d). From the GLIPH prediction set, selected 10 of the best scoring predicted TCRs that were at least two amino acids different from TCR025, and. 8 out of 10 TCRs demonstrated antigen-specific activation to Rv119515-29, with two such TCRs being significantly

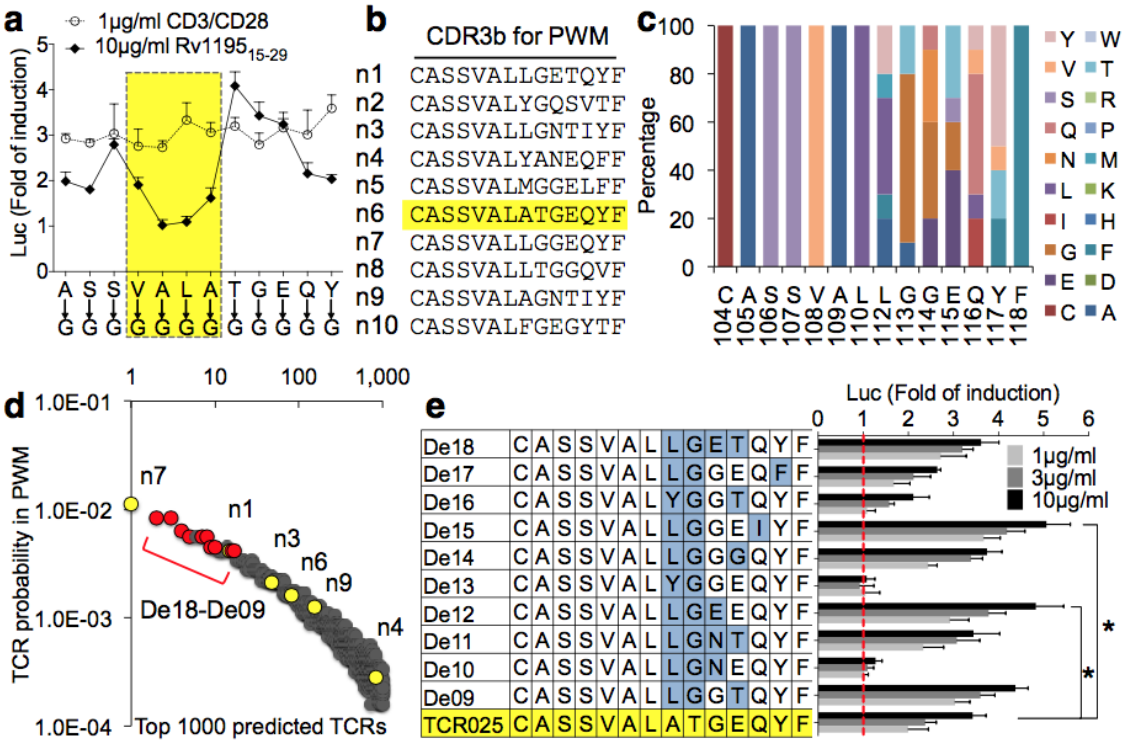


Figure 2.5: Mutagenesis validation and de novo TCR design. a, Glycine scan of CDR3β of TCR025 (Group II). Each mutant was stimulated by DRB1*15:03 restricted Rv119515-29, as well as a CD3/CD28 positive control. Mean ± s.d. (n=3) shown. b, Group II CDR3β sequences with common CDR3 length. c, Positional Weight Matrix (PWM) reports observed CDR3β positional amino acid frequencies from (b). d, Top 1000 theoretical TCRs and scores from PWM (Formula 5). Top 10 predicted TCRs (De18-De09) shown in red. Natural TCRs obtained from donors shown in yellow. e, De18-De09 were stimulated by DRB1*15:03 restricted Rv119515-29. Blue indicates modified amino acids and red dash line indicates the basal activity. Mean ± s.d. (n=3) shown. Activity compared to TCR025, *P $<$ 0.01 two-tailed Student's t-tests.

more active than TCR025 (Fig. 5e). This shows that GLIPH is able to predict new members of a specific group and even to improve sensitivity.

2.2.3 Discussion

Relying on tetramer-sorted TCR "training" dataset, we successfully developed a novel algorithm to search for and automatically cluster TCR sequences into distinct groups according to their likely specificity. This algorithm, "Grouping Lymphocyte Interactions by Paratope Hotspots" or GLIPH, combines global TCR sequence similarity, local CDR sequence similarity (motifs), spatial peptide-antigen contact propensity, V-segment bias, CDR3 length bias, shared HLA alleles among TCR contributors, and clonal expansion bias, together with other observations concerning TCR specificity from the literature, to identify and cluster TCR sequences in specificity groups – sets of TCRs that are likely to be recognizing the same or very similar peptide-MHC ligands (Extended Data Fig. 3, https://github.com/immunoengineer/gliph). The algorithm first searches in parallel for global and local similarity signatures. For local similarity signatures, it tests for enrichment of any 2, 3, and 4 amino acid continuous motifs within CDR3 β sequences while excluding conserved N-terminal V region or C-terminal J region CDR3 amino acid positions (IMGT 104, 105, 106, 117, and 118) that have never been observed to bind the antigenic peptide in crystal structures (Extended Data Fig. 1). Optionally it can also search for discontinuous amino acid motifs in CDR3 that allow any amino acid at a position, with specific 2, 3, or 4 amino acids around that position. A motif is considered enriched if it is elevated at least 10-fold over expected frequency in the naïve TCR reference pool, with a probability$<$0.001 being at that level of enrichment by chance. In parallel, it also identifies global similarity sequences with CDR3 β sequences that are identical or only differ by one amino acid and are of the same length. Next, it clusters these TCRs with identified similarities, gathering together variants of similar amino acid motifs or global CDR3 similarity into a single sequence group in the process. Finally, GLIPH confirms the significance of the formed clusters through evaluation of enrichment of independent features, including V-gene usage, CDR3 length distribution, clonal expansion

bias and donor HLA usage in each sequence group. This three-step process of motif nucleation creates a powerful statistical framework for identifying TCRs that recognize the same antigen. Optionally, GLIPH can restrict cluster members to only those TCRs that share the same V-gene. When HLA genotypes for the donors are available, GLIPH also provides a predicted HLA-restriction for each TCR sequence group generated. This HLA allele prediction can then in turn be used to score databases of candidate peptides, ranking them according to their odds of being the target antigen. While previous studies have noted that certain TCR positions are more likely to be in contact with antigen, or that distributions of short TCR CDR3β motifs systematically alter after immunization5,28, no other TCR repertoire analysis tools that we are aware of are able to formalize these observations into a unified framework for automated TCR sequence group identification, with statistical validation and HLA prediction(29-31).

GLIPH can cluster TCR sequences based on their likely antigen specificity, predict their likely HLA restriction and prioritize specific peptide antigen candidates. Our data also indicates that linear, usually continuous motifs of 3-4 amino acids in CDR3β and CDR3α, are often highly conserved in TCRs that recognize the same peptide-MHC ligand, and that in known structures, these are the residues responsible for contacting the antigenic peptide. This builds and expands on previous work by Chain and colleagues indicating that 3 amino acid motifs in the TCR CDR3 regions are often correlated with antigen specific responses28. While GLIPH requires the HLA genotypes of the subjects in order to make an HLA prediction, in our Mtb dataset of 22 individuals there were 69 unique class II HLAs, and in each evaluated sequence group we were able to predict the presenting HLA. Similarly while GLIPH cannot predict a specific target peptide a-priori, knowing the HLA and a candidate population of peptides, it is possible to prioritize the peptides by predicted binding to the candidate HLA, greatly reducing the combinations of APCs and peptides to be tested. Where no pool of candidate antigens is available, the HLA-prediction facilitates the use of the pMHC yeast display libraries developed by Garcia and colleagues(27).

An important question is - how much of the TCR repertoire will GLIPH capture? Here the limited data we have at this time is that it will be substantial, in that of the

>2000 TCR sequences that we analyzed in Fig 3A, up to 14% could be gathered into high-confidence discrete specificity groups with GLIPH, with rarefaction suggesting that more groups could be identified if more data was available. Similarly (14%; 796/5711) of the TCRs fit into specificity groups in the Mtb cohort. A partial solution to this problem will likely come from larger sample sizes, although in very diverse populations, such as the South African cohort analyzed here, there will HLA alleles that will be very rare and those only in very large population studies will there be enough matches that could generate shared TCR specificities. Nonetheless, building large TCR data bases around specific diseases, or TCR-omes, could be very valuable in assessing the responses of individuals or cohorts, even if only a fraction of the TCR sequences can be clustered using GLIPH.

In summary, we find that the GLIPH algorithm provides a way to organize TCR sequences into distinct groups of shared specificity either within an individual or most importantly, across a population of individuals. Secondly, it facilitates T cell antigen discovery as shown by our analysis of Mtb specific T cells. Third, the specificity groups GLIPH identifies allow one to read the T-cell receptor repertoire from primary sequence data, even if the peptide-MHC ligand is not known. In fact a fourth, and perhaps the most important use of GLIPH is to analyze αβ T cell responses independently of knowing the epitope specificity and MHC restriction of the set of TCRs. The number and size of such clusters provides information as to the complexity of an immune response, or the presence of an important shared immune response across individuals. Thus in analyzing the T cell response to a vaccine or infection in a given cohort, quantifying how many distinct specificity groups are active in each individual, and whether that correlates with a good or bad response, would be valuable information.

2.2.4 Methods

Antigen-specific T cells Peripheral blood mononuclear cells (PBMCs) were obtained from 28 healthy blood donations of known HLA type at HLA-A, HLA-B, and HLA-DR

$$(1)\ CI(99.9\%\ os) = \overline{y} \pm t_{0.0005}(\frac{s}{\sqrt{n}})$$

$$(2)\ s = \sqrt{\frac{\sum(y_1 - \overline{y})^2}{n-1}}$$

$$(3)\ D = \frac{\sum(n(n-1))}{N(N-1)}$$

$$(4)\ P(X = C) = \frac{\prod_i^N P_i(X=C)}{\prod_i^N P_i(X=C) + \prod_i^N P_i(X!=C)}$$

$$(5)\ s = \prod_{i=1}^{10} Pr(a_i|PWM)$$

$$(6)\ Pr_s = \frac{\sum_{n=1}^{N}(S_n|S_n \geq (s-v))}{N}$$

Figure 2.6: Mathematical methods defined in the publication for estimating selection effects and expected clone frequencies from PWMs. (1) Approximation for accelerating observed vs expected confidence interval estimates for motif frequencies (2) SD estimate at sampling depth (3) Simpson's score for feature enrichment in specificity groups (4) probability conflation score for specificity groups (5) theoretical probability of any sequence given a positional weight matrix PWM (6) Normalization of theoretical probability scores to score distribution of unselected sequences against positional weight matrix PWM.

loci and known infection status for EBV and CMV from the Stanford Blood Center. Cells were stained with fluorophore-conjugated peptide-MHC (pMHC) tetramers of MHC HLA-A*01:01, HLA-A*02:01, HLA-B*07:02, or HLA-DRB1*04:01 backgrounds. The tetramers were engineered to display peptides of either EBV, CMV, or flu through photo-exchange of a surrogate peptide in the presence of a molar excess of a replacement peptide. For EBV HLA-A2, a commercial dextramer was also used. Cells were sorted by fluorescent activated cell sorting (FACS) to collect either as single-cells or bulk populations of antigen-specific cells in RT reverse-transcriptase one-step reaction solution (Extended Data Fig. 5aSI Figure 4A). During single-cell sorting, index sorting was applied to collect activity on a number of additional markers including CD45RA and CD62L. Sorting for tetramer specific cells was conducted on a DB Aria II (BD Biosciences). For single-cell sorting Mtb-specific CD4+ T cells using activation marker CD154, PBMCs were thawed in complete RPMI 1640 medium at 2 × 106 cells/ml and recovered 12 hours before stimulation. PBMCs were stimulated with Mtb lysate (10 µg/ml) for 12 hours in the presence of 1 µg/ml purified anti-CD49d antibody and anti-CD154-PE. After stimulation, cells were harvested and stained with surface markers for sorting. For single-cell sorting cytokine-secreting cells, PBMCs were stimulated with either CFP10/ESAT-6 peptide pool or Megapool (2 µg/ml for each peptide) for 4 hours in the presence of 1 µg/ml purified anti-CD49d antibody. Cells were harvested and stained using the IL-2 or IFNγ Secretion Assay Kit (Miltenyi Biotec). Sorting was conducted on a BD FACSJazz cell sorter (BD Biosciences).

Mtb-infected study participants 22 Adolescent participants, aged 12 to 18 years, were randomly selected from a previous cohort study, which enrolled in the town of Worcester, approximately 100 km from Cape Town, South Africa, between 2005 and 200726. This study was approved by the Faculty of Health Sciences Human Research Ethics Committee of the University of Cape Town and Human Research Protection Program (HRPP) at Stanford University. Written informed consent was obtained from the parents of adolescents and assent was obtained from adolescents. Venous blood was collected for PBMC isolation, QuantiFERON® TB Gold In-tube (Qiagen) (QFT) and a tuberculin skin test (TST) was administered. All samples used in this

study were from asymptomatic QFT-positive adolescents. PBMCs were obtained by density gradient centrifugation using Ficoll and cryopreserved using freezing medium containing 90% fetal bovine serum and 10% DMSO. The 22 participants were HLA typed at Sirona Genomics (now Immucor inc.), under supervision of Dr. Michael Mindrinos.

Cell lines and reagents The Jurkat 76 T-cell line, deficient for both TCR-α and TCR-β chains, was kindly provided by Dr. Shao-An Xue (Department of Immunology, University College London). NFAT reporter stable cell line (J76-NFATRE-luc) was constructed using lentiviral transfer of pNL[NlucP/NFAT-RE/Hygro] (Promega) into Jurkat 76 cell. K562 cell line was obtained from the ATCC and cultured under standard conditions. Artificial antigen presenting cells were constructed using lentiviral transfer of different HLA alleles (gBlock ordered from IDT) into K562 cells. Anti-CD4-APC, anti-CD69-APC/Cy7, and anti-TCR α/β-FITC abs were purchased from BioLegend. Anti-CD3-PB, purified anti-CD49d and anti-CD154-PE abs were purchased from BD Biosciences.

Antigens Mtb CFP10/ESAT-6 peptide pool: 22 peptides spanning the length of the CFP10 molecule and 21 peptides spanning the length of the ESAT-6 molecule were purchased from PEPscreen (Sigma). Each peptide was 15 amino acids long and overlapped its adjacent peptide by 11 residues. Peptide was dissolved in DMSO at 100 µg/ml and then mixed together to make CFP10/ESAT-6 peptide pool. Megapool peptides, containing 300 epitopes from 90 Mtb proteins were kindly provided by Dr. Alessandro Sette (La Jolla Institute for Allergy & Immunology). Mtb whole cell lysate (strain H37Rv) was kindly provided by Bei Resources.

Sequencing of single cell TCRs Single cells were sorted into 96-well plates containing 12µl of oneSTEP RT reaction buffer. The cells were then amplified for TCR β and TCR α sequence, using multiplex primers, a DNA-nesting and multiplex process by our methods previously described21,27. During the PCR priming, DNA multiplex barcodes were attached to each amplicon such that 96-well plates of single cells were processed on a single MiSeq 2x300bp sequencing run.

Bulk sequencing of TCRs Bulk collections of tetramer specific T-cell populations were collected into RLT lysis buffer (Qiagen). RNA was extracted from the pool

and subjected to amplification and DNA multiplex barcoding through the use of previously described multiplex primer sets and a previously described plate-based multiplex priming reaction. Using this method, up to 2 96-well plates of samples could be sequenced in parallel on a single MiSeq 2x300bp sequencing run to generate nearly 21 million reads.

Computational analysis of single cell and bulk TCR sequences Fastq reads were paired-end assembled and converted to fasta. The fasta sequence files were demultiplexed to assign every read to a plate and well. All reads were separated into subsets of 10k reads or less per file. Each file was submitted for parallel analysis using the previously described VDJFasta algorithm21,27-29. For single-cell samples, the total population of reads is analyzed within each given well, identifying a single cell only if empirically determined boundary cutoffs of dominance for a single TCRβ and TCRα clone are encountered, as previously reported12. The resulting full sequence for the TCRα chain(s) and TCRβ chain are then combined with any index FACS phenotypic markers specific to these single cells.

Computational error correction of bulk TCR sequences by replicates PCR error, PCR contamination, read error and sample swaps can all contribute to error when performing bulk sequencing. To mitigate errors in bulk sequencing, RNA from each sample was split and processed as duplicate technical replicates. Comparison of clone frequencies across replicates established confidence intervals in apparent clone frequencies, allowed calculation of R2 reproducibility of clone frequencies across replicates, and enabled elimination of PCR and sequencing read errors resulting in a clone appearing only in one replicate. Any clones not encountered across replicates were rejected, assumed to be either read errors or too low in abundance for reliable recovery across replicates. As each replicate was sequenced to an average depth of 10,000 reads, this procedure resulted in the reliable recovery of all clones with frequencies 5e-4. Within a sample, TCR reads differing from another more frequent clone by only one nucleotide were assumed to be read errors of the more abundant species and were collapsed into that higher frequency read.

Structural analysis of TCR positional antigen contact probability All amino acid sequences for all solved PDB structures were downloaded and scored against the

TCR profile HMM with an e-value cutoff of $<$ 1-e5, and blasted against a reference database of MHC sequences with an e-value cutoff of $<$ 1-e10. Sequences from structures containing both an MHC and TCR were aligned. Every residue in every TCR of such sequences was annotated as potential-contact if within 5 Angstroms of peptide in the peptide-MHC complex as determined by Modeller 9.17 and confirmed manually using UCSF Chimera. Using this data, an average positional contact probability was generated for each homologous position in the TCR sequence alignment. The positional contact probabilities were used as a weighting scheme to influence importance of convergence motifs at homologous positions by GLIPH. It was observed that CDR3b contacts were limited to IMGT positions 107-116 irrespective of whether the four solved structures containing convergence group representatives in Figure 3 were withheld from the dataset while calculating contact probability.

Naïve reference repertoire generation For this study, the naïve control dataset consists of 162,165 non-redundant V-J-CDR3 sequences from CD45RA+RO- naïve T-cells from 2 individuals11, 83,910 non-redundant V-J-CDR3 sequences from CD4 naïve T-cells from 10 healthy controls, and 27,292 non-redundant V-J-CDR3 sequences from CD8 naïve T-cells from 10 healthy controls12, for a total of 268,955 unique naïve V-J-CDR3 sequences from 12 individuals. CDR3 length distributions and CDR3 3mer motif composition was comparable in all reference sets (Extended Data Fig. 10SI_Figure 7A,B).

Calculating TCR global convergence Global similarity is defined as the CDR3 hamming distance (number of CDR3 amino acid differences) between two TCRs using the same V-beta segment and having a same-length CDR3. In order to identify a global similarity cutoff below which two TCRs can be assumed to share a common specificity, GLIPH performed repeat random CDR3-length stratified resampling of an unselected naïve TCR reference set. Using a sampling depth s of TCRs equal in size to the query set, GLIPH performs a large number (default=1000) of random samplings of s naïve TCR sequences. For each sampling, each TCR in the set is compared to every other sample, and the lowest global similarity is recorded. The proportion of all TCR similarity distances is then taken as a probability of observing

TCRs of that level of global similarity by chance in absence of selection. (For more details, see Supplementary MethodsSI_Doc1).

Calculating TCR local convergence Within any set of T-cell receptors, a collection of all continuous 2mers, 3mers, 4mers and 5mers, can be extracted and evaluated for their frequency within the set. Positive selection of each observed motif can be quantified by comparison to expected motif frequency distributions obtained during repeat resampling from an unselected repertoire (default 1000 random resamplings; Extended Data Fig. 10SI_Figure 7). A fold-change of enrichment can be calculated as the observed frequency of the motif over the expected frequency of the motif as observed in repeat random samplings from the naïve distribution. A probability of non-enrichment can be calculated as the proportion of random subsample simulations that obtain an unselected sample where the motif is at an equal or higher frequency than found in the observed set. Local convergence analysis is only performed within residues with at least a 5% probability of antigen contact (positions 107-116). Amino acid motif frequencies in the TCR sets were comperable in content and highly correlated in degree, with the result being that GLIPH results are robust to the specific naïve TCR reference set utilized (Extended Data Fig. 10bSI_Figure 7B). If each motif could only be observed in a given sequence once, then the distribution of sampling motif frequency means u become normally distributed and this result is equivalent to calculating the frequencies of all motifs in the reference database, and then calculating one-sided confidence intervals for expected frequencies of any given motif in the reference database at any given sampling depth:

(1) CI(99.9% os)=$\bar{y}\pm t\_0.0005\ (s/\sqrt{n})$

where n is the sample set non-redundant CDR3 sample size, y is the motif mean frequency, and s is the SD estimate at that sampling depth for the motif, as

(2) $s=\sqrt{((\sum(y\_1-\bar{y})\hat{}2\ )/(n-1))}$ (Supplementary MethodsSI_Doc1)

Generating and Scoring GLIPH specificity groups After analyzing global convergence cutoffs and local convergence motifs, GLIPH clusters all TCRs, creating an edge between TCRs that share either global similarity below the significance cutoff(i.e. differ by less than 2 amino acids), or share a significant motif (i.e. share a motif "RSS" that is >10-fold enriched and <0.001 probability of occurring that

this level of enrichment in naïve TCR pools). Clusters can be optionally filtered for shared V-gene usage, where only edges between TCR members with the most common V-gene are kept. These resulting clusters are GLIPH specificity groups. GLIPH specificity groups can be provided a score that combines the analysis of CDR3 motifs, enrichment of common V-genes, enrichment of a limited CDR3 length distribution, enrichment of clonally expanded clones, enrichment of shared HLA in donors, and cluster size. V-gene enrichment and CDR3 length distribution enrichment analysis is performed by calculating the Simpson diversity index for V-genes/CDR3-lengths within clonal members of a GLIPH specificity groups, and calculating the probability that a random selection of TCR sequences of the same size would generate an equal or superior Simpson score as the observed score.

(3) D=($\sum$(n(n-1)))/(N(N-1))

Enrichment of clonal expansion is similarly calculated as the probability that a random set of equal size from the same dataset would have at least this same number of expanded members. When HLA data is available, enrichment of HLA is calculated for each HLA allele found in at least two members in the GLIPH convergence group, in each instance calculating the probability that this particular HLA would be as enriched in a set of this size by random chance as is observed in the selected set. Global similarity is scored as previously defined. Local similarity significance is calculated as previously described. Finally, GLIPH cluster size can be scored by evaluating the probability of a network of that size forming by random chance in an unselected repertoire sampled at equal depth. The summary score of a any GLIPH cluster is a combination of all individual scores, calculated as either a probability conflation

(4) P(X=C)=($\prod$_i^NP_i (X=C))/($\prod$_i^NP_i (X=C)+$\prod$_i^NP_i (X!=C))

or the first principal component (Supplementary MethodsSI_Doc2, SI_Doc3, SI_TableSupplementary Table 4). For specificity groups based on local motifs, V-gene usage, CDR3 length and clonal expansion appear as independent variables. However, it should be noted that for specificity groups defined by global similarity, the CDR3 length, and to some extent V-gene usage, are no longer truly independent variables. (GLIPH available at https://github.com/immunoengineer/gliph ; For more details, see Supplementary Methods, Supplementary codeSI_Doc1, SI_Doc2, SI_Doc3)

Predicting Specificity and De Novo TCR specificity design with GLIPH For a given convergence group, an N-terminal Positional Weight Matrix (PWM) can be constructed of the CDR3 by creating a left-justified alignment of CDR3 amino acid residues (Fig.ure 5bB) and tabulating the frequency of each amino acid at each homologous position (Fig.ure 5cC). A score of any TCR sequence to the PWM can then be calculated as the product of the probability of each amino acid in the scored sequence at the homologous position in the PWM, employing pseudocounts of 0.5% for any amino acid not observed in the PWM input alignment (Fig.ure 5dD). When attempting to identify new members of an existing GLIPH specificity group, only the first 10 N-terminal amino acid positions are used during scoring: this allows recognition of new TCRs of different lengths from the PWM, and leverages an observation that the conserved motifs always appear to be fixed a specific length from the N-terminus and within the first 10 amino acids (Extended Data Fig. 4dFigure 3D).

(5) s=$\prod$_(i=1)^10Pr(a_i |PWM)

where PWM is the positional weight matrix of amino acid frequencies per position, i is the amino acid position in the PWM, and a_i is the frequency of amino acid a at position i in the PWM. The resulting score s can be normalized by comparison to a large set of naïve TCRs as

(6) Pr_s=($\sum$_(n=1)^N(S_n |S_n$\geq$(s-v)))/N

where s is the PWM score of a TCR under evaluation (Formula 5), N is the set of naïve TCR database (200,000 for Extended Data Fig. 4dFigure 3D), Sn is the PWM score of naïve TCR n of set N, and Prs is the probability of that PWM score occurring in naïve TCRs. To account for V-gene mismatch, a V-gene mismatch penalty v is applied during scoring (v=-2 in Extended Data Fig. 4dFigure 3D).

When attempting to de-novo synthesize new members of an existing GLIPH specificity group, a global PWM of CDR3s of the same length is used. The top 1000 highest predicted scoring TCRs are emitted from the PWM by stochastic sampling, and TCRs with the highest scores are preferentially tested (Fig.ure 5dD, e5E). For normalization, the resulting score can be compared to a distribution of a large number of naïve sequences scored against the same PWM, to produce a probability of membership.

Lentiviral TCR transduction Plasmids for lentiviral transduction were kindly provided by Crabtree lab in Stanford University. Lentiviral transduction was done as previously described30. Briefly, TCR α chain, P2A linker and β chain fusion gene fragments were ordered from IDT and cloned into MCS of N103 vector (nLV Dual Promoter EF-1a-MCS-PGK-Puro). A GFP marker was also included through T2A linker. HEK-293T cells were plated on 10-cm dishes at 5x106 cells/plate 24 h prior to transfection. The culture medium was changed prior to transfection. Lentiviral supernatants were prepared by co-transfection of 293T cells, using 10 μg of transfer vector, 7.5 μg of envelope vector (pMD2.G), 2.5 μg of packaging vector (psPAX2) and 75 μl PEI (Sigma). The culture medium was replaced 16 hours after transfection and viral supernatant was collected 48 h later. The viral supernatants were filtered through a 0.45 μm SFCA syringe filter (Corning) and concentrated by centrifuge with 100K Amicon Ultra-15 filter (Millipore). Concentrated viruses were used for J76-NFATRE-luc cell transduction using spinoculation for 2 hours in the presence of 6 μg/ml polybrene (Sigma). Forty-eight hours after transduction, expression of the TCR was analyzed by flow cytometry and both GFP and TCR positive cells were sorted for epitope screen.

Epitope screen For peptide screen, 100 μl TCR transduced J76-NFATRE-luc cells (106/ml) were co-cultured with 100 μl HLA transduced K562 cells (106/ml) in a 96-well plate. Peptide pool or individual peptide was added to the well at 2 μg/ml. After 8 hours incubation, cells were harvested and Luciferase activity was measured using Nano-Glo® Luciferase Assay (Promega). Fold induction of luciferase activity was calculated referring to unstimulated samples.

2.2.5 Acknowledgements

This work was made possible by excellect co-authors, in particular my good fortune to be allied to my primary co-author Huang Huang, and well as the enourmous contributions by Allison Nau and Olivia Hatton. I'd also like to thank Lisa E. Wagar, Florian Rubelt, Xuhuai Ji, Arnold Han, Sheri M. Krams, Christina Pettus, Nikhil Haas, Cecilia S. Lindestam Arlehamn, Alessandro Sette, Scott D. Boyd, Thomas

J. Scriba, and theoretical guidance by Olivia M. Martinez, and Mark M. Davis for making this work possible. Further, we collectively would like to thank the Stanford Human Immune Monitoring Center for their high-throughput sequencing support for this project, Michael Mindrinos and co-workers at Sirona Genomics for the HLA typing. We especially thank The Bill and Melinda Gates Foundation, The National Institutes of Health (2U19 AI057229) and the Howard Hughes Medical Institute for financial support, Dr. Shao-An Xue (Department of Immunology, University College London) for providing Jurkat 76 T-cell line, Chiung-Ying Chang and Ruth Taniguchi for the help with lentiviral transduction, Rachel Hovde for valuable discussions regarding statistical measures, Hassan Mahomed, Willem Hanekom and members of the Adolescent Cohort Study (ACS) group for enrolment and follow-up of the Mtb-infected adolescents, and Rishi Bedi for his assistance in collecting TCR sequences from the literature. Sorting was (partially) performed on an instrument in the Shared FACS Facility obtained using NIH S10 Shared instrument grant (S10RR025518-01).

2.2.6 References

1 Arstila, T. P. et al. A direct estimate of the human alphabeta T cell receptor diversity. Science 286, 958-961 (1999).

2 Davis, M. M. & Bjorkman, P. J. T-cell antigen receptor genes and T-cell recognition. Nature 334, 395-402, doi:10.1038/334395a0 (1988).

3 Qi, Q. et al. Diversity and clonal selection in the human T-cell repertoire. Proceedings of the National Academy of Sciences of the United States of America 111, 13139-13144, doi:10.1073/pnas.1409155111 (2014).

4 Shortman, K., Egerton, M., Spangrude, G. J. & Scollay, R. The generation and fate of thymocytes. Seminars in immunology 2, 3-12 (1990).

5 Naylor, K. et al. The influence of age on T cell generation and TCR diversity. Journal of immunology 174, 7446-7452 (2005).

6 Robins, H. S. et al. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. Blood 114, 4099-4107, doi:10.1182/blood-2009-04-217604 (2009).

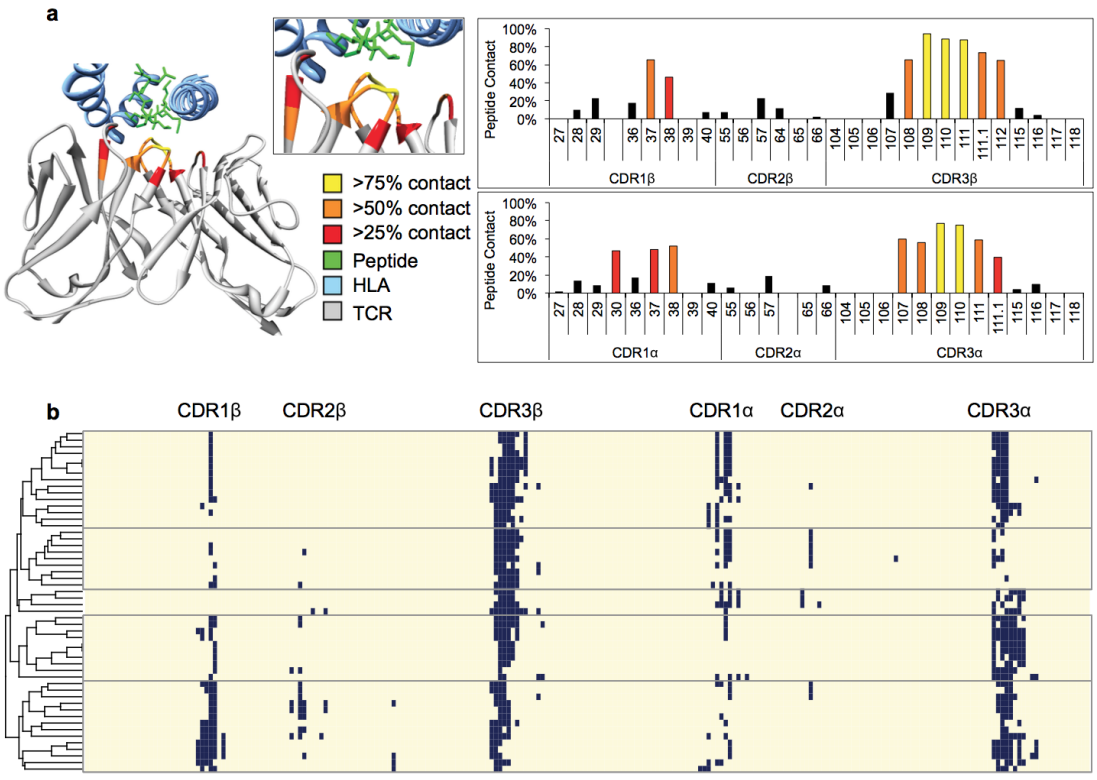


Figure 2.7: TCRs specific to common antigen show motifs within a limited region of CDR residues with high structural contact propensity. a, Probability of IMGT TCR CDR positions being within 5 Angstroms of peptide antigen, as tabulated from 52 published crystal structures of TCR-pMHC interactions (Supplementary Table 2), and displayed as a heatmap on representative TCR 2j8u. Positions with less than 25% contact probability are shown in black. b, Alignment of 52 non-redundant (<95% amino acid identity between any pair) TCR sequences from TCR-peptide-MHC PDB structure complexes. Positions within 5 Angstroms of peptide antigen are indicated in blue. Linear set of 3-5 amino acids in CDR3β observed in almost every structure, which TCRβ-CDR3 IMGT positions 108-111 being in contact in 90% of TCR structures. Minimal contacts observed by CDR1 and CDR2 of either chain. TCRs are clustered into 5 general contact modes according to contact profiles of all 6 CDRs.

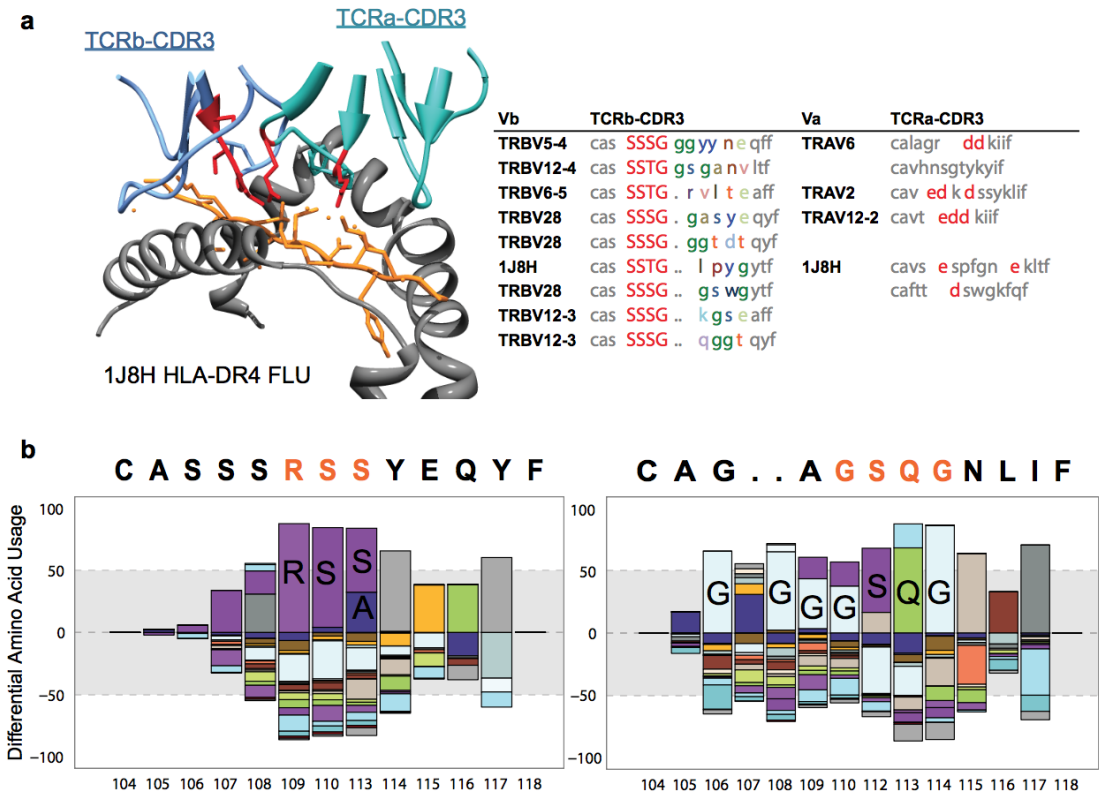


Figure 2.8: Crystal structure representative of TCR specificity groups. a, Class II single cell paired α/β sequencing with crystal structure representative indicating variable CDR3β length and discontinuous role of CDR3α. Discontinuous negative charged residues in structure 1J8H coordinate lysine positive charges in peptide; negative charged residues indicated in orange in alignment when found. b, Positional amino acid bias in flu HLA-A2 dominant motif CDR3β and CDR3α convergence group, normalized by amino acid amino acid diversity in the unselected repertoire. Enrichment of RS[S/A] motif in TCRβ compared with naïve distribution. Enrichment of SQ at IMGT positions 112, 113 in TCRα, with enrichment of Glycine at multiple positions.

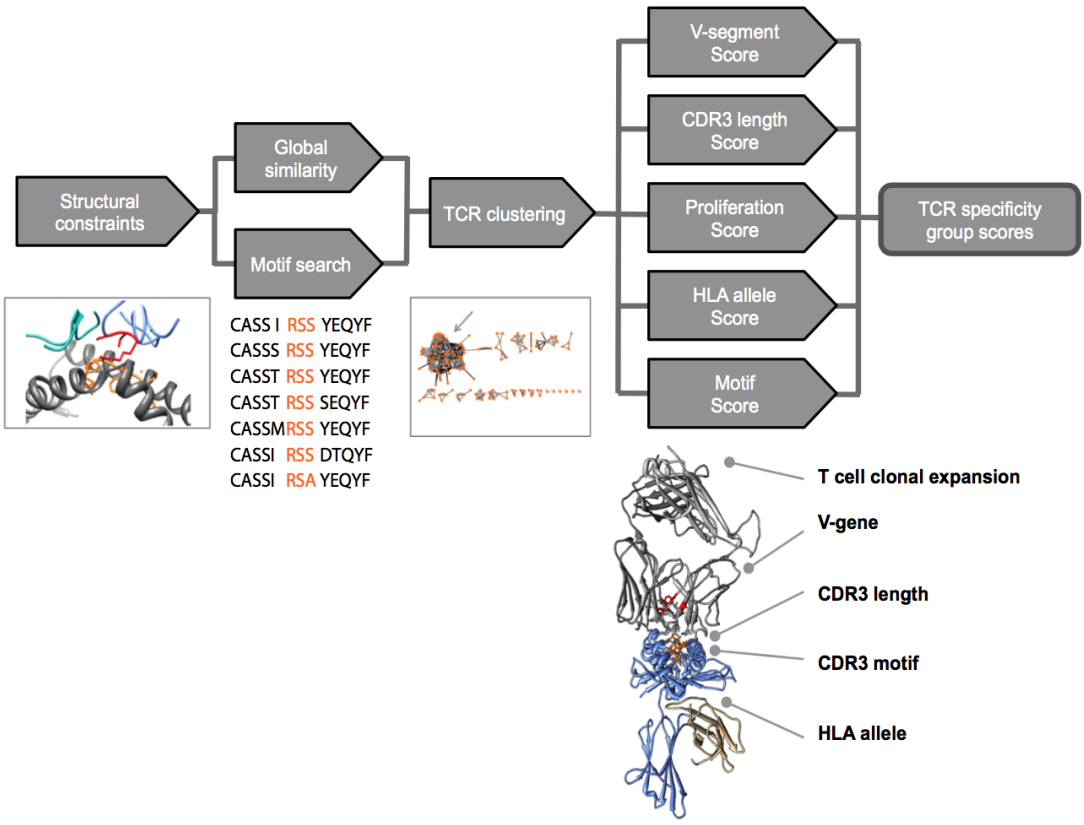


Figure 2.9: Three-step GLIPH algorithm. First, GLIPH searches for global and local (motif) CDR3 similarity in TCR CDR regions with high contact probability. Motif significance and global similarity cutoffs are established by repeat random sampling against an unbiased reference pool of TCRs. Second, all identified global and local relationships between TCRs are used to construct clusters of TCR specificity groups. Third, each specificity group is analyzed for enrichment of common V-genes, CDR3 lengths, clonal expansions, shared HLA alleles in recipients, motif significance, and cluster size. Enrichment probability is obtained by calculating the probability of obtaining at least the observed Simpson diversity index measure for that feature compared with a random sampling of equal size from the source dataset. The resulting features are combined into a specificity group score for each group.

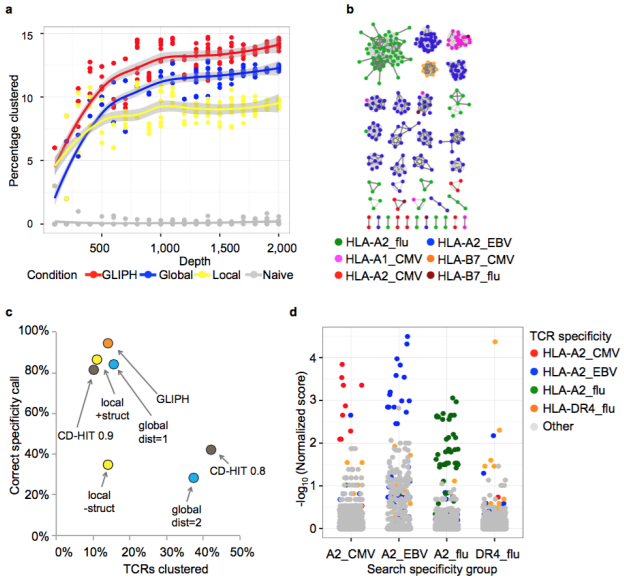


Figure 2.10: Benchmark of GLIPH subcomponents and complete algorithm on random naïve TCRs or a mixed training set pool of pMHC tetramer+ TCRs of 8 known specificities. a, GLIPH clusters up to 14.5% of tetramer+ TCRs while clustering less than 0.5% of naïve TCRs, a combination of global CDR3 similarity and local motif enrichment resulting in more clustering than either individually. b, The cluster results of applying GLIPH to the mixed pool of tetramer-sorted TCRs. Each node is a TCR, their specificity indicated by color. Edges between TCRs indicating a GLIPH-predicted shared specificity; light grey indicate shared local motif, and dark grey indicate shared global similarity. Over 95% of cluster members are grouped with other TCRs of the same specificity. c, GLIPH components evaluated for percentage of TCRs clustered vs percentage of correct specificity assignments. Global CDR3 clustering by hamming dist=1 or dist=2 reported. Global CDR3 similarity clustering by CD-HIT, with clustering cutoffs 0.8 or 0.9 reported. Local motif similarity clustering with and without structural constraints reported. Complete GLIPH, including global CDR3 identity, local CDR3 motif similarity, structural constraints and clustering scoring, resulted in 14.5% of TCRs clustering with 95% of cluster members correctly grouped with other TCRs of shared specificity. For global similarity, distance 1 resulted in effective grouping of TCRs while distance 2 resulted in predominantly mixed clusters. For local motifs, effective TCR clustering could only be obtained when structural contact probability masks were applied. Similarily, although CD-HIT was not effective at clustering TCRs by common specificity when provided the entire TCR sequences, when offered only the high contact probability CDR3s, it was able to perform effective clusters provided an appropriate clustering threshold. d, When run on replicate A containing TCRs from half of study subjects, GLIPH produced specificity groups whose positional weight matrices (PWMs) could then be used to score the TCRs from replicate B subjects (Formula 5,6). GLIPH scoring identifies new TCRs of correct specificity from new subjects.

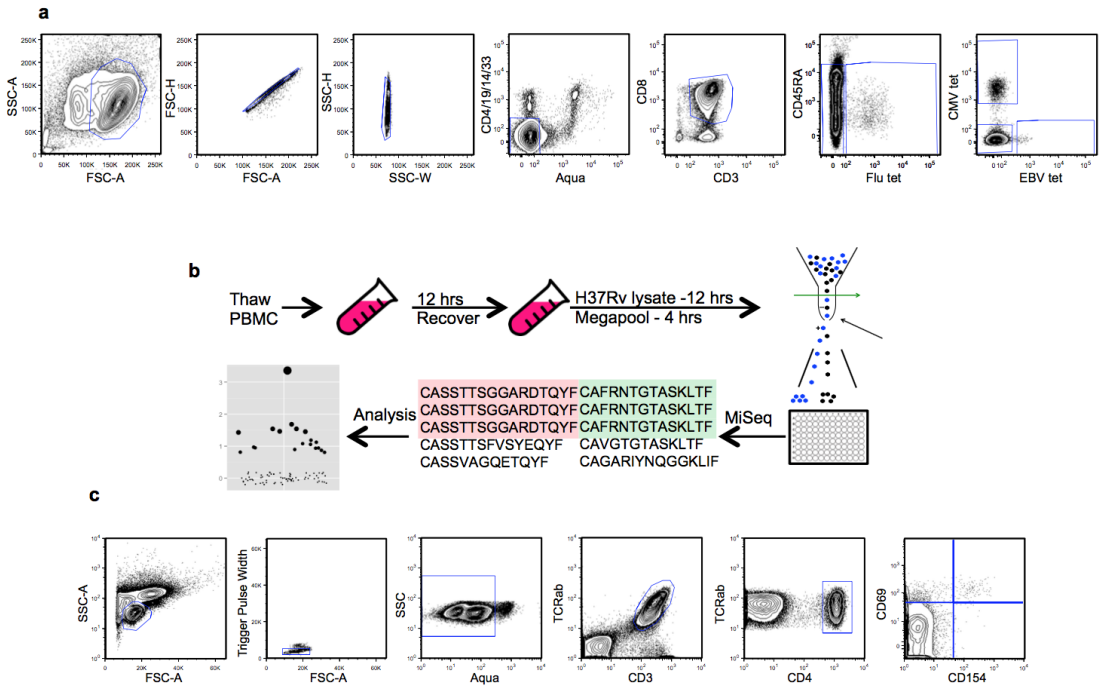


Figure 2.11: Platform for PBMC stimulation and characterization of antigen-specific TCRs. a, Gating strategy used for isolating and sorting tetramer positive T cells. b, Frozen PBMCs from QFN+ donors are thawed, recovered and stimulated with either Mtb lysate or peptide pool. Antigen-specific T cells are single-cell sorted into 96-well plate for TCR amplification using established protocol (Han, et al,). c, Gating strategy used for isolating and single-cell sorting antigen-specific T cells.

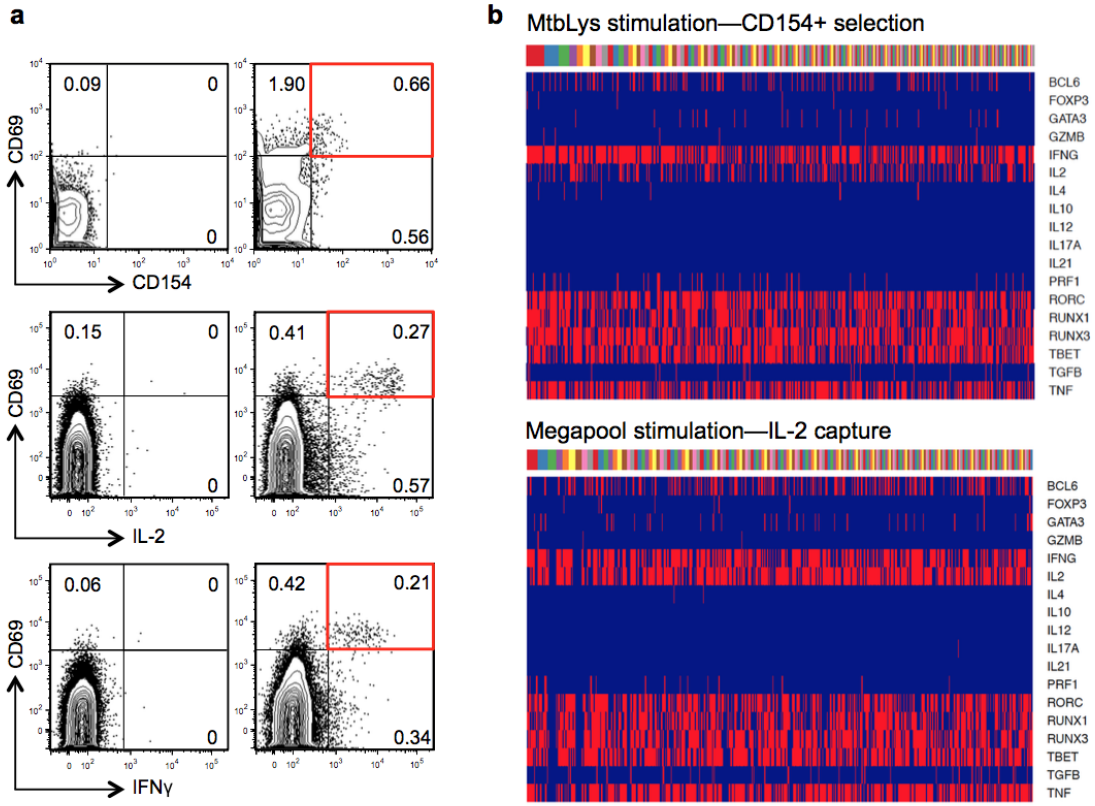


Figure 2.12: Phenotypic analysis of clonal expanded Mtb-specific CD4+ T cells. a, Gating strategy for isolating antigen-specific T cells. PBMC from one QFN+ donor (02/0259) was stimulated with Mtb lysate and then stained with activation markers CD69 and CD154. Antigen-specific CD4+ T cells were sorted by gating on CD69+CD154+ population. Alternatively, PBMC was stimulated with Megapool peptide library. Antigen-specific CD4+ T cells were isolated using cytokine capture assay, IL-2 or IFNγ. b, 18-parameter (parameters listed on right side) phenotypic analysis of Mtb-specific CD4+ T cells from all the 22 donors. Individual T cells are grouped by TCR sequence; each color on the bar above the heat maps represents a distinct and clonal expanded TCR sequence. The majority of cells presented a Th1*-like phenotype including IFNγ and IL-2 production, T-bet and RORC expression, as is characteristic of previously reported Mtb responses.

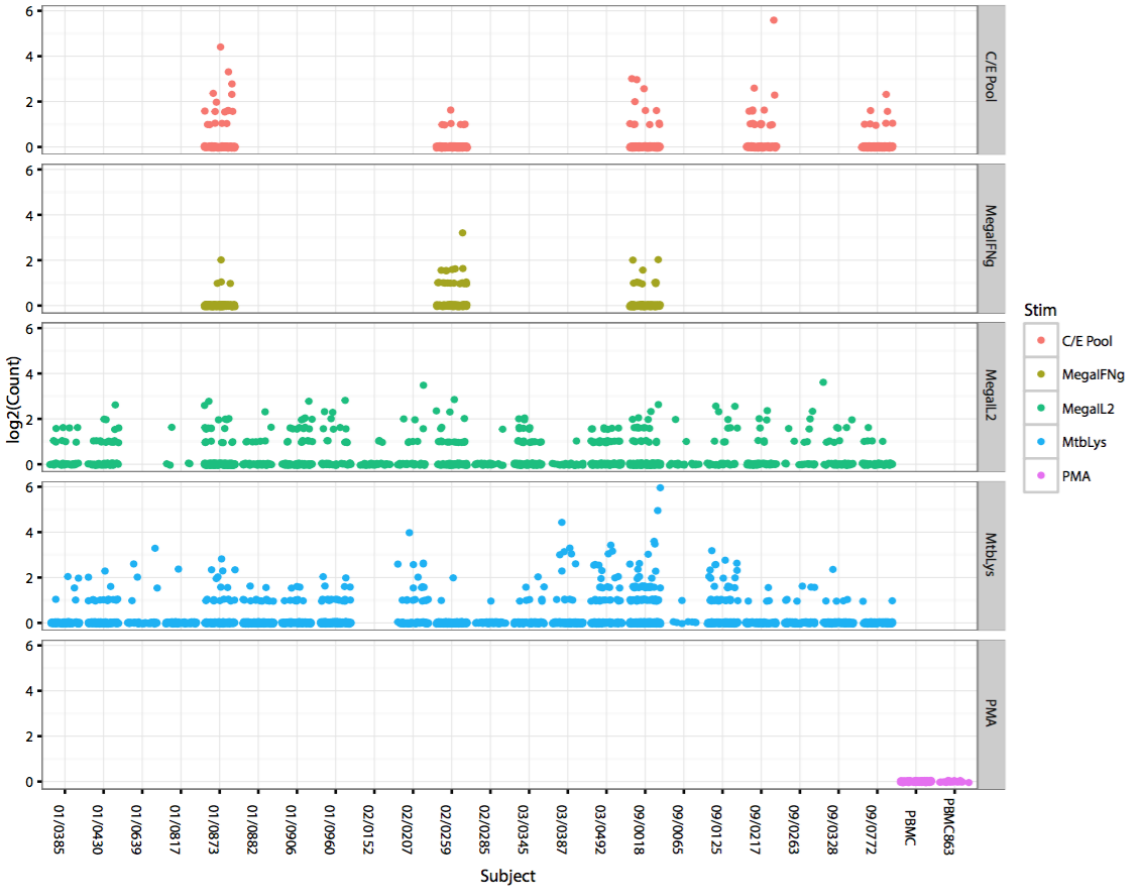


Figure 2.13: Clonal expansion of Mtb-specific CD4+ T cells. Clonal analysis of Mtb-specific CD4+ T cells from all the 22 donors using different selection strategy, including stimulation by ESAT6/CFP-10 pool (C/E Pool) or Megapool followed by cytokine capture assay and Mtb lysate stimulation followed by CD154+ selection. Each dot represents a distinct TCR sequence and the count represents the number of repeat. PMA/Ionomycin stimulation was used as a non-specific stimulation control.

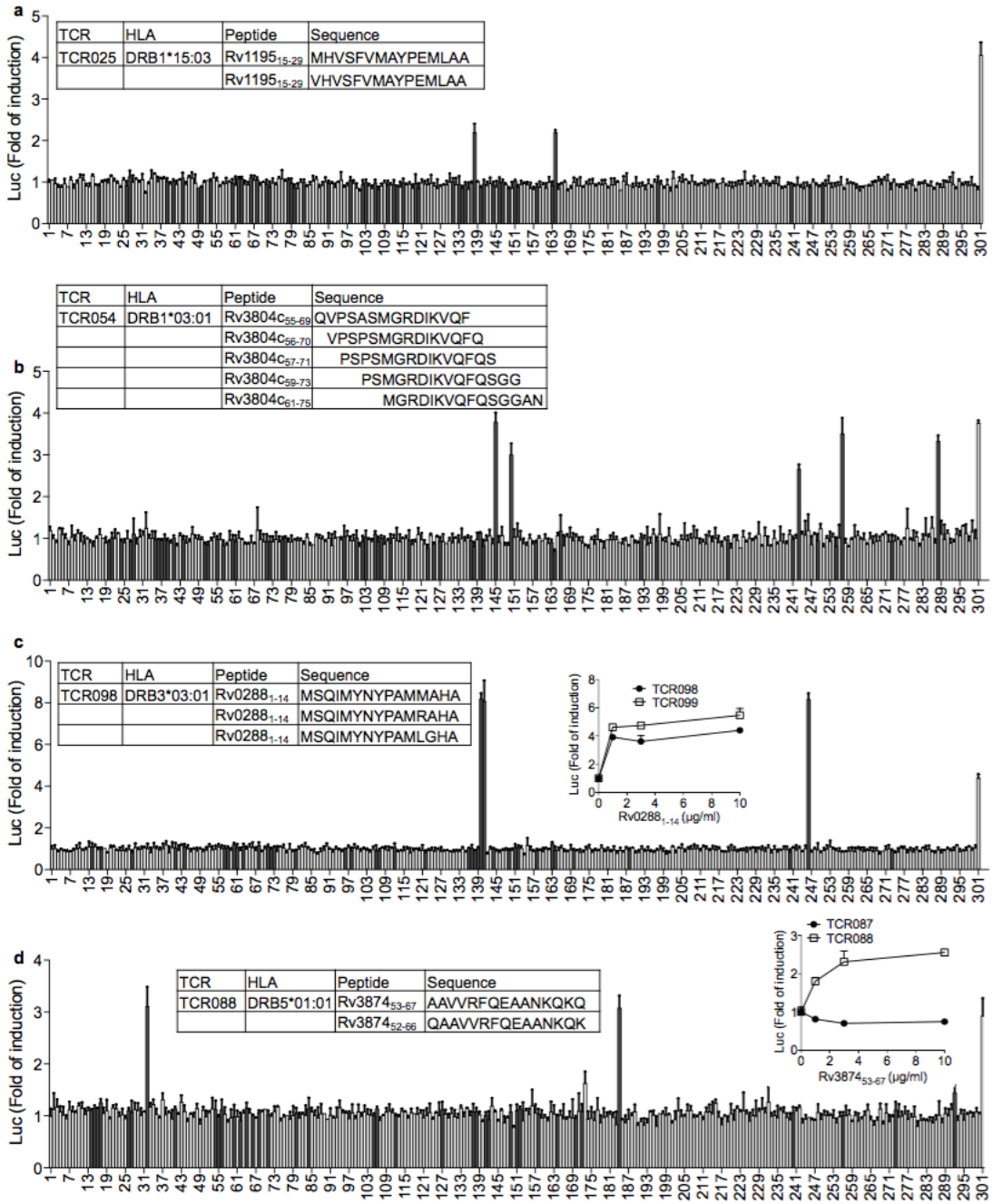


Figure 2.14: Epitope screen using luciferase assay. a, Each individual peptide from Megapool was tested against J76-NFATRE-luc cell expressing TCR025 in co-culture with K562 expressing DRB1*15:03. Column 1-300: individual peptide from Megapool, column 301: CD3/CD28 stimulation as positive control. Peptides predicted to be in the top 15 percentile of binding to each HLA by the MHC-II Consensus method are indicated by grey bars. Mean ± s.d. (n=3) are shown. The insert table shows the restricted HLA type and responding peptides. Similar screen was also done for TCR054 (b), TCR098 (c) and TCR088 (d).

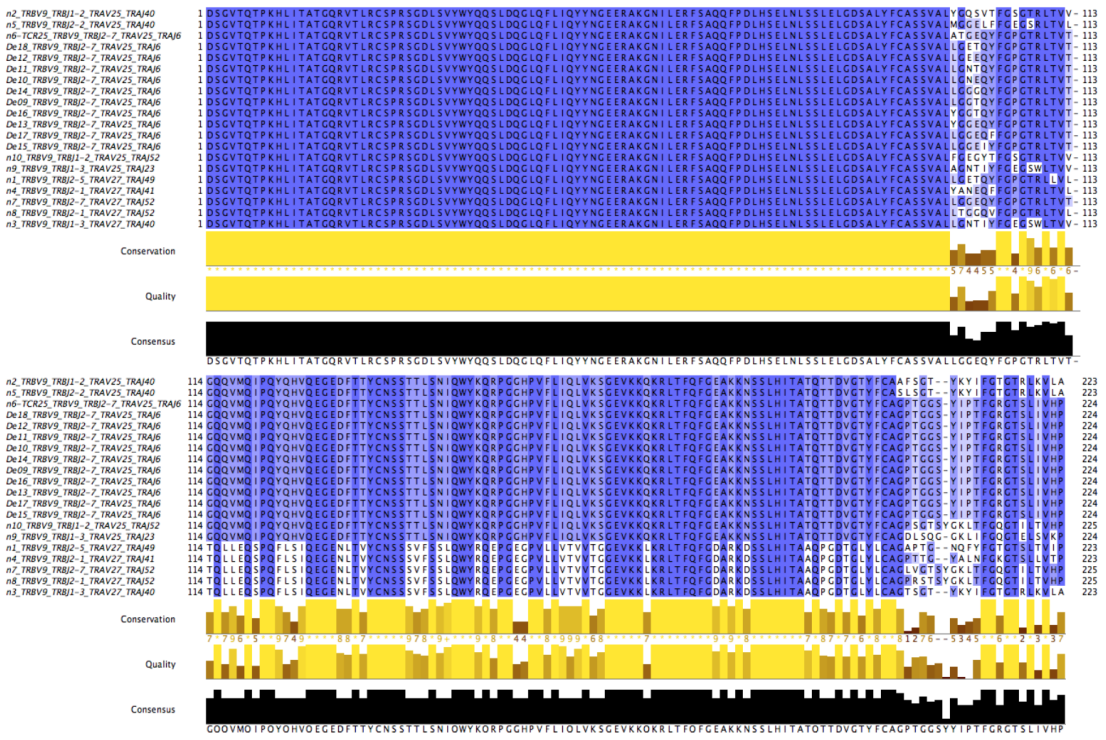


Figure 2.15: Amino acid alignment of naturally occurring and de novo Group II TCRs. Amino acid alignment presents first the TCRβ chain followed the TCR α chain for naturally occurring Group II TCRs n1-10 and de novo TCRs De9-De18. All segment identities are reported for each sequence in the sequence headers. Positional conservation is colored as dark blue if conserved, and light blue or white if variable.

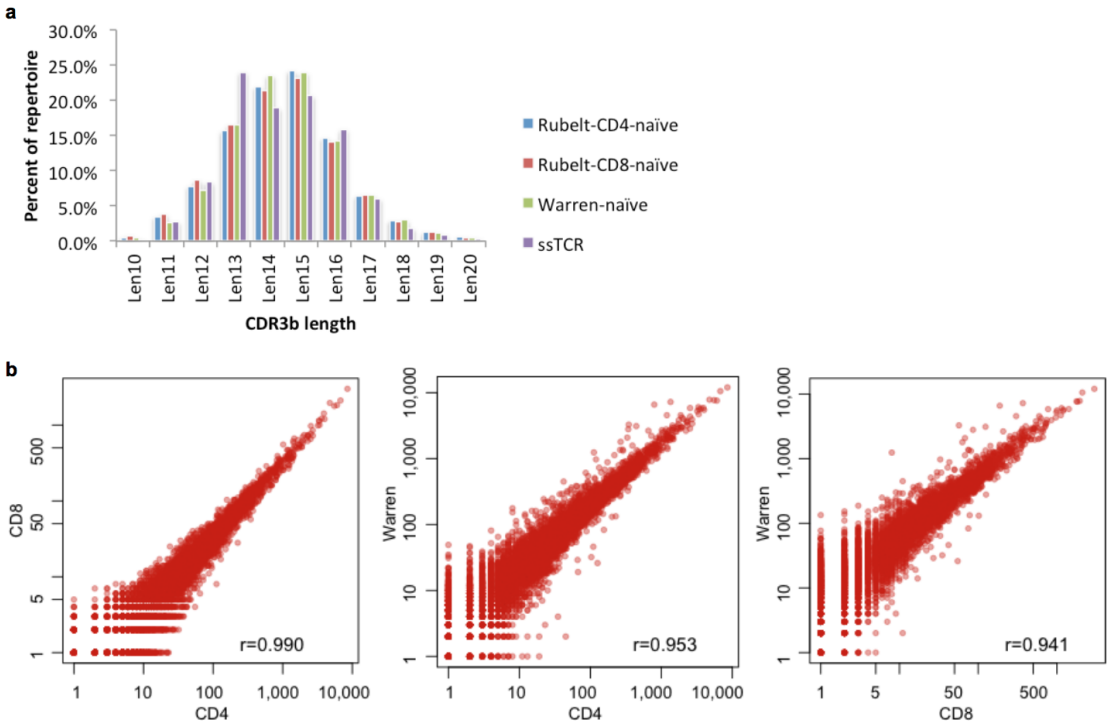


Figure 2.16: Comparison of CDR3 Length and 3mer motif composition of naïve TCR reference set. The naïve control dataset consists of 162,165 non-redundant V-J-CDR3 sequences from CD45RA+RO- naïve T-cells (Warren et al, Genome Research 2011), 83,910 non-redundant V-J-CDR3 sequences from CD4 naïve T-cells from 10 healthy controls, and 27,292 non-redundant V-J-CDR3 sequences from CD8 naïve T-cells from 10 healthy controls (Rubelt et al, ref. 8), for a total of 268,955 unique naïve V-J-CDR3 sequences. Analysis of CDR3 length distributions (a) and motif frequency distributions (b) indicates that the three naïve reference sets have very similar CDR3 length distributions and 3mer amino acid motif frequency distributions (r=0.99, r=0.95, and r=0.94 Pearson correlation coefficients for CD4xCD8, CD4xWarren, and CD8xWarren, respectively).

7 Garcia, K. C. et al. An alphabeta T cell receptor structure at 2.5 A and its orientation in the TCR-MHC complex. Science 274, 209-219 (1996).

8 Birnbaum, M. E. et al. Deconstructing the peptide-MHC specificity of T cell recognition. Cell 157, 1073-1087, doi:10.1016/j.cell.2014.03.047 (2014).

9 Rudolph, M. G., Stanfield, R. L. & Wilson, I. A. How TCRs bind MHCs, peptides, and coreceptors. Annual review of immunology 24, 419-466, doi:10.1146/annurev.immunol.23.021704.115658 (2006).

10 Berman, H. M. et al. The Protein Data Bank. Nucleic acids research 28, 235-242 (2000).

11 Warren, R. L. et al. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. Genome research 21, 790-797, doi:10.1101/gr.115428.110 (2011).

12 Rubelt, F. et al. Individual heritable differences result in unique cell lymphocyte receptor repertoires of naive and antigen-experienced cells. Nature communications 7, 11112, doi:10.1038/ncomms11112 (2016).

13 Thomas, N. et al. Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. Bioinformatics 30, 3181-3188, doi:10.1093/bioinformatics/btu523 (2014).

14 Bolotin, D. A. et al. MiTCR: software for T-cell receptor sequencing data analysis. Nature methods 10, 813-814, doi:10.1038/nmeth.2555 (2013).

15 Nazarov, V. I. et al. tcR: an R package for T cell receptor repertoire advanced data analysis. BMC bioinformatics 16, 175, doi:10.1186/s12859-015-0613-1 (2015).

16 Yu, Y., Ceredig, R. & Seoighe, C. LymAnalyzer: a tool for comprehensive analysis of next generation sequencing data of T cell receptors and immunoglobulins. Nucleic acids research 44, e31, doi:10.1093/nar/gkv1016 (2016).

17 Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658-1659, doi:10.1093/bioinformatics/btl158 (2006).

18 Chattopadhyay, P. K., Yu, J. & Roederer, M. A live-cell assay to detect antigen-specific CD4+ T cells with diverse cytokine profiles. Nature medicine 11, 1113-1117, doi:10.1038/nm1293 (2005).

19 Frentsch, M. et al. Direct access to CD4+ T cells specific for defined antigens according to CD154 expression. Nature medicine 11, 1118-1124, doi:10.1038/nm1292 (2005).

20 Lindestam Arlehamn, C. S. et al. A Quantitative Analysis of Complexity of Human Pathogen-Specific CD4 T Cell Responses in Healthy M. tuberculosis Infected South Africans. PLoS pathogens 12, e1005760, doi:10.1371/journal.ppat.1005760 (2016).

21 Han, A., Glanville, J., Hansmann, L. & Davis, M. M. Linking T-cell receptor sequence to functional phenotype at the single-cell level. Nature biotechnology 32, 684-692, doi:10.1038/nbt.2938 (2014).

22 Lindestam Arlehamn, C. S. et al. Memory T cells in latent Mycobacterium tuberculosis infection are directed against three antigenic islands and largely contained in a CXCR3+CCR6+ Th1 subset. PLoS pathogens 9, e1003130, doi:10.1371/journal.ppat.1003130 (2013).

23 Sallusto, F. Heterogeneity of Human CD4(+) T Cells Against Microbes. Annual review of immunology 34, 317-334, doi:10.1146/annurev-immunol-032414-112056 (2016).

24 Sharma, G. & Holt, R. A. T-cell epitope discovery technologies. Human immunology 75, 514-519, doi:10.1016/j.humimm.2014.03.003 (2014).

25 Wang, P. et al. Peptide binding predictions for HLA DR, DP and DQ molecules. BMC bioinformatics 11, 568, doi:10.1186/1471-2105-11-568 (2010).

26 Mahomed, H. et al. Predictive factors for latent tuberculosis infection among adolescents in a high-burden area in South Africa. The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease 15, 331-336 (2011).

27 Han, A. et al. Dietary gluten triggers concomitant activation of CD4+ and CD8+ alphabeta T cells and gammadelta T cells in celiac disease. Proceedings of

the National Academy of Sciences of the United States of America 110, 13073-13078, doi:10.1073/pnas.1311861110 (2013).

28 Glanville, J. et al. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. Proceedings of the National Academy of Sciences of the United States of America 108, 20066-20071, doi:10.1073/pnas.1107498108 (2011).

29 Glanville, J. et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proceedings of the National Academy of Sciences of the United States of America 106, 20216-20221, doi:10.1073/pnas.0909775106 (2009).

30 Hathaway, N. A. et al. Dynamics and memory of heterochromatin in living cells. Cell 149, 1447-1460, doi:10.1016/j.cell.2012.03.052 (2012).

2.2.7 Copyright

2.3 Reading specificity in the γδ T-cell receptor repertoire

Interleukin (IL)-17 plays a key role in immunity. In acute infections, a rapid IL-17 response must be induced without prior antigen exposure, and γδ T cells are the major initial IL-17 pro- ducers. In fact, some γδ T cells make IL-17 within hours after an immune challenge. These cells appear to acquire the ability to respond to IL-1 and IL-23 and to make IL-17 naturally in naïve animals. They are known as the natural T γδ 17 (nT γδ 17) cells. The rapidity of the nT γδ 17 response, and the apparent lack

of explicit T cell receptor (TCR) engagement for its induc- tion have led to the view that this is a cytokine (IL-1, IL-23)-mediated response. However, pharmacological inhibition or genetic defects in TCR signaling drastically reduce the nT γδ 17 response and/or their presence. To better understand antigen recognition in this rapid IL-17 response, we analyzed the antigen receptor repertoire of IL-1R + /IL-23R + γδ T cells, a proxy for nT γδ 17 cells in naïve animals directly ex vivo , using a barcode-enabled high through- put single-cell TCR sequence analysis. We found that regardless of their anatomical origin, these cells have a highly focused TCR repertoire. In particular, the TCR sequences have limited V gene combinations, little or no junctional diversity and much reduced or no N region diversity. In contrast, IL-23R − cells at mucosal sites similar to most of the splenic γδ T cells and small intestine epithelial γδ lymphocytes expressed diverse TCRs. This remark- able commonality and restricted repertoire of IL-1R + /IL-23R + γδ T cells underscores the importance of antigen recognition in their establishment/function.

2.3.1 Introduction

Interleukin (IL)-17 is an important cytokine in the inflamma- tory response. It induces chemokines and cytokines that mediate the maturation and release of neutrophils from the bone mar- row. Neutrophil recruitment focuses the immune response at the site of infection to reduce pathogen load, and induces the subsequent phases of the inflammatory response, which primes antigen-specific αβ T cell and B cell activation and initiates the resolution program. Although both αβ T cells and γδ T cells can make IL-17, αβ T cells producing IL-17 (Th17 cells) require antigen-specific priming and a specific cytokine environment to develop. In acute infections, a rapid IL-17 response must be ini- tiated without prior antigen exposure, and γδ T cells have been identified as the major initial IL-17 producers in infections and after immunization [reviewed in Ref. (1)].

Some naïve γδ T cells in secondary lymphoid organs undergo antigen-driven activation and differentiation to become IL-17 producers: within 24 h after immunization,

antigen-specific γδ T cells in the draining lymph node increase in numbers and show activated phenotypes (e.g., becoming CD44 hi and CD62L lo).

Forty-eight hours after immunization, activated γδ T cells express ROR γ t and after another 12 h, these cells make IL-17A and IL-17F (2, 3), these are the inducible T γδ 17 cells. Importantly, encoun- tering antigen in an immune response induces the expression of inflammatory cytokine receptors such as IL-1R and IL-23R on γδ T cells. Signaling through the T cell receptor (TCR) and the cytokine receptors can then induce sustained, high magnitude IL- 17 production (2, 4). These observations provide a mechanistic basis for the induction of a sustained antigen-specific γδ T cell IL-17 response, which is much more rapid than that of Th17 αβ T cells.

In addition to the inducible T γδ 17 cells discussed above, some γδ T cells in naïve mice, such as those in the skin dermis, the peri- toneum, intestinal lamina propria, the lung, and the spleen have an activated phenotype (CD44 hi CD62L lo) and express IL-1R and IL- 23R. These cells make IL-17 a few hours after immune challenge- these are the natural T γδ 17 (nT γδ 17) cells (1). The observation that IL-17 can be induced with IL-1 and IL-23 alone without deliberate TCR triggering has led to the supposition that the anti- gen recognition by these cells is irrelevant to their response (5).

Nonetheless, this response is inhibited by cyclosporine A (CsA) or by FK506 (2). Both compounds reduce nuclear factor of activated T cells (NFAT) activity and disrupt the calcineurin-NFAT signal- ing circuit activated by signaling through the antigen receptor (6). Furthermore, the amount of IL-17 induced by the inflammatory cytokines alone is much lower in magnitude when compared with that induced by cytokines together with TCR stimulation (2, 4), suggesting that robust IL-17 production requires combined signal- ing through the TCR and cytokine receptors. Moreover, the num- ber of rapid IL-17 responding IL-1R + γδ T cells in the intestinal lamina propria and peritoneum is markedly reduced in germ free mice, and in SPF mice treated with the antibiotic neomycin sul- fate, vancomycin but not in mice treated with metronidazole when compared with SPF mice and the numbers can be restored by SPF microbiota reconstitution. However, the presence of these IL-1R + γδ T cells requires signaling through VAV1, a guanine nucleotide exchange factor required for

the activation of γδ T cells via γδ TCR ligation (7), but not the myeloid differentiation primary response protein 88 (MyD88) or toll-like receptor 3 signaling pathways (8). In addition, the number of nT γδ 17 cells is drastically reduced in the SKG mouse (9), which carries a mutation that reduces the function of the kinase domain of the TCR-proximal signaling kinase Zap70. These observations demonstrate the importance of TCR signaling in nT γδ 17 induction and function. To evaluate the contribution of antigen recognition to their function, we seek to determine the antigen receptor repertoire of nT γδ 17 cells. To this end, we use a bar-code-enabled high throughput single-cell TCR sequencing strategy, which allows us to identify the TCR γ and δ gene pair from each cell directly ex vivo , without the bias introduced through generating T cell clones or hybridomas. This method determines the entire sequence of both the TCR γ and δ chains, including the V gene segment and CDR3 region, such that we can properly define the antigen receptor specific repertoire, rather than describing these cells solely based on their V γ or V δ usage. The results are discussed below.

2.3.2 Results

A defining feature of nT γδ 17 cells is their surface expression of IL-1R and IL-23R in naïve animals. To determine the antigen receptor repertoire of γδ T cells that are "poised" to mount a rapid IL-17 response, we analyzed skin dermal cells, and IL-23R + γδ T cells in the colon lamina propria, fat, and spleen of naïve IL-23R reporter mice (IL-23R EGFP). Peritoneal nT γδ 17 cells are charac- terized by their IL-1R expression in rapid response situations (8); therefore, we analyzed IL-1R + peritoneal γδ T cells from C57/BL6 mice that were intra-peritoneally (i.p.) infected with T. gondii 5 h prior. Representative FACS analysis and gates used to isolate these cells are shown in Figure 1 . The TCR sequences were determined from a single FACS sorted γδ T cell using a bar-code-enabled high throughput single-cell TCR sequencing strategy. We found that IL-17F + spleen γδ T cells from naïve IL-17F reporter mice (Il17f thy1.1/thy1.1) and IL-23R + spleen γδ T cells from naïve IL-23R reporter mice have similar TCR repertoires (Figure 2). This obser- vation is consistent with the

supposition that IL-23R + γδ T cells in naïve animals can be used as a proxy for nT γδ 17 cells in TCR repertoire analysis.

A striking characteristic of the TCR repertoire of IL-1R + /IL- 23R + γδ T cells is the lack of diversity. They express TCRs with limited V gene combinations, little or no junctional diversity and much reduced or no N region diversity. In particular, a single pair of TCR sequences encoded by V δ 1D δ 2J δ 1 and V γ 6J γ 1 (Group 1 sequences, Figure 2) dominates the repertoire of dermal cells, IL-23R + γδ T cells from the lung, colon, and IL-1R + γδ T cells from the peritoneum. These cells also utilize two sets of closely related TCR sequences, which consist of similar V γ 4J γ 1 rearrange- ments, paired with very similar V δ 5D δ 2J δ 2 (designated as Group 2, 3 sequences, Figure 2). Naïve spleen IL-23R + and IL-17F + T cells did not have a dominant population that expressed Group 1 sequences. Instead, cells with the Group 3 sequences were more represented. Some of these γδ T cells also expressed TCRs consist- ing of Group 3 TCR γ chains paired with a very similar V δ 4D δ 2J δ 2 TCR δ chains (designated as the Group 4 sequences, Figure 2). In contrast, reported TCR sequences identified from spleen γδ T cells and small intestine epithelial γδ lymphocytes (IELs) (14–16) and IL-23R − γδ T cell populations in the spleen, lung, and colon lamina propria analyzed here (Table S1 in Supplementary Mate- rial) are highly diverse, using different V γ 's and V δ 's, with CDR3 regions consisting of both D δ 1 and D δ 2 gene segments in all three reading frames, and N regions in each of the gene-segment junc- tions. An analysis of CDR3 paratope convergence within IL-23R − , IL-23R + , and lIL-17F + γδ T cell populations is shown in Figure 3 . Along this line, it should be noted that the antigen-specific γδ T cells, including the inducible T γδ 17 cells, also utilize diverse TCRs (2, 3, 16). In this context, ˜1/3 of the IL-23R + or IL-17F + spleen γδ T cells, and ˜1/5 of IL-23R + lung γδ T cells express TCRs with different V γ V δ genes and diverse CDR3 regions. The spleen and lungs are continuously exposed to blood-borne or air-borne environmental antigens. It is likely that the TCR repertoire of IL- 1R + /IL-23R + γδ T cells reflects both the natural and the inducible T γδ 17 cells.

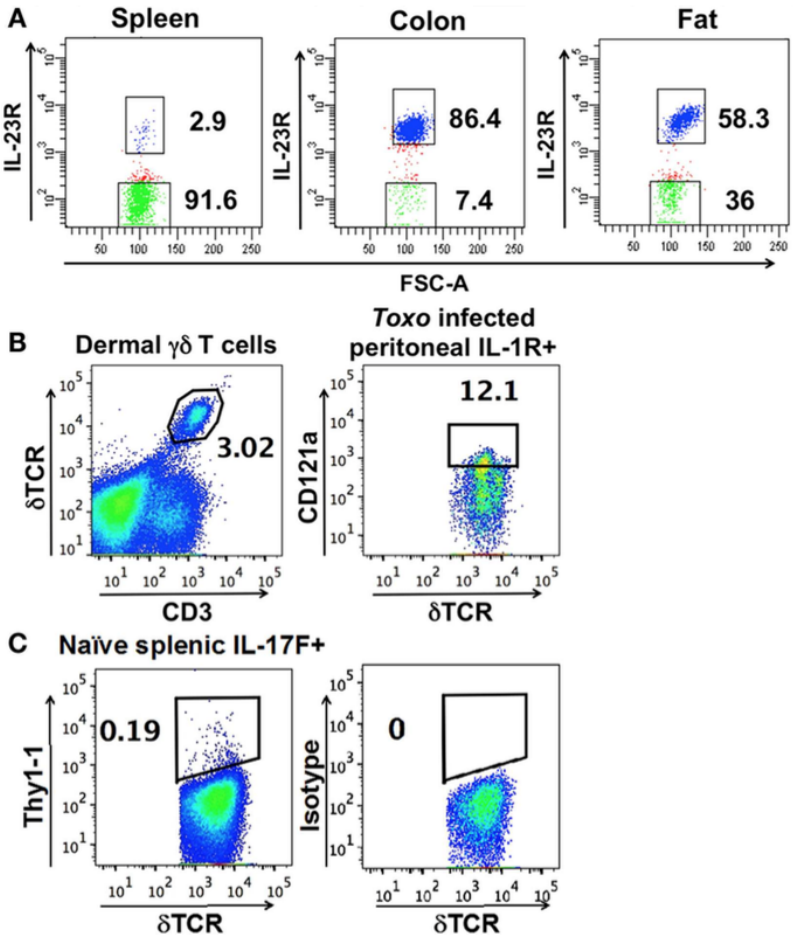
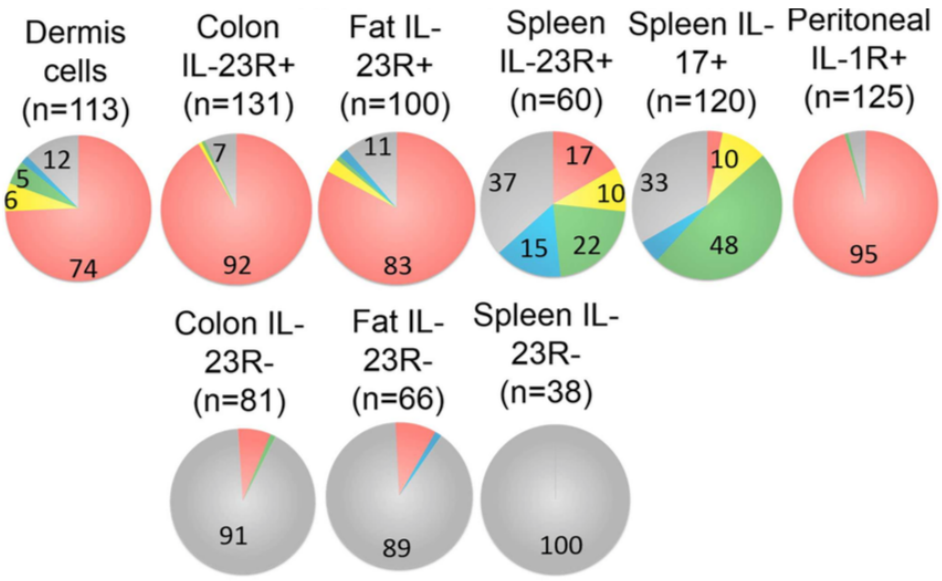


Figure 2.17: Representative FACS analysis and gates used to isolate(A) IL-23R+ (in blue) and IL-23R− γδ T cells (in green)(using FACSDiva) from IL-23R reporter mice. (B) Dermal cells, IL-1R+ γδ T cells from the peritoneum of C57BL/6 mice infected with T. gondii 5 h prior. (C) Thy1.1+ cells from the spleen of naïve IL-17F reporter mice. (B,C) are plotted using FlowJo. The number within each graph indicates the percentage of the designated population of cells out of the total γδ T cells.



	TCRγ		TCRδ	
Group 1	Vγ6Jγ1	CACWD-SSGFH	Vδ1Dδ2Jδ2	CGSD-IGG-SSWDTR
Group 2	Vγ4Jγ1	CSYG-(L)Y- SSGFH	Vδ5Dδ2Jδ1	CASGY- IGGIRA-TDKLV
Group 3	Vγ4Jγ1	CSYG-(X)Y-SSGFH	Vδ5Dδ2Jδ1	CASGY(X)-(I)GGIR(A)-(X)(T)DKLV
Group 4	Vγ4Jγ1	CSYG-(X)Y-SSGFH	Vδ6Dδ2Jδ1	CALM(ER)(X)-(I)GGIR(A)-(X)(T)DKLV
Group 5	None of the above, diverse TCRs		X indicates any amino acid	

Figure 2.18: Frequency of each major group of TCR sequences in IL-1R+/IL23R+ γδTcellpopulations.SpleenIL-17+,IL-23R+,IL-23R− γδT cells,lung,fat,andcolonlaminapropriaIL-23R+ andIL-23R− γδTcells, peritoneumIL-1R+ γδTcells5hafterintraperitoneumToxoplasmagondii infection and skin dermal γδ T cells were analyzed. Each cell population is represented by one pie chart. Each section of the pie chart represents one group of TCR sequences, color-coded as described. n, total number of analyzed sequences. The number within each section of the pie chart indicates the percentage of a given group of TCR sequences in the total number of analyzed sequences of that cell population (Table S1 in SupplementaryMaterial).Allexperimentswereperformedtwoindependent times,exceptttheanalysisofspleenIL-23R+ andIL-23R− γδTcells,which wereisolatedandanalyzedonce.TCRsequencesfromtwoindependent isolations and analyses are very similar and the combined results are shown. In two independent experiments, 58% and 82% of the total colon γδ T cells are IL23R+ ; 74% and 86% of total fat γδ T cells are IL-23R+ ; 0.1% and 0.2% of spleen cells are IL-17F+ ; 2.9% of spleen γδ T cells are IL23R+ . In the peritoneum 5 h after infection, 12 and 30% of the γδ T cells are IL-1R+.

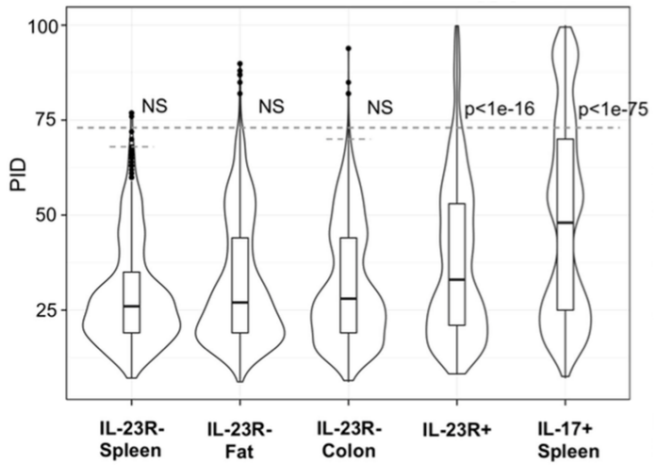


Figure 2.19: Analysis of CDR3 paratope convergence of all unique sequencesofIL-23R− γδTcellsfromthespleen,fat,colonsamples, IL-17F+ γδTcellsfromspleenandIL-23R+ γδTcells(combinedfromall anatomical sites). Percent identity among aligned γ and δ CDR3 amino acid sequences for all pairwise comparisons within each group are represented in violin/Box plot. Significance assessed by the Mann–Whitney–Wilcoxon test with Bonferroni's correction for multiple testing given a = 0.01 set to $p < 0.001$ to be considered significance. Dotted line indicates average 99th percentile percent identity for the IL-23R− T cell populations (68% ID for spleen, 73% ID for fat, 70% ID for colon).

Despite the fact that a substantial number of IL-1R + /IL-23R + γδ T cells and dermal γδ T cells express TCRs with similar V γ 4J γ 1 rearrangement (CSYG-(X)Y-SSGFHK), V γ 4 + TCR γ chain sequences are not utilized exclusively by this set of T cells. In fact, ˜50% of the IL-23R − cells also expressed TCRs with V γ 4, and more than half of these V γ 4 sequences were also expressed in IL-23R + cell populations (Figure 4).

2.3.3 Discussion

Our analysis showed that regardless of their anatomical location, IL-1R + /IL-23R + γδ T cells express a highly focused antigen recep- tor repertoire. While all major groups of TCR sequences expressed by these cells result from rearrangements with exonuclease diges- tion and P nucleotide addition (17), only Group 3 and 4 TCR

sequences have N nucleotides at the CDR3 γ and δ junctions. The N nucleotides are generated at the terminal of the combining gene segments by terminal transferase (TdT) in a template-independent manner. In mice, TdT is not expressed in developing thymocytes until 4–5 days after birth (18). Thus, γδ T cells that express Group 1 and 2 sequences are most likely generated during the fetal and/or neonatal stages. Indeed, Group 1 TCR has also been described for hybridomas derived from fetal and newborn γδ thymocytes (19) and is also present at the mucosal sites (20–22). Our observation that Group 3, 4 TCR expressing IL-1R + /IL-23R + γδ T cells are prevalent in the spleen and present in the lung and skin is con- sistent with the observation that adult precursor cells contribute to the nT γδ 17 cell pool and that these cells express V γ 4 + TCR γ chains (23–25).

