High-Throughput Methods for Characterizing the Immune Repertoire

by

Uri Laserson

BA Mathematics and Biology
New York University, 2005

Submitted to the

Medical Engineering and Medical Physics Program
Harvard-MIT Division of Health Sciences and Technology

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
Biomedical Engineering and Computational Biology

Massachusetts Institute of Technology

September 2012
[FEBRUARY 2013]

Signature of Author: _____

Certified by: _____

George M Church, PhD
Professor of Genetics
Harvard Medical School
Thesis Supervisor

Accepted by: _____

Arup Chakraborty, PhD
Director, Institute for Medical Engineering and Sciences and
the Harvard-MIT Program in Health Sciences and Technology
Robert T. Haslam Professor of Chemical Engineering,
Chemistry, and Biological Engineering

High-Throughput Methods for Characterizing the Immune Repertoire

by

Uri Laserson

Submitted to the Harvard-MIT Division of Health Sciences and Technology on 21 September 2012 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Biomedical Engineering and Applied Mathematics

ABSTRACT

The adaptive immune system is one of the primary mediators in almost every major human disease, including infections, cancer, autoimmunity, and inflammation-based disorders. It fundamentally functions as a molecular classifier, and stores a memory of its previous exposures. However, until recently, methods to unlock this information or to exploit its power in the form of new therapeutic antibodies or affinity reagents have been limited by the use of traditional, low-throughput technologies. In this thesis, we leverage recent advances in high-throughput DNA sequencing technology to develop new methods to characterize and probe the immune repertoire in unprecedented detail. We use this technology to 1) characterize the rapid dynamics of the immune repertoire in response to influenza vaccination, 2) characterize elite neutralizing antibodies to HIV, to better understand the constraints for designing an HIV vaccine, and 3) develop new methodologies for discovering autoantigens, and assaying large libraries of protein antigens in general. We hope that these projects will serve as stepping-stones towards filling the gap left by low-throughput methods in the development of antibody technologies.

Thesis Supervisor: George M Church

Title: Professor of Genetics, Harvard Medical School

Acknowledgments

Graduate school is a very human, social endeavor. The people I have come to know and rely on have made me feel like the luckiest person on earth, and while I am excited to move past graduate school, I am saddened to leave such an incredible environment. It's too difficult to describe the many ways that people have been a part of my life over my 7-year tenure in graduate school. They have been in my life as mentors, collaborators, colleagues, family, and friends. Of the literaly hundreds of people who have touched me during my time in school, here is a handful of those who have had a particularly strong influence.

I feel like one of the luckiest people that I got to spend >5 years with George Church. He gave me my first glimpse of real creativity. He believed that I could succeed in a completely new environment. He is endlessly supportive of any endeavor I would propose. And he has one of the greatest senses of humor. Every day I looked forward to coming to lab. My grandchildren's grandchildren will be telling stories about him.

Most of my lab work was performed in matrimony with Francois Vigneault, from whom I've learned an immense amount about how experiments actually get done.

Ben Larman has been one of my best friends since I met him in graduate school, and one of the most incredible scientists I have known. I am happy I had the opportunity to work with him so closely and share many a scotch.

The wonderful people I befriended and worked with in George's lab, including Marc Lajoie, Sri Kosuri, Billy Li, Harris Wang, Adrian Briggs, Nikolai Eroshenko, Jay Lee, Mike Sismour, Tara Gianoulis, Raj Chari, Dan Mandell, Mark Umbarger, Farren Isaacs, Noel Goddard, Morten Sommer, and Gautam Dantas, to name a few.

One of the most challenging but rewarding experiences I had was starting a company, and I am especially grateful to have done it with my partners Eric Boutin, Paris Wallace, Greg Porreca, and Bob Carpenter.

Two of my best friends, Michael Manapat and Yibo Ling, have been like my brothers during my time in graduate school.

And most importantly, Elizabeth Moran, without whom I probably could not function. In our (almost) decade together, I have become a better person because of her.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

I have had the privilege of sitting at the epicenter of a revolution in the way the life sciences are researched: the emergence of high-throughput DNA sequencing. Experiments that were laughably difficult at the start of my graduate training are now commonplace and unremarkable. This thesis details my efforts (along with a slew of incredible collaborators) to point this brand-new "microscope" toward the functioning of the adaptive immune system, a field that touches practically every biological/medical process in humans, and has been relatively slow to embrace high-throughput technologies.

1.1 The immune system is a pattern classifier

One of the primary functions of the immune system is to function as a molecular pattern classifier: discriminating between "self" and "non-self" and between "safe" and "dangerous" [1]. In order to be effective, the immune system must be able to 1) respond to an enormous diversity of molecular patterns and 2) respond in a timely manner to a changing array of molecular antigens. (Indeed, the ocean of pathogens trying to colonize your body is constantly trying to circumvent your immune system.) The number of possible molecular shapes on which your body must make life-or-death decisions is enormous; stored as digital data, this amount of information would easily surpass all the information stored in all the genomes of living individuals[1]. Yet remarkably, your body can respond to virtually any substance by using a small library of genetic components occupying less than 0.2% of your genome. This feat is accomplished by

[1] One arbitrary way to arrive at this calculation is the requirement to respond to every possible confirmation of every 5-mer sequence of amino acids. Say you store the backbone confirmation of every possible 5-mer peptide. This is equivalent to 8 ϕ and ψ angles, each stored with, say, 8 bits, along with the identities of the side chains (not even including rotamers), for which you need < 5 bits each: 89 bits per 5-mer. To encode every possible 5-mer, along with its tertiary structure would require 2^{89} bits, or 77 yottabytes, which could be stored in 309 Ybp of DNA at 2 bits per base [2].

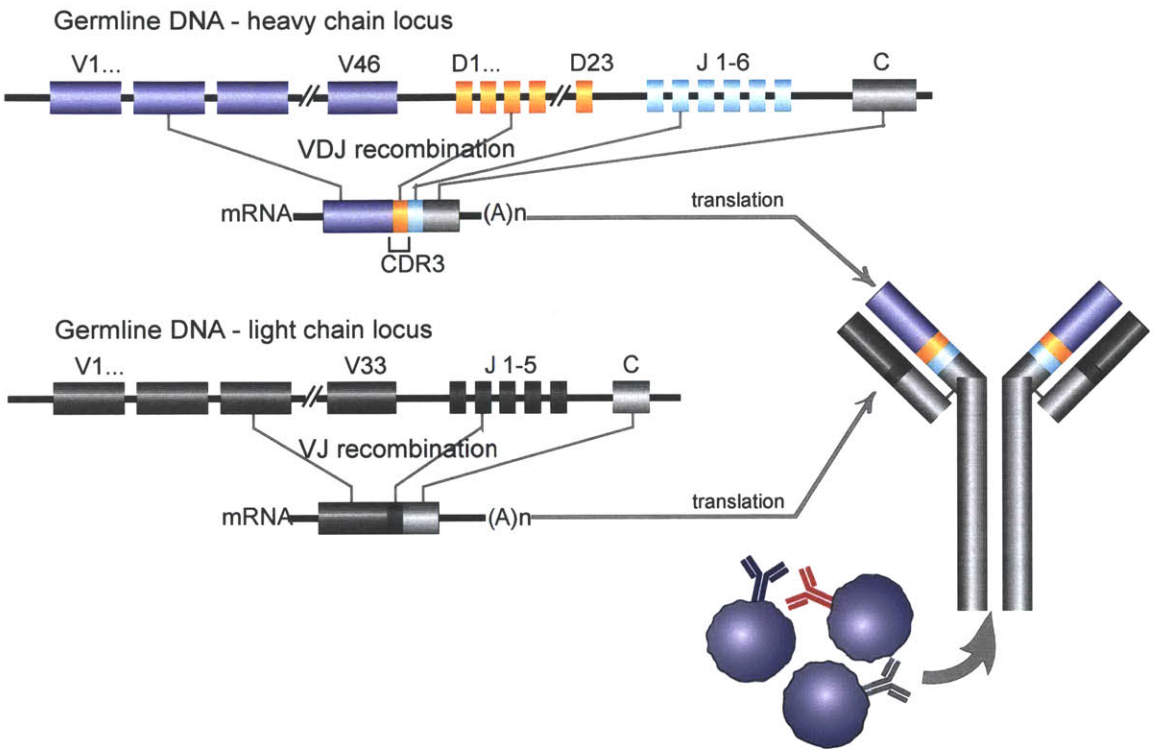


Figure 1.1: VDJ recombination

the combination of a molecular mechanism for generating a large diversity of molecules, along with evolutionary selective pressures at the population and somatic levels.

More concretely, the function of the adaptive immune system is largely mediated by a collection of lymphocytes (B and T cells) that each a express a unique, genetically-encoded receptor. In order to generate the repertoire of antibodies necessary for antigen recognition, each lymphocyte independently constructs a unique receptor through the process of VDJ recombination; each cell randomly selects a single V, D, and J gene segment through genetic recombination, introducing additional non-germline-encoded nucleotides at the junctions (Figure 1.1). This process creates the antibody diversity, the majority of which is encoded in the heavy chain complementarity determining region 3 (CDR3) [3].

1.2 The immune system stores a database of immune exposures

The immune system functions as a renewing, flowing, distributed computer (in contrast to the immune system, which operates on a fixed structure) [1]. Recognition of particular patterns is

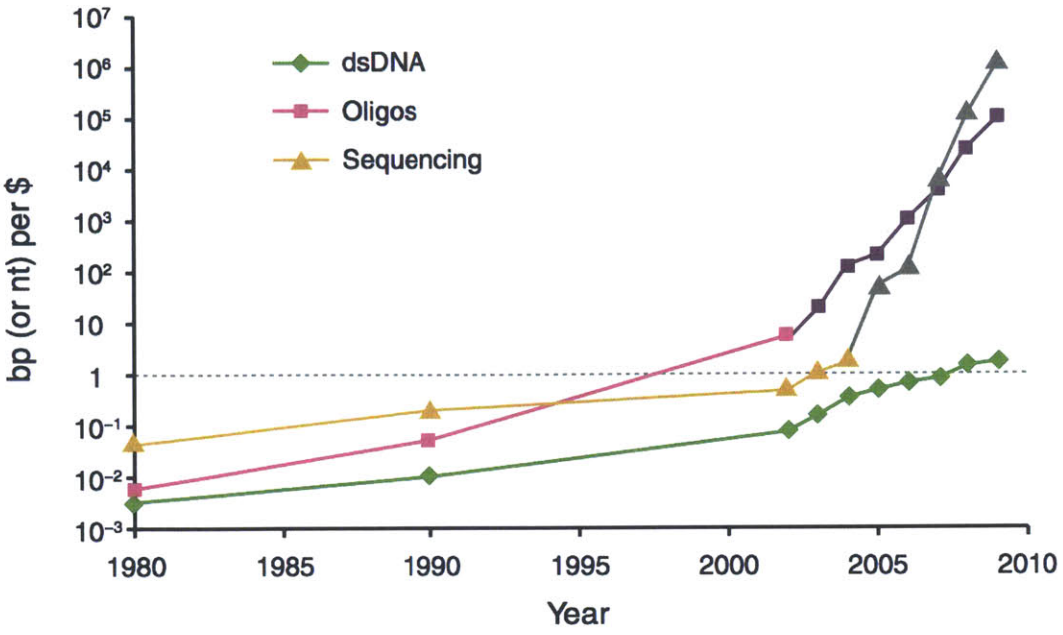


Figure 1.2: Exponential improvements in DNA sequencing. Taken from [8].

stored in individual cells that encode a protein that is a receptor for that pattern. While the supply of cells is constantly renewing itself (to maintain the ability to adapt to new threats), the immune system also stores those receptors that it deems useful and that were activated by their cognate antigens. Indeed, the memory compartment of lymphocytes contains the "fossil record" of exposures that each individual has experienced in his life [4]. This "database" of interactions is a treasure-trove of useful information. In principle, each individual carries with him the history of every disease he has had; the potential immunity he would have to new diseases; information on allergies; unique places to which he traveled; potential for autoimmunity; etc. However, because this information is stored as the sum total of millions of unique receptors, the technology to assay it did not exist until recently.

1.3 All DNA-encodable assays are now high-throughput

Since the release of the first high-throughput sequencing (HTS) platforms [5, 6], improvements in the technology have been surpassing Moore's law (Figure 1.2) [7, 8]. Because of the massive increase in data-generation capabilities, any assay that can be encoded in DNA now has a high-throughput instrument available, even dark matter detection [9].

1.4 High-throughput sequencing for the immune system

Indeed, immune repertoire sequencing is particularly well-suited to HTS, as antibodies and T cell receptors are inherently encoded in the genome. Of particular interest in our case is the relationship between the universe of antibodies and the universe of antigens. In principle, the space of possible antibodies and antigens is enormous, and it is of n^2 complexity; but luckily, the space of *interacting* pairs is thought to be sparse. Our overarching vision is to use DNA sequencing as a method to assay this space of antibody-antigen interactions in high-throughput. There has been a considerable amount of interest in applying HTS to the immune system in the last several years. However, much of that work has focused either on characterizing the immune system for its own sake or using the immune system to discover biomarkers for diseases (e.g., [10–12]). We have been particularly interested in developing methods for directly understanding how the immune system interacts with antigens. Indeed, such a capability would have significant implications for understanding the immune system, but also for designing vaccines, multiplexed diagnostics, and therapeutic discovery.

1.5 Summary of thesis work

In this thesis, I will describe four bodies of work approaching the antibody-antigen problem in three different ways (categorized using machine learning terminology).

1. *Unsupervised-learning.* In the first project, we attempt to characterize the functioning of the immune system to a controlled immune challenge, without using any of the cell-state/phenotype characterization techniques that much of immunology depends on and that is generally very low-throughput. We find evidence for an innate-adaptive spectrum in the antibody repertoire, and find that the immune system is generally incredibly dynamic at even the shortest time-scales.

2. *Supervised-learning.* As the immune system turns out to be very noisy, our next project used information about known antibodies against a known antigen: HIV. We used HTS to "fish" for variants of known, elite neutralizing HIV antibodies so that we can understand how they evolved and use the information to improve HIV vaccine design.

3. *Label-only.* The first two projects approach the antibody-antigen interaction problem from the antibody side. In the second two projects, we take the inverse approach, focused on antigens. Using a synthetic peptide library encoding the entire human peptidome, we interrogate the adaptive immune system by defining what functionalities it has, without ever determining the identities of the particular antibodies. This approach is also significant

because even raw results can lead to biological insight, since the identities of particular antigens are assayed from the start. We use the technique as a method to discover autoantigens for autoimmune diseases.

We conclude by presenting some proposed methods to directly assay interacting antibody-antigen pairs.

Chapter 2

High-Resolution Antibody Dynamics of Vaccine-Induced Immune Responses

2.1 Introduction

The immune system is able to rapidly sense and respond to a vast array of invading organisms. Its arsenal must contain components that are immediately effective against commonly-seen patterns (innate immunity) and components that are capable of responding to novel invaders (adaptive immunity). Given the acute nature and diversity of infections, the immune system must be capable of rapid excitement and contraction of a highly specific response. To achieve these goals, the immune system relies on a constantly-renewing, enormous library of antibody receptors while simultaneously storing the most useful ones (via memory cells) for rapid use when challenged by the same foreign molecules. This repertoire of immune receptors is genetically encoded in the somatically-modified genomes of billions of individual lymphocytes.

Currently, many immunology studies depend on characterizing cell-state markers (e.g., cell-surface receptors) and the ability to correlate them to encoded genetic information [13]. While it has been difficult to generate cell-state information in at large scales, recent advances in high-throughput sequencing (HTS) [7] have enabled any DNA-encodable assays to produce massive amounts of data. Indeed, HTS has enabled unprecedented views into the immune repertoire, as its diversity is naturally stored as genetically-encoded receptors among a complex collection of lymphocytes [11, 12, 14–16].

This study set out to dissect the rapid dynamics of the antibody response against a controlled immune challenge (vaccination), without the *a priori* notion of cell state markers or functions. We vaccinated three individuals a total of four times and banked blood samples at multiple time points before and after the vaccinations. Using the 454 sequencing platform, we analyzed the dynamic behavior of the immune repertoire in response to the vaccinations. We found that the

immune system is highly dynamic, and each vaccine response was qualitatively different. In contrast, we found that each individual uses the germline-encoded library of antibody components in very similar ways. Because the immune system is shaped by selective pressures at both the population and somatic levels [1], we observed that some germline components are geared towards innate action, while others are more likely to mutate and adapt to new challenges. Finally, we synthesized a collection of the strongest-responding clones one week after vaccination, and tested them for affinity against the vaccine antigens.