Group 1 TCR sequences have been described for γδ T cell hybridomas generated from lung epithelium (26), from expanded γδ T cells after Listeria monocytogenes and Bacillus subtilis infection and in models of autoimmune inflammation (27–29). In addi- tion, the rapid appearance of V γ 6 and/or V δ 1 T γδ 17 cells has been reported in various infection systems: E. coli (i.p.) (30, 31), L. monocytogenes (i.p. oral) (32, 33) and Staphylococcus aureus (i.p.) (34). V γ 6 + and V γ 4 + dermal γδ T cells making IL-17 in response to imiquimod applied topically to induce skin inflammation has also been reported (24, 25). Separated TCR γ and δ chains of Group 4 sequences were identified from CFA-induced IL-17 making γδ T cells (35, 36). Taken together, our repertoire analysis confirms and advances previous studies of TCR usage of nT γδ 17 cells by defining the precise TCR sequences of these cells and observing how constrained they are. These observations suggest that antigen encountering is important for establishing their functional attrib- utes, a finding consistent with observations that signaling through the TCR is essential for this process (2, 8, 9).

It is unclear what nT γδ 17 cells recognize. However, the iden- tification of their TCR sequences is an important step forward in characterizing the antigens of these cells. In this context, O'Brien, Born and their colleagues demonstrated that a multimeric stain- ing reagent of soluble TCR expressing the Group 1 sequences can bind L cells, NIH 3T3 cells, a keratinocyte cell line XB-2, as well as freshly isolated macrophages from naïve mice and from mice infected with Listeria (37, 38).

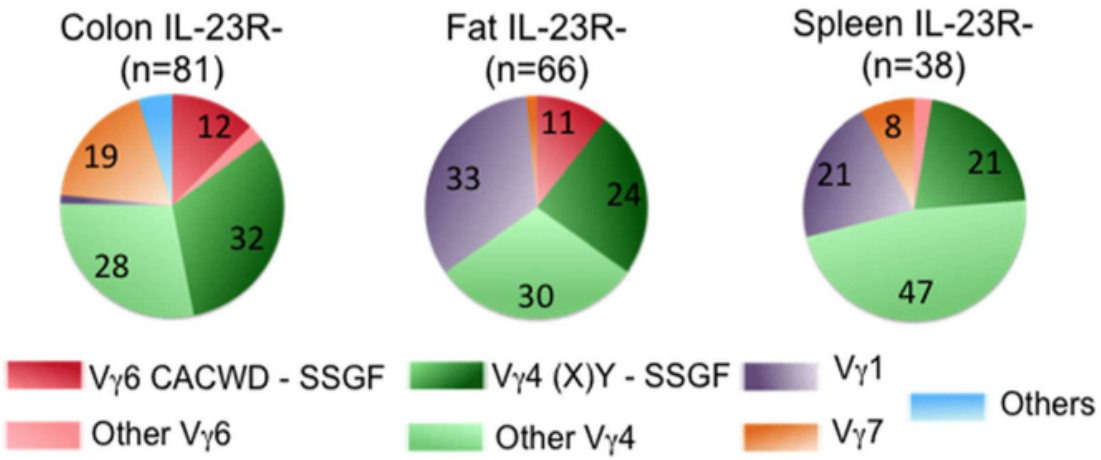


Figure 2.20: FrequencyofVγchainusageinIL-23R− cellpopulation.Each pie chart represents each cell population. n, total number of analyzed sequences. Number with each section of the pie chart, the percentage of each group of sequences color-coded as indicated in Figure 2.

While nT γδ 17 responses are well documented in the mouse, it is unclear whether or not a human counterpart exists. In this regard, human and murine γδ TCR gene sequences are very dif- ferent. Thus, it is unlikely that one would find human γδ TCRs that show the sequence equivalent of the TCRs described for the murine nT γδ 17 cells. However, one of the defining characteris- tics of adaptive immune recognition is that the antigen specificity, but not the particular antigen-specific receptor sequences, is con- served through evolution. The recognition of lysozyme by specific murine, human, and camel antibodies as well as by the adap- tive immune receptors of sea lamprey (39), and the recognition of the algae protein phycoerythrin (PE) by specific human and murine γδ TCRs (2) are such examples. Thus, differences in the TCR gene sequences among different species should not preclude the presence of nT γδ 17 cells.

It should be noted that the focused antigen receptor reper- toire described here is based on the analysis of pairs of TCR γ and δ chains, consisting of V gene segments as well as CDR3 regions. While the majority of these γδ T cells expressed V γ 6 or V γ 4, not all V γ 6 and V γ 4 expressing cells belong to this group of nT γδ 17 cells.

These observations underscore the need for cau- tion in categorizing γδ T cell function solely according to V gene usage. The approach of determining TCR sequences from a sin- gle cell directly ex vivo , as outlined here, should facilitate future analysis of the contributions of γδ T cells to a range of immune responses.

2.3.4 Methods

MICE

C57BL/6 mice were purchased from Jackson Laboratories and housed in the Stanford Animal Facility for at least 1 week before use. IL-17F Thy1.1/Thy1.1 mice (10) were bred and housed in the pathogen-free Stanford Animal Facility. IL-23R EGFP mice (11) were bred and housed in the pathogen-free Merck Research Labo- ratories, Palo Alto Animal Facility. All experiments were performed in accordance with the Institutional Biosafety Committee and the Institutional Animal Care and Use Committee.

ANTIBODIES AND CELL ISOLATION

Antibodies were purchased from either eBioscience or BD Bio- sciences unless otherwise stated. All analyses and sorting were performed on a BD Aria or Falstaff sorter. γδ T cells were enriched from mouse splenocytes or peritoneal cells by negative depletion as described (2).

To isolate Thy1.1 positive spleen γδ T cells from IL- 17F Thy1.1/Thy1.1 reporter mice, enriched γδ T cells were stained with PE-GL3, Pacific Blue-CD3e, PerCPCy5.5-Thy1.1, PerCP/Cy5.5 Mouse IgG1, κ Isotype Ctrl (OX-7 and its isotype control; BioLegend), LIVE/DEAD Aqua, APC-Cy7 conjugated anti-TCR β , CD19, CD11b, CD11c, F4/80, TER-119. APC-Cy7 and Aqua pos- itive cells are excluded from analysis. Peritoneal IL-1R positive γδ T cells were isolated from C57BL/6J mice i.p. infected with 1000 tachyzoites of Type II Me49 strain of Toxoplasma gondii 5 h prior. To isolate IL-1R (CD121a) positive cells, enriched γδ T cells were stained with PE-GL3 (pan anti- γδ TCR), PE-Cy7-CD3e (145- 2C11), APC-CD121a (JAMA-147; BioLegend), LIVE/DEAD Aqua, and APC-Cy7 conjugated anti-TCR β H57-597), CD19 (1D3), CD11b (M1/70), CD11c (N418), F4/80 (BM8), TER-119 (TER- 119).

APC-Cy7 and Aqua positive cells are excluded from analysis. Dermal split-thickness skin was obtained from C57BL/6J mice ears. Dermal sheets were prepared by incubation of split-thickness skin with 0.25% trypsin for 16 h at 4°C, and subsequent removal of the epidermis. Dermal sheets were digested with 2.5 mg/ml colla- genase and 0.3 mg/ml hyaluronidase for 45 min at 37°C to release dermal cells. Dermal cells were stained with PE-GL3, APC-Cy7- CD3e antibodies and Live/Dead Aqua. GL3 and CD3e positive dermis γδ T cells were isolated with FACS.

Two- to four-month-old female IL-23R EGFP +/ − mice were used for the isolation of IL-23R + and IL23R − γδ T cells. Five mice were combined for each type of tissue preparation. Visceral fat was directly minced in 4 mg/ml collagenase II (Worthington), 5% FBS in RPMI followed by shaking for 45 min at 37°C. Cells were further purified with 36% Percoll gradient (GE Healthcare) in PBS and spun at 2000 rpm for 5 min at room temperature. The floating layer and Percoll layer were aspirated and the resulting cell pellet was suspended in PBS, counted, and stained for flow cytometry. Colons were cleaned and washed in PBS and minced into 1 cm segments and placed into 0.5 mM EDTA in PBS. After shaking for 20 min at 37°C, the intraepithelial cell rich supernatant was discarded. Colon fragments were washed with PBS, then further minced to pieces <0.25 cm 3 in size in digestion buffer [PBS + 10% FCS + 1 mg/ml collagenase D (Sigma) + 2000 U/ml DNase I (Sigma) + Dispase (Corning, dilute 1:100)], and incubated with shaking for 20 min at 37°C. Cells were further purified with percoll gradient as described for isolating cells from fat. Isolated cells were stained with FcBlock, CD3 Percp-Cy5.5, TCR δ APC (Clone GL3), TCR β APC-Cy7, CD4- PE, CD8 α PE-Cy7, Live/Dead Aqua. IL-23R GFP + and IL-23R GFP − γδ T cells were single sorted into the wells of a 96-well plate using a FACsAria II (BD Biosciences).

BARCODE-ENABLED HIGH THROUGHPUT SINGLE-CELL TCR DETERMINATION

Single T cells are sorted into 96-well PCR plates and sequencing is performed as described (12), except murine γδ TCR specific primers are used for this study. γδ TCR primer sequences and the sequencing reaction are described in detail in Supplemental Methods in Supplementary Material. Briefly, an RT-PCR reac- tion

is carried out with TCR primers. The products are then used in a second PCR reaction, with nested primers for TCR genes. A third reaction is then performed that incorporates individual barcodes into each well. The products are combined, purified, and sequenced using the Illumina MiSeq platform. The result- ing paired-end sequencing reads are assembled and de-convoluted using barcode identifiers at both ends of each sequence by a cus- tom software pipeline to separate reads from every well in every plate. The resulting sequences are analyzed using VDJFasta (13), which we have adapted to resolve barcodes and analyze sequences with a customized gene-segment database. The CDR3 nucleotide sequences are then extracted and translated. Barcode design is shown in Figure S1 in Supplementary Material and TCR sequencing primer sequences are shown in Table S2 in Supplementary Material.

2.3.5 Acknowledgements

This work was made possible and largely my the co-authors, including primary author Yu-Ling Wei, as well as Arnold Han, Fengqin Fang, Luis Alejandro Zuniga, Jacob S. Lee, Daniel J. Cua and Yueh-hsiu Chien. Additional, we thank the National Institutes of Health for grant support (YC).

2.3.6 References

1. Chien YH, Meyer C, Bonneville M. γδ T cells: first line of defense and beyond. Annu Rev Immunol (2014) 32 :121–55. doi:10.1146/annurev-immunol-032713- 120216

2. Zeng X, Wei YL, Huang J, Newell EW, Yu H, Kidd BA, et al. γδ T cells recognize a microbial encoded B cell antigen to initiate a rapid antigen-specific interleukin-17 response. Immunity (2012) 37 (3):524–34. doi:10.1016/j.immuni.2012.06.011

3. Zeng X, Meyer C, Huang J, Newell EW, Kidd BA, Wei YL, et al. Gamma delta T cells recognize haptens and mount a hapten-specific response. eLIFE (2014) 3 :e03609. doi:10.7554/eLife.03609

4. Sutton CE, Lalor SJ, Sweeney CM, Brereton CF, Lavelle EC, Mills KH. Interleukin- 1 and IL-23 induce innate IL-17 production from gammadelta T cells,

ampli- fying Th17 responses and autoimmunity. Immunity (2009) 31 (2):331–41. doi:10.1016/j.immuni.2009.08.001

5. Kapsenberg ML. Gammadelta T cell receptors without a job. Immunity (2009) 31 (2):181–3. doi:10.1016/j.immuni.2009.08.004

6. Flanagan WM, Corthésy B, Bram RJ, Crabtree GR. Nuclear association of a T- cell transcription factor blocked by FK-506 and cyclosporin A. Nature (1991) 352 (6338):803–7. doi:10.1038/352803a0

7. Swat W, Xavier R, Mizoguchi A, Mizoguchi E, Fredericks J, Fujikawa K, et al. Essential role for Vav1 in activation, but not development of gammadelta T cells. Int Immunol (2003) 15 (2):215–21. doi:10.1093/intimm/dxg021

8. Duan J, Chung H, Troy E, Kasper DL. Microbial colonization drives expansion of IL-1 receptor 1-expressing and IL-17-producing gamma/delta T cells. Cell Host Microbe (2010) 7 (2):140–50. doi:10.1016/j.chom.2010.01.005

9. Wencker M, Turchinovich G, Di Marco Barros R, Deban L, Jandke A, Cope A, et al. Innate-like T cells straddle innate and adaptive immunity by altering antigen-receptor responsiveness. Nat Immunol (2014) 15 (1):80–7. doi:10.1038/ ni.2773

10. Lee YK, Turner H, Maynard CL, Oliver JR, Chen D, Elson CO, et al. Late devel- opmental plasticity in the T helper 17 lineage. Immunity (2009) 30 (1):92–107. doi:10.1016/j.immuni.2008.11.005

11. Awasthi A, Riol-Blanco L, Jäger A, Korn T, Pot C, Galileos G, et al. IL-23 receptor gfp reporter mice reveal distinct populations of IL-17-producing cells. J Immunol (2009) 182 (10):5904–8. doi:10.4049/jimmunol.0900732

12. Han A, Glanville J, Hansmann L, Davis MM. Linking T-cell receptor sequence to functional phenotype at the single-cell level. Nat Biotechnol (2014) 32 :684–92. doi:10.1038/nbt.2938

13. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proc Natl Acad Sci U S A (2009) 106 (48):20216–21. doi:10.1073/pnas.0909775106

14. Elliott JF, Rock EP, Patten PA, Davis MM, Chien YH. The adult T-cell receptor delta-chain is diverse and distinct from that of fetal thymocytes. Nature (1988) 331 (6157):627–31. doi:10.1038/331627a0

15. Lacy MJ, McNeil LK, Roth ME, Kranz DM. T-cell receptor delta-chain diversity in peripheral lymphocytes. Proc Natl Acad Sci U S A (1989) 86 (3):1023–6. doi:10.1073/pnas.86.3.1023

16. Shin S, El-Diwany R, Schaffert S, Adams EJ, Garcia KC, Pereira P, et al. Anti- gen recognition determinants of gammadelta T cell receptors. Science (2005) 308 (5719):252–5. doi:10.1126/science.1106480

17. Lafaille JJ, DeCloux A, Bonneville M, Takagaki Y, Tonegawa S. Junctional sequences of T cell receptor gamma delta genes: implications for gamma delta T cell lineages and for a novel intermediate of V-(D)-J joining. Cell (1989) 59 (5):859–70. doi:10.1016/0092-8674(89)90609-0

18. Bogue M, Gilfillan S, Benoist C, Mathis D. Regulation of N-region diversity in antigen receptors through thymocyte differentiation and thymus ontogeny. Proc Natl Acad Sci U S A (1992) 89 (22):11011–5. doi:10.1073/pnas.89.22.11011

19. Heilig JS, Tonegawa S. Diversity of murine gamma genes and expression in fetal and adult T lymphocytes. Nature (1986) 322 (6082):836–40. doi:10.1038/322836a0 20. Itohara S, Farr AG, Lafaille JJ, Bonneville M, Takagaki Y, Haas W, et al. Hom- ing of a gamma delta thymocyte subset with homogeneous T-cell receptors to mucosal epithelia. Nature (1990) 343 (6260):754–7. doi:10.1038/343754a0

21. Nandi D, Allison JP. Phenotypic analysis and γδ T cell receptor repertoire of murine T cells associated with vaginal epithelium. J Irnrnunol (1991) 147 :1773.

22. Heyborne KD, Cranfill RL, Carding SR, Born WK, O'Brien RL. Characterization of gamma delta T lymphocytes at the maternal-fetal interface. J Immunol (1992) 149 (9):2872–8.

23. Narayan K, Sylvia KE, Malhotra N, Yin CC, Martens G, Vallerskog T, et al. Immunological genome project consortium. Intrathymic programming of effector fates in three molecularly distinct γδ T cell subtypes. Nat Immunol (2012) 13 (5):511–8. doi:10.1038/ni.2247

24. Gray EE, Ramírez-Valle F, Xu Y, Wu S, Wu Z, Karjalainen KE, et al. Deficiency in IL-17-committed V γ 4(+) γδ T cells in a spontaneous Sox13-mutant CD45.1(+) congenic mouse substrain provides protection from dermatitis. Nat Immunol (2013) 14 (6):584–92. doi:10.1038/ni.2585 25. Cai Y, Xue F, Fleming C, Yang J, Ding C, Ma Y, et al. Differential develop- mental requirement and peripheral regulation for dermal V γ 4 and V γ 6 T17 cells in health and inflammation. Nat Commun (2014) 5 :3986. doi:10.1038/ ncomms4986

26. Roark CL, Aydintug MK, Lewis J, Yin X, Lahn M, Hahn YS, et al. Subset-specific, uniform activation among V gamma 6/V delta 1+ gamma delta T cells elicited by inflammation. J Leukoc Biol (2004) 75 (1):68–75. doi:10.1189/jlb.0703326

27. Mukasa A, Lahn M, Pflum EK, Born W, O'Brien RL. Evidence that the same gamma delta T cells respond during infection-induced and autoimmune inflammation. J Immunol (1997) 159 (12):5787–94.

28. Simonian PL, Roark CL, Diaz del Valle F, Palmer BE, Douglas IS, Ikuta K, et al. Regulatory role of gammadelta T cells in the recruitment of CD4+ and CD8+ T cells to lung and subsequent pulmonary fibrosis. J Immunol (2006) 177 (7):4436–43. doi:10.4049/jimmunol.177.7.4436

29. Mukasa A, Lahn M, Pflum EK, Born W, O'Brien RL. Evidence that the same gamma delta T cells respond during infection-induced and autoimmune inflammation. J Immunol (1997) 159 (12):5787–94.

30. Tagawa T, Nishimura H, Yajima T, Hara H, Kishihara K, Matsuzaki G, et al. Vdelta1+ gammadelta T cells producing CC chemokines may bridge a gap between neutrophils and macrophages in innate immunity during Escherichia coli infection in mice. J Immunol (2004) 173 (8):5156–64. doi:10. 4049/jimmunol.173.8.5156

31. Shibata K, Yamada H, Hara H, Kishihara K, Yoshikai Y. Resident Vdelta1+ gam- madelta T cells control early infiltration of neutrophils after Escherichia coli infection via IL-17 production. J Immunol (2007) 178 (7):4466–72. doi:10.4049/ jimmunol.178.7.4466

32. Sheridan BS, Romagnoli PA, Pham QM, Fu HH, Alonzo F III, Schubert WD, et al. γδ T cells exhibit multifunctional and protective memory in intestinal tissues. Immunity (2013) 39 (1):184–95. doi:10.1016/j.immuni.2013. 06.015

33. Hamada S, Umemura M, Shiono T, Tanaka K, Yahagi A, Begum MD, et al. IL-17A produced by gamma delta T cells plays a critical role in innate immunity against Listeria monocytogenes infection in the liver. J Immunol (2008) 181 (5):3456–63. doi:10.4049/jimmunol.181.5.3456

34. Murphy AG, O'Keeffe KM, Lalor SJ, Maher BM, Mills KH, McLoughlin RM. Staphylococcus aureus infection of mice expands a population of memory γδ T cells that are protective against subsequent infection. J Immunol (2014) 192 (8):3697–708. doi:10.4049/jimmunol.1303420

35. Roark CL, French JD, Taylor MA, Bendele AM, Born WK, O'Brien RL. Exacerbation of collagen-induced arthritis by oligoclonal, IL-17-producing gamma delta T cells. J Immunol (2007) 179 (8):5576–83. doi:10.4049/jimmunol. 179.8.5576

36. Roark CL, Huang Y, Jin N, Aydintug MK, Casper T, Sun D, et al. A canon- ical V γ 4V δ 4+ γδ T cell population with distinct stimulation requirements which promotes the Th17 response. Immunol Res (2013) 55 (1–3):217–30. doi:10.1007/s12026-012-8364-9

37. Aydintug MK, Roark CL, Yin X, Wands JM, Born WK, O'Brien RL. Detection of cell surface ligands for the gamma delta TCR using soluble TCRs. J Immunol (2004) 172 (7):4167–75. doi:10.4049/jimmunol.172.7.4167

38. Aydintug MK, Roark CL, Chain JL, Born WK, O'Brien RL. Macrophages express multiple ligands for gammadelta TCRs. Mol Immunol (2008) 45 (11):3253–63. doi:10.1016/j.molimm.2008.02.031

39. Deng L, Velikovsky CA, Xu G, Iyer LM, Tasumi S, Kerzic MC, et al. A structural basis for antigen recognition by the T cell-like lymphocytes of sea lamprey. Proc Natl Acad Sci U S A (2010) 107 (30):13408–13. doi:10.1073/pnas.1005475107

2.3.7 Copyright

2.4 Reading specificity in the B-cell receptor repertoire

IGHV polymorphism provides a rich source of humoral immune system diversity. One important example is the IGHV1-69 germline gene where the biased use of alleles that encode the critical CDR- H2 Phe54 (F-alleles) to make broadly neutralizing antibodies (HV1-69-sBnAb) to the influenza A hemagglutinin stem domain has been clearly established. However, whether IGHV1-69 polymorphism can also modulate B cell function and Ab repertoire expression through promoter and copy number (CN) variations has not been reported, nor has whether IGHV1-69 allelic distribution is impacted by ethnicity. Here we studied a cohort of NIH H5N1 vaccinees and demonstrate for the first time the influence of IGHV1-69 polymorphism on V-segment usage, somatic hypermutation and B cell expansion that elucidates the dominance of F-alleles in HV1-69-sBnAbs. We provide evidence that Phe54/Leu54 (F/L) polymorphism correlates with shifted repertoire usage of other IGHV germline genes. In addition, we analyzed ethnically diverse individuals within the 1000 genomes project and discovered marked variations in F- and L- genotypes and CN among the various ethnic groups that may impact HV1- 69-sBnAb responses. These results have immediate implications for understanding HV1-69-sBnAb responses at the individual and population level and for the design and implementation of "universal" influenza vaccine.

2.4.1 Introduction

Neutralizing antibody (nAb) responses to influenza infection and vaccination are highly variable among individ- uals throughout the population. This observation can in part be explained by differences associated with health status, exposure history, age and host variability of immune response genes1. Since protection is also correlated with nAb titers, any role that immunoglobulin (IG) germline gene polymorphism may play in this variability is important to establish, but has been difficult to investigate due to the use of numerous V, D and J genes in the genesis of immunoglobulins and

the enormous combinatorial diversity that results from the pairing of rearranged VH and VL genes. However, the discovery of biased usage of the IG heavy chain variable (IGHV) germline gene IGHV1-69 in anti-hemagglutinin stem-directed broadly neutralizing Abs (HV1-69-sBnAbs) and the finding that only the heavy chain makes contact with hydrophobic HA stem2 has provided a unique opportunity to define the molecular features of anti-influenza BnAbs and simplify immunogenetic studies to understand the contribution of allelic variation at the IGHV1-69 locus to the anti-influenza sBnAb response.

IGHV1-69 is one of the most polymorphic loci within the human IGHV gene cluster (14q32.33), exhibit- ing both allelic and copy number (CN) variation3,4. There are 14 alleles known to be associated with this gene that can be differentiated by the presence of either a phenylalanine (F) or leucine (L) at amino acid position 54 (Kabat numbering) within the apex of the CDR-H2 loop. Historically, this classification refers to the 51p1-like and hv1263-like allelic groups, respectively (Supplementary Fig. 1a). In addition to coding polymorphisms, the number of IGHV1-69 germline copies per diploid human genome can vary from 2–4 (Supplementary Fig. 1b)3,5,6, and there are 4 IGHV1-69 haplotypes with gene duplications in an earlier established American cohort5 (Supplementary Fig. 1c).

The relevance of F/L polymorphism to HV1-69-sBnAbs is the fact that almost all of these Abs originate from the IGHV1-69 F-allelic group. The conserved CDR-H2 Phe54 is a major anchor residue making direct contact with HA, and the replacement of Phe54 by Ala54 or Leu54 (L) has been shown to dramatically reduce binding affinities7,8. Importantly, in this study and in two recent studies9,10 the F/L polymorphism is shown to correlate with the frequencies of HV1-69-sBnAbs, being highest in individuals carrying F-alleles. In contrast, the predom- inant usage of the L-allele group in generation of non-neutralizing anti-gp41 Abs was recently demonstrated in a HIV-1 vaccination study11. These recent findings highlight the need to better understand how this genetic var- iability at the IGHV1-69 locus can modulate B cell repertoires as well as the extent to which this polymorphism varies across diverse human populations3,5,6,12. To address these two questions we analyzed Ab repertoires from an NIH H5N1 vaccinee cohort and samples from the 1000 Genomes Project

(1KG)13, respectively. We report the new finding that the two allele families have markedly different effects on Ab repertoire expression that is in part explained by CN variation but there are also differences in B cell expansion and somatic hypermutation. In addition, we discovered marked variance in IGHV1-69 gene duplication and CN among the different ethnic populations that will affect HV1-69-sBnAb responses to influenza vaccines and natural infections.

2.4.2 Results

Comparison of antibody responses to H5N1 vaccine among three IGHV1-69 genotypic groups. Individuals from a 2007 H5N1 vaccination trial were genotyped and phenotyped for IGHV1-69 CDR-H2 Phe54 F/L polymorphism (rs55891010; see Fig. 1a and methods). Their one month post-vaccination sera was competed against the anti-stem sBnAb F10 for binding to the pandemic H1CA0709 HA, which was not circulating when the serum samples were collected. Figure 1b shows a statistically significant difference in F10 blocking activity among the groups and was highest for the F/F group, followed in decreasing order by the F/L and L/L groups. The microneutralization titers (MN) for the F/F group were 1.67 and 2.29 fold higher than the mean values for F/L and L/L groups, respectively with a similar trend in their median values (Supplementary Fig. 2a). The post-vaccination hemagglutination inhibition titers (HAI) and the ELISA titers for H1CA0709 and H1CA0709 HA proteins were shown to not significantly differ from one another among the three IGHV1-69 genotypic groups (Supplementary Fig. 2b,d,e). In addition, when HAI and MN titers were compared within indi- viduals, there was also a trend toward lower HAI/MN ratios for the F/F and F/L groups compared to the L/L individuals (Supplementary Fig. 2c). Supplementary Fig. 3 shows that stem binding activity originally boosted by H5VN04 vaccination was generally maintained within each genotypic group over the 4-year period. The similar trends observed in the analysis of the F10 competition studies, MN titers, and HAI/MN ratios supports the con- cept that IGHV1-69 germline polymorphism has an effect on the profile of the HA-directed Ab response, with expression from F-alleles leading to a higher Ab response to the stem domain.

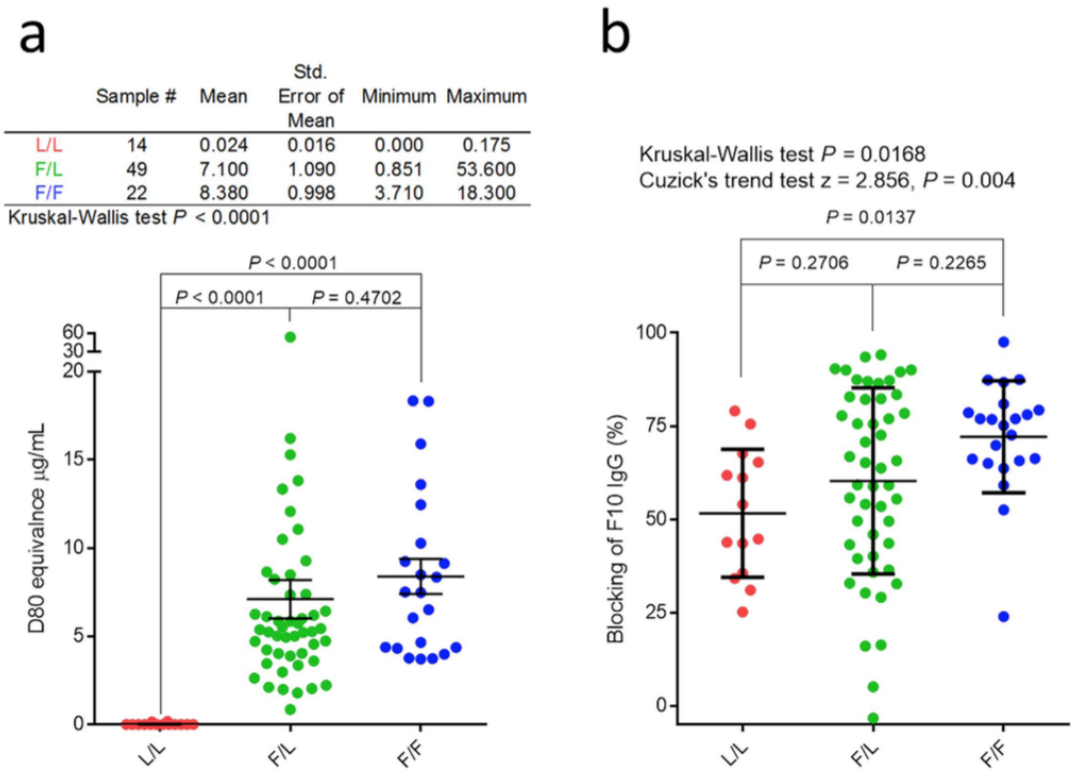


Figure 2.21: Correlation between IGHV1-69 polymorphism and Ab response to the H5 vaccine. The pre- vaccinated sera of the 85 individuals were diluted 1/1250 and analyzed for binding activities against the anti- IGHV1-69 idiotype mAb G6. Binding activities were normalized by subtracting the G6 MSD signal with the MSD signal obtained from an isotype control, and by using a standard curve made with the IGHV1-69 F-allele- based IgG Ab D8035. (b) Post-vaccination sera (diluted 1/125) were competed with the anti-stem Ab F10 IgG for binding to H1CA0709. Cuzick's trend test was used to further confirm that the occurrence of F-alleles increases the ability of serum to block F10 binding (L/L = 0, F/L = 1, F/F = 2). Error bars represent standard error of mean.

Effect on IGHV1-69 polymorphism on germline gene utilization and expressed HV1-69-sBnAb repertoires. To assess the role of IGH locus polymorphism on expressed IGHV1-69 germline gene reper- toires ≥5 × 106 PBMCs (circa 10% B cells) were analyzed from the blood samples of 18 individuals (F/F = 4, F/L = 11, L/L = 3), collected 4 years following the H5N1 vaccine trial. The IGHV-gene frequencies from inde- pendent V(D)J rearrangements were rendered non-redundant, and IgM and IgG class determinations were made by analyzing the PCR products obtained from reverse priming with IG constant region primers. Figure 2 shows that in both the unmutated IgM (naïve) and all IgG (memory) V-segment datasets, IGHV1-69 usage was at the highest frequency in the F/F group (7.7% IgM, 3.9% IgG), intermediate frequency in the F/L group (4.7% IgM, 3% IgG), and the lowest frequency in the L/L group (1.8% IgM, 1.4% IgG). The significance of the ~3-fold difference in IGHV1-69 usage between the F/F and L/L groups was further demonstrated by noting that, in the F/F group, IGHV1-69 was the 4th and 7th most frequently used IGHV germline gene in the unmutated IgM and IgG datasets, respectively, whereas in the L/L group IGHV1-69 was ranked 18th and 23rd (data not shown). This variation in IGHV1-69 germline gene utilization was also seen for putative HV1-69-sBnAbs with the highest frequencies and correlation coefficients in individuals with F/F alleles and across the IgM B cell subset (Supplementary Fig. 4a–d). We have been able to further delineate some of these HV1-69-sBnAbs signatures through functional analyses (Supplementary Fig. 5 and text). These results demonstrate that F-allele individuals have higher levels of circulat- ing IGHV1-69 Ab and HV1-69-sBnAb repertoires than L-allele individuals.

Differential effects of IGHV1-69 genotype on B cell expansion, somatic hypermutation (SHM) and evolution to HV1-69-sBnAb clones. We next investigated if other B cell functions were affected by IGHV1-69 genotype. Analysis of the naive and memory IGHV1-69 datasets within each individual's repertoire revealed additional variation in clonal expansion, SHM frequency, and IgG-to-IgM ratios among each geno- typic group. For example, the frequency of highly expanded IGHV1-69 clones (frequency > 1e-4) was greater for L/L than the F/L or F/F genotypic groups (Supplementary Fig. 6a). However, the clones of the F/F group, of which there were fewer highly

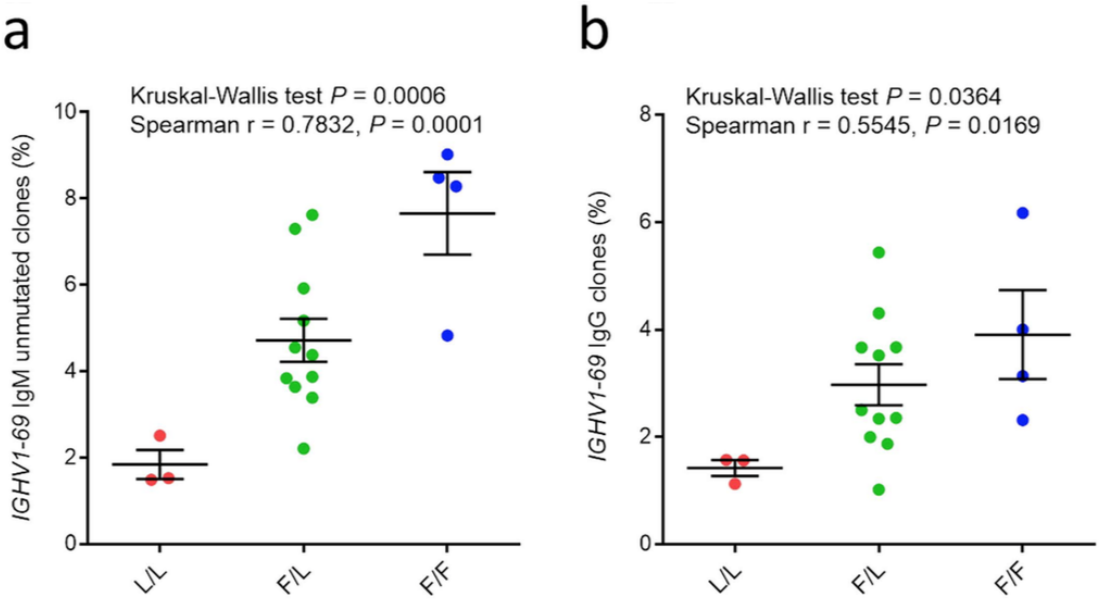


Figure 2.22: Analyzing IGHV1-69 V-segment gene utilization among the three IGHV1-69 genotypic groups. (a) The frequency of IGHV1-69 IgM clones defined by unmutated V-segments (b) the frequency of IGHV1-69 IgG clones. Error bars represent standard error of mean.

expanded clones, were also significantly more mutated than those of the L/L group (Supplementary Fig. 6b). Additionally, we note that IGHV1-69 is unusual among V-genes in that these BCRs appear at a lower frequency in memory B-cells than in naïve B-cells (Supplementary Fig. 6c) (an approximately 40% reduction)14. Interestingly, this effect was strongest in individuals of the F/F genotype. These results suggest that the capacity of the IGHV1-69 B cells to undergo expansion, SHM and Ig class switching may be different among the genotypic groups.

An expanded dataset of 57 published HV1-69-sBnAbs2,9,15 was also used to investigate the effects of allele var- iation on SHM and VDJ recombination that results in the signature CDR-H3 amino acids G95, P96 and Y99 $\pm$ 1 (Supplementary Fig. 6d). The effects of IGHV1-69 allelic variation revealed that transition from the germline L54 to the critical F54 in L/L individuals through SHM was a rare event (Supplementary Fig. 6e), as was the occurrence of HV1-69-sBnAb CDR-H3 signatures in the

IgG dataset (Supplementary Fig. 6f). The higher fre- quencies of V-segment amino acid substitutions at positions that are significantly enriched in HV1-69-sBnAbs in the L/L group (Supplementary Fig. 6g) suggests that these Abs may be evolving to compensate for the lack of Phe549. Collectively, this analysis implies that the scarcity of HV1-69-sBnAb in L-allele individuals was due to the underutilization of this allelic type by the immune system.

IGHV1-69 Copy Number and Regulatory Region SNPs. We further studied the potential correla- tion between IGHV1-69 usage and F-allele copy number (CN)12. We found a significant positive correlation between both unmutated IgM and IgG IGHV1-69 utilization and increasing IGHV1-69 F-allele copy number (Fig. 3) (Spearman, IgM r = 0.91, P $<$ 0.0001; IgG r = 0.75, P $<$ 0.0003). A strong positive correlation was also seen between CN and IgM but a weaker one for IgG HV1-69-sBnAb clonal frequencies suggesting that the IgG switch memory B cell subset is subject to additional regulation (Supplementary Fig. 4e,f). Interestingly, all L/L individuals were found to have a mean CN = 2 and they also had the lowest IGHV1-69 utilization among the three genotypic groups that include individuals whom lack gene duplication (Fig. 3 insets), suggesting that CNV only partially explains the lower IGHV1-69 utilization. For this reason we also investigated other genetic variants in strong linkage disequilibrium (LD) with IGHV1-69 alleles that could represent SNPs that may influ- ence the control of transcription or V-D-J recombination rates (e.g., variants in the 5´UTR and recombination signal sequences). SNPs within the vicinity of IGHV1-69 (+− 1.5 kb; GRCh37, chr14:107168431-107171928) in LD with the F/L variant (rs55891010) were identified using data from the 1 KG phase3 dataset for African (n = 661), Asian (n = 504), and European (n = 503) populations. Only four SNPs had an r2 $>$ 0.8 in at least one of the three populations (Supplementary Fig. 7a). Two of the identified SNPs represented additional coding variants within IGHV1-69, and the remaining two occurred upstream of the leader sequence ATG start codon. The SNP rs10220412 was found to reside in the 5´ UTR of IGHV1-69 and within a promoter initiator element16 which is also a binding motif of the B cell associated protein RUNX3, that has been shown to bind to this region in a lymphoblastoid cell line ChIP-seq dataset17 (Supplementary Fig. 7b). RUNX3 has been shown to be

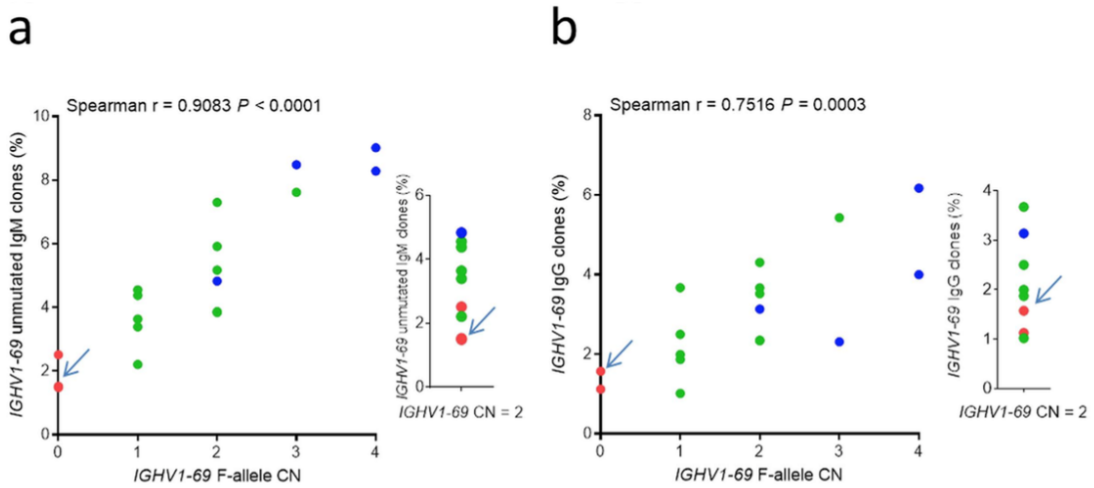


Figure 2.23: Correlating IGHV1-69 F-allele copy number with IGHV1-69 utilization. (a) Correlating F-allele CN with the frequency of IGHV1-69 IgM clones defined by unmutated V-segments. (b) Correlating F-allele CN with the frequency of IgG clones. The insets in both panels (a) and (b) describe IGHV1-69 clone frequency in individuals that lack IGHV1-69 gene duplication (Arrows point to overly of two L/L individuals).

elevated following EBV infection or activation by PMA of primary B cells and is proposed to have a role in B cell prolifer- ation18,19. These findings suggest that genetic factors beyond CN can influence IGHV1-69 transcript frequencies and that the rs10220412 SNP is a candidate that may affect Ab gene transcription in L-allele individuals by hin- dering the association of RUNX3 to this variant RUNX3/Inr site.

IGHV1-69 polymorphism has broad effects on the expressed IGHV repertoire. The underutiliza- tion of IGHV1-69 germline genes in L/L individuals led us to investigate whether the F/L polymorphism was also associated with different usage frequencies of other V-genes in naïve and memory repertoires. In Fig. 4a,b V-gene frequencies were averaged across individuals within each IGHV1-69 genotypic group using data from the unmu- tated IgM and all IgG V-segments, respectively, and aligned accord- ing to their relative position in the IGH locus on chr14. In addition to IGHV1-69, IGHV2-70 utilization was also significantly different in both the unmutated IgM and IgG datasets with repertoire frequencies being highest for the F/F and lowest for the L/L group (Spearman, IgM r = 0.64, P = 0.0046; IgG r = 0.57, P = 0.0131)

(Fig. 4c,d). This is likely explained by the fact that IGHV1-69 and IGHV2-70 reside on the same duplicated genomic segment of IGHV, and thus exhibit correlated increases in CN3 (Supplementary Fig. 1b). However, we also found evidence of more spatially separated associations between IGHV1-69 locus polymorphism and other IGHV genes. For example, in the IgG subset, IGHV4-30- 4/31 usage was shown to have a significant positive correlation with the occurrence of L-alleles ($r = -0.53$, $P = 0.0223$) (Insert Fig. 4c (compare red bars)). Although IGHV4-30-4/31 does not achieve significance in the IgM dataset, it was apparent that IGHV4-30-4/31 was part of a cluster of IGHV genes which include IGH4-30-2, IGHV3-30/33rn, IGHV4-28 and IGHV3-23 (Inset Fig. 4c,d), all of which were defined by weak to moderate negative correlation coefficients ($r = -0.17$ to -0.53) and exhibit the highest usage frequencies in L/L individ- uals. To further assess V-gene usage differences in L/L individuals, we compared V-gene repertoire frequencies between the L/L group and a combined F/L-F/F group using a t-test, and visualized these differences using heatmaps (Supplementary Fig. 8a). This analysis revealed that, in the unmutated IgM dataset, IGHV3-30/33rn and IGHV4-30-2 were consistently more highly expressed in the L/L group compared to the F/L-F/F group ($P < 0.05$), whereas IGHV1-24, IGHV1-69, IGHV2-70 and IGHV3-49 were significantly underrepresented in L/L individuals ($P < 0.05$). In the IgG subset, IGHV4-30-2 and IGHV4-30-4/31 were significantly overrepresented in the L/L group ($P < 0.05$), and again IGHV1-69 was significantly underrepresented ($P < 0.05$; Supplementary Fig. 8b). Taken together, we observe that multiple clusters of V-genes within the IGHV locus are positively or negatively correlated with IGHV1-69 genotype.

IGHV1-69 F/L polymorphism, copy number and gene duplication among different ethnic groups. To further investigate the association between F/L polymorphism and CN we examined published rs55891010 genotypes13 and CN3 in 288 samples from 3 broad ethnic groups (African, Asian, and European) of the 1 KG Project13. Consistent with observations from our H5N1 vaccinee cohort (Fig. 5a upper table), in the combined set of 1 KG samples we found a strong association between rs55891010 genotypes and CN, with higher mean IGHV1-69 CN in F/F (mean = 2.53) and F/L (mean = 2.46) individuals compared to individuals of the L/L genotype (mean =

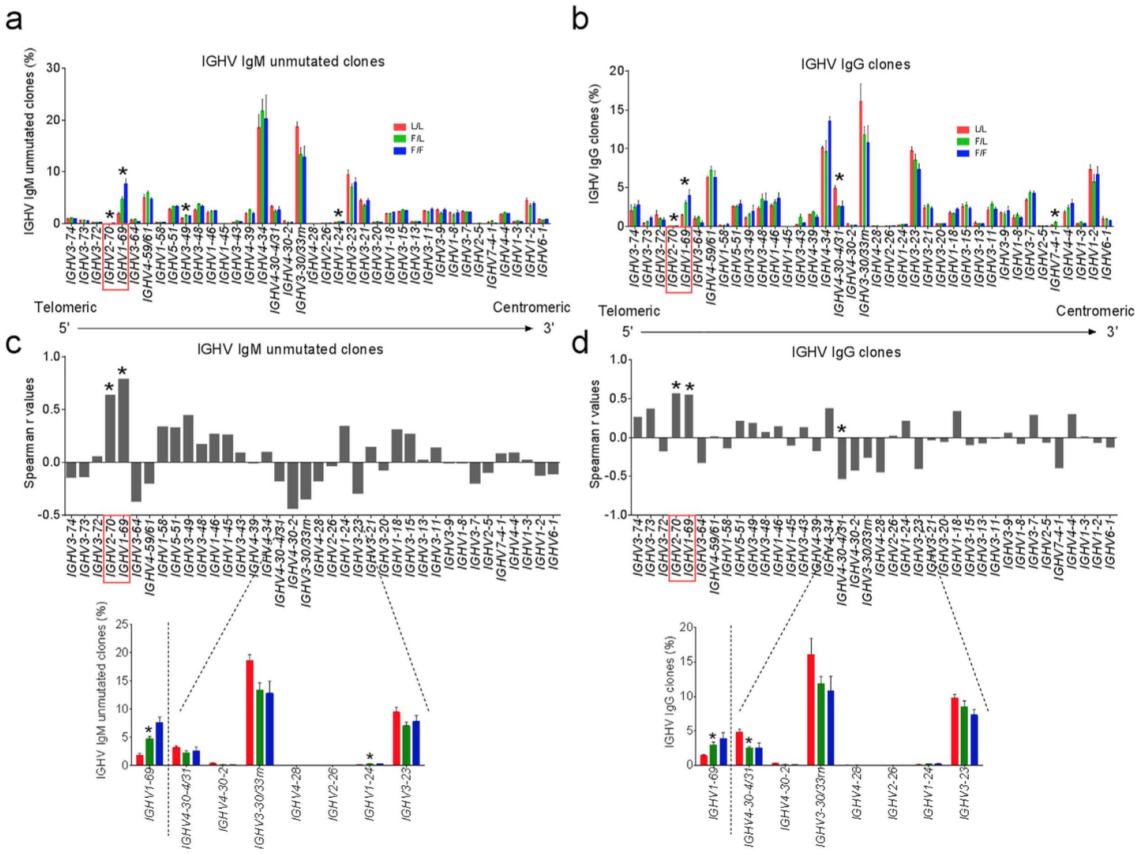


Figure 2.24: The antibody repertoire of the three IGHV1-69 genotypic groups. V-gene frequencies were averaged for the L/L group (n = 3), F/L group (n = 11), and F/F group (n = 4) from the datasets of IgM clones characterized by unmutated V-segments (a) and IgG clones (b). The majority of the functional V-genes were tabulated according to their respective positions in the IGH locus (further detailed in Supplementary Fig. 11). Asterisks denote V-genes utilized differently among the three genotypic groups as determined by Kruskal- Wallis test ($P < 0.05$). Error bars represent standard error of mean. In panels (c,d) Spearman correlation coefficients are derived for the data presented in panels (a,b) with L/L = 0, F/L = 1, and F/F = 2. Asterisks indicate statistically significant correlations ($P < 0.05$). Red rectangles point to the location of IGHV1-69 and IGHV2-70, for which their usages were significantly different among the three genotypic groups, being the highest in the F/F group and lowest in the L/L group, in both the unmutated IgM and IgG datasets. The inset panels are enlarged cropped sections from Panel (a,b) of the IGHV4-30-4/31-to-IGHV3-23 region that is negatively correlated with F-alleles.

2; Fig. 5a). We next partitioned these 1 KG samples by ethnicity, which revealed dramatic population differences in frequency of IGHV1-69 genotypes (Fig. 5a). A significant relationship between IGHV1- 69 genotype and CN was found in Europeans, with mean IGHV1-69 CN estimates of 2.55, 2.39, and 2, for F/F, F/L, and L/L genotypic classes, respectively. This relationship, however, was not clear in the Asian population, as none of the F/F samples in our analysis were found to have greater than 2 copies of IGHV1-69 (Fig. 5a,b). Additionally, in the African population, as noted previously3, CN was higher on average overall, but in contrast to Europeans, there was a much larger fraction of F/L individuals, and the majority of these samples were found to have 3 copies of IGHV1-69. The CN trends were also corroborated by a second IGHV1-69 gene duplication assay (Supplementary Fig. 9a,b). The African group was also defined by a marked low frequency of L/L individuals (Fig. 5a,b).

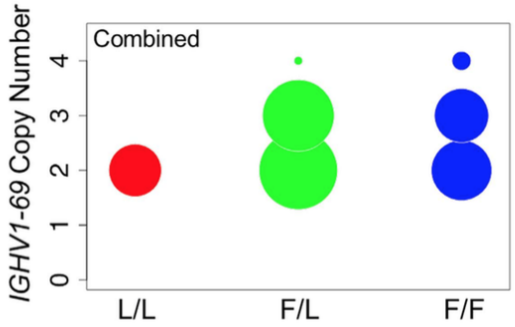
Next we expanded our analysis of the IGHV1-69 rs55891010 polymorphism to include all samples of the 1 KG cohort (Supplementary Fig. 10). We found that the frequency of the L/L genotype varied considerably across human populations, with the lowest frequencies occurring in samples of African ancestry, and the highest in South Asian populations; as expected, opposite trends were noted for the F/F genotype. Taken together, these analyses indicate that interrelationships among IGHV1-69 F/L genotype, CN and IGHV loci genomic architec- ture likely exhibit population-specific patterns that may have broad implications for mounting broadly protective HV1-69-sBnAb responses.

2.4.3 Discussion

There is growing evidence that IG polymorphism may have a critically important role in Ab responses20. Allelic variation exists in many IGHV, IGKV, and IGLV germline genes and the results reported herein demonstrate that these genetic differences at the IGHV1-69 locus can modulate the nAb response. In a cohort of NIH H5N1 vaccinees21,22, we found higher anti-stem competition and MN titers but not HAI titers in the F/F group. The stem competition results are in agreement with the reports by Pappas9 following seasonal influenza vaccination and Wheatley10 following

a

Ethnicity	Genotype L/L	F/L	F/F
Combined (n=288)	2 (n=38)	2.46 (n=155)	2.53 (n=95)
African (n=78)	2 (n=1)	2.87 (n=53)	3.04 (n=23)
Asian (n=85)	2 (n=13)	2.08 (n=47)	2 (n=25)
European (n=125)	2 (n=24)	2.39 (n=54)	2.55 (n=47)
NIH cohort (n=18)	2 (n=3)	2.73 (n=11)	3.25 (n=4)



b

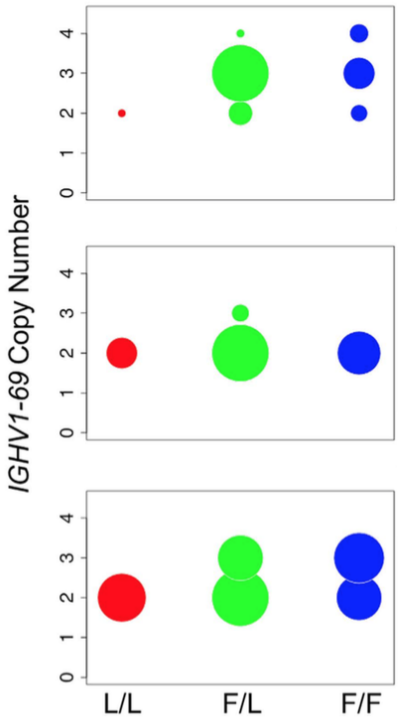


Figure 2.25: IGHV1-69 F/L polymorphism and CN variations among various ethnicities. (a) Table including mean IGHV1-69 copy number estimates after partitioning by IGHV1-69 rs55891010 genotype, provided for the total combined population, for three broad ethnic groups and the NIH cohort samples that were analyzed by NGS. Bubble plots corresponding to IGHV1- 69 CN for each genotypic class in the total combined population (a, lower) and individual ethnic groups (b). In each plot, the area of a given circle is proportional to the number of individuals observed for that particular combination of IGHV1-69 CN and rs55891010 genotype relative to the number of samples analyzed in each group (e.g., Combined, African, Asian, or European).

a H5 DNA vaccine. Thus, the strong correlation between IGHV1-69 polymorphism and HV1-69-sBnAb responses provided support for establishing a deeper understanding of the immunobiology of IGHV1-69 B cells. Likewise, it is important to gain knowledge on the extent to which the distribution of F/F, F/L and L/L individuals vary among the population. Here we demonstrated for the first time the influence of IGHV1-69 polymorphism on V-segment usage, SHM and B cell expansion that explain the dominance of F-alleles in HV1- 69-sBnAbs. We also provided evidence that F/L polymorphism is associated with shifted repertoire usage of other IGHV genes. Marked variations in F- and L- genotypes and IGHV1-69 CN are also demonstrated among various ethnic groups.

Ab repertoire analysis revealed that IGHV1-69 utilization increased from L/L to F/L to F/F individuals in both the naive and memory subsets. The correlation observed between F-allele CN and IGHV1-69 gene usage only partially explained this trend. However the higher IGHV1-69 gene usage in F-allele individuals whom lack gene duplication over L/L individuals suggested that IGHV1-69 polymorphism in non-coding regions also had an effect on germline gene expression. We found a strong LD between the F/L polymorphism and SNP rs10220412, located in the 5´ UTR of IGHV1-69, within the promoter initiator element16 and an annotated binding motif of the B cell associated protein RUNX317. We suggest that the non-consensus mutation in this RUNX3 binding motif found in L/L individuals may affect Ab transcription and/or B cell proliferation18,19. Additional analyses aimed at investigating the role of SNP rs10220412 will be essential to fully understand the observed relationship between the IGHV1-69 polymorphism, germline gene utilization and circulating Ab repertoires. Other genetic factors that were not identified in our studies may also be involved.

Our results also show that both IGHV1-69 F- and L-allele B cells share similar frequency of unmutated IgM clones defined by CDR-H3 HV1-69-sBnAb signatures, however only F-allele B-cells show these elevated signa- tures in the memory compartment while L-allele B-cells are recessive in their ability to readily resolve HA neutralization in this manner. Indeed, estimates can be made to the frequency of HV1-69-sBnAb precursor B-cells (Supplementary Fig. 6). While 1.8% of the repertoire of

an F/F genotype individual and 0.6% of the repertoire of an F/L individual will have an appropriate progenitor CDR-H3 signature rearranged on a IGHV1-69-F54 V-gene, an L/L genotype repertoire will have roughly one in ten thousand (0.01%) B-cells configured with the minimal signatures that appear necessary for HV1-69-sBnAbs generation. These results strongly suggest that L/L individ- uals must reconcile sB-nAb responses through the use of other VH germline genes.

The overall variance in IGHV1-69 utilization among the three genotypic groups was associated with differ- ential usage of other IGHV germline genes. Interestingly, our analyses point to a cluster of IGHV genes circa IGHV3-30 (Fig. 4: IGHV4-30-4/31, IGH4-30-2, IGHV3-30/33rn, IGHV4-28 and IGHV3-23) that are positively correlated with the presence of L-alleles. Although long-range haplotypes spanning the length of the IGHV locus have not been characterized, one possible explanation for this correlation is that L/L individuals could have a dis- tinct IGHV locus architecture that promotes increased VDJ recombination events that preferentially utilize IGHV germline genes within the region of IGHV3-30. This could result in shifting of germline gene usage within the Ab repertoire. This positive correlation with L-alleles is of particular interest as there are several reports of sBnAbs that are based on the IGHV3-30 V-gene23–25. Further studies might show that L/L individuals carry additional copies of the IGHV3-30 and IGHV3-23 genes, both of which are known to vary in copy number3,20. Alternatively, it is also plausible that IGHV1-69 duplication haplotypes contain architectural features that reduce the accessibil- ity and therefore utilization of genes in the IGHV3-30 region.

Further characterization of CN frequencies in the context of ethnic background using 1 KG samples fur- ther supports our observation that gene duplication events ($CN > 2$) rarely occur in L/L individuals (Fig. 5, Supplementary Fig. 9). However, remarkable population-specific diversity was observed in CN frequencies in the F/L and F/F genotypes. The majority of the African individuals bear IGHV1-69 gene duplication while in the Asian groups gene duplication hardly occurs. Indeed, nearly every CNV studied in IGHV to date has been shown to exhibit population-specific patterns3,26,27. This remarkable population-specific diversity is also demonstrated with the frequencies of the F/L genotypes. Indeed, the frequency of L/L individuals

is miniscule in the African group (~2%), while to the other extreme L/L individuals comprise 41.3% of the South Asian population (Fig. 5, Supplementary Fig. 10). These results imply that individuals of different ethnicities may vary in their capacity to elicit HV1-69-sBnAbs by natural infection or vaccination.

In conclusion, our studies of IGHV1-69 polymorphism have provided an important entry point to further understand the impact of IGH polymorphism on host nAb responses. We have discovered several important biological associations that had not been previously described. We also uncovered a marked association between IGHV1-69 polymorphism and ethnicity. The combined genotypic, molecular and phenotypic analyses as pre- sented in this study offer a new approach to understand vaccine responsiveness at the individual and population level. Indeed, IGHV1-69 genotypes were shown to correlate with anti-HA stem titers and circulating HV1- 69-sBnAb repertoires. Clearly defined associations between B cell genotype and phenotype are likely to emerge from these studies, and together with serologic analyses of pre-28,29 and post- exposure Ab responses should lead to important advances in our knowledge of B cell responses to influenza. We propose that the establishment of a complete human IGHV haplotype map is needed to catalogue and map genomic variation in the IGHV locus from larger cohorts of individuals of different ethnicities. This advance will naturally lead to the development of complementary high-throughput genotyping tools that may prove useful for predicting vaccine responsiveness at the individual and population levels. This progress will be particularly important for the development and moni- toring of next generation "universal" influenza vaccines30–34.

2.4.4 Methods

Ethics statement. All the experiments were performed in accordance with the approved guidelines and reg- ulations of the Institutional Review Boards (IRBs). Specifically, the samples from H5N1 vaccinee cohort at NIH/ NIAID were collected under NIH IRB approved Study # 06-I-0235 (Clinical Trials.gov identifier: NCT00383071), titled "A Phase II Vaccine Dose Finding Pilot Study for Development of an Anti-Influenza A (H5Nl) Intravenous Hyper-Immune Globulin". Informed consent was

obtained from all participants. Experiment protocols involving human samples were approved by DFCI IRB and performed under the regulation of the DFCI Legacy # 11–093.

Details of the H5 vaccination cohort. In the H5N1 vaccination study (Clinical Trials.gov identifier: NCT00383071), the effect of varying amounts of H5 protein and various numbers of vaccinations was tested using the (rgA/H5N1 Vietnam/1203/04 X A/PR/8/34) manufactured by Sanofi Pasteur Inc, Swiftwater, PA. The study concluded that additional vaccinations increased HAI and MN titers, but not increasing vaccine dose. The Kruskal-Wallis test indicated that three groups do not differ in the number of vaccinations ($P = 0.37$), therefore variance in antibody response to the H5N1 vaccine among the three genotypic groups (F/F, F/L, and L/L) is not predicted to be the result of differences number of vaccinations.

Competing sera with F10 for binding to H1CA0709. MSD 384-well standard plate coated overnight with 6.25 ng of H1CA0709 (A/California/07/2009, Influenza Reagent Resource (IRR) FR-559)) were washed 1 time with PBS and blocked for 1 h at 37 °C with 2% BSA/PBS. After blocking, sera (pre or post) diluted 1/125 in 2% milk/PBST was added to the plate for 45 min at 37 °C, after which 62.5 ng/mL of F1035 labeled with sulfo-tag was added for an additional 45 min at 37 °C. Following washes with PBST, read buffer was added and the plate was read using a Sector Imager 2400 instrument. The threshold of 100% inhibition was derived from the wells processed with the dilution buffer, and the threshold of 0% inhibition was obtained from binding of F10 to wells containing dilution buffer.

MSD ELISA assay for analyzing binding of post-vaccination sera H1CA0709 and H1CA0709 HA1. MSD 384-well high bind plates coated overnight with 25 ng of H1CA0709 and H1CA0709 HA1 (IRR FR-695) were washed 1 time with PBS and blocked for 1 h at 37 °C with 2% BSA/PBS. After blocking, post-vaccination serum samples diluted 1/7500 for H1CA0709 and 1/1500 for H1CA0709 HA1 in 2% milk/ PBST were added to the plates for 45 min at 37 °C. Following washes with PBST Goat-anti-human IgG sulfo-tag antibody diluted in 2% milk/PBST was added for 45 min at 37 °C. Following washes with PBST, read buffer was added and the plate was read using a Sector Imager 2400 instrument.

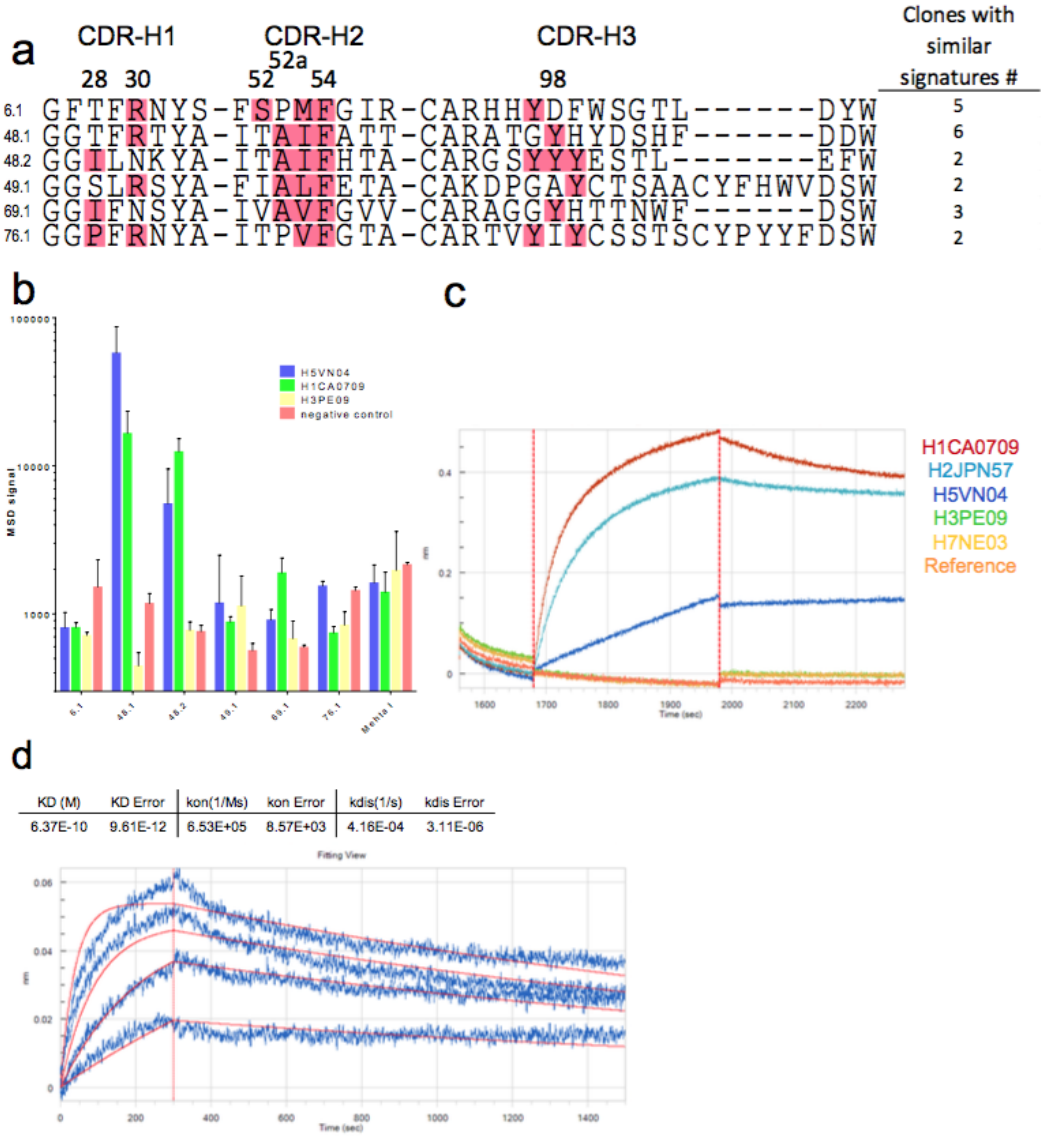


Figure 2.26: The binding activities of synthesized HV1-69-sBnAb precursor clones to hemagglutinins. (a) Six clones characterized by notable HV1-69-sBnAb signatures (maroon residues) were cloned into phagemid light chain shuffle libraries (kappa/lambda). For each clone, also shown is the number of clonally related or duplicated clones (b) The 1:1 mixed lambda and kappa phagemid libraries were selected against H5VN04 and the rescued bulk phagemid libraries were analyzed for binding activities against H5VN04, H1CA0709, H3PE09, and against a negative control protein, using 1e13 phagemid particles/mL. Mehta I was used as a control phagemid library. (c) Binding kinetic profile of clone 48.1 as scFv-Fc against 10 μg/ml of H1CA0709, H2JPN57, H5VN04, H3PE09, and H7NE03. (d) Detailed binding kinetics of clone 48.1 against H2JPN57.

Next generation sequencing. RNA was extracted from cryopreserved PBMCs by using RiboPure Kit. cDNA was generated by using Superscript RT II kit with 100–600 ng of RNA and oligo-dT primer. VH libraries were generated from cDNA template using Taq polymerase with ThermoPol buffer for multiplex PCR. Primers used for PCR were an equimolar mix of eight forward primers and two reverse primers previously described in36. Reactions were carried out using 0.2μM both forward and reverse primer mixes and 2μL cDNA in 50μL total volume. Cycling conditions were as follows: 92 °C denaturation for 3 min, 92 °C for 1 min, 50 °C for 1 m, 72 °C for 1m for 4 cycles, 92°C for 1min, 55°C for 1m, 72°C for 1m for 4cycles, 92°C for 1min, 63°C for 1m, 72°C for 1m for 20 cycles, 72 °C for 7 min, hold at 4 °C. PCR reactions performed in replicates were pooled and purified by spin column using a QiaQuick PCR Purification Kit. The cDNA products were ligated with indexed Illumina adapters using Thruplex DNAseq. The adapter ligated libraries were quantified by qPCR (Kapa Biosystems cat# KK4824), pooled in equal quantities, and then sequenced using MiSeq 250 bp paired-end reads by the Dana-Farber Cancer Institute Molecular Biology Core Facilities. All kits were used according to the manufacturer's protocols.

Genotyping assays. Individuals of the H5N1 vaccination trial were genotyped by real time PCR approach and by ELISA assay with the anti-IGHV1-69 idiotype mouse mAb G6, which does not bind to IGHV1-69 Abs displaying CDR-H2 Leu5437. This combined genotyping/phenotyping approach enabled us to classify with confi- dence 85 of the 86 individuals as follows: F/F (G6 +) = F/L (G6 +) = 49, and 14 L/L (G6−).

Real time PCR genotyping approach. Two allelic-group-specific TaqMan (Applied Biosystems) probes were designed to distinguish codon variants of IGHV1-69, encoding either the CDR-H2 Leu54 (L-alleles) or Phe54 (F-alleles) alleles, allowing the estimation of the number of copies of each allele in each sample (F-alleles, Forward = TGGACAAGGGCTTGAGTGGAT; Reverse = CCCTGGAACTTCTGTGCGTAGT; Reporter Sequence = CCCTATCTTTGGTACAGC. L-alleles, forward = TGGACAAGGGCTTGAGTGGAT; Reverse = C CCTGGAACTTCTGTGCGTAGT; Reporter Sequence = CCTATCCTTGGTATAGCA, all are 5´-3´). For

twenty individuals which donated blood 4 years post vaccination, polymorphonuclear cells were isolated by a stand- ard dextran approach38 from which genomic DNA was extracted using QIAGEN's DNeasy Blood & Tissue Kit. TaqMan PCR assays were performed at least two times with at least three replicates using Applied Biosystems 7300 Real-Time PCR System, and for each sample allelic CN counts were estimated using the delta-delta-CT method. The other 66 individuals were genotyped with the TaqMan probes by using PCR amplified IGHV1-69 gene product obtained from residual circulating DNA found in the serum. Circulated DNA was isolated from 100–200µl of pre-vaccination, 1 month post-vaccination, and 4 years post-vaccination sera by using ZR-96 Quick-gDNA Blood kit by ZYMO research, and the IGHV1-69 gene was amplified using QIAGEN's HotStarTaq Plus Master Mix Kit, with 3ul of the circulated DNA eluent, and forward primer that anneals in the V-segment intron domain and a reverse primer that anneals in the V-segment RSS domain. Cycling conditions were as fol- lows: 95°C min, [94°C 1min, 50°C 1min, 72°C 1min]×50, 72°C 10min, 4°C 16h. Real time PCR analysis was performed by diluting the PCR products 1-to-1e6 and the TaqMan probes described above were used in order to determine IGHV1-69 allele composition. Samples were analyzed using CFX384 instrument by Biorad, and Cq and END RFU thresholds for F/F, F/L, and L/L were generated from the sera samples of the 20 individuals which were genotyped using genomic DNA. For samples that did not amplify, a second PCR was performed or residual DNA was isolated again. For confidence purposes both the pre and post-vaccination sera were analyzed, and in situations where there was no agreement residual DNA was isolated again from the pre- and post-vaccination samples and the real time PCR was repeated.

MSD ELISA assay for analyzing binding of pre-vaccination sera to G6. MSD 384-well high bind plates coated overnight with 6.25 ng of G6 and with isotype control mAb 1D4, were washed 1 time with PBS and blocked for 1 h at 37 °C with 2% BSA/PBS. After blocking pre-vaccination sera diluted 1/1250 in 2% milk/PBST as well as serial dilutions of the IGHV1-69 F54 allele-based IgG Ab D8035 were added to the plate for a period of 1 h at 37 °C. After washing the plates with PBST, Goat-anti-human-IgG sulfo-tag antibody diluted in 2%milk PBST was added for 1 h at 37

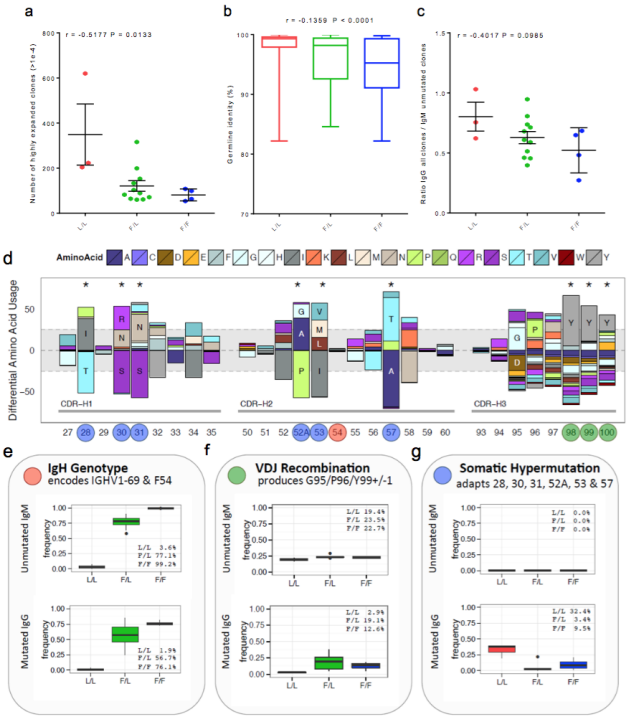


Figure 2.27: Variances in IGHV1-69 clonal expansion, IgG to IgM ratio, and expansion in-situ evolution of HV1-69-sBnAb precursors. The three genotypic groups were analyzed for the number of highly expanded clones (frequency of >1e-4) (a); their percent germline identity (b); and the ratio of IgG to IgM clones defined by unmutated V-segments (c). Spearman correlation coefficients (r) were used to summarize the association. Error bars represent standard error of mean; (d-f) in order to deconstruct the in-situ evolution of HV1-69-sBnAb precursors, the positional amino acid variability of 57 HV1-69- sBnAbs were compared to the total IGHV1-69 repertoires of F/F individuals. (d) Numerous amino acids in the rearranged VH segment are over-represented in the HV1-69-sBnAbs. Positions under significant selection are indicated by asterisks. HV1-69-sBnAb precursor signature variation can be classified into genotype derived allele variation, VDJ recombination- derived CDR-H3 variation, and somatic hypermutation-derived positional arming (highlighted as red (e), green (f) and blue (g), respectively). (e) IGH genotype dictates the abundance of total IGHV1-69 and IGHV1-69+F54 BCRs in the repertoire. In both unmutated IgM and class- switched IgG, F54 IGHV1-69 receptors are abundant in F-allele bearing individuals and nearly absent in L/L homozygotes. (f) Generation of HV1-69-sBnAbs CDR3 signatures, dominated by Tyr98-100, a preference for Gly95 and Pro96, and an aversion to Asp95 though VDJ recombination. In the unmutated IgM repertoire, all genotypes have similar frequencies of progenitor HV1-69-sBnAbs CDR-H3 signatures, but only remain elevated in class-switched IgG memory in individuals with F/L or F/F Phe54 genotypes, while loss of HV1-69-sBnAbs CDR-H3 in memory is L-recessive. (g) V-segment somatic hypermutation favors six specific positional amino acid changes in HV1-69-sBnAbs. These substitutions are absent in the unmutated IgM, but are expanded in IgG memory. L-allele homozygotes exhibit elevated SHM at these sites.

°C. Following washes with PBST, read buffer was added and the plate was read using Sector Imager 2400 instrument. G6 binding activities were normalized by subtracting the G6 MSD signal from the MSD signal obtained for 1D4, and by using a standard curve generated from D80 binding activities to G6.

Analysis of IGHV1-69 SNP and copy number/duplication data for samples derived from the 1000 Genomes Project. To investigate potential relationships between genotypic variation at the IGHV1- 69 F/L SNP variant (rs55891010) and gene copy number, as well as IGHV1-69 regulatory polymorphisms, we used available CNV and SNP data from two previous studies3,13. CNV genotype calls were previously estimated for 425 individuals from a broad set of human populations based on standard PCR and targeted TaqMan qPCR assays unique to sequence characterized within the IGHV1-69 duplication haplotype (see3 and Supplementary Fig. 9a). In total, 288 samples had both CNV and SNP data for comparisons across three broad ethnic groups (African, n = 78; East Asian, n = 85; European, n = 125) represented by 7 subpopulations (Yoruba in Ibadan, Nigeria; Luhya in Webuye, Kenya; Japanese in Tokyo, Japan; Han Chinese in Bejing, China; Toscani in Italy; British in England and Scotland; Finnish in Finland). Additional analyses of allele and genotype frequency dif- ferences at rs55891010 in a larger sample of human populations were conducted using genotypes downloaded from the 1 KG project for 5 broad ethnic groups (African, n = 661; East Asian, n = 504; South Asian, n = 489; American, n = 347; European, n = 503).

Antibody repertoire sequence analysis. Sequences were demultiplexed by Illumina DNA barcodes, converted from fastq to fasta and paired-end assembled. Barcode and primer sequences in the V-segment were then removed from all reads. Sequences were then processed using the AbGenesis VDJFasta pipeline to iden- tify V,D,J segments, isotype, and translated CDRs, as previously described39. Samples were rendered clonally non-redundant to include one unique representative of each V + J + CDR-H3 sequence as described in14. Somatic hypermutation burden on each clone was determined by number of non-templated base mismatches to the iden- tified V-gene and J-gene, using the closest allele in IMGT4.

Expression of HV1-69-sBnAb precursor clones. The selected six VH genes were synthesized by Genewiz (Plainfield, NJ) with a 5´ SfiI and a 3´ BspEI restriction sites. The VH genes were cloned into two pFarber phagemid scFv display kappa and lambda light chain shuffle libraries and were transformed into electrocompe- tent TG-1 cells resulting in 1E6-to-1E7 transformants. The phagemid particles were rescued by standard approach utilizing VCSM13 helper phage and phagemid particles were purified by peg-precipitation. Phagemid preps were panned from 1:1 mixed lambda and kappa libraries for H5VN04 (protein sciences (Meriden, CT) by adding 5E12 of phagemid particles. Clone 48.1 was also successfully expressed as scFv-Fc with F10's light chain using the pcDNA3.4 vector which was transfected into 293T cells using standard polyethylenimine approach.

MSD assay for bulk phagemids. MSD 384-well high bind plate coated overnight with 25 ng H1CA0709, 25 ng H3PE09 (H3 A/Perth/16/2009, IRR FR-472), 25 ng H5VN04 (H5 A/Vietnam/1203/2004, IRR FR-39), and 25 ng human PDL1 (irrelevant control protein), were washed 1 time with PBS and blocked for 1 h at 37 °C with 2% BSA/PBS. After blocking, peg precipitated bulk phagemid libraries from first round of selection were diluted in 2% milk PBST (1E13 phagemid particles/mL) were added to the wells and plates were incubated at 37 °C for 1 hr. Plates were washed with PBST three times, then sulfo-tagged anti M13 diluted to in 2% milk PBST was added and the plate was incubated for 1 hr at 37 °C and washed again as above. Read buffer was added and plates were read on an MSD Sector Imager.

Binding kinetic assays. Binding kinetics of 48.1-scFv-Fc was performed using Octet RED (Fortebio, USA) with anti-human Fc sensors. Kinetic cycle consisted of: 1) Equilibrium—sensors were dipped in wells containing kinetic buffer (0.02%/tween, 0.1%/BSA, PBS) for 60 sec. 2) Loading—sensors were dipped in wells containing supernatant of cells transfected with 48.1 (Supplementary Fig. 5c) or 0.05μg/ml of purified 48.1 scFv for 300 sec (Supplementary Fig. 5d). 3) Baseline—sensors were dipped into wells containing kinetic buffer for 180 sec. 4) Association—sensors were dipped into wells containing 10μg/ml HA proteins H1CA0709, H2JPN57 (H2 A/Japan/305/1957, IRR FR-700), H5VN04 (protein sciences), H3E09 and H7N3 (H7 A/Netherlands/219/2003, IRR FR-71) (Supplementary Fig. 5c) or serial dilutions

of H2JPN57 (Supplementary Fig. 5d) for 300 sec. 5) Dissociation—sensors were dipped into wells containing kinetic buffer for 300 sec (Supplementary Fig. 5c) or 1200 sec (Supplementary Fig. 5d). Association rate (Kon) dissociation rate (kdis) and equilibrium dissociation (KD) constants were calculated using 1:1 fitting model as provided by the Fortebio Data analysis software.

Statistical analysis. Overall comparisons between the F/F, F/L, and L/L genotype classes were performed by using the Kruskal-Wallis test and subsequent pairwise comparisons were conducted by applying Dunn's pro- cedure to control the overall Type I error rate. Cuzick's trend test was used to detect the trend over F/F, F/L, and L/L genotype classes by assigning scores 0, 1, 2 to the three groups; moreover, Spearman's correlation coefficient was used to summarize the association. Statistical analyses were performed by using Prism 6 (Graphpad Software, Inc.) and R software (www.r-project.org).

2.4.5 Acknowledgements

This work was made possible by my co-authors, including first author Yuval Avnir, as well as Corey T. Watson, Eric C. Peterson1, Aimee S. Tallarico, Andrew S. Bennett, Kun Qin, Ying Fu, Chiung-Yu Huang, John H. Beige, Felix Breden, Quan Zhu & Wayne A. Marasco. MAb G6 was kindly provided by Dr. Roy Jefferis, University of Birmingham, Medical School, Birmingham, United Kingdom. This work was supported by the following agencies. WAM – NIH AI074518, AI109223 and DARPA W911NF-10-0266. FB – Natural Sciences and Engineering Research Council of Canada. JHB—Intramural Research Programs of NIAID, National Institutes of Health; NCI, National Institutes of Health Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

2.4.6 References

1. Franco, L. M. et al. Integrative genomic analysis of the human immune response to influenza vaccination. eLife 2, e00299, doi: 10.7554/eLife.00299 (2013).

2. Avnir, Y. et al. Molecular signatures of hemagglutinin stem-directed heterosubtypic human neutralizing antibodies against influenza A viruses. PLoS pathogens 10, e1004103, doi: 10.1371/journal.ppat.1004103 (2014).

3. Watson, C. T. et al. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. American journal of human genetics 92, 530–546, doi: 10.1016/j. ajhg.2013.03.004 (2013).

4. Lefranc, M. P. Immunoglobulins: 25 years of immunoinformatics and IMGT-ONTOLOGY. Biomolecules 4, 1102–1139, doi: 10.3390/biom4041102 (2014).

5. Sasso, E. H., Willems van Dijk, K., Bull, A. P. & Milner, E. C. A fetally expressed immunoglobulin VH1 gene belongs to a complex set of alleles. The Journal of clinical investigation 91, 2358–2367, doi: 10.1172/JCI116468 (1993).

6. Milner, E. C., Hufnagle, W. O., Glas, A. M., Suzuki, I. & Alexander, C. Polymorphism and utilization of human VH Genes. Annals of the New York Academy of Sciences 764, 50–61 (1995).

7. Lingwood, D. et al. Structural and genetic basis for development of broadly neutralizing influenza antibodies. Nature 489, 566–570, doi: 10.1038/nature11371 (2012).

8. Throsby, M. et al. Heterosubtypic neutralizing monoclonal antibodies cross-protective against H5N1 and H1N1 recovered from human IgM + memory B cells. PloS one 3, e3942, doi: 10.1371/journal.pone.0003942 (2008).

9. Pappas, L. et al. Rapid development of broadly influenza neutralizing antibodies through redundant mutations. Nature 516, 418–422, doi: 10.1038/nature13764 (2014).

10. Wheatley, A. K. et al. H5N1 Vaccine-Elicited Memory B Cells Are Genetically Constrained by the IGHV Locus in the Recognition of a Neutralizing Epitope in the Hemagglutinin Stem. Journal of immunology 195, 602–610, doi: 10.4049/jimmunol.1402835 (2015).

11. Williams, W. B. et al. Diversion of HIV-1 vaccine-induced immunity by gp41-microbiota cross-reactive antibodies. Science , doi: 10.1126/science.aab1253 (2015).

12. Sasso, E. H., Johnson, T. & Kipps, T. J. Expression of the immunoglobulin VH gene 51p1 is proportional to its germline gene copy number. The Journal of clinical investigation 97, 2074–2080, doi: 10.1172/JCI118644 (1996).

13. Abecasis, G. R. et al. A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073, doi: 10.1038/ nature09534 (2010).

14. Glanville, J. et al. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. Proceedings of the National Academy of Sciences of the United States of America. 108, 20066–20071, doi: 10.1073/pnas.1107498108 (2011).

15. Whittle, J. R. et al. Flow cytometry reveals that H5N1 vaccination elicits cross-reactive stem-directed antibodies from multiple Ig heavy-chain lineages. Journal of virology. 88, 4047–4057, doi: 10.1128/JVI.03422-13 (2014).

16. Roy, A. L., Sen, R. & Roeder, R. G. Enhancer-promoter communication and transcriptional regulation of Igh. Trends in immunology 32, 532–539, doi: 10.1016/j.it.2011.06.012 (2011).

17. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome research 22, 1760–1774, doi: 10.1101/gr.135350.111 (2012).

18. Spender, L. C., Cornish, G. H., Sullivan, A. & Farrell, P. J. Expression of transcription factor AML-2 (RUNX3, CBF(alpha)-3) is induced by Epstein-Barr virus EBNA-2 and correlates with the B-cell activation phenotype. Journal of virology 76, 4919–4927 (2002).

19. Brady, G. & Farrell, P. J. RUNX3-mediated repression of RUNX1 in B cells. Journal of cellular physiology 221, 283–287, doi: 10.1002/ jcp.21880 (2009).

20. Watson, C. T. & Breden, F. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. Genes and immunity 13, 363–373, doi: 10.1038/gene.2012.12 (2012).

21. Sui, J. et al. Wide prevalence of heterosubtypic broadly neutralizing human anti-influenza A antibodies. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 52, 1003–1009, doi: 10.1093/cid/cir121 (2011).

22. Beigel, J. H., Voell, J., Huang, C. Y., Burbelo, P. D. & Lane, H. C. Safety and immunogenicity of multiple and higher doses of an inactivated influenza A/H5N1 vaccine. The Journal of infectious diseases. 200, 501–509, doi: 10.1086/599992 (2009).

23. Wyrzucki, A. et al. Alternative recognition of the conserved stem epitope in influenza A virus hemagglutinin by a VH3-30-encoded heterosubtypic antibody. Journal of virology. 88, 7083–7092, doi: 10.1128/JVI.00178-14 (2014).

24. Corti, D. et al. A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins. Science 333, 850–856, doi: 10.1126/science.1205669 (2011).

25. Nakamura, G. et al. An in vivo human-plasmablast enrichment technique allows rapid identification of therapeutic influenza A antibodies. Cell host & microbe 14, 93–103, doi: 10.1016/j.chom.2013.06.004 (2013).

26. Sasso, E. H., Buckner, J. H. & Suzuki, L. A. Ethnic differences in VH gene polymorphism. Annals of the New York Academy of Sciences 764, 72–73 (1995).

27. Shin, E. K. et al. Polymorphism of the human immunoglobulin variable region segment V1-4.1. Immunogenetics. 38, 304–306 (1993).

28. Andrews, S. F. et al. High preexisting serological antibody levels correlate with diversification of the influenza vaccine response. Journal of virology 89, 3308–3317, doi: 10.1128/JVI.02871-14 (2015).

29. Henry Dunand, C. J. et al. Preexisting human antibodies neutralize recently emerged H7N9 influenza strains. The Journal of clinical investigation 125, 1255–1268, doi: 10.1172/JCI74374 (2015).

30. Krammer, F. & Palese, P. Universal influenza virus vaccines: need for clinical trials. Nature immunology. 15, 3–5, doi: 10.1038/ni.2761 (2014).

31. Krammer, F., Pica, N., Hai, R., Margine, I. & Palese, P. Chimeric hemagglutinin influenza virus vaccine constructs elicit broadly protective stalk-specific antibodies. Journal of virology 87, 6542–6550, doi: 10.1128/JVI.00641-13 (2013).

32. Wohlbold, T. J. et al. Vaccination with soluble headless hemagglutinin protects mice from challenge with divergent influenza viruses. Vaccine 33, 3314–3321, doi: 10.1016/j.vaccine.2015.05.038 (2015).

33. Yassine, H. M. et al. Hemagglutinin-stem nanoparticles generate heterosubtypic influenza protection. Nature medicine 21, 1065–1070, doi: 10.1038/nm.3927 (2015).

34. Impagliazzo, A. et al. A stable trimeric influenza hemagglutinin stem as a broadly protective immunogen. Science , doi: 10.1126/ science.aac7263 (2015).

35. Sui, J. et al. Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses. Nature structural & molecular biology 16, 265–273, doi: 10.1038/nsmb.1566 (2009).

36. Ippolito, G. C. et al. Antibody repertoires in humanized NOD-scid-IL2Rgamma(null) mice and human B cells reveals human-like diversification and tolerance checkpoints in the mouse. PloS one 7, e35497, doi: 10.1371/journal.pone.0035497 (2012).

37. Potter, K. N., Li, Y., Mageed, R. A., Jefferis, R. & Capra, J. D. Molecular characterization of the VH1-specific variable region determinants recognized by anti-idiotypic monoclonal antibodies G6 and G8. Scandinavian journal of immunology 50, 14–20 (1999).

38. Clark, R. A. & Nauseef, W. M. In Current Protocols in Immunology (John Wiley & Sons, Inc., 2001).

39. Glanville, J. et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proceedings of the National Academy of Sciences of the United States of America 106, 20216–20221, doi: 10.1073/pnas.0909775106 (2009).

2.4.7 Copyright

polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. Scientific reports, 6.