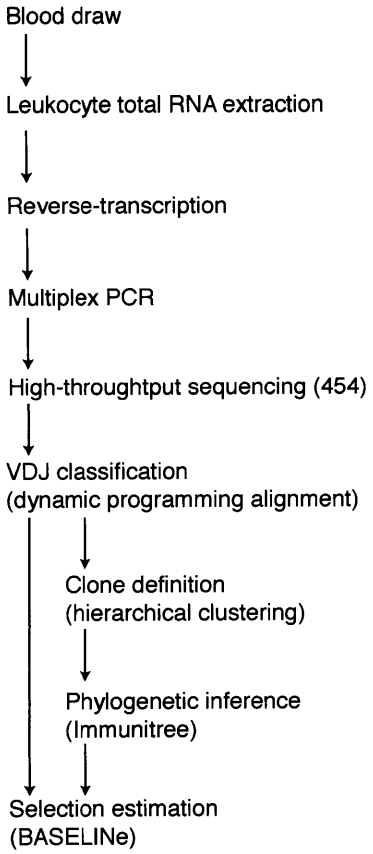
2.2 Results

2.2.1 Vaccination time-course design

We characterized the antibody repertoire of three Personal Genome Project (PGP) subjects, GMC, IB, and FV, in response to vaccination. In 2008, GMC was vaccinated against seasonal influenza, hepatitis A, and hepatitis B; in 2009, GMC, IB, and FV all received the seasonal flu vaccine. Blood samples were collected before and after the vaccination, as specified in 2.1. One of the goals of the study was to track the response of the immune system exclusively through genetic information. Therefore, lymphocytes were not sorted for particular subsets or activation states; total RNA from ficolled PBMCs was extracted and processed as described below. Each sequencing library of B cell antibody genes was generated using gene-specific reverse transcription and PCR. Each sample was uniquely bar-coded during the process and subjected to 454 sequencing and analysis.

2.2.2 Reproducibility and quantitation

Through the course of 7 runs of 454 sequencing, we obtained 4.3 million reads that successfully aligned to the IMGT germline reference database (Figure 2.1. Our initial experiments focused on characterizing the reproducibility of our library preparation method and calibrating our computational pipeline. We sequenced one library twice (generating sequencing replicates SR1 and SR2) and also sequenced an independent library from the same RNA sample (technical replicate TR1). Between these three sequencing runs, 477118 unique clones were identified of which only 3% were shared between the three runs and 14% were observed in at least two runs (Figure 2.2a). However, those shared clones accounted for 59% and 71% of all reads, suggesting that the highly expressed clones are actually sampled significantly between replicate runs. This was further validated by a strong correlation between technical replicate samples, confirming technical reproducibility of our approach (Figure 2.2b). Furthermore, resampling our data showed that 10^5 reads are sufficient to properly characterize a sample and obtain high



		Heavy GMC		IB		FV		Light GMC	
Raw reads		3 752 117		1 220 302		883 079		945 606	
Filtered reads*...		2 261 155	60%	1 008 912	83%	703 192	80%	348 810	37%
...with isotype		1 462 059		873 110		590 291		...by locus	
	IgM	658 154	45%	428 070	49%	174 874	30%	Igκ 185 891	53%
	IgG	273 057	19%	168 321	19%	174 692	30%	Igλ 162 919	47%
	IgA	415 102	28%	156 908	18%	206 089	35%		
	IgD	115 668	8%	119 587	14%	34 553	6%		
	IgE	78	0%	224	0%	83	0%		
...by time point									
2008 Oct 07, 12:00	−2 w	392 810	17%						
2008 Oct 21, 13:00	+1 h	220 473	10%						
2008 Oct 22, 12:00	+1 d	239 055	11%					131 777	38%
2008 Oct 24, 12:00	+3 d	160 865	7%						
2008 Oct 28, 12:00	+1 w	244 824	11%					99 518	29%
2008 Nov 04, 12:00	+2 w	198 288	9%						
2008 Nov 11, 12:00	+3 w	203 192	9%					117 449	34%
2008 Nov 18, 12:00	+4 w	99 632	4%						
2009 Dec 07, 12:00	−8 d	12 808	1%	287 570	29%	48 816	7%		
2009 Dec 13, 12:00	−2 d	47 186	2%	84 061	8%	103 505	15%		
2009 Dec 15, 11:00	−1 h	59 416	3%	77 580	8%	93 648	13%		
2009 Dec 15, 13:00	+1 h	54 841	2%	70 300	7%	95 609	14%		
2009 Dec 16, 12:00	+1 d	54 044	2%	69 660	7%	7 609	1%		
2009 Dec 18, 12:00	+3 d	28 666	1%	85 843	9%	48 387	7%		
2009 Dec 22, 12:00	+1 w	66 294	3%	172 661	17%	143 051	20%		
2009 Dec 29, 12:00	+2 w	9 056	0%	47 791	5%	67 145	10%		
2010 Jan 05, 12:00	+3 w	82 882	4%	62 477	6%	47 833	7%		
2010 Jan 12, 12:00	+4 w	86 823	4%	50 969	5%	47 589	7%		
Number of clones...		725 202		526 838		174 593		4 598	
...with ≥3 reads		91 672		41 459		20 881		1 525	
...with ≥10000 reads		13		1		4		6	
...seen in ≥2 time points		58 941		12 569		11 850		771	
...seen in all time points		98		87		72		389	

*Size-selected, VJ-filtered

Figure 2.1: Overview of vaccination experiment. The experimental/computational pipeline is summarized on the left. The resulting sequencing data is summarized on the right, broken down by isotypes, time points, individual, and clone.

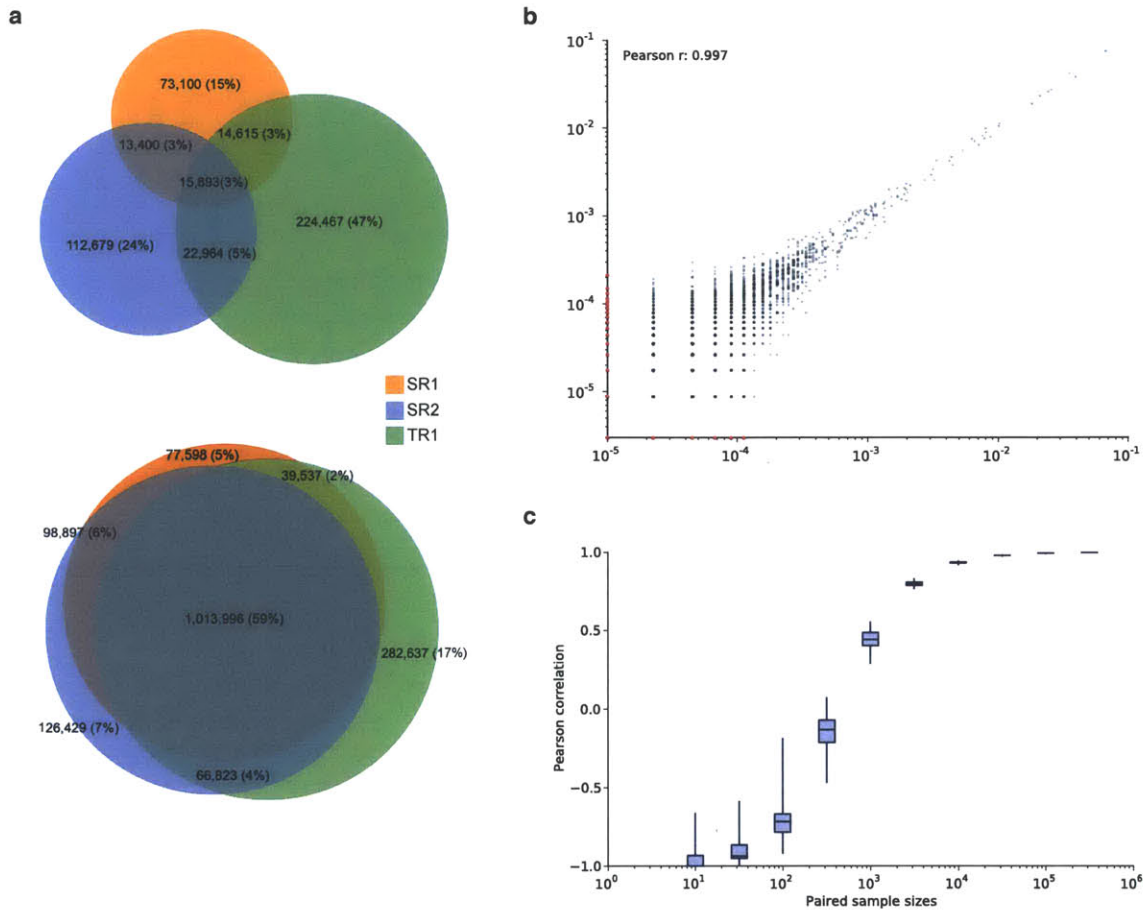


Figure 2.2: Reproducibility in vaccination experiment. (a) Venn diagrams showing overlapping clones (top) and the same overlaps weighted by number of reads (bottom). (b) Correlation between technical replicates. Axis scales are clone frequencies; red points are zero-valued on that axis. (c) Correlation between paired random samplings of reads of the indicated size.

correlations between replicates (Figure 2.2c).

2.2.3 Characteristics of the static heavy chain repertoire

Overall usage of the individual V and J components was highly non-uniform within a given individual (Figure 2.3). The most frequently observed V segments were IGHV3-23 (11% of all reads), IGHV3-30 (8%), IGHV4-59 (7%), and IGHV1-69 (6%) while the most frequent J segments were IGHJ4 (41%) and IGHJ6 (31%), consistent with previous studies [16, 17]. Nevertheless, utilization of the germline-encoded VDJ gene library was quite similar between individuals and across time. Indeed, the Spearman correlation between VJ-usage vectors was consistently high across time points and individuals (Figure 2.4) and VJ-usage time series are

remarkably stable as well (Figure 2.5). Finally, we built a neighbor-joining tree using the V-usage vectors of each of the samples to see how V-usage is structured. For the most part, V-usage clustered first by individual, and then by isotype, implying that while V usage is grossly similar across individuals, each individual still has a unique signature (Figure 2.6).

For a majority of the reads, we were able to genetically discern the antibody isotype. We found that IgM antibodies were the most abundant (43% of all reads), followed by IgA (27%), IgG (21%), IgD (9%), and IgE (0.01%) (Figure 2.1). However, the isotype usage varied significantly between time points (Figure 2.7).

Mutation levels were also measured across each of the reads. As expected, mutation rates were higher in the CDR regions of the antibodies, and were much higher in IgG and IgA antibodies (Figure 2.8). We further processed our reads through the BASELINe pipeline that estimates selection pressure on the antibodies [18]. Framework regions (FWR) were universally negatively selected, while CDR regions showed either neutral to slightly negative selection on average; however, CDR selection values were always more positive than FWR selection values (Figure 2.9).

The CDR3 length distribution we observed was consistent with both TCR data [19] as well as IMGT/LIGM data [20] (Figure 2.10). The 5th and 95th percentiles of the observed CDR3 lengths are 36 nt and 75 nt with median length 54 nt (with longest observed CDR3 at 140 nt).

Antibodies can be present at vastly different quantities, depending on cell types and whether they have been activated and are proliferating. Because the VDJ recombination process introduces so much diversity, the CDR3 sequence effectively functions as a natural barcode for a particular clone [21]. To functionally define antibody clones, we perform clustering of the CDR3 sequences and define two reads as derived from the same clone if their CDR3 sequences are highly similar, since it is unlikely that two independent B cells will generate the same nucleotide sequences. In total, we observe $>$1.4M clones across all of our data; however, only 150k clones had at least 3 reads each and only 24 clones with $>$10k reads each. Separately, approximately 84k clones were seen in two separate time points, while only 257 heavy-chain clones were seen in every time point (for a given individual)(See Figure 2.1 for more information).

We also found that the propensity for a clone to become activated (estimated assuming a binomial distribution) is reproducibly biased by VJ usage (Figure 2.11). The V regions most likely to become activated are dominated by IGHV4- and IGHV5-family genes, and the three individuals have highly correlated biases in the VJ-activation probabilities (Spearman correlation of 0.7). Nevertheless, we find that there is virtually no correlation between whether a VJ combination is likely to become activated and whether it is highly used (Figure 2.12). Taken together, this provides evidence that the antibody repertoire is shaped by selective forces at both population and somatic timescales, and individual antibodies occupy their own innate-adaptive

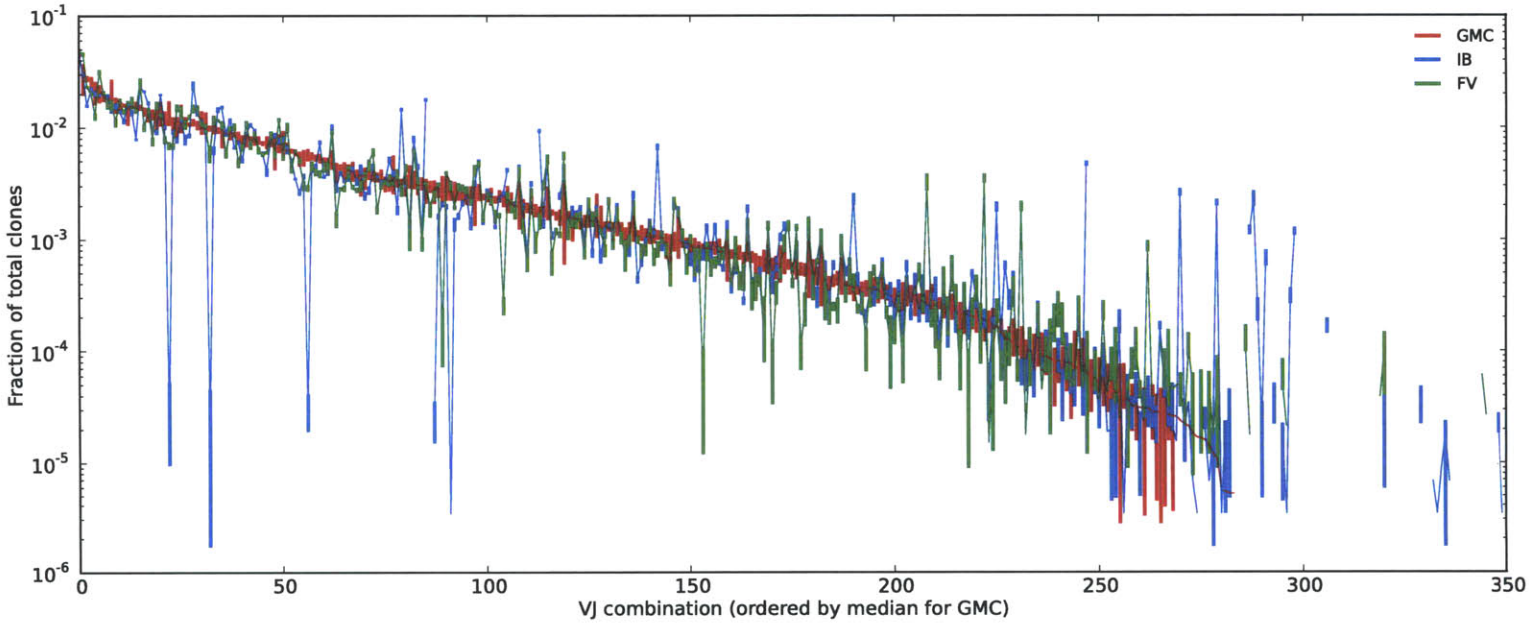


Figure 2.3: VJ usage. For each sample from each individual, the number of clones with a particular VJ combination was computed. The values are not weighted by reads, so that each count represent a single recombination event. For each possible VJ combination, a distribution of frequencies was computed for each individual. Each bar represents the 25th–75th percentile value across the different sample, while the line tracks the median values. The VJ combination are ordered by median for GMC.

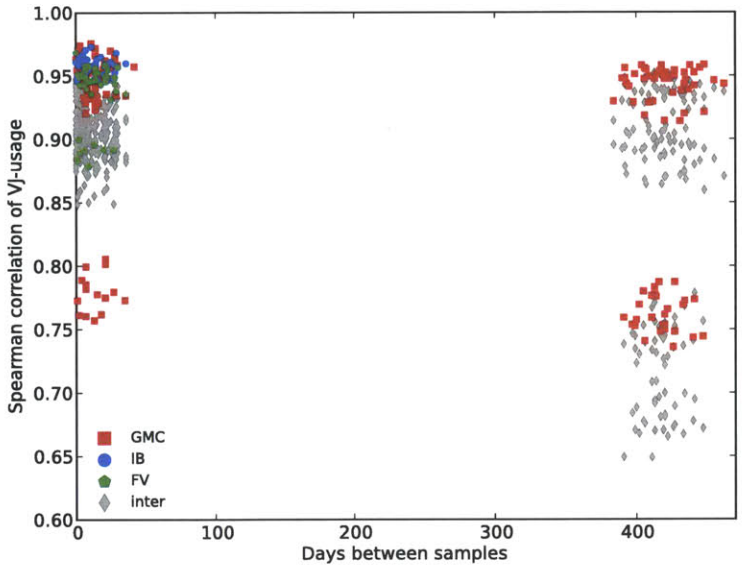


Figure 2.4: For each sample, a VJ-usage vector is formed. The Spearman correlation is computed between every pair of vectors; intra-individual comparisons are shown with the indicated color, while inter-individual comparisons are shown in gray. Note that intra- and inter-individual comparisons both show comparable correlations.

spectrum [1]. More precisely, utilization of the VDJ germline library may be optimized for naive interactions with common pathogens at the population scale, while the propensity of any given germline gene to somatically mutate may be optimized for the evolvability of the target organisms.