2.5 Reading convergence of phenotypes

The mechanisms through which successful immunotherapy induces possible deletion, replacement, or reprogramming of T cells are unknown. By evaluating the expression of T-cell– related genes, and using appropriate multivariate statistical approaches, our data show that successful immunotherapy can induce previously unidentified CD4+ T-cell subtypes during treatment that could help to predict an "immune-tolerant" clinical phenotype identified after cessation of treatment. The ability to use "anergic" transcriptional phenotypes in single T cells to predict successful "immune tolerance" induction in the clinic setting, as suggested by our findings, could lead to transformative impacts in the field of immunotherapy.

Allergen immunotherapy can desensitize even subjects with poten- tially lethal allergies, but the changes induced in T cells that underpin successful immunotherapy remain poorly understood. In a cohort of peanut-allergic participants, we used allergen-specific T-cell sorting and single-cell gene expression to trace the transcriptional "road- map" of individual CD4+ T cells throughout immunotherapy. We found that successful immunotherapy induces allergen-specific CD4+ T cells to expand and shift toward an "anergic" Th2 T-cell phe- notype largely absent in both pretreatment participants and healthy controls. These findings show that sustained success, even after im- munotherapy is withdrawn, is associated with the induction, expan- sion, and maintenance of immunotherapy-specific memory and naive T-cell phenotypes as early as 3 mo into immunotherapy. These results suggest an approach for immune monitoring participants undergoing immunotherapy to predict the success of future treatment and could have implications for immunotherapy targets in other diseases like cancer, autoimmune disease, and transplantation.

2.5.1 Introduction

Allergen immunotherapy (IT) is a process in which small amounts of allergen are given over time to the allergic indi- vidual until they can safely tolerate high amounts of allergen with no signs of clinical symptoms (1–9). In the regimen of oral IT for peanut-allergic patients (identified by an allergic reaction during a standardized blinded food challenge to peanut), small amounts of peanut flour protein are ingested and escalated to a servings-worth of peanut protein (4 g of peanut protein) over a period of 2–3 y (6-7). Most patients require continuous frequent (e.g., daily) exposure to such therapy for beneficial clinical outcomes. Mecha- nistic studies of oral IT to food allergens, although limited to date, show that plasma markers such as IgE and IgG4 immunoglobulins, skin test markers, component testing, and basophil activation tests are only weakly predictive of long-term clinical success (10–16). T cells are critical upstream regulators of allergic sensitization that are required to help B cells to synthesize IgE/IgG4 immunoglob- ulins, which then can activate or inhibit basophils and mast cells (10–17). Moreover, successful IT is associated with the develop- ment of regulatory T cells (Tregs) that are thought to dampen allergic reactivity to offending allergens (7). We therefore focused on finding T-cell markers of immune tolerance that could be de- tected early in the peripheral blood during IT. CD4+ T cells can be relatively long-lived (compared with plasma proteins and ba- sophils) and changes detected early in populations of T cells could perhaps predict longer-lasting successful IT. For example, in one of the first studies in peanut allergen IT to withdraw therapy for more than 10 wk, we previously showed that despite negative skin tests to peanut, and high IgG4/low specific IgE levels to peanut, and decreased basophil reactivity to the allergen, some patients who withdrew from therapy for 3–6 mo were still reactive upon rechallenge with peanut (7). However, that study was limited be- cause of uncertainty whether the T cells monitored were peanut specific. Furthermore, many other T-cell markers which could play a role in immune tolerance (18) were not analyzed (7).

Therefore, in the current study, we focused our research on peanut-specific T cells by using tetramer technology, and we performed multiplex transcriptional profiling on single CD4+ T cells. We hypothesized that allergic individuals have a complex

set of antigen-specific and nonspecific CD4+ T lymphocytes, including allergic, non-allergic, anergic, and regulatory subtypes that undergo induction and/or transitions during IT. In addition, we posit that there may be certain subsets of CD4+ T lymphocytes that could predict more permanent vs. transient vs. refractory outcomes in IT. To test this hypothesis, we chose a model of antigen-specific oral IT for peanut allergens, and studied periph- eral blood samples from the peanut-allergic and control partic- ipants in a small, phase 1 IT clinical study (7), while excluding potentially confounding variables, such as simultaneous exposure to multiple antigens and use of concomitant immunosuppressive agents. To enable the recovery of comparable allergen-specific T-cell populations, we limited the analyses to those participants whose specific HLA haplotypes matched available allergen-specific HLA-dextramer sorting reagents. Allergen-specific T cells were collected at four time points during the first 18 mo of IT (Fig. 1). In accordance with a previously published protocol (7), par- ticipants who had no signs of clinical reactivity on a standardized food challenge (i.e., to 4 g of peanut protein) at 24 mo were withdrawn from peanut IT for an additional 3 mo, and tested for any signs of allergic reactivity at 27 mo with the same standardized food challenge. Participants were defined functionally as either "immune tolerant" (i.e., no allergic reaction with the food chal- lenge at 27 mo and thus possibly representing a more permanent clinical outcome), "desensitized" (i.e., any allergic reaction with the food challenge at 27 mo and thus possibly representing a more transient clinical outcome), or "refractory" (i.e., daily allergic symptoms to less than 300 mg peanut protein for at least 3 mo). Importantly, immunophenotyping of samples from these partici- pants was done by blinded laboratory staff and were conducted before clinical outcomes of the participants had been determined, therefore permitting us to determine, once the study was com- pleted, whether specific CD4+ T-cell subtypes could serve as possible predictors of immune tolerance.

Prior transcriptional profiling studies of lymphocytes have been based on analyses of bulk cellular populations, making it impossible to discern the cell-fate pathways and clonal re- latedness of individual T cells or even clusters of T cells of a common phenotype (7, 19, 20). By contrast, functional pheno- typing a single-cell level allows one to discern and quantify in- dividual cell phenotypes among complex mixtures of

T cells. We sorted dextramer+ and dextramer− CD4+ T cells for single-cell gene-expression profiling (21) to investigate their ontogeny in vivo. Transcript profiling was limited to 22 markers using Flu- idigm Biomark technology (22). Biological controls were obtained to compare healthy controls vs. subjects with peanut allergy at pretreatment, healthy controls vs. patients undergoing IT treat- ment, patients at pretreatment vs. during IT treatment, and dextramer+ vs. dextramer− CD4+ T cells (Figs. 2 and 3). We first performed univariate analysis comparing gene profiles of the sorted CD4+ cells, which demonstrated significance ($P < 0.00057$, Table 1) for many individual immune markers such as IL-13, CD25, IL-17A, IL-4, and ITGα4β7 between comparison groups. Interestingly, during IT some markers (CD28, CD27) seemed to "normalize" to healthy control levels but others, such as CCR7, CD25, and forkhead box P3 (FOXP3), remained significantly different ($P < 0.00057$, Table 1).

We next performed multivariate analysis of multiple immune markers simultaneously to detect possible novel CD4+ T lymphocyte subtypes. Phenotype clustering of single-cell gene-expression profiles obtained over the course of IT revealed distinct phe- notypic clusters of CD4+ T cells, with marker combinations characteristic of Th2 "allergic" (IL4+/IL13+), "nonallergic" (IFN-γ+), "regulatory" (FOXP3+/CD25+/IL10+), and "anergic" (CD28-/CD38-/IFN-γ/IL4-/IL13-/IL10-) CD4+ T-cell subsets (Figs. 4 and 5). In summary, our data show, for the first time to our knowledge, that during the course of IT, antigen-specific CD4+ T cells of diverse T-cell receptor (TCR) clonal origin expanded in frequency, and transitioned from allergic and regulatory to anergic and nonallergic phenotypes, changes that were associated with decreased allergic symptoms and the de- velopment of operationally defined immune tolerance.

2.5.2 Results

CD4+ T-cell Transcriptional Profiling. We performed transcriptional profiling of individual dextramer+ and dextramer− CD4+ T lym- phocytes throughout the course of IT in vivo, using a regimen of peanut oral IT to test our hypothesis. IT was given to peanut- allergic participants, who had no other known allergies, under a

published protocol (7), and peripheral blood was collected from these participants at different time points before treatment (pretreatment time points) and during IT at 3 mo (IT-1), 6–7 mo (IT-2), 9–10 mo (IT-3), and 11–18 mo (IT-4) (Fig. 1). One IT-3 blood draw was performed at 9 mo and the other was performed at 10 mo, whereas one IT-4 blood draw was performed at 11 mo and the other at 18 mo. Participants from whom blood was drawn pretreatment are the same individuals from whom blood was drawn during IT. CD4+ lymphocytes from each participant were labeled with dextramers specific for the peanut-derived an- tigen Ara h 2 23 (Fig. 1), the most widely recognized peanut antigen among allergic individuals (23) and dextramer+ and dextramer− CD4+ T cells were sorted separately into single-cell wells, followed by profiling of genes expressed in T cells like CD69, Ki67, CD28, CD38, CD27, CD127, IL-4, IL-13, IFN-γ, ITGα4β7, FOXP3, and IL-10 and others (Table S1) to generate heat maps and determine immunophenotyping of CD4+ T-cell subtypes (Fig. S1) (24). t tests of individual gene expression for dextramer+ CD4+ T cells between healthy controls vs. pretreatment (all pretreatment time points), healthy controls vs. IT treatment (all IT time points), pretreatment vs. IT treatment, and dextramer+ vs. dextramer− CD4+ T cells, identified several shared significant markers ($P < 0.00057$) across two or more comparisons, particularly CD28, IL-10, FOXP3, IL-17a, ITGα4β7, IL-13, CCR7, CCR8, and CD25 (Table 1). The most frequent statistically significant changes ($P < 0.00057$) were detected in the pretreatment vs. IT treatment comparison. In addition, there were several markers that were statistically different between dextramer+ and dextramer− CD4+ T cells (Table 1). Notably, the elbow method for gap statistics performed on all data (including all healthy, pretreatment, and IT cells) identified seven clusters of CD4+ T cells with distinct gene- expression patterns (Fig. 2A). The elbow method for gap statistics looks at the percentage of variance explained as a function of the number of clusters in a data set, seeking to choose a number of clusters so that adding more clusters does not significantly im- prove the modeling of the data (25). Each of the seven clusters identified had a particular "phenotype" assigned using the ex- pression level, or absence, of specific transcripts (Fig. 2 B and C, Table 2, and Fig. S2). Within most clusters, there were allergen- positive CD4+ T lymphocytes and, to a lesser extent, negatively sorted (dextramer−) cells

(which represent a cell population that is over 90% nonspecific because these cells do not have a TCR cognate for the peanut peptide-MHC used in the staining reagent) (26) (Fig. 2B). Comparison of the dextramer+ vs. dextramer− composition of each cluster by t tests showed statistically signif- icant ($P < 0.01$) different proportions of antigen-specific CD4+ T cells in each cluster, except cluster 7 (Fig. 2B). Cluster 1, IFNγ- expressing cluster 2, and nonallergic cluster 3 were primarily composed of dextramer− or "antigen-nonspecific" cells. The al- lergic cluster 4 and regulatory cluster 5 consisted exclusively of dextramer+ or "antigen-specific" cells, whereas anergic memory cluster 6 was primarily composed of dextramer+, antigen-specific cells. IL-10 expressing cluster 7 exhibited a roughly equal mixture of antigen specific and antigen nonspecific cells (Fig. 2B).

Dextramer+ cells are strongly enriched for allergy-associated phenotypes, demonstrating the specificity of the dextramer reagents used (Fig. 2B). Nonspecific binding of dextramer reagents would assume to sample randomly from antigen-specific and antigen- nonspecific T cells, and therefore to resemble the phenotypic distribution of the background population and not to enrich spe- cifically for allergy-related phenotypes. However, the phenotypic characterization of the dextramer+ cells is consistent with what is currently assumed regarding the phenotypes of allergen-specific T cells, including being greatly enriched IL4+/IL13+ cells as well as FOXP3+/IL10+/CD25+ cells compared with dextramer-cell types (Fig. 2B).

Differential Cluster Gene Expression. Gene-expression profiling al- lowed the phenotypic categorization of particular CD4+ T-cell clusters, including cluster 5, which exhibited increased IL-10, CD25, and FOXP3 expression, features associated with Tregs (27, 28), cluster 4, with increased IL-4 and IL-13 and low CD27, features linked to allergic Th2 cells (2), cluster 6, with low CD28 and Ki-67, features linked to anergic T cells (29), cluster 3, with high CD27 and low IL-4 and IL-13 expression, features linked to nonallergic cells (2), and cluster 2, with increased IFN-γ, a feature linked to Th1 cells (30) (Fig. 2 B and C, Table 2, and Fig. S2). Interestingly, only the anergic cluster 6 had a memory phenotype, exhibiting low CD45RA expression (Fig. 2 B and C, Table 2, and Fig. S2).

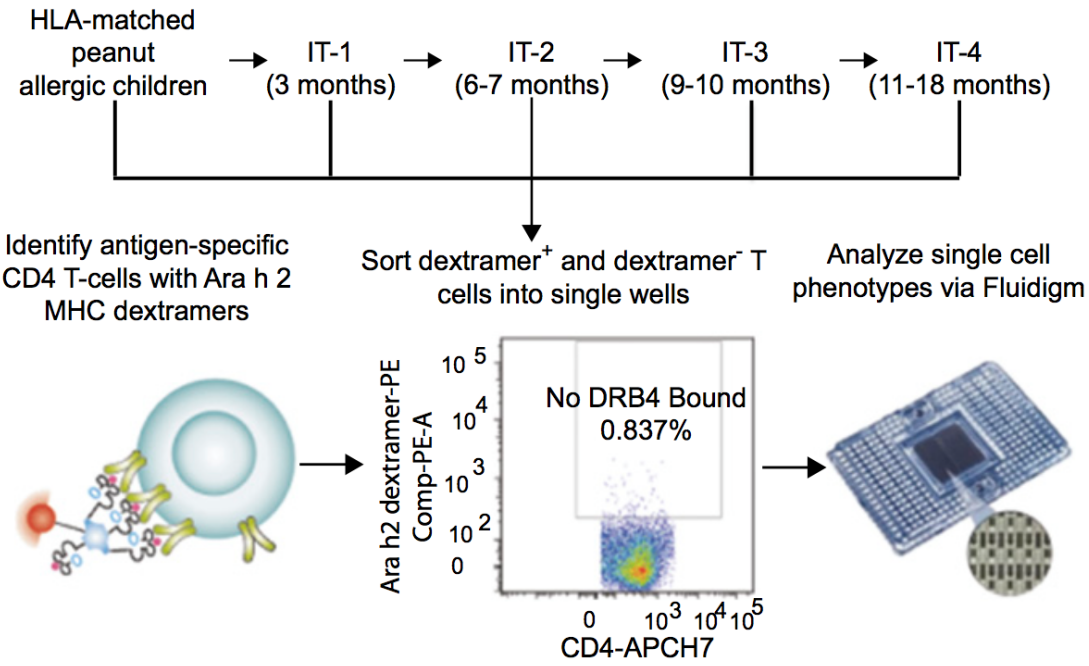


Figure 2.28: Allergen-specific CD4+ T cells are collected before and throughout IT treatment and analyzed for gene expression at the single-cell level. Schematic of the overall oral IT study and the time-point analysis. Allergen- specific CD4+ T cells were sorted from peripheral blood monocytes from HLA-DR4+ and HLA-DR15+ peanut-allergic participants by HLA-matched Ara h 2-MHC dextramer reagents on FACS into individual wells, followed by transcriptional profiling of a targeted phenotypic marker panel by Fluidigm Biomark.

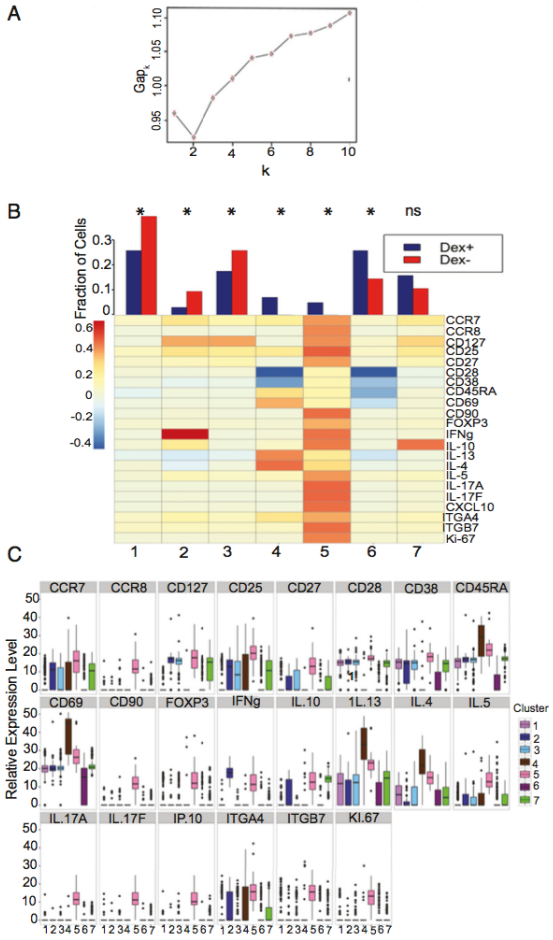


Figure 2.29: CD4+ T cells form clusters with distinct gene expression, with allergen-specific cells preferentially occupying IL4+/IL13+/CD69+ cluster 4 and FOXP3+/IL10+/CD25+ cluster 5. (A) Gap statistics graph il- lustrating the elbow method for determining the number of k-means clusters that best represents the variance observed in single-cell gene-expression data. Seven clusters were chosen. (B) Fraction of dextramer+ and dextramer− CD4+ T cells which belong to each cluster, and a heat map profile of relative gene ex- pression for each cluster. *$P < 0.0073$, ns = not sig- nificant ($\chi 2$ tests with Bonferroni-corrected P-value cutoff for multiple testing). (C) Interquartile ranges of relative gene expression for each of the seven clusters.

We next used principal components analysis (PCA) to globally visualize gene expression of single CD4+ T cells. PCA allows for data visualization by reducing the dimensionality of the data by deriving principal components (PCs) that account for the varia- tion in the data. Plotting the individual cells along the first 2–3 PCs showed a clear separation of distinct CD4+ T-cell clusters, including separation and clustering of allergic cluster 4, regula- tory cluster 5, and anergic memory cluster 6 (Fig. 3 A and B). Plotting along PC2 and PC3 continued to show distinct separation of CD4+ T-cell clusters, despite accounting for a smaller percentage of the data's variance than PC1 and PC2 (Fig. 3B).

IT Changes Cluster Proportion. Because previous studies showed that IT induces phenotypic changes in bulk T cells (31), we vi- sualized the changing composition of all seven clusters in anti- gen-specific CD4+ T cells from individual participants during the course of IT. Compared with healthy controls, IT participants had significantly increased frequencies of antigen-specific CD4+ T cells over the duration of treatment ($P < 0.01$) (Fig. 4A). In Fig. 4, it appears that whereas healthy individuals have a distribution of phenotypes in their allergen-specific CD4+ T cells that are similar to allergic patients, it is the number of these cells that is signifi- cantly different at some time points during IT (Fig. 4A). The al- lergic individuals have less cluster 4 (IL4+/IL13+) cells at baseline, but more cluster 1 (also IL4+/IL13+) cells than healthy controls.

Antigen-specific CD4+ T-cell clusters were recognized that were associated with IT vs. pretreatment (Fig. 4B). At IT-1 (3 mo into IT), compared with pretreatment, there was an increase in CD4+ T cells in the allergic cluster 4 and anergic memory cluster 6, and a decrease in CD4+ T cells in regulatory cluster 5 (Fig. 4B). As IT progressed, the anergic memory cluster 6 markedly in- creased, which was associated with a decrease in allergic cluster 4, beginning at IT-2 (6–7 mo into IT), and a reduction of allergy symptoms (Fig. 4B). Interestingly, at IT-2 there was an increase in nonallergic cluster 3 and a maintenance of IL-10-expressing cluster 7. Importantly, over time, the anergic memory cluster 6 increased at later IT time points (IT 3 and 4) compared with earlier IT time points (pretreatment, IT-1, and IT-2) (Fig. 4B).

Further, CD4+ T cells in regulatory cluster 5 decreased as IT progressed, and were undetectable at IT-4 (at 11–18 mo into IT) compared with pretreatment.

Antigen nonspecific cells did not exhibit significant changes in cluster distribution during IT, indicating that IT induced changes predominately among antigen-specific CD4+ T cells (Fig. S3A). When observing the distribution of antigen-specific CD4+ T cells in an individual participant, the relative contribution of anergic memory cluster 6 continued to increase as IT progressed (Fig. S3B). Peanut-allergic participants at pretreatment and healthy controls maintained an equal distribution of cells across the different clusters over time, although some nonstatistically sig- nificant ($P > 0.05$) fluctuations in clusters were observed (Fig. S3 C and D).

Importantly, we juxtaposed the aggregated clinical symptoms of the participants undergoing IT with the same time points in which immune monitoring occurred (Fig. 4C). As a participant ingests the food allergen as part of their IT regimen, we found that allergic symptoms such as skin rash, abdominal pain, and respiratory symptoms decreased in frequency over the time course of IT. This decrease in allergic reaction frequency over time in IT is associated with the concomitant increase in the anergic memory CD4 antigen-specific immunophenotype (Fig. 4 B and C).

Furthermore, other immune markers were examined in these same participants, including skin tests, basophil activation tests, and levels of IgG4 and IgE antibodies. Tests in healthy controls did not show significant activation of their basophils upon stim- ulation with peanut compared with media (less than 3% differ- ence in percentage of CD63+ cells for all patients), nor did healthy controls have high levels of IgE specific to peanut (<0.35 kUA/L for all individuals) or IgG4 specific to peanut (<0.14 mgA/L for all individuals). Although there were differences in basophil activation, IgE levels, skin tests, and IgG4 levels among allergic individuals, we found that in the small group of subjects analyzed, none of these markers were specifically associated with immune- tolerant vs. desensitized vs. refractory clinical phenotypes (Fig. S4). Importantly, changes in certain T-cell phenotypes occurred before changes in levels of IgE or IgG4, or in results of basophil activation (Fig. S5).

Distinct Clustering Linked to Tolerance. We next tested whether the changes in phenotypes of antigen-specific CD4+ T cells could be used to predict individual

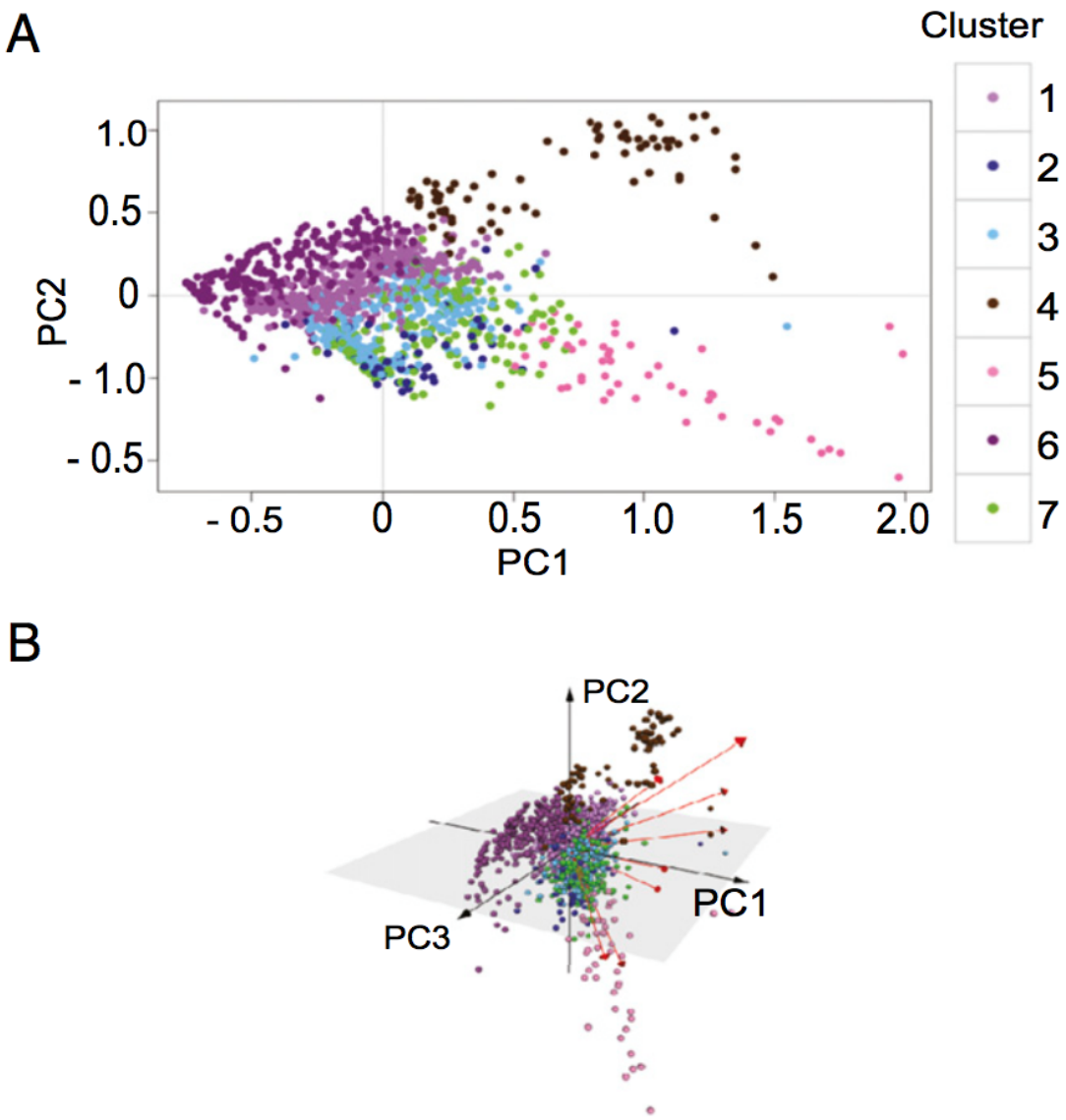


Figure 2.30: Antigen-specific CD4+ T cells form seven cluster phenotypes with distinct gene expression. PCA representation of variations in cell pheno- types, plotting cells along (A) PC1 and PC2; (B) PC1, PC2, and PC3. The cells of all participants at all time points are colored according to the CD4+ T-cell clusters to which they were assigned using k-means. Variance accounted for by each PC: PC1 = 32.8%, PC2 = 15.8%, PC3 = 8.83%, PC4 = 6.34%.

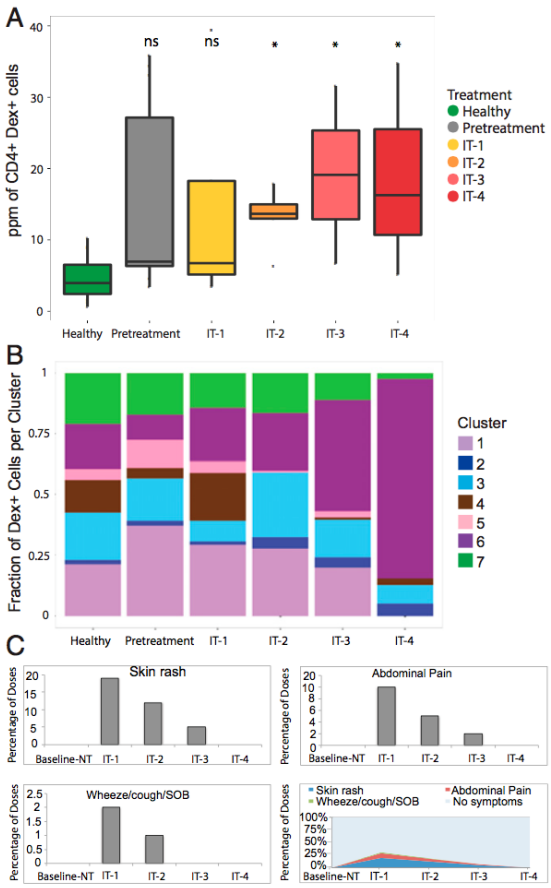


Figure 2.31: Cluster 6 T-cell receptor nonresponsive (CD28-/CD38-) noninterleukin secreting (IL13-/IL4-/IL10-/IL5-/IFNγ-) antigen-specific CD4+ T cells expand and allergy symptoms diminish over the course of IT. (A) Interquartile ranges for ppm of antigen-specific CD4+ T cells from pooled healthy (n = 7), pre- treatment (n = 5), IT-1 (n = 5), IT-2 (n = 5), IT-3 (n = 2), and IT-4 (n = 2) participants. *P $<$ 0.01, ns = not significant (t tests comparing each time point to healthy controls with Bonferroni-corrected p-value cutoff for mul- tiple testing). (B) The fractional proportion of dextramer+ CD4+ T cells in each phenotype cluster from all participants at each IT time point IT-1 (n = 5), IT-2 (n = 5), IT-3 (n = 2), and IT-4 participants (n = 2), pretreatment (n = 5), and healthy controls (n = 7). (C) Percentage of doses resulting in allergy symptoms observed within 2 h of daily peanut ingestion within each IT time frame drawing from all IT participants.

clinical outcomes following IT. To do this, we visualized transitions in CD4+ T-cell phenotypes during the course of IT among individual participants for whom we later determined clinical phenotypes of refractory, immune tol- erant, or "desensitized." PCA permitted us to detect distinct T-cell clusters over time during IT, with distinct individual gene- expression patterns for T cells in each cluster at each IT time point (Fig. 5 A and B and Table 1). At pretreatment, there was a diversity of clusters represented for all participants. Interestingly, in both the refractory and immune-tolerant patients, there were no allergic cells at pretreatment but at IT-1 there was a significant transition toward allergic cluster 4 cells for the refractory clinical phenotype with an indistinct scattering of cells across several clusters as IT progressed (Fig. 5 A and B). Whereas cells remained partly scattered across clusters in the desensitized clinical phenotype during IT, there was some transition toward nonallergic and anergic memory clusters as IT progressed (Fig. 5B). In contrast, for the immune-tolerant clinical phenotype, cells transitioned distinctly to nonallergic cluster 3 and anergic memory cluster 6 during IT (Fig. 5B).

A comparison of levels of each biomarker per cluster demon- strated significant differences ($P < 0.00057$) in the pretreatment vs. IT groups and a shift over time toward the anergic memory cluster (Table 1). The phenotypic shift distance, a measure of variation in all markers for each cell from one IT time point to the next, revealed that immune-tolerant and desensitized clinical phenotypes had statistically significant ($P < 0.001$) reduced vari- ance in the later stages of IT (Fig. 5C). The refractory clinical phenotype, however, exhibited greater phenotypic shifting throughout IT ($P < 0.001$) (Fig. 5C). Overall, antigen-specific CD4+ T cells in immune-tolerant and desensitized individuals appeared to settle on anergic memory or nonallergic immunophenotypes during IT, whereas those in the refractory clinical phenotype continued to represent several different phenotypes.

CD4+ T-Cell TCR and Gene Expression. Because successful IT in- duced significant increases in anergic memory and nonallergic CD4+ T-cell clusters, we sought to define the characteristics of CD4+ T-cell clones in a representative immune-tolerant clinical phenotype. We sequenced TCRs and determined additional gene expression

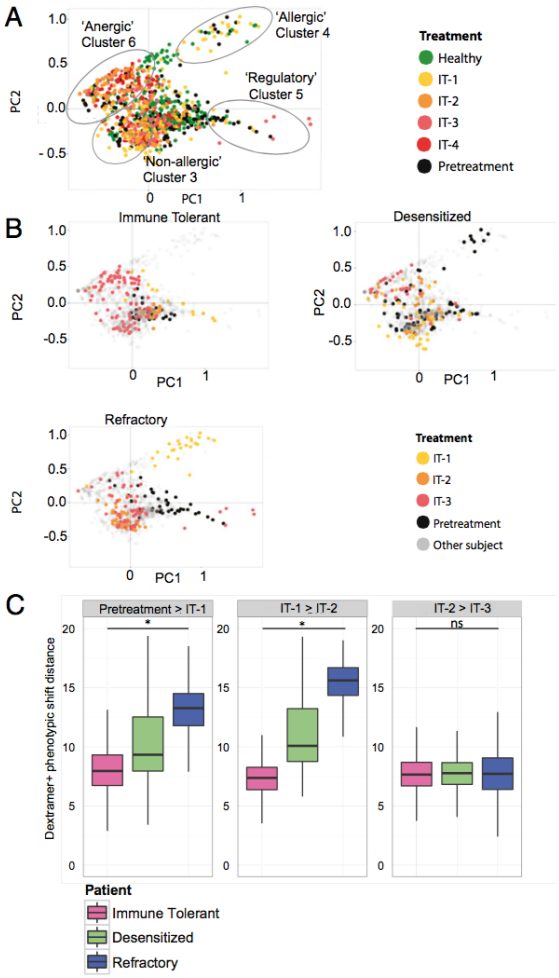


Figure 2.32: Temporal expression of CD4+ T-cell clusters reveals individualized patterns associated with clinical phenotypes. (A) Two-dimensional PCA of antigen-specific CD4+ T cells from all healthy subjects (green), and all pre- treatment time points (black), IT-1 (yellow), IT-2 (orange), IT-3 (pink), and IT-4 (red) time points for IT participants. Variance accounted for by PCs: PC1 = 32.8%, PC2 = 15.8%. (B) Representative phenotypic shifts across IT time points from one immune tolerant subject, one refractory subject, and one desensi- tized subject, displayed via 2D-PCA for antigen-specific CD4+ T cells from immune-tolerant, desensitized, or refractory participants over the course of IT. Gray dots represent cells from other participants. (C) Interquartile range of "phenotypic shift" distances for dextramer+ CD4+ T cells from individual participants between IT time points. The one immune-tolerant subject, one refractory subject, and one desensitized subject are the same individuals that appear in B. Higher shift values indicate greater average change in CD4+ allergen-specific phenotype between time points. Phenotype shifts were for months 0–18 during IT, whereas "desensitized/immune tolerant/refractory" status was determined at month 27. *P $<$ 0.001, ns = not significant (one-way ANOVAs). The calculations were performed from every cell in one time point to every cell in the next time point within an individual. The total number of cell–cell comparisons are summarized in Table S2.

in individual antigen-specific and nonspecific CD4+ T cells (21) from an immune-tolerant participant at IT-2, when several of the phenotypic transitions began to emerge (Figs. 4B and 5B). By doing this, we gained insight into the clonal TCR and additional gene-expression changes linked to tolerance induction.

Both antigen-specific and nonspecific CD4+ T cells were ob- served to be poly-clonal, indicating no particular TCR bias (Fig. 6). TGF-β1, which has been linked to tolerance induction and Th2 inhibition, was expressed by several clones (Fig. 6) (32, 33). Runt-related transcription factor, RUNX1, which represses Th2 programming (34), was expressed in multiple clones (Fig. 6). Interestingly, GATA-3, the transcriptional regulator of Th2 de- velopment, was expressed in several antigen-nonspecific clones, often in association with TGF-β1, and not IL-13, a Th2 cytokine (Fig. 6). The Th1 cytokine, TNFα, was expressed by many clones, often concurrently with TGF-β1, but not necessarily with T-BET, the transcriptional regulator of Th1 development (Fig. 6). Ex- pression of TNFα is consistent with presence of Th1 cells; however, it is difficult to draw definitive conclusions based on a limited number of cells. Few other cytokine transcripts were expressed by the selected clones, including limited expression of IFN-γ, IL-13, IL-12, and IL-21. The follicular helper T-cell (Tfh) lineage commitment factor BCL-6, was expressed by some an- tigen-specific clones, raising the possibility that Tfh may have a role in tolerance induction during IT (Fig. 6). Additionally, one FOXP3-expressing clone was observed (Fig. 6). Although all dextramer+ cells were unique, in 1,000 repeat random sub- samplings of 13 CDR3 sequences from 5´RACE-derived, naive TCR repertoire, four motifs (GLT, PTG, LTD, and RVA) were found to be significantly elevated above expectation in the dextramer+ set ($P < 0.01$). These four motifs occur in partially overlapping regions in the middle of TCRb CDR3, shared across 4 of the 13 recovered dextramer+ single cells. Overall, during successful IT, it is possible to observe marked expression of tolerogenic TGF-β1 and indistinct lineage commitment or cyto- kine expression by CD4+ T cells.

dex +/-	Vb	Jb	Va	Ja	CDR3b	CDR3a	FOXP3	GATA3	IFNG	IL13	IL12	IL21	TBET	TGFB1	TNF	BCL6	RUNX1	RUNX3
dex-	V12-3	J1-1	V41	J42	CASSDQHGVDTEAFF													
dex-	V18	J2-3	V6	J24	CASSPPTGGFPDTQYF	CALDLPTDSWGKLQF		■										
dex-	V19	J2-3	V19	J23	CASSTLASGAFTDTQYF									■	■			
dex-	V20-1	J2-1	V21	J52	CSAREGLQRDNEQFF	CAVRGAGGTSYGKLTF								■	■			
dex-	V20-1	J2-5			CSASGTEETQYF										■		■	
dex-	V20-1	J2-7	V13-1	J3	CSARGSGTGTYEQYF	CAARYSSASKIIF												
dex-	V20-1	J2-3	V12-1	J4	CSARSLSGTQGSQYF	CVVSRSGGYNKLIF												
dex-	V24-1	J2-1	V8-2	J22	CATRTSGRNEQFF	CVVSDGSARQLTF		■									■	■
dex-	V25-1	J2-4	V13-1	J41	CASSHRDIQYF	CAASISGYALNF		■						■				
dex-	V27	J1-2	V10	J7	CASSSQRGANGYTF	CVVSANCGNNRLAF		■									■	
dex-	V27	J2-1			CASRRSLGAGYNEQFF													
dex-	V28	J2-3	V2	J34	CASSLTESTDTQYF	CAVEDGNTDKLIF								■	■			
dex-	V28	J2-2	V35	J26	CASSQQGTNNGELFF	CAGRYNYGQNFVF									■			
dex-	V29-1	J2-5	V27	J24	CSVGSQGAEETQYF	CAGGTDSWGKLQF	■	■						■			■	
dex-	V29-1	J2-7	V26-1	J52	CSVKIDDRVYEQYF	CIVLGTSYGKLTF									■		■	
dex-	V30	J2-4	V9-2	J42	CAWRVSGRRQNIQYF	CALSDPFYGGSQGNLIF									■			
dex-	V4-1	J1-5	V13-1	J24	CASSQDRGGNQPQHF	CAASIPRSATDSWGKLQF								■				
dex-	V4-3	J1-2			CASSQDQGPYGYTF			■						■	■		■	
dex-	V5-1	J2-5			CASSPQGAGETQYF				■			■			■			
dex-	V5-1	J1-1	V2	J16	CASSRDRQNTEAFF						■	■		■	■			■
dex-	V5-4	J2-7	V26-2	J52	CASSSPRQGSYEQYF	CILPLAGGTSYGKLTF											■	
dex-	V6-1	J2-5	V13-1	J49	CASSELGGRVQETQYF	CAASKKNTGNQFYF					■	■			■			
dex-	V6-5	J1-1	V8-2	J11	CASSYEADPLNTEAFF	CAVRFSGYSTLTF					■			■	■			
dex-	V7-2	J2-3	V12-3	J53	CASSLNRDSPRF	CAMPPRDSGGSNYKLTF									■			
dex-	V7-2	J2-1	V27	J49	CASSKTSGSSYNEQFF													
dex-	V7-6	J1-4			CASSLDRRPTNEKLFF			■						■				
dex-	V7-8	J2-4	V12-1	J40	CASSLTPGGAKNIQYF	CVVNAAAGTYKYIF								■				
dex+	V15	J2-4	V8-2	J5	CATSLAGGAKNIQYF	CVVSNTGRRALTF					■				■			
dex+	V19	J2-2	V1-2	J29	CASREQGFAGELFF	CAVRDRGNTPLVF			■						■			
dex+	V20-1	J2-3	V26-1	J39	CSARVAGLTDTQYF	CIVRVTNAGNMLTF							■	■	■			
dex+	V20-1	J2-7			CSAQAGGYEQYF									■			■	
dex+	V20-1	J2-3			CSARVRGLSGRSGTDTQYF									■			■	
dex+	V20-1	J2-4			CRHNGGRARGNIQYF					■								
dex+	V4-3	J2-5			CASSPLRVAEPQETQYF											■	■	
dex+	V6-6	J2-7	V26-2	J34	CASSGQEYPYEQYF									■				
dex+	V7-2	J1-6	V6	J42	CASSLVPTGGLGNSPLHF	CALALGGSQGNLIF							■	■	■			
dex+	V7-2	J2-7			CASSPGQGGGYYEQYF									■				
dex+	V7-2	J2-1	V19	J24	CASSLVSGGPYNEQFF	CASLTDSWGKFQF												
dex+	V7-9	J1-1	V34	J31	CASSPRWDTEAFF	CGVNNNARLMF												
dex+	V9	J2-3	V19	J24	CASSAPTGLTDTQYF	CALSEAKTDSWGKLQF												

Figure 2.33: Single-cell TCR sequencing and gene expression during successful IT demonstrates tolerogenic gene expression without defined lineage com- mitment. TCR sequencing of CDR3 regions and Vβ, Vα, Jα, and Jβ use, and RT-PCR of transcript expression, for individual sorted dextramer+ or dextramer− CD4+ T cells at IT-2 from one participant later determined to be immune tol- erant. Shaded boxes represent gene expression of FOXP3, GATA3, IFN-γ, IL-13, IL-12, IL-21, T-BET, TGF-β1, TNFα, BCL-6, RUNX1, or RUNX3. Horizontal line sep- arates data from dextramer− cells (above the line) and dextramer+ cells (below the line).

2.5.3 Discussion

Our goal in this study was to identify mechanisms involved in IT by using single-cell gene profiling, combined with multivariate statistical analyses. Quantifying single-cell gene expression has applications across many biological fields (35–37). High-throughput transcriptional profiling of single cells and computational modeling enabled us to track, on an unprecedented level, the molecular details of CD4+ lymphocytes during IT in vivo. We found evidence of both antigen-specific and -nonspecific CD4+ lymphocytes belonging to seven different phenotypic clusters with distinct gene-expression profiles. Many of our computational anal- yses of phenotypic transitions were possible only with data derived from single cells. In particular, our data showed significant distinct transitions in antigen-specific CD4+ T cells that were not observed in antigen-nonspecific CD4+ T cells. Notably, shifts in T-cell populations appeared before significant changes in basophil ac- tivation, IgE levels, or IgG4 levels, and were more predictive than such tests of an individual participant's clinical outcome (Figs. 4 and 5 and Fig. S4). These findings demonstrate the potential importance of monitoring T cells during IT and suggest that single-cell approaches may be useful for future studies on the effects of IT and the immune monitoring of individuals un- dergoing IT. Further, significant differences in CD28 and CD38 expression between experimental groups suggest that it would be interesting in future studies to monitor other cell types, such as B cells and natural killer cells, which express the same markers as the T cells we monitored in our study.

Although Ara h 2 is only one among multiple clinically rele- vant peanut allergens, it is of great significance that this single representative allergen allows for complex insights into the cel- lular mechanisms of the allergic process. It is significant that strong statistical statements and predictive observations can be made using Ara h 2 as a single representative allergen. It is possible that the population of dextramer− cells could contain some CD4+ T cells specific against other peanut allergen peptides displayed on the same or other HLA molecules. How- ever, given the frequency of the known immunodominant peanut-specific cells (5–20 cells per million) (Fig. 4A), even with 1,000 different peanut epitopes just as immunodominant as the one we investigated, with each represented at 20 parts per million, we would expect less than

1 in 50 T cells would be allergen-specific from the circulating CD4+T-cell repertoire, and thus would not affect our statistical results in any meaningful way. Further, it is worth noting that although cells were obtained from the pe- ripheral blood rather than the gut, studies in celiac disease have indicated that several days after ingesting bread, gliadin-specific T cells appear in the peripheral blood (38). Therefore, we speculate that daily exposure to peanuts through oral IT may contribute, via egress of peanut-specific T cells from the gut to the blood, to elevations in the frequencies of peanut-specific T cells in the peripheral blood of patients during IT.

Although there was a relatively small sample size of participants in our study, we were able to follow their single cells sequentially over time and therefore perform detailed, multidimensional comparisons against a previous time point. To address the small sample size, we conducted studies in many control settings (i.e., health controls vs. allergic controls, pretreatment vs. treatment, negative-sorted vs. positive tetramer-sorted cells) to best de- termine the biological significance of our findings. Sequential single-cell gene-expression measurements in CD4+ T lymphocytes during the course of IT identified unique sets of T-cell clusters that were significantly linked to the immune-tolerant clinical pheno- type. As IT progressed, anergic memory and nonallergic antigen- specific CD4+ T lymphocytes were significantly induced only in the immune-tolerant individuals. A regulatory CD4+ T-lympho- cyte cluster was also observed as having CD45RA expression. Perhaps there is meaningful variation in the intermediate region of the CD45RA expression and further studies will be done to characterize these potential regulatory populations. The regulatory cluster expressed CD127, suggesting that it is likely an induced regulatory T-cell subset, not a thymic-derived regulatory subset (3, 27, 39. At the same time points, the desensitized clinical phenotype exhibited a greater dispersion of cells across the different clusters, with only some transitions toward nonallergic and anergic memory clusters. Notably, the desensitized subject was the only one of these three subjects to have allergic cells pretreatment. This might be predictive of clinical outcome. In contrast, in the refractory clinical phenotype, cells spread indistinctly across several different clusters as IT progressed. Of interest is a shift in the refractory patient from no allergic cells at pretreatment to a large number of allergic cells at IT-1 (correlating with the peak of

Gene	Healthy vs. pretreatment	Healthy vs. IT	Pretreatment vs. IT	Dextramer+ vs. dextramer−
CCR7	0.5963	0.0003	0.0001	0.1180
CCR8	0.0015	0.0410	<0.0001	<0.0001
CD127	0.7228	0.5096	0.2863	0.5007
CD25	0.1125	<0.0001	<0.0001	0.0057
CD27	0.0127	0.5430	0.0011	0.0222
CD28	<0.0001	0.3085	<0.0001	<0.0001
CD38	<0.0001	0.0442	<0.0001	0.3122
CD45RA	0.0383	<0.0001	<0.0001	0.1107
CD69	0.0119	0.2122	<0.0001	<0.0001
CD90	0.0001	0.2854	<0.0001	<0.0001
FOXP3	0.3602	<0.0001	<0.0001	0.4124
IFNγ	0.0414	0.4841	0.1087	0.4301
IL-10	0.9659	0.0171	0.0244	0.0416
IL-13	0.0074	<0.0001	<0.0001	<0.0001
IL-4	0.6381	0.0055	0.0020	<0.0001
IL-5	0.4080	0.1204	0.0249	<0.0001
IL-17A	0.0066	0.0337	<0.0001	<0.0001
IL-17F	0.0051	0.0321	<0.0001	<0.0001
CXCL10	0.0050	0.0209	<0.0001	<0.0001
ITGA4	0.0023	0.7393	0.0003	0.0284
ITGB7	0.0001	0.0446	<0.0001	0.0076
KI-67	0.0006	0.0291	<0.0001	<0.0001

Table 2.1: Ryan Hovde Glanville PNAS 2015 Table1.

allergic symptoms) be- fore transitioning toward anergic and nonallergic phenotypes at later IT time points. This spike in allergic cells coinciding with IT-1, which was not seen in either the immune-tolerant or desen- sitized patient, may be predictive of the refractory clinical phe- notype. By using current single-cell sorting technology, we were able to distinguish distinct anergic and nonallergic cellular phenotypes that could not have been identified using traditional immunophenotyping methods (39). Defining phenotypic transition patterns in antigen-specific and antigen-nonspecific lymphocytes during successful IT might allow us to infer the timing of expression changes in key genes associated with induction of tolerance.

We defined the anergic memory phenotype by the lack of co- stimulatory receptor CD28 and the proliferation-associated anti- gen Ki-67, and low-to-absent expression of the early activation marker CD69, indicating a nonproliferating, nonactivated phe- notype with diminished CD28 signaling potential, which has been linked to anergic CD4+ T cells (40). However, our current single- cell gene-expression study was

technically limited by the number of markers that could be used for the comprehensive detection of anergic cells. Using a more expansive panel of anergy markers (i.e., Cbl-b, GRAIL, program-death-1, and cytotoxic T-lymphocyte- associate protein-4) (20, 41–45) may elucidate in greater detail the transitioning phenotypes of anergic cells during IT.

Although the generation of long-lived memory lymphocytes would seem to be an essential feature of IT, we observed several changes in the naive antigen-specific and -nonspecific CD4+ T cells. Most importantly, the predominant anergic cluster 6 lymphocytes induced during IT had a memory phenotype and were increased substantially over all other clusters. Although there is background variability between subjects at pretreatment time points, the expansion of the anergic cluster 6 can be seen to be much larger than any background variation (Fig. 4) and was statistically significant compared with background ($P < 0.01$). This may indicate that repeated antigen treatments during the course of IT may be critical for anergy induction in memory CD4+ T cells, and for the promotion of a tolerogenic environment that inhibits pathogenic CD4+ T-cell recruitment and aberrant re- sponses (43). Although we do not usually think of anergic cells as a population that proliferates, there are some plausible explanations of this observed phenomenon. A change in IL-2 production could account for this expansion. Also, it is plausible that through the course of IT other nonanergic cell types are activated and pro- liferate before becoming anergic. These cells may then act upon subsequent allergen exposure, suppressing the allergic response. It would be interesting in future studies to further elucidate the mechanism of this observed expansion, and to define in greater detail the phenotype of these cells, measuring the presence of cytokines (such as IL-2) and markers of effector memory T cells (such as CD62L).

Our results also suggest that successful IT can induce poly- clonal antigen-specific and -nonspecific CD4+ T cells, with in- distinct lineage commitment, possibly indicating their functional plasticity and transitory phenotype. During successful IT, several clones expressed TGF-β1, which has been linked to anergy and tolerance induction during IT, typically by way of Tregs (31). Further, clones expressing GATA-3 also largely coexpressed TGF-β1, which inhibits GATA-3–driven Th2 differentiation

Color:	Light purple	Royal blue	Light blue	Brown	Pink	Dark purple	Green
Clusters:	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7
	Nonspecific	"IFNγ-Expressing"	Nonallergic	Allergic	Regulatory	"Anergic memory"	"IL-10-Expressing"
Antigen-specific:	Negative	Negative	Positive	Positive	Positive	Both	Both
Markers							
CD69	0	0	0	↑↑ increase	↑ increase	**↓↓ decrease**	0
CD38	0	0	0	**↓↓↓ decrease**	↑ increase	**↓↓↓↓↓ decrease**	decrease
CD28	0	0	0	**↓↓↓↓↓↓↓↓↓ decrease**	↑ increase	**↓↓↓↓↓↓↓↓↓↓ decrease**	decrease
CD45RA	0	0	0	**↑↑ increase**	↑ increase	**↓↓↓↓↓ decrease**	0
CD25	↑ increase	↑↑ increase	↑ increase	**↑ increase**	**↑↑↑↑ increase**	0	↑ increase
FoxP3	0	0	0	0	**↑↑ increase**	0	0
IL-4	0	decrease	0	**↑↑↑ increase**	↑ increase	decrease	decrease
IL-13	decrease	decrease	**↓ decrease**	**↑↑ increase**	↑ increase	↓ decrease	decrease
CCR7	↑ increase	↑↑ increase	↑ increase	↑↑increase	**↑↑↑ increase**	0	↑↑ increase
ITGA4	↑ increase	↑↑ increase	↑ increase	**↑↑ increase**	**↑↑ increase**	0	↑ increase
IL-10	negative	↑↑ increase	negative	0	**↑↑↑ increase**	decrease	**↑↑↑ increase**
IFN-γ	negative	**↑↑↑↑↑ increase**	negative	0	↑↑↑ increase	decrease	decrease
CD27	↑ increase	↑ increase	**↑ increase**	↓ decrease	↑↑ increase	0	↑ increase
IL-5	0	↑ increase	0	↑ increase	↑↑ increase	↑ increase	↑ increase
CCR8	0	0	0	decrease	↑↑ increase	decrease	0
CD127	0	↑↑ increase	↑↑ increase	decrease	↑↑ increase	0	↑↑ increase
CD90	0	0	0	0	↑↑↑ increased	0	0

Notes: (*i*) All expression is relative to 18S housekeeping control, 0 = no change. (*ii*) Most distinguishing markers for each cluster phenotype (shaded/bold). (*iii*) ↑ or ↓ indicate relative magnitude of marker expression increase or decrease. (*iv*) Seventeen out of 22 markers shown here because these were the markers significantly associated with cell clusters.

Table 2.2: Ryan Hovde Glanville PNAS 2015 Table2.

(33). This may be another indication of transitory CD4+ T-cell phenotypes within a successfully treated individual, with allergic Th2 cells undergoing transcriptional reprogramming to tolerogenic TGF- β1-expressing cells. Additionally, several clones expressing TNFα also coexpressed TGF-β1, again indicating functional plasticity and transitory phenotypes of CD4+ T cells during IT.

The results of the TCR sequencing gives confidence that the cells we sequenced are representative of a larger pool of antigen- specific T cells and do not display only clones that were greatly expanded. Initially when performing our phenotype analysis we did have some concerns that there might be large clonal ex- pansions that would bias our analysis of T-cell phenotypes to- ward the most expanded T cells. However, in our TCR sequencing analysis, we determined that the recovered populations are quite diverse. This implies that the total number of allergen-specific clones is considerably larger than the number of T cells that we characterized, and that the single-cell phenotyping analysis per- formed was sampling phenotypes from almost entirely non- redundant clones. Further, evidence of specific shared and over- lapping CDR3 motifs could possibly indicate that a diverse population of clones converges to recognize pMHC complexes with similar selected amino acid motifs. In future

Participant type	Number of participants	IT time points, mo	Frequency of $CD4^+Dex^+$, median (range)	Age, median (range)	Sex	HLA type
Healthy controls	7	−3, 0	22.5 (3–55)	38 (32–47)	Males = 1 Females = 6	1501 = 4 DRB4 = 3
Pretreatment participants	5	−6, −3, 0	27.5 (9–48)	10 (8–15)	Males = 1 Females = 4	1501 = 2 DRB4 = 3
IT participants	5	3, 6–7, 9–10, 11–18	44 (14–78)	10 (8–15)	Males = 1 Females = 4	1501 = 2 DRB4 = 3

Table 2.3: Ryan Hovde Glanville PNAS 2015 Table3.

studies, this suggests that such motifs could be used to recognize allergen- specific T cells from primary sequence (21). Although we only acquired TCR sequencing data during treatment, it would be interesting in future studies to match TCR sequencing data ob- tained during treatment to TCR sequencing data acquired before treatment begins to answer important questions pertaining to clonal transformation during IT treatment.

Studies of antigen-specific IT have implicated FOXP3-expressing Treg induction in association with immune tolerance in allergy, diabetes, and multiple sclerosis (7, 46–49), and a subversion of Th2 or Th1 responses in allergy and diabetes, respectively (47, 48, 50). However, we observed a reduction in allergic Th2 cells and a reduction in antigen-specific regulatory cells in the peripheral blood over time during IT. Preferential deletion of antigen-spe- cific Th2 cells has previously been observed following IT (1) and decreased Treg over time in IT has been reported (7, 15). Further repertoire analysis of CD4+ T cells during IT may reveal the preferential deletion or reprogramming of T cells based on fac- tors like functional avidity or structural TCR avidity. Addition- ally, several IT studies, particularly for allergy, have identified changes in antigen-specific IgE and IgG responses, regulatory B cells, and basophils associated with tolerance induction (13, 17, 39, 46, 51).

In conclusion, our study is the first, to our knowledge, to show complex phenotypic transitions in CD4+ T cells during IT. By analyzing the gene-expression patterns of individual CD4+ lymphocytes, we were able to reconstruct the pathways of these cells as they transitioned to “tolerant” or “nontolerant” states. This approach has yielded previously unidentified insights into the po- tential mechanisms of tolerance induction during oral IT. Although our study has used oral IT in food allergy as a model

treatment, the mechanisms we have uncovered and the methods we have applied could be relevant to other forms of IT and disease states associated with modulation of the immune system, such as cancer, autoimmune diseases, and transplantation. The data presented here provide important insights into the changes in gene expression and T-cell phenotypes that may occur during successful IT.

2.5.4 Methods

Study Design. The study was designed to discern the changes in CD4+ T-cell phenotypes that underpin the induction of immune tolerance in oral IT. To this end, we isolated CD4+ T cells from IT participants and healthy controls (individuals who have no known allergies and have not taken drugs that are known to influence peripheral T-cell response) at predefined time points (Table 3), sorted individual T cells, and performed single-cell gene-expression analyses and TCR sequencing and phenotyping on these isolated cells. The protocol for this study was reviewed and approved by the Institutional Re- view Board (IRB) of Stanford University. Written informed consent was obtained for all participants before entering the study.

Participants and IT. Participants with matching HLA types, HLA-DRB1*1501 and HLA-DRB4, compatible with peanut-derived Ara h 2 peptide dextramers from a previous recent study (7), were enrolled in this pilot study. Double- blind, placebo-controlled food challenges (DBPCFCs) occurred at screening and clinical reactivity was determined. Clinical reactivity is defined as any sign of allergic reaction. The oral IT protocol was conducted in a hospital setting with trained staff. Healthy controls were proven to be nonallergic via serum IgE less than 0.35 kU/L and no clinical symptoms consistent with atopy and no positive food challenges. Peanut allergy participants were confirmed using DBPCFCs conducted under an IRB-approved protocol at Stanford University School of Medicine. Par- ticipant demographics can be found in Table 3. Further subject information, including eligibility criteria, and further IT details, including food challenge dosing protocol, have been previously published (7).

Cell Preparation and Sorting. From each subject, 20–40 mL of blood was obtained during the afternoon and placed on a rotator overnight before being processed in the

morning the following day. Basophil activation assays were performed as previously described (52). Specific IgE and IgG4 were measured (Stanford Clinical Laboratories). CD4+ T cells were isolated using the Human CD4+ T-cell enrichment kit (StemCell Technologies), according to the manufacturer's protocol. CD4+ T cells were activated with phorbol myristate acetate (PMA) at 20 ng/mL and Ionomycin at 1 μg/mL at 37 °C for 1.5 h to up-regulate cytokine expression. After washing twice, CD4+ T cells were stained with PE-labeled Ara h 2 dextramers [DRB1*1501 and DRB4; Ara h 2 (120-139, RQQEQQFKRELRNLPQQCGL)] (Immudex) at room temperature for 45 min. Cells were then stained with anti-CD3 V500, anti-CD4 APC-H7, anti-CD8 FITC, anti-CD14 FITC, anti-CD19 FITC (BD Biosciences), anti-CD45RA brilliant violet 421 (BioLegend), anti-CD56 FITC, and anti-CD294 Alexa Fluor 647 (BD Biosciences) for 20 min at 4 °C. Cells were washed and incubated with anti-PE magnetic beads (Miltenyi Biotec), according to the manufacturer's protocol, and a 1/20 fraction was saved for analysis. The other fraction was passed through a magnetic column (Miltenyi Biotec). The bound, PE-labeled cells were flushed and collected. Cells in the bound fraction and the fraction not passed through the column were stained with 7-AAD (BD Biosciences) for 10 min before flow cytometry. Dextramer-stained T cells were sorted as single cells into individual wells of a 96-well plate. Cells were stored in reverse transcriptase reaction buffer at −80 °C until use. Flow cytometric data were acquired on a BD Fluorescence Activated Cell Sorting (FACS) Aria (BD Biosciences) and the data analyzed using FlowJo (FlowJo).

Fluidigm. This assay was performed by the Human Immune Monitoring Center at Stanford University. For single sorted cells, RT-PCR was performed directly in a 96-well PCR plate (ABI) containing lysis buffer (Invitrogen) by using SuperScript III One-Step RT-PCR System with PlatinumTaq (CellDirect kit, Invitrogen). PreAmp was performed on a thermocycler using the TaqMan PreAmp Master Mix Kit (Invitrogen) added to cDNA and 0.2× pooled Taq- man assays. RT enzyme was inactivated and the Taq polymerase reaction was started by bringing the sample to 95 °C for 2 min. The cDNA was pre- amplified for 18 cycles by denaturing at 95 °C for 15 s, annealing at 60 °C for 4 min. The resulting cDNA product was diluted 1:2 with 1×

TE buffer (Invi- trogen). Next, 2× Applied Biosystems Taqman Master Mix, Fluidigm Sample Loading Reagent, and preamplified cDNA were mixed and loaded into the 48.48 Dynamic Array (Fluidigm) sample inlets, followed by loading 10× as- says into the assay inlets. Manufacturer's instructions for chip priming, pipetting, mixing, and loading onto the BioMark system were followed. Real-time PCR was carried out with the following conditions: 10 min at 95 °C, followed by 50 cycles of 15 s at 95 °C and 1 min at 60 °C. Data were analyzed using Fluidigm software. All reactions were performed in duplicate or trip- licate, and Ct values were normalized to the 18S positive control.

TCR Sequencing and Phenotyping. TCR sequencing and phenotypic analyses were performed as previously described (21). Briefly, PCR sequence and gene- expression analysis from single cells were obtained by a series of nested PCRs for multiple Vα, Vβ, Jα, Jβ, Cα, and Cβ regions and multiple genes, including FOXP3, GATA-3, IFN-γ, IL-13, IL-12, IL-21, T-BET, TGF-β1, TNFα, BCL6, RUNX1, and RUNX3. Bar-coding PCRs were used to track PCR products from individual cells that were sequenced using the Illumina MiSeq platform (Illumina Inc.).

Initial Data Transformation. Initial Fluidigm output consists of marker expression levels ranging from 0 to 40, which represent the number of amplification cycles necessary to obtain a threshold level of the marker. A score closer to 0 indicates a greater expression level, and a score closer to 40 indicates a lesser expression level. Any markers not expressed at a high enough level to be detected after 40 cycles receive a score of N/A (nonapplicable). We trans- formed them by subtracting each score from 40, so that higher transformed scores correspond to higher expression level. All N/A values were set to 0.

Normalization. To control for interplate variance, we normalized using the expression of 18S, a housekeeping gene, to establish a baseline level of activity for each plate and adjust accordingly. We found the median 18S expression of all of the cells on each plate and then found the median of these 35 medians. An "activity level" for each plate was calculated by dividing the median 18S expression of that plate by the median-of-medians. Then, each marker ex- pression level for each cell on the plate was divided by this activity level.

Cluster Analysis. We started by using the gap statistic to compare within- to between-cluster sums of squares for various numbers of clusters (25). After plotting these and looking for an elbow, we divided the cells into seven clus- ters. Before clustering, the data for each marker were centered by the overall median and scaled by the range. Cells were then clustered using k means. To ensure that all presenting phenotypes were represented and contributed to cluster designation, all cells (both dextramer+ and dextramer−) were included for cluster analysis at all time points. All statistical analysis and graphs were performed using R. K-means clustering (53) was performed using the R pack- age "kmeans" (https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans. html). Heat map visualization was performed using the R package "heatmap 2" (www.inside-r.org/packages/cran/gplots/docs/heatmap.2). The agglomerative clustering method used for dendrogram construction was complete-linkage clustering, using a Euclidean metric. The clustering method passed to the clusGap function was kmeans, with k ranging from 1 to 7.

PCA was performed using the single-cell gene-expression data to visualize the relationship among the individual cells. PCA is an unsupervised method that generates a new set of unrelated variables (PCs) that represent the most variation in the data set (54). PCA was run using the "prcomp" function in R (stat.ethz.ch/R-manual/R-patched/library/stats/html/prcomp.html) to attribute the variance in the data to a reduced set of variables (PCs). For 3D-PCA, we projected the high-dimensional immunophenotype data onto the first three PCs and mapped each element into a 3D viewer using the R "pca3d" package (cran.r-project.org/web/packages/pca3d/index.html). For 2D-PCA we used the R package "ggplot2" (cran.r-project.org/web/packages/ggplot2/index.html) and mapped each element.

Variance/cluster composition bar graphs were plotted using the R package "ggplot2." The fractional cluster composition at each stage of IT was cal- culated as the number of cells in that stage of IT that were statistically de- termined to sort into each cluster (k-means) divided by the total number of cells, yielding a fraction between 0 and 1.

Phenotypic Shift Distance. Within each participant, we calculated the degree of phenotypic shift of all observed CD4+ dextramer+ peanut-specific T cells between progressive time points. Phenotypic shift is reported as the root- mean-square deviation (RMSD) (55) of all phenotypic markers between each cell at an IT time point to each cell observed at the following time point. All phenotypic shift distance calculations were performed in Perl (Version 3.20.1) (https://www.perl.org).

Statistics. To determine which markers were informative in determining the effect of IT on individual T cells, we used t tests to compare: cells from healthy controls against cells from pretreatment participants, cells from healthy controls against cells from IT participants (all IT time points), cells from pre- treatment participants against cells from IT participants (all IT time points), and dextramer+ cells vs. dextramer− cells. t tests were performed using R. Comparison of phenotypic shift distances by one-way analysis of variance (ANOVA) was performed with Prism software (GraphPad). Comparison of proportion of dextramer+ and proportion of dextramer− cells belonging to each cluster was done using $\chi 2$ tests in Excel. Comparison of number of dextramer+ cells at each IT time point was performed in Excel using t tests. Bonferroni-corrected P-value cutoffs were computed by setting a significance level of $\alpha = 0.05$ and dividing by the number of tests performed.

2.5.5 Acknowledgements

We thank Dr. William Kwok (Benaroya Research In- stitute) for providing the Ara h 2 tetramers, Stephan Haley and Tina Jakobsen (Immudex Corporation) for Ara h 2 dextramer construction, and Dr. Steve Quake, Dr. Wendy Davidson, Dr. Audrey Lau, and Dr. Marshall Plaut for manuscript review. We thank the Sean N. Parker Center for Allergy Research at Stanford University and the Myra Reinhard Foundation for their support. This work was funded by National Institutes of Health and National Institute of Allergy and Infectious Disease Grant U19AI104209.

2.5.6 References

1. Larché M, Akdis CA, Valenta R (2006) Immunological mechanisms of allergen-specific immunotherapy. Nat Rev Immunol 6(10):761 – 771.

2. Wambre E, et al. (2012) Differentiation stage determines pathologic and protective allergen-specific CD4 + T-cell outcomes during specific immunotherapy. J Allergy Clin Immunol 129(2):544 – 551, 551.e1 – 551.e7.

3. Akdis M, Akdis CA (2009) Therapeutic manipulation of immune tolerance in allergic disease. Nat Rev Drug Discov 8(8):645 – 660.

4. Bedoret D, et al. (2012) Changes in antigen-specific T-cell number and function during oral desensitization in cow ' s milk allergy enabled with omalizumab. Mucosal Immunol 5(3):267 – 276.

5. Burks AW, et al. (2011) NIAID-sponsored 2010 guidelines for managing food allergy: Applications in the pediatric population. Pediatrics 128(5):955 – 965.

6. Turcanu V, Maleki SJ, Lack G (2003) Characterization of lymphocyte responses to peanuts in normal children, peanut-allergic children, and allergic children who acquired tolerance to peanuts. J Clin Invest 111(7):1065 – 1072.

7. Syed A, et al. (2014) Peanut oral immunotherapy results in increased antigen-induced regulatory T-cell function and hypomethylation of forkhead box protein 3 (FOXP3). J Allergy Clin Immunol 133(2):500 – 510.

8. Herold KC, Bluestone JA (2011) Type 1 diabetes immunotherapy: Is the glass half empty or half full? Sci Transl Med 3(95):95fs1.

9. Francis JN, Till SJ, Durham SR (2003) Induction of IL-10 + CD4 + CD25 + T cells by grass pollen immunotherapy. J Allergy Clin Immunol 111(6):1255 – 1261.

10. Vickery BP, et al. (2014) Sustained unresponsiveness to peanut in subjects who have completed peanut oral immunotherapy. J Allergy Clin Immunol 133(2):468 – 475.

11. Tang ML, et al. (2015) Administration of a probiotic with peanut oral immunother- apy: A randomized trial. J Allergy Clin Immunol 135(3):737 – 44.e8.

12. Sewell WAC, Dore PC (2004) Premature testing of allergen-specific IgE post-anaphy- laxis may cause false negative results. J Allergy Clin Immunol 113(2):S241.

13. Jones SM, et al. (2009) Clinical efficacy and immune regulation with peanut oral immunotherapy. J Allergy Clin Immunol 124(2):292 – 300, 300.e1 – 300.e97.

14. Brinkmann V, Heusser CH (1993) T cell-dependent differentiation of human B cells into IgM, IgG, IgA, or IgE plasma cells: High rate of antibody production by IgE plasma cells, but limited clonal expansion of IgE precursors. Cell Immunol 152(2):323 – 332.

15. Stapel SO, et al.; EAACI Task Force (2008) Testing for IgG4 against foods is not rec- ommended as a diagnostic tool: EAACI Task Force Report. Allergy 63(7):793 – 796.

16. Vasquez-Ortiz M, et al. (2014) Ovalbumin-specific IgE/IgG4 ratio might improve the prediction of cooked and uncooked egg tolerance development in egg-allergic chil- dren. Clin Exp Allergy 44:579 – 588.

17. van Neerven RJ, et al. (1999) Blocking antibodies induced by specific allergy vacci- nation prevent the activation of CD4 + T cells by inhibiting serum-IgE-facilitated al- lergen presentation. J Immunol 163(5):2944 – 2952.

18. Wambre E, James EA, Kwok WW (2012) Characterization of CD4 + T cell subsets in allergy. Curr Opin Immunol 24(6):700 – 706.

19. Standifer NE, Burwell EA, Gersuk VH, Greenbaum CJ, Nepom GT (2009) Changes in autoreactive T cell avidity during type 1 diabetes development. Clin Immunol 132(3): 312 – 320.

20. Burton BR, et al. (2014) Sequential transcriptional changes dictate safe and effective antigen-specific immunotherapy. Nat Commun 5:4741.

21. Han A, Glanville J, Hansmann L, Davis MM (2014) Linking T-cell receptor sequence to functional phenotype at the single-cell level. Nat Biotechnol 32(7):684 – 692.

22. Hong JW, Quake SR (2003) Integrated nanoliter systems. Nat Biotechnol 21(10): 1179 – 1183.

23. Burks AW, et al. (1992) Identification and characterization of a second major peanut allergen, Ara h II, with use of the sera of patients with atopic dermatitis and positive peanut challenge. J Allergy Clin Immunol 90(6 Pt 1):962 – 969.

24. Maecker HT, et al. (2012) New tools for classification and monitoring of autoimmune diseases. Nat Rev Rheumatol 8(6):317 – 328.

25. Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Series B Stat Methodol 63(2):411 – 423.

26. DeLong JH, et al. (2011) Ara h 1-reactive T cells in individuals with peanut allergy. J Allergy Clin Immunol 127(5):1211 – 8.e3.

27. Pellerin L, Jenks JA, Bégin P, Bacchetta R, Nadeau KC (2014) Regulatory T cells and their roles in immune dysregulation and allergy. Immunol Res 58(2-3):358 – 368.

28. Lechner O, et al. (2001) Fingerprints of anergic T cells. Curr Biol 11(8):587 – 595.

29. Aslam A, Chan H, Warrell DA, Misbah S, Ogg GS (2010) Tracking antigen-specific T-cells during clinical tolerance induction in humans. PLoS One 5(6):e11028.

30. Szabo SJ, Sullivan BM, Peng SL, Glimcher LH (2003) Molecular mechanisms regulating Th1 immune responses. Annu Rev Immunol 21:713 – 758.

31. Steinke JW, Lawrence MG (2014) T-cell biology in immunotherapy. Ann Allergy Asthma Immunol 112(3):195 – 199.

32. Tran DQ (2012) TGF- β : The sword, the wand, and the shield of FOXP3(+) regulatory T cells. J Mol Cell Biol 4(1):29 – 37.

33. Gorelik L, Fields PE, Flavell RA (2000) Cutting edge: TGF- β inhibits Th type 2 devel- opment through inhibition of GATA-3 expression. J Immunol 165(9):4773 – 4777. 34. Komine O, et al. (2003) The Runx1 transcription factor inhibits the differentiation of naive CD4 + T cells into the Th2 lineage by repressing GATA3 expression. J Exp Med 198(1):51 – 61.

35. Shalek AK, et al. (2014) Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature 510(7505):363 – 369.

36. Schleifman EB, et al. (2014) Targeted biomarker profiling of matched primary and metastatic estrogen receptor positive breast cancers. PLoS One 9(2):e88401.

37. Pollen AA, et al. (2014) Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat Biotechnol 32(10):1053 – 1058.

38. Anderson RP, Degano P, Godkin AJ, Jewell DP, Hill AV (2000) In vivo antigen challenge in celiac disease identifies a single transglutaminase-modified peptide as the dominant A-gliadin T-cell epitope. Nat Med 6(3):337 – 342.

39. Akdis M, Akdis CA (2014) Mechanisms of allergen-specific immunotherapy: Multiple suppressor factors at work in immune tolerance to allergens. J Allergy Clin Immunol 133(3):621 – 631.

40. Fathman CG, Lineberry NB (2007) Molecular mechanisms of CD4 + T-cell anergy. Nat Rev Immunol 7(8):599 – 609.

41. Rudd CE, Taylor A, Schneider H (2009) CD28 and CTLA-4 coreceptor expression and signal transduction. Immunol Rev 229(1):12 – 26.

42. Nishikawa H, Sakaguchi S (2014) Regulatory T cells in cancer immunotherapy. Curr Opin Immunol 27:1 – 7.

43. Walker LS (2013) Treg and CTLA-4: Two intertwining pathways to immune tolerance. J Autoimmun 45:49 – 57.

44. Dalai SK, Mirshahidi S, Morrot A, Zavala F, Sadegh-Nasseri S (2008) Anergy in memory CD4 + T cells is induced by B cells. J Immunol 181(5):3221 – 3231.

45. David A, et al. (2014) Tolerance induction in memory CD4 T cells requires two rounds of antigen-specific activation. Proc Natl Acad Sci USA 111(21):7735 – 7740.

46. Scadding GW, et al. (2010) Sublingual grass pollen immunotherapy is associated with increases in sublingual Foxp3-expressing cells and elevated allergen-specific immu- noglobulin G4, immunoglobulin A and serum inhibitory activity for immunoglobulin E-facilitated allergen binding to B cells. Clin Exp Allergy 40(4):598 – 606.

47. Hjorth M, et al. (2011) GAD-alum treatment induces GAD65-specific CD4 + CD25highFOXP3 + cells in type 1 diabetic patients. Clin Immunol 138(1):117 – 126.

48. Suárez-Fueyo A, et al. (2014) Grass tablet sublingual immunotherapy down-regulates the TH2 cytokine response followed by regulatory T-cell generation. J Allergy Clin Immunol 133(1):130 – 8.e1, 2.

49. Loo EW, Krantz MJ, Agrawal B (2012) High dose antigen treatment with a peptide epitope of myelin basic protein modulates T cells in multiple sclerosis patients. Cell Immunol 280(1):10 – 15.

50. Savilahti EM, Savilahti E (2013) Development of natural tolerance and induced de- sensitization in cow ' s milk allergy. Pediatr Allergy Immunol 24(2):114 – 121.

51. Santos AF, et al. (2014) Basophil activation test discriminates between allergy and tolerance in peanut-sensitized children. J Allergy Clin Immunol 134(3):645 – 652.

52. Gernez Y, et al. (2011) Basophil CD203c levels are increased at baseline and can be used to monitor omalizumab treatment in subjects with nut allergy. Int Arch Allergy Immunol 154(4):318 – 327.

53. Do JH, Choi DK (2008) Clustering approaches to identifying gene expression patterns from DNA microarray data. Mol Cells 25(2):279 – 288.

54. Ringnér M (2008) What is principal component analysis? Nat Biotechnol 26(3):303 – 304. 55. Coutsias EA, Seok C, Dill KA (2004) Using quaternions to calculate RMSD. J Comput Chem 25(15):1849 – 1857

2.5.7 Copyright

Chapter 3

Heredity, Environment and Receptor Convergence

With an established ability to identify clones in Chapter 1, and then identify convergence groups of related clones that perform the same function in Chapter 2, it becomes of considerable interest to establish whether polymorphim in human populations at the V,D,J and C loci of the BCR and TCR repertoires could influence the formation of clones and convergence groups. One example, the genotype-specific formation of broadly neutralizing IGHV1-69 influenza stem binders, has been discussed in Chapter 2.4. Here we expand on that analysis into additional pathogens, additional resources for characterization of allele polymorphism at these complex loci, and strategies for integrating polymorphism data into functional response analysis and the search for correlates of protection.

3.1 Introduction

Antibodies (Abs) produced by immunoglobulin (IG) genes are the most diverse proteins expressed in humans. While part of this diversity is generated by recombination during B-cell development and mutations during affinity maturation, the germ-line

IG loci are also diverse across human populations and ethnicities. Recently, proof-of-concept studies have demonstrated genotype– phenotype correlations between specific IG germ-line variants and the quality of Ab responses during vaccination and disease. However, the functional consequences of IG genetic variation in Ab function and immunological outcomes remain underexplored. In this opinion article, we outline interconnections between IG genomic diversity and Ab-expressed repertoires and structure. We further propose a strategy for integrating IG genotyping with functional Ab profiling data as a means to better predict and optimize humoral responses in genetically diverse human populations, with immediate implications for personalized medicine.

3.1.1 The molecular basis for antibody diversity

Antibodies (Abs) have long been appreciated as key constituents of the adaptive immune response. Their function is to enable selective recognition and mediate immune responses to novel foreign antigens. This is accomplished through the somatic generation of vast repertoires of hundreds of millions of unique Ab receptors that can be selected, matured, and ultimately participate in the formation of long-term memory during B-cell development and activation. As a consequence of this diversity, even after nearly a century of research, the complexity of the Ab response within and between individuals is only beginning to be delineated at the molecular and genetic levels.

Hundreds of variable (V) and dozens of diversity (D) and joining (J) immunoglobulin (IG) germ-line gene segments across three primary loci in the human genome comprise the necessary building blocks of the expressed Ab heavyand light-chain repertoires [1]. Whereas the heavy chain is encoded by genes at the IG heavy-chain locus (IGH), the light chain can be encoded by genes at either the IG kappa (IGK) or IG lambda (IGL) chain loci [1]. The naïve Ab repertoire is formed by assembling variants of these building blocks using a specialized V(D)J recombination process that somatically joins various V, D, and J segments (or V and J at IGK and IGL). The introduction and deletion of P and N nucleotides at V(D)J junctions and the

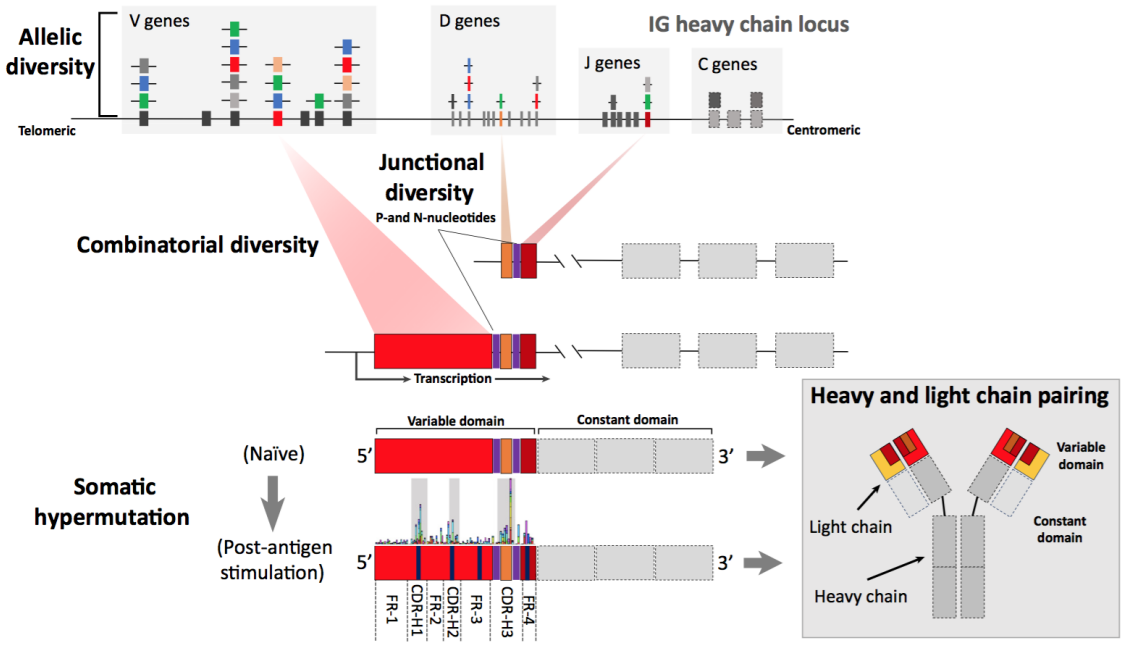


Figure 3.1: Sources of antibody diversity. The germline IGH locus, consists of tandemly arrayed IGH V, D, J, and constant (C) gene segments. For a subset of these segments, multiple alleles are shown, representing population-level allelic diversity. During the initial formation of the naïve repertoire, single V, D and J gene segments on one of two chromosomes in a given B cell are somatically recombined; at each of these steps, P and N nucleotides are added at the D–J and V–D junctions ('junctional diversity'), respectively. This process, known as V(D)J rearrangement, is the basis for 'combinatorial diversity'. Two identical heavy chains and two identical light chains are ultimately paired through disulfide bonds to form a functional Ab; thus, additional diversity in the expressed Ab repertoire comes from 'heavyand light-chain pairing'. Following Ag stimulation, 'somatic hypermutations' introduce additional variation in the variable domain of the Ab (vertical purple bars), with the aim of improving binding affinity. Mutations that arise via SHM can occur across all FRs and CDRs, but these are most prevalent in CDRs.

pairing of different heavy and light chains dramatically increase diversity (Figure 1) [2]. Considering these processes alone, a given baseline or primary naïve repertoire can theoretically sample from different Abs [3]. The extraordinary diversity of the naïve repertoire ensures that it will likely contain a naïve Ab with at least weak initial binding against a vast array of antigens.

Even so, this impressive baseline diversity can be subsequently augmented when a B cell encounters and is stimulated by an antigen to undergo somatic hypermutation (SHM; Figure 1), resulting in lineages of tens of thousands of clonally derived affinity maturation variants of the initial Ab. Specifically, SHM introduces somatic mutations throughout the variable portion of the Ab, including targeted hotspots residing within the antigen-contacting hypervariable complementarity-determining regions (CDRs). This process ultimately increases the affinity and specificity of the Ab for binding the target epitope, facilitating a highly focused antigen-specific response.

While the prevailing paradigm for investigating B-cell and Ab-mediated responses has placed emphasis on the importance of the unique molecular mechanisms cited earlier in the generation of key functional Abs, there is a growing appreciation for the fact that IG genes are highly variable at the germ-line level, exhibiting extreme allelic polymorphism and gene copy number variation (CNV) between individuals and across populations [4–9]. Recent studies have begun to highlight that, in addition to diversity introduced during V(D)J recombination, heavyand lightchain pairing, and SHM, IG germ-line variation (e.g., allelic variation; Figure 1) plays a vital part in determining the development of the naïve repertoire, with downstream impacts on signatures observed in the memory compartment, and the capacity of an individual to mount an Ab response to specific epitopes [10–16].

3.1.2 IgH haplotype diversity in human populations

Recent genomic sequencing indicates that IG loci, specifically IGH, may be among the most polymorphic in the human genome [17]. Across IGH, IGK, and IGL, there are currently >420 alleles cataloged in the ImMunoGeneTics information system database

(IMGT) [18–21] that have been described from germ-line DNA in the human population, with an enrichment of nonsynonymous variants (Table 1). Although the validity of some alleles in IMGT has been called into question [22], the number of polymorphic alleles continues to grow [11,23,24], especially as IG gene sequencing is conducted in increasing numbers of non-Caucasian samples [7,9,25]. A recent study conducted in 28 indigenous South Africans identified 122 non-IMGT IGHV alleles [9]. In addition to IG allelic variation and single nucleotide polymorphisms (SNPs), CNVs, including large deletions, insertions, and duplications ($8–75 Kb in length), are also prevalent in IG regions (Table 1). Using IGH as an example, up to 29 of the 58 functional/open reading frame (ORF) IGHV genes may vary in genomic copy number CNVs of IGH D (diversity) and constant (C) region genes are also known [11,12,29]. Until recently, primarily due to technical difficulties associated with the complex genomic architecture of the IG loci, none of the known CNVs in IGHV had been sequenced at nucleotide resolution [7]; many likely remain undescribed at the genomic level.

The high prevalence of IG allelic and locus structural diversity translates into extreme levels of inter-individual haplotype variation [4–7]. For example, recent comparisons of the two available completed assemblies for the IGHV gene region ($1 Mb in length) revealed that two human chromosomes can vary by >100 Kb of sequence, with >2,800 SNPs, and CNVs of 10 IGHV functional/ORF genes [7,17]. In population sequencing experiments, extreme examples of heterozygosity have been noted, with evidence of some individuals carrying more than one allele at every IGHV coding gene [9]. Supporting earlier genetic mapping data [4,5], more recent analysis of inferred haplotypes from Ab repertoire data surveyed in nine individuals revealed that all 18 haplotypes characterized were unique [6]. Furthermore, at the population level, of the few SNPs and CNVs screened within IGH, allele and genotype frequencies have been shown to vary considerably between ethnic backgrounds [7–9,15], with evidence of selection [7]. Despite the evidence for elevated germ-line diversity, genomic resources for IG loci continue to lag behind other regions of the genome [26]. Because of this, the comprehensive and accurate genotyping of IG polymorphisms remains a significant challenge [26,30], and as a result, the full extent of IG polymorphism and the

implications for human health are yet to be uncovered [26]. However, it is plausible that population-level diversity in the IG loci, particularly in IGH, will rival that of other complex immune gene families, such as the human leukocyte antigen (HLA) and killer cell IG-like receptor (KIR) genes. These genes are also characterized by extreme haplotype diversity, due to CNV and coding region variation [31,32]; HLA genes, for example, have thousands of known alleles [31]. In contrast to IG genes, HLA and KIR have been studied more extensively across human populations, and have demonstrated critical roles in disease [31,32].