2.2.4 Antibody repertoire dynamics

In the hope of capturing at least one immunological event, we coordinated our experiments with clinically-indicated vaccinations. Each individual was given the seasonal flu vaccine, and GMC was also given boosters to hepatitis A/B in 2008. None of the subjects were naive to the antigens at the time of vaccination (through either prior vaccination or infection). Each read was assigned to a clone and a timepoint, allowing us to compute time series. The clone frequencies were tracked across all 38 time points to produce >20M clone-frequency measurements. In contrast to the relative stability of the VJ usage, antibody clones were highly dynamic and variable across individuals (Figures 2.13 and 2.14).

Responses to each of the four vaccination events were qualitatively different: IB produced a "textbook" response with large proliferating clones 7 days after vaccination; FV was likely

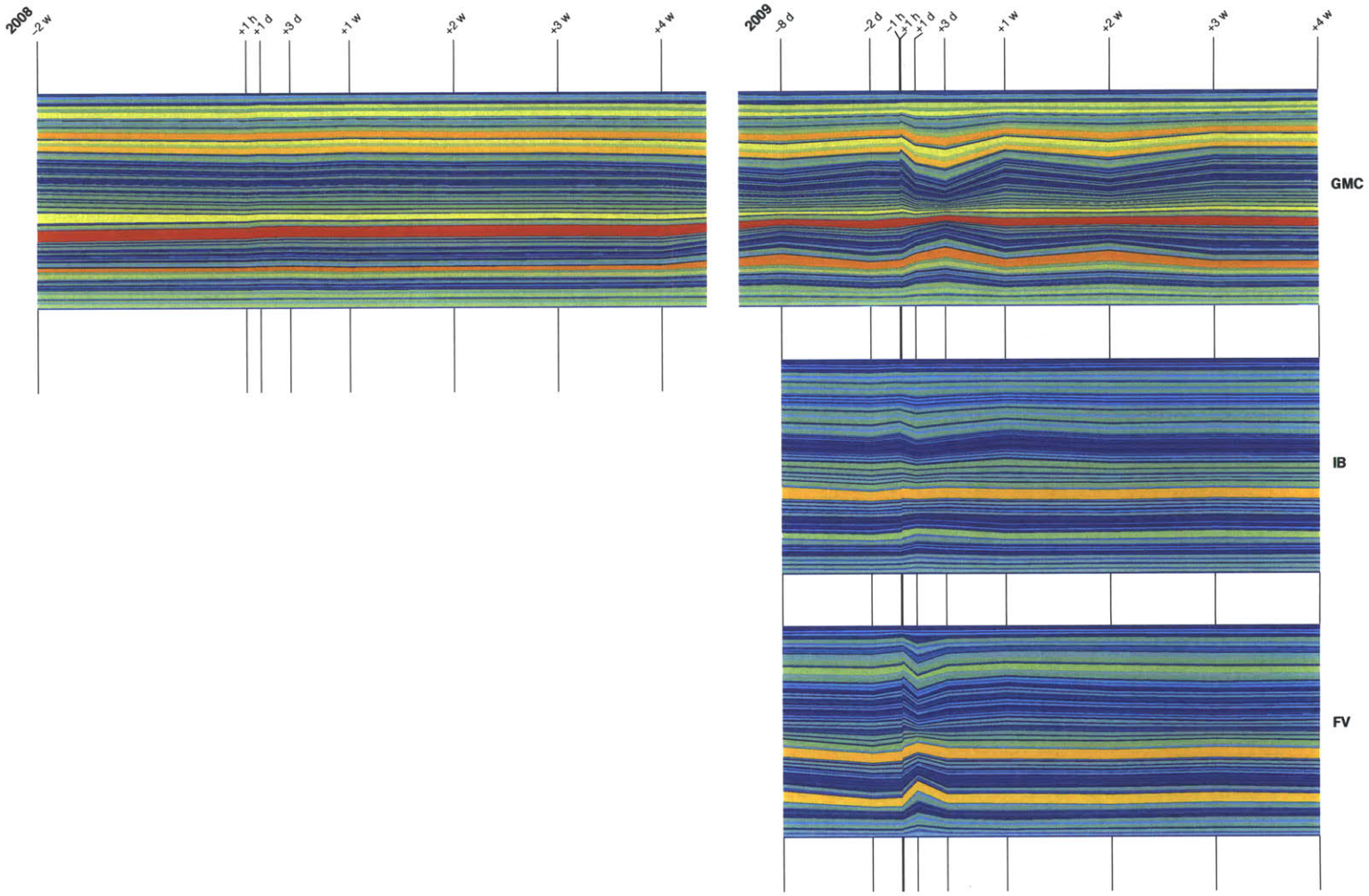


Figure 2.5: VJ-usage dynamics. Streamgraph of VJ-usage for each individual. Time is listed on the x-axis, with time points relative to vaccinations marked at the grid lines. Each stream/layer corresponds to a particular VJ combination, and its thickness at a given time point is proportional to its frequency at that time point. All streams add up to 100% usage at each time point.

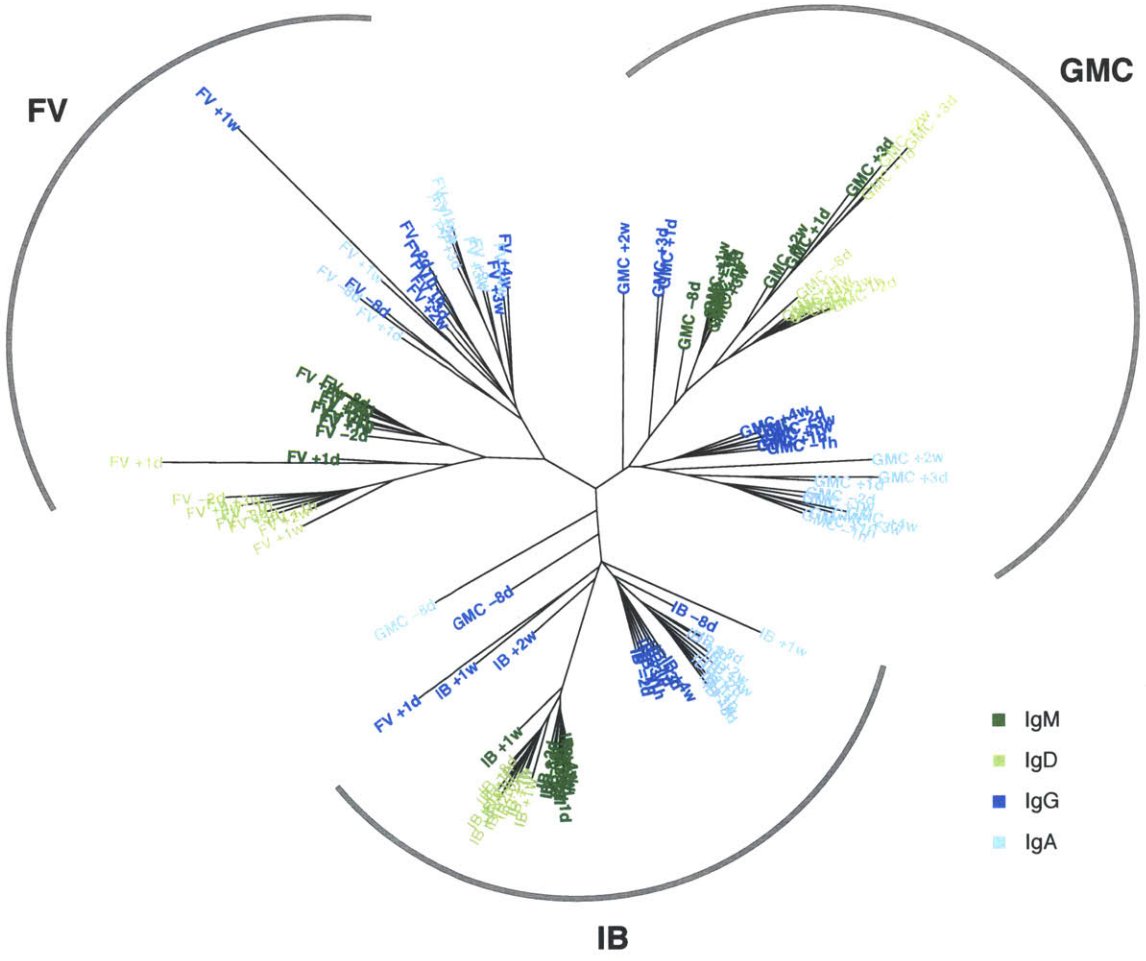


Figure 2.6: NJ tree of VJ-usage vectors. VJ-usage vectors are calculated for each individual-isotype combination, and clustered using the neighbor-joining algorithm. Each isotype is colored according to the legend. The tree naturally clusters by individual and then by isotype.

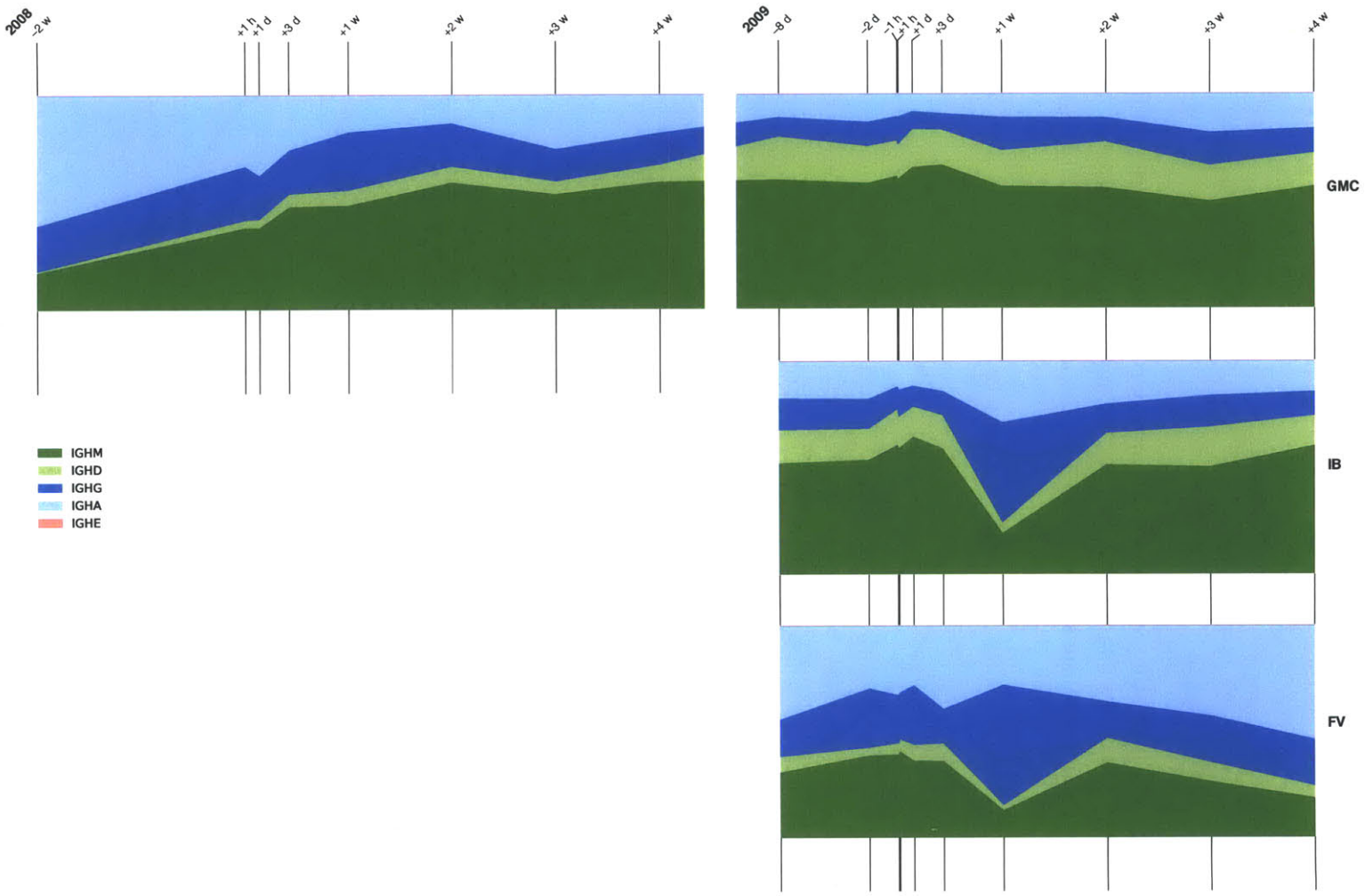


Figure 2.7: Isotype dynamics. Streamgraph of isotype usage at each timepoint for each individual.

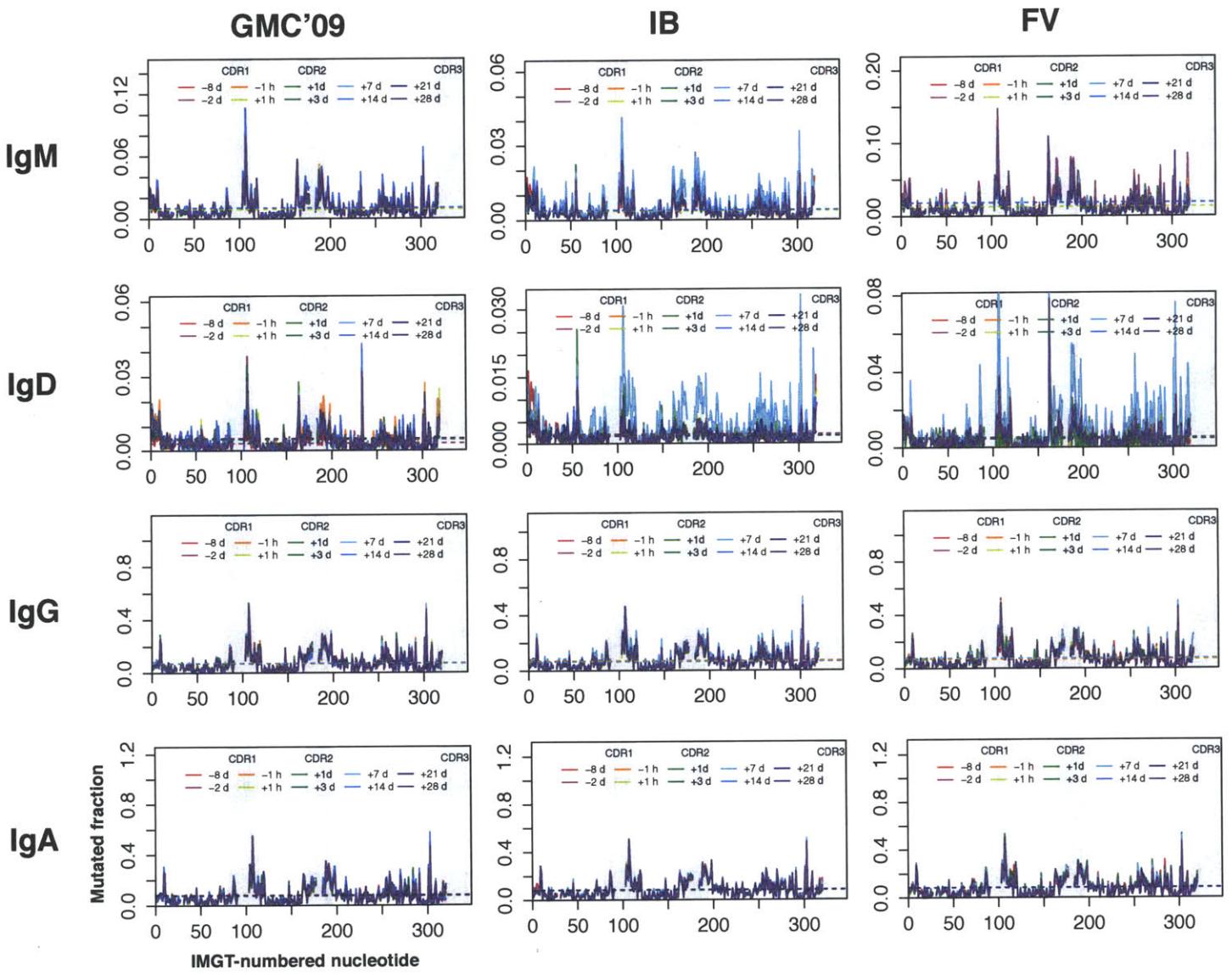


Figure 2.8: Antibody mutation patterns. The mutation density for the indicated subset of reads is computed along the length of the gene. Note that different scales are used for different suplots.