3.1.3 Germline influence on the expressed antibody repertoire

Our limited knowledge of IG population diversity has hindered our ability to comprehensively test for direct connections between IG germ-line polymorphisms, variation in the repertoire generated after recombination, amino acid variation in the Ab produced, and ultimately Ab function. Advances in high-throughput sequencing technology now enable extensive characterization of the expressed Ab repertoire [33–35], creating opportunities for beginning to investigate the heritability of the Ab response at fine-scale resolution. Applications of these methods, collectively referred to as repertoire sequencing ('IgSeq' or 'RepSeq'), have already led to a wealth of new discoveries in a range of contexts [33,36]. These include general observations that key features of the Ab repertoire show extensive variability between healthy individuals [10,11,13,14,37], and a limited overlap of B-cell receptor clones between individuals, even monozygotic (MZ) twins [10,13,14]. However, RepSeq studies have also revealed that these inter-individual differences are not necessarily random, but likely have a strong underlying genetic component, providing initial support for the importance of germ-line IG polymorphism in determining the naïve and Ag-stimulated Ab repertoire. For example, several recent studies have revealed that V, D, and J gene usage in the naïve repertoire is much more highly correlated between MZ twins than between unrelated individuals [10,13,14], and that IG gene usage patterns are consistent across time points within a given individual [38]. A role for genetic factors can be seen for other repertoire features in twins as well, including the degree of SHM [13],

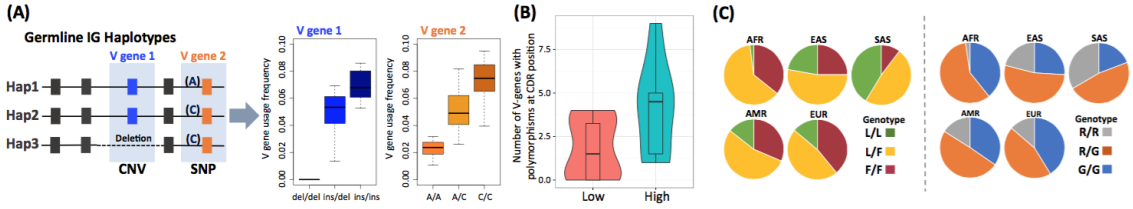


Figure 3.2: Impacts of IG Germ-Line Polymorphism on Ab Repertoire/Structural Diversity. (A) Hypothetical examples of associations between IG gene region CNV (V gene 1 insertion/deletion) and SNP (noncoding regulatory variant, A/C) genotypes and V gene usage frequencies in the expressed Ab repertoire. (B) Violin plot showing nonsynonymous polymorphism rates in CDR positions with high (>0.6; 'high', blue) or low (<0.25; 'low', red) frequency of contact with antigen. The Y axis records, for each CDR-H1 and CDR-H2 position, the number of IMGT IGHV genes that have alleles with nonsynonymous polymorphisms at that position. The positional probability of antigen contact was calculated for each CDR position as the percentage of 150 crystal structures of antibody–antigen complexes from the protein database (PDB) where any atom of that residue is within 5 Å of any antigen atom. Allelic variation is enriched in antigencontact sites, in that the number of IGHV genes with alleles containing nonsynonymous polymorphisms is greater for high contact probability positions. (C) Genotype frequency differences between five human ethnic groups [Africans (AFR); East Asians (EAS); South Asians (SAS); Central/South American (AMR); and Europeans (EUR)], published by the 1000 Genomes Project [80]*, at two SNPs in IGHV1-69 that have been shown to encode functional residues critical for neutralizing Abs against the influenza HA stem (F54 and L54 amino acid-associated alleles; SNP rs55891010; left panel), and 'NEAT2' domain of Staphylococcus aureus (R50 and G50 alleles; SNP rs11845244; right panel). In the left panel, the F allele encodes the functional critical phenylalanine residue, and in the right panel, the primary glycine residue is encoded by the G allele. Interestingly, in both cases, the frequency of individuals lacking alleles encoding the critical residues varies among populations, with the L/L and R/R genotypes showing the lowest frequencies in Africans, and the highest frequencies in South Asians. rs55891010 and rs11845244 are in linkage disequilibrium, and thus R50 and L54 amino acids (and likewise, G50 and F54) tend to co-occur in alleles of IGHV1-69. This explains similarities in genotype frequency estimates between the two SNPs in each population.

and the distribution of CDR-H3 length and clone convergence [10,13,14]. Intriguingly, although existing data suggest that features in the memory compartment are more stochastic, likely reflective of random recruitment and transient proliferation, certain genes and repertoire features exhibit predictable patterns even in memory B cells [10,13,14,39].

Studies of repertoire heritability are consistent with a number of examples for which germ-line IG polymorphisms have been explicitly linked to features in the expressed Ab repertoire [12,15,40–42] (see Figure IA in Box 1 for hypothetical examples of IG genotype effects on the repertoire). Sasso et al. [40] reported the first direct connection to IG genotype, reporting that CNV of IGHV1-69 was tightly correlated with its relative usage in tonsillar B cells. Our own work has also demonstrated this relationship, but uncovered associations for IGHV1-69 coding and potentially noncoding polymorphism as well as CNV [15]. Inferred deletions of IGHD genes have also been shown to associate with variation in D–J pairing frequencies, demonstrating that germ-line effects on the repertoire extend beyond V genes [12]. An interesting aspect of IGH CNVs is that, in addition to observed effects of these variants on the genes within the CNV event, they also can impact the usage of genes elsewhere in the locus [12,15]. For example, we recently observed apparent long-range effects of IGHV1-69 CNV in the naïve and memory repertoire, in that individuals with fewer IGHV1-69 germ-line copies and reduced usage showed consistently higher usage of IGHV genes over 200 Kb away [15]. The mechanisms underlying the observed effects of CNVs in human IG loci remain technically difficult to assess experimentally, but it has been speculated that these large changes in locus architecture (i.e., deletions and insertions) could alter regulatory systems related to V(D)J recombination [12,15], for example, by modifying the chromatin landscape, cis-regulatory elements and transcription factor binding, and/or the physical locations of the IG V, D, and J genes. All of these factors are known to be key determinants of IG gene accessibility and usage frequencies in mice [43,44].

A role for noncoding polymorphisms is also strongly supported by early work conducted in the human IGK region which directly showed that a variant associated with Haemophilus influenzae infection susceptibility in the recombination signal

sequence (RSS) of IGKV2-29 significantly decreased gene rearrangement frequency [42]. RSSs, which are critical for the recruitment of RAG1/2 proteins, have also been demonstrated to impact IGHV gene usage in mice [43,44]. Moreover, extensive work in the murine IG gene loci has uncovered important roles for other key cis-regulatory sequences and transcription factors as well [45,46]. Such analyses have not yet been comprehensively conducted in humans, and as a result, our knowledge of the IG regulatory elements involved in the formation of the expressed Ab repertoire is restricted to canonical RSS, promoter, enhancer elements, and class switch regions. However, even for these well-known noncoding regulatory regions, limited data on human population-level variation exist, and thus the broader consequences of polymorphism in these elements on Ab repertoire variability have not been explored.

Although direct links between repertoire variability and human IG CNVs and noncoding polymorphisms remain limited to the few examples discussed above, additional evidence from expressed Ab repertoire studies in unrelated individuals also highlights the clear potential for these variants to have pervasive impacts on Ab repertoire features, particularly gene usage in the naïve compartment. Most demonstrable is the fact that many of the genes with the most variability in naïve repertoire usage across individuals are also known to be in CNV, including examples of the complete absence of genes in the expressed Ab repertoires of some donors [6,10–12]. In addition, allele-specific usage in the naïve Ab repertoires of individuals heterozygous at a given IGHV gene has been demonstrated, also clearly suggesting a role for noncoding variation and CNV [11]. Moreover, although effects of germ-line IG polymorphism may be most evident on a per gene basis, it is worth noting that findings from MZ twins demonstrated that certain CDR-H3 features are highly heritable [13,14]. This indicates that even strong genetically determined biases on individual V, D, and J gene usage [and thus their nonrandom combination during V(D)J rearrangement] could also be directly linked to variation observed within CDR-H3. This is an important point given that CDR-H3 variation has classically been considered independent of the germ line [13,14].

In addition to effects of IG polymorphism on gene usage, functional CDR variants can also be directly encoded in the genome. For example, across the $267 coding

alleles cataloged in IMGT for functional and ORF IGHV genes, 60% of the 382 polymorphisms are nonsynonymous (Table 1), including sites located in CDR-H1 and CDR-H2 with predicted relevance to Ab functional residue diversity (see Figure IB in Box 1). Although the CDR-H3 loop, formed at the V (D)J junction, is the most diverse region of an Ab and is a principal determinant of specificity [47,48], there is a growing appreciation for the importance of residues outside of CDR-H3 in antigen recognition and binding [15,49–51]. For example, recent analyses have shown that the median length of CDR-H2, which is solely encoded by germ-line V gene sequence, is substantially longer than that of CDR-H3, and typically forms the same number of interactions with antigen [52]. Specifically, analyses of antigen-binding region (ABRs; which roughly correspond to CDRs, but differ slightly in their boundaries) have shown that Abs contain a median of six, six, and four contact residues in the heavy-chain CDR-H3, H2, and H1 ABR regions, respectively. In addition, the overall percentage of energetically important Ag-binding residues within each ABR follows the same rank order, with $31%, 23%, and 14% for H3, H2, and H1, respectively. Similar trends were noted for light-chain ABRs as well [52]. In addition, considering that many known nonsynonymous sites reside outside of CDRs (Table 1), it is worth highlighting the fact that there are also examples demonstrating indirect effects of framework region variants on Ag binding [53,54].

3.1.4 Shared antibody signatures across individuals

A critical question is whether the germ-line effects on the repertoire outlined above can also partially account for inter-individual variation of the Ab-mediated response in disease and clinical phenotypes. The initial observation from RepSeq studies that essentially no Ab clones were shared among individuals, including MZ twins, posed a challenge to comparative Ab repertoire analysis: how could correlates of protection be identified in the Ab repertoire if every individual was responding with different Abs? However, an answer began to emerge with the observation that in multiple settings, including viral and bacterial infection, different individuals have been shown to respond to a given antigen with Abs that share convergent amino acid signatures

Color:	Light purple	Royal blue	Light blue	Brown	Pink	Dark purple	Green
Clusters:	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7
	Nonspecific	"IFNγ-Expressing"	Nonallergic	Allergic	Regulatory	"Anergic memory"	"IL-10-Expressing"
Antigen-specific:	Negative	Negative	Positive	Positive	Positive	Both	Both
Markers							
CD69	0	0	0	↑↑ increase	↑ increase	**↓↓ decrease**	0
CD38	0	0	0	**↓↓↓ decrease**	↑ increase	**↓↓↓↓↓ decrease**	decrease
CD28	0	0	0	**↓↓↓↓↓↓↓↓↓ decrease**	↑ increase	**↓↓↓↓↓↓↓↓↓↓ decrease**	decrease
CD45RA	0	0	0	**↑↑ increase**	↑ increase	**↓↓↓↓↓ decrease**	0
CD25	↑ increase	↑↑ increase	↑ increase	**↑ increase**	**↑↑↑↑ increase**	0	↑ increase
FoxP3	0	0	0	0	**↑↑ increase**	0	0
IL-4	0	decrease	0	**↑↑↑ increase**	↑ increase	decrease	decrease
IL-13	decrease	decrease	**↓ decrease**	**↑↑ increase**	↑ increase	↓ decrease	decrease
CCR7	↑ increase	↑↑ increase	↑ increase	↑↑increase	**↑↑↑ increase**	0	↑↑ increase
ITGA4	↑ increase	↑↑ increase	↑ increase	**↑↑ increase**	**↑↑ increase**	0	↑ increase
IL-10	negative	↑↑ increase	negative	0	**↑↑↑ increase**	decrease	**↑↑↑ increase**
IFN-γ	negative	**↑↑↑↑↑ increase**	negative	0	↑↑↑ increase	decrease	decrease
CD27	↑ increase	↑ increase	**↑ increase**	↓ decrease	↑↑ increase	0	↑ increase
IL-5	0	↑ increase	0	↑ increase	↑↑ increase	↑ increase	↑ increase
CCR8	0	0	0	decrease	↑↑ increase	decrease	0
CD127	0	↑↑ increase	↑↑ increase	decrease	↑↑ increase	0	↑↑ increase
CD90	0	0	0	0	↑↑↑ increased	0	0

Notes: (*i*) All expression is relative to 18S housekeeping control, 0 = no change. (*ii*) Most distinguishing markers for each cluster phenotype (shaded/bold). (*iii*) ↑ or ↓ indicate relative magnitude of marker expression increase or decrease. (*iv*) Seventeen out of 22 markers shown here because these were the markers significantly associated with cell clusters.

Table 3.1: Watson Glanville Trends in Immunology 2017 Table1.

[13,49,54–58]. These convergent Abs are often encoded by common V genes or sets of V genes, and specific amino acid residues in their CDRs enable them to converge upon a common binding solution against a shared antigen. Critically, in some cases evaluated, convergent signatures include amino acid residues that are directly encoded in the germ line. The occurrence of such convergent Ab responses highlights the potential for tracking common immune responses across individuals, and understanding the role of genetic factors, even when each individual creates unique Abs. Importantly, the implications of this line of thinking could be broad, as IG gene biases have been observed in contexts other than infection, including autoimmunity and cancer [59,60]. Moreover, IG gene biases may also extend to usage patterns of D and J genes, light-chain genes, and heavyand light-chain V gene pairing frequencies [56,61,62].

3.1.5 Polymorphism enriched in antigen contact sites

There are now many instances for which functional contributions of biased IG genes have been traced back to specific germ-line-encoded residues, including sites that are

polymorphic in the human population [15,16,50,53–55,63–65]. These examples illuminate a direct role of the IG germ line in disease-associated Ab responses. In the case of stem-directed broadly neutralizing Abs (BnAbs) against influenza hemagglutinin (HA), the most prevalent Abs use the heavy-chain gene IGHV1-69 [66–70]. These IGHV1-69 BnAbs recognize an overlapping epitope of group 1 influenza A viruses and only amino acids from IGHV make contact with HA. Importantly, of the 14 known alleles at IGHV1-69, only those encoding a critical phenylalanine at position 54 (F54) within CDR-H2 have a major role in shaping the BnAbs response [16,15,55,71]. Although IGHV1-69 F54-encoding alleles are dominant, there is a growing list of additional HA-directed BnAbs that also show IG germ-line biases [51,56,72–74], including those also known to be polymorphic with respect to coding variants and CNVs.

Interestingly, there are additional instances of biased IGHV1-69 allele usage in other disease contexts, with both overlapping and contrasting patterns to that observed for influenza. For example, F54 alleles are predominantly observed in IGHV1-69-expressing B cells associated with chronic lymphoid leukemia (CLL), whereas alleles encoding a leucine (L54) at this position are primarily used by non-neutralizing anti-gp41 Abs in HIV-1 [63,64]. Moreover, it has been shown that IGHV1-69 F54 alleles, in comparison with L54 alleles, have lower usage in the memory B-cell pool [10,15]. This observation may be similar to trends noted for IGHV4-34, which is also significantly underrepresented in the memory compartment of healthy individuals [10], and presumes to reflect a selective pressure against autoreactive Abs [75,76].

Other polymorphic positions in the framework regions of IGHV1-69, in conjunction with CDRH2 54, have also recently been shown to influence Ab binding of Middle East respiratory syndrome coronavirus (MERS-CoV) [53] and the Staphylococcus aureus NEAr iron transporter 2 (NEAT2) domain [54]. In the example of NEAT2, neutralizing Abs encoded by IGHV1-69 alleles carrying an arginine (R) at position 50 in place of glycine (G) showed significantly reduced NEAT2 binding [54]. Interestingly, based on publicly available data, the frequencies of critical alleles within polymorphic positions of IGHV1-69 vary across populations (see Figure IC in Box 1).

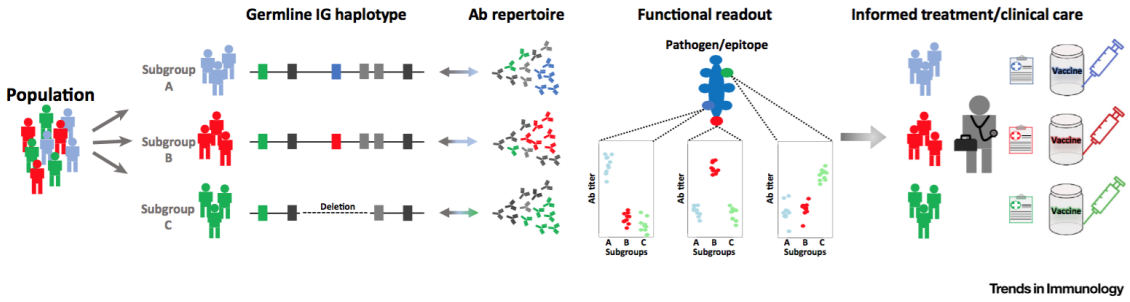


Figure 3.3: In the proposed paradigm, a population cohort is partitioned into subgroups based on functional genotypes/haplotypes that are directly associated with subgroup-specific signatures in the expressed repertoire and other relevant phenotypes (e.g., Ab titer; clinical outcome) associated with the Ab response to a given antigen/epitope. This partitioning can be used to inform tailored clinical care and treatment (e.g., vaccination regime). Ab, antibody.

3.1.6 Relating genotype, repertoire & outcomes

Considering the aforementioned evidence, we argue that the antigen-specific Ab repertoire is likely influenced by the host genotype. Although the genetic bases for repertoire and germ-line gene biases have not been comprehensively investigated, several recent studies provide a strategy for systematically integrating data on IG polymorphism and Ab responses at the population and molecular levels to provide unique insight into Ab signatures associated with disease.

We have begun to explore this idea in detail at the IGHV1-69 locus in the context of influenza vaccination [15]. Providing strong proof-of-concept, by initially focusing on observed IGHV1-69 allelic usage bias against a critical broadly neutralizing epitope, we genotyped the IGHV1-69 F54/L54 allele and copy number frequencies in a cohort of 85 H5N1 vaccines, including 18 individuals with accompanying Ab repertoire data [15]. Drawing directly on aspects of repertoire heritability reviewed above, we found robust connections between these polymorphisms and repertoire gene usage in both the unmutated IgM (naïve) and IgG memory repertoires, with IGHV1-69 germ-line gene usage increasing with the number of copies of F54 alleles. In addition to usage frequencies, IGHV1-69 genotype also associated with IGHV1-69 B-cell expansion, SHM, and Ig class switching. It is important to note that these genotype

effects extended to levels of circulating anti-HA stem BnAbs postvaccination, with individuals carrying only germline-encoded CDR-H2 L54 alleles having lower IGHV1-69 BnAbs. Furthermore, with direct repertoire sequencing, we were able to specifically demonstrate that only carriers of the IGHV169 F54 alleles expressed convergent anti-BnAb signatures. These results are bolstered by similar observations recently made by two other groups that also carried out IGHV1-69 F54/ L54 allele genotyping in their cohorts [16,55]. Altogether, these data demonstrate that genetically determined baseline differences in the Ab repertoire can set the stage for disease-related responses.

A crucial aspect of this story (which is expected to emerge in other cases as well) is that the frequency of IGHV1-69 F54 alleles and CNV varies considerably across populations [7,15]. Specifically, the number of individuals that would be predicted to lack the capacity to generate effective IGHV1-69 BnAbs was much higher in some populations. However, we and others have shown that individuals lacking IGHV1-69 F54 alleles likely utilize other germ-line genes in place of IGHV1-69 [51,55]. This finding in particular both highlights the complexity of the Ab response and demonstrates that the integration of genotyping information can help provide a more nuanced interpretation of the signatures discovered in the expressed repertoire. Moreover, it suggests that efforts should be made to study these complex responses in larger and more diverse cohorts, including individuals from presently understudied populations.

Building on findings in these initial studies [15,16,55], we propose a framework for integrating genotypic information into future studies of the Ab response in wellness and disease (Figure 2, Key Figure). The general strategy is as follows: (i) identify IG gene biases observed in a diseaserelated or epitope-specific response; (ii) characterize this response at the population level by performing comprehensive genotyping of coding, noncoding, and gene copy number variants at and around the locus of interest (and others if there is rationale); (iii) perform repertoire sequencing and analysis of the response in all relevant B-cell subsets to identify all Ab convergence groups with allele bias; and (iv) evaluate genotype–phenotype linkages of the functional Ab response and specific Ab convergence groups.

3.1.7 Conclusing remarks

We see a growing body of evidence to support the link between IG polymorphism and phenotype that may have important clinical applications (see Outstanding Questions). The most obvious of these correlations include potential effects of CNV and SNPs in non-translated and translated IG gene regions on expressed repertoire variability in naïve and memory B cell subsets. Some of these polymorphisms could be expected to more broadly impact variation in protective Ab responses [77] and quality of the memory B-cell pool. We anticipate that IG polymorphism will contribute to differences in expression of common (public) and unique (private) antibody signatures that are associated with protective responses in disease and in response to vaccination. We propose a model for the future in which cataloging these public signatures for biased gene use, V(D)J associations, SHMs, and heavy-light chain pairing in the context of IG germ-line variation should begin to provide us with information to advance our understanding of the immunogenetic potential of an individual's baseline naïve repertoire (Figure 2), particularly when more complete data sets of biased Ab signatures to specific epitopes become available. Based on existing genetic data, it is probable that similar IG haplotypes will associate with overlapping signatures in baseline repertoire profiles, even if not to the degree of repertoire similarity observed in MZ twins. This IG polymorphism, as we and others have begun to show, may further influence the evolution of antigen-experienced B cells and plasma cells, where other genetic polymorphisms in the IG loci and environmental exposures come into play in continuing to shape affinity, epitope specificity, and fate. In addition, class-switched memory B-cell compartments will vary over time [37], and could be quantitated in the type and size of clonotypes with both public and private signatures against immunodominant epitopes9.

Together, this knowledge should pave the way to using molecular and genetic signatures for mapping an individual's exposure history, current wellness state, and immune potential against future antigenic threats. For example, characterization of genotypes that specifically lead to common BnAb signatures in the repertoire should be useful for tailoring vaccines to responsive genotypes with the goal of achieving 100% 'universal vaccine' responsiveness at the population level (Figure 2). In addition, such

information could lead to advances in the use of anti-idiotypic antibody and chimeric antigen receptor T-cell therapies that are directed against germ-line gene expressing B-cell clonotypes that are directly involved in autoimmune disease and hematologic malignancies [78,79]. We face tall hurdles to moving this paradigm forward, the greatest being the completion of a comprehensive catalogue of human IG haplotype variation [26]. However, with ever expanding advances in immunologic and genomic technologies, we believe that such integrative approaches are within our reach, and have the potential to transform our understanding of Ab-mediated immune responses in the clinical and research arenas.

3.1.8 Acknowledgements

This work was supported by the Gates Foundation, the National Institute of Allergy & Infectious Disease of the US National Institutes of Health (NIH) under awards U01-AI074518, R56-AI109223, and R01-AI121285 to W.A.M.

3.1.9 References

1. Lefranc, M.-P. and Lefranc, G. (2001) The Immunoglobulin Facts- book, Academic Press

2. Tonegawa, S. (1983) Somatic generation of antibody diversity. Nature 302, 575–581

3. Schroeder, H.W. (2006) Similarity and divergence in the develop- ment and expression of the mouse and human antibody reper- toires. Dev. Comp. Immunol. 30, 119–135

4. Chimge, N.-O. et al. (2005) Determination of gene organization in the human IGHV region on single chromosomes. Genes Immun. 6, 186–193

5. Li, H. et al. (2002) Genetic diversity of the human immunoglobulin heavy chain VH region. Immunol. Rev. 190, 53–68

6. Kidd, M.J. et al. (2012) The inference of phased haplotypes for the immunoglobulin H chain V region gene loci by analysis of VDJ gene rearrangements. J. Immunol. 188, 1333–1340

7. Watson, C.T. et al. (2013) Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and join- ing genes and characterization of allelic and copy-number varia- tion. Am. J. Hum. Genet. 92, 530–546

8. Sasso, E.H. et al. (1995) Ethnic differences in polymorphism of an immunoglobulin VH3 gene. J. Clin. Invest. 96, 1591–1600

9. Scheepers, C. et al. (2015) Ability to develop broadly neutralizing HIV-1 antibodies is not restricted by the germline IG gene reper- toire. J. Immunol. 194, 4371–4378

10. Glanville, J. et al. (2011) Naive antibody gene-segment frequen- cies are heritable and unaltered by chronic lymphocyte ablation. Proc. Natl. Acad. Sci. U. S. A. 108, 20066–20071

11. Boyd, S.D. et al. (2010) Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. J. Immunol. 184, 6986–6992

12. Kidd, M.J. et al. (2015) DJ pairing during VDJ recombination shows positional biases that vary among individuals with differing IGHD locus immunogenotypes. J. Immunol. 196, 1158–1164

13. Wang, C. et al. (2015) B-cell repertoire responses to varicella- zoster vaccination in human identical twins. Proc. Natl. Acad. Sci. U. S. A. 112, 500–505

14. Rubelt, F. et al. (2016) Individual heritable differences result in unique lymphocyte receptor repertoires of naïve and antigen- experienced cells. Nat. Commun. 6, 1–12

15. Avnir, Y. et al. (2016) IGHV1-69 polymorphism modulates anti- influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. Sci. Rep. 6, 20842

16. Wheatley, A.K. et al. (2015) H5N1 vaccine-elicited memory B cells are genetically constrained by the IGHV locus in the recognition of a neutralizing epitope in the hemagglutinin stem. J. Immunol. 195, 602–610

17. Watson, C.T. et al. (2014) Sequencing of the human IG light chain loci from a hydatidiform mole BAC library reveals locus-specific signatures of genetic diversity. Genes Immun. 16, 24–34

18. Pallarès, N. et al. (1999) The human immunoglobulin heavy vari- able genes. Exp. Clin. Immunogenet. 16, 36–60

19. Lefranc, M.-P. et al. (2014) IMGT1, the international ImMunoGe- neTics information system1 25 years on. Nucleic Acids Res. 43, D413–D422

20. Pallarés, N. et al. (1998) The human immunoglobulin lambda variable (IGLV) genes and joining (IGLJ) segments. Exp. Clin. Immunogenet. 15, 8–18

21. Barbié, V. and Lefranc, M.P. (1998) The human immunoglobulin kappa variable (IGKV) genes and joining (IGKJ) segments. Exp. Clin. Immunogenet. 15, 171–183

22. Wang, Y. et al. (2008) Many human immunoglobulin heavy-chain IGHV gene polymorphisms have been reported in error. Immunol. Cell Biol. 86, 111–115

23. Gadala-Maria, D. et al. (2015) Automated analysis of high- throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. Proc. Natl. Acad. Sci. U. S. A. 112, E862–E870

24. Corcoran, M.M. et al. (2016) Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. Nat. Commun. 7, 13642

25. Wang, Y. et al. (2011) Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants. Immunogenetics 63, 259–265

26. Watson, C.T. and Breden, F. (2012) The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. Genes Immun. 13, 363–373

27. Milner, E.C. et al. (1995) Polymorphism and utilization of human VH genes. Ann. N. Y. Acad. Sci. 764, 50–61

28. Shin, E.K. et al. (1993) Polymorphism of the human immunoglob- ulin variable region segment V1-4.1. Immunogenetics 38, 304–306

29. Bottaro, A. et al. (1991) Pulsed-field electrophoresis screening for immunoglobulin heavy-chain constant-region (IGHC) multigene deletions and duplications. Am. J. Hum. Genet. 48, 745–756

30. Luo, S. et al. (2016) Estimating copy number and allelic variation at the immunoglobulin heavy chain locus using short reads. PLoS Comput. Biol. 12, 1–21

31. Trowsdale, J. and Knight, J.C. (2013) Major histocompatibility complex genomics and human disease. Annu. Rev. Genomics Hum. Genet. 14, 301–323

32. Parham, P. and Moffett, A. (2013) Variable NK cell receptors and their MHC class I ligands in immunity, reproduction and human evolution. Nat. Rev. Immunol. 13, 133–144

33. Georgiou, G. et al. (2014) The promise and challenge of high- throughput sequencing of the antibody repertoire. Nat. Biotech- nol. 32, 158–168

34. Boyd, S.D. and Joshi, S.A. (2014) High-throughput DNA sequencing analysis of antibody repertoires. Microbiol. Spectr. 2, 1–13

35. Yaari, G. and Kleinstein, S.H. (2015) Practical guidelines for B-cell receptor repertoire sequencing analysis. Genome Med. 7, 121

36. Jackson, K.J.L. et al. (2013) The shape of the lymphocyte recep- tor repertoire: lessons from the B cell receptor. Front. Immunol. 4, 1–12

37. Galson, J.D. et al. (2015) In-depth assessment of within-individual and inter-individual variation in the B cell receptor repertoire. Front. Immunol. 6, 1–13

38. Laserson, U. et al. (2014) High-resolution antibody dynamics of vaccine-induced immune responses. Proc. Natl. Acad. Sci. U. S. A. 111, 4928–4933

39. Vollmers, C. et al. (2013) Genetic measurement of memory B-cell recall using antibody repertoire sequencing. Proc. Natl. Acad. Sci. U. S. A. 110, 13463–13468

40. Sasso, E.H. et al. (1996) Expression of the immunoglobulin VH gene 51p1 is proportional to its germline gene copy number. J. Clin. Invest. 97, 2074–2080

41. Sharon, E. et al. (2016) Genetic variation in MHC proteins is associated with T cell receptor expression biases. Nat. Genet. 48, 995–1002

42. Feeney, A.J. et al. (1996) A defective Vkappa A2 allele in Navajos which may play a role in increased susceptibility to Haemophilus influenzae type b disease. J. Clin. Invest. 97, 2277–2282

43. Feeney, A.J. (2009) Genetic and epigenetic control of V gene rearrangement frequency. Adv. Exp. Med. Biol. 650, 73–81

44. Choi, N.M. et al. (2013) Deep sequencing of the murine IgH repertoire reveals complex regulation of nonrandom V gene rear- rangement frequencies. J. Immunol. 191, 2393–2402

45. Volpi, S.A. et al. (2012) Germline deletion of Igh 30 regulatory region elements hs 5, 6, 7 (hs5-7) affects B cell-specific regula- tion, rearrangement, and insulation of the Igh locus. J. Immunol. 188, 2556–2566

46. Verma-Gaur, J. et al. (2012) Noncoding transcription within the Igh distal VH region at PAIR elements affects the 3D structure of the Igh locus in pro-B cells. Proc. Natl. Acad. Sci. U. S. A. 109, 17004–17009

47. Xu, J.L. and Davis, M.M. (2000) Diversity in the CDR3 region of V H is sufficient for most antibody specificities. Immunity 13, 37–45

48. Mahon, C.M. et al. (2013) Comprehensive interrogation of a minimalist synthetic CDR-H3 library and its ability to generate antibodies with therapeutic potential. J. Mol. Biol. 425, 1712–1730

49. Thomson, C.A. et al. (2008) Germline V-genes sculpt the binding site of a family of antibodies neutralizing human cytomegalovirus. EMBO J. 27, 2592–2602

50. Bryson, S. et al. (2016) Structures of preferred human IgV genes- based protective antibodies identify how conserved residues contact diverse antigens and assign source of specificity to CDR3 loop variation. J. Immunol. 196, 4723–4730

51. Fu, Y. et al. (2016) A broadly neutralizing anti-influenza antibody reveals ongoing capacity of haemagglutinin-specific memory B cells to evolve. Nat. Commun. 7, 12780

52. Kunik, V. and Ofran, Y. (2013) The indistinguishability of epitopes from protein surface is explained by the distinct binding prefer- ences of each of the six antigen-binding loops. Protein Eng. Des. Sel. 26, 599–609

53. Ying, T. et al. (2015) Junctional and allele-specific residues are critical for MERS-CoV neutralization by an exceptionally potent germline-like antibody. Nat. Commun. 6, 8223

54. Yeung, Y.A. et al. (2016) Germline-encoded neutralization of a Staphylococcus aureus virulence factor by the human antibody repertoire. Nat. Commun. 7, 13376

55. Pappas, L. et al. (2014) Rapid development of broadly influenza neutralizing antibodies through redundant mutations. Nature 516, 418–422

56. Joyce, M.G. et al. (2016) Vaccine-induced antibodies that neu- tralize group 1 and group 2 influenza A viruses. Cell 166, 609–623

57. Parameswaran, P. et al. (2013) Article convergent antibody signatures in human dengue. Cell Host Microbe 13, 691–700

58. Strauli, N.B. and Hernandez, R.D. (2016) Statistical inference of a convergent antibody repertoire response to influenza vaccine. Genome Med. 8, 60

59. Johansen, J.N. et al. (2015) Intrathecal BCR transcriptome in multiple sclerosis versus other neuroinflammation: equally diverse and compartmentalized, but more mutated, biased and over- lapping with the proteome. Clin. Immunol. 160, 211–225

60. Bomben, R. et al. (2010) Expression of mutated IGHV3-23 genes in chronic lymphocytic leukemia identifies a disease subset with peculiar clinical and biological features. Clin. Cancer Res. 16, 620–628

61. Forconi, F. et al. (2013) The IGHV1-69/IGHJ3 recombinations of unmutated CLL are distinct from those of normal B cells. Blood 119, 2106–2109

62. Zhu, D. et al. (2013) Biased immunoglobulin light chain use in the Chlamydophila psittaci negative ocular adnexal marginal zone lymphomas. Am. J. Hematol 88, 379–384

63. Hwang, K.K. et al. (2014) IGHV1-69 B cell chronic lymphocytic leukemia antibodies cross-react with HIV-1 and hepatitis C virus antigens as well as intestinal commensal bacteria. PLoS One 9, e90725

64. Williams, W.B. et al. (2015) HIV-1 vaccines. Diversion of HIV-1 vaccine-induced immunity by gp41-microbiota cross-reactive antibodies. Science 349, aab1253

65. Liu, L. and Lucas, A.H. (2003) IGH V3-23*01 and its allele V3-23*03 differ in their capacity to form the canonical human antibody combining site specific for the capsular polysaccharide of Haemophilus influenzae type b. Immunogenetics 55, 336–338

66. Throsby, M. et al. (2008) Heterosubtypic neutralizing monoclonal antibodies cross-protective against H5N1 and H1N1 recovered from human IgM+ memory B cells. PLoS One 3, e3942

67. Wrammert, J. et al. (2011) Broadly cross-reactive antibodies dominate the human B cell response against 2009 pandemic H1N1 influenza virus infection. J. Exp. Med. 208, 181–193

68. Ekiert, D.C. et al. (2009) Antibody recognition of a highly conserved influenza virus epitope. Science 324, 246–251

69. Kashyap, A.K. et al. (2008) Combinatorial antibody libraries from survivors of the Turkish H5N1 avian influenza outbreak reveal virus neutralization strategies. Proc. Natl. Acad. Sci. U. S. A. 105, 5986–5991

70. Corti, D. et al. (2011) A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglu- tinins. Science 333, 850–856

71. Lingwood, D. et al. (2012) Structural and genetic basis for devel- opment of broadly neutralizing influenza antibodies. Nature 489, 566–570

72. Nakamura, G. et al. (2013) An in vivo human-plasmablast enrich- ment technique allows rapid identification of therapeutic influenza A antibodies. Cell Host Microbe 14, 93–103

73. Kallewaard, N.L. et al. (2016) Structure and function analysis of an antibody recognizing all influenza A subtypes. Cell 166, 596–608

74. Wu, Y. et al. (2015) A potent broad-spectrum protective human monoclonal antibody crosslinking two haemagglutinin monomers of influenza A virus. Nat. Commun. 6, 7708

75. Pugh-Bernard, A.E. (2001) Regulation of inherently autoreactive VH4-34 B cells in the maintenance of human B cell tolerance. J. Clin. Invest. 108, 1061–1070

76. Cappione, A.J. et al. (2004) Lupus IgG VH4.34 antibodies bind to a 220-kDa glycoform of CD45/B220 on the surface of human B lymphocytes. J. Immunol. 172, 4298–4307

77. Lee, J. et al. (2016) Molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination. Nat. Med. 22, 1456–1464

78. Fesnak, A.D. et al. (2016) Engineered T cells: the promise and challenges of cancer immunotherapy. Nat. Rev. Cancer 16, 566–581

79. Chang, D.K. et al. (2016) Humanized mouse G6 anti-idiotypic monoclonal antibody has therapeutic potential against IGHV1-69 germline gene-based B-CLL. MAbs 8, 787–798

80. Auton, A. et al. (2015) A global reference for human genetic variation. Nature 526, 68–74

3.1.10 Copyright

3.2 Quantifying heritability in the adaptive receptor repertoires

This study, produced in 2011 the year before I began the PhD program at Stanford, remains one of my favorites ever produced. Simple and elegant, it emerged as a consequence of a dare - could we, when blinded of the identity of sets of PBMC samples from pairs of monozygotic twins, distinguish A) repertoires from twins vs those of unrelated subjects, B) sample replicates within twin pairs, and C) identify the MS-affected subject in the discordant twin pair in the study. The results of this analysis provided strong guidance for my understanding of the natural antibody repertoire.

A diverse antibody repertoire is essential for an effective adaptive immune response to novel molecular surfaces. Although past studies have observed common patterns of V-segment use, as well as variation in V-segment use between individuals, the relative contributions to variance from genetics, disease, age, and environment have remained unclear. Using high-throughput sequence analysis of monozygotic twins, we show that variation in naive V H and D H segment use is strongly determined by an individual's germline genetic background. The inherited segment-use profiles are resilient to differential environmental exposure, disease processes, and chronic lymphocyte depletion therapy. Signatures of the inherited profiles were observed in class switched

germ-line use of each individual. However, despite heritable segment use, the rearranged complementarity-determining region-H3 repertoires remained highly specific to the individual. As it has been previously demonstrated that certain V-segments exhibit biased representation in autoimmunity, lymphoma, and viral infection, we anticipate our fi ndings may provide a unique mechanism for stratifying individual risk pro fi les in speci fi c diseases.

3.2.1 Introduction

Specific biases in the antibody repertoire have been found in many diseases, from viral infections to cancers to autoimmune disorders (1 – 15). Although it is possible that heritable variation in the composition of the antibody repertoire could alter inherent risk to speci fi c diseases, the diversity of the antibody repertoire has hindered direct characterization of heritable in fl uences.

Early twin studies provided some evidence of genetic variation affecting reactive titers from the antibody repertoire. Multiple studies observed both total Ig and antigen-speci fi c titers to be more correlated in monozygotic twins than dizygotic twins or unrelated individuals (16 – 18). In some cases of monozygotic twins discordant for autoimmune diseases, the healthy twin often shared high autoantibody reactive titers with their affected twin (16, 19, 20).

Early sequencing studies were able to identify some systematic biases in the antibody repertoire with limited sampling depth. The fi rst sequencing studies to characterize V(D)J diversi fi cation mechanisms identi fi ed the gene segment recombination process, but also implied a repertoire too diverse to exhaustively interrogate by traditional sequencing technologies (21). Complete characterization of V-segment loci established $\sim$ 50 V H ,40V $\varkappa$, and 30 V λ segments in an individual, with a number of allelic variants for the majority of segments (22 – 24). Evaluation of use across individuals revealed biased V-gene representation that preceded selection (25 – 27). Quantitative PCR of V-gene families showed family use largely stable over time, with fl uctuations in use correlated to antigen-speci fi c responses (28). In the

T-cell receptor (TCR) repertoire, TCRB-V use was more highly correlated in healthy monozygotic twins than unrelated individuals (29, 30).

Recent developments in high-throughput sequencing of antibody diversity have enabled direct analysis of repertoires from entire organisms (31, 32). Such high-throughput repertoire studies have suggested both stochastic and heritable mechanisms involved in generating the antibody repertoire (32 – 34). A highthroughput study of 12 human samples identi fi ed variation in Vsegment use between individuals (33). Preferential use of some alleles suggested a potential heritable mechanism for repertoire variation, but they represented minor contributors to the total repertoire. Without longitudinal studies, it has remained unclear whether the observed variation was because of genetics, differences in antigen exposure, or natural fl uctuation of V-gene use in the repertoire over time. Longitudinal studies of repertoire development in the zebra fi sh model organism have established strong correlations in the early repertoire that grew more divergent in adult fi sh (34). The authors attributed variation in the adult fi sh to stochastic clonal expansion from a common underlying repertoire (34).

To clarify the impact of heritable mechanisms governing B-cell receptor diversity, we performed a blinded high-throughput sequencing evaluation of antibody repertoire diversity in two middle-aged monozygotic twin pairs, one pair discordant for multiple sclerosis (MS) and the affected twin chronically treated with lymphocyte-depleting and immunomodulatory agents. By controlling for genetic variation, the selected samples provide a means of addressing the impact of genetics on repertoire formation and environment on repertoire drift. In the discordant twin pair, chronic immunotherapy in the MS-affected sibling provides a unique opportunity to evaluate heritable in fl uences of repertoire re-establishment after chemical ablation.

3.2.2 Results

Peripheral blood mononuclear cells (PBMCs) were obtained from both siblings in two monozygotic twin pairs (twin group A: twin A1 and twin A2; twin group B: twin B1 and twin B2). Each PBMC sample was divided into biological replicates before

cell lysis and RNA extraction. B-cell repertoires of the twin pairs were ampli fi ed separately using single isotype-speci fi c3 ´ primers and the 5 ´ SMARTer RACE universal primer mix (Table S1). All samples were assigned multiplex identi fi er (MID) barcodes and sequenced with 454 GS FLX high-throughput technology. Titanium long-read chemistry was used to observe the somatic hypermutation (SHM) load of entire variable domains in a single read.

A total of 3,316,360 reads were obtained across all twin samples (Table S2). When fi ltered for reads containing an identi fi able V-segment, J-segment, and complementarity-determining region (CDR) 3 in a single frame, 153,500 ± 13,400 heavy-chain sequences and 101,400 ± 16,200 light-chain sequences were obtained per individual. To remove chimeric sequences, ambiguous segment assignments, and other common sources of sequencing error (35), the VDJFasta probabilistic classi fi er was used to fi lter the sequence output (36). After fi ltering, 143,800 ± 13,000 heavy-chain and 99,800 ± 16,200 light-chain sequences were available for each of the four twins (Table S2).

Both B-cell proliferation and PCR can cause reads from a single V(D)J rearrangement event to appear multiple times in a sequence dataset. To avoid bias from overcounting dominant clones, all sequences bearing the same V-gene, J-gene, CDR3 length and CDR3 amino acid composition differing by no more than two amino acids were clustered and treated as individual, nonredundant sequences. After clustering, each twin retained 15,500 ± 4,600 unique heavy-chain clones and 10,600 ± 2,400 unique light-chain clones. This target depth was determined to produce highly reproducible (Pearson ' s r $>$ 0.99) V-gene pro fi les in simulated sampling experiments (Fig. S1).

Sequence analysis approximated interrogation of the naive compartment without cell sorting by considering only sequences that were IgM + and SHM low [contain less than 5-bp mutations over at least 270 V-gene bases from the closest reference International Immunogenetics (IMGT) V-gene germ line] (Fig. S2). The optimal SHM cutoff between naive and memory repertoire was determined by analyzing 379,637 CD27 − and 483,940 CD27 + antibody sequences obtained from FACS-sorted B-cells

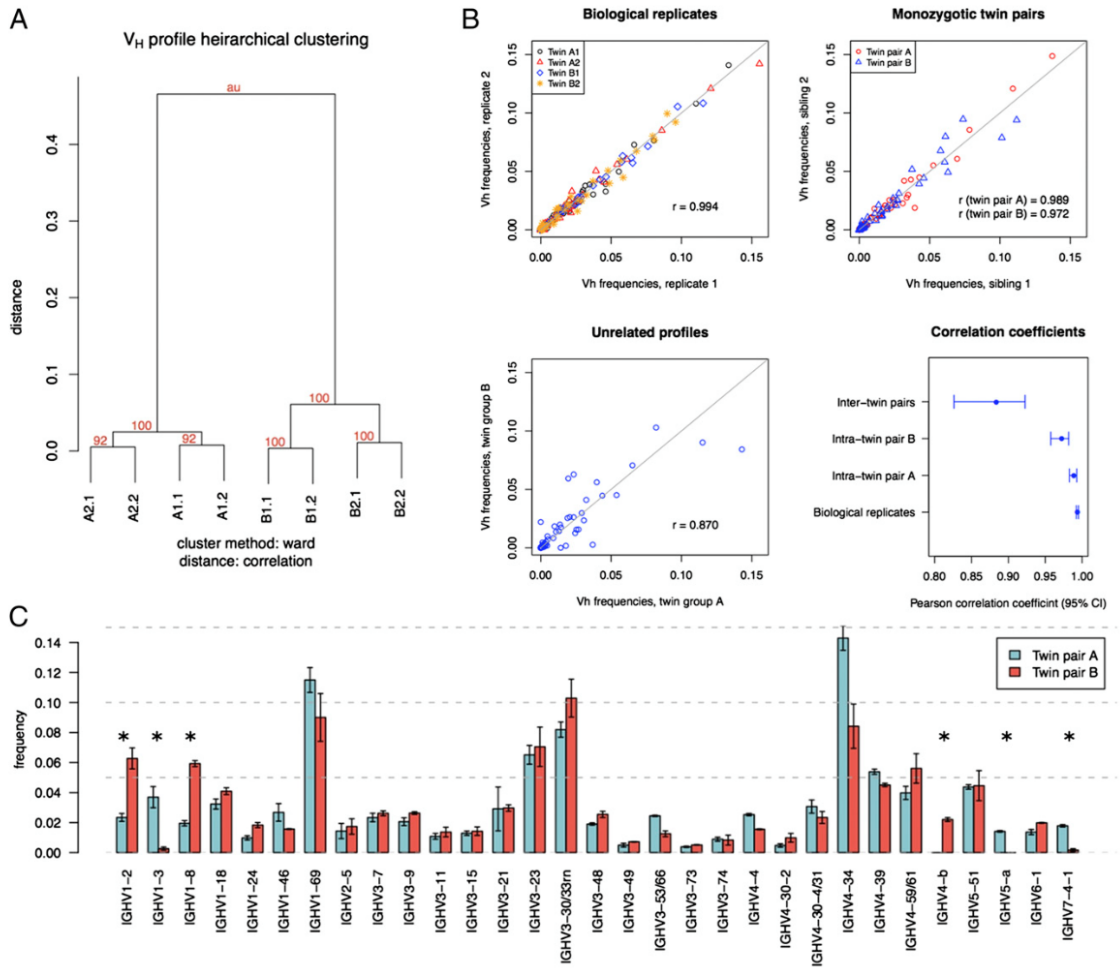


Figure 3.4: Heritable variation observed in naive VH-gene use profiles. (A) Hierarchical clustering of naive VH-gene profiles from twins and biological replicates. Genetically distinct groups can be distinguished with high confidence by approximately unbiased multiscale bootstrapping (AU 100%). (B) VH-gene use profile correlation between biological replicates (r = 0.994), monozygotic twins (r = 0.989, r = 0.972), and unrelated individuals (r = 0.870). (C) Significant heritable differences in V-gene profiles between twin pairs observed for germ lines IGHV1-2, IGHV1-3, IGHV1-8, IGHV4-b, IGHV5-a, and IGHV7-4–1 (*P $<$ 4.7e-04).

from three healthy controls (Fig. S2). This " in silico cell sorting " procedure resulted in 9,700 ± 5,600 unique IgM + SHM low heavychain clones and 4,800 ± 1,200 unique SHM low light-chain clones per twin.

In a blinded evaluation, naive V H -gene pro fi les could correctly discriminate genetically distinct twin pair samples with high con fi dence [100% Approximately Unbiased (AU) bootstrap probability of correct classi fi cation (37, 38)], and correctly assign biological replicates to individuals within twin pairs (92% AU for Twin group A and 100% for Twin group B) (Fig. 1 A).

V-gene use differed between unrelated individuals but not within twins (Fig. 1 B). Although V-gene pro fi les from unrelated donors exhibited commonalities in the characteristic human germ-line V H use (Pearson ' s correlation coef fi cient r = 0.87), intratwin group A (A1 vs. A2) and intratwin group B (B1 vs. B2) assessments were nearly as correlated as biological replicates from the same individual (r = 0.989, r = 0.972, and r = 0.994, respectively) (Fig. 1 B). The effect did not appear to be because of an abnormality of V-gene use in one of the twin pairs: additional comparisons between the V-gene pro fi les of the three healthy controls (female Hispanic age 47, female Asian age 44, female Caucasian age 54) provided an average Pearson ' s correlation coef fi cient between unrelated individuals as 0.768 ± 0.093.

Although all V H segments showed highly correlated use between twins, a substantial proportion (6/31) were observed to have signi fi cantly different use (P < 8.4e-4) compared with the unrelated individuals from the other twin pair (Fig. 1 C). Nested ANOVA attributed the proportion of variance to intertwin variability greater than 95% for all signi fi cant V H segments, compared with an average of 31.7% (SD 35.3%) for nonsigni fi cant V H segments. The twin pairs were observed to use different alleles for three of the differentially used segment loci, IGHV12, IGHV1-3, and IGHV7-4-1 (Table S3). Differential utilization of IGHV4-b and IGHV5-a was characterized by complete absence of the segment in one twin pair but not the other. Twins could also be distinguished by heritable variation in their V $\varkappa$ and V λ pro fi les, although the effects were less pronounced than that observed in the heavy chain (Fig. S3). Of 45 evaluated, 4 light chain segments showed signi fi cantly different use between twins: IGKV1-NL1, IGKV1-27, IGKV2-28, and IGLV3-10 (Fig. S3).

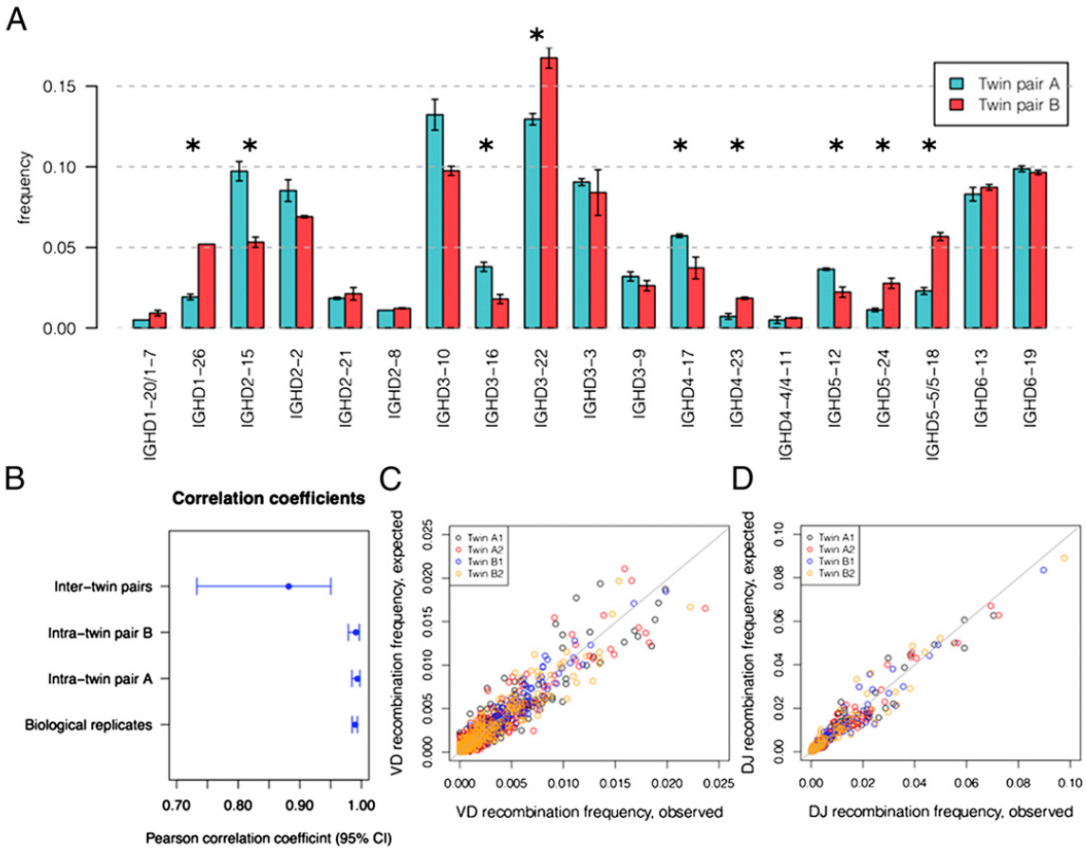


Figure 3.5: Limits of heritable variation observed in naive V(D)J recombination repertoire. (A) Significant heritable differences in D-gene profiles between twin pairs observed for nine D-segments (*P $<$ 5.3e-4). (B) Correlation coefficients for D-segment use within biological replicates, within twin pairs, and between twin pairs. Reported with 95% confidence intervals. (C) V-D observed recombination frequencies are highly correlated to expected frequencies (r $= 0.95 \pm 1.7$). (D) D-J observed recombination frequencies are highly correlated to expected frequencies (0.97 ± 0.01).

Signi fi cant heritable variation was also observed in the Dsegment repertoire (Fig. 2 A). Although not all D-segments can be reliably classi fi ed (Table S4), the subset that was classi fi ed showed high correlations within twins (r = 0.993 twin pair A, r = 0.991 twin pair B), and substantial differences between twins (r = 0.882) (Fig. 2 B). Of D-segments that could be classi fi ed and represented, at least 0.5% of the repertoire of any individual, 9 of 19 showed signi fi cantly different use between twin pairs (P $<$ 5.3e-4) (Fig. 2 A). In contrast, J-segments did not show signi fi cant differences in use across twin pairs (Fig. S3).

Although the V-gene and D-gene use exhibited heritable variation between the twin pairs, the V(D)J recombination process does not show preferential linkage, instead occurring largely in proportion to underlying abundance of the segments. A comparison of observed vs. expected V-D and D-J recombination frequencies showed high correlations (V-D r = 0.95 $\pm$ 1.7; D-J r = 0.97 $\pm$ 0.01) and few examples of biased selectivity in the recombination process (Fig. 2 C and D), in agreement with previous reports (32).

During the blinded analysis, three individuals had evidence of SHM in 30 – 40% of their IgM V-segments, but a single individual was observed to have a nearly complete absence of V-gene mutations in their IgM repertoire (Fig. 3 A). TCRB repertoire sequencing for each of the four twins served as a control to con fi rm that the observed somatic hypermutation rates in three of the four samples was well in excess of the PCR and sequencing read error rate expected from the process (Fig. 3 A). The fi nding was further corroborated by FACS analysis, where CD20 + IgM + B-cells of this individual were found to consist almost entirely of naive cells (97.5% CD27 $-$) (Fig. 3 B).

Unblinding of the samples revealed this to be the MS-affected and chronically immunosuppressed patient twin B1. Age 57 at the time of sampling, twin B1 was diagnosed with relapsing remitting MS at age 31 and had progressed to a secondary progressive MS course. Since diagnosis, twin B1 had been treated with immunomodulatory (IFN B1, glatiramer acetate) agents that may promote B regulatory response (39) or alter the composition of B-repertoires (40), but also immunosuppressive agents

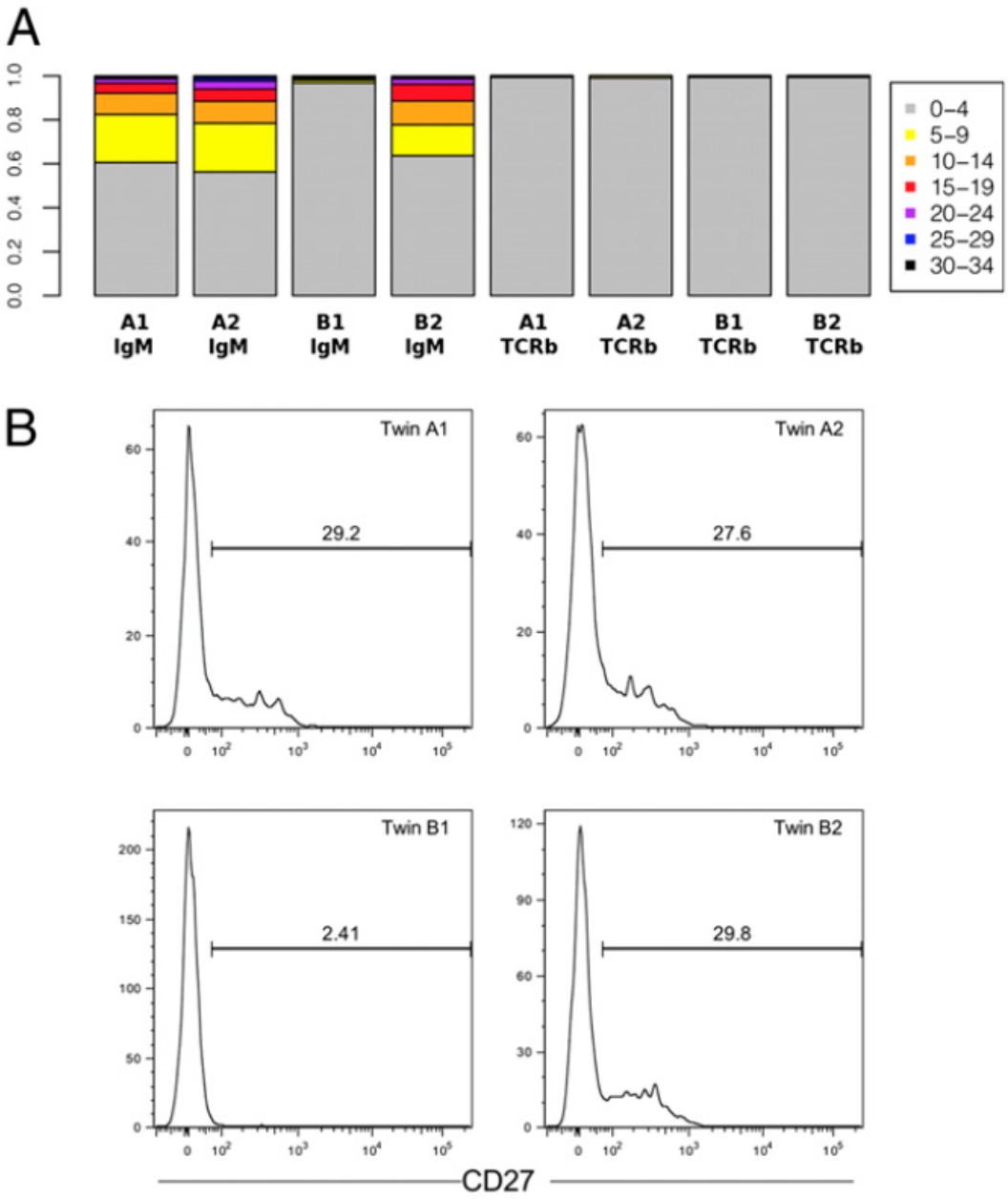


Figure 3.6: Impact of chronic immunosuppressive therapy on twin B1. (A) SHM load in variable domains of IgM repertoire. In healthy twins A1, A2, and B2, 30–40% of unique clones contain >4 bp SHM. In treated twin B1, >97% of the IgM repertoire appears naive. Of the sequences from the TCRB repertoire of each twin, 10,000 were sequenced as a negative control for SHM detection. (B) Histograms of CD27+ cells gated on IgM+/CD20+. Healthy twins A1, A2, and B2 have 70–72% CD27− cells, but treated twin B1 has 97.5% CD27− cells.

		A1			A2			B1			B2		
		IgM	IgG	IgA	IgM	IgG	IgA	IgM	IgG	IgA	IgM	IgG	IgA
A1	IgM	9131	308	28						2	2		1
	IgG	8%	3761	143							1		1
	IgA	0.5%	4%	5604				7			1		2
A2	IgM	0.0%	0.0%	0.0%	10096	53	41			2	1		
	IgG	0.0%	0.0%	0.0%	4%	1438	87		1				
	IgA	0.0%	0.0%	0.0%	1%	6%	3163				1		
B1	IgM	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	17946	39	23	3		
	IgG	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	6%	652	20			
	IgA	0.1%	0.0%	0.0%	0.1%	0.0%	0.0%	1%	3%	1565			
B2	IgM	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	8967	32	38
	IgG	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2%	1414	61
	IgA	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2%	4%	2089

Figure 3.7: Unique V(D)J clone overlap between IgM and class-switched IgG and IgA. Unique counts shown in upper diagonal, percent overlap [(A ⋂ B)/min(A,B)] in the lower diagonal, and unique clones in sample along diagonal. In the upper diagonal, counts for twins A are in green, for twins B are in orange. Log10 heatmap colors sequence counts. Percentage of overlap in the bottom diagonal is dark gray if over 1%, light gray if 0.1–1%, and white if less than 0.1%.

(mitoxantrone, methotrexate, mycophenolate, cyclophosphamide) that directly deplete the B-cell repertoire by inhibiting DNA synthesis. Twin B1 had been off any of these medications for 10 mo at the time of sampling. Remarkably, although chronic lymphocyte depletion therapy for twin B1 had caused a dramatic loss of somatically hypermutated IgM cells (Fig. 3 A), naive V-gene, D-gene, and J-gene use remained highly correlated to her monozygotic sibling (Figs. 1 and 2, and Fig. S3). This resiliency of the naive V-segment pro fi le to depletion suggests that naive V-segment use is predetermined by hereditary in fl uences that precede repertoire formation.

The highly diverse CDR-H3 is considered to be the key determinant of antigen speci fi city (41). To assess whether heritable fi ve-gene and D-gene use also contributes to convergent (CDRH3) generation, we analyzed the amino acid sequences CDR-H3 in monozygotic twins. At the sampling depth obtained in this study, CDR-H3 repertoires appear highly personal, sharing few common clones between individuals and bearing no more similarity to an identical twin than an unrelated individual. No more than nine common clones were observed between individuals, but $> 99.9\%$ of clones were unique to an individual (Fig. 4 A). Within an individual, $2 - 8\%$ of clonally derived CDR-H3s are present in both IgM and classed switched repertoires at the depth obtained in this study (Fig. 4 A). Clonal variants were also found: in an individual, 27% of CDR-H3 clones had a related clone that differed by less than three amino acids (Fig. S4). However, when comparing across samples from different individuals, the next closest CDR-H3 differed by at least four residues (Fig. S4). Notably, CDR-H3 repertoires produced by identical twins neither produced identical CDR-H3 sequences nor produced CDR-H3 repertoires that were detectably more similar in amino acid composition than that produced by an unrelated individual.

Both intratwin pairs share nearly identical naive V H -gene use but each had lived apart and been exposed to different antigens. Antigenic exposure activates and matures B cells for improved binding to speci fi c antigens. This fi nding establishes a repertoire of speci fi c antigen-experienced receptors that will respond rapidly to secondary antigenic challenge. To assess the balance between genetically predetermined V-gene use pro fi les and antigenic pressure, we analyzed the rearranged Ig transcripts from the antigen-experienced class-switched compartments.

In the three healthy twins (twin A1, A2, B2), class-switched Vgene use was signi fi cantly correlated ($P < 0.0005$) with their respective naive V-gene pro fi les, but with far more variation than observed in the naive repertoire (Fig. 5). Even with increased variation, the majority of segments appear at similar frequencies in class-switched and naive repertoires, with 76 of 83 switched V-segments (91.6%) appearing between 0.5 to 2-fold of naive use, and only 10 appearing at ratios signi fi cantly different from expectation (IGHV1-69, IGHV1-8, IGKV1-8, IGKV1D-8, IGKV3-11, IGKV4-1, IGLV2-5, IGLV2-23, IGLV3-16, IGLV3-19; single t test, null ratio = 1, P

$<$ 6.0e-04) (Fig. 5). These segment-speci fi c ratios were found in V H ,V $\varkappa$,andV λ , showed some isotype variation (Fig. S5), and were similar across healthy twins A1, A2, and B2. However, the affected twin B1 ' s memory use (indicated by red circles) differed substantially (Fig. 5).

3.2.3 Discussion

Through the sequence analysis of antibody repertoires in monozygotic twins, we are able to distinguish a fi xed naive gene segment use, a derived stochastic memory, and a highly diverse and personal CDR-H3 repertoire. The presence of heritable naive gene-segment pro fi les provides an essential basis for the understanding of B-cell repertoire genesis. We observe monozygotic twins to share nearly identical naive gene segment use pro fi les. V H -gene and D H -gene segment pro fi les showed population variation, differing between unrelated twin pairs. Environmental in fl uences, autoimmunity (42), and lymphocyte depletion introduced almost no variation between the common segment use pro fi les of a twin pair. These observations exclude the possibility that naive V-gene and D-gene use is governed by environment or homeostasis, but rather appears as a function of genetic origins.

We were able to discriminate between the naive pro fi le and the activated pro fi le by sequence analysis. By fi ltering out noise from redundant clones and the af fi nity matured repertoire, the underlying heritable naive repertoire became available for inspection. Parallel analysis of the TCRB repertoire provided an important control to distinguish the rate of sequencing error from the underlying somatic hypermutation rate in our samples. The use of this in silico cell-sorting methodology will facilitate rapid characterization of the naive repertoire without FACS for future large-scale repertoire disease-association studies.

A fi xed heritable naive V-gene repertoire and a stochastically derived memory repertoire complement observations made in previous studies. In the TCRB-V repertoire of monozygotic twins (30), correlations in healthy twins but not discordant twins can be explained by a divergent memory response mixed with a common underlying naive repertoire. In zebra fi sh (34), the early repertoire was highly correlated, but

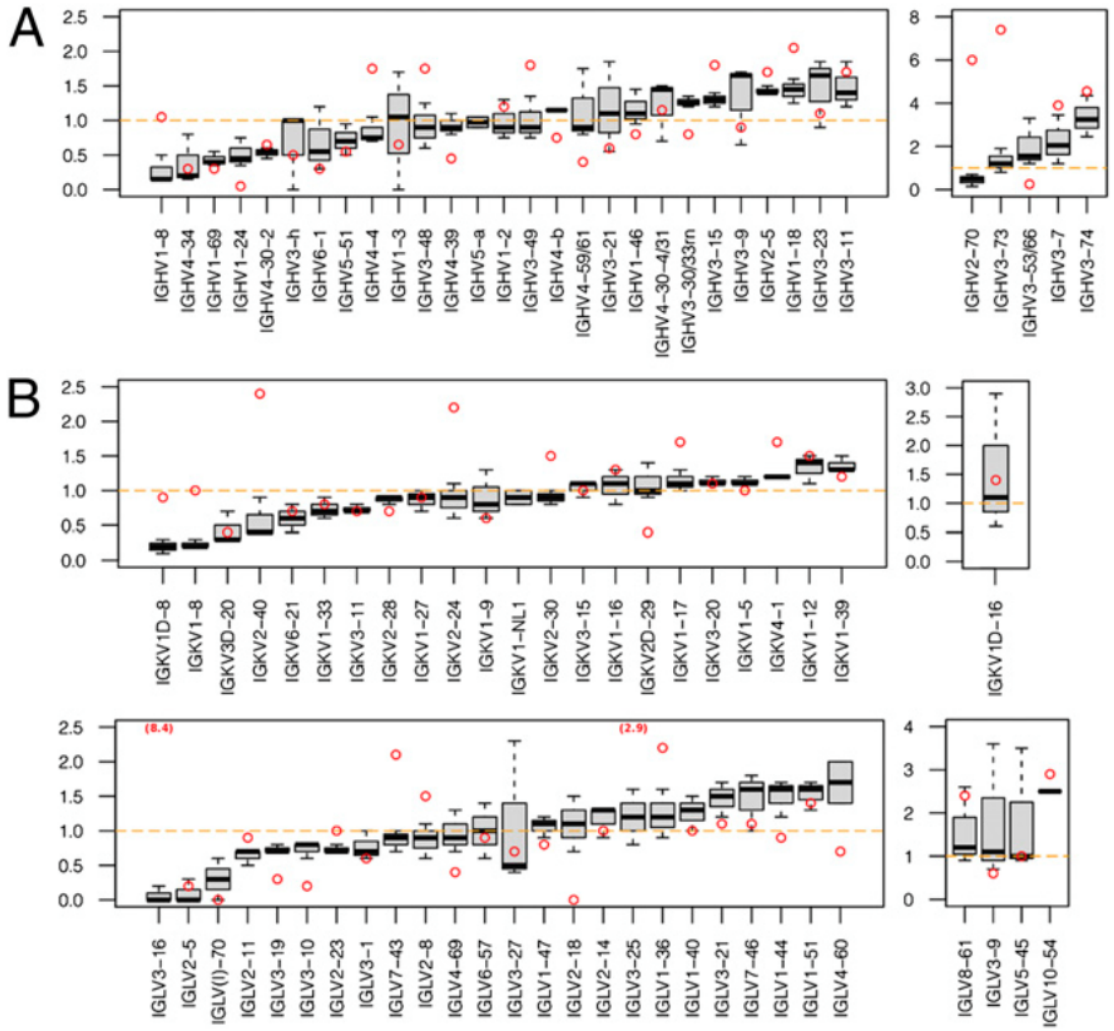


Figure 3.8: Ratio of V-gene use in activated to naive repertoire. (A) Ratio of heavy-chain V-gene use between class-switched IgA and IgG to naive IgM. Twins A1, A2, and B2 show significant correlation between naïve and classswitched V-gene use (Pearson's r: P = 3.6e-4, 3.3e-4, 2.0e-8, respectively). The MS Twin B1, indicated by red open circles, shows a different ratio in V-gene use between class-switched IgA and IgG to naive IgM. The null ratio of 1 is indicated with an orange line. Elevated V-genes are indicated in a separate scale to the right. Individual isotypes IgM, IgG, and IgA are shown in Fig. S5. (B) Same as above for V$\varkappa$ and Vλ: (Pearson's r: P $<$ 1e-10 for A1, A2, and B2).

grew more divergent with age. The result is consistent with an observation of an initial naive repertoire becoming increasingly mixed with a derived but stochastic memory. By separating the analysis of naive from antigen-experienced subcompartments of the repertoire, we demonstrate that the naive repertoire remains highly correlated within twin pairs, even as the activated and memory compartments diversify in a stochastic manner.

Previous studies have suggested allele use as a source of heritable segment use variation (33, 43). Studies in V-gene use variation in inbred mouse strains suggest that the regulatory mechanisms may represent a more general mechanism for maintaining V-gene diversity in a population (44). A complete characterization of heritable mechanisms should consider haplotype architecture, cis -regulatory elements, copy-number variation, and even transregulatory elements that may in fl uence segment selection (33, 45).

Despite common segment-use pro fi les, the CDR-H3 repertoires remained highly personal to each individual in the study. Very few public clones (32) were encountered between individuals, with no increase in common clones observed between twins (Fig. 4 A), or even a statistically signi fi cant generation of more biochemically similar CDR-H3 repertoires (Fig. S4). Although he sequencing performed in this study did not exhaustively sample all B-cell receptors in each individual, our sampling suggests that even with identical biased segment pro fi les, the unbiased V(D)J recombination processes (Fig. 2 B) are able to generate far more diversity than can be displayed on an individual ' s B cells. As a consequence, even with common genesegment pro fi les, twins are likely to respond with different speci fi c antibodies to a common environmental exposure.

As previously observed (46), memory V-segment use is correlated to naive use despite antigen-driven selection between random environmental antigens and personal CDR-H3 repertoires. The memory compartment, although stochastic in antigen-driven selection, is sampling repeatedly from a very stable naive distribution. Over time, the memory compartment can be expected to recapitulate the naive distribution from which it was sampling. A limited number of V-segments exhibited systematic underor overrepresentation in class-switched B cells. This bias suggests an active

mechanism favoring or impeding memory retention of B cells using these V-segments. These biases may hint at unique properties of speci fi c V-genes.

Subsets of V-segments are already known to exhibit biased representation in infection, autoimmunity, and B-cell lymphoma (1 – 8). The underrepresented V H 1 – 69 segment is known to have a distinctive hydrophobic CDR-H2 that may aid in early viral response (8), but also predispose toward polyspeci fi cty and chronic lymphocytic leukemia (47 – 49). The V H 4 – 34 V-segment has known autoreactivity to red blood cell antigens and is overrepresented in cold agglutinin disease (5, 10, 11, 13 – 15, 50), dysregulated in systemic lupus erythematosis (7, 51), and overrepresented in B-cell lymphomas (1, 3, 52). Past studies have identi fi ed germinal center exclusion as a mechanism for underrepresentation of the inherently autoreactive V H 4 – 34 segment (1 – 5, 7, 8). Whatever the origins, it is likely that inherited variation in use of segments under special selection by the adaptive immune system will impact immune function. Identifying such pro fi le-risk associations would provide a key component to understanding complex disease pathogenesis.

In our study, one monozygotic twin pair was discordant for MS, a disease in which B cells appear to contribute greatly to the disease process (9). A previous study examining this twin pair ' s genome, epigenome, and RNA demonstrated the absence of factors in these compartments suitable to explain discordance (42). Here we establish heritable naive V-gene pro fi les that may contribute to, but are not altered by, autoimmune pathology. In future studies, personal, stable, naive V-gene use may act as a useful control to aid in detecting pathology-driven repertoire variation in the B-cell memory compartment.

3.2.4 Methods

Sample Collection. The study was approved by the University of California, San Francisco Institutional Review Board. Peripheral blood was obtained from two pairs of adult, monozygotic twin pairs: Pair A (A1 and A2), female, Caucasian non-Hispanic origin (Western European and Mediterranean backgrounds); and Pair B (B1 and B2), female, Ashkenazi Jewish origin. Twin pair B is discordant for MS (affected twin:

B1). At the time of sampling, monozygotic twin pair A was 54 y old, and monozygotic twin pair B was 57 y old. Twin B1 was diagnosed with relapsing remitting MS at age 31 and had progressed to a secondary progressive MS course. Since diagnosis, twin B1 had been treated with numerous immunomodulatory (IFNβ 1, glatiramer acetate) and immunosuppressive (mitoxantrone, methotrexate, mycophenolate, cyclophosphamide) agents, but had been off any of these medications for 10 mo at the time of sampling. Monozygosity was con fi rmed for twin pair A by SNP analysis covering all chromosomes and for twin pair B by complete genome sequencing in a previous study (1).

Amplicon Generation. PBMCs from each individual were split into biological replicates before RNA extraction. Total RNA was isolated using All Prep kit (Qiagen) and reverse-transcribed using SMARTer RACE cDNA Ampli fi cation Kit (Clontech). Isotype repertoires were ampli fi ed using single validated isotype speci fi c3 ´ primers (Table S1). Products were isolated and puri fi ed using a Gel Puri fi cation kit (Qiagen). Puri fi ed products were quanti fi ed using Quant-iT PicoGreen dsDNA kit (Invitrogen).

Sequencing. Gel-puri fi ed IgM, Ig $\varkappa$,Ig λ , IgG, IgA, and TCRB 5 ´ RACE products were taken into the 454 Rapid library construction protocol to uniquely ligate Lib-L adaptors containing 10-base MIDs/barcodes to each sample. Pooled libraries were subjected to emulsion PCR and bidirectional sequencing using the GS FLX Titanium Lib-L chemistry.

Sequencing Depth Determination. A target depth of 5,000 reads ensured a highly reproducible gene-segment pro fi le ($r > 0.99$). Parameters were determined by simulated sampling from a dataset of 863,577 5 ´ RACE, 454 Titanium sequences from three healthy controls (Fig. S1).

Antibody Sequence Analysis. To remove chimeric sequences and other common sources of sequencing error (35), V, J, and CH1 segment use was determined by probabilistic classi fi cation among National Center for Biotechnology Information blastn solutions to an IMGT reference database (36) (Tables S4 and S5). D-segments were only classi fi ed if found between a V-segment and J-segment (Table S5). Parameters, validation, and performance are reported in Tables S4 and S5. CDRs

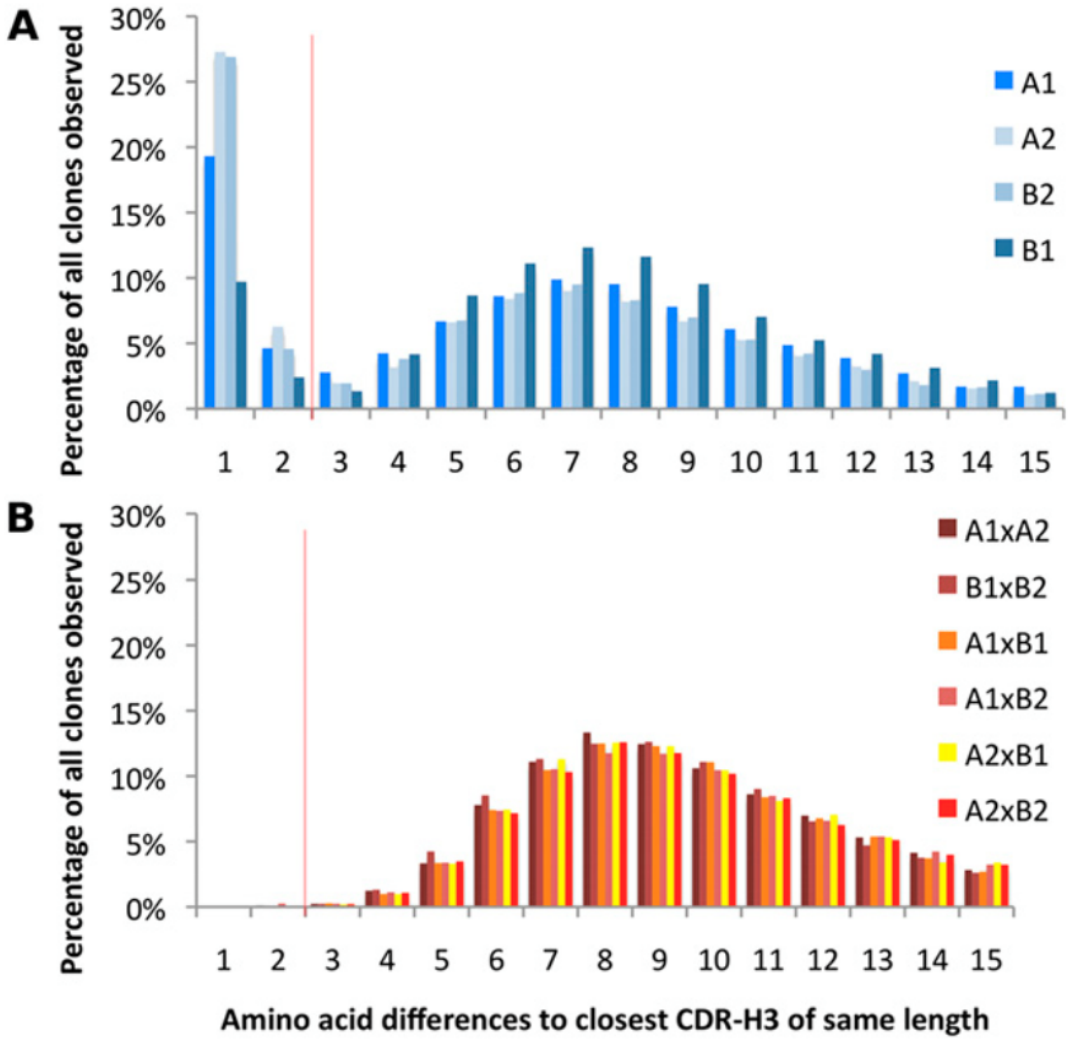


Figure 3.9: Related clone distances. (A) Amino acid distance to next closest CDR-H3 within each twin. (B) Amino acid distance to next closest CDR-H3 across samples from different individuals. To avoid bias from overcounting expanded clones, sequences were clustered and only one representative of each unique V(D)J rearrangement was selected for subsequent analysis. Clonal nonredundancy was established by considering only one representative from a set of reads bearing the same V-segment, J-segment, CDR3 length, and CDR3 amino acid composition that differed by less than two residues from any other CDR-H3 in the dataset. Parameters were determined by comparing distances to the next closest clone between all clones within an individual's samples (assumed to contain clonal variants), and between different individual's samples (that cannot share clonal variants). A and B illustrate that the great majority of variants are single amino acid changes away in CDR-H3, but the minimum distance to the closest unrelated clone is at least 4 amino acid changes. A cutoff of 2 amino acids maximizes separation between related and unrelated clones: clustering the majority of redundant clones without falsely clustering unrelated clones. Although it is true that a limited number of distantly related clones may not be grouped in some cases, by using a cutoff of 2, the majority of bias is eliminated from the dataset. The most

	Primer
IgM	5′-GATGGAGTCGGGAAGGAAGTCCTGTGCGAG-3′
IgG	5′-GGGAAGACSGATGGGCCCTTGGTGG-3′
IgA	5′-CAGGCAKGCGAYGACCACGTTCCCATC-3′
Igκ	5′-CATCAGATGGCGGGAAGATGAAGACAGATGGTGC-3′
Igλ	5′-CCTCAGAGGAGGGTGGGAACAGAGTGAC-3′
TCRB	5′-GCTCAAACACAGCGACCTCGGGTGGGAACAC-3′
5′RACE long	5′-CTAATACGACTCACTATAGGGCAAGCAGTGGTATCAACGCAGAGT-3′
5′RACE short	5′-CTAATACGACTCACTATAGGGC-3′

Table 3.2: Glanville PNAS 2011 Table1.

were identi fi ed with pro fi le Hidden Markov Models trained on IMGT segments (53) (Table S6). Reads not passing all criteria were fi ltered from subsequent analysis. Code available at http://sourceforge.net/projects/vdjfasta/.

Clone Clustering. Unlike the TCR repertoire (2), the antibody repertoire contains a large number of highly expanded and somatically mutated clonal variants. To avoid bias from overcounting expanded clones, sequences were clustered, and only one representative of each unique V(D)J rearrangement was selected for subsequent analysis. Clones were considered redundant and removed if they shared the same V-gene, J-gene, CDR3 length, and CDR3 amino acid sequence that differed by less than 3 amino acids. Parameters were determined through the analysis of clone distances within individuals (assumed to contain clonal variants), and across individuals (that cannot share clonal variants). For details, see Fig. S4.

In Silico Cell Sorting. The amount of V-segment mutations were used to classify reads as either naive or activated: over 270 bp of V-gene, $0 - 4$bp mismatches to reference were considered naive, and > 4 bp mismatches were consider activated. Parameters were determined through analysis of 379,637 CD27 $-$ sequences and 483,940 CD27 + obtained from CD27 FACSsorted CD20 + cells from three healthy controls (Fig. S2), as well as 40,000 TCRB sequences from each of the twins (Fig. 3 A). For details, see Fig. S2 .

Statistical Analysis. Pairwise V-segment pro fi le comparisons were performed between biological replicates, within monozygotic twin pairs, and between monozygotic twin pairs with Pearson ' s correlation coef fi cient. V-segment pro fi le clustering

Twin	Chain	BioRep	Reads	Ig	H VDJ	L VJ	H SHM$^-$	L SHM$^-$	H SHM$^+$	L SHM$^+$
A1	IgM/IgK/IgL	1	231,111	210,586	30,810	34,065	7,661	7,299	4,759	17,340
A1	IgM/IgK/IgL	2	246,010	226,525	38,026	32,882	9,242	6,800	5,507	17,138
A1	IgG	1	95,177	64,662	10,303	NA	940	NA	7,209	NA
A1	IgG	2	93,019	76,133	19,684	NA	1,216	NA	13,872	NA
A1	IgA	1	48,944	43,444	10,781	NA	197	NA	4,883	NA
A1	IgA	2	46,736	41,676	11,058	NA	181	NA	4,714	NA
B1	IgM/IgK/IgL	1	292,499	268,200	39,927	41,877	16,771	27,271	511	4,820
B1	IgM/IgK/IgL	2	323,398	299,665	48,406	50,987	19,318	31,786	386	6,435
B1	IgG	1	67,173	36,890	2,776	NA	339	NA	1,677	NA
B1	IgG	2	63,004	39,297	6,271	NA	890	NA	4,036	NA
B1	IgA	1	72,307	62,841	15,466	NA	443	NA	7,191	NA
B1	IgA	2	61,104	54,522	13,576	NA	665	NA	5,057	NA
B2	IgM/IgK/IgL	1	248,289	226,474	32,221	33,671	7,938	11,639	4,535	13,753
B2	IgM/IgK/IgL	2	240,321	220,742	34,066	34,315	8,876	12,070	4,248	12,924
B2	IgG	1	66,738	43,203	7,185	NA	89	NA	5,227	NA
B2	IgG	2	61,694	46,468	12,013	NA	198	NA	9,112	NA
B2	IgA	1	73,152	64,127	16,577	NA	56	NA	7,979	NA
B2	IgA	2	61,714	54,874	14,609	NA	90	NA	6,403	NA
A2	IgM/IgK/IgL	1	300,087	275,152	43,392	41,318	9,970	10,566	7,437	19,017
A2	IgM/IgK/IgL	2	299,270	277,293	46,717	41,564	11,021	10,281	7,641	18,713
A2	IgG	1	85,610	53,739	7,398	NA	100	NA	5,776	NA
A2	IgG	2	69,304	49,881	10,789	NA	200	NA	8,306	NA
A2	IgA	1	67,749	60,346	16,886	NA	150	NA	8,153	NA
A2	IgA	2	61,950	55,569	16,991	NA	70	NA	7,676	NA

Table 3.3: Glanville PNAS 2011 Table2.

was performed with the R package pvclust using Wards agglomerative method and correlation distance. Cluster con fi dence was addressed with AU multiscale multistep resampling bootstrap con fi dence values using 10,000 bootstrapping replicates (37, 38). AU $>$ 95% was considered signi fi cant. Nested ANOVA was used to compare individual V-segments within and between monozygotic twins and to estimate their respective variance components. A single sample t test of log 2 ratios was used to evaluate biased use in class-switched vs. naive V-segment use. Given the number of naive segments evaluated (95: 34 V H ,20V $\varkappa$,25V λ , and 19 D H), we used the Bonferroni correction multiple comparisons correction such that P values $<$ 0.05/95 = 5.3e-4 were considered to be statistically signi fi cant. Analyses were performed in R version 2.11.1.

3.2.5 Acknowledgements

We thank the individuals who agreed to serve as subjects for this study, Dilduz Telman for 454 emulsion processing, and Stacy Caillier for expert sample p reparation. This work was supported by National Institutes of Health Grant R01NS26799, National Multiple Sclerosis Society Grant RG 2901, and a grant fro m Small Ventures USA Inc.

3.2.6 References

1. Bhat NM, et al. (2004) B cell lymphoproliferative disorders and VH4-34 gene encoded antibodies. Hum Antibodies 13:63 – 68.

2. Chan CH, Hadlock KG, Foung SK, Levy S (2001) V(H)1-69 gene is preferentially used by hepatitis C virus-associated B cell lymphomas and by normal B cells responding to the E2 viral antigen. Blood 97:1023 – 1026.

3. Marasca R, et al. (2001) Immunoglobulin gene mutations and frequent use of VH1-69 and VH4-34 segments in hepatitis C virus-positive and hepatitis C virus-negative nodal marginal zone B-cell lymphoma. Am J Pathol 159:253 – 261.

4. Owens GP, et al. (2007) VH4 gene segments dominate the intrathecal humoral im- mune response in multiple sclerosis. J Immunol 179:6343 – 6351.

5. Pugh-Bernard AE, et al. (2001) Regulation of inherently autoreactive VH4-34 B cells in the maintenance of human B cell tolerance. J Clin Invest 108:1061 – 1070.

6. Thorsélius M, et al. (2006) Strikingly homologous immunoglobulin gene rearrange- ments and poor outcome in VH3-21-using chronic lymphocytic leukemia patients in- dependent of geographic origin and mutational status. Blood 107:2889 – 2894.

7. van Vollenhoven RF, et al. (1999) VH4-34 encoded antibodies in systemic lupus erythematosus: A speci fi c diagnostic marker that correlates with clinical disease characteristics. J Rheumatol 26:1727 – 1733.

8. Wang TT, Palese P (2009) Universal epitopes of in fl uenza virus hemagglutinins? Nat Struct Mol Biol 16:233 – 234.

9. Hauser SL, et al.; HERMES Trial Group (2008) B-cell depletion with rituximab in re- lapsing-remitting multiple sclerosis. N Engl J Med 358:676 – 688.

10. Zheng NY, et al. (2004) Human immunoglobulin selection associated with class switch and possible tolerogenic origins for C delta class-switched B cells. J Clin Invest 113: 1188 – 1201.

11. Pascual V, et al. (1992) VH restriction among human cold agglutinins. The VH4-21 gene segment is required to encode anti-I and anti-i speci fi cities. J Immunol 149: 2337 – 2344.

12. Pascual V, Capra JD (1992) VH4-21, a human VH gene segment overrepresented in the autoimmune repertoire. Arthritis Rheum 35:11 – 18.

13. Børretzen M, Chapman C, Stevenson FK, Natvig JB, Thompson KM (1995) Structural analysis of VH4-21 encoded human IgM allo- and autoantibodies against red blood cells. Scand J Immunol 42:90 – 97.

14. Thompson KM, et al. (1991) Human monoclonal antibodies against blood group antigens preferentially express a VH4-21 variable region gene-associated epitope. Scand J Immunol 34:509 – 518.

15. Silberstein LE, et al. (1991) Variable region gene analysis of pathologic human autoantibodies to the related i and I red blood cell antigens. Blood 78:2372 – 2386.

16. Kakoulidou M, et al. (2004) The autoimmune T and B cell repertoires in monozygotic twins discordant for myasthenia gravis. J Neuroimmunol 148:183 – 191.

17. Kohler PF, Rivera VJ, Eckert ED, Bouchard TJ, Jr., Heston LL (1985) Genetic regulation of immunoglobulin and speci fi c antibody levels in twins reared apart. J Clin Invest 75: 883 – 888.