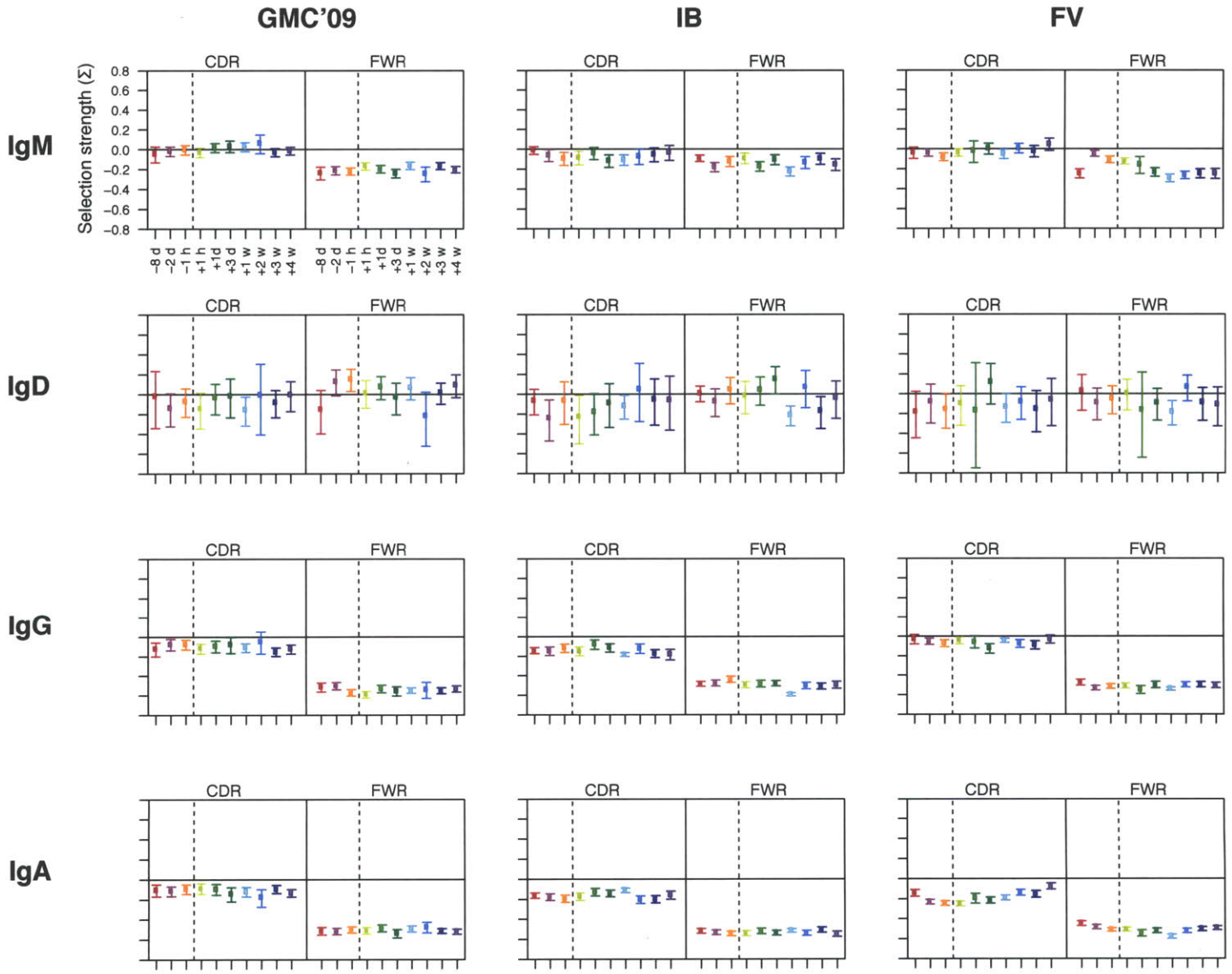


Figure 2.9: Antibody selection estimation. For each set of antibody sequences, the selection pressure has been estimated with BASELINe [18].

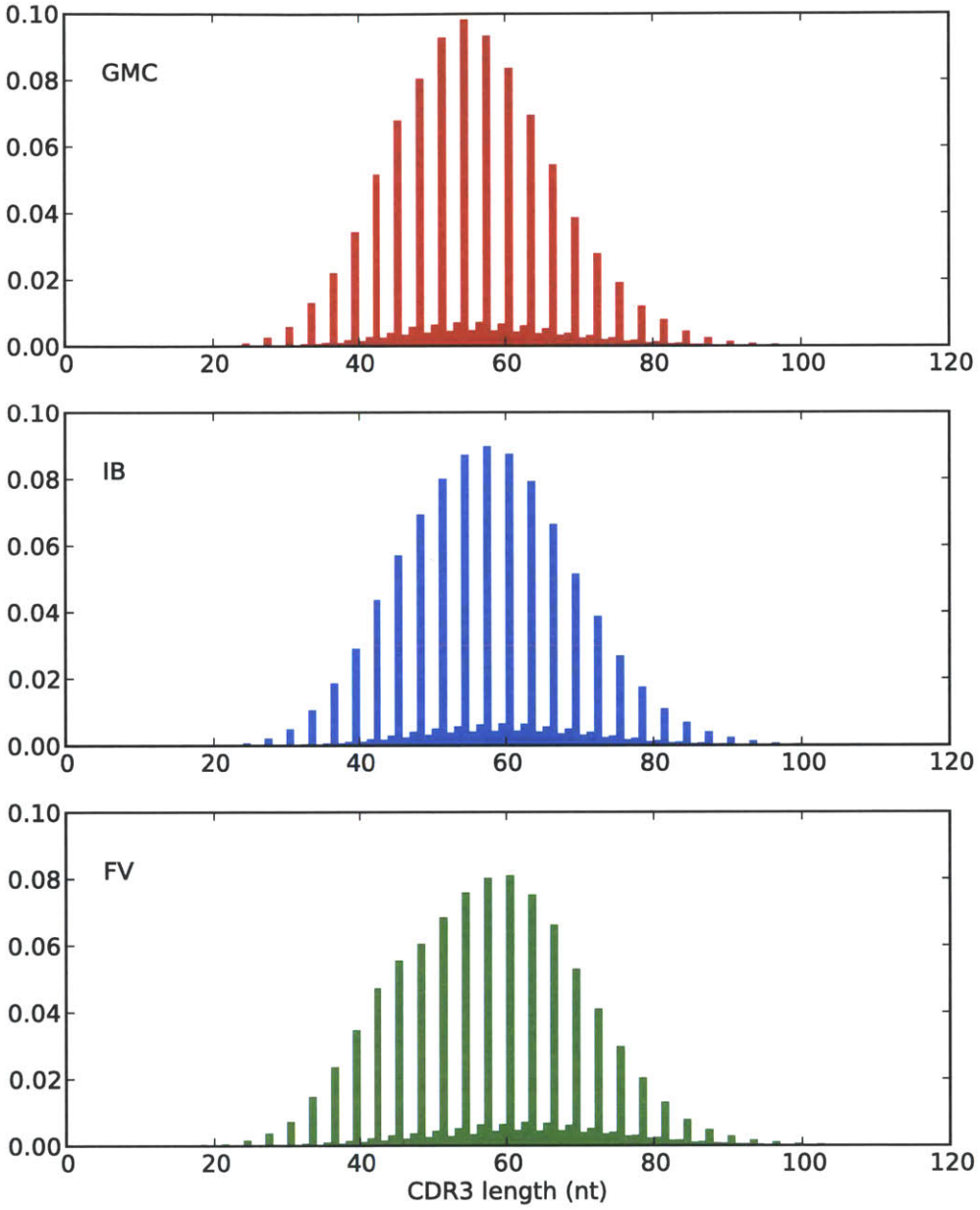


Figure 2.10: CDR3 length distributions. The CDR3 is defined according to the IMGT, as the segment spanning the second conserved cysteine through the conserved tryptophan.