18. Sjöberg K, et al. (1992) Genetic regulation of human anti-malarial antibodies in twins. Proc Natl Acad Sci USA 89:2101 – 2104.

19. Brix TH, Kyvik KO, Hegedüs L (2000) A population-based study of chronic autoim- mune hypothyroidism in Danish twins. J Clin Endocrinol Metab 85:536 – 539.

20. MacGregor AJ, et al. (1995) Rheumatoid factor isotypes in monozygotic and dizygotic twins discordant for rheumatoid arthritis. J Rheumatol 22:2203 – 2207. 21. Tonegawa S (1983) Somatic generation of antibody diversity. Nature 302:575 – 581.

22. Huber C, et al. (1993) The V kappa genes of the L regions and the repertoire of V kappa gene sequences in the human germ line. Eur J Immunol 23:2868 – 2875.

23. Kawasaki K, et al. (1995) The organization of the human immunoglobulin lambda gene locus. Genome Res 5:125 – 135.

24. Matsuda F, et al. (1998) The complete nucleotide sequence of the human immuno- globulin heavy chain variable region locus. J Exp Med 188:2151 – 2162.

25. Cox JP, Tomlinson IM, Winter G (1994) A directory of human germ-line V kappa segments reveals a strong bias in their usage. Eur J Immunol 24:827 – 836.

26. Rao SP, et al. (1999) Biased VH gene usage in early lineage human B cells: Evidence for preferential Ig gene rearrangement in the absence of selection. J Immunol 163: 2732 – 2740.

27. Suzuki I, P fi ster L, Glas A, Nottenburg C, Milner EC (1995) Representation of re- arranged VH gene segments in the human adult antibody repertoire. J Immunol 154: 3902 – 3911.

28. Van Dijk-Härd I, Lundkvist I (2002) Long-term kinetics of adult human antibody repertoires. Immunology 107:136 – 144.

29. Loveridge JA, Rosenberg WM, Kirkwood TB, Bell JI (1991) The genetic contribution to human T-cell receptor repertoire. Immunology 74:246 – 250.

30. Roth MP, et al. (1994) TCRB-V gene usage in monozygotic twins discordant for mul- tiple sclerosis. Immunogenetics 39:281 – 285.

31. Weinstein JA, Jiang N, White RA, 3rd, Fisher DS, Quake SR (2009) High-throughput sequencing of the zebra fi sh antibody repertoire. Science 324:807 – 810.

32. Arnaout R, et al. (2011) High-resolution description of antibody heavy-chain reper- toires in humans. PLoS ONE 6:e22365.

33. Boyd SD, et al. (2010) Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. J Immunol 184:6986 – 6992.

34. Jiang N, et al. (2011) Determinism and stochasticity during maturation of the ze- bra fi sh antibody repertoire. Proc Natl Acad Sci USA 108:5348 – 5353. 35. Nguyen P, et al. (2011) Identi fi cation of errors introduced during high throughput sequencing of the T cell receptor repertoire. BMC Genomics 12:106.

36. Glanville J, et al. (2009) Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proc Natl Acad Sci USA 106:20216 – 20221.

37. Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. Syst Biol 51:492 – 508.

38. Suzuki R, Shimodaira H (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22:1540 – 1542.

39. Kala M, et al. (2010) B cells from glatiramer acetate-treated mice suppress experi- mental autoimmune encephalomyelitis. Exp Neurol 221:136 – 145.

40. van Boxel-Dezaire AH, et al. (2010) Major differences in the responses of primary human leukocyte subsets to IFN-beta. J Immunol 185:5888 – 5899.

41. Xu JL, Davis MM (2000) Diversity in the CDR3 region of V(H) is suf fi cient for most antibody speci fi cities. Immunity 13:37 – 45.

42. Baranzini SE, et al. (2010) Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. Nature 464:1351 – 1356.

43. Posnett DN, et al. (1994) Level of human TCRBV3S1 (V beta 3) expression correlates with allelic polymorphism in the spacer region of the recombination signal sequence. J Exp Med 179:1707 – 1711.

44. Yancopoulos GD, Malynn BA, Alt FW (1988) Developmentally regulated and strain- speci fi c expression of murine VH gene families. J Exp Med 168:417 – 435.

45. Feeney AJ (2009) Genetic and epigenetic control of V gene rearrangement frequency. Adv Exp Med Biol 650:73 – 81.

46. Tian C, et al. (2007) Evidence for preferential Ig gene usage and differential TdT and exonuclease activities in human naïve and memory B cells. Mol Immunol 44: 2173 – 2183.

47. Sui J, et al. (2009) Structural and functional bases for broad-spectrum neutralization of avian and human in fl uenza A viruses. Nat Struct Mol Biol 16:265 – 273. 4

8. Corti D, et al. (2010) Heterosubtypic neutralizing antibodies are produced by in- dividuals immunized with a seasonal in fl uenza vaccine. J Clin Invest 120:1663 – 1673.

49. Lerner RA (2011) Rare antibodies from combinatorial libraries suggests an S.O.S. component of the human immunological repertoire. Mol Biosyst 7:1004 – 1012.

50. Pascual V, et al. (1991) Nucleotide sequence analysis of the V regions of two IgM cold agglutinins. Evidence that the VH4-21 gene segment is responsible for the major cross-reactive idiotype. J Immunol 146:4385 – 4391.

51. Cappione A, 3rd, et al. (2005) Germinal center exclusion of autoreactive B cells is defective in human systemic lupus erythematosus. J Clin Invest 115:3205 – 3216.

52. Bhat NM, Bieber MM, Young LW, Teng NN (2001) Susceptibility of B-cell lymphoma to human antibodies encoded by the V4-34 gene. Crit Rev Oncol Hematol 39:59 – 68.

53. Johnson LS, Eddy SR, Portugaly E (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics 11:431

3.2.7 Copyright

3.3 Convergent heritable antibody responses against Staphylococcus aureus

In chapter 2.4, Reading specificity in the B-cell receptor repertoire, we presented the analysis of an IGHV1-69 antibody convergence group that specifically recognized the broadly neutralizing step epitope of influenza hemagglutinin trimer. As part of that study, it was established that genotype variation at the IGHV1-69 locus influenzed the ability of the hosts to generate productive vaccination responses. For structural reasons, as specific CDR-H2 position is implicated in this genotype variability in immune response. The study established a strong case for the mysterious balancing selection

of allelic polymorphism active at the IgH loci. In the following study we demonstrate an additional example of genotype-specific naive encoded responses against common pathogens: staphylococcus aureus.

Staphylococcus aureus is both an important pathogen and a human commensal. To explore this ambivalent relationship between host and microbe, we analysed the memory humoral response against IsdB, a protein involved in iron acquisition, in four healthy donors. Here we show that in all donors a heavily biased use of two immunoglobulin heavy chain germlines generated high affinity (pM) antibodies that neutralize the two IsdB NEAT domains, IGHV4-39 for NEAT1 and IGHV1-69 for NEAT2. In contrast to the typical antibody/antigen interactions, the binding is primarily driven by the germline-encoded hydrophobic CDRH-2 motifs of IGHV1-69 and IGHV4-39, with a binding mechanism nearly identical for each antibody derived from different donors. Our results suggest that IGHV1-69 and IGHV4-39, while part of the adaptive immune system, may have evolved under selection pressure to encode a binding motif innately capable of recognizing and neutralizing a structurally conserved protein domain involved in pathogen iron acquisition.

3.3.1 Introduction

Staphylococcus aureus is a major human pathogen that can cause significant morbidity and mortality with a wide range of clinical manifestations(1). These include bacteremia, pneumonia and infective endocarditis as well as osteoarticular, skin and soft tissue, and device-related infections1. The clinical burden is further exacerbated by the increasing occurrence of antibiotic resistance, in particular the raise of methicillin-resistant strains (MRSA)2. At the same time S. aureus is a human commensal that is carried persistently (20–30%) or transiently (Z50%) on the skin and in the nares of the general population1, with the majority of individuals never experiencing an overt infection episode. This remarkable commensal relationship, likely of evolutionary origins, affords the opportunity to study the immune response to a bacterial pathogen to which humans are exposed on a continuous or recurrent basis over their lifetime.

To explore the role of the immune response in this host/ microbe interaction, we focused our attention on the mechanism used by S. aureus to obtain the iron necessary for colonization and pathogenesis3. S. aureus steals iron from haemoglobin, the most abundant iron source within vertebrates, through the concerted activity of the proteins in the iron-regulated surface determinant (Isd) locus4. IsdB in particular, a surface-exposed protein covalently anchored to the cell wall5, functions as a central component of this pathway by removing heme from haemoglobin and transferring it to other Isd proteins, which in turn import and degrade it to release iron in the bacterial cytoplasm4. IsdB contains two structurally conserved NEAT (NEAr iron Transporter) domains than bind haemoglobin and heme, respectively6–9. NEAT domains represent a structurally conserved heme-binding protein fold encoded in the genomes of several other Gram-positive human pathogens such as Bacillus anthracis, Streptococcus pyogenes, Clostridium perfringens and Listeria monocytogenes10,11. Importantly, S. aureus strains lacking IsdB or with IsdB mutants unable to bind haemoglobin, display a reduction in virulence in animal models of staphylococcal infection6,12. It was also previously shown that a recombinant anti-IsdB antibody was able to confer protection against S. aureus infections in animal models13.

High-serum titres against lsdB, as well as other S. aureus proteins, are readily observed in healthy adults14. While their biological significance and possible role in protection against infection remains to be elucidated, it has been shown that serum titres against IsdB are elevated during infection15,16. Therefore, to gain a better understanding of the functionality of these antibodies and to explore the relationship between the human immune system and the commensal pathogen S. aureus, we used single-B cell cloning, phage display libraries, high-throughput sequencing and epitope mapping17, structural and mutagenesis methodologies to characterize in detail the humoral immune response to the staphylococcal protein IsdB.

3.3.2 Results

Persistence of IsdB-reactive memory B cells. We first determined the presence of IsdB-reactive B cells in blood samples collected from a donor (D3) at months 1, 3

and 15 using flow cytometry (FACS), single-cell cloning (Supplementary Fig. 1) and high-throughput sequencing techniques. By FACS, we observed the persistence of a distinct IsdB-reactive memory B cell population (˜0.06%) within the IgM negative peripheral memory compartment (Fig. 1a). The majority of the IsdB-reactive memory B cells collected at three different time points expressed clonally related B cell receptor (BCR) transcripts: 25 of the 31 unique IsdB-reactive clusters obtained from this donor contained sibling transcripts isolated from at least two different time points (Fig. 1b). Longitudinal lineage analysis of the heavy chain variable region of these clusters indicates that the immune system maintains a repertoire of continually evolving antibodies against IsdB, presumably as a consequence of steady or intermittent exposure to low levels of antigen due to the commensal relationship between humans and S. aureus (Supplementary Fig. 2).

Molecular characterization of the anti IsdB antibodies. To investigate the nature of the interaction of these related antibodies with their target antigens we cloned the heavy and light chain BCR transcripts from single IsdB-reactive IgM CD19 + CD27 + memory B cells obtained from the peripheral blood of four healthy donors with high-antibody titre to IsdB (D1-4) (Supplementary Figs 1 and 3). Overall, 75 unique antibodies representing 438 single-cell BCR transcripts were confirmed to bind both recombinant IsdB as well as IsdB on the surface of iron-starved S. aureus cells. Next, we determined their epitope binning, epitope mapping, affinity and ability to block haemoglobin binding to IsdB (Fig. 1c; Supplementary Data 1 and Supplementary Figs 4–7). We found the majority of the antibodies to be directed against epitopes on the conserved core of IsdB (NEAT1–Linker–NEAT2) (Fig. 1c; Supplementary Fig. 8a) with two prominent sets of antibodies (bins C and P) that block haemoglobin binding to IsdB, targeting NEAT1 and NEAT2, respectively. Antibodies belonging to both sets are found in all four donors (Fig. 1c; Supplementary Data 1). Surprisingly, among the antibodies that bind NEAT1, there is a strong bias (7 out of 12 antibodies) towards using the immunoglobulin heavy chain variable gene (IGHV) 4-39 and immunoglobulin kappa light chain variable gene (IGKV) (Supplementary Fig. 9a), whereas the antibodies that bind NEAT2 are invariably derived from the IGHV1-69 germline and IGKV light chains (Supplementary Fig. 9b), irrespective of

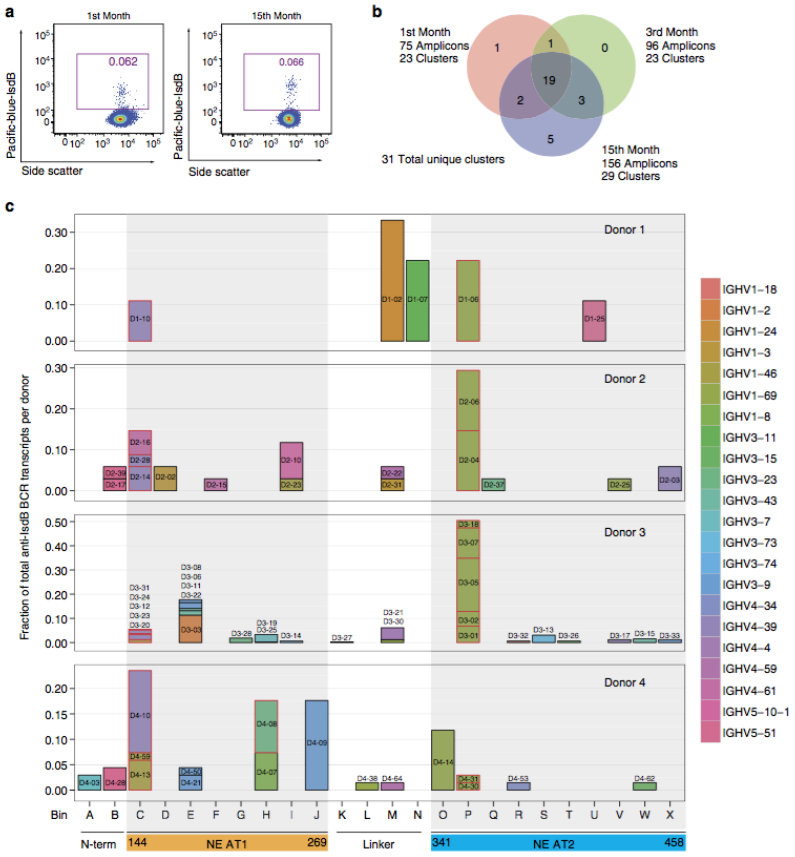
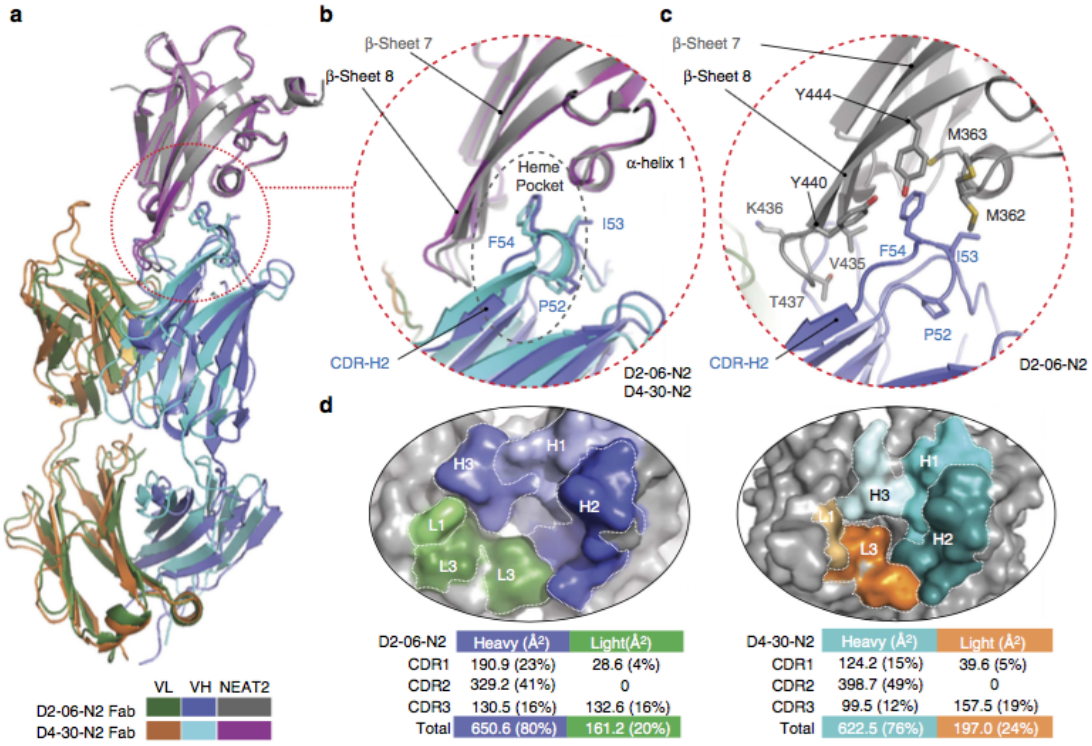


Figure 3.10: Characterization of anti-IsdB antibodies in the human memory B cell repertoire. (a) A persistent population of IsdB þ memory B cells from the peripheral blood mononuclear cells (PBMC) of a donor (D3) was observed over a 15-month period. By FACS, B0.06% of the IgM CD19 + CD27 + memory B cells in the total memory B cell repertoire of this donor bind IsdB. (b) Most of the cloned BCR transcripts of the IsdB þ memory B cells collected at month 1, 3 and 15 are clonally related. BCR sequences from single-cell cloning of IsdB + memory B cells were clustered based on heavy chain V-gene usage and CDR-H3 sequences. In total, we identified 31 unique clusters from donor D3 over the three collection time points. The Venn diagram shows that sibling clones within a cluster can be isolated at multiple time points. (c) Two distinct sets (bin C and bin P) of function-blocking antibodies specifically target NEAT1 and NEAT2, respectively. Single-cell cloning was performed at three different time points for donor D3 and one time each for donors D1, D2, and D4. In total, 75 unique antibodies targeting IsdB were identified and characterized. Shown here are the results of a comprehensive epitope binning analysis of 67 antibodies. Each reformatted clone is shown as a box and coloured according to its VH germline usage. The height of the box indicates the number of clustered BCR transcripts represented for each reformatted clone. There are in total 9, 34, 327, and 68 anti-IsdB single-cell BCR transcripts for D1, D2, D3 and D4, respectively. Each column of clones represents an epitope bin and this is overlaid on top of a linear representation of the IsdB molecule with NEAT1 in orange, and NEAT2 in blue. Clones that are able to fully block haemoglobin binding are outlined with a red box.

the donor of origin. Remarkably, we found several of these antibodies to have affinities in the single digit pico-molar range at 37 °C (Fig. 2e; Supplementary Fig. 6).

Characterization of IGHV1-69-derived NEAT2 binders (Bin P). To elucidate the binding mechanism between the IGHV1-69derived antibodies and NEAT2, we determined two crystal structures of Fabs from two different donors in complex with NEAT2 (Fig. 2a and Table 1; Supplementary Fig. 10a,b). Both structures, denoted by D2-06-N2 (3.22Å) and D4-30-N2 (3.21 Å), surprisingly exhibited a nearly identical NEAT2 binding mode (Fig. 2a), with the heavy chain variable domain (VH), particularly complementary determining region (CDR)-H2, dominating the interaction in both structures (Fig. 2b). Specifically, the V-gene encoded CDR-H1 and CDR-H2 contribute 64% of the total buried surface area (BSA) for both structures (Fig. 2d). This heavy reliance on CDR-H1 and CDRH2 is unusual for antibody/antigen interactions as highly diverse, VDJ recombination-generated CDR-H3s are typically considered to be the most important CDR for antigen binding18,19. The CDR-H2s of the Fabs engage the NEAT2 domain in two major modes. First, the b7-turn-b8 of NEAT2 slides into a groove at the interface of heavy and light chain variable regions, forming major contacts with the stem of the CDR-H2 loop. Second, F54 (Kabat numbering) of CDR-H2 protrudes into the hydrophobic heme pocket of NEAT2, made up of M362, M363 and F366 in the ahelix 1, V435 on the b7 and Y440 and Y444 on the b8 of IsdB (Fig. 2b,c). While this group of NEAT2 binding antibodies was initially found to block haemoglobin binding to IsdB, which likely occurs by steric hindrance as IsdB NEAT1 and NEAT2 are proposed to be spatially adjacent to each other based on their homology to the solved crystal structure of haemoglobin bound to IsdH NEAT2–linker–NEAT3 (ref. 20), the structural data reveals a second very effective way to block the activity of IsdB as antibody binding to the heme pocket precludes the possibility of concurrent heme binding9. Given the highly conserved nature of S. aureus NEAT2, particularly at the binding interface (Supplementary Fig. 8), we predict that these IGHV1-69-derived antibodies will be able to recognize and neutralize IsdB encoded by the vast majority if not all S. aureus strains (4,112 strains analysed in this study).



D2-06-N2	Heavy (Å²)	Light(Å²)
CDR1	190.9 (23%)	28.6 (4%)
CDR2	329.2 (41%)	0
CDR3	130.5 (16%)	132.6 (16%)
Total	650.6 (80%)	161.2 (20%)

D4-30-N2	Heavy (Å²)	Light (Å²)
CDR1	124.2 (15%)	39.6 (5%)
CDR2	398.7 (49%)	0
CDR3	99.5 (12%)	157.5 (19%)
Total	622.5 (76%)	197.0 (24%)

e

	CDR-H2 (Kabat)							Antibody		NEAT2		Antibody	
Clones	50	51	52	52a	53	54	55	K_D	F54A	V435R	D390A/K436A/T437A	Germline	Hb block
VH1-69*01	G	I	I	P	I	F	G	(pM)	(Folds)	(ELISA)	(ELISA)		
D1-06-N2	–	–	V	–	–	–	–	9–17	135	3%	3%	ELISA+	ND
D2-04-N2	–	–	–	–	–	–	–	20–25	>40,000	3%	3%	ND	ND
D2-06-N2	–	–	S	–	–	–	–	13–16	>62,000	3%	3%	24.1 nM	Yes
D3-01-N2	–	–	V	–	–	–	–	10–22	>10,000	4%	3%	40.5 nM	Yes
D3-02-N2	–	S	–	–	–	–	–	2,480–2,500	>400	4%	4%	665 nM	Yes
D3-05-N2	A	–	V	–	–	–	–	63–110	>9,000	3%	3%	56.9 nM	Yes
D3-07-N2	–	–	–	–	–	–	–	156–255	>3,900	3%	4%	ELISA+	ND
D3-18-N2	–	–	S	–	–	–	–	29–50	>20,000	3%	3%	30.8nM	Yes
D4-30-N2	–	–	M	–	–	–	A	21–31	No Binding	4%	4%	ELISA+	ND
D4-31-N2	–	–	–	–	–	–	–	58–292	>34,000	4%	4%	716 nM	Yes

Figure 3.11: Germline-encoded binding of IGHV1-69 to the NEAT2 domain of IsdB. (a) Crystal structure of IGHV1-69-derived Fabs from two donors (D2-06-N2 and D4-30-N2) in complex with NEAT2. The Fabs of D2-06-N2 and D4-30-N2 show a near identical binding mechanism to NEAT2 as evidenced by the superimposed structures. To facilitate the crystallization process, a sandwiching Fab from an antibody (D3–13) that binds NEAT2 at a non-overlapping epitope was used. For clarity, the sandwiching Fab is removed from the figure, but is included in the Supplementary Data (Supplementary Fig. 10a,b). (b) Both IGHV1-69-derived antibodies use the conserved F54 on CDR-H2 to interact with the heme-binding pocket of NEAT2. The stem of the CDR-H2 loop also mediates major contacts with the b7-turn-b8 loop of NEAT2. (c) The heme pocket residues of NEAT2 which interact with the conserved F54 on CDR-H2 are highlighted in the complex with D2-06-N2. They are M362, M363 and F366 in a-helix 1, V435 on the b-strand 7, and Y440 and Y444 on the b-strand 8. (d) CDR-H2 dominates the interaction in terms of BSA in both structures. Structural analysis shows that 75–80% of the BSA is attributed to the heavy chain, and 20–25% to the light chain. In particular, the CDR-H2 contributes 41 and

We next used mutagenesis to determine if the other eight IGHV1-69-derived antibodies in this set bind IsdB in a similar manner, as suggested by the fact that they all share the IGHV1-69 framework and a conserved F54 in CDR-H2. Figure 2e shows that a single mutation at the surface-exposed F54 position (F54A) resulted in 4100-fold loss in affinity for all of the antibodies, highlighting that its importance for binding is shared by every antibody in this set. To further substantiate the common binding mode, we generated NEAT2 variants with mutations at the binding interface residues. Mutations in the heme pocket (V435R) and at the base of the b7-turn-b8 motif (D390A/ K436A/T437A; Fig. 2c) consistently disrupted binding to every antibody in the NEAT2 binding set without affecting binding of antibodies that also bind NEAT2 but belong to different epitope bins (Fig. 2e). We observed slight differences in the extent to which mutations at several NEAT2 residues impacted the binding of the antibodies (Supplementary Fig. 11a,c), presumably due to subtle differences in how the distinct CDR-H3 and CDR-L3 of each antibody contribute to binding IsdB. Overall, the combination of structural and mutagenesis data indicates that IGHV1-69derived antibodies from four donors bind NEAT2 in similar fashion by primarily using CDR-H2 germline residues.

Having established the prominent role of CDR-H2 in binding NEAT2 of IsdB, we determined if these antibodies have other commonalities by first examining the contribution of the individual CDR-H3 and J-region residues based on the two crystal structures and then by comparing the CDR-H3 amino acid usage of the 10 isolated binders to that of IGHV1-69-derived antibodies in memory repertoires of twelve healthy donors (Supplementary Fig. 12). For the heavy chain, we did not observe any common residue on CDR-H3 and JH that contributes substantially to the binding to NEAT2. Sequence analysis of the 10 binders also did not reveal any particular preference in immunoglobulin heavy chain joining segment (IGHJ) usage as IGJH1, 3 and 4 were all used (Supplementary Data 1). There seems to be a bias for charged residues (D, K or R) at position 95 and glycine at position 96, but both positions only have minor contributions to the overall binding based on the structures (Supplementary Fig. 12). As for the light chain, which overall contributes only 20–24% to the binding surface, there is no apparent preference for the IGKV or immunoglobulin

kappa light chain joining segment (IGKJ) usage based on the 10 binders. Interestingly, the CDR-L3s of all the binders are 11 amino acids long, likely resulting from a direct fusion of IGKV and IGKJ genes. On the basis of the crystal structures, the two aromatic residues at position 94 and 96, which form a distinct motif (F/WP-W/Y), are responsible for the majority of the CDR-L3 contribution to the binding. This motif was also found in another three binders, and a similar X-P-X motif was also found in four of the remaining five binders, suggesting a potential preference for light chain having a specific pattern at position 94– 96 in pairing with IGHV1-69-derived heavy chain to bind NEAT2.

Germline-reverted variants of IGHV1-69-derived antibodies. The shared binding mode of the IGHV1-69-derived antibodies led us to hypothesize that this germline has inherent potential to recognize the NEAT2 domain of IsdB. This hypothesis was tested by reverting the heavy chain V-gene region (framework 1 to framework 3) of multiple clones from each donor to their respective germline precursor sequences and testing their binding to IsdB NEAT2 by ELISA and biosensor. While we observed significant losses in monovalent affinity for all the germlinereverted clones (4600 fold compared with the originally isolated clones), all of them were still able to bind IsdB NEAT2 by ELISA with four germline-reverted clones, D2-06-N2, D3-01-N2, D3-05-N2 and D3-18-N2, having surprisingly high-monovalent affinity (24 nM–60 nM) to NEAT2 (Fig. 2e). In contrast, non-matured antibodies from na¨ıve B cells typically bind antigen with high micro-molar affinity and binding can only be reliably detected in avidity-driven assays21,22. This data further supports the hypothesis that the IGHV1-69 germline possesses inherent capability to bind IsdB NEAT2. Also, these germline-reverted antibodies retain the ability to block haemoglobin binding (Fig. 2e; Supplementary Fig. 7).

Allelic preference of IGHV1-69-derived antibodies. Moreover, we determined that the binding of IGHV1-69-derived antibodies to IsdB is strongly influenced by the allelic variation at position 50 (Fig. 3a). We found that the presence of R50, in contrast to G50 or A50 as in the antibodies described here, completely abolished the binding (Fig. 3b,c), presumably due to the steric clash between the extended side-chain of R50 and the IsdB b7-turn-b8 of NEAT2 (ref. 23; Fig. 3d). Data from

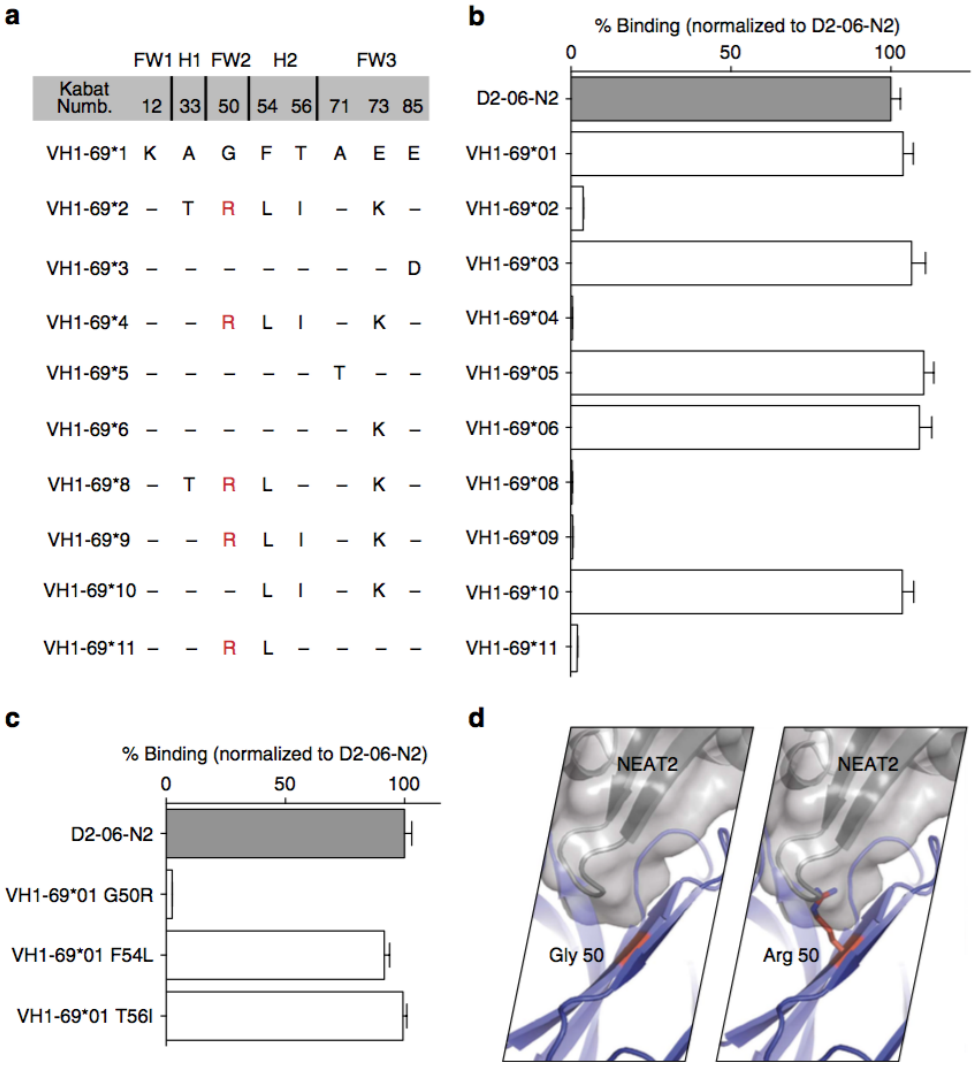
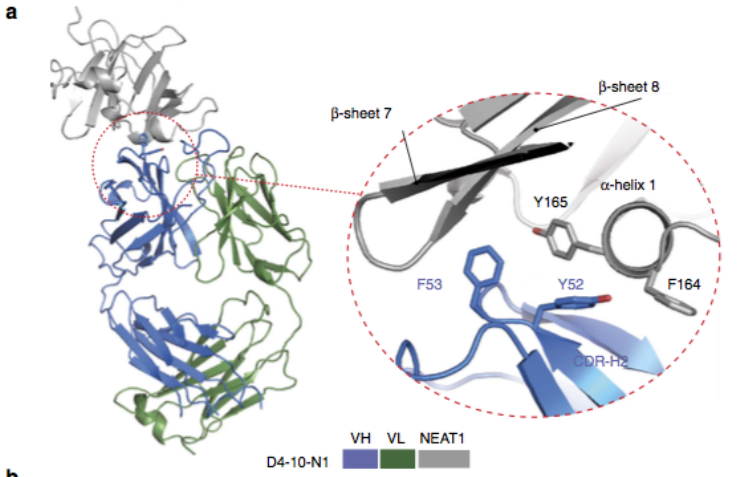


Figure 3.12: Allelic specificity of IGHV1-69-derived NEAT2 binders. (a) Amino acid differences among functional alleles of IGHV1-69. (b) The VH of clone D2-06-N2 was germline-reverted to all alleles with amino acid differences, and tested for binding to IsdB by ELISA. All alleles with a G50R substitution lost binding. ELISA data is an average of three independent experiments. Error bars are defined as s.d. (c.) Three individual variants (G50R, F54L and T56I) of D2-06-N2 (IGHV1-69*01 germline-reverted) were generated and their binding to IsdB was tested. Only variant G50R showed significant loss of binding. ELISA data is an average of three independent experiments. Error bars are defined as s.d. (d) Analysis of the structure illustrates how a change from G to R (most frequent rotamer) at position 50 is expected to cause a steric clash in the binding to NEAT2.

the 1000 Genomes Project show that the R50 polymorphism, which abolishes NEAT2 binding, is present in 38% of the IGHV1-69 allele with a predicted homozygous rate of 14% in the general population24. This raises the possibility that there is a population-level difference in the ability to neutralize NEAT2-mediated heme-iron acquisition and therefore different susceptibility to S. aureus infection25,26.

Characterization of IGHV4-39-derived NEAT1 binders (Bin C). We also characterized a second class of antibodies that are derived from IGHV4-39 and bind to the NEAT1 domain of IsdB. We determined the crystal structure of the D4-10-N1 Fab in complex with NEAT1 (3.17Å; Fig. 4a and Table 1; Supplementary Fig. 10c). The structure reveals that binding is again dominated by the heavy chain, particularly by CDR-H2, which contributes 45% of the BSA (Fig. 4b). Specifically, D4-10-N1 utilizes CDR-H2 residues Y52 and F53 to interact with residues Y165 of NEAT1, targeting the same binding region that is responsible for the interaction between haemoglobin and NEAT1 (ref. 6; Supplementary Fig. 13) and therefore providing a mechanistic explanation on how antibodies in this group block haemoglobin binding. Remarkably, CDR-H2 F53 protrudes into a hydrophobic pocket of NEAT1 that is structurally homologous to the heme pocket of NEAT2. Therefore this resembles the IGHV169 CDR-H2 interaction with NEAT2 (Supplementary Fig. 14). All antibodies in the NEAT1-binding group have a conserved aromatic residue (Y or F) at positions 52 and 53, and lost binding to NEAT1 when these residues were mutated to A (Fig. 4c). Correspondingly, mutations of NEAT1 at residue Y165 abolished binding for every antibody in this group without disrupting the binding of antibodies that also bind NEAT1 but belong to different epitope bins (Fig. 4c and Supplementary Fig. 11b,d). Collectively, the structural and mutational data strongly suggest that all of the antibodies in this set interact with NEAT1 in a similar fashion. Similar to NEAT2, the sequence of NEAT1 is also highly conserved (Supplementary Fig. 8), therefore we expect these IGHV4-39-derived antibodies to be able to recognize and neutralize IsdB encoded by the vast majority if not all S. aureus strains.

Among the seven IGHV4-39-derived NEAT1 binders there are no apparent preferences on the usage of IGKV germlines, IGKJ and IGHJ regions (Supplementary Data 1). Analysis of the amino acid usage in CDR-H3 (Supplementary Fig. 15) revealed



D4-10-N1	Heavy (Å²)	Light (Å²)
CDR1	118.4 (15%)	53.9 (7%)
CDR2	362.8 (45%)	0
CDR3	161.4 (20%)	112.6 (14%)
Total	643.5 (79%)	166.5 (21%)

c

	CDR-H2 (Kabat)				Antibody		NEAT1	Antibody	
Clones	52	53	54	55	K_D	52,53 -> AA	Y165R	Germline Hb block	
VH4-39*01	Y	Y	S	G	(pM)	(ELISA)	(ELISA)		
D1-10-N1	–	F	–	–	37–121	5%	3%	6.7 nM	Yes
D2-14-N1	F	F	–	–	4–12	2%	3%	6.1 nM	Yes
D3-12-N1	–	F	–	–	217–1,152	2%	3%	32.5 nM	Yes
D3-23-N1	F	F	–	–	3–7	3%	3%	56.6 nM	Yes
D3-24-N1	–	–	–	–	24–36	2%	5%	160 nM	Yes
D3-03-N1	F	F	–	–	686	2%	3%	1170 nM	ND
D4-10-N1	–	F	–	–	<1	9%	3%	90.7 nM	Yes

Figure 3.13: Germline-encoded binding of IGHV4-39 to the NEAT1 domain of IsdB. (a) Crystal structure of an IGHV4-39-derived Fab (D4-10-N1) in complex with NEAT1. The two aromatic resides (Y52 and F53) in CDR-H2 interact with the a-helix1 of NEAT1 which is normally involved in binding haemoglobin. IGHV4-39 CDR-H2 F53 of Fab D4-10-N1 protrudes into a hydrophobic pocket of NEAT1, which is structurally homologous to the heme binding pocket of NEAT2. Crystallization was facilitated by the use of a sandwiching Fab from an antibody (D3–19) that binds NEAT1 at a non-overlapping epitope (bin H). For clarity, the sandwiching Fab is removed from the figure but is included in the Supplementary Data (Supplementary Fig. 10c). (b) CDRH2 dominates the interaction in terms of BSA. Structural analysis shows that 79% of the BSA is attributed to the heavy chain, and 21% to the light chain. The CDR-H2 contributes about 45% of total BSA. (c) Mutational analysis confirms the structural data and demonstrates that all IGHV derived antibodies in this set bind NEAT1 with a similar mechanism. The KD for all antibodies in this set was determined by SPR-based biosensor binding analysis to recombinant full-length IsdB at 37 °C (KD range, nZ2). The binding of antibody variants at positions 52 and 53 of CDR-H2 to wild type IsdB and the binding of antibodies to NEAT1 variant Y165R (a-helix 1) were evaluated by ELISA (percentage binding relative to binding between original isolated antibodies and wild type IsdB, one representative set of results out of three independent experiments is shown). Every clone was reverted to

a strong underrepresentation of R at position 94, as S, T and K were used instead. This may affect the typical salt-bridge connection between R94 and D101, which structurally supports the CDR-H3 loops. Four out of the seven binders are missing the typical pairing of K/R94 and D101. In addition, we also observed that P or G, which can alter the backbone conformation, are preferred at position 95. Therefore it is plausible that even though both residues do not mediate direct binding based on the crystal structure, they could co-operate to uniquely orient the CDR-H3 residues in these IGHV4-39-derived NEAT1 binders. Charged residues were also found preferentially at position 99 and 100 among the binders. However, both residues do not have any direct contact with NEAT1 in the crystal structure of D4-10-N1. The sequence/structure analysis did not reveal any common contact residue among the CDR-H3s of the seven IGHV4-39IGKV-derived binders.

Germline-reverted variants of IGHV4-39-derived antibodies. Given the prominent role of the IGHV4-39 germline-encoded CDR-H2 in binding IsdB NEAT1, we next measured the affinity of heavy chain V-gene germline-reverted (framework 1–framework 3) antibodies for all IGHV4-39 antibodies. Remarkably, all IGHV4-39 germline-reverted antibodies exhibited very high affinities (with KD values at 37 °C in the singleto triple-digit nanomolar range; Fig. 4c), supporting the idea that the IGHV4-39 germline has intrinsic potential for recognizing IsdB NEAT1. These germline-reverted antibodies can also block haemoglobin binding (Fig. 4c; Supplementary Fig. 7). This feature appeared to be specific for IGHV4-39, as reverting selected antibodies to two other highly homologous germlines23, IGHV4-30*04 and IGHV4-61*01, resulted in significant loss of binding for the antibodies evaluated (Fig. 5). Unlike the IGHV1-69 NEAT2 binders, allelic variation did not appear to affect the ability of IGHV4-39-derived clones to bind NEAT1 (ref. 23; Fig. 5).

IGHV4-39 encoded NEAT1 binders from naı¨ve B cells. To expand the breadth of our findings we first tested serum samples from 36 donors (including the original 4 donors) against the two NEAT domains of IsdB and show that there are detectable titres against both NEAT domains (Supplementary Fig. 16). Moreover, these titres were reduced by pre-blocking the NEAT domains with antibodies that bind the

haemoglobin and heme-binding sites, suggesting that antibodies that target the functional domains of IsdB are present in the serum of all donors tested (Supplementary Fig. 16).

Next, given that the majority of the IsdB NEAT domain binding was primarily driven by germline-encoded CDR-H2, we investigate if antibodies from na¨ıve B cells can recognize IsdB in a similar manner as the one described above and asked if such antibodies could be found in additional donors. Using individually barcoded IGHV4-39 specific primers, we selectively amplified the IGHV4-39 variable heavy chain gene from the cDNA of sorted CD19+ CD27IgM+ naive B cells of 36 individuals; this allowed us to unequivocally match binders with their respective donors. A single-chain Fv (scFv) phage display library was then constructed by pairing the individually barcoded IGHV4-39 VH gene with the pooled na¨ıve IGKV families 1–4 genes from all donors (schematics are shown in Fig. 6a). Sequencing of the starting phage library confirmed that the heavy chain of more than 90% of the clones was encoded by IGHV4-39, with minor contaminations from other IGHV4 family members. After four rounds of panning against IsdB NEAT1 domains, the binding of 960 individual phage clones against IsdB and its variants was evaluated by ELISA. About 90% of the clones showed specific binding to full-length IsdB and IsdB NEAT1 domain. Sequence analysis determined that three of the phage clones with unique CDR-H3 (D14-1, D15-1 and D16-1) represented the majority of the binders (96%); this could be due to their superior affinity as scFv's (not determined) or to a growth bias introduced through the phage amplification process. Despite the presence of these three dominant clones, we were able to isolate a total of 16 clones with unique CDR-H3 sequences from 13 different donors (Supplementary Fig. 17).

The binding characteristics of the isolated clones were then evaluated. Remarkably all the unique phage clones lost binding to IsdB NEAT1 variant Y165R, suggesting that all the clones bind the haemoglobin-binding pocket on NEAT1 (Supplementary Fig. 17). Next, one clone from seven different donors was randomly selected and reformatted as IgG (Fig. 6b and Supplementary Fig. 18). Consistent with the phage binding results, all seven reformatted IGHV4-39-derived antibodies maintained the ability to recognize IsdB NEAT1 and all lost binding to the IsdB Y165R variant as

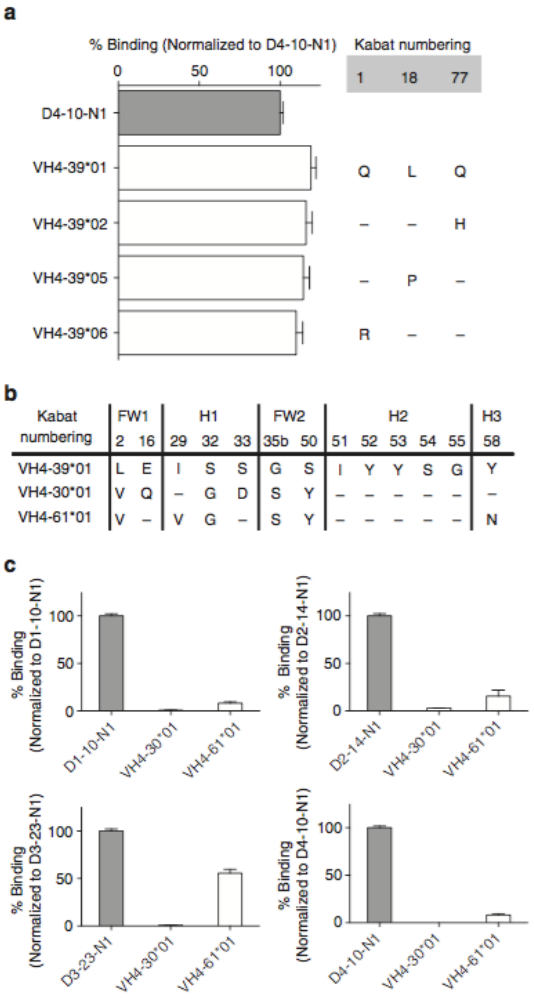
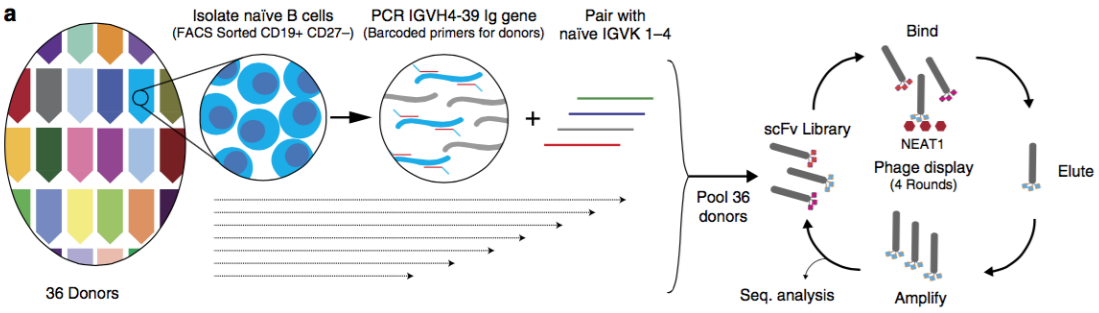


Figure 3.14: Germline and allelic specificity of IGHV4-39-derived NEAT1 binders. (a) Clone D4-10-N1 was reverted to all allelic variants with amino acid differences relative to IGHV4-39*01. No differences in IsdB binding were observed. Results shown are an average of three independent experiments. Error bars are defined as s.d. (b) IGHV4-39*01 has high sequence homology to IGHV4-30*04 and IGHV4-61*01 (they only differ by six amino acids in the variable region), and they all have the critical Y52 and Y53 residues in the CDR-H2. (c) The VH of four clones, one from each donor, were reverted to both IGHV4-30*04 and IGHV4-61*01, and their binding to IsdB was tested by ELISA. All IGHV4-30*01-derived variants were unable to bind IsdB, while most of the IGHV4-61*01-derived variants exhibited significantly loss of binding. ELISA data is an average of three independent experiments. Error bars are defined as s.d.

measured by Elisa (Supplementary Fig. 18). We confirmed binding of these antibodies to IsdB in a monovalent based biosensor assay at 37°C, while their Y52A/Y53A variants lost the ability to bind IsdB (Supplementary Fig. 18). Overall, these results showed that the na¨ıve IGHV4-39 clones isolated from the na¨ıve B cells of 13 additional donors bind IsdB NEAT1 in a manner similar and consistent with the binding interaction described above for the antibodies isolated from the memory B cells of four donors. This further strengthens the suggestion that IGHV4-39 possesses inherent affinity toward the NEAT1 domain of IsdB of S. aureus.

We next used the sequence information from these phagederived NEAT1 binders to further examine the immunoglobulin gene use looking for any additional features that may be common toallIGHV4-39-derivedbinders.First,thena¨ıvenatureofthese phage binders allows DH gene usage to be more reliably identified; we did not observe any preference for specific DH genes (Supplementary Fig. 17). Meanwhile, similar to what was observed for the NEAT1 binders isolated from memory B cells, we identified a strong underrepresentation of R at position 94 (G, S and T are used instead), as 14 of 16 phage binders (88%) do not have the typical salt-bridge pair of R94-D101. This is in contrast to the starting library, where 77% of the phage clones have R94 and D101 (Supplementary Fig. 19). In addition, the remaining two clones possessing the R94-D101 pair have P at position 95. This was also frequently observed in the memory B-cell-derived binders (Supplementary Fig. 15). Besides position 94 and 95, there is no apparent preference for amino acid usage in the CDR-H3 (Supplementary Fig. 19). Unexpectedly, IGHJ3 was exclusively used for all of the phage isolated binders. This strong bias of IGJH3 was not observed in our seven NEAT1 binders isolated from memory B cells, as all IGHJ1-6 genes were used. It is possible that this bias was introduced by displaying antibodies as scFv or by an inherent bias of the phage selection process. As for the light chain, IGKV3-20 was the most frequently used variable kappa light chain (VK) germline in these binders (11 out of 16), similar to what we observed from the memory B cells-derived NEAT1 binders (3 out of 7).



b

Donor#	VH germline	CDR-H3	VH FW Mutation#	VK germline	Antibody binding to 1 uM of IsdB (Biosensor) Parental	52,53->AA	ELISA binding IsdB Y165R
D6-1	IGHV4-39*01	CATYYYDSSGYYFPDAFDIW	1	IGKV3-20	+	–	–
D14-1	IGHV4-39*01	CASWGGNSGVDAFDIW	0	IGKV3-20	+	–	–
D16-1	IGHV4-39*07	CARPHSSSYDSLGAFDIW	1	IGKV3-20	+	–	–
D18-1	IGHV4-39*01	CAGIGSGSYYNPPDAFDIW	1	IGKV3-15	+	–	–
D28-1	IGHV4-39*01	CAGRAGSGSFDDAFDIW	0	IGKV3-20	+	–	–
D32-1	IGHV4-39*01	CAGLSSSSGYANDAFDIW	0	IGKV3-11	+	–	–
D40-1	IGHV4-39*01	CATYYGSGSGAFDIW	0	IGKV3-20	+	–	–

Figure 3.15: Naı¨ ve IGHV4-39-derived antibodies from naı¨ ve B cells. (a) Schematics of na¨ıve IGHV4-39 antibody phage library IgM þ na¨ıve B cells were isolated individually by FACS from 36 donors. Total Ig RNA was converted into cDNA using an IgM then uniquely barcoded IGHV4-39 primers for each donor were used to selectively amplify the IGHV4-39 VH gene from the cDNA. The amplified VH genes were then pooled together and paired with the light chain variable genes from IGKV families 1–4 amplified from the same set of donors to generate the single-chain Fv library. Antibody libraries were then displayed on phage and 4 rounds of panning against recombinant IsdB NEAT1 were performed. (b) Binding characterization of IGHV4-39 encoded na¨ıve NEAT1-binding antibodies from seven different donors. Heavy and light chain germlines usage, CDR-H3 sequence identities and number of variable heavy chain framework nucleotide mutation of the seven NEAT1 binders are shown. The observed single framework mutation in selected clones may have been introduced by the amplification process during library generation. The binding of the parental antibodies and their Y52A/Y53A variants to full-length IsdB was determined by SPR-based biosensor binding analysis at 37 °C. The binding of the parental antibodies to the full-length IsdB Y165R variant was determined by ELISA (n 1/4 2).

3.3.3 Discussion

In this study, we characterized in detail the endogenous humoral response in healthy individuals against IsdB, a prominent molecule in the iron-acquisition pathway necessary for colonization and pathogenesis of the commensal bacterium S. aureus. We found that the human immune system maintains a sizable repertoire of continually evolving IsdB-reactive memory B cells by comparing the repertoire of a NEAT-2 binder over a 15-month period (Supplementary Fig. 2). For IsdB, high-serum titres are representative of antibodies that bind to a variety of epitopes on IsdB. Among them, we identified two prominent sets of neutralizing antibodies that target the specific fold of the highly conserved NEAT domains of IsdB with a dedicated, highly specific V-gene response for each NEAT domain, IGHV4-39 for NEAT1 and IGHV1-69 for NEAT2. Interestingly these neutralizing antibodies are not derived from the IGHV3 family which encode antibodies targeted by the S. aureus virulence factor Protein A27,28. These antibodies bind with very high affinity and neutralize the activity of IsdB by occupying the structurally homologous active regions on NEAT1 (Hb binding) and NEAT2 (heme binding) (Supplementary Fig. 14), respectively. The affinities of these antibodies, with dissociation constants measured at 37 °C in the low pM range in all four donors, are beyond the proposed in vivo affinity maturation limit of 100 pM to 1nM (ref. 29). This perhaps results from the continuous or intermittent exposure to very low levels of antigen as a consequence of the commensal nature of S. aureus. Low antigen levels may only be recognized by B cell clones expressing these high affinity BCRs, thereby favoring their specific stimulation and subsequent selection in the context of the continually maturating B cell memory compartment.

We have isolated, by single-cell cloning, IGHV1-69-derived antibodies against NEAT2 and IGHV4-39-derived antibodies against NEAT1 from memory B cells from four donors and additional IGHV4-39-derived antibodies against NEAT1 from na¨ıve B cells from 13 donors using a phage display approach. We show by a combination of structural and mutagenesis data that these antibodies bind the respective NEAT domains in a similar fashion. The binding interaction is primarily driven by the germline-encoded hydrophobic CDRH-2 motifs of IGHV1-69 and IGHV4-39. In fact, we also show that germline-reverted versions of these antibodies demonstrate

specific binding to their respective NEAT domains and maintain the ability to block the interaction with haemoglobin. Sequence analysis of the CDR3s of the binders revealed positional preferences of certain amino acid residues, for example a strong underrepresentation of the canonical R at position 94 of the heavy chain for NEAT1 binders and the presence of aromatic residues at positions 94 and 96 of the light chain for NEAT2 binders. However, there is no conserved residue in CDR-H3 among the binders that mediates major contacts with the NEAT domains. Given the large size of the human antibody repertoire30,31 and the importance of the highly diverse CDR-H3 for binding19, antigen recognition that is predominantly driven by the invariant V-gene encoded CDR-H2 is rarely observed. Similarly, the occurrence of multiple antibodies isolated from different individuals that bind the same epitope with a similar mechanism is very rare and suggests a predetermined structural bias in the immune repertoire. Several examples have been reported in the literature to suggest a germline preference for antibody binding to certain antigens32,33. One noteworthy example is the recognition of the I/i antigen (N-acetyllactosamine) on red blood cells by IGHV4-34-derived IgM antibodies34,35. In absence of a co-crystal structure, mutagenesis studies suggested an unconventional binding mode as the germline-restricted interaction with the antigen is mediated by framework 1 residues and the C-terminal region of CDR-H3 (refs 36,37). Another example of germline-restricted response is represented by antibodies against the capsular polysaccharide of Haemophilus influenzae type b. A strong preference for VH3 families usage was identified, in particular IGHV3-23 and IGHV3-15 (refs 38,39). A subset of these antibodies has a binding interface characterized by IGHV3-23 with short CDR-H3 and IGKV2D-29 with an R residues inserted in the V–J junction39,40. A structural model and limited mutagenesis suggested potential interactions from all CDRs with the short CDR-H3 and long CDR-L1 forming the floor of a groove flanked by CDR1s and CDR2s of the heavy and light chain41. In both cases, in the absence of definitive structural characterization, the binding modes appear to be clearly different from the CDR-H2-driven interaction described in our study.

Detailed structural analysis of the binding mechanism is needed to illustrate the fine details of these germline-restricted recognitions; however, such structural data

has largely not been available. The most prominent published body of work with detailed structural information is represented by two sets of IGHV1-69-derived broadly neutralizing antibodies that bind two distinct sites on the hemagglutinin molecule (HA) of the influenza virus. The first set of antibodies targets the receptor binding site on the globular head of HA42,43 (Supplementary Fig. 20) while the second set binds a structurally conserved epitope on the stem region of the molecule44–46 (Supplementary Fig. 21). Four antibodies from different donors that target the receptor binding site of HA, all with a heavy chain derived from IGHV1-69, were shown to recognize an overlapping epitope on the head of the molecule. This led to the suggestion of a IGHV1-69 germline preference for binding this region of HA. However, the structural data clearly shows that their binding modes are quite different from each other (Supplementary Fig. 20). This seems to stem from the fact that the binding interaction and orientation is driven by the different CDR-H3s of the different antibodies.

For the antibodies targeting the stem region of HA, IGHV1-69derived antibodies isolated from three different donors were shown to consistently use a signature motif on CDR-H2 (I53/M53 and F54) and Y98 from CDR-H3 to target the same hydrophobic groove on HA47,48. Previous work has also shown that germline-reverted versions of these broadly neutralizing antibodies were able to engage HA49, suggesting that the IGHV1-69 CDR-H2 motif is well suited to recognize a specific epitope on the HA stem50. Our results parallel the work on antibodies against the stem region of HA and show that the germline-encoded structural motif on IGHV1-69 CDR-H2 is not only well suited to bind the hydrophobic heme pocket of IsdB NEAT2 of S. aureus, but is also the essential determinant of this binding (Figs 2d and 3b). Unlike the IGHV1-69-HA-stem binding mechanism, our data shows that CDR-H2 is so crucial for binding that IGHV1-69-derived antibodies from different healthy donors exhibit a nearly identical binding mechanism to NEAT2 (Supplementary Fig. 21). Moreover, to our knowledge, our work illustrates for the first time that the IGHV4-39 CDR-H2 motif is also particularly well suited to recognize the haemoglobin-binding domain of IsdB NEAT1, uncovering another example of a germline specialized in binding to a common antigen. Altogether, the fact that both sets of antibodies were found in all

four donors tested and the results showing that the germline-reverted version of these antibodies maintain surprisingly high affinity for IsdB, further support the conclusion that the CDR-H2 motifs of IGHV1-69 and IGHV4-39 represent a particularly good fit to bind and neutralize the active sites on NEAT1 and NEAT2 of S. aureus, respectively. Because of the high degree of sequence identity among IsdB molecules encoded by S. aureus strains, we expect these antibodies to be broadly neutralizing across these strains. In addition, NEAT domains are structurally conserved with many homologues encoded in the genomes of Gram-positive bacteria in the phylum Firmicutes10 (Supplementary Fig. 22), therefore we speculate that IGHV1-69 and IGHV4-39 antibodies against other significant human pathogens such as Bacillus anthracis, Streptococcus pyogenes, Clostridium perfringens and Listeria monocytogenes may be already present in humans, or may be induced on antigen exposure.

Overall, our study reveals that two human germlines, IGHV1-69 and IGHV4-39, have inherent affinity against the specific folds of structurally conserved NEAT domains of a bacterial commensal pathogen. The results expand the concept of germline-restricted usage of antibodies suggested mostly for viral pathogens50–52 and extend this to a bacterial pathogenic protein. In addition, our work also illustrates for the first time how germline V-gene encoded residues can be so dominant in driving antibody binding that resulting antibodies from different individuals exhibit nearly identical binding mechanism. The data suggest that existing human V-genes may represent not only V(D)J recombination scaffolds for the antibodies of the adaptive repertoire, but also innate-like proto-receptor scaffolds to recognize certain unique structural motifs presented by infectious pathogens, such as the influenza virus and S. aureus. This may allow a proportion of the adaptive immune repertoire to activate rapidly and provide protection against a pathogen at the earliest encounter, without the time required for lengthy affinity maturation. Given the presumptive evolutionary advantage of these responses53, it is possible that pathogens to which humans are exposed seasonally or recurrently due to a commensal relationship may have exerted evolutionary pressure to promote the retention or expansion of specific V-gene segments in the human repertoires.

3.3.4 Methods

Isolation and clustering of anti-IsdB antibodies. Blood samples of 50–100 ml were collected from healthy consented donors. Drawing of blood samples was approved by the Pfizer Institutional Review Board. The blood sample was first diluted 1:1 with PBS/2% FBS/1 mM EDTA and centrifuged at 120g for 10 min with the brake off. The plasma fraction was removed and replaced with an equivalent volume of PBS/2% FBS/1 mM EDTA. The sample was further diluted 1:1 with PBS/2% FBS/1 mM EDTA and layered on top of Ficoll-Paque PLUS (GE Healthcare). After centrifugation at 2,000 r.p.m. for 20 min with brake off, PBMC were collected from the interface and washed twice with PBS/2% FBS/1 mM EDTA. ACK lysis buffer (Thermo Fisher Scientific) was added to remove the red blood cells. After washing the cells with PBS/2% FBS/1 mM EDTA, B cells from the PBMC were enriched using EasySep Human Pan-B Cell Enrichment kit (Stemcell Technologies) according to the manufacturer's protocol. After the enrichment, B cells were washed with PBS/2% FBS/1 mM EDTA and labelled at 8 °C for 2–4 h with anti-human CD3-PerCP/Cy5.5 (1:20 dilution, UCHT1), anti-human CD16 PerCP/Cy5.5 (1:20 dilution, 3G8), anti-human CD19-AlexaFluor488 (1:20 dilution, HIB19), anti-human IgM-Phycoerythrin (1:20 dilution, MHM-88), anti-human CD27-Allophycocyanin (1:20 dilution, O323), 7-AAD (1:100 dilution) (all from Biolegend) and 40 nM recombinant IsdB conjugated to Pacific-blue according to the manufacturer's protocol (Thermo Fisher Scientific). Cells were then pelleted, washed and resuspended in PBS/2% FBS for FACS sorting. CD3, CD16-, 7AAD-, IgM-, CD19+, CD27+, IsdB+ memory B cells were either bulk sorted for high-throughput sequencing or single-cell sorted into 96-well PCR plates for cloning. Individual cells were collected into each well of a 96-well plate containing 3.5 ml of Quick extraction buffer (Epicenter Bio) and immediately frozen over dry ice. Reverse transcription mixture containing 0.5 ml of reverse transcription primers mix (Supplementary Table 1), 5 ml of 2x reaction buffer and 1 ml of enzyme mix from SuperScript III One-Step RT-PCR system (Thermo Fisher Scientific) was added to the cell solution and reverse transcription was carried out at 55 °C for 30 min. Then, 20 ml of PCR mixture containing 0.15 mM each of the leader region primer mix, 0.25 mM each of constant reverse primer mix (Supplementary Table 1), 10 ml of the 2 x

reaction buffer and 0.5 ml of enzyme mix from the SuperScript III One-Step RT-PCR system were added directly to the tube, which contained 10 ml of the reverse transcription reaction product. PCR reaction conditions were 94°C for 2min, 40 cycles of 94°C for 15s, 55°C for 30s and 68°C for 1 min, a final extension step of 68 °C for 5 min. Separate nested PCR reactions were then performed to further amplify the transcripts of heavy chain and light chain (Vkappa þ Vlambda) using Taq polymerase system (Thermo Fisher Scientific) according to the manufacturer's protocol. Specifically, 0.2 mM each of variable region forward primer mix and 0.3 mM each of constant region reverse primer mix (Supplementary Table 1), and 2 ml of the first PCR reaction product as template were used in the nested PCR. Reaction conditions were 94 °C for 2 min, 40 cycles of 94°C for 15s, 56°C for 30s and 72°C for 1min, and a final extension step of 72 °C for 10 min. PCR amplicons were then gel-purified and sequenced. DNA sequences were analysed using an in-house developed software that identifies V-gene usage, J gene usage and CDRs identities. Antibodies (mAbs) sequences were then further triaged through a clustering algorithm based on VH and VL gene usages, CDR-H3 length and amino acid composition to identify unique clones and potential cluster of sibling clones for recombinant antibody expression.

High-throughput sequencing of memory B cells. High-throughput sequencing was performed as previously described54. In brief, total RNA was isolated from bulk sorted CD3-, CD16-, 7AAD-, IgM-, CD19+, CD27+, IsdB+ memory B cells (see above for labelling conditions) using RNeasy micro kit according to the manufacturing protocol (Qiagen). RNA quality was assessed using an Agilent Bioanalyzer. Total RNA was reverse-transcribed into cDNA using the SMARTer RACE kit according to the manufacturing protocol (Clontech) and cDNA was used as template for five IgG-VH and two IgA-VH PCR reactions. The PCR reaction was carried out using a modified 50 SMARTer RACE 10 x Universal Primer Mix (UPM) (Clontech) with the Lib-L-specific adaptor (Roche) sequence attached to the short oligo in the UPM (50CCTATCCCCTGTGTGCCTTGGCA GTCTCAGCTAATACGACTCACTATAGGGC-30) and a IgG or IgA isotype specific 30 primers (IgG, 50-GGGAAGACSGATGGGCCCTTGG -TGG-30; IgA, 50-CAGGCAKGCGAYGACCACGTTCCCATC-30) with the Lib-L-specific adaptor

(50-CCATCTCATCCCTGCGTGTCTCCGACTCAG-30) and a six-nucleotide barcode sequence attached at the 30 end. The reaction was carried out for 31 cycles following the manufacturer's recommendation. Barcoded IgG-VH (B640 base pairs (bp)) and IgA-VH (B630 bp) transcript libraries were purified using AMPure XP (Beckman Coulter), quantified using PicoGreen (Thermo Fisher Scientific) and pooled at equimolar amounts. The final multiplexed library pools were subjected to emulsion PCR and unidirectional sequencing using the GS FLX Titanium Lib-L chemistry (Roche).

Lineage analysis of the memory B cells. Lineage analysis and the tree topology were performed with the nucleotide sequences generated from 454 highthroughput sequencing of the donor's memory B cell repertoire and from the corresponding single-cell cloning. The VDJ segments and VH CDR3 of each clone were first identified by using VDJFasta31. Sequences were clustered into each clonal lineage using the VDJFasta single linkage method described previously54. For tree topology representation of the memory B cell repertoire, somatic hypermutations and isotype in each sequence were identified by using VDJFasta. Sequences were aligned with MUSCLE algorithm55. Identical sequences at the nucleotide level were collapsed to a single sequence. To avoid erroneous connections due to DNA amplification error, any less frequent clonal lineage member, which has connectivity of 1 and is only one nucleotide different from a more frequent neighbour, was grouped into the higher frequency neighbour. A lineage tree topology was then generated using the nucleotide alignment, somatic hypermutation level and isotype identity. The diagram was generated using Cytoscape.

Expression and purification of mAbs and protein reagents. Human IgGs and Fabs were expressed using the Expi293 system (Thermo Fisher Scientific). Human IgGs were purified with MabSelect (GE Healthcare). Fabs were histidine tagged and purified with Ni Sepharose Excel (GE Healthcare). Isd protein sequences used in this study are based on that of S. aureus strain USA300. The genes of Isd proteins were directly synthesized by GeneArt (Thermo Fisher Scientific) and cloned into either the pET-20 or the pET-47 expression vector (Novagen). Isd proteins and variants (truncations and point mutants) were expressed in BL21 DE3 cells (Thermo Fisher

Scientific) under the control of lac operator with an N-term histidine tag and C-term Flag Tag. These constructs were purified with Ni Sepharose Excel (GE Healthcare). The NEAT1 and NEAT2 domains utilized in crystallization were expressed with an N-terminal histidine tag followed by a HRV 3C site using the pET-47b vector (Novagen). The NEAT domains were purified with Ni Sepharose Excel, and then subjected to HRV 3C cleavage overnight at 4 °C for tag removal. Digested protein was then reapplied to Ni Sepharose Excel and the flow through was collected, removing the tag and undigested protein. NEAT domains were then complexed with their respective Fab (excess NEAT) and underwent size exclusion chromatography using a Superdex 200 column in AKTA (GE Healthcare). The sandwich Fab was later added at a 1:1 ratio to this complex. Human haemoglobin (Hb) was purified from fresh red blood cells obtained from Bioreclamation LLC. Briefly, haemoglobin from cell lysates was purified by anion exchange chromatography using Q-sepharose XL (GE healthcare), followed by size exclusion chromatography with Superdex 200 Prep Grade in AKTA (GE healthcare)56.

Epitope binning. Epitope binning for the anti-IsdB mAbs was carried out as previously described17. Briefly, anti-IsdB antibodies were individually amine-coupled as single spots onto a SensEYE G-COOH (Ssens bv) sensor chip to generate a 96-mAb array using a continuous flow microspotter (CFM) (Wasatch Microfluidics Inc). The printed sensor chip was then docked into an surface plasmon resonance (SPR) imager reader (MX96, IBIS Technologies bv) to perform interaction analysis of an analyte's binding towards the entire array of 96 antibodies simultaneously. Epitope binning experiments were performed using a classical sandwich assay format where each binding cycle comprised three steps; 35 nM rIsdB was injected for 3 min, 20 mg ml -1 antibody analyte was injected for a further 3 min, and then the surfaces were regenerated using a 30-s injection of 75 mM phosphoric acid. Ninety-six mAb analytes were injected over a 96-mAb array per unattended run. Binding data were analysed in SPRint software v. 6.15.2.1 and analysed in Wasatch Microfluidics' binning software for heat map generation, sorting and node plotting.

Enzyme-linked immunosorbent assay (ELISA). Recombinant Isd proteins and their variants (purified as described above) or commercially available alpha-toxin

(Calbiochem) and SEB (Toxin Technologies) at 2 ug ml -1 in PBS were coated directly onto 96-well Maxisorp, Nunc-immunoplates (ELISA plate, Thermo Scientific) by incubation overnight at 4 °C. Alternatively, the molecules were captured through their FLAG tag with a previously coated anti-FLAG antibody (F1904, Sigma-Aldrich) for 2 h at RT. Following blocking with PBS/1% BSA/0.01% Tween-20 (ELISA buffer) for one hour at RT, the plates were washed six times with PBS/0.05% Tween-20 on an automated plate washer. Diluted human serum samples or mAbs serial dilutions in ELISA buffer where the added and incubated with shaking for one hour at RT. Plates were washed as above and an HRP-conjugated goat anti-human IgG (H þ L) (Jackson Immunoresearch, 109-001-003) diluted 1:15,000 in ELISA buffer was added and incubation continued with shaking for one hour at RT. After a final wash as above, the plates were developed by addition of 3,30 ,5,50 -Tetramethylbenzidine (TMB) peroxidase substrate (KPL) and the colorimetric reaction was stopped by addition of 5% phosphoric acid (Aqua Solutions). Absorption was read at 450 nm on a Spectra MAX 340 plate reader (Molecular Devices).

Biosensor assay to determine the Hb-blocking effect of mAbs. The anti-IsdB mAbs were tested for their ability to block the rIsdB/Hb interaction as previously described17. Briefly, mAbs were captured at 15 mg ml -1 via anti-species sensors (10 min), 32 nM rIsdB was bound (10 min) followed by 1 mM Hb. Anti-species sensors were regenerated with 75 mM phosphoric acid. An isotype-matched negative control mAb (against an irrelevant target) was used to assess any non-specific cross-reaction of rIsdB or Hb.

Cell-based assay to determine the Hb-blocking effect of mAbs. The binding of human Hb to endogenously-expressed IsdB on S. aureus cells was used to assess the blocking effect of anti-IsdB mAbs as previously described57 with the following modifications. The S. aureus DSpA strain was used. The antibodies, at a concentration of 600 nM, were pre-incubated with S. aureus cells for 10 min at room temperature before purified human Hb was added to give a final Hb concentration of 150 nM. Final detection of Hb was performed by standard Western blotting with a biotinylated primary antibody (sheep polyclonal anti-Hb, biotinylated, Abcam ab95152, used at

2 mg ml -1) followed by a streptavidinconjugated secondary reagent (Streptavidin-IRDye 800 CW, 1 mg ml -1; Odyssey 926–32230, 1:4,000 dilution). A Licor Odyssey system was used to image the blot and to quantify the band intensities, data were expressed as per cent of maximum Hb binding in absence of antibodies.

Affinity measurements of anti-IsdB mAbs. Active concentrations of the recombinant IsdB antigens were determined using a calibration-free concentration analysis on a Biacore T200 biosensor equipped with CM5 sensor chip (GE Healthcare), as described previously58. This assay relied on the use of a high affinity anti-IsdB antibody that could be regenerated easily; thus D1–06 was chosen as the reaction surface and was amine-coupled onto flow cell 2 at a high density (8,700 RU) to promote mass transport limited binding. Flow cell 1 was left blank (activated and blocked, without any protein coupled) to serve as a reference surface. Surfaces were generated with 75 mM phosphoric acid. These experiments returned apparent activities of 31–34% for both the full-length recombinant IsdB and IsdB NEAT 2 domain, and these 'active' concentrations were used as input values for the kinetic experiments instead of the 'nominal' protein concentrations (as determined by light absorbance at A280 nm with appropriate extinction). Kinetic experiments were performed in a running buffer of PBS/0.01% Tween20 using a one-shot kinetic method59 on a ProteOn XPR36 equipped with GLC sensor chips (BioRad). The surfaces for these experiments were prepared at 25 °C. Briefly, ligand channels were minimally activated using a 2–3 min injection of a freshly mixed aqueous solution of 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide (EDC) and sulfo-N-hydroxysuccinimide (SNHS) at final concentrations of 1 mM EDC and 0.25 mM SNHS, antibodies were amine-coupled for 3 min at 15 mg ml -1 in 10 mM sodium acetate pH 4.5 buffer, and excess reactive esters were blocked with a 3 min injection of 1 M ethanolamine pH 8.5. Final amine-coupled levels ranged from 400–1,000 response unit (RU) per antibody, witho3% variation along the six spots within each ligand channel. The temperature was then adjusted to 37 °C to study the interaction of full-length recombinant IsdB or IsdB NEAT2 domain as analytes, which were injected for 3 min along the analyte channels as a five-membered serial dilution along with a buffer sample to provide an in-line buffer blank for data processing purposes. The dissociation phase was monitored for up to

4 h. Alternatively, the interaction analysis was performed in a 36-ligand array format using a kinetic titration injection methodology, as described previously60. The same surfaces were also used to study analytes in a short and long injection methodology. The top analyte concentration used for the kinetic experiments, regardless of the injection method used, was adjusted as appropriate for the antibodies being studied, and varied from active concentrations of B1 mM (for the weak affinity binders) to 20 nM (for the high affinity binders). Analyte injections were performed in duplicate and all experiments were repeated on different chips to generate up to three independent measurements per interaction. Binding data were analysed using ProteOn Manager software; the sensorgrams from the reaction spots were interspot-referenced and double-referenced and fit globally to a simple Langmuir model with mass transport to deduce the equilibrium dissociation constant (KD 1/4 kd/ka) for each rIsdB/ antibody binding interaction.

Binding of the germline-reverted antibodies were performed on a Biacore T200 SPR biosensor (GE Healthcare). Briefly, an anti-human Fc sensor chip was prepared by activating all flow cells of a Biacore CM4 sensor chip with a 1:1 (v/v) mixture of 400mM EDC and 100mM NHS for 7min, at a flow rate of 10 ml min 1. An anti-human Fc reagent (Southern Biotech 2014-01) was diluted to 50 mg ml -1 in 10 mM sodium acetate pH 4.5 and injected on all flow cells for 7 min at 20 ml min -1. All flow cells were blocked with 100 mM ethylenediamine in 150 mM Borate buffer pH 8.5 for 7 min at 10 ml min -1. The running buffer for this immobilization procedure was 10 mM HEPES, 150 mM NaCl, 0.05% (v/v) Tween-20, pH 7.4. Kinetics experiments were performed at 37 °C using a running buffer of 10 mM sodium phosphate, 150 mM NaCl, 0.01% (v/v) Tween-20, pH 7.4. Anti-IsdB mAbs were captured on downstream flow cells (flow cells 2, 3 and 4) at concentrations that ranged from 8 to 20mgml-1 at a flow rate of 10mlmin-1 for 2 min. Flow cell 1 was used as a blank reference surface. Following capture mAbs, analyte (buffer, or IsdB) was injected at 30 ml min -1 on all flow cells for 2 min. Multiple IsdB analyte concentrations were tested. The IsdB analyte concentrations were 1.6, 8.0, 40, 200 and 1,000 nM. After the analyte injection, dissociation was monitored for 5 min. Following analyte binding and dissociation all flow cells were regenerated with three 1-minute injections of 75 mM Phosphoric Acid.

The double-referenced sensorgrams were fit globally to a 1:1 Langmuir with mass transport binding model using the Biacore T200 evaluation software.

Crystal structure determination and analysis. IsdB NEAT1 or NEAT2 in complex with their respective Fabs were subjected to Index (Hampton Research), JCSG þ (Qiagen) and PEG/Ion (Hampton Research) crystallization screens using the Mosquito robot. A number of initial hits were refined using vapour diffusion followed by microbatch using Al's oil (Hampton Research). D2-06-N2 complex crystallized with 20% PEG 3350 and 0.1 M sodium citrate at pH 5.4. The D2-06-N2 crystal was flash frozen with liquid nitrogen using mother liquor with 20% glycerol. D4-30-N2 complex was crystallized with 15% PEG 10 K and 0.1 M Tris pH 8.0, and flash frozen with mother liquor with 20% ethylene glycol. D4-10-N1 complex crystallized with 17% PEG 3350 and 0.2 M ammonium citrate tri basic pH 7.0, and flash frozen with mother liquor with 20% glycerol. Data collection was performed at Advance Light Source beamline 5.0.2. (Lawrence Berkeley National Labs). Images collected were indexed and scaled with HKL2000 (ref. 61), and Phaser62 was used for molecular replacement. The models were further refined with a combination of Coot 0.7 (ref. 63), CCP4i 6.5.0 (ref. 64), Phenix 1.9 (ref. 65), and autoBUSTER66.

Generation of naı¨ ve IGHV4-39-derived phage antibody library. Blood samples of 25–50 ml were collected from healthy consented donors and PBMC were isolated as described earlier. A total of 1.0–1.9 million of CD19 + (1:20 dilution, HIB19), CD27 (1:20 dilution, O323) na¨ıve B cells were FACS isolated from the PBMC of each donor. Total RNAs from each donor were individually obtained using the RNeasy mini kit according to the manufacturing protocol (Qiagen). First strand cDNA was synthesized using a human IgM heavy chain constant reverse primer (50 -GAAGGCAGCAGCACCTGTGAG-30) and a human IgK light chain constant reverse primer (50 -TGGAGGGCGTTATCCACCTTCC-30) in a reverse transcriptase reaction (SuperscriptIII, Thermo Fisher Scientific). Light chain cDNA from all donors were pooled and variable kappa light chain genes were amplified as previously described using the VK family 1–4 primers and JK reverse primers67. Heavy chain cDNA from each donor was first individually amplified for 20 cycles using an IGHV4-39 specific leader region primer (50-TTCCTCCTGCTGGTG GCG-30) and

an IgM constant region reverse primer (50 -AAGTGATGGAGTCGG GAAGGAAG-30). PCR conditions were according to the manufacturing protocol of Pfu Ultra (Agilent). A uniquely barcoded nesting primer for each donor was generated by adding nine unique nucleotides combinations, which are the different combinations of codons encoding for amino acids glycine-glycine-serine, in front of each IGHV4-39 specific forward primer (50-CAGCTGCAGCTGCAGGA GTC-30). The IGHV4-39 VH gene from each donor was then individually amplified for 20 cycles using 2 ml of first PCR product as template, and the barcoded IGHV4-39 forward primer and JH reverse primers67. Pooled VK genes and VH genes were then sequentially ligated into a single-chain Fv antibody phage display vector. SS320 cells (Lucigen) were transformed with the assembled scFv library vector in thirty parallel 50 ml electroporation reactions according to the manufacturing protocol. The size of the starting library was ˜6 x 10ˆ9. Antibodydisplaying phages were recovered with M13KO7 helper phage (New England Biolab) according to previously published methods68. Briefly, overnight phage cultures were spun down at 12,000 g for 15 min. Supernatant was collected and incubated with 1:5 volumes of PEG 8000/2.5 M NaCl (Teknova) at room temperature for 20 min. The mixture was then spun down at 15,000 g for 10 min. Supernatant was removed and PBS was added to dissolve the phage pellet. Dissolved phage solution was spun down at 15,000 g for 10 min to remove any insoluble material. Phage supernatant was used immediately or frozen at -80 °C for storage.

Selection of NEAT1 binders from antibody phage library. For panning of phage library, 2–4 mg ml -1 of recombinant IsdB NEAT1 protein in PBS was first coated overnight at 4 °C onto 24 wells of a Maxisorp plate. Plates were then blocked with either Superblock or StartingBlock (Thermo Fisher Scientific). After washing off the blocking solution with PBS/0.05%Tween, 10ˆ13 phage particles (100 ml per well) in PBS/1% BSA/0.05% Tween were then incubated with plate-bound NEAT1 protein overnight at 4 °C in the first round. The amount of phage input were subsequently reduced to 5 x 10ˆ12 particles in round 2, 1012 particles in round 3 and 5 x 10ˆ11 particles in round 4. Phage incubations for round 2 to round 4 were performed at room temperature for 2–4 h. After phage incubation, plates were washed with PBS/0.05%Tween for 5–20 times. Bound phages were recovered by incubating the

well with 120 ml of 100 nM HCl for 20 min and immediately followed by neutralizing with 16 ml of 1 M TRIS pH 9.2. The eluted phages were then used to infect XL-1Blue (Agilent) cells growing at log phase (OD B0.3–0.6) for phage propagation and subsequent round of panning68. After four rounds of panning, infected E. coli were plated on LB carbenicillin plates. For screening, single E. coli colony were picked and individually inoculated in growth media (2YT/100 ug ml -1 carbenicillin/10^9 M13KO7) overnight at 37 °C to produce phage. Phage cultures were then spun down and one sixth dilution of the phage supernatant in PBS/0.5% BSA/0.05% Tween was used in ELISA to test the binding of the phage clone to NEAT1. Phage ELISA conditions were similar to the ELISA conditions described earlier, but anti M13-IgG-HRP conjugate (GE Healthcare, 27942101) was used as the detection reagent. Clones that were reactive to IsdB NEAT1 and not binding negative control proteins were then sequenced. Selected clones with unique HC sequences from different donors were reformatted as human IgG1 for further testing.

Sequence analysis of NEAT domains of S. aureus IsdB variants. Protein sequences from 4,152 S. aureus genomes annotated as IsdB were downloaded from the March 2015 release of the PATRIC pathogen database69. Additional filtering was performed to exclude partial sequences and incorrect protein or taxonomic classification: sequences with lengths of outside of the range 645±30 amino acids were excluded as well as sequences from the isolates S. aureus F87619 and S. aureus M21126. A multiple sequence alignment of the remaining 4,112 filtered protein sequences was generated using the MUSCLE algorithm55. The alignment was manually refined for the sequences of isolates from S. aureus subsp. aureus E1410, S. aureus RF122 and S. aureus O11 using Jalview70. The conservation score was computed as the frequency of the most commonly aligned residue at each position in the alignment. The multiple sequence alignment of the representative NEAT domain sequences was generated using the MUSCLE algorithm55 with the default alignment parameters.

Data availability. The accession number for the structures of D2-06-N2, D4-30-N2, and D4-10-N1 in the Protein Data Bank are 5D1Q, 5D1X and 5D1Z, respectively. The data that support the findings of this study are available within the article or from the corresponding authors on request.

3.3.5 Acknowledgements

This work was made possible and largely driven by my co-authors, including first authors Yik Andy Yeung, Davide Foletti, and Xiaodi Deng, as well as authors Yasmina Abdiche, Pavel Strop, Steven Pitts, Kevin Lindquist, Purnima D. Sundar, Marina Sirota, Adela Hasa-Moreno, Amber Pham, Jody Melton Witt, Irene Ni, Jaume Pons, David Shelton, Arvind Rajpal & Javier Chaparro-Riggers. We thank Mark Gilbert for help with flow cytometry, Lora Zhao and Dilduz Telman for help with high-throughput sequencing, Jennifer Zhang for help with antibody expression, Andrea Rossi for help with structural alignments, Wenwu Zhai for help with na¨ıve B cells isolation, and Daniel McDonough for help with high-throughput phage ELISA. We also thank the Rinat protein expression and purification group for providing reagents and cells for antibodies expression. We thank the Rinat biosensor group for binding analysis. We thank Hong Liang and John Lin for thorough review of the manuscript.

3.3.6 References

3.3.7 Copyright

3.4 Polymorphism in human adaptive receptor repertoire segments

Multiple efforts have been made to characterized the allelic polymorphism of the human V(D)J segment loci. One effort attempted to apply analysis to the 1000genomes project data to generate a polymorphim database. While in general we consider this a useful strategy, the resulting approach produced a degree of novel alleles that was not

consistent with the allele frequencies obtained when analyzing individual repertoires or long-range sequencing of individual haplotypes. We wrote the following response to the publication to as a means of generating a dialog about best practices in populating a new generation of allele variation databases for immunology.

It was with great interest that we read the recently published article by Yu et al. (1), which proposes a solution to the problem of building complete and accurate databases of germline Ig (IG) and TCR genes and alleles. This highlights one of the most formidable challenges in the immunogenetics field, as it has become apparent in recent years that existing germline databases (GLDB) are neither complete (i.e., lack existing alleles) (2–7) nor accurate in some cases (i.e., contain nonexisting alleles) (8). The impacts of this problem have most prominently come into focus in the context of IG/TCR expressed repertoire sequence datasets, the analysis and interpretation of which critically depend on the use of accurate GLDBs. Indeed, recent GLDB improvements via the inclusion of previously undetected IG alleles in repertoire sequence analysis have demonstrated the potential for direct consequences on human health research (9). In this comment, we enumerate some difficulties inherent in employing the data used by Yu et al. (1) to build a GLDB, and argue that a broad-based collaborative effort using a variety of data types is needed to achieve the goal of a complete yet reliable GLDB.

In their article (1), the authors develop a pipeline for identifying novel IG/TCR alleles from single nucleotide polymorphism (SNP) genotype data, and apply it to diverse population samples of the 1000 Genomes Project (G1K) (10–12) to build "Lym1K," a GLDB covering the human TCR β (TRB), TCR α (TRA), IGH, and IGK and IGL, summarized as "IGL" chain loci. Across the variable (V), diversity (D) and joining (J) genes in these loci, the authors report the discovery of 8750 germline alleles not currently curated in the international ImMunoGeneTics information system (13, 14). At face value, this finding is profound, and suggests potential for augmenting IG/TCR GLDBs using existing or newly generated genotype data. However, we are concerned about the accuracy of the underlying data, and the fact that erroneous genotypes/haplotypes used as input will result in incorrectly inferred IG/TCR alleles. We argue that users of such an approach should exercise due caution.

There are at least three potential caveats concerning the use of G1K data and similar short-read sequencing datasets for variant discovery, genotyping, and the downstream inference of novel IG/TCR alleles. These include:

1) the repetitive nature and structural complexity of IG and TCR loci;

2) the unknown extent of haplotype diversity and prevalence of large copy number variants (CNVs) involving genes in these regions; and

3) the use of source material derived from immortalized B cell lines.

The variant calls produced by G1K are only as reliable as the underlying short-read sequencing technologies used. Mapping of short-read data (for G1K Phase 3, reads can be as short as 70 bp) can be confounded in complex genomic loci (15–17), such as IG and TCR, which are characterized by a highly repetitive sequence architecture and extensive haplotype diversity (5, 18–23). Each of the IG and TCR loci consist of ∼40 or more phylogenetically related functional/open reading frame V, D (in IGH and TRB), and J genes, which exhibit high sequence homology that in some cases can reach 100% (e.g., for alleles at IGHV3-30 and related paralogs) (13, 14). Importantly, due to the fact that germline allele databases are incomplete, the degree of "allele sharing" between genes within IG/TCR loci is not fully understood. This would be expected to be a serious issue in the IGK locus, as nearly every V gene resides within two tandem duplication blocks, between which direct gene conversion events have been described (18). The repetitive nature of sequences in these loci creates the potential for mismapping of reads and ultimate assignment of variants to the incorrect genes.

A second critical consideration is that variant calls made using standard short-read data and bioinformatics pipelines are restricted to loci present in the genome reference assembly used for read mapping. This is important, as haplotype variability, in the form of large CNVs and SNPs is common in the IG and TCR loci (5, 18–23). Therefore, a single reference assembly poorly represents standing haplotype variation in any given population being screened. As noted by Yu et al. (1), there are in fact many genes missing from the current reference assemblies (e.g., GRCh37 and GRCh38), and thus by definition, it is impossible to make reliable genotype and allele calls for these genes. In total, using IGH as an example, there are at least 16

known functional/open reading frame V genes and >220 kbp of genomic sequence present in haplotypes in the human population that are not represented in GRCh37 (the assembly used by G1K for Phase 3 read mapping and variant calling) (5). In some populations, these alternate haplotypes represent the major allele, indicating that the majority of samples screened would carry a sequence absent from GRCh37 (5). The technical effects of this "missing sequence" are not known, but it would be expected that reads representing alternate haplotypes and genes in any given individual would have the potential to be incorrectly mapped to off-target genes that are present in the reference. Directly related to this, the presence of CNVs in a sample can cause other problems for short-read mapping and downstream genotype inference. For example, heterozygous gene deletions (hemizygotes) can masquerade as homozygotes for a given SNP or coding allele, whereas paralogous sequence variants between close gene duplicates can result in artifactual heterozygote calls (24, 25).

Furthermore, it is important to take the source of genomic DNA used by a study into account. In the case of G1K, DNA was extracted from lymphoblastoid cell lines, i.e., B cells immortalized by EBV. Therefore, a fraction of the IG loci in these lines has undergone V(D)J recombination, which can lead to a reduction or complete loss of reads (lower read depth) overlapping proximal V, D, and distal J genes within a given sequencing library (5, 26). Low read coverage can directly impact the reliability of variant discovery and genotype calling (11, 12). Additionally, hypermutated memory B cells can be the target of EBV transformation (27), which will result in the presence of non-germline IG gene mutations in DNA isolated from lymphoblastoid cell lines, resulting in potential false-positive allele calls; evidence of somatic hypermutation at IG genes that have undergone V(D)J recombination has been directly observed in G1K samples (5). Although not directly applicable to G1K data, it should also be noted that similar issues concerning the reliability of genotyping due to somatic rearrangements have been reported for TCR loci using DNA isolated from blood (28). Requiring variants to be present in multiple individuals or conducting analyses in family-based datasets could potentially help mitigate this issue, but the reliability of such an approach would need to be demonstrated, as somatic mutations at hot-spots likely recur.

Unfortunately, because population and genomic resources in the IG and TCR gene regions remain limited, the true impacts of the potential caveats laid out above remain difficult to assess. However, as part of the Phase 3 data release, G1K has used quality control metrics from low-coverage data across >2,600 human samples to directly assess the "accessibility" of every base in the genome to sequencing technologies used currently by the consortium [see Refs. (12) and (29)]. Using this approach, certain bases have been masked as having potentially higher false-positive and -negative variant call rates. Using IGH as an example, >25% of bases within the coding exons of 62/83 IGHV, D, and J genes in GRCh38 fall within this category, even when using the least stringent ("pilot") criteria established by G1K. Although this does not by definition mean calls made in these regions are incorrect, we would argue it implies that their reliability is difficult to assess at this time. Indeed, G1K found that variant calls at these masked bases also had higher failure rates using alternative variant discovery/genotyping methods (12).

Taken together, the caveats discussed above suggest that databases constructed from alleles inferred from short-read genomic data should be carefully vetted, bearing in mind that even a single incorrect genotype within an IG/TCR gene can impact the reliability of haplotype phasing and allele inference for that gene. Therefore, we urge users to critically examine and consider both the features of the data underlying a given allele call, such as read lengths, coverage depth, library construction methods, cohort sample size, and the source of DNA, as well as the bioinformatics methods and the genic and sequence content of the genome reference assembly used for read mapping and variant calling. It is likely that all of these will impact the reliability of the allele database constructed, and importantly, may be more or less critical depending on the locus or gene/allele in question.

Finally, in addition to understanding factors related to the underlying data used, systems for thorough validation and benchmarking should be implemented to ensure low error rates. Such efforts have proven critical for the development of allele calling and genotyping methods using short-read data in other immune loci of comparable complexity (e.g., KIR and HLA) (30, 31). A basic cross-referencing of variant calls to other databases may be a useful strategy in certain circumstances, but would be

expected to be problematic if variants in that database are not mutually exclusive from the variant call set used for allele inference. For example, dbSNP (32), used by Yu et al. (1) for filtering of calls from their pipeline, contains SNPs directly submitted by G1K, and thus an overlap of G1K IG/TCR variants and dbSNP would be expected, not offering an unbiased form of validation. Furthermore, if a database cross-referencing approach is used, the secondary database must be reliable, and may itself require careful filtering. For example, there are 62 G1K SNPs across 24/44 IGHV genes (GRCh37) that are cataloged by dbSNP, but are flagged as "suspect" variants potentially representing false-positives.

We hope that this debate can motivate a concerted effort on the part of our community to find sustainable strategies to improve and complement the current IG/TCR GLDBs. Over the coming years, in addition to population genome sequencing efforts by short-read platforms, data from long-read technologies and inferred alleles from expressed repertoire sequencing efforts will become generally available. It is clear that a multitude of approaches can and will be taken to create reference GLDBs from these data, but we should recognize that the quality of a GLDB cannot be measured by its allele count. Instead, we consider it to be the most productive path to set our current focus on the creation of GLDBs containing high-confidence, independently confirmed genes/alleles, even if stringent confirmation requirements result in the exclusion of rare alleles. In addition, the community should strive to develop statistics that describe the uncertainty associated with an individual allele to provide a transparent measure for users. Ultimately, however, it is worth considering that studies requiring the precise germline sequence of a specific donor may necessitate direct sequencing of the individual, instead of relying on a reference database. Ideally, in line with the principles of the Reproducible Research Standard (33), both databases and their underlying datasets should be available under a free and open licensing scheme to facilitate further development. Finally, it is important to note that the issues discussed here are not limited to human GLDBs, and will apply to other species, including murine and nonhuman primate models (7, 34). We are convinced that a community effort toward achieving these goals has the potential to greatly enhance the analysis of repertoire sequencing studies across the field and provide more detailed and reliable

insights into adaptive immune responses in the context of infection, autoimmunity, and malignancies.

3.4.1 Copywrite

3.4.2 References

Yu Y., R. Ceredig, C. Seoighe. 2017. A database of human immune receptor alleles recovered from population sequencing data. J. Immunol. 198: 2202–2210.

Wang Y., K. J. Jackson, B. Gäeta, W. Pomat, P. Siba, W. A. Sewell, A. M. Collins. 2011. Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants. Immunogenetics 63: 259–265.

Boyd S. D., B. A. Gaëta, K. J. Jackson, A. Z. Fire, E. L. Marshall, J. D. Merker, J. M. Maniar, L. N. Zhang, B. Sahaf, C. D. Jones, et al. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. J. Immunol. 2010. 184: 6986–6992.

C., R. K. Shrestha, B. E. Lambson, K. J. Jackson, I. A. Wright, D. Naicker, M. Goosen, L. Berrie, A. Ismail, N. Garrett, et al. 2015. Ability to develop broadly neutralizing HIV-1 antibodies is not restricted by the germline Ig gene repertoire. J. Immunol. 194: 4371–4378.

Watson C. T., K. M. Steinberg, J. Huddleston, R. L. Warren, M. Malig, J. Schein, A. J. Willsey, J. B. Joy, J. K. Scott, T. A. Graves, et al. 2013. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. Am. J. Hum. Genet. 92: 530–546.

Gadala-Maria D., G. Yaari, M. Uduman, S. H. Kleinstein. 2015. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. Proc. Natl. Acad. Sci. USA 112: E862–E870.

Corcoran M. M., G. E. Phad, V. B. Néstor, C. Stahl-Hennig, N. Sumida, M. A. A. Persson, M. Martin, G. B. Karlsson Hedestam. 2016. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. Nat. Commun. 7: 13642.

Wang Y., K. J. Jackson, W. A. Sewell, A. M. Collins. 2008. Many human immunoglobulin heavy-chain IGHV gene polymorphisms have been reported in error. Immunol. Cell Biol. 86: 111–115.

Xochelli A., A. Agathangelidis, I. Kavakiotis, E. Minga, L. A. Sutton, P. Baliakas, I. Chouvarda, V. Giudicelli, I. Vlahavas, N. Maglaveras, et al. 2015. Immunoglobulin heavy variable (IGHV) genes and alleles: new entities, new names and implications for research and prognostication in chronic lymphocytic leukaemia. Immunogenetics 67: 61–66.

Abecasis G. R., D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, G. A. McVean1000 Genomes Project ConsortiumAbecasis G. R., D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, G. A. McVean. 2010. A map of human genome variation from population-scale sequencing. [Published erratum appears in 2011 Nature 473: 544.] Nature 467: 1061–1073.

Abecasis G. R., A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, G. A. McVean1000 Genomes Project ConsortiumAbecasis G. R., A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, G. A. McVean. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65.

Auton A., L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis1000 Genomes Project ConsortiumAuton A., L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis. 2015. A global reference for human genetic variation. Nature 526: 68–74.

Lefranc M.-P. L. G. 2001. The Immunoglobulin factbook, Vol. 262. Academic Press, London.

Giudicelli V., D. Chaume, M.-P. Lefranc. 2005. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res. 33: D256–D261.

Musumeci L., J. W. Arthur, F. S. G. Cheung, A. Hoque, S. Lippman, J. K. V. Reichardt. 2010. Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. Hum. Mutat. 31: 67–73.

Li H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics 30: 2843–2851. Abstract/F

Reumers J., P. De Rijk, H. Zhao, A. Liekens, D. Smeets, J. Cleary, P. Van Loo, M. Van Den Bossche, K. Catthoor, B. Sabbe, et al. 2011. Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. Nat. Biotechnol. 30: 61–68.

Watson C. T., K. M. Steinberg, T. A. Graves, R. L. Warren, M. Malig, J. Schein, R. K. Wilson, R. A. Holt, E. E. Eichler, F. Breden. 2015. Sequencing of the human IG light chain loci from a hydatidiform mole BAC library reveals locus-specific signatures of genetic diversity. Genes Immun. 16: 24–34.

Mackelprang R., C. S. Carlson, L. Subrahmanyan, R. J. Livingston, M. A. Eberle, D. A. Nickerson. 2002. Sequence variation in the human T-cell receptor loci. Immunol. Rev. 190: 26–39.

Mackelprang R., R. J. Livingston, M. A. Eberle, C. S. Carlson, Q. Yi, J. M. Akey, D. A. Nickerson. 2006. Sequence diversity, natural selection and linkage disequilibrium in the human T cell receptor alpha/delta locus. Hum. Genet. 119: 255–266.

Li H., X. Cui, S. Pramanik, N.-O. Chimge. 2002. Genetic diversity of the human immunoglobulin heavy chain VH region. Immunol. Rev. 190: 53–68.

Chimge N.-O., S. Pramanik, G. Hu, Y. Lin, R. Gao, L. Shen, H. Li. 2005. Determination of gene organization in the human IGHV region on single chromosomes. Genes Immun. 6: 186–193.

Kidd M. J., Z. Chen, Y. Wang, K. J. Jackson, L. Zhang, S. D. Boyd, A. Z. Fire, M. M. Tanaka, B. A. Gaëta, A. M. Collins. 2012. The inference of phased

haplotypes for the immunoglobulin H chain V region gene loci by analysis of VDJ gene rearrangements. J. Immunol. 188: 1333–1340.

Estivill X., J. Cheung, M. A. Pujana, K. Nakabayashi, S. W. Scherer, L.-C. Tsui. 2002. Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome. Hum. Mol. Genet. 11: 1987–1995.

Ho M. R., K. W. Tsai, C. H. Chen, W. C. Lin. 2011. dbDNV: a resource of duplicated gene nucleotide variants in human genome. Nucleic Acids Res. 39:D920–D925.

Luo S., J. A. Yu, Y. S. Song. 2016. Estimating Copy Number and Allelic Variation at the Immunoglobulin Heavy Chain Locus Using Short Reads. PLOS Comput. Biol. 12: e1005117.

Kozbor D., J. C. Roder. 1981. Requirements for the establishment of high-titered human monoclonal antibodies against tetanus toxoid using the Epstein-Barr virus technique. J. Immunol. 127: 1275–1280.

Schwienbacher C., A. De Grandi, C. Fuchsberger, M. F. Facheris, M. Svaldi, M. Wjst, P. P. Pramstaller, A. A. Hicks. 2010. Copy number variation and association over T-cell receptor genes–influence of DNA source. Immunogenetics 62: 561–567.

1000 Genomes Project. webpage: http://www.internationalgenome.org/faq/why-only-85-genome-assayable/.

Dilthey A., C. Cox, Z. Iqbal, M. R. Nelson, G. McVean. 2015. Improved genome inference in the MHC using a population reference graph. Nat. Genet. 47: 682–688.

Norman P. J., J. A. Hollenbach, N. Nemat-Gorgani, W. M. Marin, S. J. Norberg, E. Ashouri, J. Jayaraman, E. E. Wroblewski, J. Trowsdale, R. Rajalingam, et al. 2016. Defining KIR and HLA Class I Genotypes at Highest Resolution via High-Throughput Sequencing. Am. J. Hum. Genet. 99: 375–391.

Sherry S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, K. Sirotkin. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29: 308–311.

Stodden V. 2009. The legal framework for reproducible scientific research: licensing and copyright. Comput. Sci. Eng. 11: 35–40.

Collins A. M., Y. Wang, K. M. Roskin, C. P. Marquis, K. J. Jackson. 2015. The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains. Philos. Trans. R. Soc. Lond. B Biol. Sci. 370: 20140236.

3.5 Copyright

Chapter 4

Discoveries in natural repertoires

4.1 Introduction

Repertoire sequencing and analysis technologies have availed the adaptive immune system to direct inspection. In this chapter, we present the results of a series of studies made possible by reading the repertoire.

4.2 B cell exchange across the blood-brain barrier in multiple sclerosis

In multiple sclerosis (MS) pathogenic B cells likely act on both sides of the blood-brain barrier (BBB). However, it is unclear whether antigen-experienced B cells are shared between the CNS and the peripheral blood (PB) compartments. We applied deep repertoire sequencing of IgG heavy chain variable region genes (IgG-VH) in paired cerebrospinal fluid and PB samples from patients with MS and other neurological diseases to identify related B cells that are common to both compartments. For the first time to our knowledge, we found that a restricted pool of clonally related B cells participated in robust bidirectional exchange across the BBB. Some clusters of related IgG-VH appeared to have undergone active diversification primarily in the CNS, while others have undergone active diversification in the periphery or in both

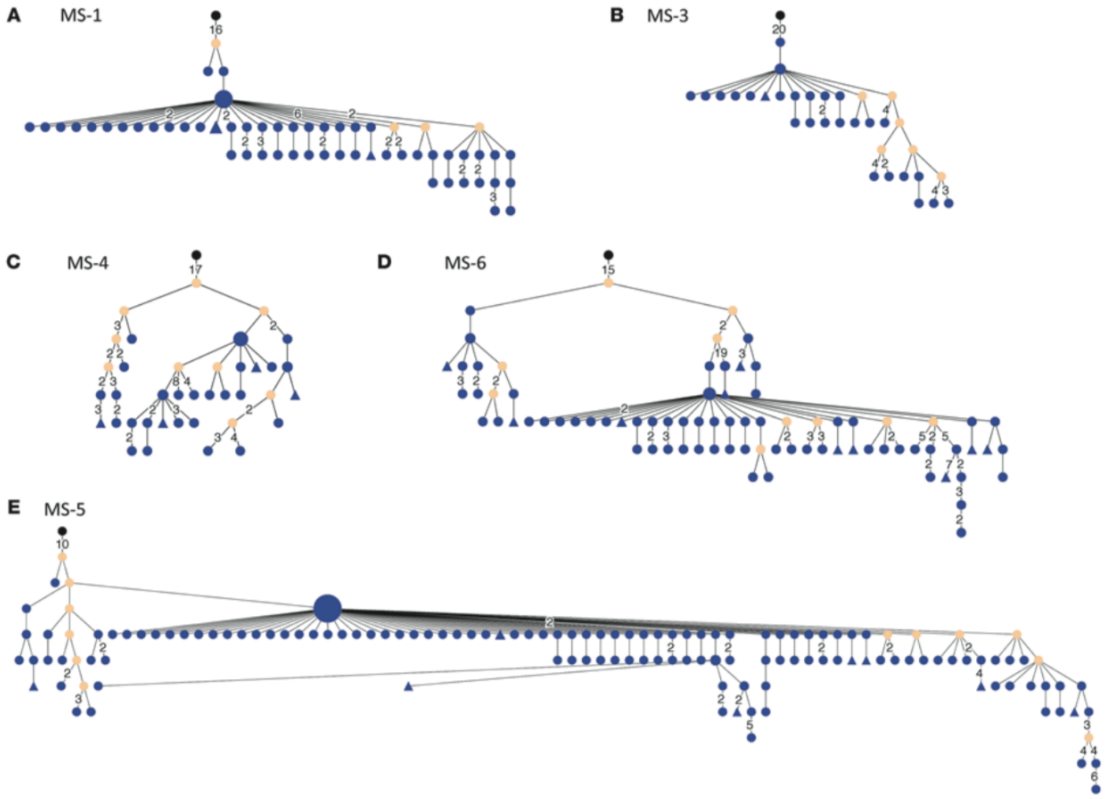


Figure 4.1: In MS, B cell receptors are subject to extensive intrathecal SHM. Nucleotide sequences, represented by clusters shown in Figure 1, were selected from the IgG-VH sequence database and used to generate lineage trees using IgTree software (see Methods). (A–E) Representative trees for CSF-restricted clusters for patients (A) MS-1, (B) MS-3, (C) MS-4, (D) MS-6, and (E) MS-5. The corresponding IGHV, IGHJ, and most com- mon H-CDR3 AA sequences are listed in Supplemental Table 2. In the lineage trees, each round node represents at least one unique IgG-VH sequence ranging from at least the 5´ end of H-CDR1 to the 3´ end of H-CDR3; larger nodes represent up to hundreds of identical sequences. Putative germline sequences were determined using SoDA (https://dulci.org/soda/; ref. 36) and are labeled as black, and hypothetical intermedi- ates calculated by IgTree are labeled as beige. The numbers represent mutational steps between nodes; only mutational steps >1 are indicated; thus, unlabeled branches represent a single mutation. Triangular nodes contain 2 or more singleton sequences in leaves.

compartments in parallel. B cells are strong candidates for autoimmune effector cells in MS, and these findings suggest that CNS-directed autoimmunity may be triggered and supported on both sides of the BBB. These data also provide a powerful approach to identify and monitor B cells in the PB that correspond to clonally amplified populations in the CNS in MS and other inflammatory states.

4.2.1 Copyright

4.3 Seroconversion signatures and convergent antbodies in influenza

B cells produce a diverse antibody repertoire by un- dergoing gene rearrangements. Pathogen exposure induces the clonal expansion of B cells expressing antibodies that can bind the infectious agent. To assess human B cell responses to trivalent seasonal influenza and monovalent pandemic H1N1 vaccina- tion, we sequenced gene rearrangements encoding the immunoglobulin heavy chain, a major determi- nant of epitope recognition. The magnitude of B cell clonal expansions correlates with an individual's secreted antibody response to the vaccine, and the expanded clones are enriched with those expressing influenza-specific monoclonal antibodies. Addition- ally, B cell responses to pandemic influenza H1N1 vaccination and infection in different people show a prominent family of convergent antibody heavy chain gene rearrangements specific to influenza antigens. These results indicate that microbes can induce specific signatures of immunoglobulin gene rearrange- ments and that pathogen exposure can potentially be assessed from B cell repertoires

4.3.1 Introduction

Human B cells generate a vast diversity of antibodies by rearranging the genes encoding immunoglobulins V (variable), D (diversity), and J (joining) in their genomes (Tonegawa, 1983). For decades, most monitoring of human antibody responses to infections or vaccines has been performed by serological measurements that can evaluate antibody specificities but has provided only limited insight into the underlying changes in clonal populations of B cells, or the gene rearrangements responsible for the antibodies. More recently, single-cell sorting and cloning of antibody genes, as well as optimized culture systems and hybridoma generation, have given greater insight into the specificity and breadth of reactivity of the antibodies produced by influenza-specific B cells and molecular under- standing of the genes encoding such antibodies (Li et al., 2012; Wrammert et al., 2011; Wrammert et al., 2008; Yu et al., 2008). High-throughput DNA sequencing methods now permit detailed monitoring of B cell repertoires in humans and are starting to be applied extensively to the study of vac- cine responses (Boyd et al., 2009; DeKosky et al., 2013; Jiang et al., 2013; Krause et al., 2011; Liao et al., 2011; Wu et al., 2011).

It is largely unknown whether different people use similar anti- body genes in their responses to common pathogen-associated antigens. With a few exceptions, such as the antibody responses to repetitive polysaccharide antigens (Ademokun et al., 2011; Park et al., 1996; Scott et al., 1989; Silverman and Lucas, 1991), there has been little evidence of similarity between different humans' responses to most pathogens. Indeed, anti- bodies would themselves be expected to exert a selection pres- sure upon the pathogens they target and thus cause pathogens to avoid expressing antigens that are recognized by human anti- body genes.

We conducted a detailed study of B cell clonal expansions in response to influenza vaccination and used deep sequencing to identify within a week of vaccination clonal expansion signa- tures that correlate with the magnitude of the serological response in vaccinated individuals. Comparison of expanded clones to influenza-specific plasmablasts identified by single- cell sorting from the same subjects demonstrated substantial overlap between these populations. More surprisingly, we iden- tified shared convergent antibody responses to the H1N1 2009 influenza strain among different

people in response to both vaccination and infection. These results represent an example of a signature in immunoglobulin gene rearrangements specific to the pathogen that elicited them and suggest that features of an individual's history of pathogen exposure can be identified by sequence analysis.

4.3.2 Results

Deep Sequencing of Rearranged IGH from the Trivalent Inactivated Seasonal Influenza Vaccine Response

To take an overview of B cell responses induced by vaccination, we carried out deep sequencing of immunoglobulin heavy chain (IGH) from peripheral-blood B cells of 14 healthy young individ- uals vaccinated with the 2007 or 2008 trivalent inactivated seasonal influenza vaccine (TIV) (Moody et al., 2011). Seven indi- viduals were "seroconverters," who raised at least a 4-fold in- crease in titer above baseline to two or more vaccine antigens, as measured by ELISA against purified hemagglutinins (HAs). The other seven individuals were "nonseroconverters," who failed to increase their vaccine-specific antibody to meet these criteria (Table S1 , available online) (Moody et al., 2011).

Twelve replicate IGH libraries were prepared from inde- pendent genomic DNA template aliquots from cryopreserved peripheral-blood mononuclear cells for each individual at each of three time points: prevaccination, day 7 postvaccination, and day 21 postvaccination (Figure 1 A). On average, 35,436 IGH sequences were analyzed for each individual. Sequencing depth was relatively evenly distributed across the time points; there were an average of 11,661 IGH sequences prevaccination, 12,200 at day 7, and 11,564 at day 21.

B Cell Clonal Signatures from Deep Sequencing Correspond to Serological Measures of Vaccine Response

Clonally related B cell lineages were identified by the presence of identical, or near identical, IGH in independent replicate sequence libraries from genomic DNA for each time point. This approach ensures that high expression of antibody gene mRNA in individual cells, or amplification bias, is not misinter- preted as evidence

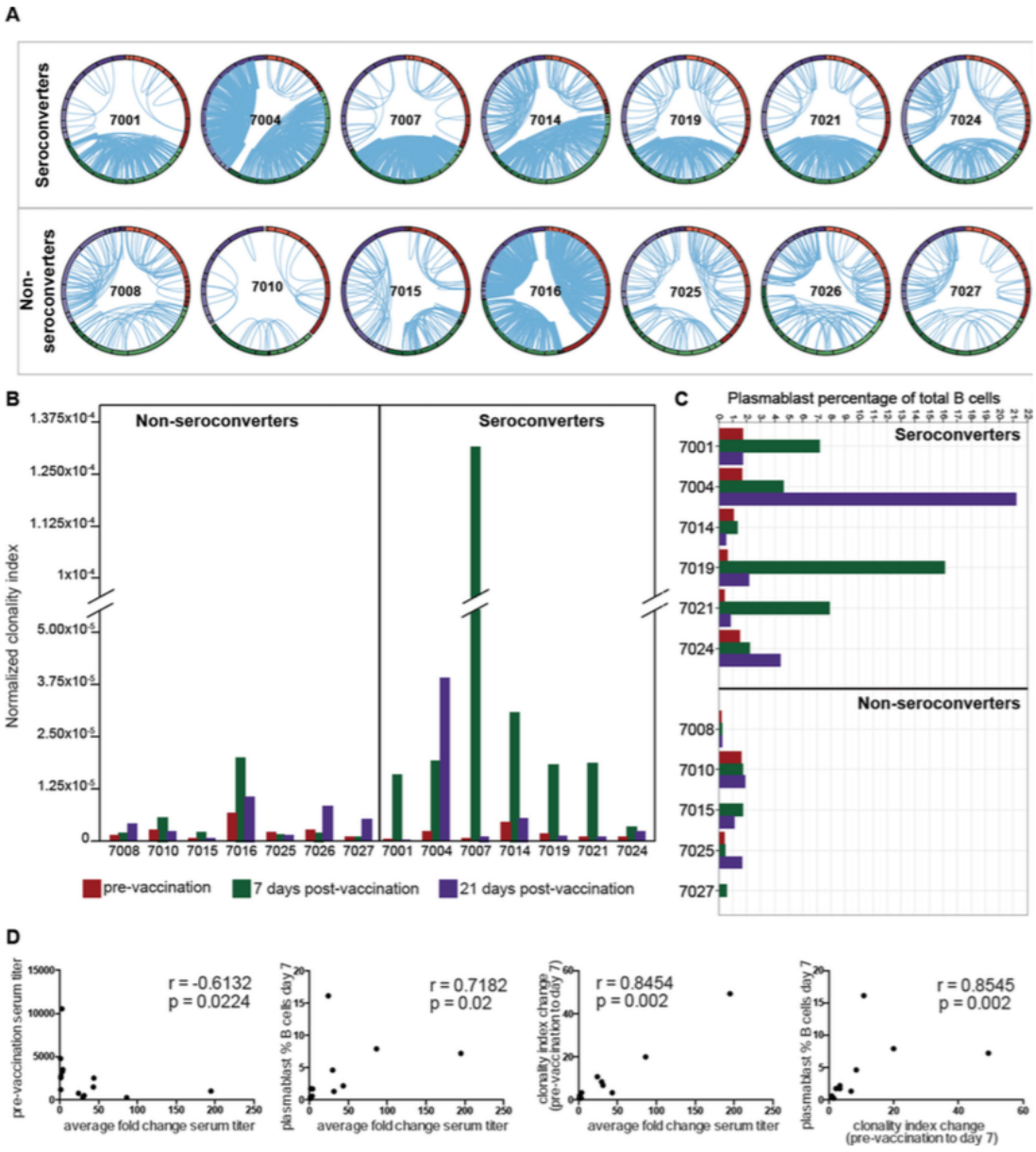


Figure 4.2: Quantitation of clonal B cells in the blood following vaccination predicts seroconversion (A) Replicate IGH libraries were generated from peripheral blood B cells for 14 individuals (Table S1) at three time points: pre-vaccination (red arc), day 7 (green arc) and day 21 post- vaccination (purple arc). Replicates are shown as bands within each arc and lines connect clonally related VDJs from independent replicates. Detailed IGH repertoires for each individual are shown in Figure S1A. Figure S1B presents criteria for definition of clonal lineages. (B) Normalized clonality index scores; pre-vaccination (red), 7 days (green) and 21 days post-vaccination (purple). (C) Plasmablast percentages of total B cells (Table S3); pre-vaccination (red) and at days 7 (green) and 21 (purple) post-vaccination. (D) Correlation between metrics and serological antibody response.

of clonal B cell populations. Most sero- converters showed a response with one to three larger clones and variable numbers of smaller clones, although subject 7024, who showed a predominance of smaller lineages, was an exception. The median number of expanded clones for sero- converters at day 7 was 69 (range 39–92), whereas the median for nonseroconverters was 25 (range 8–85).

To compare the clonal signatures between samples, we used a previously described clonality index (Wang et al., 2014). The clonality index is a scale-independent normalized measure that reflects the probability that two independent sequences derive from clonally related B cells (Figure 1 B). Each of the seroconvert- ers showed a prominent increase in clonality on day 7 in compar- ison to prevaccination (Movie S1). Measured changes in clonality from prevaccination to day 7 ranged from 3.38-fold to 247.57- fold among seroconverters. Nonseroconverters showed a mixture of modestly increased (four individuals) and modestly decreased (three individuals) indices (Figure 1 B).

The change in B cell clonality by day 7 was positively cor- related with the fold change in antibody titer against vaccine HA antigens, as measured by ELISA 21 days postvaccination (Spearman r = 0.8454, p value = 0.002; Figure 1 D). Nonseroconverter 7016 showed a partial response with a 1.58-fold titer increase for combined TIV antigens and a 2.98-fold increased clonality index, but there was a strong response to one vaccine component (5.39-fold change to A/Brisbane/10/2007/H3).

Plasmablast counts in blood after vaccination have been re- ported to correlate with serological responses (Liao et al., 2011; Sasaki et al., 2008). Day 7 plasmablast frequencies (Table S3) in the samples studied here showed significant correlation with the clonality index (r = 0.8545, p = 0.002, Figure 1 D) and a slightly weaker correlation with changes in serum antibody titer (r = 0.7182, p = 0.02, Figure 1 D). However, the difference be- tween these two correlations did not meet statistical significance (p = 0.1214, Steiger's dependent variable correlation) (Steiger, 1980). Consistent with prior literature, prevaccination titers were negatively correlated with vaccine-stimulated titer changes (Figure 1 D) (Sasaki et al., 2008).

Clones from IGH Deep Sequencing Comprise a Subset of Antigen-Specific Plasmablasts

To directly evaluate the influenza specificity of the detected B cell expansions and assess whether these represent members of the plasmablast pool, we compared the data to IGH from flow-sorted plasmablasts isolated from the same day 7 samples for five of the seroconverters (7001, 7004, 7014, 7021, and 7024). A total of 398 sorted plasmablasts were expressed as recombi- nant monoclonal antibodies (mAbs); their antigen specificities were evaluated as part of a previous study (Moody et al., 2011), and 59.8% were influenza reactive.

Analysis revealed that 24.4% of sorted plasmablasts be- longed to clonally expanded lineages containing IGH identified by sequencing (Figure 2 A; Table S2). Sixty-six percent of the plasmablast-derived IGHs in these lineages were from influenza HA-binding mAbs (Figure 2 B; Table S2). Conversely, 10.2% of the day 7 expanded clones were in lineages that included plasmablast IGHs. Subject 7001, who contributed 247 mAbs (Moody et al., 2011), shared 19.6% of day 7 clonally expanded B cell lineages with plasmablast IGHs, and 82.4% of these included influenza-specific mAbs (Figure 2 B; Table S2). This indicates that at day 7 postvaccination, there is partial overlap of both the total population of clonally expanded B cells and plasmablasts sorted on the basis of their immunophenotype.

Somatic Hypermutation of Vaccine-Stimulated B Cell Repertoires

Seroconverters and nonseroconverters differed in the somatic mutation of expanded B cell clones at day 7 postvaccination (mean mutation was 6.4% and 4.3%, respectively, $p = 0.0041$, t test). Prior to vaccination, the groups shared similar levels of mutation in expanded clones (3.2% and 3.4%, $p = 1.00$) and non- clonal IGHV sequences (1.5% and 1.3%, $p = 0.8048$). At day 21, the expanded lineages of the seroconverters retained a higher somatic point mutation (5.7%), whereas nonseroconverters returned to the prevaccination mutation frequency of 3.5% ($p = 0.1649$).

Time Course of Response to H1N1 Single Antigen Vaccination

To obtain a more detailed view of the dynamics of the B cell response to influenza vaccine in a somewhat simpler vaccination context, we undertook IGH sequencing from a daily time course of the response to the single-agent pandemic H1N1 2009 influ- enza vaccination in an additional healthy subject (BFI-278) (Fig- ure 3 A; Movie S2). Prior to vaccination, few clonally expanded B cell lineages were detected

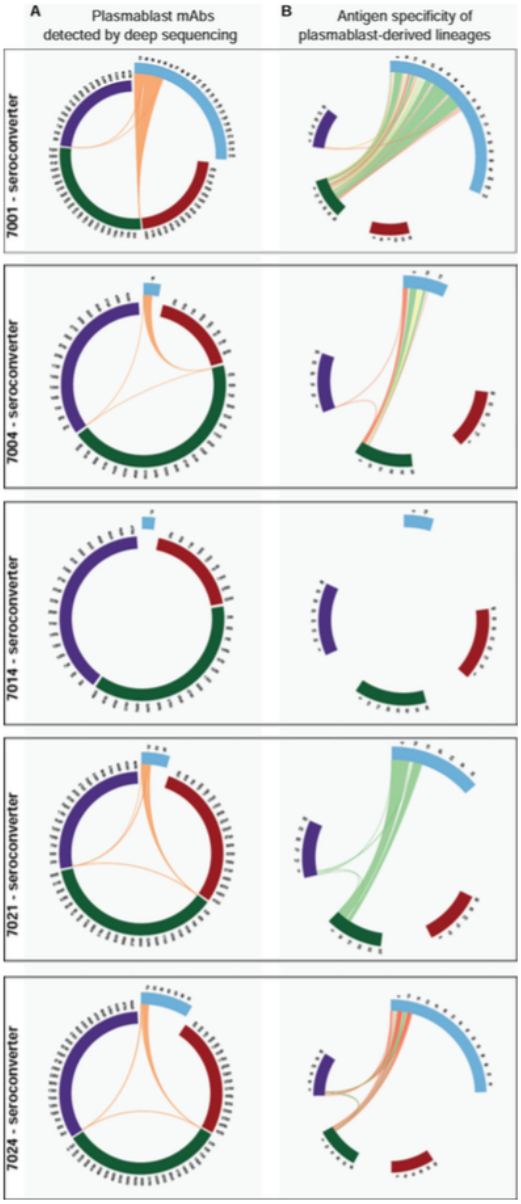


Figure 4.3: Antigen specificity of vaccine-induced B cell clonal populations (A) B cell lineages with members from mAbs derived from day 7 sorted plasmablasts (blue arc) and deep sequenced IGH from peripheral B cells prior to (red) and at days 7 (green) and 21 (purple) post TIV vaccination. Lines join lineage members. (B) Antigen binding of plasmablast-derived mAbs (Table S2); influenza antigen (green), unknown/untested (yellow), non-influenza antigen (red). Arcs are 'zoomed' to shared lineages. 7014 had no shared lineages.

(five expanded clones, on average 0.06% of the total). After vaccination, a clonal response peaked at day 7 and decreased toward prevaccination levels by day 10 (Figure 3 A). Overall, 256 clonal lineages were detected postvac- cination; 11 lineages were strongly induced by vaccination, and each contributed more than 0.1% of the total rearrangements. Vaccine-induced clonal lineages reached their peak at day 7, when they accounted for 7.1% of total IGH. Many of the vac- cine-induced expansions continued to be detected by day 9, although some were observed only at a single time point (Fig- ure 3 A). Clonal lineages utilizing IGHV3-7 paired with IGHJ6 were overrepresented in the vaccine-induced proliferation, and 6 of the 11 prominent lineages used these genes with an 18 amino acid CDR3 (Figure 3 B).

The postvaccination clonally expanded IGHs were somatically mutated, and members of the 11 prominent lineages showed an average of 5.68% IGHV mutation (Figure 3 B). The full IGH repertoire detected in this individual had a mean IGHV mutation level of 1.18%.

Identification of Convergent IGH Rearrangements in the H1N1 Vaccine Response

The dominant IGH lineages in BFI-278's response to single- antigen H1N1 vaccination (Figure 3) had stereotyped features: IGHV3-7, IGHJ6, and an 18 amino acid CDR3. These lineages were compared to previously reported H1N1-responding B cell repertoires. Surprisingly, we observed response convergence with constrained CDR3 sequences seen in two other studies (Wrammert et al., 2011; Krause et al., 2011). The CDR3 se- quences of 4K8 and 2K11 (Krause et al., 2011) differed from member BFI-278's dominant expansion by a single amino acid residue (Figures 4 A and 4B; Table S4). 48K and 2K11 were iso- lated from one individual after pandemic H1N1 vaccination, and both had activity against 2009 H1N1 and human and swine influenza strains from 1918, 1930, 1976, and 1977 by hemagglu- tination activation inhibition (HAI) (Krause et al., 2011). Paired with the light chain of antibody 4K8, a mAb generated from the IGH isolated from BFI-278 showed HAI with a titer of 20 in tripli- cate assays against the influenza A H1N1 California/7/09 strain, but no detectable activity against the other strains in the panel.

Stereotypic IGHs were also present in H1N1-stimulated clones from a single donor 15 days after acute pandemic H1N1 infection (Wrammert et al., 2011). One mAb,

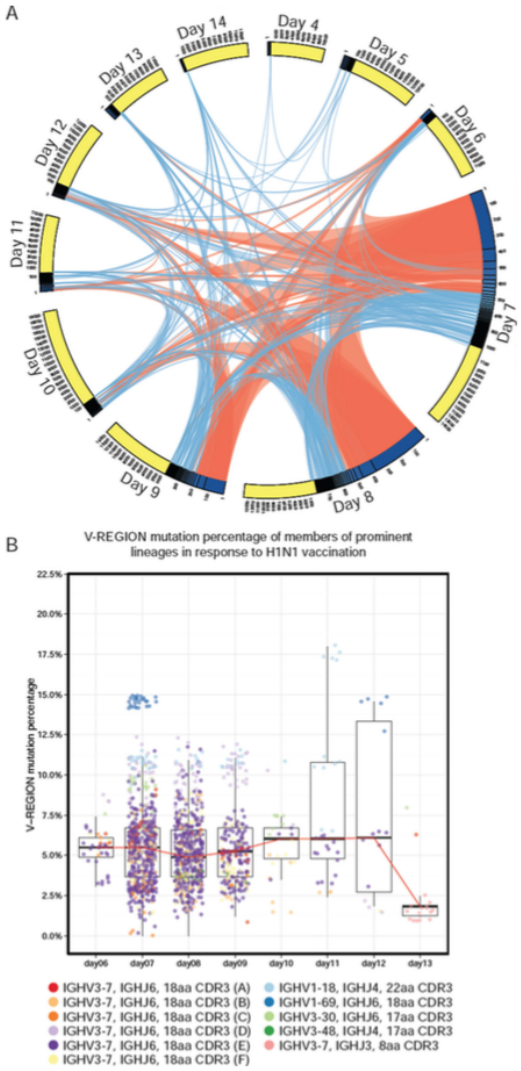


Figure 4.4: Time course of B cell clonal expansions induced by inactivated H1N1 vaccine (A) Vaccine induced B cell clonal lineage relationships during H1N1 vaccine response (complete IGH repertoires at Figure S2). Each arc represents a post-vaccination time point sub-divided into expanded (blue) and unexpanded (yellow) compartments. Lines join lineage members at later time points. Orange lines indicate that a lineage was strongly induced at the originating time point ($>$ 0.1% of total IGH). (B) IGHV percentage mutation for strongly induced lineages. Median (red line), quartiles shown as box and whiskers and points are jittered to prevent over-plotting.

70-1B03, differed from IGH isolated from BFI-278 by two CDR3 residues and was reported to be cross-reactive between pandemic H1N1 and the 2009 annual TIV vaccine antigens (Figure 4 C) (Wrammert et al., 2011). Two further IGHs, 70-5E04H and 70- 1F05H, were derived from H1N1-infection-induced plasma- blasts but had no detectable binding to the panel of antigens tested (Wrammert et al., 2011).

We examined somatic mutations in IGHV to assess additional evidence of molecular convergence in these antibodies. Overall, the IGHV of convergent H1N1-specific lineage members from BFI-278 showed an average of 5.24% mutation, and the most highly mutated sequence showed 12.27% mutation. Somatically mutated positions in CDR1 and CDR2 were shared with 4K8 and 2K11 mAbs (Figure 4 D). Shared substitutions could result from intrinsic mutation hotspots favoring mutation at these locations. This is very unlikely for the observed convergent mutations, as the CDR1 and CDR2 mutations in the BFI-278 H1N1 antibody sequences were uncommon in two unrelated, nonvaccinated data sets: IGHV3-7 IGH from the prevaccination of 14 TIV 2007 and 2008 subjects (Figure 4 D) and IGHV3-7 IGH from 27 healthy subjects who had not been recently vaccinated (Wang et al., 2014) (both $p < 0.0001$, Pearson's dependent groups). Impor- tantly, the two nonvaccination data sets were highly similar in their mutations ($r = 0.9311$), and both were distinct from the H1N1 lineages ($p = 0.1794$). Evaluation of mutated positions in mAbs 48K and 2K11 for whether they were more likely to have been drawn from the mutation distribution of the H1N1 lineage or from the two nonvaccination data sets showed that 4K8 car- ried mutations similar to those of the H1N1 lineage (log-likelihood ratio =

7.98 for H1N1 and

22.59 for the IGHV3-7 background), whereas 2K11 did not show higher similarity to either distribution (

41.97 for H1N1 and

42.52 for the IGHV3-7 background).

The response of BFI-278 to the next year's 2010 TIV, which included pandemic H1N1, was also examined. The 2010 TIV response included 28 IGHs with stereotypic features at day 7 postvaccination (0.36%). These differed from the prior year's stereotypic clones by at least four CDR3 residues, from the H1N1-stimulated B cells from

Wrammert et al. (2011) by more than five positions, and from the H1N1-specific mAbs from Krause et al. (2011) by six positions. The 2010 TIV IGH also lacked convergent somatic mutations and therefore appeared to represent distinct B cell clones.

IGH Rearrangements with Stereotypic Features Prior to 2009

To evaluate whether the convergent B cell clones were present at high frequencies in BFI-278's repertoire prior to pandemic H1N1 antigen exposure, we studied previously reported sam- ples collected 14 months apart in 2006 and 2008 (Boyd et al., 2009). Only four IGHs from 2006 (0.03%) and ten IGHs from 2008 (0.06%) used IGHV3-7 and IGHJ6 with an 18 amino acid CDR3. The CDR3s of the pre-2009 IGHs differed from the H1N1 response clones by at least six amino acids and appeared to be unrelated.

The contribution of stereotypic IGH chains in independent populations was also analyzed in 151,217 unique IGHs pooled from IGH sequencing of 27 additional adults (2008 and 2009) (Wang et al., 2014). IGH using IGHV3-7 and IGHJ6 represented 1.01% of total rearrangements, and 115 of these IGHs had an 18 amino acid CDR3. The 115 rearrangements did not share similar CDR3 sequences with IGH from BFI-278's H1N1 vaccination; all had less than 85% sequence identity. Among 496,104 IGH sequences from the 14 TIV 2007 and 2008 subjects, a single IGH (seroconverter 7014, 2008 TIV) differed from a rearrange- ment from individual BFI-278's H1N1 2009 response by one amino acid residue. Examination of over 500,000 IGH sequences collected from 41 individuals prior to 2009 therefore revealed only a single example of the convergent IGH seen in 2009 in BFI-278 and in the Wrammert et al. (2011) and Krause et al. (2011) studies, indicating that antibody lineages capable of binding the pandemic H1N1 strain were indeed present in human B cell repertoires but were relatively rare.

4.3.3 Discussion

Deep sequencing of immunoglobulin libraries enables tracking of known antigen-specific B cell populations (Liao et al., 2011; Liao et al., 2009; Wu et al., 2011

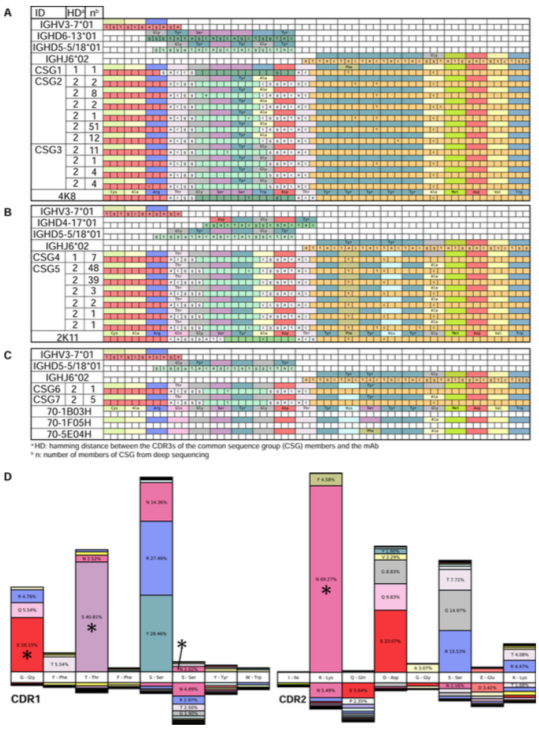


Figure 4.5: Convergent immunoglobulin heavy chain rearrangements identified in H1N1 vaccine responses in different individuals Multiple sequence alignment of closely related stereotypic H1N1 clones isolated from difference sources. Nucleotide positions are labeled when they differ from the germline genes (top of each block). Nucleotides inferred to be derived from non-templated nucleotide addition during V(D)J rearrangement are not colored. Amino acids positions are labeled when they differ from those in the mAb (bottom of each block). Each block is labeled with the germline, the conserved sequence group (CSG) and the mAb. Members of CSGs are detailed in Table S4. (A) IGH related to 4K8 mAb isolated (Krause et al., 2011) (B) IGH related to 2K11 (Krause et al., 2011) (C) IGH related to CDR3s previously identified by Wrammert et al. (Wrammert et al., 2011). (D) The percentage of total sequences with CDR1 and CDR2 somatic mutation amino acid substitutions in convergent H1N1 clones (upper) compared to 4,784 IGH using IGHV3-7 from the pre-vaccination repertoire of 14 subjects pre-pandemic (lower). The germline sequence is depicted in the white boxes (middle). Asterisks (*) mark amino acid substitutions shared with mAbs 48K or 2K11 (Krause et al., 2011).

). In addition, replicate libraries of B cells from an individual can identify clonally expanded popula- tions arising in response to immune stimuli, such as vaccination or infection. Recently, influenza vaccine responses in humans have been associated with increased numbers of plasmablast immunophenotype B cells in the blood; these B cells peak at approximately 7 days postexposure, and the majority express antibodies specific to influenza antigens and often feature signif- icant clonal expansion of particular plasmablast lineages (Brok- stad et al., 1995; Cox et al., 1994; Halliley et al., 2010; Sasaki et al., 2007; Wrammert et al., 2008). Prior deep-sequencing analysis of responses to influenza vaccination has identified sequences present at elevated levels in libraries amplified from cDNA of peripheral-blood B cells (Jiang et al., 2013). Our results reveal a strong correlation between the levels of B cell clonal expansion detected in the blood and the serological response to vaccination. Vaccine-induced clonal lineages measured by replicate library sequencing peaked at day 7 after having increased 118-fold over background lineages and persisted at detectable levels until about 2 weeks postvaccination.

To evaluate the antigen specificity of the clones detected by sequencing, we compared our data to the sequences of mAbs generated from plasmablasts at day 7 postvaccination individ- uals who seroconverted to seasonal TIV (Moody et al., 2011). Almost 25% of the IGH utilized by the plasmablast-derived mAbs shared clonal lineages with those we identified, and 66% of these were specific to influenza. This cross-identification confirms that a relatively simple evaluation of clonal expansions in the vaccine response by IGH repertoire sequencing highlights antigen-specific lineages. In addition, the incomplete overlap suggests that there are low-frequency plasmablast lineages that are not identified as clonally expanded and that there may be other clonally expanded B cell populations that do not have a plasmablast immunophenotype, such as influenza-specific memory B cells, which have been previously reported to signifi- cantly increase their levels in the blood at day 7 after influenza vaccination (Wrammert et al., 2008).

Unexpectedly, the response to monovalent H1N1 vaccination showed striking stereotypic features (use of IGHV3-7, IGHJ6, and an 18 amino acid CDR3) in the most prominent postvaccination clonally expanded B cell lineages. We looked for evidence

of this stereotypic IGH response in other individuals responding to influenza vaccination or infection and expected that the likeli- hood of finding "convergent evolution" in antibody responses to influenza would be low, given that the immunoglobulin repertoire is potentially very large, that deep sequencing still permits relatively shallow sampling of each individual's repertoire (Boyd et al., 2009; Glanville et al., 2011), and that selection pres- sures on influenza strains avoid common immune responses. To our surprise, IGH from two independent H1N1 response studies (Krause et al., 2011; Wrammert et al., 2011) shared the stereo- typic features. The CDR3s of these rearrangements differed by only 1 or 2 of 18 amino acids, highlighting a striking degree of convergence in the antibody responses in these unrelated indi- viduals. Paired with the light chain from the previously reported 48K mAb (Wrammert et al., 2011), a recombinant mAb of the BFI-278 stereotypic IGH lineage showed HAI specific to the pandemic H1N1 strain. Notably, in a search of over 500,000 IGH sequences from the B cell repertoires of 41 subjects studied from 2006 to 2009, we identified a single IGH with the same convergent features, highlighting the low prior frequency of such clones that are preferentially stimulated by the 2009 pandemic H1N1 influenza strain (Brokstad et al., 1995; Sasaki et al., 2007).

The use of IGH involving IGHV3-7, IGHJ6, and an 18 amino acid CDR3 was a recurrent pattern of the response of BFI-278 (Figure 3) to successive years of H1N1 influenza vaccination, but importantly, the stereotyped B cell clones in 2009 were distinct from those that appeared in 2010 after vaccination with TIV containing the same pandemic H1N1 antigens. This suggests the recruitment of new B cells with stereotypic IGH chains rather than a recall response dominated by the highest-frequency members of the prior year's clones. This favoring of new clones could be due to the fact that serum antibody derived from the 2009 vaccination decreases restimulation of the dominant 2009 B cell clones but doesn't prevent stimulation of B cells that recognize somewhat different epitopes of the viral antigens.

In summary, high-throughput DNA sequencing of peripheral- blood B cells provides a highly informative measure of clonal expansions, and such a measure correlates with serological vaccine responses. B cell clones that expand after vaccination

show substantial but incomplete clonal overlap with vaccine- specific single-cell plasmablasts. Responses to vaccination with the 2009 pandemic H1N1 influenza strain revealed a domi- nant pattern of antibody responsiveness that was convergent in different individuals. Antibodies with convergent features were rare prior to 2009. Overall, if the detection of convergent anti- body signatures can be generalized to other antigens and in- fectious diseases, it may be feasible to use IGH repertoire sequencing to assess an individual's history of antigenic expo- sures and infections more broadly (Glanville et al., data not shown). The results also lend some support to the idea that evolutionary selection of the germline IGH genes may predis- pose the adaptive immune response to follow common paths of response to common pathogens. Further evaluation in the context of different vaccines and pathogen infections will be able to determine whether these findings are the exception or the rule in human antibody responses.

4.3.4 Methods

Collection of Specimens All subject recruitment was performed with informed consent. The Duke University institutional review board (IRB) approved protocols for the study of response to 2007 and 2008 seasonal TIV. The Stanford University IRB approved the study of the 27 additional prevaccination subjects. Subjects re- cruited at Duke University were given either 2007–2008 or 2008–2009 FluZone TIV (Sanofi Pasteur) as described in Table S1 and detailed in the Supplemental Information . Blood was drawn before vaccination and on days 7 and 21 after challenge. Subjects recruited at Stanford University were 27 healthy individ- uals aged 20–89 years (Wang et al., 2014). Peripheral-blood samples from these patients were collected prior to vaccination in two successive years (2008 and 2009). Subject BFI-278, recruited at Stanford University, was administered the 2009 monovalent inactivated vaccination to influenza A (H1N1) in January 2010. Blood was drawn before vaccination and daily on days 4–14 after chal- lenge. Ten months after receiving the monovalent H1N1 vaccine, BFI-278 was administered a seasonal TIV that included the influenza

A/California/07/2009- like pandemic H1N1 antigenic component. Blood was drawn prior to 2010 TIV vaccination and on days 7 and 21 postvaccination.

Sample Preparation and PCR Amplification for Deep Sequencing Peripheral-blood mononuclear cells (PBMCs) were isolated by centrifugation of blood layered over Histopaque-1077 (Sigma-Aldrich) and cryopreserved (Moody et al., 2011). Column purification (QIAGEN) was used for isolating genomic DNA template. IGHs were amplified from genomic DNA template with the previously described primer design for 454 instrument sequencing as detailed in the Supplemental Information (Boyd et al., 2009).

Deep Sequencing of IGH Amplicon library pools were quantified by real-time PCR (Roche). Sequencing data presented here were derived from the 454 instrument with the use of Titanium chemistry; long-range amplicon pyrosequencing began from the "B" primer in the manufacturer's protocol (Roche). The complete data sets from the samples generated are available at dbGaP accession number phs000760.v1.p1.

Analysis of Sequence Data Sequences from each input specimen were demultiplexed with the sample and replicate library barcodes as detailed in the Supplemental Information . Alignment of rearranged IGH sequences to germline V, D, and J and determi- nation of V-D junctions and D-J junctions were performed with iHMMune-align (Gae¨ ta et al., 2007). IGHs were assigned to clonal lineages by means of clustering on CDR3 nucleotide similarity for IGHs that shared the same IGHV, IGHJ, and CDR3 length. CDR3 sequence similarity was measured by Hamming distance, and clusters were assigned with a Hamming distance of 95% identity to any existing sequence and at least 80% identity within the cluster (Figure S1).

Scale-Independent Normalized Measure of Overall Clonality To compare the degree of B cell clonal expansion detected in data sets with varying sequencing depth, we used a scale-independent metric we have pre- viously described (Wang et al., 2014). This clonality metric can be considered the probability that any two randomly selected sequences drawn from inde- pendent replicates are members of the same clonal lineage.

Analysis of Hypermutation Spectra CDR1 and CDR2 sequences, as defined by IMGT criteria, were extracted from iHMMune-align results and translated to amino

acid sequences. CDR1 and CDR2 position count matrices of mutations were converted to substitution frequency vectors. Pearson's correlation and Pearson's test for dependent correlations were used for assessing whether there were differences between mutation spectra for different data sets. Log-likelihood ratios were used for exploring single-sequence mutation distributions, allowing comparison of the likelihood that the observed substitutions were drawn from the mutation distributions of different data sets.

Single-Cell Plasmablast Sorting, mAb Expression, and Quantification of Plasmablast Frequencies PBMCs were cryopreserved with standard methods, and single-cell sorting was performed as previously described (Moody et al., 2011). Flow cytometry was carried out prior to single-cell IGH V(D)J PCR and expression of recombi- nant IgG1 mAbs as previously described (Liao et al., 2009; Moody et al., 2011) and detailed in the Supplemental Information . In addition, samples collected prior to vaccination and days 7 and 21 postvaccination were assessed for plasmablast and B cell phenotypes.

Serum Antibody Binding Tests Plasma samples were evaluated by ELISA with purified HAs and split vaccine preparations as detailed in the Supplemental Information . Five-parameter curve fits were used for data analysis. Endpoint titers were calculated as 3-fold above assay background, and the assay cutoff was a 1:25 dilution. Expressed mAbs were tested for binding to influenza antigen by ELISA, as previously described (Moody et al., 2011). Reactivity to influenza antigens was also studied with a standardized custom Luminex assay.

HAI of mAbs HAI assays of mAbs were performed as described elsewhere (Davtyan et al., 2011; Wang et al., 2006) and are detailed in the Supplemental Information . Working stocks of influenza were standardized to a HA titer of 8 HA units per 50 m l for assays for A/California/4/2009 (H1N1), A/Brisbane/59/2007 (H1N1), A/Brisbane/10/2007 (H3N2), B/Brisbane/60/2008, and B/Florida/4/ 2006. The HAI titer was defined as the reciprocal of the highest dilution of anti- body that inhibits red blood cell hemagglutination by influenza virus.

Accession numbers

The dbGaP accession number for the metadata and sequences reported in the paper is phs000760.v1.p1

4.3.5 Acknowledgements

This work was made possible by my co-authors, including first author Katherine J.L. Jackson, Yi Liu, Krishna M. Roskin, Ramona A. Hoh, Katie Seo, Eleanor L. Marshall, Thaddeus C. Gurley, M. Anthony Moody, Barton F. Haynes, Emmanuel B. Walter, Hua-Xin Liao, Randy A. Albrecht, Adolfo Garcıa-Sastre, Javier Chaparro-Riggers, Arvind Rajpal, Jaume Pons, Birgitte B. Simen, Bozena Hanczaruk, Cornelia L. Dekker, Jonathan Laserson, Daphne Koller, Mark M. Davis, Andrew Z. Fire, and Scott D. Boyd The authors thank Sally Mackey for project, regulatory, and data management; research nurses Sue Swope and Cynthia Walsh; phlebotomist Michele Ugur; and research assistant Kyrsten Spann for scheduling and conducting the study visits. This work was supported by NIH grants U19AI090019, P01AI089618, AI0678501, U19AI067854, and 1U19AI089987 and grants from the Ellison Medical Foundation to M.M.D. and S.D.B. We would like to thank Thomas Kepler for helpful discussions about data analysis of sequences derived from the single-cell-sorting experiments.

4.3.6 References

Ademokun, A., Wu, Y.C., Martin, V., Mitra, R., Sack, U., Baxendale, H., Kipling, D., and Dunn-Walters, D.K. (2011). Vaccination-induced changes in human B-cell repertoire and pneumococcal IgM and IgA antibody at different ages. Aging Cell 10 , 922–930.

Boyd, S.D., Marshall, E.L., Merker, J.D., Maniar, J.M., Zhang, L.N., Sahaf, B., Jones, C.D., Simen, B.B., Hanczaruk, B., Nguyen, K.D., et al. (2009). Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. Sci. Transl. Med. 1 , 12ra23.

Brokstad, K.A., Cox, R.J., Olofsson, J., Jonsson, R., and Haaheim, L.R. (1995). Parenteral influenza vaccination induces a rapid systemic and local immune response. J. Infect. Dis. 171 , 198–203.

Cox, R.J., Brokstad, K.A., Zuckerman, M.A., Wood, J.M., Haaheim, L.R., and Oxford, J.S. (1994). An early humoral immune response in peripheral blood following parenteral inactivated influenza vaccination. Vaccine 12 , 993–999.

Davtyan, H., Ghochikyan, A., Cadagan, R., Zamarin, D., Petrushina, I., Movsesyan, N., Martinez-Sobrido, L., Albrecht, R.A., Garcı´a-Sast re, A., and Agadjanyan, M.G. (2011). The immunological potency and therapeutic poten- tial of a prototype dual vaccine against influenza and Alzheimer's disease. J. Transl. Med. 9 , 127.

DeKosky, B.J., Ippolito, G.C., Deschner, R.P., Lavinder, J.J., Wine, Y., Rawlings, B.M., Varadarajan, N., Giesecke, C., Do¨ rner, T., Andrews, S.F., et al. (2013). High-throughput sequencing of the paired human immunoglob- ulin heavy and light chain repertoire. Nat. Biotechnol. 31 , 166–169.

Gae¨ ta, B.A., Malming, H.R., Jackson, K.J., Bain, M.E., Wilson, P., and Collins, A.M. (2007). iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. Bioinformatics 23 , 1580–1587.

Glanville, J., Kuo, T.C., von Bu¨ dingen, H.-C., Guey, L., Berka, J., Sundar, P.D., Huerta, G., Mehta, G.R., Oksenberg, J.R., Hauser, S.L., et al. (2011). Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. Proc. Natl. Acad. Sci. USA 108 , 20066–20071.

Halliley, J.L., Kyu, S., Kobie, J.J., Walsh, E.E., Falsey, A.R., Randall, T.D., Treanor, J., Feng, C., Sanz, I., and Lee, F.E. (2010). Peak frequencies of circulating human influenza-specific antibody secreting cells correlate with serum antibody response after immunization. Vaccine 28 , 3582–3587.

Jiang, N., He, J., Weinstein, J.A., Penland, L., Sasaki, S., He, X.-S., Dekker, C.L., Zheng, N.-Y., Huang, M., Sullivan, M., et al. (2013). Lineage Structure of the Human Antibody Repertoire in Response to Influenza Vaccination. Sci. Transl. Med. 5 , 171ra119.

Krause, J.C., Tsibane, T., Tumpey, T.M., Huffman, C.J., Briney, B.S., Smith, S.A., Basler, C.F., and Crowe, J.E., Jr. (2011). Epitope-specific human influ- enza antibody repertoires diversify by B cell intraclonal sequence divergence and interclonal convergence. J. Immunol. 187 , 3704–3711.

Li, G.-M., Chiu, C., Wrammert, J., McCausland, M., Andrews, S.F., Zheng, N.-Y., Lee, J.-H., Huang, M., Qu, X., Edupuganti, S., et al. (2012). Pandemic H1N1 influenza vaccine induces a recall response in humans that favors broadly cross-reactive memory B cells. Proc. Natl. Acad. Sci. USA 109 , 9047–9052.

Liao, H.-X., Chen, X., Munshaw, S., Zhang, R., Marshall, D.J., Vandergrift, N., Whitesides, J.F., Lu, X., Yu, J.-S., Hwang, K.-K., et al. (2011). Initial antibodies binding to HIV-1 gp41 in acutely infected subjects are polyreactive and highly mutated. J. Exp. Med. 208 , 2237–2249.

Liao, H.-X., Levesque, M.C., Nagel, A., Dixon, A., Zhang, R., Walter, E., Parks, R., Whitesides, J., Marshall, D.J., Hwang, K.-K., et al. (2009). High-throughput isolation of immunoglobulin genes from single human B cells and expression as monoclonal antibodies. J. Virol. Methods 158 , 171–179.

Moody, M.A., Zhang, R., Walter, E.B., Woods, C.W., Ginsburg, G.S., McClain, M.T., Denny, T.N., Chen, X., Munshaw, S., Marshall, D.J., et al. (2011). H3N2 influenza infection elicits more cross-reactive and less clonally expanded anti-hemagglutinin antibodies than influenza vaccination. PLoS ONE 6 , e25797.

Park, M.K., Sun, Y., Olander, J.V., Hoffmann, J.W., and Nahm, M.H. (1996). The repertoire of human antibodies to the carbohydrate capsule of Streptococcus pneumoniae 6B. J. Infect. Dis. 174 , 75–82.

Sasaki, S., He, X.-S., Holmes, T.H., Dekker, C.L., Kemble, G.W., Arvin, A.M., and Greenberg, H.B. (2008). Influence of prior influenza vaccination on anti- body and B-cell responses. PLoS ONE 3 , e2975.

Sasaki, S., Jaimes, M.C., Holmes, T.H., Dekker, C.L., Mahmood, K., Kemble, G.W., Arvin, A.M., and Greenberg, H.B. (2007). Comparison of the influenza virus-specific effector and memory B-cell responses to immunization of chil- dren and adults with live attenuated or inactivated influenza virus vaccines. J. Virol. 81 , 215–228.

Scott, M.G., Tarrand, J.J., Crimmins, D.L., McCourt, D.W., Siegel, N.R., Smith, C.E., and Nahm, M.H. (1989). Clonal characterization of the human IgG antibody repertoire to Haemophilus influenzae type b polysaccharide. II. IgG antibodies contain VH genes from a single VH family and VL genes from at least four VL families. J. Immunol. 143 , 293–298.

Silverman, G.J., and Lucas, A.H. (1991). Variable region diversity in human circulating antibodies specific for the capsular polysaccharide of Haemophilus influenzae type b. Preferential usage of two types of VH3 heavy chains. J. Clin. Invest. 88 , 911–920.

Steiger, J.H. (1980). Tests for comparing elements of a correlation matrix. Psychol. Bull. 87 , 245–251. Tonegawa, S. (1983). Somatic generation of antibody diversity. Nature 302 , 575–581.

Wang, C., Liu, Y., Xu, L.T., Jackson, K.J.L., Roskin, K.M., Pham, T.D., Laserson, J., Marshall, E.L., Seo, K., Lee, J.-Y., et al. (2014). Effects of aging, cytomegalovirus infection, and EBV infection on human B cell repertoires. J. Immunol. 192 , 603–611.

Wang, S., Taaffe, J., Parker, C., Solo´ rzano, A., Cao, H., Garcı´a-Sastre, A., and Lu, S. (2006). Hemagglutinin (HA) proteins from H1 and H3 serotypes of influenza A viruses require different antigen designs for the induction of optimal protective antibody responses as studied by codon-optimized HA DNA vaccines. J. Virol. 80 , 11628–11637

Wrammert, J., Koutsonanos, D., Li, G.-M., Edupuganti, S., Sui, J., Morrissey, M., McCausland, M., Skountzou, I., Hornig, M., Lipkin, W.I., et al. (2011). Broadly cross-reactive antibodies dominate the human B cell response against 2009 pandemic H1N1 influenza virus infection. J. Exp. Med. 208 , 181–193.

Wrammert, J., Smith, K., Miller, J., Langley, W.A., Kokko, K., Larsen, C., Zheng, N.-Y., Mays, I., Garman, L., Helms, C., et al. (2008). Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. Nature 453 , 667–671.

Wu, X., Zhou, T., Zhu, J., Zhang, B., Georgiev, I., Wang, C., Chen, X., Longo, N.S., Louder, M., McKee, K., et al.; NISC Comparative Sequencing Program (2011). Focused evolution of HIV-1 neutralizing antibodies revealed by struc- tures and deep sequencing. Science 333 , 1593–1602.

Yu, X., Tsibane, T., McGraw, P.A., House, F.S., Keefer, C.J., Hicar, M.D., Tumpey, T.M., Pappas, C., Perrone, L.A., Martinez, O., et al. (2008). Neutralizing antibodies derived from the B cells of 1918 influenza pandemic survivors. Nature 455 , 532–536.

4.3.7 Copyright

4.4 Dietary gluten triggers convergent T cells in celiac disease

Celiac disease is an intestinal autoimmune disease driven by dietary gluten and gluten-specific CD4+ T-cell responses. In celiac patients on a gluten-free diet, exposure to gluten induces the appearance of gluten-specific CD4+ T cells with gut-homing potential in the pe- ripheral blood. Here we show that gluten exposure also induces the appearance of activated, gut-homing CD8+ αβ and γδ T cells in the peripheral blood. Single-cell T-cell receptor sequence analysis indicates that both of these cell populations have highly focused T- cell receptor repertoires, indicating that their induction is antigen- driven. These results reveal a previously unappreciated role of an- tigen in the induction of CD8+ αβ and γδ T cells in celiac disease and demonstrate a coordinated response by all three of the major types of T cells. More broadly, these responses may parallel adaptive immune responses to viral pathogens and other systemic auto- immune diseases.

4.4.1 Introduction

Celiac disease (CD) is a common autoimmune disease with an estimated prevalence of 1% among people of European ancestry. It is characterized by small intestinal mucosal injury and nutrient malabsorption in genetically susceptible individuals due to dietary gluten ingestion. CD4+ T cells bearing αβ T-cell receptors (TCRs) are critical in the pathogenesis of the disease, as it occurs almost exclusively in HLA-DQ2– or HLA-DQ8– positive individuals (1, 2). CD-associated gluten peptide CD4+

T-cell epitopes have been discovered, and HLA-DQ2/8– restricted gluten-reactive CD4+ T cells have been identified in individuals with CD (3–5). Nonetheless, no gluten-induced en- teropathy is seen in humanized mouse models expressing HLA-DQ2 and a gluten-specific TCR (6, 7), suggesting that CD4+ T cells alone are unable to induce tissue damage in CD (1, 2).

An increase in intestinal intraepithelial lymphocytes (IELs), composed of both CD8+ αβ T cells and γδ T cells, is a hallmark of CD. IELs are responsible for the detrimental consequences of CD, including tissue damage and lymphoma development. CD8+ TCRαβ+ IELs (CD8+ IELs) function as effectors in protective immunity to pathogens (8), and in CD they assume a natural killer (NK)-like phenotype to kill intestinal epithelial cells in a manner independent of TCR specificity (9). In rare instances, IELs in CD may transform into enteropathy-associated T-cell lymphoma (EATL), an aggressive lymphoma with a very poor prognosis (10). EATL cells have been shown to have clonal TCRαβ or TCRγδ rearrangements, indicating that either CD8+ IELs or γδ IELs may give rise to lymphoma (11, 12).

Despite intense efforts, gluten-specific IELs in CD have not been readily identified, and there is no significant genetic asso- ciation of CD with any HLA class I alleles. Moreover, the cy- tolytic function of IELs in CD can be induced irrespective of their TCR specificity (9). Thus, although the link between dietary gluten and the CD4+ response is well-established, the link be- tween dietary gluten and the recruitment and activation of CD8+ or γδ IELs in celiac disease is unknown. Furthermore, the role of the antigen specificity of IELs in CD is unclear. Here we find that CD8+ and γδ T cells bearing gut-homing receptors are induced by gluten ingestion in CD patients in parallel with gluten-specific CD4+ T cells, and they bear TCR sequences that indicate an antigen-focused response. This indicates that antigen-specific responses of all three of these major T-cell types play a role in this disease.

4.4.2 Results

Celiac disease requires the continuous presence of dietary glu- ten. Reintroducing dietary gluten to celiac patients who are on a gluten-free diet induces large numbers

of gluten-specific CD4+ T cells in the peripheral blood 6 d later (4, 5, 13). These cells express the β7 integrin receptor, indicating that they will home to the intestine (5). They also express the activation marker CD38 and lack the expression of CD62L, consistent with an effector phenotype (14). This is generally thought to represent the initi- ation of an immune response to gluten, and captures activated gluten-reactive CD4+ effector T cells en route from mesenteric lymph nodes or gut-associated lymphoid tissue to the intestine. In an effort to better characterize the context of this immune response, we studied peripheral blood T cells in celiac patients undergoing gluten challenge by time-of-flight mass cytometry (CyTOF) (15), which allows for the independent assessment of many more cellular parameters (currently >40) than fluores- cence-based flow cytometry. Indeed, we observed an increase in gluten peptide/HLA-DQ2 tetramer-positive CD4+ T cells in the peripheral blood in all five HLA-DQ2+ celiac patients on day 6 following gluten challenge (Fig. 1 A and C). Unexpectedly, we also observed a large increase in the number of peripheral blood CD8+ αβ and γδ T cells expressing the intestinal epithelial- homing markers αE (CD103) and β7 integrins (16) and the ac- tivation marker CD38 (Fig. 1 A and B and Table S1) at this same time point. These cells were not detected in healthy HLA-DQ2+ controls, who underwent oral gluten challenge after at least 1 mo on a gluten-free diet.

The kinetics with which these CD8+ and γδ T cells appear is the same as that of gluten-specific CD4+ T cells, peaking at day 6 after gluten challenge and declining to the baseline by day 14 (Fig. 1C). A similar response was also detected in two celiac patients who underwent rechallenge after returning to a gluten- free diet for at least 1 mo (Fig. 1 A and B and Table S1).

The magnitude of the peripheral blood gluten-specific CD4+ T-cell response is known to be quite variable (4). Similarly, the extent of the αEβ7+CD38+ T-cell response varied between patients, ranging from 0.37% to 10.17% of total peripheral blood CD8+ and from 0.06% to 18.61% of total peripheral blood γδ T cells (Fig. 1B and Table S1). One celiac patient (celiac 2) had αEβ7+CD38+CD8+ and γδ T cells above background levels on day 0, but showed a further increase following gluten

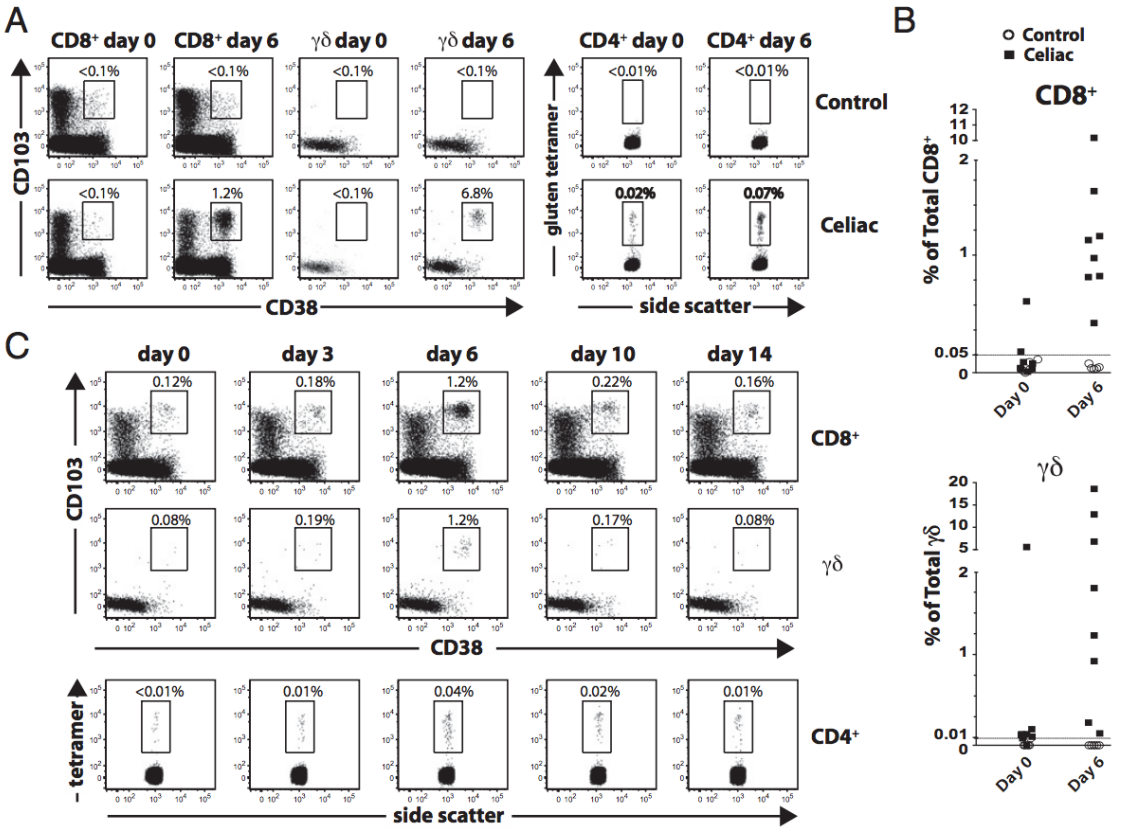


Figure 4.6: nduction of activated, gut-homing CD8 αβ and γδ T cells in peripheral blood of celiac patients following oral gluten challenge. (A) Representative FACS analysis of CD8+ αβ and γδ T-cell (Left) and CD4 + T-cell (Right) response to oral gluten challenge in CD vs. nonceliac control. Expansion of CD103+ (αE integrin), CD38+, and gluten tetra- mer+ CD4+ T-cell populations is seen on day 6 in CD. Most CD38 CD103 cells also express β7 integrin; only CD103 staining is depicted here. (B) Relative frequency of αEβ7+CD38+ CD8+ T cells as a percentage of total CD8+ cells (Top) and relative fre- quency of αEβ7+CD38+ γδ cells as a percentage of total γδ T cells (Bot- tom). (C) Time course showing rel- ative percentage of CD38+CD103+ CD8+ (Top), CD38+CD103+ γδ (Mid- dle), and gluten tetramer+ CD4+ (Bottom) in the same patient at the indicated time points following oral gluten challenge. Parallel recruit- ment of CD38+CD103+ and gluten tetramer+ cells peaks on day 6 before returning to baseline.

challenge. The individual with the lowest detectable response (celiac 6) was an HLA-DQ8+ celiac patient whose disease was diagnosed in- cidentally by intestinal biopsy, had equivocal antibody test results, and has always been clinically asymptomatic to gluten. Three individuals with active celiac disease, as determined by ongoing symptoms and positive autoantibody titers, were found to have αEβ7+CD38+CD8+ and γδ T-cell proportion below background levels of 0.05% and 0.01%, respectively (Fig. S1). This aspect is similar to the absence of gluten-specific CD4+ T cells in periph- eral blood of patients with active celiac disease (4, 5). Also, al- though plasma cells secreting anti-gluten and autoantibodies are present in celiac intestinal lesions (17–19), we did not detect a similar increase in intestinal-homing B cells (not shown). This is consistent with reports indicating that tissue transglutaminase- specific B cells were undetectable in the peripheral blood of celiac patients (19, 20). In summary, dietary gluten induces the activa- tion and concomitant peripheral blood presence of CD4+ and CD8+ αβ T cells and γδ T cells with gut-homing potential in celiac patients who have been on a gluten-free diet (Fig. 1 and Table S1).

Gluten-reactive CD4+ T cells in the peripheral blood of celiac patients have been shown to be CD38+CD62L−, suggesting that they are gut-bound effector cells (7). CyTOF analysis showed that αEβ7+CD38+CD8+ T cells are CD38+, CD45RO+, CD27−, CD28low, CD62L−, and CCR7low (Fig. 2). This phenotype closely resembles the phenotype of CD8+ T cells isolated from duodenal tissue biopsy specimens of patients with active celiac disease (Fig. 2). CD8+ T cells of this phenotype have been re- ported to represent differentiated effectors and, accordingly, αEβ7+CD38+CD8+ T cells resemble peripheral blood effector memory CD8+ T cells (Fig. S2) (15, 21, 22). αEβ7+CD38+ γδ cells are predominantly CD45RO+ and CD27−, mirroring intestinal γδ cells from celiac biopsies (Fig. S3). CD45RO+, CD27− γδ T cells are thought to be memory cells (23).

The fact that gluten ingestion induces the activation of gluten- specific CD4+ T cells in CD is well-established. However, whether or not the CD8+ and γδ IELs induced in the intestine are responding to specific antigens is unknown. To address this question, we performed single-cell TCR sequencing, which provides a nonbiased means to assess the TCR repertoire with- out requiring expansion of T-cell clones

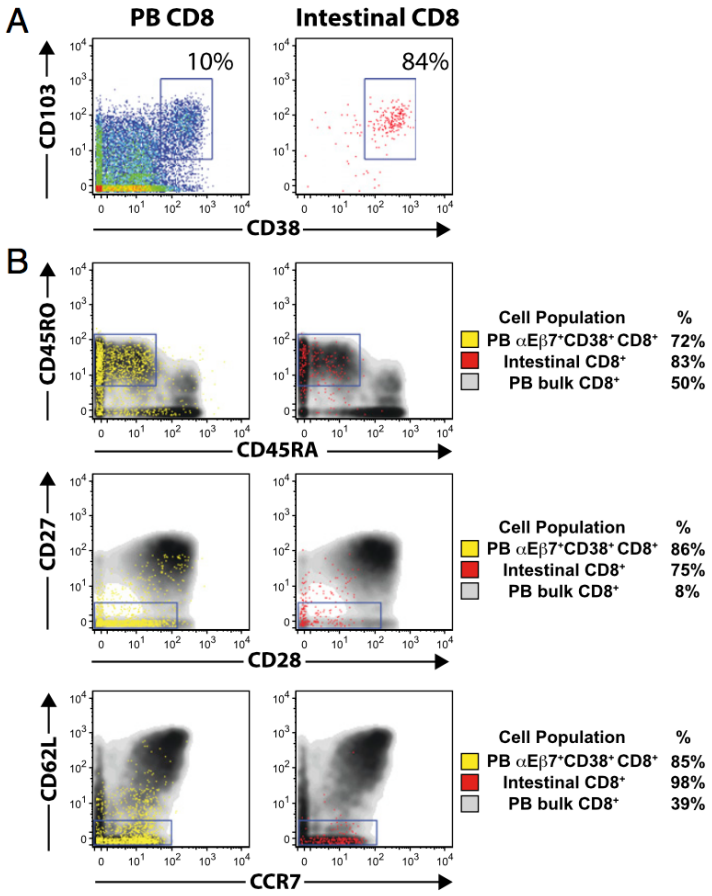


Figure 4.7: Peripheral blood αEβ7+CD38+CD8+ T cells induced by oral gluten challenge express surface markers of effector memory cells and resemble intestinal epithelial CD8+ T lymphocytes from celiac mucosal biopsies. (A) CyTOF analysis of total peripheral blood (PB) CD8+ from a gluten-challenged individual (Left) and total intestinal CD8+ T cells from a celiac patient with active disease (Right) with respect to CD103 and CD38 expression. (B) CyTOF analyses of peripheral blood αEβ7CD38+CD8+ T cells (yellow) and total in- testinal CD8+ T cells (red) are overlaid on total peripheral blood CD8+ T cells. Peripheral blood αEβ7+CD38+CD8+ and celiac intestinal CD8+ cells are pre- dominantly CD38+CD45RO+CD45RA−CD27−CD28lowCD62L−CCR7−, consistent with an effector memory phenotype.

in culture (24). Single T cells were sorted into 96-well PCR plates from peripheral blood samples of celiac patients following gluten challenge. TCRβ or TCRγ genes were amplified by a series of nested PCRs, and PCR products were directly sequenced.

We were able to perform sequencing on single T cells with high efficiency. We sorted and sequenced 90 single tetramer- positive CD4+ T cells recognizing the gluten epitope DQ2-α-II from the blood of two celiac patients on day 6 after oral gluten challenge (Table S2). Sequences were successfully obtained from 77/90 (86%) of wells into which single T cells were sorted. Consistent with published sequences of DQ2-α-II–reactive T cells from blood and tissue (25), the majority (79%) of unique TCRβ sequences of individual DQ2-α-II–tetramer+ T cells used TRBV7-2 and most (74%) contained the described dominant arginine in position 5 of the CDR3β loop (Table S2), thus val- idating our methodology.

We then sequenced αEβ7+CD38+CD8+ and γδ T cells iso- lated from celiac patients on day 6 following gluten challenge. αEβ7+CD38+CD8+ T cells, sequenced in five celiac patients, and αEβ7+CD38+ γδ T cells, sequenced in three celiac patients, were found to have a high degree of clonal expansion that was not observed in CD8+CD45RO+ control T cells (Fig. 3). αEβ7 T cells were sequenced in celiac patients who underwent repeat gluten challenge to determine whether both challenges would elicit a similar responding TCR repertoire. Indeed, identical TCRβ and TCRδ clones and similarity in frequency of common clones were found in the two gluten challenges of these patients who underwent repeat challenge after returning to a gluten-free diet for at least 1 mo (Fig. S4).

We next evaluated sequences from αEβ7+CD38+CD8+ and γδ T cells to determine whether we could observe a convergence of TCR features among distinct TCR sequences. To evaluate con- vergence, we analyzed the nonredundant, unique TCR reper- toire of αEβ7+CD38+ T cells. For a particular MHC–peptide, specific CD8+ T-cell responses are often biased toward the use of a particular TCRVβ gene (26). We initially examined TCRVβ gene use. Even individuals of significantly different genetic backgrounds share similar frequency of V gene use in their TCR repertoire, indicating that skewing within a particular population of cells is not attributable to genetic variation in baseline V gene use (27). When assessing the nonredundant

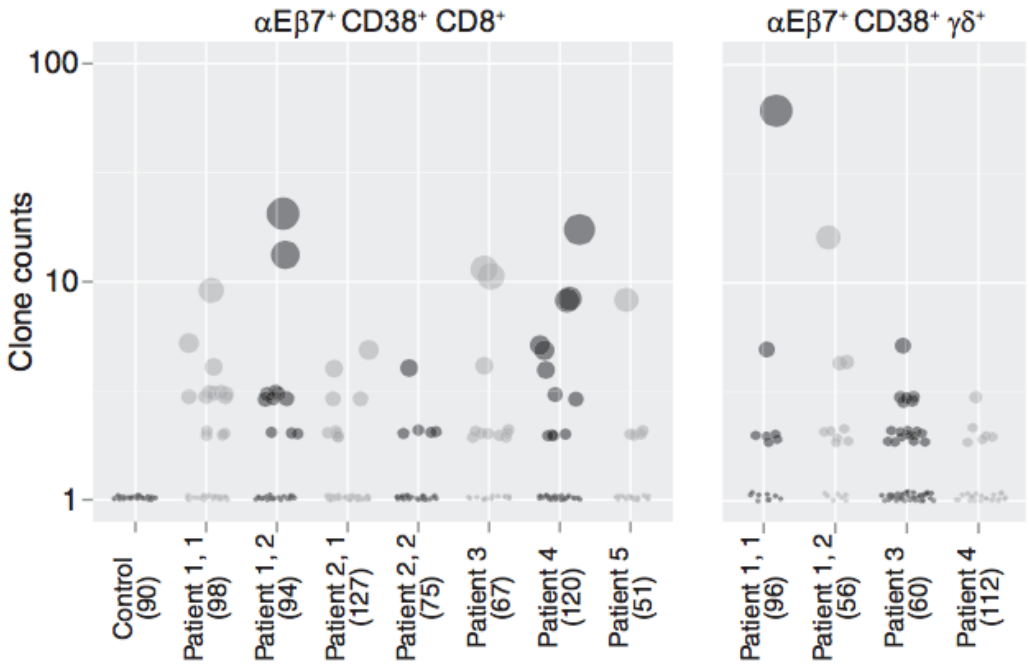


Figure 4.8: Single-cell TCR sequencing of peripheral blood αEβ7+CD38+CD8+ and αEβ7+CD38+ γδ T cells reveals clonal expansion upon gluten challenge in celiac disease. αEβ7+CD38+CD8+ TCRs were sequenced in five separate patients following gluten challenge, two of whom underwent rechallenge. αEβ7+CD38+ γδ TCRs were sequenced in three patients, one of whom un- derwent rechallenge. Each individual dot represents a distinct TCR clone. The size of dots and the position along the y axis, plotted on a log scale, indicate the relative frequency of a particular clone. The total number of clones se- quenced in each patient is indicated in parentheses.

TCRβ repertoire of αEβ7+CD38+CD8+ T cells in celiac samples, we found significant overrepresentation of particular V regions in multiple celiac samples compared with unselected healthy controls (Fig. S5 A and B).

Most of the peptide specificity of TCRβ is determined by the CDR3 loop, which is usually positioned over the antigenic pep- tide (28, 29). We then determined whether convergence could be observed within CDR3β motifs, focusing on groups using TCRVβ genes that were overrepresented in a nonredundant sampling within a particular individual and had members that were clonally expanded. Strikingly, in αEβ7+CD38+CD8+ T cells, we found four separate examples where identical TCRβ proteins used different DNA sequences (Fig. 4 A and B). In three of these instances, the identically convergent TCRβ occurred in the same patient, and represented a dominantly expressed TCRβ in that individual. In the other instance, the identical TCRβ occurred in different patients (Fig. 4A).

Additionally, within TCRVβ sequences using TRBV7-8, TRBV7-9, and TRBV28, we could identify characteristic amino acid motifs in the center of the CDR3β that were very common within celiac αEβ7+CD38+CD8+ T cells compared with healthy reference CDR3β sequences (30) (Fig. 4). For instance, the GN motif at positions 6–7 within the CDR3 region of TCRβ clones using TRBV7-9 was highly enriched, occurring in 16 out of 40 unique (nonredundant) TCRβ clones, while occurring in only 12/ 9,584 of TCRβ clones using TRBV7-9 within the reference da- tabase ($P < 0.0001$) (Fig. 4 A and C). In patient 4, this motif occurred in 14 of 19 unique TCRβ clones, and 5 of these unique clones converged on two identical TCRβs. This motif also oc- curred in two other patients, who converged upon an identical TCRβ. TCRβ clones using TRBV7-8 similarly converged on a GT motif at position 6–7, which occurred in 17 out of 29 unique TCRβ clones, in contrast to only 43/4,546 TRBV7-8– containing TCRβ clones within the reference database ($P < 0.0001$) (Fig. 4 B and C). In all instances where the same TCR was formed using distinct VDJ rearrangements within the same patient, there were at least two nucleotide changes within the CDR3, making a PCR or sequencing error improbable.

We applied a similar analysis to αEβ7+CD38+ γδ T cells. In- testinal γδ T cells are appreciated to be heavily biased toward TRDV1 use (31). Consistent with this, the

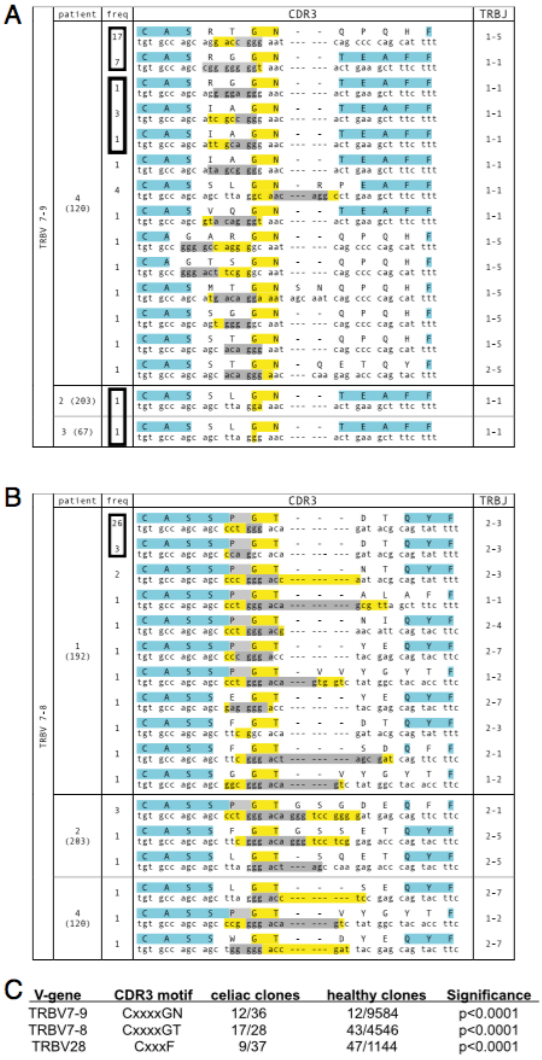


Figure 4.9: Convergent αEβ7+CD38+CD8+TCRβ CDR3 motifs are found among clones within the same celiac patient and across different patients following gluten challenge. (A and B) Convergent motifs CxxxxGN (A) and CxxxxGT (B) are seen in TCRβ clones using TRBV7-9 and TRBV7-8, respectively. The fre- quency of each clone is indicated and the total number of T cells sequenced in the patient is indicated in parentheses. The protein sequence with the cor- responding DNA sequence is shown. Within the protein sequence, yellow indicates absolutely conserved amino acids, gray indicates relatively con- served (≥50%) amino acids, and blue indicates conserved amino acids that are encoded within the V or J genes. Within the DNA sequence, nucleotides in yellow are formed through N or P addition, whereas nucleotides in gray are encoded by D genes. Boxes around frequency numbers highlight distinct clones sharing identical protein sequences. (C) Convergences of motifs seen in TCRβ clones using TRBV7-9, TRBV7-8, and TRBV28 are statistically significant compared with reference control TCRβ sequences.

majority (80%, 150/ 188) of unique αEβ7+CD38+ TCRδ sequences from CD patients use TRDV1 (Table S3). We analyzed CDR3δ sequences using TRDV1 to determine whether convergent motifs could be seen in celiac patients. For comparison, we sequenced TCRδ from bulk small intestinal γδ T cells from a person without celiac disease and bulk blood γδ T cells from nine different con- trol patients, obtaining 18,579 unique TCRδ sequences using TRDV1. The most highly expanded sequence, which was present in 76/152 total sequences, shared the CxxxxxPxLGD motif with five other unique CDR3δ sequences across two patients. This motif was rare in reference sequences, occurring in only 50/ 18,579 unique sequences ($P < 0.0001$; Fig. 5 A and C). We also found that the amino acid motif CxxxxxxxxYWGI was highly enriched within TCRDV1+ CDR3δ in αEβ7+CD38+ γδ cells compared with reference TCRDV1+ γδ T-cell sequencing, oc- curring in all three celiac patients at a total frequency of 14/152 unique sequences while only present in 115/18,579 unique reference sequences ($P < 0.0001$; Fig. 5 B and C).

The high clonality of αEβ7+CD38+ T cells, the similarity of TCR repertoire upon a second gluten challenge, and the con- servation of CDR3 motifs in different T-cell clones suggest that both CD8+ αβ and γδ T cells are activated in an antigen-specific manner in response to dietary gluten.

4.4.3 Discussion

In CD, dietary gluten induces the infiltration of T cells in the small intestine and the destruction of intestinal epithelial cells. We find that along with the induction of gluten-specific CD4+ cells, the reintroduction of dietary gluten to celiac patients on a gluten-free diet induces the peripheral appearance of large numbers of activated CD8+ and γδ T cells expressing gut-homing markers. These findings are consistent with the supposition that these T cells are activated and imprinted with gut-homing po- tential in secondary lymphoid organs by dendritic cells presenting gut-derived antigens (32). Like peripheral blood gluten-specific CD4+ T cells, these cells express surface markers consistent with memory or effector cells, indicating that they are programmed as such before gut recruitment. This suggests that at least some of the pathogenic IELs in

A

	patient	freq	CDR3 (AA) (	TRDJ
TRDV 1	1 (152)	76	C A L G G L P T L G D T P - - - T D K L I F tgt gct ctt ggg ggt ctt cct act ctg ggg gat acg ccc --- --- --- acc gat aaa ctg atc ttt	1
		2	C A L A P L P T L G D T G - - - P D K L I F tgt gct ctt gcc cct ctt cct act ctg ggg gat acg ggc --- --- --- ccc gat aaa ctc atc ttt	1
		2	C A L G F P P V L G D P Y - - - T D K L I F tgt gct ctt ggg ttc cct cca gta ctg ggg gat ccg tac --- --- --- acc gat aaa ctc atc ttt	1
		1	C A L G V L P T L G D Q G - - P T D K L I F tgt gct ctt ggg gtt ctt cct acg ctg ggg gat cag ggg --- --- ccc acc gat aaa ctc atc ttt	1
	3 (112)	5	C A L G E L P L L G D T L R S Y T D K L I F tgt gct ctt ggg gag ctc ccc tta ctg ggg gat aca ttg agg tcg tac acc gat aaa ctg atc ttt	1
		2	C A L G A L P T L G D R G - - - V D K L I F tgt gct ctt ggg gcg ctt cct aca ctg ggg gat agg ggg --- --- --- gtc gat aaa ctg atc ttt	1

B

	patient	freq	CDR3 (AA) (	TRDJ
TRDV 1	1 (152)	2	C A L G E R R P S Y W G I R R G P - - - L I F tgt gct ctt ggg gaa cga cgc cca agt tac tgg ggg ata cgg cgt gga ccc --- --- ctg atc ttt	4
		1	C A L G E A R P S Y W G I R T - T D K L I F tgt gct ctt ggg gaa gca agg cct tcc tac tgg ggg ata cg a acg--- acc gat aaa ctc atc ttt	1
		1	C A L G E Y L P R Y W G I H G - T D K L I F tgt gct ctt ggg gaa tat cta ccc cgg tac tgg ggg ata cac ggg --- acc gat aaa ctc atc ttt	1
		1	C A L G E R L P S Y W G I S Y - T D K L I F tgt gct ctt ggg gaa aga ctt ccc tct tac tgg ggg atc tcg tac --- acc gat aaa ctc atc ttt	1
		1	C A L G E L S P R Y W G I G Y - T D K L I F tgt gct ctt ggg gaa ctt tct cca cgg tac tgg ggg ata ggg tac --- acc gat aaa ctc atc ttt	1
		1	C A L G E R F R G Y W G I Q Y - T D K L I F tgt gct ctt ggg gaa cgg ttc cgg ggg tac tgg ggg ata cag tac --- acc gat aaa ctc atc ttt	1
	3 (112)	2	C A L G E L I R G Y W G I R Y - T D K L I F tgt gct ctt ggg gaa cta ata agg ggg tac tgg ggg ata cgg tac --- acc gat aaa ctc atc ttt	1
		1	C A L G E L L P R Y W G I G G - T D K L I F tgt gct ctt ggg gaa ctg ctt ccc cgg tac tgg ggg ata ggt ggt --- acc gat aaa ctc atc ttt	1
		1	C A L G E R G P R Y W G I A Y - T D K L I F tgt gct ctt ggg gaa cgc gga cca cgg tac tgg ggg ata gcg tac --- acc gat aaa ctc atc ttt	1
		1	C A L G E L R R I Y W G I R - - I D K L I F tgt gct ctt ggg gaa cta cgt agg ata tac tgg ggg ata cga --- --- atc gat aaa ctc atc ttt	1
	4 (60)	1	C A L G Q R I P S Y W G I A G S T D K L I F tgt gct ctt ggc cag agg att cct tct tac tgg ggg ata gcg ggt agc acc gat aaa ctc atc ttt	1
		1	C A L G D F L P S Y W G I R G - T D K L I F tgt gct ctt ggg gat ttc cta ccc tcc tac tgg ggg ata agg ggc --- acc gat aaa ctc atc ttt	1
		1	C A L G E I Y R G Y W G I R A G - D K L I F tgt gct ctt ggg gaa att tac cgt ggg tac tgg ggg ata cgc gcc gga --- gat aaa ctc atc ttt	1
		1	C A L G E T T L S Y W G I R Y - T D K L I F tgt gct ctt ggg gaa acc acg ctt tcc tac tgg ggg ata cgg tac --- acc gat aaa ctc atc ttt	1

C

V-gene	CDR3 motif	celiac clones	healthy clones	Significance
TRDV1	CxxxxxPxLGD	6/150	50/18759	p<0.0001
TRDV1	CxxxxxxxxYWGI	14/150	115/18759	p<0.0001

Figure 4.10: Convergent αEβ7+CD38+TCRδ CDR3 motifs are found among clones within the same celiac patient and across different patients following gluten challenge. (A and B) Convergent motifs CxxxxxPxLGD (A) and CxxxxxxxxYWGI (B) are seen in TCRδ clones using TRBV1. The frequency of each clone is in- dicated and the total number of T cells sequenced in the patient is indicated in parentheses. The protein sequence with the corresponding DNA sequence is shown. Within the protein sequence, yellow indicates absolutely conserved amino acids, gray indicates relatively conserved (≥50%) amino acids, and blue indicates conserved amino acids that are encoded within the V or J genes. Within the DNA sequence, nucleotides in yellow are formed through N or P addition, whereas nucleotides in gray are encoded by D genes. (C) Con- vergences of motifs seen in TCRδ clones using TRBV1 are statistically signifi- cant compared with reference control TCRδ sequences.

CD are purposefully activated and recruited to the gut. Importantly, these cells respond with a very focused TCR repertoire, indicating that they are selected in an antigen-specific manner before entering the intestine.

The presence of inflammation has long been postulated to promote the loss of tolerance, and prevailing models of CD pathogenesis propose that IELs are activated as a result of in- flammation that is initiated by gluten-specific CD4+ cells. The inflammatory cytokine IL-15 is up-regulated within celiac intestinal mucosa, and has been implicated in promoting inflammation through diverse means, including impairing regulatory T-cell generation promoting NK-like function of CD8+ IELs, and enabling the expansion of IELs (9, 33). CD8+ IELs have been shown to demonstrate cytotoxicity through stimulation by IL-15 and activation through NK receptors including CD94 and NKG2D (9, 34). Whereas αEβ7+CD38+CD8+ T cells clearly show markers of effector cells and are capable of IFN-γ production, they largely do not express perforin, CD57, or higher levels of NKG2D (Fig. S6). Therefore, it is possible that tissue factors, including IL-15, are further required for cytotoxicity.

The function of γδ IELs is more poorly understood. In human CD, both cytotoxic and anti-inflammatory functions have been attributed to subsets of γδ IELs (35, 36). In mice, γδ IELs appear to be constitutively activated with high cytotoxic potential at baseline (37). However, they express both activating and in- hibitory NK receptors, and it has been suggested that the com- bination of these NK receptors can keep the effector functions of γδ IELs in check but enable them to be readily switched on. Thus, both CD8+ and γδ IEL cell populations may ultimately mediate tissue destruction through NK receptors and require tissue-derived factors. However, we find that αEβ7+CD38+ T cells express markers of differentiated effector cells before gut recruitment, and their appearance parallels the appearance of gluten-reactive CD4+ T cells in blood, rather than occurring later. Also, although increased numbers of IELs and mildly in- creased levels of IL-15 are present in celiac patients on a gluten-free diet (38), the recruitment we describe precedes significant intestinal inflammation and tissue damage, which only reliably occur histologically after 2–4 wk of continuous gluten exposure (39). These findings suggest that IELs in CD are not simply activated as bystanders as a consequence of gut inflammation.

As celiac IEL populations are induced by gluten, a long- standing question has been whether their TCRs recognize gluten. Despite extensive study, gluten-derived peptide epitopes recog- nized by CD8+ T cells in CD have not been apparent, and there is no significant genetic association of CD with particular HLA class I alleles. Therefore, it is generally thought that IELs do not mediate tissue damage through gluten recognition. Nevertheless, one group has identified a class I gluten epitope recognized by CD8+ T cells isolated from CD intestinal mucosa (40). If the αEβ7+CD38+CD8+ T cells we describe are responding to gluten, this would imply a rapid and efficient cross-presentation of gluten on MHC class I. Besides gluten, other possibilities for IEL ligands include self-antigens or infectious pathogens. The possibility of self- antigen recognition is supported by the very selective destruction of intestinal epithelial cells and the presence of autoantibodies, in- cluding antibodies to tissue transglutaminase (10, 41, 42). The role of an infectious cofactor in CD has been proposed based on epidemiologic data showing that neonatal infection seems to predispose individuals to the development of CD (43).

This process through which these three T-cell subsets are synchronously mobilized and recruited to intestinal tissue clearly has implications in immunity to infections. The development of autoimmunity in CD likely represents a misdirected application of processes that are meant to be protective. Due to the well- established dependence of CD on the CD4+ T-cell response, the coordinated T-cell response we describe here presumably depends upon gluten-specific CD4+ T cells. In this context, multiple aspects of the effector CD8+ T-cell responses to viruses have been shown to depend upon CD4+ T-cell help, including the primary effector response, the generation of memory, and re- cruitment to sites of infection (44–47). This process has been termed "licensing," referring to the ability of CD4+ T cells to license cognate effector CD8+ T-cell responses. Here we specu- late that CD4+ T cells may be "licensing" self-antigen–specific CD8+ T cells to become activated and recruited to the intestine, subsequently leading to tissue damage. This process may share mechanisms with the processes that have been described to co- ordinate CD4+ and effector T-cell responses to viruses.

Like CD, most autoimmune diseases with HLA associations are associated with MHC class II alleles, including type 1 di- abetes, multiple sclerosis, rheumatoid arthritis, and ulcerative colitis (48). Despite the association of these diseases with class II alleles rather than class I alleles, CD8+ effector T cells play an important role in the pathogenesis of these diseases. For in- stance, although type 1 diabetes is strongly associated with class II alleles, autoreactive CD8+ T cells are extensively found in inflamed diabetic islets and are appreciated to be the primary effectors driving tissue damage (49–51). Thus, the scenario we outline above for celiac disease may be generalizable to other forms of autoimmunity, in that an initial misdirected CD4+ T-cell response may license effector CD8+ and γδ T cells to cause tissue destruction at a particular site.

The mobilization of specific lymphocytes into the peripheral blood 6 d after antigenic challenge, as has been reported in CD (4, 5) and in the context of influenza vaccination (52), has provided an invaluable window into antigen-specific responses in human subjects. It will be interesting to see whether other such migrations are occurring at specific times in other auto- immune diseases. We also suggest that the analysis of activated T cells with gut-homing markers in the peripheral blood on day 6 after gluten challenge may be a superior method to diagnose CD in individuals currently on a gluten-free diet. An estimated 1.6 million Americans follow a gluten-free diet without an established diagnosis of CD (53). Available tests, including antibody levels and intestinal biopsy results, can be completely normal in CD patients on a gluten-free diet. Consequently, such individuals are often asked to continually eat gluten-containing foods for 2–4 wk before testing (39). This is often intolerable and precludes an accurate diagnosis. Our study shows promise in the reliable clinical diagnosis of CD with only short-term gluten exposure.

4.4.4 Methods

Gluten Challenge. All human sample collection was performed with informed consent under Stanford University Institutional Review Board oversight Volunteers underwent oral gluten challenge as described (4). At time of participation, all volunteers adhered

to a strict gluten-free diet for at least 1 mo. After an initial blood draw, volunteers consumed four slices of white bread per day for 3 consecutive days (days 1, 2, and 3) and returned for a second blood draw on day 6. All celiac patient volunteers had a clinical diagnosis of celiac disease established by small intestinal biopsy in addition to serologic antibody testing. Five of six celiac volunteers were HLA-DQ2.5+. One celiac volunteer was HLA-DQ8+ according to clinical testing. Healthy HLA-DQ2.5+ volunteers were either parents of children with celiac disease or individuals who endorsed gluten intolerance. Patients were tested for HLA- DQ2.5 by PCR (SI Methods). All healthy volunteers had a negative clinical diagnostic workup for celiac disease, and were able to comply with a gluten- free diet for at least 1 mo before participation.

Tetramer Analysis and Flow Cytometry. All FACS experiments were performed on ARIAII or LSRII instruments (Becton Dickinson). Water-soluble MHC–DQ2 molecules with covalently tethered peptides were produced in a baculovirus expression system (54). Two different MHC–DQ2.5 molecules with engi- neered biotinylation sites were produced with tethered deamidated T-cell epitopes of α-gliadin, DQ2-α-I (QLQPFPQPELPY) and DQ2-α-II (PQPELPYPQPE). Proteins were biotinylated, purified, and stored in PBS, 50% (vol/vol) glycerol at −20 °C. Tetramers were prepared by incubating protein with streptavidin– fluorophore conjugates (eBioscience) at a 4:1 molar ratio. Tetramer staining was performed at room temperature for 1 h using 10 mg/mL tetramer. Anti- body clones used for flow cytometry are in SI Methods.

Intestinal Biopsy Preparation. Small intestinal biopsies were obtained with informed consent from celiac patients undergoing endoscopy at Stanford University Hospital and processed as described (55). See SI Methods.

CyTOF Staining and Data Acquisition. CyTOF and data acquisition were performed as described (16) on cryopreserved peripheral blood mononuclear cells or freshly isolated intestinal lymphocytes. See SI Methods.

CyTOF Antibody Labeling. Purified antibodies (lacking carrier proteins) were labeled 100 μg at a time according to instructions provided by DVS Sciences with

heavy metal-preloaded maleimide-coupled MAXPAR chelating poly- mers via Pre-Load Method version 2.1 (16).

Single-Cell Sorting and TCR Sequencing. Single-cell sorting was performed using an ARIAII cell sorter (Becton Dickinson). TCR sequences from single cells were obtained by a series of three nested PCRs as described (24). The full method and TCR sequence analysis are described in SI Methods.

4.4.5 Acknowledgements

ACKNOWLEDGMENTS. We thank members of the M.M.D. and Y.-h.C. laboratory for helpful discussions. We thank Ludvig Sollid for critical reading of the manuscript and helpful suggestions. We are indebted to all volunteers who participated in this study. We thank Jennifer Iscol and the Celiac Community Foundation of Northern California for help with volunteer recruitment. We thank the Human Immune Monitoring Core and the Stanford Shared FACS Facility for the use of equipment. A.H. was supported by a National Institutes of Health (NIH) T32 Gastroenterology Training Grant. E.W.N. was supported by a fellowship through the American Cancer Society. C.K., Y.-h.C., and M.M.D. are funded by NIH grants: DK063158 (C.K.), AI057229-07 (M.M.D.), and AI090019 (M.M.D.). M.M.D. is an Investigator of the Howard Hughes Medical Institute.

4.4.6 References

1. Fallang LE, et al. (2009) Differences in the risk of celiac disease associated with HLA- DQ2.5 or HLA-DQ2.2 are related to sustained gluten antigen presentation. Nat Im- munol 10(10):1096 – 1101.

2. Sollid LM, Qiao SW, Anderson RP, Gianfrani C, Koning F (2012) Nomenclature and listing of celiac disease relevant gluten T-cell epitopes restricted by HLA-DQ mole- cules. Immunogenetics 64(6):455 – 460.

3. Lundin KE, et al. (1993) Gliadin-speci fi c, HLA-DQ(alpha 1*0501,beta 1*0201) restricted T cells isolated from the small intestinal mucosa of celiac disease patients. J Exp Med 178(1):187 – 196.

4. Brottveit M, et al. (2011) Assessing possible celiac disease by an HLA-DQ2-gliadin tetramer test. Am J Gastroenterol 106(7):1318 – 1324.

5. Ráki M, et al. (2007) Tetramer visualization of gut-homing gluten-speci fi c T cells in the peripheral blood of celiac disease patients. Proc Natl Acad Sci USA 104(8):2831 – 2836.

6. de Kauwe AL, et al. (2009) Resistance to celiac disease in humanized HLA-DR3-DQ2- transgenic mice expressing speci fi c anti-gliadin CD4 + T cells. J Immunol 182(12): 7440 – 7450.

7. Du Pré MF, et al. (2011) Tolerance to ingested deamidated gliadin in mice is main- tained by splenic, type 1 regulatory T cells. Gastroenterology 141(2):610 – 620.

8. Abadie V, Discepolo V, Jabri B (2012) Intraepithelial lymphocytes in celiac disease immunopathology. Semin Immunopathol 34(4):551 – 566.

9. Meresse B, et al. (2004) Coordinated induction by IL15 of a TCR-independent NKG2D signaling pathway converts CTL into lymphokine-activated killer cells in celiac disease. Immunity 21(3):357 – 366.

10. Jabri B, Sollid LM (2009) Tissue-mediated control of immunopathology in coeliac disease. Nat Rev Immunol 9(12):858 – 870.

11. Chan JK, et al. (2011) Type II enteropathy-associated T-cell lymphoma: A distinct aggressive lymphoma with frequent γδ T-cell receptor expression. Am J Surg Pathol 35(10):1557 – 1569.

12. Tack GJ, et al. (2012) Origin and immunophenotype of aberrant IEL in RCDII patients. Mol Immunol 50(4):262 – 270.

13. Anderson RP, Degano P, Godkin AJ, Jewell DP, Hill AV (2000) In vivo antigen chal- lenge in celiac disease identi fi es a single transglutaminase-modi fi ed peptide as the dominant A-gliadin T-cell epitope. Nat Med 6(3):337 – 342.

14. du Pré MF, et al. (2011) CD62L(neg)CD38 + expression on circulating CD4 + T cells identi fi es mucosally differentiated cells in protein fed mice and in human celiac dis- ease patients and controls. Am J Gastroenterol 106(6):1147 – 1159.

15. Newell EW, Sigal N, Bendall SC, Nolan GP, Davis MM (2012) Cytometry by time- of- fl ight shows combinatorial cytoki ne expression and virus-speci fi ccellniches within a continuum of CD8 + T cell phenotypes. Immunity 36(1):142 – 152.

16. Gorfu G, Rivera-Nieves J, Ley K (2009) Role of beta7 integrins in intestinal lymphocyte homing and retention. Curr Mol Med 9(7):836 – 850.

17. Dieterich W, et al. (1997) Identi fi cation of tissue transglutaminase as the autoantigen of celiac disease. Nat Med 3(7):797 – 801.

18. Sulkanen S, et al. (1998) Tissue transglutaminase autoantibody enzyme-linked im- munosorbent assay in detecting celiac disease. Gastroenterology 115(6):1322 – 1328.

19. Di Niro R, et al. (2012) High abundance of plasma cells secreting transglutaminase 2-speci fi c IgA autoantibodies with limited somatic hypermutation in celiac disease intestinal lesions. Nat Med 18(3):441 – 445.

20. Marzari R, et al. (2001) Molecular dissection of the tissue transglutaminase autoan- tibody response in celiac disease. J Immunol 166(6):4170 – 4176.

21. Sallusto F, Lenig D, Förster R, Lipp M, Lanzavecchia A (1999) Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. Nature 401(6754):708 – 712.

22. Appay V, et al. (2002) Memory CD8 + T cells vary in differentiation phenotype in different persistent virus infections. Nat Med 8(4):379 – 385.

23. De Rosa SC, et al. (2004) Ontogeny of gamma delta T cells in humans. J Immunol 172(3):1637 – 1645. 24. Su LF, Kidd BA, Han A, Kotzin JJ, Davis MM (2013) Virus-speci fi c CD4(+) memory- phenotype T cells are abundant in unexposed adults. Immunity 38(2):373 – 383.

25. Qiao SW, et al. (2011) Posttranslational modi fi cation of gluten shapes TCR usage in celiac disease. J Immunol 187(6):3064 – 3071.

26. Kedzierska K, Turner SJ, Doherty PC (2004) Conserved T cell receptor usage in primary and recall responses to an immunodominant in fl uenza virus nucleoprotein epitope. Proc Natl Acad Sci USA 101(14):4942 – 4947.

27. Ramakrishnan NS, Grunewald J, Janson CH, Wigzell H (1992) Nearly identical T-cell receptor V-gene usage at birth in two cohorts of distinctly different ethnic

origin: In fl uence of environment in the fi nal maturation in the adult. Scand J Immunol 36(1): 71 – 78.

28. Kjer-Nielsen L, et al. (2003) A structural basis for the selection of dominant alphabeta T cell receptors in antiviral immunity. Immunity 18(1):53 – 64.

29. Garboczi DN, et al. (1996) Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. Nature 384(6605):134 – 141.

30. Warren RL, et al. (2011) Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured rep- ertoire size of at least 1 million clonotypes. Genome Res 21(5):790 – 797.

31. Chowers Y, Holtmeier W, Harwood J, Morzycka-Wroblewska E, Kagnoff MF (1994) The V delta 1 T cell receptor repertoire in human small intestine and colon. J Exp Med 180(1):183 – 190.

32. Sigmundsdottir H, Butcher EC (2008) Environmental cues, dendritic cells and the programming of tissue-selective lymphocyte traf fi cking. Nat Immunol 9(9):981 – 987.

33. DePaolo RW, et al. (2011) Co-adjuvant effects of retinoic acid and IL-15 induce in- fl ammatory immunity to dietary antigens. Nature 471(7337):220 – 224.

34. Meresse B, et al. (2006) Reprogramming of CTLs into natural killer-like cells in celiac disease. J Exp Med 203(5):1343 – 1355.

35. Jabri B, et al. (2000) Selective expansion of intraepithelial lymphocytes expressing the HLA-E-speci fi c natural killer receptor CD94 in celiac disease. Gastroenterology 118(5): 867 – 879.

36. Bhagat G, et al. (2008) Small intestinal CD8 + TCRgammadelta + NKG2A + intraepithelial lymphocytes have attributes of regulatory cells in patients with celiac disease. J Clin Invest 118(1):281 – 293.

37. Fahrer AM, et al. (2001) Attributes of gammadelta intraepithelial lymphocytes as suggested by their transcriptional pro fi le. Proc Natl Acad Sci USA 98(18):10261 – 10266.

38. Di Sabatino A, et al. (2006) Epithelium derived interleukin 15 regulates intraepithelial lymphocyte Th1 cytokine production, cytotoxicity, and survival in coeliac disease. Gut 55(4):469 – 477.

39. Lef fl er D, et al. (2013) Kinetics of the histological, serological and symptomatic re- sponses to gluten challenge in adults with coeliac disease. Gut 62(7):996 – 1004.

40. Mazzarella G, et al. (2008) Gliadin activates HLA class I-restricted CD8 + T cells in celiac disease intestinal mucosa and induces the enterocyte apoptosis. Gastroenterology 134(4):1017 – 1027.

41. Meresse B, Malamut G, Cerf-Bensussan N (2012) Celiac disease: An immunological jigsaw. Immunity 36(6):907 – 919.

42. Sollid LM, Jabri B (2013) Triggers and drivers of autoimmunity: Lessons from coeliac disease. Nat Rev Immunol 13(4):294 – 302.

43. Sandberg-Bennich S, Dahlquist G, Källén B (2002) Coeliac disease is associated with intrauterine growth and neonatal infections. Acta Paediatr 91(1):30 – 33.

44. Nakanishi Y, Lu B, Gerard C, Iwasaki A (2009) CD8(+) T lymphocyte mobilization to virus-infected tissue requires CD4(+) T-cell help. Nature 462(7272):510 – 513.

45. Janssen EM, et al. (2003) CD4 + T cells are required for secondary expansion and memory in CD8 + T lymphocytes. Nature 421(6925):852 – 856.

46. Shedlock DJ, Shen H (2003) Requirement for CD4 T cell help in generating functional CD8 T cell memory. Science 300(5617):337 – 339.

47. Sun JC, Bevan MJ (2003) Defective CD8 T cell memory following acute infection without CD4 T cell help. Science 300(5617):339 – 342.

48. Trowsdale J (2011) The MHC, disease and selection. Immunol Lett 137(1-2):1 – 8.

49. Coppieters KT, et al. (2012) Demonstration of islet-autoreactive CD8 T cells in insulitic lesions from recent onset and long-term type 1 diabetes patients. J Exp Med 209(1): 51 – 60.

50. Wang B, Gonzalez A, Benoist C, Mathis D (1996) The role of CD8 + T cells in the ini- tiation of insulin-dependent diabetes mellitus. Eur J Immunol 26(8):1762 – 1769.

51. Wong FS, Visintin I, Wen L, Flavell RA, Janeway CA, Jr. (1996) CD8 T cell clones from young nonobese diabetic (NOD) islets can transfer rapid onset of diabetes in NOD mice in the absence of CD4 cells. J Exp Med 183(1):67 – 76.

52. Wrammert J, et al. (2008) Rapid cloning of high-af fi nity human monoclonal anti- bodies against in fl uenza virus. Nature 453(7195):667 – 671.

53. Rubio-Tapia A, Ludvigsson JF, Brantner TL, Murray JA, Everhart JE (2012) The preva- lence of celiac disease in the United States. Am J Gastroenterol 107(10):1538 – 1544.

54. Quarsten H, et al. (2001) Staining of celiac disease-relevant T cells by peptide-DQ2 multimers. J Immunol 167(9):4861 – 4868.

55. Shacklett BL, Critch fi eld JW, Lemongello D (2009) Isolating mucosal lymphocytes from biopsy tissue for cellular immunology assays. Methods Mol Biol 485:347 – 356

4.4.7 Copyright

4.5 IgE allergen-specific memory storage during immunotherapy

Specific immunotherapy (SIT) is the only treatment with proved long-term curative potential in patients with allergic disease. Allergen-specific IgE is the causative agent of allergic disease, and antibodies contribute to SIT, but the effects of SIT on aeroallergen-specific B-cell repertoires are not well understood. Objective: We sought to characterize the IgE sequences expressed by allergen-specific B cells and track the fate of these B-cell clones during SIT. Methods: We used high-throughput antibody gene sequencing and identification of allergen-specific IgE with combinatorial antibody fragment library technology to analyze immunoglobulin repertoires of

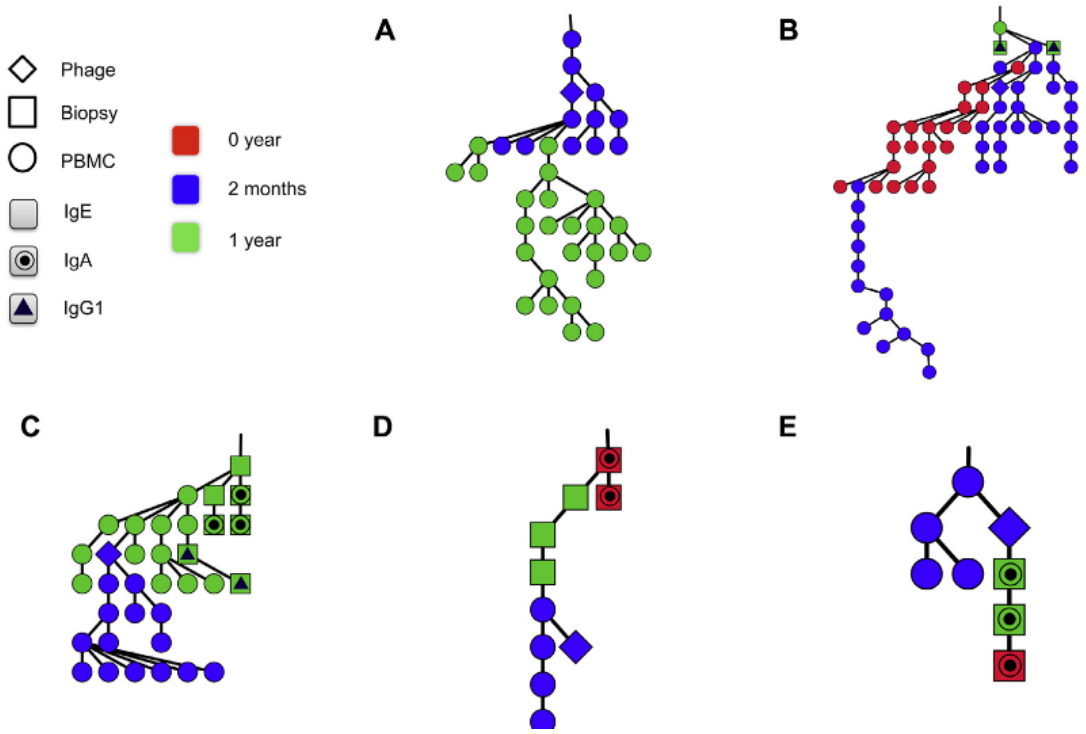


Figure 4.11: Antibody heavy chain somatic mutation trees describing putative clonal evolution relationships between members of allergen-specific B-cell clones. Clone sets related to scFvs IT5-rD13 (A) , IT6-nD16 (B) , IT6-rD128 (C) , IT6-nD17 (D) , and IT6-rD111 (E) are shown. IgE sequences identified as allergen specific in phage display experiments are represented by diamonds , whereas clone members identified by means of deep sequencing of nasal biopsy specimens are shown as rectangles and clone members identified by means of deep sequencing of PBMC specimens are displayed as circles . Colors indicate the time point at which sequences were identified: red , 0 months; blue , 2 months; and green , 1 year of SIT. Symbols indicate the isotype expressed by the clone member. Nodes with no internal symbol represent IgE clone members, nodes with triangles represent IgG 1 , and nodes with a circled dot represent IgA members. Additional exam- ples of trees are shown in Fig E8 .

blood and the nasal mucosa from ae roallergen-sensitized subjects before and during the first year of subcutaneous SIT. Results: Of 52 distinct allergen-specific IgE heavy chains from 8 allergic donors, 37 were also detected by using high-throughput antibody gene sequencing of blood samples, nasal mucosal samples, or both. The allergen-specific clones had increased persistence, higher likelihood of belonging to clones expressing other switched isotypes, and possibly larger clone size than the rest of the IgE repertoire. Clone members in nasal tissue showed close mutational relationships. Conclusion: In the future, combining functional binding studies, deep antibody repertoire sequencing, and information on clinical outcomes in larger studies might aid assessment of SIT mechanisms and efficacy. (J Allergy Clin Immunol 2016;137:1535-44.)

4.5.1 Copyright

4.6 High-fat diet insulin resistance induces repertoire changes

The development of obesity-associated insulin resistance is associated with B-lymphocyte accumulation in visceral adipose tissue (VAT) and is prevented by B-cell ablation. To characterize potentially pathogenic B-cell repertoires in this disorder, we performed high-throughput immunoglobulin (Ig) sequencing from multiple tissues of mice fed high-fat diet (HFD) and regular diet (RD). HFD significantly changed the biochemical properties of Ig heavy-chain complementarity-determining region-3 (CDRH3) sequences, selecting for IgA antibodies with shorter and more hydrophobic CDRH3 in multiple tissues. A set of convergent antibodies of highly similar sequences found in the VAT of HFD mice but not RD mice showed significant somatic mutation, suggesting a response shared between mice to a common antigen or antigens. These

findings indicate that a simple high-fat dietary intervention has a major impact on mouse B-cell repertoires, particularly in adipose tissues.

4.6.1 Copyright

4.7 Affinity maturation targets CDR3 then other CDRs

Affinity maturation occurs through two selection processes: the choice of appropriate clones (clonal selection), and the internal evolution within clones, induced by somatic hyper-mutations, where high affinity mutants are selected for. When a final population of immunoglobulin sequences is observed, the genetic composition of this population is affected by a combination of these two processes. Different immune induced diseases can result from the failure of regulation of clonal selection or of the regulation of the within clone affinity maturation. In order to understand each of these processes separately, we propose a mixed lineage tree/sequence based method to detect within clone selection as defined by the effect of mutations on the average number of offspring. Specifically, we measure the imbalance in the number of leaves in lineage trees branches following synonymous and non-synonymous (NS) mutations. If a mutation is positively selected, we expect the number of leaves in the sub-tree below this mutation to be larger than in the parallel sub-tree without the mutation. The ratio between the number of leaves in such branches following NS mutations can be used to measure selection within a clone. We apply this method to the sampled Ig repertoire from multiple healthy volunteers and show that within clone selection is positive in the CDR2 region and either positive or negative in the CDR3 and FWR3 regions. Selection occurs already at the IgM isotype level mainly in the DH gene region, with a strong negative selection in the join region. This is followed in the later memory stages in the CDR2 region. We have not studied here the FWR1 and CDR1

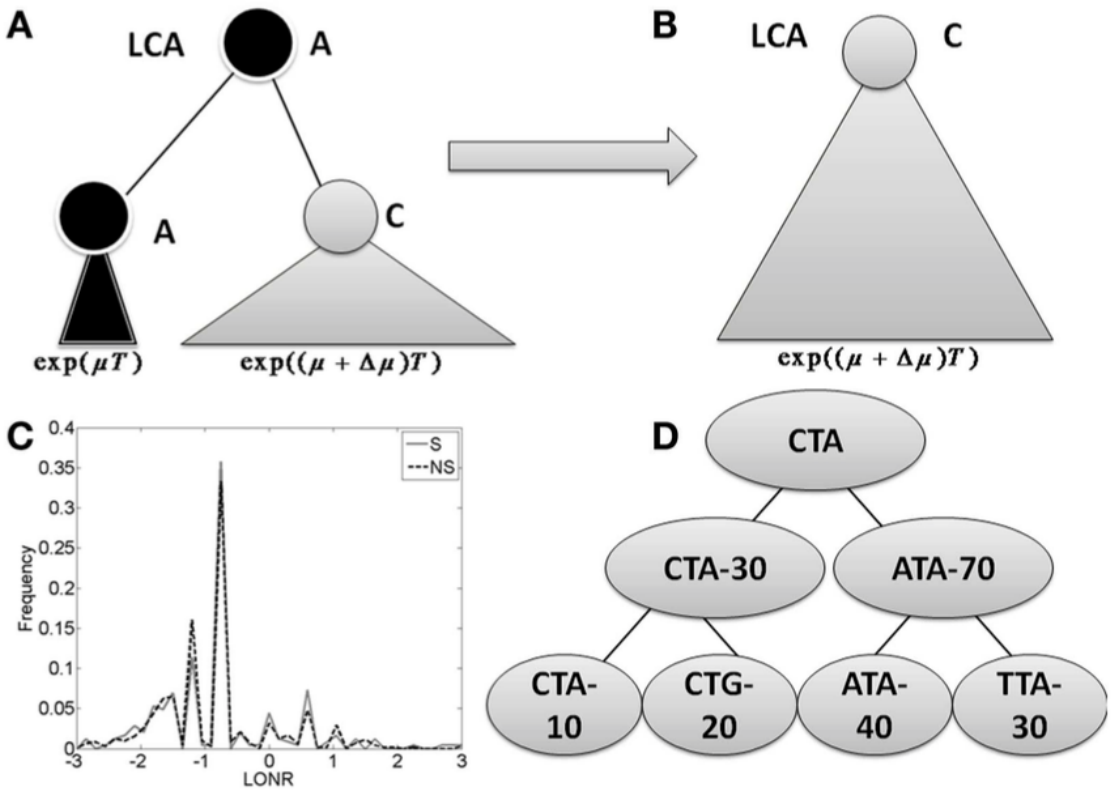


Figure 4.12: The branch imbalance framework and examples. (A) Schematic view of a branch corresponding to a mutation event. Following a mutation, the population can be expanded (or reduced), the advantage will lead to an exponentially growing difference in the number of offspring in parallel branches descending from the same internal origin. T is the time from the mutation to the sampling time. (B) After some time, one branch will take over the entire sample, and the information carried in the ratio between the branches will be lost. (C) LONR values histogram for one simulated sequence pool, simulated under naive multiplication from unique ancestral sequence. While the average is not 0, there is no difference between branches following S and NS mutations. (D) Example of tree. In the left branch, a mutation occurred from CTA to CTG, and the ratio between the mutated and un-mutated branches number of offspring is 20/10. In the right branch, a mutation from ATA to TTA occurred, with a ratio of 30/40. In the root, a mutation from CTA to ATA occurred with a ratio of 70/30.

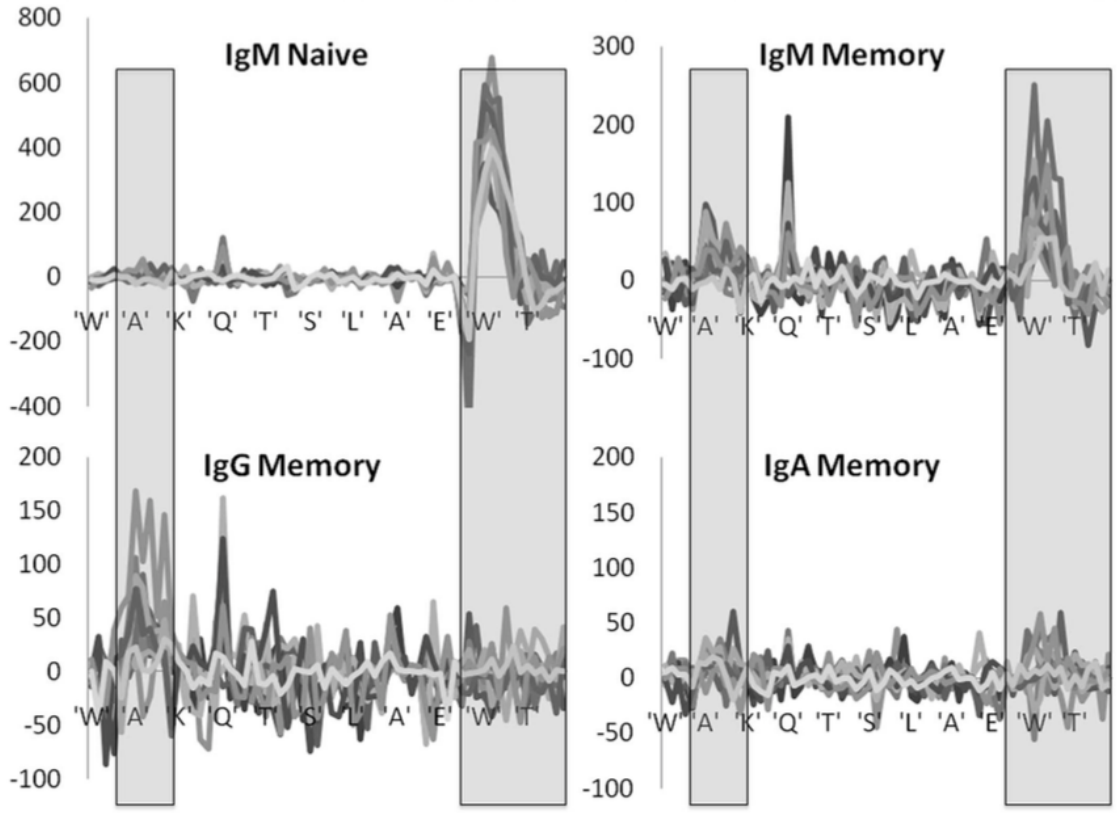


Figure 4.13: Position specific effects. Total LONR score per amino acid (averaged over the 3-nt composing the amino acid). The LONR is drawn for the four isotypes discussed in previous figures: naïve IgM, memory IgM, IgG, and IgA. Each line represents a different donor. The highlighted regions represent the CDR2 (to the left) and CDR3 (to the right). The sequence at the bottom is a typical sequence. The zone of very strong negative selection at the beginning of the CDR3 is the junction region. One can observe a switch from a very strong positive selection in the naïve IgM isotype focused on the CDR3 to a much weaker selection in the IgG and IgA focused on the CDR2. Note that the scales of the y axes are different between the plots.

regions. An important advantage of this method is that it is very weakly affected by the baseline mutation model or by sampling biases, as are most synonymous to NS mutations ratio based methods.

4.8 Amino acid content restricts Dh frame usage

The Ab repertoire is not uniform. Some variable, diversity, and joining genes are used more frequently than others. Nonuniform usage can result from the rearrangement process, or from selection. To study how the Ab repertoire is selected, we analyzed one part of diversity generation that cannot be driven by the rearrangement mechanism: the reading frame usage of D H genes. We have used two high-throughput sequencing methodologies, multiple subjects and advanced algorithms to measure the D H reading frame usage in the human Ab repertoire. In most D H genes, a single reading frame is used predominantly, and inverted reading frames are practically never observed. The choice of a single D H reading frame is not limited to a single position of the D H gene. Rather, each D H gene participates in rearrangements of differing CDR3 lengths, restricted to multiples of three. In nonproductive rearrangements, there is practically no reading frame bias, but there is still a striking absence of inversions. Biases in D H reading frame usage are more pronounced, but also exhibit greater interindividual variation, in IgG + and IgA + than in IgM + B cells. These results suggest that there are two developmental checkpoints of D H reading frame selection. The first occurs during VDJ recombi- nation, when inverted D H genes are usually avoided. The second checkpoint occurs after rearrangement, once the BCR is expressed. The second checkpoint implies that D H reading frames are subjected to differential selection. Following these checkpoints, clonal selection induces a host-specific D H reading frame usage bias. The Journal of Immunology , 2013, 190: 5567–5577.

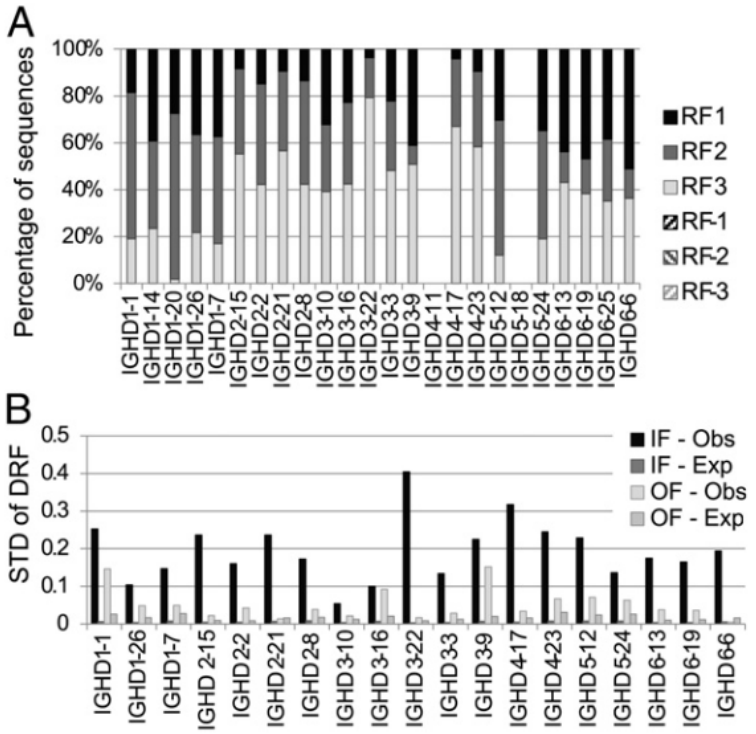


Figure 4.14: DRF usage in the Illumina data. (A) Fraction of sequences using each DRF in each D H . Each column is a D H gene. Each shade is a DRF. One can clearly see the similar DRF usage as observed in Fig. 3. Differences may arise due to the fact that different donors were recruited for each sequencing. (B) SD of DRF usage in IF and OF sequences. Each column is the SD of the DRF usage in a given gene, as computed in Fig. 4B, using the Illumina-based sequences. The expected SD was computed as 1 over the square root of the sample size. The OF RF SD is only slightly higher than the one expected by the size of the sample, because the sample size for OF rearrangements was limited. The IF SD of DRF usage approaches the maximal possible in this case [0.57 = root of (1/3)], and much more than expected by the sample size. One can thus conclude that most of the variance in the DRF usage is determined by selection. D H genes for which there were not enough OF or IF sequences (, 100) were not incorporated in the analysis.

4.8.1 Copyright

4.9 Detecting selection by branch imbalance in lineage trees

Incremental selection within a population, defined as limited fitness changes following mutation, is an important aspect of many evolutionary processes. Strongly advantageous or deleterious mutations are detected using the synonymous to non-synonymous mutations ratio. However, there are currently no precise methods to estimate incremental selection. We here provide for the first time such a detailed method and show its precision in multiple cases of micro-evolution. The proposed method is a novel mixed lineage tree/sequence based method to detect within population selection as defined by the effect of mutations on the average number of offspring. Specifically, we propose to measure the log of the ratio between the number of leaves in lineage trees branches following synonymous and non-synonymous mutations. The method requires a high enough number of sequences, and a large enough number of independent mutations. It assumes that all mutations are independent events. It does not require of a baseline model and is practically not affected by sampling biases. We show the method's wide applicability by testing it on multiple cases of micro-evolution. We show that it can detect genes and inter-genic regions using the selection rate and detect selection pressures in viral proteins and in the immune response to pathogens.

The detection of selection is a crucial issue in population bi- ology, evolution theory and ecology. It also has important clinical implications. While multiple sequence based meth- ods have been proposed to detect selection (15,64–69), most of them are focused on strongly advantageous or deleterious mutations. We have here proposed a method best adapted to the detection of slightly advantageous or deleterious mutations in micro-evolution.

The basic concept behind the here reported LONR mea- sure is to test for the systematic increase of the population size following non-synonymous mutations in a given region. An advantage of the LONR is that each mutation is counted once independently of the total number of sequences that end up containing this mutation. Thus, it is practically un- affected by sampling biases or by the expansion of specific sub-populations.

While multiple tree shape based methods were developed (18–21), these methods often cannot detect the direction of selection, and cannot detect which region in the sequence is selected. Moreover, many of these tree shapes are sensi- tive to sampling effects making them impractical to use in realistic situations (70).

We have here proposed a new method that can clearly detect positive and negative selection or their combination, based on the effect each mutation has on the number of off- spring in the tree under the branch where the mutation has occurred. This method can only be applied where the muta- tion rate is high enough, and the selection is weak enough for alleles with disadvantageous mutations to exist in the population. Specifically, the mutation rate multiplied by the fixation time of mutations should be much larger than one. The code and a short manual are supplied in the Supplementary Materials.

Such a range exists for example in the population dynam- ics of mitochondria within host species, in viral dynamics and in the affinity maturation process in germinal centers. We have here studied all these cases and have shown that in- deed selection can be detected in all cases studied. Other ap- plications of this method can be the evolution of the Y chro- mosome and the changes in Short Tandem Repeats (STR) frequencies in it or the evolution of bacteria in an infection in the population.

This method is precise in the domain of a large number of mutations per sequence (>10) and large samples (>300). In this domain, the method proved to have many important applications, such as the detection of selection in genes (and actually the direct detection of genes), the detection of viral proteins passing positive and negative selection and under- standing the selection process in a B cell immune response. We have shown that while the ribosomal RNA has a very strong positive selection, some genes pass positive selection, and others negative selection. In B cells, we have shown

that while CDR mutations are always positively selected, FWR mutations are selected against in the majority of the popula- tions, but actually strongly positively selected in the extreme cases.

The comparison between S and NS mutations is only the most basic distinction between different types of muta- tions. Other possibilities exist, especially change/no-change of some amino-acid property, such as size or hydropho- bicity. Such methods would test for selection for specific changes and not selection for mutation in general. In other words, the proposed methodology can be used to estimate whether changes in a given property increase or decrease the number of offspring, compared with a change not altering this properties. In other words, different definition of the mutations of interest and the baseline can be defined, and used to detect selection for other features.

The main limitation of the current score is that it is blind to strong selection. Once a mutation is fixed in (or com- pletely removed from) the population, we will not observe the polymorphism at this site that allows us to compare the branch sizes.

4.9.1 Copyright

4.10 Comparative analysis of the mammalian IgH repertoires

Next-Generation Sequencing combined with bioinformatics is a powerful tool for analyzing the large number of DNA sequences present in the expressed antibody repertoire and these data sets can be used to advance a number of research areas including antibody discovery and engineering. The accurate measurement of the immune repertoire sequence composition, diversity and abundance is important for understanding the repertoire response in infections, vaccinations and cancer immunology and could also be useful for elucidating novel molecular targets. In this study 4 individual domestic cats (Felis catus) were subjected to antibody repertoire sequencing with total

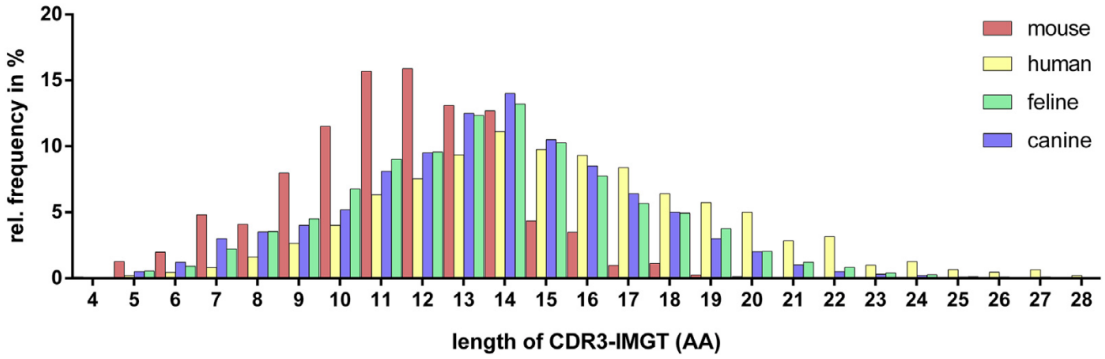


Figure 4.15: In MS, B cell receptors are subject to extensive intrathecal SHM. Nucleotide sequences, represented by clusters shown in Figure 1, were selected from the IgG-VH sequence database and used to generate lineage trees using IgTree software (see Methods). (A–E) Representative trees for CSF-restricted clusters for patients (A) MS-1, (B) MS-3, (C) MS-4, (D) MS-6, and (E) MS-5. The corresponding IGHV, IGHJ, and most com- mon H-CDR3 AA sequences are listed in Supplemental Table 2. In the lineage trees, each round node represents at least one unique IgG-VH sequence ranging from at least the 5′ end of H-CDR1 to the 3′ end of H-CDR3; larger nodes represent up to hundreds of identical sequences. Putative germline sequences were determined using SoDA (https://dulci.org/soda/; ref. 36) and are labeled as black, and hypothetical intermedi- ates calculated by IgTree are labeled as beige. The numbers represent mutational steps between nodes; only mutational steps >1 are indicated; thus, unlabeled branches represent a single mutation. Triangular nodes contain 2 or more singleton sequences in leaves.

number of sequences generated 1079863 for VH for IgG, 1050824 VH for IgM, 569518 for VK and 450195 for VL. Our analysis suggests that a similar VDJ expression patterns exists across all cats. Similar to the canine repertoire, the feline repertoire is dominated by a single subgroup, namely VH3. The antibody paratope of felines showed similar amino acid variation when compared to human, mouse and canine counterparts. All animals show a similarly skewed VH CDR-H3 profile and, when compared to canine, human and mouse, distinct differences are observed. Our study represents the first attempt to characterize sequence diversity in the expressed feline antibody repertoire and this demonstrates the utility of using NGS to elucidate entire antibody repertoires from individual animals. These data provide significant insight into understanding the feline immune system function.

4.10.1 Copyright

Chapter 5

Applications in engineering synthetic repertoires

5.1 Introduction

In Chapter 1-4, multiple observational studies were presented that provided insights into the function of the adaptive repertoires in-vivo. In the process, the data informs new theories for what features cause repertoires to be more or less effective and their core functions of affinity and specificity. A powerful test environment of such theories are synthetic repertoires. Interventional and mechanistic, the generation of an in-vitro repertoire by synthetic means provides a powerful means of evaluating specific selection effects on a per amino acid per position basis, without the constraints of the natural repertoire structure and many counfounding processes of positive and negative selection that occur in-vivo.

5.2 Synthetic repertoires with unbiased landscapes

In this study, we generated a synthetic library where each position of the antibody CDR-H3 loop had an equal probability of each amino acid. The library provided a unique opportunity to ask what the effects of amino acid variability would be on the capacity for antibodies to fold and be selected against antigen. At heart, the libraries

help answer to what degree the amino acid variability that we observe in nature are present as a consequence of genetic constraints of V(D)J recombination and SHM, or rather biases at the amino acid level affecting the ability of the molecules to fold.

We have generated large libraries of single-chain Fv antibody fragments (N10^10 transformants) containing unbiased amino acid diversity that is restricted to the central combining site of the stable, well-expressed DP47 and DPK22 germline V-genes. Library WySH2A was constructed to examine the potential for synthetic complementarity-determining region (CDR)-H3 diversity to act as the lone source of binding specificity. Library WySH2B was constructed to assess the necessity for diversification in both the H3 and L3. Both libraries provided diverse, specific antibodies, yielding a total of 243 unique hits against 7 different targets, but WySH2B produced fewer hits than WySH2A when selected in parallel. WySH2A also consistently produced hits of similar quality to WySH2B, demonstrating that the diversification of the CDR-L3 reduces library fitness. Despite the absence of deliberate bias in the library design, CDR length was strongly associated with the number of hits produced, leading to a functional loop length distribution profile that mimics the biases observed in the natural repertoire. A similar trend was also observed for the CDR-L3. After target selections, several key amino acids were enriched in the CDR-H3 (e.g., small and aromatic residues) while others were reduced (e.g., strongly charged residues) in a manner that was specific to position, preferentially occurred in CDR-H3 stem positions, and tended towards residues associated with loop stabilization. As proof of principle for the WySH2 libraries to produce viable lead candidate antibodies, 114 unique hits were produced against Delta-like ligand 4 (DLL4). Leads exhibited nanomolar binding affinities, highly specific staining of DLL4+ cells, and biochemical neutralization of DLL4–NOTCH1 interaction.

5.2.1 References

Mahon, M Lamburt, J Glanville, et al. “Comprehensive interrogation of a minimalist synthetic CDR-H3 library and its ability to generate antibodies with therapeutic potential.” Journal of Molecular Biology (2013).

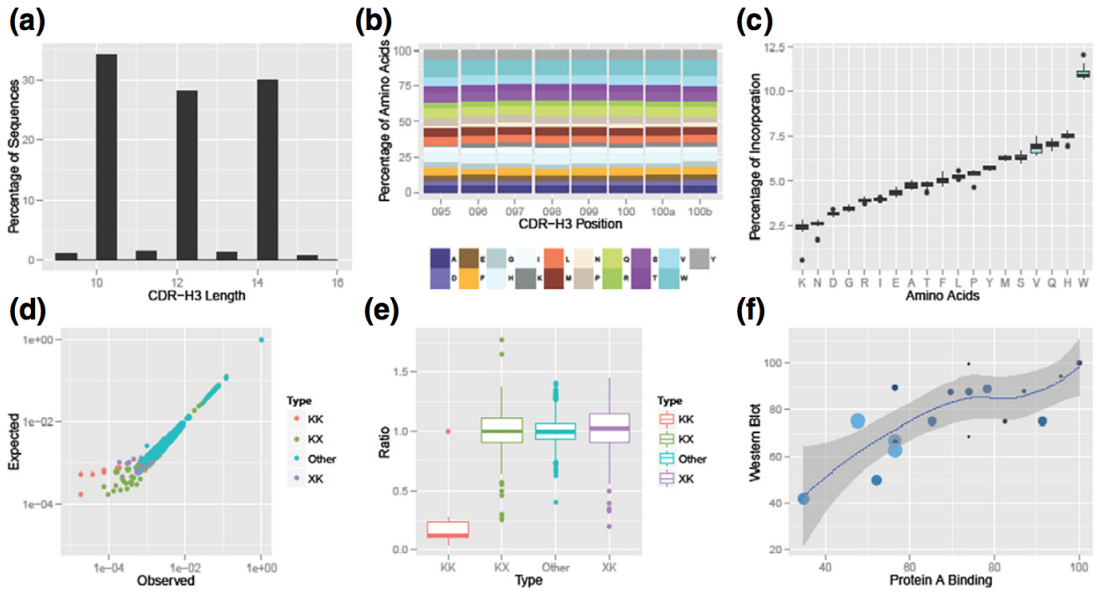


Figure 5.1: Quality control analyses for sub-libraries WySH2A and WySH2B. (a) Length distribution observed in CDR-H3s of lengths 10, 12, and 14. While the majority of each library falls into the correct length bin, pyrosequencing of 465,242 clones in total identified many examples with incorrect amino acid length in CDR-H3. This length mismatch involves both addition and omission of codons, suggesting that the fidelity of codon incorporation is imperfect. (b) Amino acid incorporation profile in clones from CDR-H3 length 10 (only the 8 randomized positions shown, for clarity). Amino acid biases were very similar in every randomized position. Profiles showed no substantial differences between library lengths (other than the length differences). (c) Amino acid incorporation trends for CDR-H3 lengths 10, 12, and 14. While equal representation of all amino acids (excluding cysteine) was expected at each diversified position, some clear deviations are observed. (d) Analysis of design fidelity. Plotting percentage observed against percentage expected shows a strong correlation (r2 = 0.992), suggesting that the amino acid distributions found at each position in all sequences are highly reproducible. In this plot, lysine (K) codons were found to be significant outliers, while all others were very uniform. (e) Plotting the ratio of observed versus expected for lysine-containing codon pairs in comparison to all others shows that KK pairs (or AAA-AAA in DNA sequence) suffer from particularly poor incorporation rates. (f) A random selection of between 16 and 24 scFvs from each of the 12 WySH2A and 6 WySH2B sub-libraries (360 clones in total) were tested for expression in Western blot and also for binding to Protein A in ELISA. Each point represents one sub-library and is plotted according to percentage performance in each analysis. Points are colored according to CDR-H3 length, running from shortest ( ) to longest ( ). In comparison of the two data sets, r2=0.5939, suggesting a correlation between scFv expression rate and functionally folded library content. For both WySH2A and B libraries, increased CDR-H3 length leads to a decreased frequency of expressed clones.

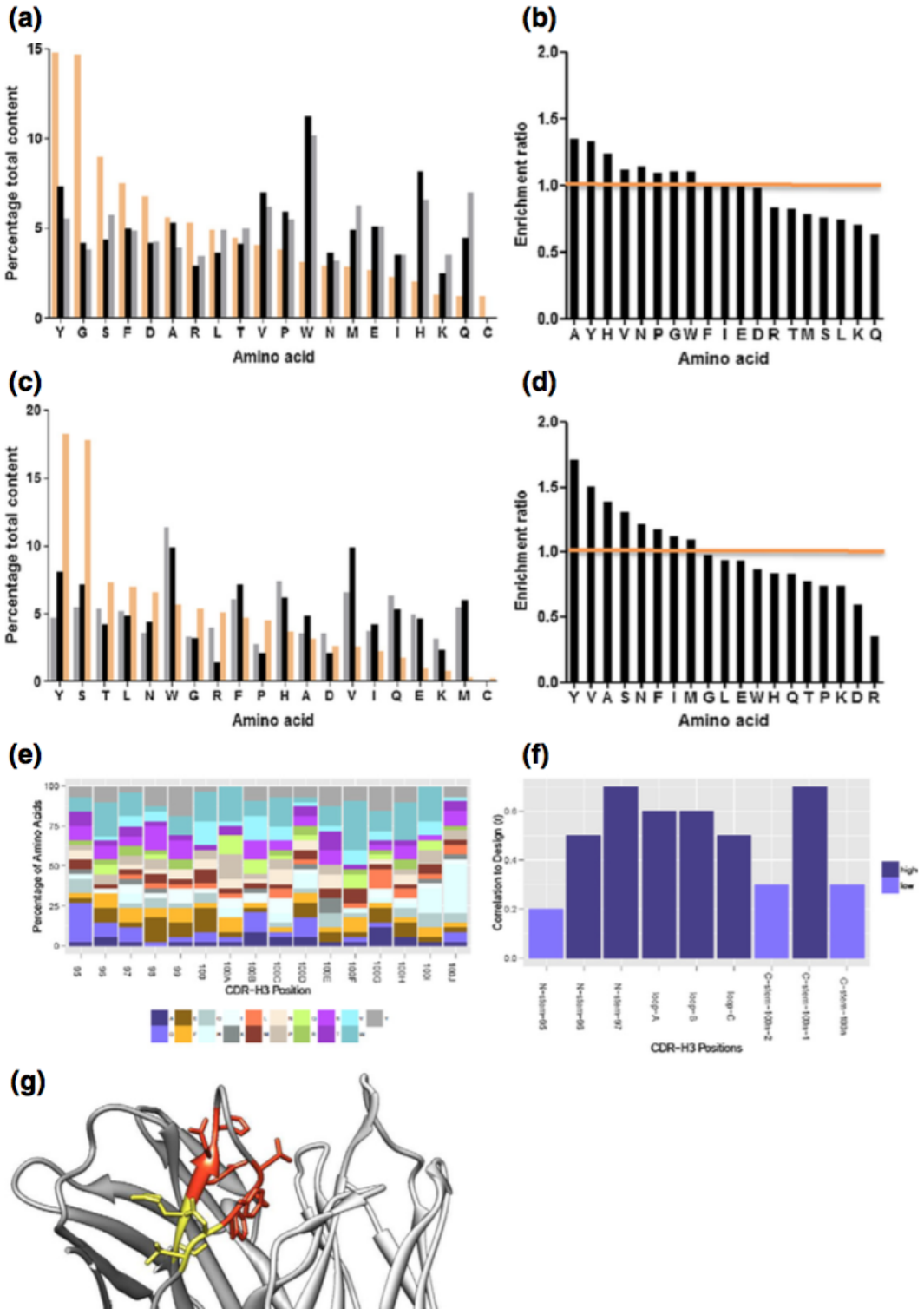


Figure 5.2: Analyses of percentage amino acid content in CDRs of unselected WySH2 clones, WySH2 hits, and naturally occurring sequences. (a) A total of 407 unselected WySH2 CDR-H3 sequences ( ) are compared to sequences of 243 hits from WySH2 (■) and 9242 CDR-H3 sequences found in natural human repertoire databases( ). Amino acids are arranged in descending order from left to right, corresponding to their frequency in the natural human repertoire. (b) Amino acid ER for

5.3 Synthetic repertoires with natural sequence landscapes

We present a method for synthetic antibody library generation that combines the use of high-throughput immune repertoire analysis and a novel synthetic technology. The library design recapitulates positional amino acid frequencies observed in natural antibody repertoires. V-segment diversity in four heavy (VH) and two kappa (V) germlines was introduced based on the analysis of somatically hypermutated donor-derived repertoires. Complementarity- determining region 3 length and amino acid designs were based on aggregate frequencies of all VH and V sequences in the data set. The designed libraries were constructed through an adaptation of a novel gene synthesis technology that enables precise positional control of amino acid composition and incorporation frequencies. High-throughput pyrosequencing was used to monitor the fidelity of construction and characterize genetic diversity in the final 3.6 × 10ˆ10 transformants. The library exhibited Fab expression superior to currently reported synthetic approaches of equivalent diversity, with greater than 93% of clones observed to successfully display both a correctly folded heavy chain and a correctly folded light chain. Genetic diversity in the library was high, with 95% of 7.0×105 clones sequenced observed only once. The obtained library diversity explores a comparable sequence space as the donor- derived natural repertoire and, at the same time, is able to access novel recombined diversity due to lack of segmental linkage. The successful isolation of low- and subnanomolar-affinity antibodies against a diverse panel of receptors, growth factors, enzymes, antigens from infectious reagents, and peptides confirms the functional viability of the design strategy.

5.3.1 Introduction

Antibodies represent the largest class of biother- apeutic molecules in the clinic. Their ability to impart high specificity and long half-life has made them a very attractive therapeutic modality for targeting extracellular ligands and receptors. Tradi- tional methods of generating monoclonal leads by hybridoma technology are increasingly

replaced by in vivo (transgenic mice) and in vitro (yeast and phage display) methods that generate human leads.1,2 In vitro methods employ libraries generated from donor-derived B cells, synthetically derived diversity, or semisynthetic approaches that incorpo- rate diversity from a combination of these two approaches.

The most common display method for these libraries relies on surface display of filamentous phage3 because of the ability to generate fairly large libraries (10^7–10^11). The number of unique anti- bodies displayed by phage is limited by the transformation efficiency of Escherichia coli, a re- quirement for the library generation process. In contrast to natural repertoires, the size of an in vitro library is fixed after generation. Therefore, it is widely accepted that large libraries yield more specific binding solutions,4 and the trend has been towards the development of larger libraries: libraries of 10^11 transformants are not uncommon.5

Given the finite number of clones within a library, the goal of library design should be to maximize functional diversity in the antibody population. Library diversity is often defined as the total number of unique antibodies encoding different amino acid combinations in6their complementarity-determining regions (CDRs). Functional diversity considers only the subset of these unique antibodies that are able to fold appropriately and display a stable heterodimer capable of recognizing antigen in the context of the display system being utilized. As it is possible to generate genetic diversity that cannot produce a properly folded antibody, there needs to be an effort to maximize the overlap between genetic and functional diversity during in vitro repertoire design. Traditional library designs have balanced these needs by obtaining diversity from two fundamen- tally different sources: natural diversity or synthetic diversity.

Natural antibody repertoires benefit from high functional diversity but exhibit constrained genetic diversity. A high proportion of well-folded proteins capable of recognizing antigens are selected for during B-cell maturation and during positive, negative and antigen-driven selection. Biased amino acid usage in the natural repertoires7 has been demonstrated to directly impact the success of functional antibody

response.8–10 However, natural clonal proliferation results in many redundant high-frequency clones,11 and the V(D)J (V, variable gene segment; D, diversity gene segment; J, joining gene segment) diversification mechanisms produce cor- relations between adjacent amino acids that con- strain accessible genetic diversity.12

For exploring diversity beyond that available from the natural repertoire, synthetic methods have been developed to control framework usage and intro- duce artificial CDR variation not observed in nature. Two fundamentally distinct methodologies have been developed to insert synthetic codon diversity into antibody CDR positions. The first employs degenerate nucleotide codons,13–17 and the second uses trinucleotide phosphoramidite (TRIM)-based oligonucleotide synthesis.18,19 Both produce librar- ies with variable positions diversified according to the composition of a limited number of amino acid cocktails. In general, synthetic antibody library design approaches allow for additional control over framework selection and access to nonnatural CDR diversity but suffer from common errors in molecular assembly and difficulty exerting fine control over amino acid composition. Without natural selection mechanisms to screen for function- al diversity, synthetic library designs are vulnerable to generating nonfunctional antibodies.20 Synthetic designs limiting diversity to a single CDR have succeeded in demonstrating expression of up to 90% of clones.17 However, published synthetic libraries with diversity in multiple CDRs have reported much lower proportions, with between 23% and 39% of unproductive clones.15,18,21 Screening methods have been proven effective in reducing stop codons and frameshifts in synthetic repertoires but add additional library preparation steps.16,21

From the earliest synthetic libraries, it has been observed that biases in the natural repertoire are selected for in the functional binders recovered from synthetic repertoires.9,13,22 Many of these biases are considered to represent functional elements that contribute to folding and stability.9,15,23,24 Including known biases from the natural antibody repertoires has been demonstrated to contribute to successful synthetic designs.9,15 A number of design strategies have aimed to balance the benefits of genetic and functional diversity by combining features of natural and synthetic repertoires, such as introduc- tion of synthetic diversity only in CDR3 while leaving CDR1 and CDR2 entirely germ line,18 introducing synthetic diversity into CDR1

and CDR2 while introducing natural diversity into CDR3,25 and diversifying multiple CDRs with minimalist degenerate codons strongly biased to- wards a few common amino acids observed in natural paratopes.9 More recent libraries have attempted diversifying all six CDRs by synthetic methods, controlling which amino acids are found at each position to be more consistent with those found in nature.21

The ability to mimic nature in synthetic libraries has been limited by both the quality of synthesis technologies and the depth of our understanding of the natural repertoire. Recent advances in high- throughput sequencing and repertoire analysis have made it possible to survey the antibody functional diversity landscape at great depth.6 In this study, we present a synthetic library design derived from high-throughput sequencing of a human donor- derived natural repertoire. To act on the design, we modified a novel solid-phase gene synthesis tech- nology that enables precise positional control of amino acid composition and frequency.26 We evaluate the genetic and functional diversity of the design with Sanger sequencing, high-throughput sequencing, expression analysis, and antigen pan- ning. We compare the sequence space explored in these libraries with previously reported natural and synthetic libraries.

5.3.2 Results

Design of synthetic fitness libraries A total of four heavy-chain and two light-chain Fab libraries were designed based on the amino acid variation observed in natural antibody repertoires.6,27–30 Germ-line frameworks VH1–69, VH3–23, VH3–30, VH4–34, V 1–39 (K02), and V 3–20 (A27) were selected based on three criteria: thermostability of V-gene families,18 germline frequency in natural repertoires,6,28–30 and heterodimer compatibility observed in productive binders.6 The synthetic fitness Fab (SF-Fab) design aimed to minimize variation that might cause instability or misfolding by mimicking natural diversity. Diversity was determined through the analysis of amino acid frequencies in functional binders from public data- bases and high-throughput sequencing analysis of a somatically diversified natural repertoire derived from over 650 human donors.6 While both periph- eral blood mononuclear

cells and public databases have considerable redundancies in their composi- tion, bias was minimized by rendering the data sets clonally nonredundant by including only one representative from each clonally related population of sequences derived from a common V(D)J rearrangement event for the final analysis. The design of VH1-69/V 1–39 Fab library is shown in Fig. 1. All six CDRs were diversified using amino acid variation observed in nature, while framework positions were left entirely human germ line. In CDRs, diversity was incorporated in a subset of positions observed to tolerate variation in somatically hypermutated natural antibodies and to be structurally proximal to antigen contact. The CDR-H3 design considered a cassette of the 10 lengths that are most often observed in nature (lengths 7–16; positions 95–100j, Kabat), avoiding very short and long CDR-H3s that have been reported to be underrepresented in recovered binders.22 In each length, the diversity of amino acids was introduced based on their frequen- cy at each position within that length. All heavy-chain libraries received germ-line-specific CDR-H1 and CDR-H2 diversity designs but shared a common CDR-H3 design. At all variable positions, amino acid frequencies were rounded off to increments of 5%, cysteines were excluded from the design, and rounding off was corrected by the slight over- incorporation of tyrosine when appropriate.10 An example of designed versus observed amino acid frequency in VH1–69 CDR-H3 (length 13) and V 1– 39 CDR3 is shown in Tables 1a and 1b.

Generation of SF-Fab libraries The library construction technology reported in this study was based on a novel ligation gene synthesis technology (Slonomics®), which relies on a set of universal building blocks ("anchors") that is used in a series of enzymatic reactions to generate any possible DNA sequence triplet per triplet.26 The technology was modified in order to allow complex codon mixture incorporations in highly defined gene libraries. Instead of ligating a single building block to the growing chain, a defined mixture of anchors is used to introduce a set of triplets (codons) at a certain position. In the following ligation step, a complementary mixture of anchors is added either to return to the constant wild-type sequence or to generate a further variable position (Supplementary Fig. 1). Based on this principle, any variable motif can be introduced into any position of a DNA sequence. Precise amino acid composition is controlled by adjusting codon frequencies in an anchor mix. A total

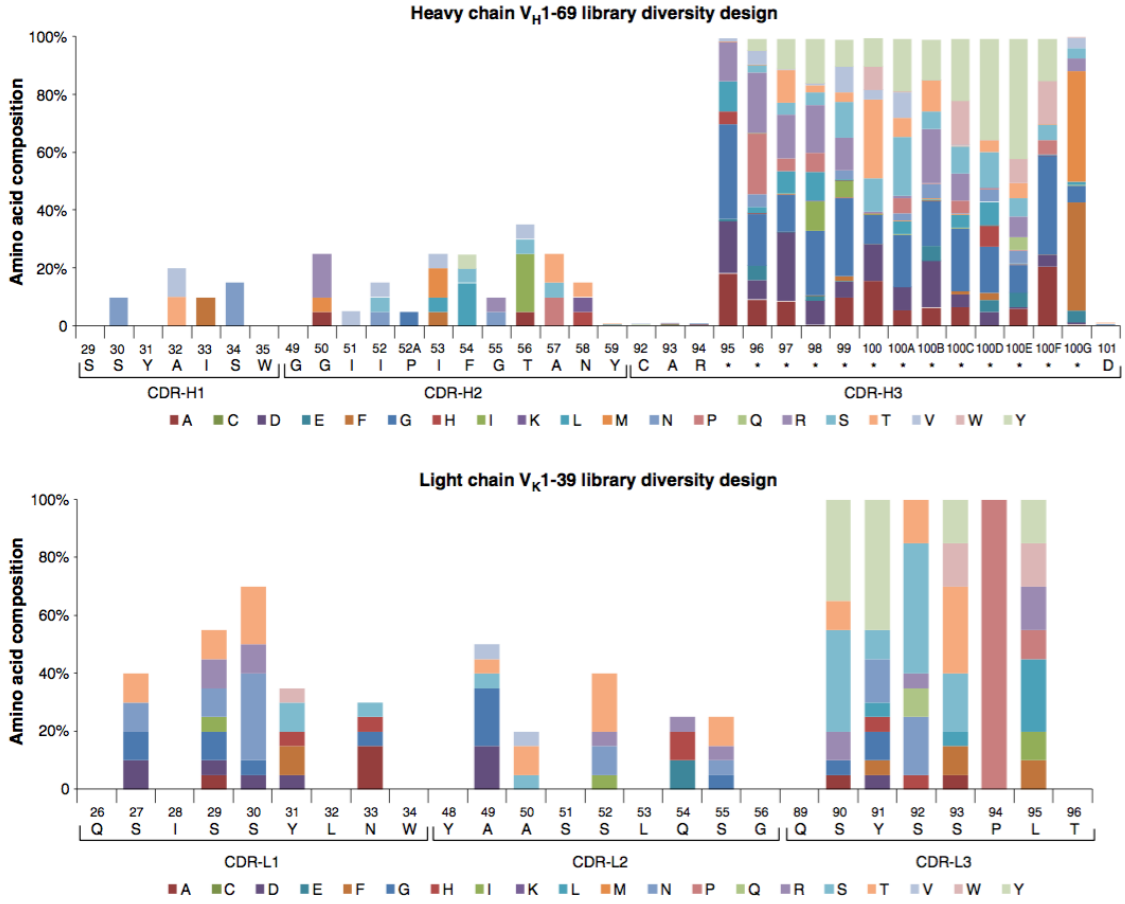


Figure 5.3: Library design of sublibrary VH1–69/V 1–39. Biased amino acid frequencies observed from affinity-matured nature repertoires were used to introduce diversity into all six CDRs. Non-germ-line amino acid variation for each CDR position is indicated for all CDRs: Kabat labeling and germline reference residue are shown for each position. Amino acids were incorporated at the frequencies observed in nonredundant populations of functional binders from public databases and high-throughput sequencing of the donor-derived repertoires,6 rounding off to the nearest 5%. The CDR-H3 design considered a cassette of the 10 lengths most often observed in nature (lengths 7–16, Kabat positions 95–100j). All heavy- chain libraries shared the same H3 cassette along with their germ-line-specific CDR-H1 and CDR-H2 diversity. The CDR- H3 length 13 design is shown.

of 100 "g of DNA was obtained for each library. All four VH libraries were paired with each of two V libraries to create eight SF-Fab libraries (SF-Fab1 to SF-Fab8). After a total of 800 transformations, an estimated 3.6 × 1010 successful tranformant-containing Fab antibodies were obtained.

454 sequencing of the libraries The quality of both pre-transformation and post- transformation SF-Fab libraries was monitored via Roche Titanium 454 high-throughput pyrosequen- cing. Total filtered sequences obtained from each library before and after transformation are summa- rized in Supplementary Table 1. In the sequence set, we observed 0.00069 framework base miscalled per framework nucleotide. However, the accumulation of errors was such that if one error was located in the framework of a single sequence, more errors were more likely to be observed in other areas of that same sequence (Supplementary Fig. 5). This positive correlation of errors in "low quality" sequences leads us to drop any sequence where a known error could be detected. Short sequences, sequences with errors in framework positions, and sequences with detectable indel errors were removed from subse- quent analysis. Initial pre-transformation sequenc- ing was used to identify and correct deviations from design introduced during library synthesis. Final sequencing was used to evaluate sequence compo- sition of the library. Each V-segment library was compared to the design with a minimum of 17,400 full-length translated reads, and a single library (VH1–69/V 3–20) was sequenced to a much greater depth of 532,479. Amino acid composition of every variable position created during synthesis was compared to the expected design. The resulting amino acid incorporations were highly correlated (Pearson's r = 0.992) to the corresponding design for both high (N50%) and low (5%) expected incorpo- ration frequencies (Fig. 2a). Across all variable positions, codons were inserted in proportion to their expected frequency, although a marginal trend towards underincorporation of phenylalanine co- dons and overincorporation of arginine codons was observed (Fig. 2b). Amino acid usage remained unaltered before and after transformation (Pearson's r N 0.999 for all CDR-H3 lengths). Stop codon usage was observed to be at most 0.024% per position (3139 stop codons of 13,079,597 generated CDR-H3 amino acids observed). In the final library, 99.6% of H3 designs were free of stop codons (3103 CDR-H3s with at least one stop codon

out of 736,929 CDR-H3s observed). The stop codon rate was 15-fold lower than that observed in 504,550 CDR-H3 sequences obtained from a synthetic CDR-H3 library produced using TRIM synthesis (0.36% per position observed over 8,387,818 CDR-H3 amino acids generated).

Distance of CDRs to natural repertoire Sequence variation in CDR1 and CDR2 was compared between SF-Fab libraries, Dyax synthetic libraries, HuCAL GOLD synthetic libraries, and donor-derived natural libraries (Fig. 3a). For all antibodies in each repertoire, the number of amino acid changes in the CDRs to the closest ImMuno- GeneTics (IMGT) human germ-line V gene was evaluated.31 Antibody variable domain sequences from the donor-derived IgM natural library were obtained from high-throughput sequencing in a previous study.6 Antibody variable domain sequences from published Dyax synthetic libraries were simulated based on reported amino acid frequencies at diversified positions.18,25 SF-Fab sequences were obtained by Roche 454 sequencing in this study. The Dyax library with synthetic diversity in CDR1 and CDR2 produced antibodies with more than 99% of the sample library carrying at least six nonhuman germ-line mutations. The HuCAL GOLD synthetic libraries encoded no variation into CDR1 and CDR2. Our synthetic design produces a V-gene mutational load compa- rable to that generated by somatic hypermutation in the natural repertoire, with less than 5% of the library being germ-line reference, less than 5% containing over five non-germ-line mutations, and the remaining 90% exploring combinations of one to five mutations in the V-segment CDR1 and CDR2. All synthetic libraries produce repertoires with less framework mutations than those observed in natural repertoires.

CDR-H3 amino acid distance to the natural human repertoire was compared between SF-Fab, HuCAL GOLD synthetic, unbiased TRIM synthetic, and natural donor-derived libraries (Fig. 3b). The antibody sequences in each library were compared to an independent natural reference set of 1.4 million nonredundant CDR-H3 amino acid sequences obtained from CD20+CD27 fluorescence-activated cell-sorted peripheral blood mononuclear cells from 16 human donors of diverse age, gender, and ethnographic origins (unpublished results). For each CDR-H3 sequence in a

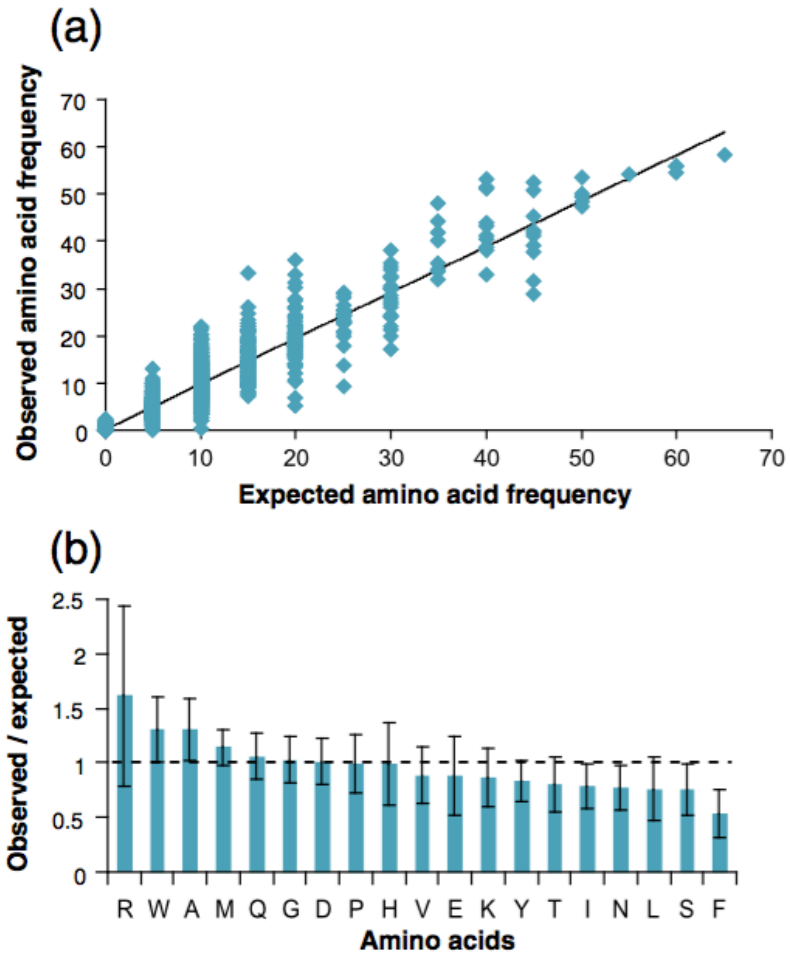


Figure 5.4: (a) Amino acid incorporation, observed versus expected. All amino acids at all variable positions were compared to the design with a minimum of 20,000 full- length translated reads from high-throughput sequencing. The observed amino acid frequencies were highly corre- lated (Pearson's r=0.992) to the design for both high (N50%) and low (5%) expected incorporation frequencies. (b) When averaged across all variable positions in all sublibraries, amino acid incorporation resembles expected with marginal overincorporation of arginine and under- incorporation of phenylalanine.

library repertoire, the percentage of amino acid identity to the most similar same-length CDR-H3 in the reference set was obtained. For the donor-derived SF-Fab and TRIM repertoires, the query sequences were obtained by pyrosequencing. For HuCAL GOLD, the reference sequences were simulated from the amino acid frequencies reported for the published library design.21 For any CDR-H3 amino acid sequence in the SF-Fab library, the closest CDR-H3 found in the natural reference set was 64% identical on average (length was normalized to avoid overcounting most frequent lengths). Other synthetic libraries pro- duced CDR-H3 repertoires that are more distant from the natural repertoire: 46% and 54% identity for unbiased TRIM and HuCAL GOLD, respective- ly. The SF-Fab design produces a CDR-H3 popula- tion with amino acid distance comparable to that observed in donor-derived natural repertoires (63% identical with closest reference). The trend was observed for all CDR-H3 lengths, with the strongest effects observed for longer CDR-H3 lengths. Data for CDR-H3 length 14 is shown in Fig. 3b.

Overlap and frequency of library clones Each chain of each sublibrary was sequenced once before transformation and twice after transforma- tion. All reads were translated and aligned to a reference profile hidden Markov model and were filtered by expected framework composition sur- rounding the CDRs. In the recovered sequences, clonal overlap in the sequencing results of CDR-L3 and CDR-H3 repertoire was evaluated (Fig. 4).

In the light chain, 50,000 randomly sampled CDR- L3 sequences were compared between a pre-trans- formation sample, v0, and two post-transformation sequencing sets, v1 and v2 (Fig. 4a). In all samples, the majority of clones were observed multiple times, with between 5250 and 5309 unique clones observed in each sample of 50,000 sequences. Over 75% of clones in each sample were observed in another sample, with 62% (3329) of CDR-L3s from any sample found in all three samples. When extended to all 294,272 CDR-L3 sequences obtained from post-transformed libraries, 13,948 unique amino acid CDR-L3s were observed, with 56.5% being observed more than once and over 98.8% of all 14,112 possible CDR-L3s theoretically available by the design recovered during sequencing. For all unique clones, the observed clone frequency obtained from sequencing was compared to the theoretical clone frequency

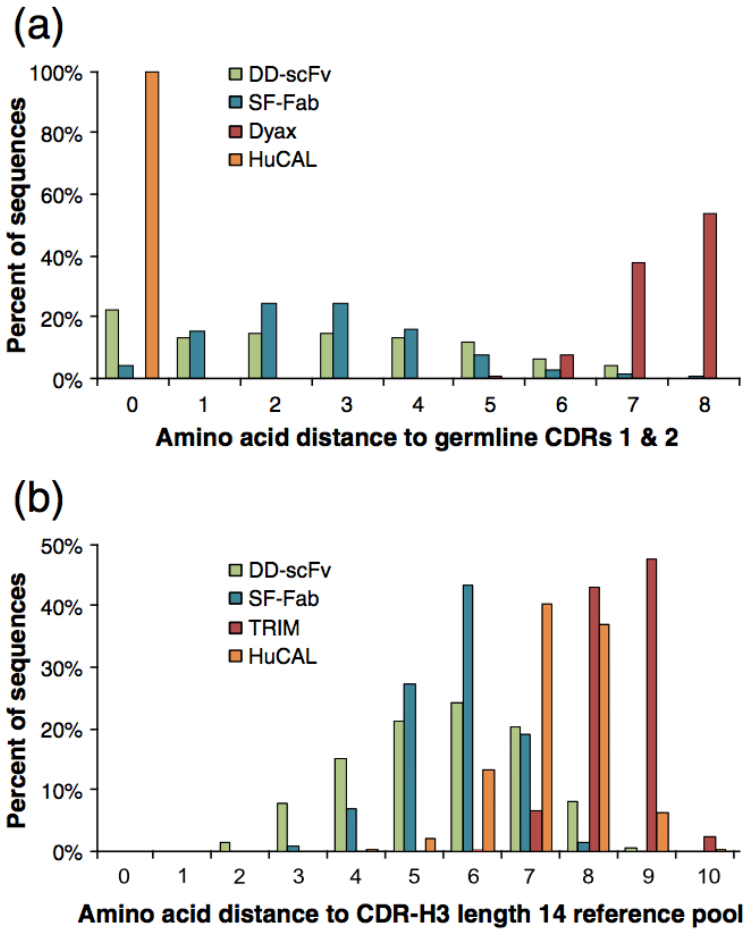


Figure 5.5: (a) Sequence variation in CDR1 and CDR2 was compared among SF-Fab libraries, previously published synthetic libraries, and a donor-derived library. For each sequence in each repertoire, the number of CDR1 and CDR2 amino acid changes to IMGT reference is reported. For each library, the percentage of sequences in each mutation category is reported. (b) CDR-H3 length 14 amino acid distance to a natural human repertoire reference data set of 1.4 million sequences from 16 healthy donors (unpublished results). The SF-Fab CDR-H3 diver- sity largely overlaps that of the donor-derived repertoire in amino acid composition.

expected from the design (Fig. 4b). Theoretical clone frequency describes how often one expects to find a specific clone given a known library design. When compared to an actual library of finite size, it can be used to predict the probability of a clone being present. For example, a 10! 12 clone is expected to appear once for every 1012 sequences. In a 1010 library, it has a 1% chance of being present. In general, any clone with a theoretical frequency less than the inverse of the library size is expected to either not appear in the library or appear as a singleton. Conversely, clones are likely to appear multiple times in a library if the theoretical clone frequency is greater than the inverse of the library size. High-frequency clones can have a strong negative impact on total library size by "occupying" space that could otherwise be occupied by singletons.

Clones expected to be of high frequency in the design were recovered more frequently during sequencing. Common CDR-L3 clones were ob- served to be highly represented in the repertoire, with 10! 3 CDR-L3s representing 30% of the library, 10! 4 CDR-L3s representing 52% of the library, and 10! 5 CDR-L3s accounting for 16% of the library. Singleton CDR-L3 clones, recovered only once during sequencing, had lower theoretical frequen- cies than CDR-L3s recovered multiple times. Overall, observed clone frequencies were highly correlated to theoretical clone frequencies across all observed frequencies (Pearson's r=0.733).

In the heavy chain, 100,000 randomly sampled CDR-H3 sequences were compared in the same manner as described above (Fig. 4c). In the CDR-H3 samples, the overwhelming majority of clones were observed only once, with 90,164, 98,409, and 98,596 unique clones observed in sets v0, v1, and v2, respectively. Almost all clones observed were unique to a single sample, with between 97.0% and 98.5% of clones being unique to each sample and only 188 CDR-H3s found in common between all three samples. Overlap between pre-transforma- tion and post-transformation samples was minimal, with 0.8% of overlapping clones observed between samples. Post-transformation samples, obtained from the same sample preparation, exhibited 2.4% overlap. In all 737,285 translated CDR-H3 post- transformation sequences, 696,090 unique CDR-H3 clones were observed, with 5.0% observed more than once. For all clones observed more than once, and a random subset of 1000 CDR-H3 singletons, the observed clone

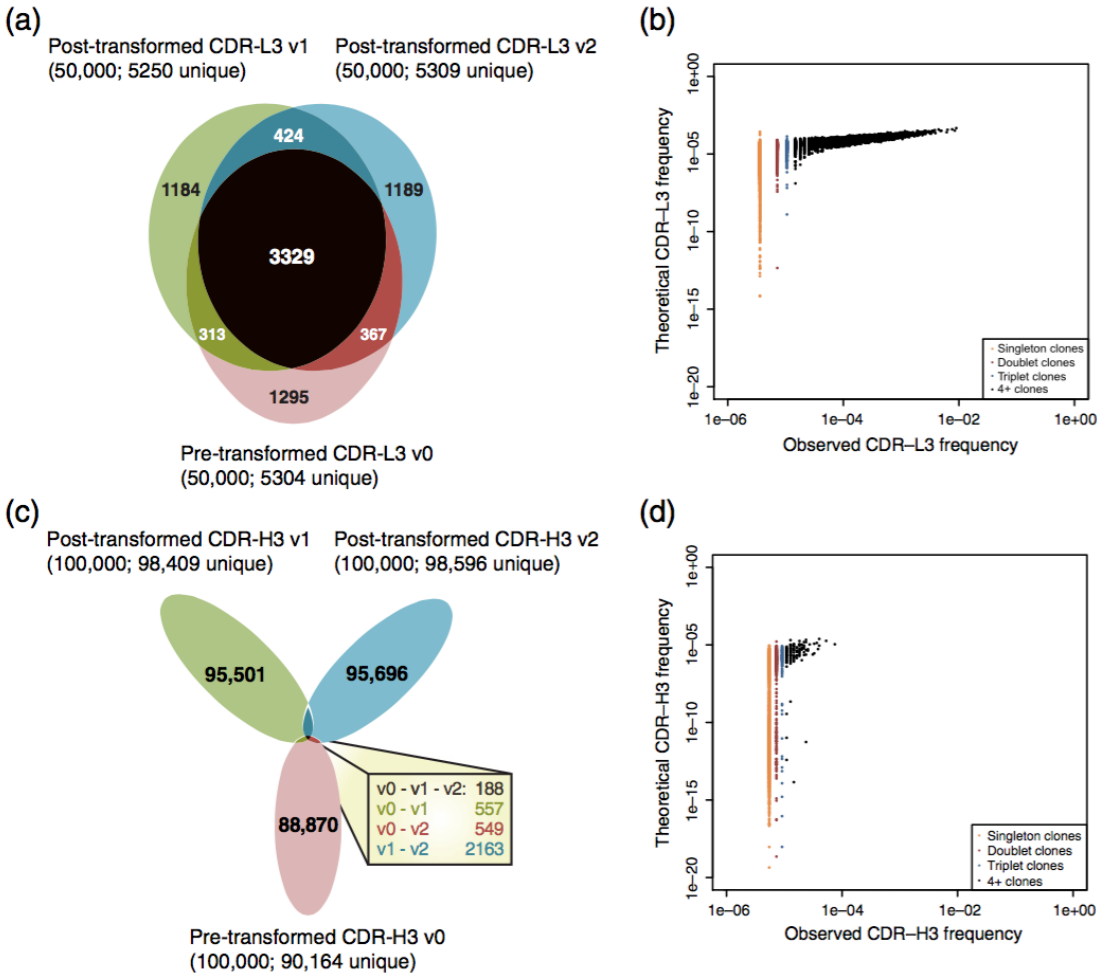


Figure 5.6: Clonal frequency and overlap in CDR-H3 and CDR-L3 repertoires. (a) Overlap of CDR-L3 repertoire observed in 50,000 sequences from one pre-transformation population (v0) and two post-transformation populations (v1 and v2). The majority of CDR-L3 clones were recaptured in all three samples. (b) Observed versus theoretical CDR-L3 frequency of 13,948 unique clones recovered from 294,272 sequences. High-frequency clones dominate the repertoire. (c) Overlap from CDR-H3 repertoire observed in 100,000 sequences from v0, v1, and v2. Over 98% of clones were only observed once. (d) Observed versus theoretical CDR-H3 frequency of clones recovered from 747,285 sequence$2s. All clones appear with frequencies b10! 4, and low-frequency clones dominate the repertoire.

frequency obtained from se- quencing was compared to the theoretical clone frequency expected from the design (Fig. 4d). In contrast with the CDR-L3, almost all clones were observed at low frequency, and no clones were predicted or observed to occur at a frequency higher than 10! 4. Clones recovered five or more times during sequencing had correlated high theoretical frequencies (10! 4 to 10! 5) and were expected from the design. Clones observed more than once represented 1.5% of the total library. The majority of these clones were from the shortest CDR-H3 lengths that have the fewest variable positions and are therefore inherently expected to produce higher-frequency clones. We also observed a very short CDR-H3 clone not expected by design but repre- senting 0.13% of all observed clones. The clone represented a failed insertion of CDR-H3, establish- ing the ligation efficiency of the CDR-H3 insertion step at greater than 99%.

CDR-H3 linkage profile During V(D)J recombination, adjacent amino acids are "linked" on the segment and continue to appear more frequently at adjacent positions in the V(D)J rearranged repertoire. This correlation be- tween adjacent amino acids has been observed to greatly constrain theoretical diversity in natural CDR-H3 repertoires. The SF-Fab synthetic method allows complete recombination between all amino acids at every adjacent position and is therefore not expected to exhibit similar linkage constraints. We compared the degree of linkage in equal-sized sample populations of 15,084 CDR-H3 length 10 sequences obtained from donor-derived and SF-Fab libraries. At every position and for every amino acid, we compared the fraction of all possible adjacent amino acid triplets observed at each position in the CDR-H3 (Fig. 5). For example, alanine is among the amino acids observed at greater than 1% usage at CDR-H3 Kabat position 97 in the natural repertoire. At position 96, glutamate, lysine and glutamine are also observed at greater than 1% usage; the same is true for proline, serine and tyrosine at position 98. In the population of 15,084 sequences from the data set from the donor-derived library, four of nine possible triplet combinations of these residues are found, while E–A–P, E–A–S, K–A–Y, Q–A–S, and Q–A–Y are never found. Given the independent amino acid frequencies at each position, these combinations would have been expected. Therefore, in the subset of residues displayed in the sample figure, linkage only provides for 44% of

all possible combinations of this subset of neighbor amino acid triplets for alanine at position 97. Total triplet linkage was calculated as the fraction of all possible amino acid triplets observed at every position in the CDR-H3 for each amino acid observed with at least 1% frequency at a given position. When all amino acids at all positions were evaluated in the data set of 15,084 randomly chosen sequences, only 49% of all possible combinations were observed in the donor- derived repertoire (Fig. 5a). In contrast, the synthetic design produces a repertoire in which every residue is found recombined with almost every other residue (Fig. 5b) with 96% of all possible triplets observed in a similar random size-matched subset of 15,084 sequences.

Expression of functional Fab antibodies The expression of full functional Fab antibodies in the library was accessed by ELISA from 1790 random clones. The soluble Fab antibodies were captured onto the plates by anti-His antibody (His tag on heavy chain) and detected by horseradish peroxidase (HRP)- conjugated anti-human kappa and/or anti-human Fab antibodies. On average, 93% of all tested clones in the library expressed correctly folded heavy-chain and light-chain Fab fragments (Table 2). Higher average expression efficiency was observed for V 3– 20 (96%) than for V 1–39 (91%). Heavy-chain VH3–23 produced higher average expression efficiency (96%), while the other three VH libraries performed compa- rably (92%, 92%, and 93% for VH1–69, VH3–30, and VH4–34, respectively). The positive Fab expression rate for all sublibraries was higher than that reported in past synthetic Fab designs (61%, 70%, and 77%).18,21,32 The expression level was estimated by ELISA along with a titrated standard control (purified human Fab antibody) based on 280 random clones from the library (Supplementary Fig. 2). In this subset, 94% of analyzed clones expressed soluble Fab fragments. It was observed that in the expressing clones, 91% expressed more than 240 ng/ml and the remaining 9% expressed between 60 and 240 ng/ml of soluble Fab as estimated from the standard curve. The number of Fabs per phage was determined according to an ELISA method33,34 by calculating the ratio of phage in the linear range derived from anti- kappa/anti-M13 (pIII) capture antibodies. The results show that there is an average of 0.4–0.5 Fab molecules per phage (Supplementary Fig. 3).

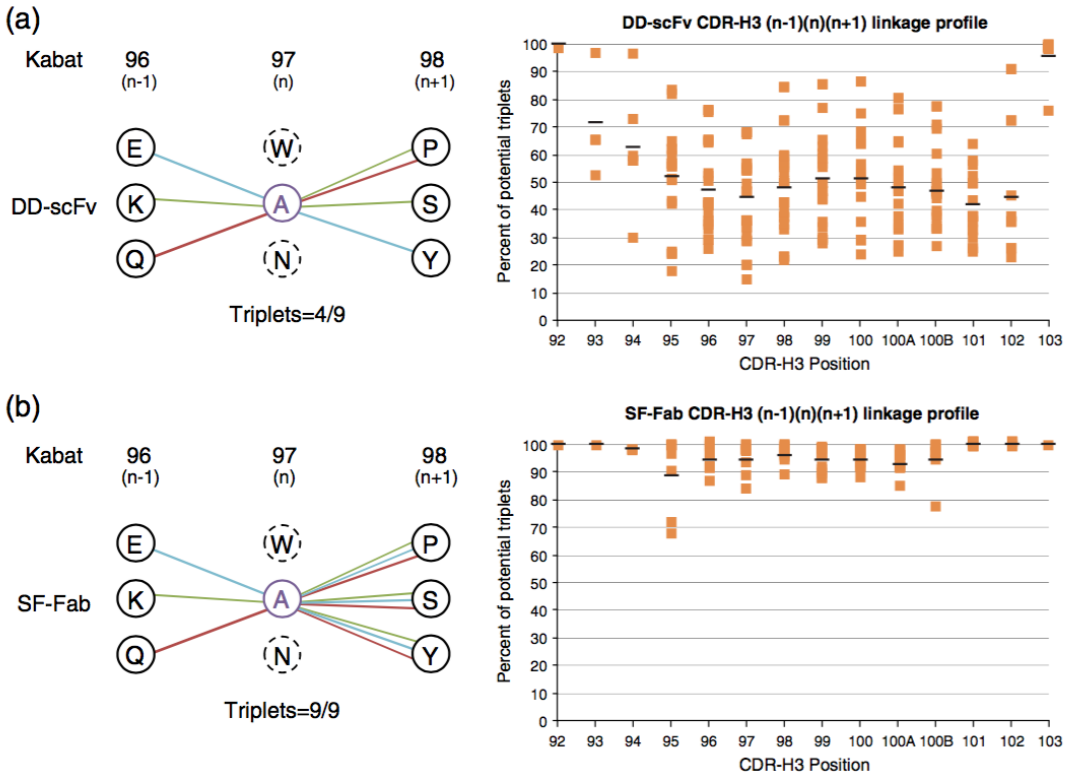


Figure 5.7: Lossofdiversityduetolinkagebetweenadjacentaminoacidpositions inCDR-H3.Ontheleft,asu acids observed at Kabat positions 96, 97, and 98 is indicated. Colored lines indicate combinations of adjacent residues observed in sequences from reference sets of 15,084 CDR-H3 length 10 sequences. On the right, all adjacent (n ! 1)(n)(n + 1) amino acid triplets observed for each residue at each position in CDR-H3 are indicated. (a) A donor-derived repertoire exhibits linkage between amino acid composition at adjacent positions: 49% of all potential (n ! 1)(n)(n + 1) adjacent amino acid triplets are observed. (b) Over 96% of all potential (n!1)(n)(n+1) adjacent amino acid triplets observed in the synthetic SF-Fab library sequences.

Selection of antibodies against a panel of antigens The library produced positive binders against a diverse panel of 10 antigens including receptors, growth factors, an enzyme, a peptide, and a viral antigen (Table 3). In three to four rounds of panning, ELISA-positive clones were recovered against all antigens, with systematic enrichment of ELISA- positive clones observed between rounds (Supple- mentary Fig. 4). Recovery varied between antigens, with ELISA-positive clones represented between 5% and 76% of clones sampled from the third or the fourth round of panning. Sequencing of ELISA- positive clones recovered 5–180 unique antibodies per antigen. The binding kinetics of nonredundant binders were determined by biosensor analysis, with accurate kinetic and affinity parameters able to be determined for 6 of 10 antigens (Fig. 6a and b). Typical results of the single analyte concentration and multiple analyte concentration methods appear in Fig. 6c and d, respectively. Affinities ranged from over 100 nM to under 1 nM. Without any affinity maturation, three of the antigens panned (CD152, H1N1 hemagglutinin, and hIgG2) produced hits ranging from 20 nM to 300 pM. Of the samples selected for follow-up kinetic analysis, it was revealed that 5 anti-H1N1, 13 anti-hFab, and 9 anti-CD152 antibodies had a KD value of less than 10 nM, with the best affinities being 1.4 nM, 2.5 nM, and 0.3 nM, respectively (Fig. 6a).

Sequence analysis of recovered binders Some VH-segment libraries were more effective at producing hits against the panel of antigens tested. Libraries utilizing the VH1–69 framework produced 309 unique binders, VH3–23 produced 106 unique binders, VH3–30 produced 79 unique binders, and VH4–34 produced 20 unique binders. Light-chain bias was not observed overall (237 V 1–39 versus 249 V 3–20 binders recovered), although V 1–39 was observed more frequently in the highest-affinity binders. Detailed analysis of recovered binders from VH1–69 demonstrated that amino acid usage in CDR-H1 and CDR-H2 resembled the design (Fig. 7a). All amino acids encoded in the V-gene design were recovered in binders. The positional amino acid frequency in binders was highly correlated with the design (Pearson's r = 0.995). Arginine at CDR-H2 Kabat position 55 was underrepresented (0.6% observed versus 3.8% expected), and proline at CDR-H2 Kabat position 57 was over-represented (28% observed versus 16% expected). Binders were recovered from all

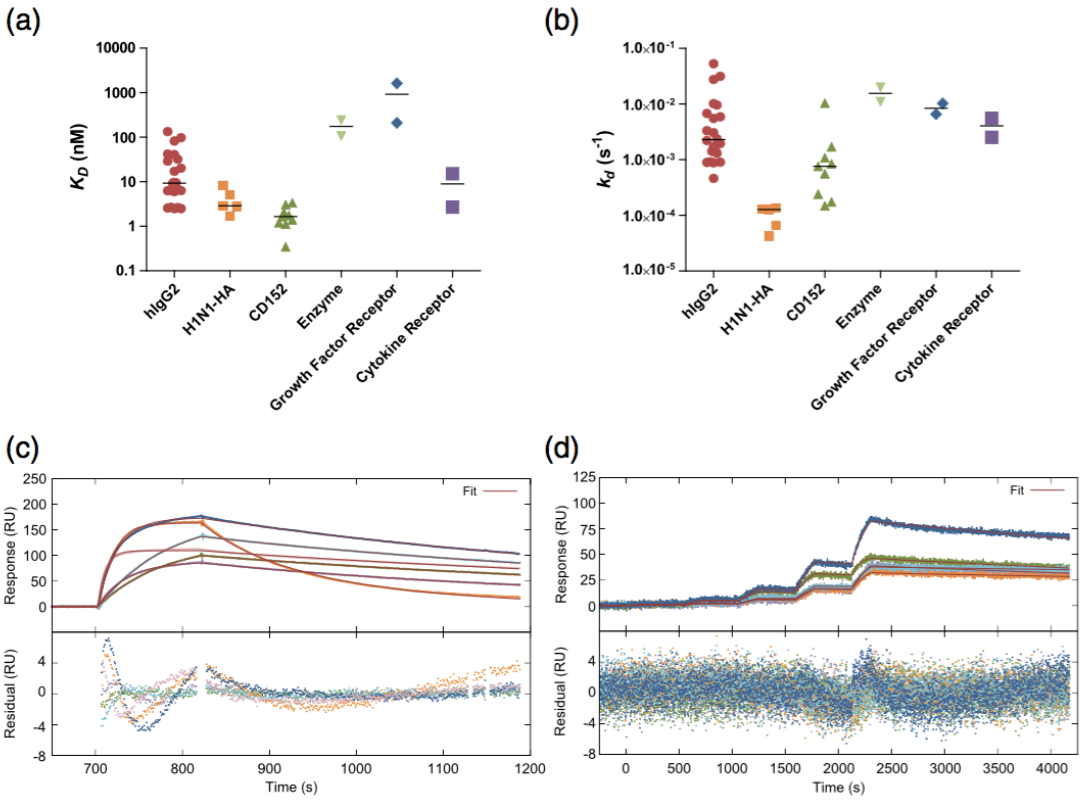


Figure 5.8: Affinity estimation of Fabs selected against hIgG2, H1N1-HA, CD152, an enzyme, a growth factor receptor, and a cytokine receptor. The median value of (a) KD (in nanomolar) and (b) kd (per second) is shown as a horizontal bar for each target where the number of Fabs tested were 22, 5, 9, 2, 2, and 2 for hIgG2, H1N1-HA, CD152, an enzyme, a growth factor receptor, and a cytokine receptor, respectively. Representative biosensor data for kinetic and affinity analyses using a single analyte concentration for an hFab binding to captured anti-hIgG2 samples (c) and representative biosensor data for kinetic and affinity analyses using multiple analyte concentrations for partially purified Fabs binding to captured H1N1-HA (d).

CDR-H3 lengths included in the design, with the number of binders for each CDR-H3 length recovered being significantly correlated with the underlying frequency of the CDR-H3 length in the library (Pearson's one-sided p b 0.001) (Fig. 7b). The library design allowed theoretical clone frequencies to be calculated for any sequence. This makes it possible to estimate the frequency of a panned binder prior to the enrichment that occurred during panning. Further analysis of VH1–69 clone frequencies shows that the majority of recovered binders were rare clones expected to occur only once in library. Over 69% of recovered binders had an expected CDR-H3 frequency of less than 10! 9 (Fig. 7c), indicating that the clone should appear at most once in each initially transformed library. The distribution of clone frequencies in the library was equivalent to that observed in recovered clones, and panned hits were recovered from rare and common clones without bias due to the underlying frequency (Fig. 7c).

5.3.3 Discussion

A major goal for library design is to maximize both genetic diversity and functional diversity. Natural antibody repertoires exhibit high functional diversity but cannot access all theoretical genetic diversity. Synthetic repertoires generate high genet- ic diversity but have not been selected for functional success. Incorporating signatures of selection from natural repertoires into synthetic designs has been demonstrated to improve functional diversity. Pio- neering work by Zemlin et al. identified characteristic amino acid CDR3 profiles in natural repertoires.7 Recent developments in next-genera- tion sequencing have allowed direct observation of millions of diverse antibody variants that have survived positive, negative, and antigen-driven selection. Analysis of antibody repertoires by these methods have improved our understanding of biases in germ-line usage, germ-line heterodimer pairing, somatic hypermutation, and position-specific amino acid usage of the highly diverse CDR3.6 A sufficiently precise synthetic technology could use the biases in natural repertoires as design features in synthetic libraries.

In contrast to the oligonucleotide-based diversity generation methodologies, the novel library synthe- sis technology utilized here makes it possible to control codon

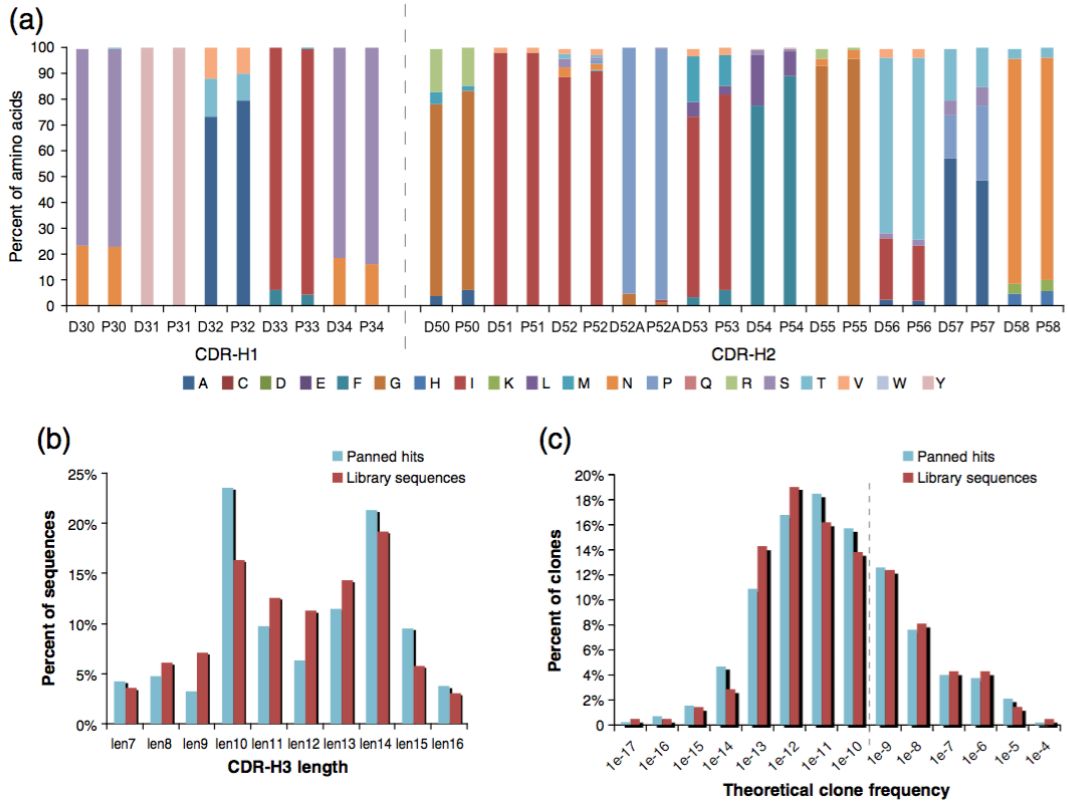


Figure 5.9: Amino acid and clone frequency analyses of 306 recovered binders from the VH1–69 sublibraries. (a) Comparison of the amino acid usage in CDR-H1 and CDR-H2 between sequenced library (D) and binders recovered from panning (P). All amino acids in design are recovered in binders, and positional frequency is correlated. (b) CDR-H3 length distribution between sequences recovered from library and 399 unique sequences recovered from panning against specific antigens. (c) Comparison of theoretical clonal frequencies in library and recovered binders. The majority (69%) of recovered binders were derived from low-frequency clones.

and frequency while avoiding a majority of stop codons, frameshifts, and other common errors of gene and library assembly. Reducing stop codons and frameshifts will increase the proportion of functional diversity in any library. However, the principal advantage of the novel synthetic approach is improved precise control over positional amino acid frequencies. This novel library synthesis technology enables the generation of synthetic antibody libraries from designs of significant complexity. In this study, the synthesis approach was used to design antibody diversity directly after selected, biased amino acid frequencies observed in natural repertoires.

A combination of novel synthetic technology and natural mimic design resulted in a library with a high proportion of correctly folded, functional antibodies. In 1790 clones evaluated, 93.4% were observed to display both a correctly folded heavy-chain Fab and a correctly folded light-chain Fab on their surface coats. The observed expression effi- ciency was substantially higher than the 61–77% reported for other large synthetic library designs of comparable diversity.18,21 Lower stop codon and frameshift rates contribute to the improved expres- sion performance but do not completely explain the difference. Rothe et al. reported 77% of clones expressing Fab, but only 9% frameshifts, suggesting that a subset of sequences may have been misfolded without obvious errors in translation. Steipe et al. have made the general observation that there is a correlation between the stability of an antibody and sequence proximity to other known, well-folding antibodies.23 We observed unbiased synthetic de- signs to produce a repertoire very distant from the natural repertoire (Fig. 3). Synthetic designs that mimic natural biased overusage of tyrosine and glycine create synthetic repertoires that are closer to the natural distribution but clearly remain distinct.10 The SF-Fab libraries were observed to generate antibody populations that were closer in amino acid composition to natural repertoires than previous synthetic libraries. It is possible that increased proximity to the natural antibody sequence land- scape may also contribute to the proportion of antibodies that are able to avoid misfolding and recognize antigen.

The SF-Fab design combines the high expression of the natural repertoire with the genetic diversity benefits of synthetic design. The synthetic ligation process reproduces the natural amino acid frequen- cies without the constraints of linkage. In

a synthetic repertoire, all possible combinations of amino acids at adjacent positions are found, with the frequency of each combination being the product of their independent frequencies. In contrast, a natural repertoire is generated by the combinatorial rear- rangement of a finite number of V, D, and J segments. Because adjacent positions on a given segment are linked in the germ line, the resulting repertoire exhibits only a subset of all possible amino acid combinations occurring between adja- cent positions. Eliminating linkage allows greater recombinant diversity, a reduction in redundant clones, and a more dispersed genetic diversity capable of thorough exploration of the natural sequence landscape. In contrast with previously observed donor-derived repertoires,6 no popula- tions of high-frequency (10! 2 to 10! 4) clones were observed during deep sequencing of the synthetic repertoire, and a very low recapture rate was observed for CDR-H3 clones even when 7.0 × 105 clones were evaluated. This reduction of high- frequency, redundant clones contributes to a more diverse library.

The linkage-free design and in-depth characteri- zation of the library provides a unique means of interrogating the underlying mechanics of phage display. Previous publications suggested that very rare clones may be at a disadvantage in panning recovery.35 The SF-Fab library design and high- throughput characterization allowed calculation of a theoretical frequency for all clones that correlated to observed frequencies obtained during sequencing. When applied to over 300 recovered binders against a panel of antigens, the resulting calculation in- dicates that rare clones are recovered from the library without selective disadvantage. The obser- vation is consequential for future library designs, as it supports further efforts to maximize nonredun- dant diversity during library design.

The combined use of immune repertoire sequenc- ing, natural mimic design, and novel library synthesis technologies has enabled a new class of library generation. Natural mimic designs combine many of the beneficial properties of natural and synthetic libraries. As a result of carefully designed diversity, natural mimic libraries are capable of being more diverse than the natural repertoire while also being closer to human germ line than are current synthetic designs. The library exhibits improved expression efficiency over that reported for past synthetic designs and fewer high-frequency

clones than that observed in donor-derived reper- toires and produces high-affinity binders against a variety of antigens. Future applications of natural mimic design can be used to generate specialized libraries with unique repertoire characteristics.

5.3.4 Methods

Library construction Constant parts of the sequence were amplified from a wild-type-containing plasmid, introducing flanking BpiI, Eco31I, or Esp3I sites for later assembly with the variable parts. Variable regions were synthesized de novo using a modification of the Slonomics® procedure for gene synthesis.26 Mixtures of anchor molecules were generated in a defined ratio to represent all combinations of two adjacent positions within a variable region. These mixtures were connected to the growing DNA chain by ligation. The reaction product was purified by immobilization on a streptavidin-coated surface. New overhangs for the next reaction cycle were then generated by restriction with Eam1104I, and a new mixture of anchor molecules was added. After three to seven reaction cycles, product pairs were combined to generate transposition intermediates. These were either combined in a second round to form long variable regions or assembled with constant parts by restriction and ligation. Length variants were synthesized separately, quantified by PAGE, mixed in a defined ratio, and assembled as one pool. Heavy-chain and light-chain fragments were synthesized separately in a 1-"g scale and combined by restriction with Esp3I in the intergenic region and subsequent ligation. Final ligation product was amplified to 100 "g by PCR using Phusion DNA polymerase (NEB) in up to 1000 parallel reactions.

Fab antibody phage display libraries A total of 100 "g of DNA from each of eight libraries was cut with SfiI and NotI, respectively, purified, and cloned into pRN8910 vector containing a Flag tag and a 10-His tag. To obtain phage antibody libraries, we transformed TG1-competent cells (Cat. No. 200123, Stratagene) by electroporation using BTX 1-mm-gap cuvettes. For individual library, 100 transformations were performed in parallel reactions. A total of estimated 3.6 $\times$ 1010 transformants containing Fab antibodies were obtained from a total of 800 electroporations.

SF-Fab libraries	Numbers of clones tested	Numbers of ELISA-positive clones	% of Fab expression
IGHV1–69/IGKV1–39	282	249	88.3
IGHV1–69/IGKV3–20	188	181	96.3
IGHV3–23/IGKV1–39	188	181	96.3
IGHV3–23/IGKV3–20	192	184	95.8
IGHV3–30/IGKV1–39	282	244	86.5
IGHV3–30/IGKV3–20	188	185	98.4
IGHV4–34/IGKV1–39	282	259	91.8
IGHV4–34/IGKV3–20	188	177	94.1
Average (total %)			93.4

ELISA-positive clones are determined by an optical density value at 405 nm over three times more than the negative control.
"% of Fab expression" means the rate of soluble, well-folded Fab expression from each sublibrary.
Total percentage of functional Fab expression (93.4%) is calculated as the normalized average of sublibraries.

Table 5.1: Zhai Glanville JMB 2011 Table1.

454 amplicon sequencing PCR amplification of the VH and V regions of the SF-Fab libraries (1–8) was carried out using Amplicon Fusion Primers containing the GS FLX Titanium Primer A or B sequence for the Lib-A chemistry, six-base barcodes, and specific primers complementary to the constant regions of the vector. A 1:4 mixture of Platinum Taq (Invitrogen) and Pfu (Promega) polymerases was used for the amplification with15cyclesof94°Cfor45s,56°Cfor45s,and72°Cfor 1 min followed by 72 °C for 10 min. Gel-purified PCR products were quantified using the Quant-iT PicoGreen assay (Invitrogen), and amplicon lengths were assessed using the Agilent Bioanalyzer 2100 DNA Chip 7500. The VH and V amplicons of the corresponding SF-Fab libraries were pooled at a ratio of 2.3:1. The pools were diluted and subjected to emulsion PCR and bidirectional sequencing using reagents and protocols for the GS FLX Titanium Lib-A chemistry (Roche). SF-Fab1–4 and SF-Fab6–8 were pooled in equal quantities and sequenced in one region of a two-region GS FLX Titanium Run. SF-Fab5 was sequenced to a greater depth over two regions of a 2-region GS FLX Titanium Run.

Sequence QC Pass-filter reads were obtained from standard 454 signal processing. Reads with perfect matches to multiplex identifiers were translated and aligned to a reference profile hidden Markov model.6 Short reads and reads with detectable insertion or deletion errors were eliminated, leaving full-length reads covering at least 90 amino acids of variable domains in a single frame. Synonymous substitution errors were ignored entirely, and sequences with substitution errors in framework boundary

positions of CDRs were excluded from analysis (Supplementary Fig. 5). Although the approach is able to detect and eliminate the majority of errors, two error types remain: nonsynonymous substitutions in CDRs and double-indel errors within a single CDR. Previous reports establish 454 error rates as 0.08% substitutions per base, 0.18% inserts per base, and 0.13% deletions per base.36 Given the reported error rates for the Roche GS20, the number of errors expected to escape filters can be estimated. Nonsynonymous mutations per CDR length 10=((0.08% mut/bp sub-error)×(75.5% nonsynon/mut))×3 bases per codon×CDR length 10=1.8% per CDR3 length 10. CDR compound indels per CDR length 10 = ((30 bases × 0.18% insert rate)×(29 bases×0.13% deletion rate))=0.2% per CDR length 10. The estimated error rate will have minimal impact on assessments of CDR length distributions or position-specific amino acid frequencies. Clonal frequen- cies can be expected to be underestimated by 2% when an error causes a loss in observed clone copy number.

Amino acid usage Amino acid usage was recorded as a position-specific scoring matrix, with the observed frequency of each amino acid at each position in a gapless global alignment of sequences of the same length (each CDR-H3 length had a separate position-specific scoring matrix). Observed amino acid usage was compared to the expected one at every designed variable position by Pearson's r. Amino- acid-specific bias was evaluated by taking the observed/ expected ratio of incorporation at any position, irrespec- tive of absolute frequency or sequence context.

Clonality assessment Observed clone frequency is the frequency that a clone appears in a sampling of a sequence population. Observed clone frequencies can be calculated as the number of observed copies of that clone ni over the total number of observed clones N. The observed clone frequency should be considered unreliable when ni = 1 and increasingly reproducible as ni grows greater than 1. Rare clone frequencies can be determined only up to $1/(2 \times \text{depth})$.

Theoretical clone frequency is the expected frequency for a clone, given a known distribution of CDR-H3 lengths, a known composition of positional amino acid frequencies in the design, and an assumption of positional independence (linkage-free design). Theoretical clone frequency can be calculated as

(formula)

Antigens	Antigen class	No. of clones tested	No. of ELISA-positive clones	No. of unique Fabs	k_d (1/s)	K_D (nM)
H1N1-HA	Viral r-protein	380	212	28	4.2×10^{-5} to 1.3×10^{-4}	1.7–8.3
IgG	Hu-IgG2	756	328	180	9.1×10^{-4} to 5.3×10^{-2}	2.5–134
VEGF	Growth factor	188	21	19	n/q	n/q
Growth factor R	Receptor	940	130	42	6.5×10^{-3} to 1.0×10^{-2}	210–1630
NGF	Growth factor	846	367	72	n/q	n/q
Enzyme	Enzyme	658	32	30	1.1×10^{-2} to 2.0×10^{-2}	106–243
Neurotrophin R	Tyrosine kinase	380	38	9	n/t	n/t
cGRP	Peptide	376	29	17	n/q	n/q
Cytokine R	Receptor	188	120	5	6.5×10^{-3} to 1.0×10^{-2}	2.7–15.2
CD152	Receptor	188	142	19	5.5×10^{-3} to 2.5×10^{-3}	0.3–3.4

n/q, not quantified; n/t, not testable.
VEGF, vascular endothelial growth factor; NGF, nerve growth factor; cGRP, calcitonin gene-related peptide.

Table 5.2: Zhai Glanville JMB 2011 Table2.

where fi is the frequency of clone i, Prob(i) is the probability (frequency) of the CDR length observed in sequencei,Prob($i,p|PFM)istheprobabilityofobserving amino acid $ at position p of sequence i given the position frequency matrix PFM for length .

Evaluating natural repertoire sequence space V-segment natural repertoire distance was calculated as the number of amino acid changes in CDR1 and CDR2 compared to the closest germ-line framework from IMGT. CDR-H3 natural repertoire distance was calculated as the number of amino acid changes in CDR-H3 compared to the closest CDR-H3 in a large natural reference database. The natural CDR-H3 repertoire reference set included 1.43 million nonredundant amino acid sequences from 16 donors diverse in age, gender, and ethnographic origin (unpublished results). The sequences included all residues between Kabat 94 and Kabat 101, noninclusive. For a given CDR, distance to reference was the number of amino acid mismatches to the most similar CDR of the same length in the reference set. For a collection of sequences, the distance is the distribution of distances of all members, with distances normalized by CDR length.

Linkage Correlations between adjacent amino acids in natural and synthetic CDR designs were quantified by evaluating the fraction of all potential triplets actually observed in a reference set of 100,000 sequences. For each reference global gapless alignment of CDR4 length 14, all amino acids observed with at least 1% frequency

at each position were recorded. For each amino acid at each position n, the fraction of all potential neighbor combinations at positions (n ! 1) and (n + 1) was recorded.

Selection of SF-Fab libraries The rescue of phage antibody particles with helper phage VCSM13 was performed according to previously published methods.37 For the selection, 1013 colony- forming units of phage was incubated with either biotinylated antigens (solution-phase selection) or anti- gens coated onto immune tubes (solid-phase selection). Most antigens were biotinylated at a ratio of 1–5 molecules of EZ-link Sulfo-NHS-LC-LC-biotin per molecule of antigen according to the manufacturer's instruction (Pierce, Cat. No. 21338). For solution-phase selection, the biotinylated antigens were captured by 100 "l of streptavidin-coated superparamagnetic beads (Dynabeads® M-280 streptavi- din, Cat. No. 112.06D from Invitrogen) in 1-ml volume of Dulbecco's phosphate-buffered saline–0.1% bovine serum albumin in Eppendorf tubes. Following 5× washing, the preabsorbed libraries were incubated with the antigens captured onto the beads. The protein antigens were used for solution-phase selection at a starting concentration of 200 nM and gradually reduced to 5 nM along with increasing of panning rounds. For the peptides, 1 "M, 500 nM, 250 nM, and 25 nM were used. The specific antibodies to various antigens were selected and screened by ELISA and Biacore after three to four rounds of biopanning. The unique antibodies were confirmed and clustered by sequencing.

Soluble Fab expression Individual colonies were inoculated in 1.5 ml of 2YT– ampicillin containing 2% glucose media in 96-well plates and cultured at 30 °C overnight. The cultures were centrifuged at 4000 rpm for 20 min, and the cell pellets were resuspended in 1.5 ml of 2YT–ampicillin media supplemented with 1 mM IPTG for the induction. After continuing to culture at 30 °C for 5 h with shaking at 200 rpm, the cell pellets were resuspended in 300 "l of Hepes-buffered saline. After freezing and thawing three times, cell lysates were filtered and the soluble Fabs were screened and tested via ELISA and Biacore. For the interested hits, the Fab antibodies were reformatted, expressed, and purified in 293 cells. We routinely obtained 20–30 mg of antibodies per liter of the culture.

ELISA For the screening after panning, biotinylated antigens were captured on pre-immobilized streptavidin plates at room temperature (RT) for 1 h, and non-biotinylated antigens in phosphate-buffered saline (PBS) were coated on MaxiSorp ELISA plates at 4 °C overnight. Plates were then blocked with 1% bovine serum albumin–PBS at RT for 1 h. The soluble Fabs prepared from the culture were added to plates and incubated at RT for 2 h. Following five washes with PBS plus Tween 20, the bound Fab antibodies were detected with either anti-human kappa or anti-His HRP antibody, and the optical density was measured at a wavelength of 405 nm. Clones giving a positive signal in ELISA (over three times of the background) were further analyzed via sequencing and Biacore. For the test of Fab expression, 50 "l of anti-His antibody (2 "g/ml) was coated onto ELISA plates at 4 °C overnight. Following washing and blocking, soluble Fabs were added and detected by HRP-conjugated anti-kappa and/or anti- human Fab antibody.

Biosensor analysis for monovalent antigens The kinetic and affinity analyses for monovalent antigens (hFab, an enzyme and a growth factor receptor) were performed on a Biacore 2000 biosensor equipped with a CM5 sensor chip. The SF-Fabs prepared as described above were captured from undiluted superna- tant via an amine-coupled anti-Flag capture molecule. The antigen was flowed as the analyte at a single concentration or, in other cases, as a dilution series using a kinetic titration methodology. In both cases, guidelines for accurate kinetic analysis were employed, and data were double-referenced as described in Myszka38 and Karlsson et al.39 Figure 6c shows representative data and fits for the single analyte concentration method where the monova- lent antigen was hFab. Data were fit using a 1:1 Langmuir model with mass transport.

Biosensor analysis for multivalent antigens The kinetic and affinity analyses for multivalent antigens (CD152 Fc Fusion and H1N1-HA) were per- formed on surface-plasmon-resonance-based biosensors (Biacore 2000 and ProteOn XPR36, respectively). The CD152 Fc Fusion was captured via an anti-human Fc surface that was amine-coupled onto a Biacore CM5 chip, and partially purified Fab was flowed as the analyte using a kinetic titration methodology. Biotinylated H1N1-HA was captured

via neutravidin that was amine-coupled onto a ProteOn GLH sensor chip, and partially purified Fabs were analyzed in parallel using an array-based kinetic titration methodology as described in Abdiche et al.40 Data were fit using a 1:1 Langmuir model with mass transport. Supplementary materials related to this article can be found online at doi:10.1016/j.jmb.2011.07.018

5.3.5 Acknowledgements

We would like to thank Dr. Javier Chaparro-Riggers for referral to Sloning BioTechnology GmbH. We are very grateful to Gabriella Huerta for her excellent technical assistance in 454 pyrosequencing and to Dr. Jan Berka for his valuable discussions in high-throughput sequencing. We would also like to thank Dr. Christoph Erkel and Dr. Katja Siegers for contributions to the further development of the synthetic technology for gener- ation of antibody libraries and Constance Benedict for her assistance with the preparation of figures.

5.3.6 References

1. Carter, P. J. (2006). Potent antibody therapeutics by design. Nat. Rev., Immunol. 6 , 343 – 357.

2. Nelson, A. L., Dhimolea, E. & Reichert, J. M. (2010). Development trends for human monoclonal antibody therapeutics. Nat. Rev., Drug Discov. 9 , 767 – 774.

3. Smith, G. P. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. Science , 228 , 1315 – 1317.

4. Figini, M., Marks, J. D., Winter, G. & Griffiths, A. D. (1994). In vitro assembly of repertoires of antibody chains on the surface of phage by renaturation. J. Mol. Biol. 239 ,68 – 78.

5. Lloyd, C., Lowe, D., Edwards, B., Welsh, F., Dilks, T., Hardman, C. & Vaughan, T. (2009). Modelling the human immune response: performance of a 1011 human antibody repertoire against a broad panel of therapeutically relevant antigens. Protein Eng. Des. Sel. 22 , 159 – 168.

6. Glanville, J., Zhai, W., Berka, J., Telman, D., Huerta, G., Mehta, G. R. et al. (2009). Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proc. Natl Acad. Sci. USA , 106 , 20216 – 20221.

7. Zemlin, M., Klinger, M., Link, J., Zemlin, C., Bauer, K., Engler, J. A. et al. (2003). Expressed murine and human CDR-H3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. J. Mol. Biol. 334 , 733 – 749.

8. Ippolito, G. C., Schelonka, R. L., Zemlin, M., Ivanov, I. I., Kobayashi, R., Zemlin, C. et al. (2006). Forced usage of positively charged amino acids in immunoglobulin CDR-H3 impairs B cell development and antibody production. J. Exp. Med. 203 , 1567 – 1578.

9. Fellouse, F. A., Esaki, K., Birtalan, S., Raptis, D., Cancasci, V. J., Koide, A. et al. (2007). High-through- put generation of synthetic antibodies from highly functional minimalist phage-displayed libraries. J. Mol. Biol. 373 , 924 – 940.

10. Birtalan, S., Zhang, Y., Fellouse, F. A., Shao, L., Schaefer, G. & Sidhu, S. S. (2008). The intrinsic contributions of tyrosine, serine, glycine and arginine to the affinity and specificity of antibodies. J. Mol. Biol. 377 , 1518 – 1528.

11. Jiang, N., Weinstein, J. A., Penland, L., White, R. A., III, Fisher, D. S. & Quake, S. R. (2011). Determinism and stochasticity during maturation of the zebrafish antibody repertoire. Proc. Natl Acad. Sci. USA , 108 , 5348 – 5353.

12. Mora, T., Walczak, A. M., Bialek, W. & Callan, C. G., Jr (2010). Maximum entrop y models for antibody diversity. Proc. Natl Acad. Sci. USA , 107 , 5405 – 5410.

13. Barbas, C. F., 3rd, Bain, J. D., Hoekstra, D. M. & Lerner, R. A. (1992). Semisynthetic combinatorial antibody libraries: a chemical solution to the diversity problem. Proc. Natl Acad. Sci. USA , 89 , 4457 – 4461.

14. Sidhu, S. S., Li, B., Chen, Y., Fellouse, F. A., Eigenbrot, C. & Fuh, G. (2004). Phage-displayed antibody libraries of synthetic heavy chain complementarity determining regions. J. Mol. Biol. 338 , 299 – 310.

15. Hackel, B. J., Ackerman, M. E., Howland, S. W. & Wittrup, K. D. (2010). Stability and CDR composition biases enrich binder functionality landscapes. J. Mol. Biol. 401 ,84 – 96.

16. Ge, X., Mazor, Y., Hunicke-Smith, S. P., Ellington, A. D. & Georgiou, G. (2010). Rapid construction and characterization of synthetic antibody libraries without DNA amplification. Biotechnol. Bioeng. 106 , 347 – 357.

17. Silacci, M., Brack, S., Schirru, G., Marlind, J., Ettorre, A., Merlo, A. et al. (2005). Design, construction, and characterization of a large synthetic human antibody phage display library. Proteomics , 5 , 2340 – 2350.

18. Knappik, A., Ge, L., Honegger, A., Pack, P., Fischer, M., Wellnhofer, G. et al. (2000). Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs random- ized with trinucleotides. J. Mol. Biol. 296 ,57 – 86.

19. Virnekas, B., Ge, L., Pluckthun, A., Schneider, K. C., Wellnhofer, G. & Moroney, S. E. (1994). Trinucleotide phosphoramidites: ideal reagents for the synthesis of mixed oligonucleotides for random mutagenesis. Nucleic Acids Res. 22 , 5600 – 5607.

20. Sidhu, S. S. & Fellouse, F. A. (2006). Synthetic therapeutic antibodies. Nat. Chem. Biol. 2 , 682 – 688.

21. Rothe, C., Urlinger, S., Lohning, C., Prassler, J., Stark, Y., Jager, U. et al. (2008). The human combinatorial antibody library HuCAL GOLD combines diversification of all six CDRs according to the natural immune system with a novel display method for efficient selection of high-affinity antibodies. J. Mol. Biol. 376 , 1182 – 1200.

22. Sheets, M. D., Amersdorfer, P., Finnern, R., Sargent, P., Lindquist, E., Schier, R. et al. (1998). Efficient construction of a large nonimmune phage antibody library: the production of high-affinity human single- chain antibodies to protein antigens. Proc. Natl Acad. Sci. USA , 95 , 6157 – 6162.

23. Steipe, B., Schiller, B., Pluckthun, A. & Steinbacher, S. (1994). Sequence statistics reliably predict stabilizing mutations in a protein domain. J. Mol. Biol. 240 ,188 – 192.

24. Fellouse, F. A., Barthelemy, P. A., Kelley, R. F. & Sidhu, S. S. (2006). Tyrosine plays a dominant functional role in the paratope of a synthetic antibody derived from a four amino acid code. J. Mol. Biol. 357 , 100 – 114.

25. Hoet, R. M., Cohen, E. H., Kent, R. B., Rookey, K., Schoonbroodt, S., Hogan, S. et al. (2005). Generation of high-affinity human antibodies by combining donor- derived and synthetic complementarity-determining- region diversity. Nat. Biotechnol. 23 , 344 – 348.

26. Van den Brulle, J., Fischer, M., Langmann, T., Horn, G., Waldmann, T., Arnold, S. et al. (2008). A novel solid phase technology for high-throughput gene synthesis. BioTechniques , 45 , 340 – 343.

27. Abhinandan, K. R. & Martin, A. C. (2008). Analysis and improvements to Kabat and structurally correct num- bering of antibody variable domains. Mol. Immunol. 45 , 3832 – 3839.

28. Hood, L. E. (2008). Wu and Kabat 1970: a transform- ing view of antibody diversity. J. Immunol. 180 , 7055 – 7056.

29. Johnson, G. & Wu, T. T. (2000). Kabat database and its applications: 30 years after the first variability plot. Nucleic Acids Res. 28 , 214 – 218.

30. Wu, T. T. & Kabat, E. A. (1970). An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implica- tions for antibody complementarity. J. Exp. Med. 132 , 211 – 250.

31. Lefranc, M. P., Giudicelli, V., Ginestoux, C., Bodmer, J., Muller, W., Bontrop, R. et al. (1999). IMGT, the international ImMunoGeneTics database. Nucleic Acids Res. 27 , 209 – 212.

32. Yang, H. Y., Kang, K. J., Chung, J. E. & Shim, H. (2009). Construction of a large synthetic human scFv library with six diversified CDRs and high functional diversity. Mol. Cell , 27 , 225 – 235.

33. Junutula, J. R., Bhakta, S., Raab, H., Ervin, K. E., Eigenbrot, C., Vandlen, R. et al. (2008). Rapid identification of reactive cysteine residues for site- specific labeling of antibody-Fabs. J. Immunol. Methods , 332 ,41 – 52.

34. Mazor, Y., Van Blarcom, T., Carroll, S. & Georgiou, G. (2010). Selection of full-length IgGs by tandem display on filamentous phage particles and Escherichia coli fluorescence-activated cell sorting screening. FEBS J. 277 , 2291 – 2303.

35. Kretzschmar, T., Zimmermann, C. & Geiser, M. (1995). Selection procedures for nonmatured phage antibodies: a quantitative comparison and optimization strategies. Anal. Biochem. 224 , 413 – 419.

36. Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L. & Welch, D. M. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. 8 , R143

37. Marks, J. D. & Bradbury, A. (2004). Selection of human antibodies from phage display libraries. Methods Mol. Biol. 248 , 161 – 176.

38. Myszka, D. G. (1999). Improving biosensor analysis. J. Mol. Recognit. 12 , 279 – 284.

39. Karlsson, R., Katsamba, P. S., Nordin, H., Pol, E. & Myszka, D. G. (2006). Analyzing a kinetic titration series using affinity biosensors. Anal. Biochem. 349 , 136 – 147.

40. Abdiche, Y. N., Lindquist, K. C., Pinkerton, A., Pons, J. & Rajpal, A. (2011). Expanding the ProteOn XPR36 biosensor into a 36-ligand array expedites protein interaction analysis. Anal. Biochem. 411 , 139 – 151.

5.3.7 Copyright

5.4 Engineering in-vivo synthetic repertoires

Chicken immune responses to human proteins are often more robust than rodent responses because of the phylogenetic relationship between the different species. For

discovery of a diverse panel of unique therapeutic antibody candidates, chickens therefore represent an attractive host for human-derived targets. Recent advances in monoclonal antibody technology, specifically new methods for the molecular cloning of antibody genes directly from primary B cells, has ushered in a new era of generating monoclonal antibod- ies from non-traditional host animals that were previously inaccessible through hybridoma technology. However, such monoclonals still require post-discovery humanization in order to be developed as therapeutics. To obviate the need for humanization, a modified strain of chickens could be engineered to express a human-sequence immunoglobulin variable region repertoire. Here, human variable genes introduced into the chicken immunoglobulin loci through gene targeting were evaluated for their ability to be recognized and diversi- fied by the native chicken recombination machinery that is present in the B-lineage cell line DT40. After expansion in culture the DT40 population accumulated genetic mutants that were detected via deep sequencing. Bioinformatic analysis revealed that the human targeted constructs are performing as expected in the cell culture system, and provide a measure of confidence that they will be functional in transgenic animals.

5.4.1 Introduction

Historically, therapeutic monoclonal antibodies have been derived from immunized mice and phage display technologies. However, antigens that are conserved throughout mammalian evolution are typically weakly or non-antigenic in mice. In some cases, the fail- ure to elicit an immune response in mice has been obviated by immunizing chickens (1–3). Early attempts to use chicken-derived antibodies were thwarted by the lack of technology to derive mon- oclonal antibodies from non-murine animals. A fusion partner for chicken B cells was identified to create an avian version of the classical murine hybridoma technology (4) although it has not gained wide usage and phage display has been used more frequently to isolate chicken monoclonals (5–11). We developed technology to isolate antigen-specific monoclonal antibodies from immunized chickens using a single lymphocyte screening and recovery method, the gel-encapsulated microenvironment (GEM) assay (see US Patents 8030095 and

841517382). The GEM assay involves placing a single antibody-secreting lymphocyte in prox- imity with reporters (which can be cells or beads). The secreted antibody diffuses locally within the GEM and has the opportunity to bind to the reporters. Bound antibody can be detected either directly through the use of a secondary antibody or by eliciting a response in the reporter that generates a visual signal. Each GEM may contain multiple types of reporters which can be differenti- ated from each other based on color. Selected GEMs are isolated and antibody genes are amplified through RT PCR and cloned into a mammalian expression vector, usually in scFv format.

The advantage of producing antibodies to conserved epitopes in chickens prompted the development of humanization protocols to obviate the immune response in patients to the avian V regions of chimeric antibodies (7, 8). An alternative approach to eliminat- ing the anti-animal response in patients is to engineer the animal to produce human immunoglobulins (12). We are currently creating a line of chickens that will produce antibodies with fully human V regions. Human V regions will be recovered from these birds using GEMs. We will then combine the human V segments with human constant regions to produce fully human antibodies with therapeutic potential. The human V region sequences have been designed to replace the equivalent chicken coding regions while leaving most of the endogenous IgH and IgL regulatory sequences intact.

Diversification of chicken immunoglobulin genes is achieved through gene conversion (GC) and somatic hypermutation (SHM) (13). In humans, diversification is achieved through V(D)J recombination and SHM. Because of the phylogenetic distance between humans and chickens and the known differences in the mechanism of diversity generation, it was prudent to evaluate the genetically modified V regions in vitro before investing in the much longer timeline to produce genetically modified birds.

A preliminary evaluation of expression and diversification of human immunoglobulin V regions in DT40 cells was previously reported (14). Briefly, chicken VL and VH loci were knocked out in DT40 and replaced with human VK (VK3-15) and VH

(VH3- 23) genes. To achieve GC of human genes in chicken B cells, human pseudogene arrays were inserted upstream of the func- tional human VK and VH regions. The sequences of the VK and VH functional genes served as the starting template for the design of the human pseudogenes. Proper expression of chimeric IgM comprises human variable regions and chicken constant regions were shown. Sanger-based sequencing of selected DT40 genetic variants confirmed that the human pseudogene arrays contributed to the generation of diversity through GC at both the Igl and Igh loci. Although these data showed that engineered pseudogene arrays contribute to human antibody sequences in chicken B cells, a more thorough repertoire analysis was not possible as only a relatively small number of events were analyzed.

Here, we have used next-generation sequencing methods to study much more comprehensively the repertoire generated by a long-term, non-selected culture of DT40 cells harboring targeted human V genes, analyzing well over 1 million sequences for each of the heavy and light chains. We are now able to show that the engineered locus can produce a diverse pool of human antibody sequences in chicken B cells.

5.4.2 Results

READ QUALITY ANALYSIS Analysis was performed by processing all reads through VDJFasta. Sequences were assigned closest segments with a probabilistic clas- sifier. All reads were translated into six-frames of translation and analyzed for Ig content by profile-Hidden–Markov model scor- ing with VDJFasta, with a 1e-10 cutoff for significance. Pass-cutoff frames were aligned using the pHMMs and analyzed for framestate and coverage. Over 96% of reads contained full-length clones, with over 1 million HuVH reads and 2 million HuVK reads available for downstream analysis (Table 2).

COMPARISON TO INSERTED ARRAYS OF V GENES Nucleotide and translated CDRs were extracted from pro- file Hidden–Markov model alignments, using minimum profile annealing cutoffs to ensure high fidelity CDR capture [see Ref. (15)]. CDRs were compared to a reference database containing all V genes and pseudogenes

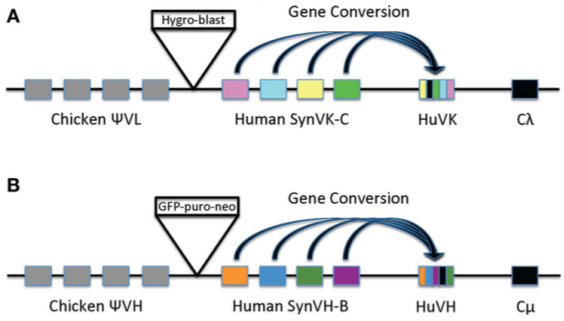


Figure 5.10: Diagrams of light chain (A) and heavy chain (B) loci in cell line 1208-1. (A) In the light chain, the endogenous rearranged chicken VL and its promoter in DT40 was replaced by an array of human SynVK-C pseudogenes and a rearranged functional HuVK gene driven by the chicken VL promoter.The chicken ΨVL pseudogene array, constant region (Cλ), J–C intron,and3′ flankingDNAareintact.Aβ-actin-hygromycin,β-actin- blasticidin resistance cassette (box labeled Hygro-blast) was placed between the chicken and human pseudogene arrays as part of the transfection process. (B) In the heavy chain, the endogenous rearranged chickenVH and350bpofitspromoterregionwerereplacedbytheSynVH-B humanpseudogenearray,thechickenVH promoter,andarearranged functionalhumanVH gene.Theupstream-chickenΨVH pseudogenearray, the chicken JH-Cμ intron, and constant regions are intact. A β-actin-EGFP, β-actin-puromycin, β-actin-neomycin selectable marker cassette (box labeled GFP-puro-neo) was placed between the chicken and human pseudogene arrays as part of the transfection process. Gene conversion in both heavy and light chains is depicted as blocks of sequences (colored blocks) being transferred from the pseudogenes to the HuVK and HuVH functional genes.

that were included in the targeted array. Counts of exact match to reference database were stored for all CDRs.

SEQUENCE COMPLEXITY The extracted sequences were highly redundant in both the heavy chain and light chain data sets, with the single functional human V gene being seen predominately in its respective group. The non- mutated VK represented 81% of all the full-length sequence reads and the non-mutated VH 57% of the total. These non-mutated sequences are referred to as the "reference" sequences (one for VH and one for VK). For the heavy chain, 9125 unique clones were found at a minimum 2× sequence depth; for the light chain, 7671 unique clones were found.

If the sequences are counted at a 1× sequence depth a total of 21,403 unique heavy chains and 33,848 unique light chain genes were seen.

IDENTIFYING GENE CONVERSION AND SOMATIC HYPERMUTATION EVENTS Each framework and CDR was analyzed separately, with an exact- match assignment performed to reference synthetic human frame- works designed into the transgenic organism. A control search was also performed with all known native IgL and IgH chicken segments, but they were never encountered in the repertoire. GC events were scored if a sequence found in the VK (Table 3) or VH (Table 4) pool could be traced back to particular pseudo- genes present in the array. In the cases where gene converted sequences are shared by multiple pseudogenes, one GC event is counted and all possible donor pseudogenes are indicated. Indi- vidual GC events were counted at the 1× sequence depth since it is unlikely that stretches of nucleotides would occur through sequencing or read error. Regardless, all CDR events classified as GC occurred more than twice in the data set. Within each CDR a high proportion of the unique sequences matched perfectly with the reference sequence. Those that deviated from the reference sequence in ways that could not be clearly attributed to GC are labeled as "SHM or fusion" and this category includes single or multiple point mutations as well as possible complex events (i.e., multiple sequential GC). Sequencing errors would be expected to show up in this category, possibly inflating the observed events.

EVALUATION OF POTENTIAL CONTRIBUTION OF ENDOGENOUS CHICKEN PSEUDOGENES The IgL and IgH knockouts were made by deleting portions of the functional VJC and VDJ regions, respectively, by homologous recombination. Since endogenous chicken pseudogenes remain upstream of the inserted human V gene array, they could in prin- ciple contribute to repertoire diversity. We specifically checked for such events by creating a library of all known chicken pseudogenes and running the analysis as we did with the library containing our human pseudogenes. Evidence of endogenous pseudogenes participating in GC events was never observed.

GENE CONVERSION AND SHM IN FRAMEWORK REGIONS Since some diversity was incorporated into the frameworks of the inserted VK pseudogenes, it

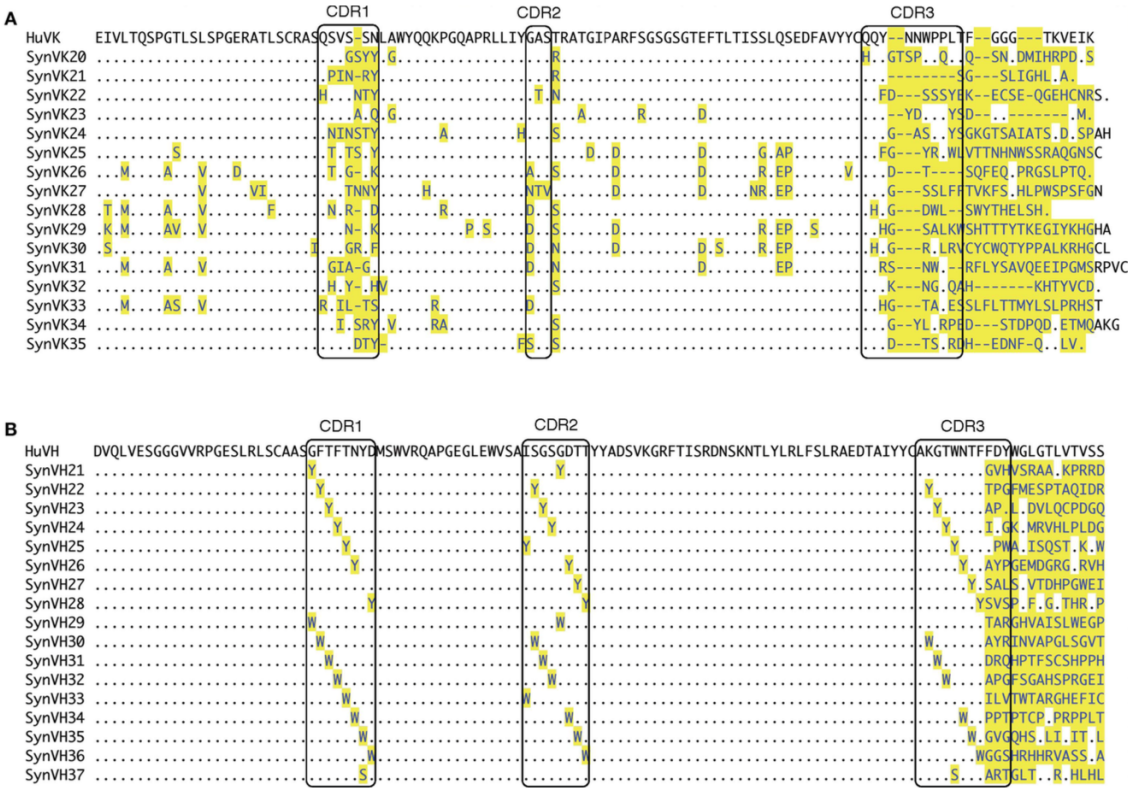


Figure 5.11: Human light and heavy chain pseudogene sequences. (A) Alignment of SynVK pseudogenes in the SynVK-C construct. Top line shows the sequence of the functional HuVK gene that is mutated by the SynVK pseudogenes. The CDRs (boxes; IMGT nomenclature) were derived from human EST databases. Some pseudogenes also contain framework changes derived from the ESTs. At the 3´ end of the pseudogenes, the sequence of the flanking DNA downstream of CDR3 in each pseudogene is shown. This flanking sequence is part of the 100 bp spacer sequence inserted between each pair of pseudogenes. (B) Alignment of the SynVH pseudogenes in the SynVH-B construct. Top line shows the sequence of the functional HuVH gene. The CDRs consist of a tyrosine/tryptophan/serine scan. The framework regions contain no changes.

was possible to identify GC events as well as SHM events in these regions (Table 5). Fewer GC attributable sequences were found in the VK frameworks as com- pared to the VK CDRs; however, this may be due simply to the lower framework diversity that was incorporated in the pseudogene design.

IDENTIFICATION OF MULTIPLE GENE CONVERSION EVENTS In some cases, we were able to identify sequences with contri- butions from two different pseudogenes, and these are termed paired-fusion events (Tables 6–8). Partial GC was analyzed using the parsimonious assumption of single-conversion events within the CDR as a source of non-100% identity match to the SynVH pseudogene segment reference database. Custom software was written to generate all non-redundant fusion events that can emerge between pairwise interactions of the reference database sequences. Paired-fusion events are highly biased in their relative occurrence, as an analysis of the most commonly encountered rearrangements demonstrates. Paired-fusion events were not fea- sible to determine for SynVK due to the sequence complexity inherent in this array.

POSITIONAL VARIATION PROFILING OF THE REPERTOIRE Analysis of positional amino acid variation was performed by con- verting a total alignment of non-redundant amino acid sequences into a positional weight matrix (PWM), with reference residue frequency omitted to emphasize non-reference residue variation (Figure 3). The number of amino acids observed at each posi- tion cannot be attributed to GC events including both complete replacements and single paired fusions. These observations suggest that the diversity generated by GC is augmented by SHM.

5.4.3 Discussion

The DT40 cell line has been used extensively to better understand the nature of immunoglobulin diversification in chickens, includ- ing the mechanism of GC (18–21). The cell line has also been used to develop chicken antibodies to novel targets using in vitro selec- tion strategies (22, 23). We have inserted human V gene arrays into the chicken immunoglobulin loci of DT40. In principle, the human V genes in DT40

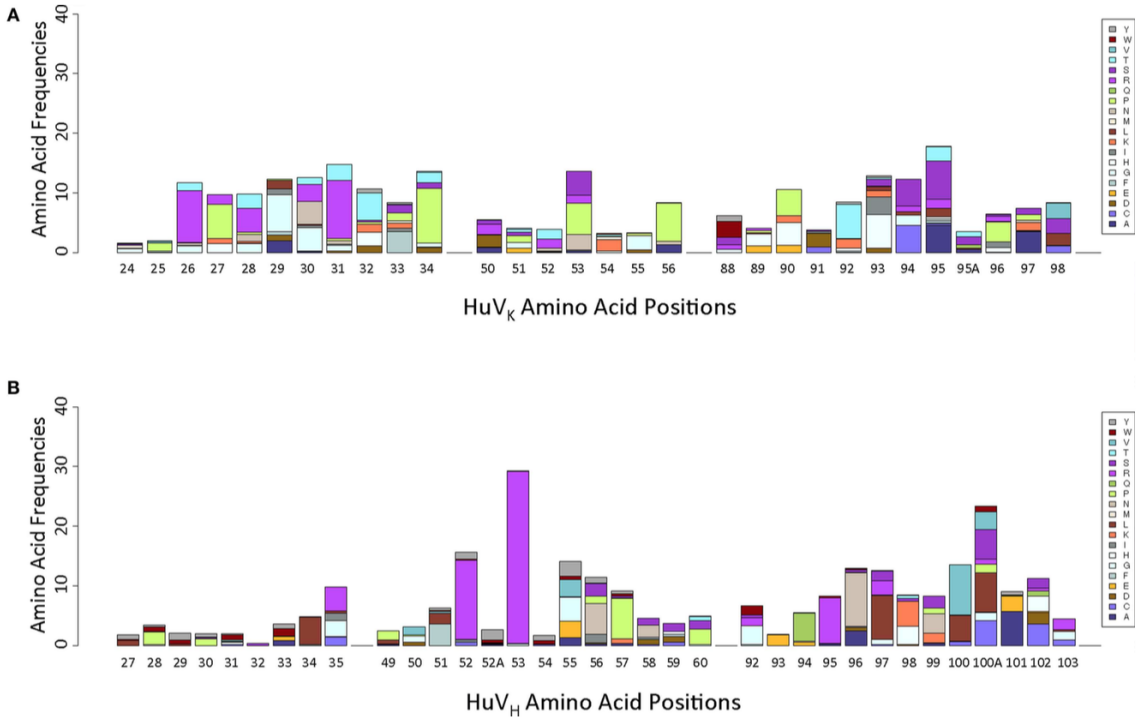


Figure 5.12: Positional weight matrices of non-reference residue variation, by Kabat position. (A) HuVK variation, indicated positionally. (B) HuVH variation, indicated positionally.

cells could be diversified in vitro to pro- vide an unselected library of immunoglobulin sequences from which antigen-specific antibodies could be extracted. However, most therapeutic antibodies are derived from immunized ani- mals producing affinity-matured, antigen-specific antibodies. In the current context, we have used DT40 cells to provide in vitro proof of concept that arrays of human-derived immunoglobu- lin gene sequences can be diversified by chicken B cells. Sub- sequently, these sequences will be introduced into chickens to provide genetically engineered animals that can be immunized to produce affinity-matured, antigen-specific antibodies with ther- apeutic potential. Thus, our purpose with DT40 is to determine whether targeted synthetic human V gene arrays can be used as a substrate for genetic diversification in chicken cells in a way that mirrors what is known regarding the native chicken immunoglobulin loci. Affirmative data in the DT40 culture system inspires confidence that the effort required on the arduous path to gen- erating a genetically engineered chicken will be rewarded with a transgenic animal that performs as expected.

We have previously shown that our heavy and light chain arrays can be diversified by both GC and SHM in DT40 cells (14). This analysis involved conventional sequencing of a few hundred clones sampled from a large population of DT40 cells. Some examples of expected diversification events were seen, but most events were likely missed at that depth. Next generation sequence technol- ogy allows for identification of very rare events in a non-selected, non-biased cellular population. Indeed, we were able to show in the current work that every pseudogene in our array was used by some cell in the population. The finding of paired-fusion events, wherein GC occurs using two different pseudogenes in succession is expected in a fully functional locus. It has been estimated that wild-type chicken B cells undergo 1–2 independent GC events on average during affinity maturation (24).

We were also able to confirm our previous conclusion that for both heavy and light chains, GC is more prevalent in CDR1 and CDR2 than CDR3, which is heavily skewed toward SHM. This finding is also consistent with previously published results (25). Nonetheless, a mix of both templated and non-templated mutations is seen in all CDRs, resulting in a repertoire with amino acid diversity at every CDR position. It will be interesting to see if a similar bias can be seen in the SynV chicken, in which

processes of cellular selection may affect the repertoire in a way that is not seen in DT40, which has no selection pressure for surface Ig expression or specificity.

While it is reasonable to use deep sequencing on our DT40 population to determine whether certain types of events have occurred, caution should be used in interpreting the observed frequency results because of the nature of the long-term DT40 culture system. In such a system, clones with particular sequences could have a growth advantage, or alternatively, a particular muta- tion could occur very early in the expansion of the culture and then subsequent mutations could occur in addition, potentially creating a large number of "unique clones," which carry the orig- inal mutation. For instance, in our VHCDR2 data, we find a very large number of clones bearing a S53R substitution (Figure 3). The high frequency observed could be the result of a true muta- tional hotspot that mutated many times independently, or simply the result of a single random mutation that occurred early in the expansion of the population. One sequence bearing this substitu- tion is highly redundant in our sample, second only to the starting sequence; this is consistent with the existence of a large subpopu- lation wherein secondary mutations could have occurred. Further, if S53R is an aberration, it skewed our data to make it appear that SHM in HC CDR2 is extremely high relative to GC, which may not be the correct interpretation.

In summary, the DT40 culture system, coupled with deep sequencing methodologies, is an excellent tool for the functional testing of arrays of synthetic human V genes designed to be diver- sified and affinity-matured in vivo. Our deep sequencing results confirm that arrays of human V genes can be targeted into the immunoglobulin loci of chicken cells and the host machinery can diversify those genes over time in a manner that recapitu- lates in vivo GC in the B cells of wild-type chickens. Furthermore, when rare events are included in the analysis, it is clear that even in a relatively small population of cells, all of the introduced pseudo- genes are capable of participating in GC, and that codons for non-templated amino acid residues are generated through SHM in every CDR of both light chain and heavy chain. These data support the concept of introducing constructs containing all nec- essary genetic elements required for diversification in the B cell compartment, contributing to a functionally diverse repertoire of human-sequence antibodies in a transgenic chicken. Once made, this

bird will be the most evolutionarily divergent host of any human-Ig transgenic animal currently available, and will be par- ticularly well suited to generating novel antibodies to therapeutic targets that are conserved among mammals.

5.4.4 Methods

CULTURE OF CHICKEN DT40 CELLS CARRYING HUMAN V GENES A derivative of the chicken B cell line DT40 was made in which the chicken immunoglobulin variable regions were replaced with human variable regions in both the IgL and IgH loci (14). In both loci, the active functional allele was targeted, thereby switching the cells from expressing normal chicken surface IgM to the expres- sion of chimeric IgM, consisting of human variable regions and chicken constant regions. A derivative of DT40, cell line 1208-1, was produced by serial transfection with knockout constructs fol- lowed by site-specific insertion of constructs for the expression of human V regions. To take advantage of the GC machinery in DT40, upstream arrays of human-sequence pseudogenes were included in the transgenes to provide the donor sequences for mutating the single functional human kappa (HuVK) and human heavy chain (HuVH) regions (Figure 1). Pseudogene arrays were synthesized by Bio Basic (Markham, ON, Canada). These pseudogenes were based on the sequences of the functional HuVK and HuVH regions, with diversity incorporated into the complementarity determin- ing regions (CDR), and in some cases, the framework regions as well (Figure 2). The pseudogenes were thus designed de novo and not based on the endogenous pseudogenes found in the human genomic heavy and light chain loci. We refer to the HuVK pseudo- genes as the SynVK array and the HuVH pseudogenes as the SynVH array. Diversity in the SynVK array was derived from human EST sequences, whereas the SynVH array was made by scanning substitution of CDR positions with tyrosine, tryptophan, or ser- ine residues. Furthermore, additional AID hotspots (nucleotides WRC/GYW) were incorporated into the SynVK-C construct, as silent changes. In the 1208-1 cell line, construct SynVH-B was inserted at the heavy chain locus, followed by insertion of the SynVK-C construct at the light chain locus. The sequences of the pseudogene arrays are shown in Figure 2.

The 1208-1 cell line was propagated for 10 weeks with both SynVK-C and SynVH-B transgenes to allow mutations to accumu- late prior to harvesting genomic DNA for sequencing (additionally, the precursor cell line carrying only the SynVH-B construct was cultured for 3weeks before transfection of the SynVK-C con- struct). The culture was expanded to 1.85e8 cells and gDNA was purified by Qiagen DNeasy kit.

GENERATION OF AMPLICONS FOR SEQUENCING Purified gDNA from the 1208-1 DT40 cell line was sent for further processing to Genewiz, Inc. The HuVK and HuVH regions were amplified using the primers in Table 1. Amplicons were sequenced by Genewiz, Inc. (South Plainfield, NJ, USA) on the Illumina MiSeq 2x250 platform (Illumina, Inc., San Diego, CA, USA). Raw data files are available online at the NCBI sequence read archive (SRA), project PRJNA275158, accession number SRP055184.

SEQUENCE DATA ANALYSIS High throughput sequencing reads were analyzed using VDJ- Fasta (http://www.distributedbio.com/vdjfastadocs/), a general antibody repertoire algorithm suitable for interpretation of engi- neered antibody diversity. In order to control for possible residual chicken content, we used a combination natural chicken and human synthetic segment classification database. In order to ana- lyze the repertoire comprehensively in a manner unbiased by the underlying GC mechanics, we used general profile Hidden– Markov models to identify immunoglobulin content and align sequences in a consistent manner independent of nucleotide composition. Kabat positional annotations were transferred from Hidden–Markov model columns to every aligned sequence in the resulting database, enabling consistent annotation of frameworks and CDR boundaries (15–17).

5.4.5 Acknowledgement

This work was made possible by the great team at Crystal Bioscience.

5.4.6 Copywrite

5.5 Engineering de-novo TCRs with optimized activity

In Chapter 2.2, we presented a method for generating de novo TCRs with prescribed specificity (Figure 2.5). The operates by first taking a specificity group of similar TCRs, grouped by GLIPH. Those sequences are then aligned, and a Positional Weight Matrix (PWM) is generated from the alignment, representing the amino acid variability at each homologous position (Figure 2.5 B, C). Once created, using Formula (5) (Figure 2.6), it is possible to estimate the probability of any sequence to emerge from that PWM. When emitted randomly, the top 1000 scoring TCRs were found to contain some TCRs already found in nature and contributing to the convergence group, but also many other novel TCRs not found in nature. When these were tested for cell-based activation activity, 80% of these de novo TCRs were functionally active and 20-40% were more active than those observed in nature. This can be expected from the method: the theory is that as one approaches the consensus of a PWM, the functional property (in this case, activity) of the clones also optimizes.

The approach is a powerful validation of the ability of GLIPH to predict specificity of TCRs. However, likely more valuable are the implications for engineering - it is now possible to apply a semi-rational approach to optimize the function of T-cell receptors. This has the potential to provide guidance in TCR engineering, with applications in acheiving clarity of fundamental mechanisms as well as direct applications in generating therapeutic TCRs.

Appendix A

Bibliography

Glanville, J., Zhai, W., Berka, J., Telman, D., Huerta, G., Mehta, G.R., Ni, I., Mei, L., Sundar, P.D., Day, G.M. and Cox, D., 2009. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proceedings of the National Academy of Sciences, 106(48), pp.20216-20221.

Qi, Q., Liu, Y., Cheng, Y., Glanville, J., Zhang, D., Lee, J.Y., Olshen, R.A., Weyand, C.M., Boyd, S.D. and Goronzy, J.J., 2014. Diversity and clonal selection in the human T-cell repertoire. Proceedings of the National Academy of Sciences, 111(36), pp.13139-13144.

Han, A., Glanville, J., Hansmann, L. and Davis, M.M., 2014. Linking T-cell receptor sequence to functional phenotype at the single-cell level. Nature biotechnology, 32(7), p.684.

Birnbaum, M.E., Mendoza, J.L., Sethi, D.K., Dong, S., Glanville, J., Dobbins, J., Özkan, E., Davis, M.M., Wucherpfennig, K.W. and Garcia, K.C., 2014. Deconstructing the peptide-MHC specificity of T cell recognition. Cell, 157(5), pp.1073-1087.

Glanville, J., D'Angelo, S., Khan, T.A., Reddy, S.T., Naranjo, L., Ferrara, F. and Bradbury, A.R.M., 2015. Deep sequencing in library selection projects: what insight does it bring?. Current opinion in structural biology, 33, pp.146-160.

Jacob Glanville, Huang Huang, Allison Nau, Olivia Hatton, Lisa E. Wagar, Florian Rubelt, Xuhuai Ji, Arnold Han, Sheri M. Krams, Christina Pettus, Nikhil Haas,

Cecilia S. Lindestam Arlehamn, Alessandro Sette, Scott D. Boyd, Thomas J. Scriba, Olivia M. Martinez, and Mark M. Davis. Identifying specificity groups in the T-cell receptor repertoire. Nature 2017 (in press)

Gamma/delta convergence Wei, Yu-Ling, et al. "A highly focused antigen receptor repertoire characterizes γδ T cells that are poised to make IL-17 rapidly in naive animals." Frontiers in immunology 6 (2015).

Avnir, Y., Watson, C.T., Glanville, J., Peterson, E.C., Tallarico, A.S., Bennett, A.S., Qin, K., Fu, Y., Huang, C.Y., Beigel, J.H. and Breden, F., 2016. IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. Scientific reports, 6.

Ryan, J.F., Hovde, R., Glanville, J., Lyu, S.C., Ji, X., Gupta, S., Tibshirani, R.J., Jay, D.C., Boyd, S.D., Chinthrajah, R.S. and Davis, M.M., 2016. Successful immunotherapy induces previously unidentified allergen-specific CD4+ T-cell subsets. Proceedings of the National Academy of Sciences, 113(9), pp.E1286-E1295.

Watson, C.T., Glanville, J. and Marasco, W.A., 2017. The Individual and Population Genetics of Antibody Immunity. Trends in Immunology.

Glanville, Jacob, et al. "Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation." Proceedings of the National Academy of Sciences 108.50 (2011): 20066-20071.

Yeung, Y.A., Foletti, D., Deng, X., Abdiche, Y., Strop, P., Glanville, J., Pitts, S., Lindquist, K., Sundar, P.D., Sirota, M. and Hasa-Moreno, A., 2016. Germline-encoded neutralization of a Staphylococcus aureus virulence factor by the human antibody repertoire. Nature Communications, 7, p.13376.

Corey T. Watson, Frederick A. Matsen IV, Katherine J. L. Jackson, Ali Bashir, Melissa Laird Smith, Jacob Glanville, Felix Breden, Steven H. Kleinstein, Andrew M. Collins and Christian E. Busse J Immunol May 1, 2017, 198 (9) 3371-3373; DOI: https://doi.org/10.4049/jimmunol.1700306

HV Büdingen, T Kuo, S Marina, C Belle, L Apeltein, J Glanville et al. “B cell exchange across the blood-brain barrier in multiple sclerosis.” The Journal of Clinical Investigation 122.12 (2012): 4533.

Jackson, K.J., Liu, Y., Roskin, K.M., Glanville, J., Hoh, R.A., Seo, K., Marshall, E.L., Gurley, T.C., Moody, M.A., Haynes, B.F. and Walter, E.B., 2014. Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. Cell host & microbe, 16(1), pp.105-114.

Han, A., Newell, E.W., Glanville, J., Fernandez-Becker, N., Khosla, C., Chien, Y.H. and Davis, M.M., 2013. Dietary gluten triggers concomitant activation of CD4+ and CD8+ αβ T cells and γδ T cells in celiac disease. Proceedings of the National Academy of Sciences, 110(32), pp.13073-13078.

Levin, M., King, J.J., Glanville, J., Jackson, K.J., Looney, T.J., Hoh, R.A., Mari, A., Andersson, M., Greiff, L., Fire, A.Z. and Boyd, S.D., 2016. Persistence and evolution of allergen-specific IgE repertoires during subcutaneous specific immunotherapy. Journal of Allergy and Clinical Immunology, 137(5), pp.1535-1544.

Pham, T.D., Chng, M.H.Y., Roskin, K.M., Jackson, K.J.L., Nguyen, K.D., Glanville, J., Lee, J.Y., Engleman, E.G. and Boyd, S.D., 2017. High-fat diet induces systemic B-cell repertoire changes associated with insulin resistance. Mucosal immunology.

Benichou, J., Glanville, J., Prak, E.T.L., Azran, R., Kuo, T.C., Pons, J., Desmarais, C., Tsaban, L. and Louzoun, Y., 2013. The restricted DH gene reading frame usage in the expressed human antibody repertoire is selected based upon its amino acid content. The Journal of Immunology, 190(11), pp.5567-5577.

Liberman, G., Benichou, J.I., Maman, Y., Glanville, J., Alter, I. and Louzoun, Y., 2016. Estimate of within population incremental selection through branch imbalance in lineage trees. Nucleic acids research, 44(5), pp.e46-e46.

Steiniger, S.C., Glanville, J., Harris, D.W., Wilson, T.L., Ippolito, G.C. and Dunham, S.A., 2017. Comparative analysis of the feline immunoglobulin repertoire. Biologicals, 46, pp.81-87.

Mahon, C.M., Lambert, M.A., Glanville, J., Wade, J.M., Fennell, B.J., Krebs, M.R., Armellino, D., Yang, S., Liu, X., O'Sullivan, C.M. and Autin, B., 2013. Comprehensive interrogation of a minimalist synthetic CDR-H3 library and its ability to generate antibodies with therapeutic potential. Journal of molecular biology, 425(10), pp.1712-1730.

Zhai, W., Glanville, J., Fuhrmann, M., Mei, L., Ni, I., Sundar, P.D., Van Blarcom, T., Abdiche, Y., Lindquist, K., Strohner, R. and Telman, D., 2011. Synthetic antibodies designed on natural sequence landscapes. Journal of molecular biology, 412(1), pp.55-71.

Leighton, P.A., Schusser, B., Yi, H., Glanville, J. and Harriman, W., 2015. A diverse repertoire of human immunoglobulin variable genes in a chicken B cell line is generated by both gene conversion and somatic hypermutation. Frontiers in immunology, 6.