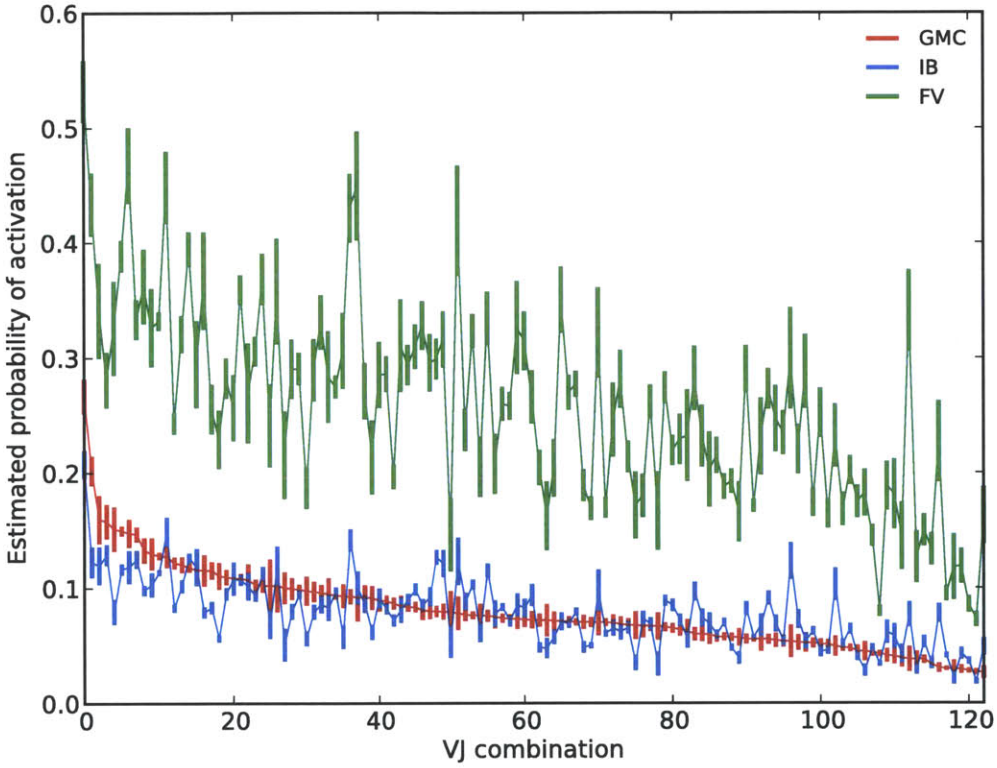


Figure 2.11: Probability of clone activation by VJ-usage. Each clone is labeled according to its VJ usage and whether it is "naive" (IgM-only with low mutation) or "activated" (IgG or IgA with high mutation). The probability of observing naive or activated clones is estimated assuming a binomial distribution. The line plots the estimated probability of activation, while the bars represent ± 1 standard deviation.

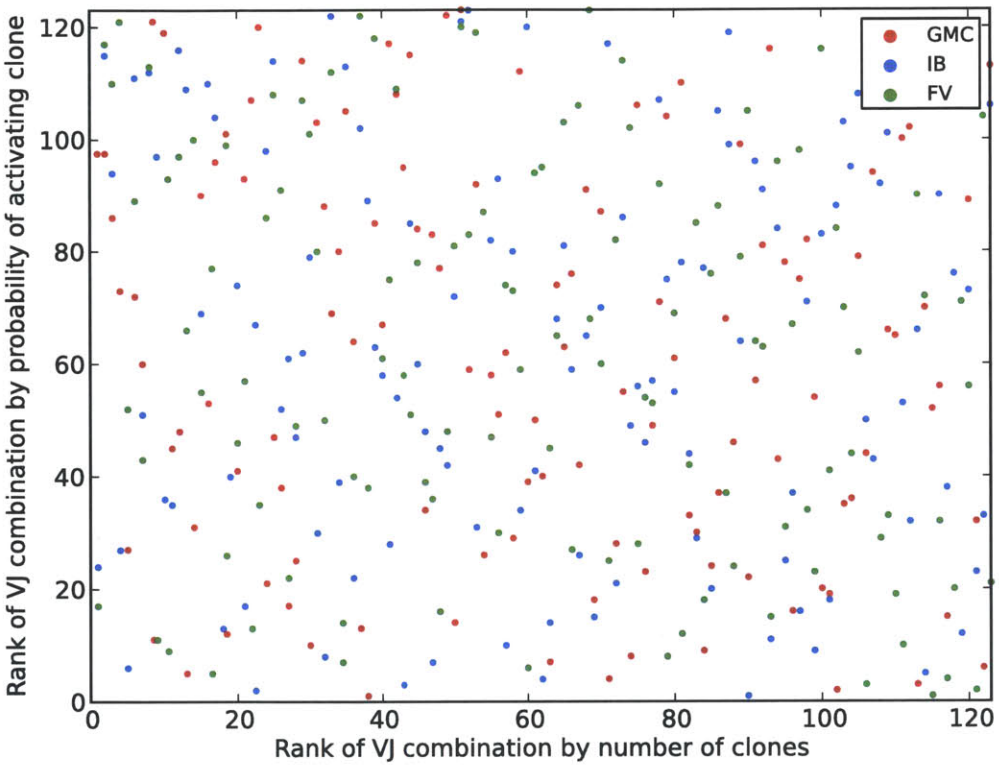


Figure 2.12: Probability of clone activation versus VJ-usage. For each VJ combination, we plot its VJ-usage frequency against its probability of becoming activated. This is filtered on VJ combinations for which we obtain at least 100 clones that are classifiable as "naive" or "activated".

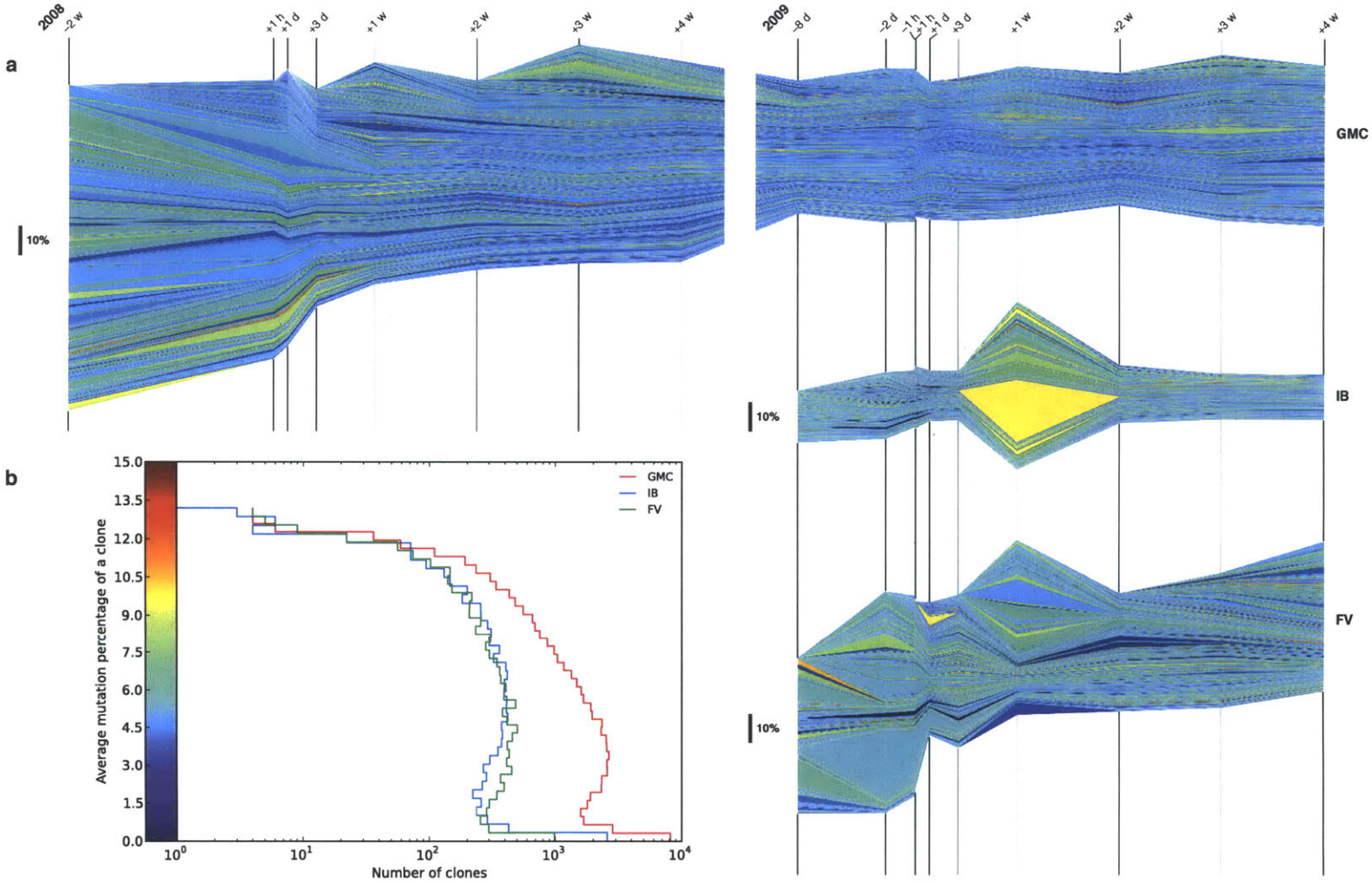


Figure 2.13: Vaccination clone dynamics colored by mutation. (a) Each layer represents a clone. Time is shown by the grid lines on the x-axis, and labeled relative to the two vaccination events. The thickness of each layer is proportional to the frequency of that clone at that time point. Each clone is colored based on the average mutation level of the corresponding reads (see colorbar for part (b)). Only clones seen in at least two time points are shown here. (b) Histogram of the average mutation level of all the clones. Each clone is counted once (i.e., clones are not weighted by the number of corresponding reads).

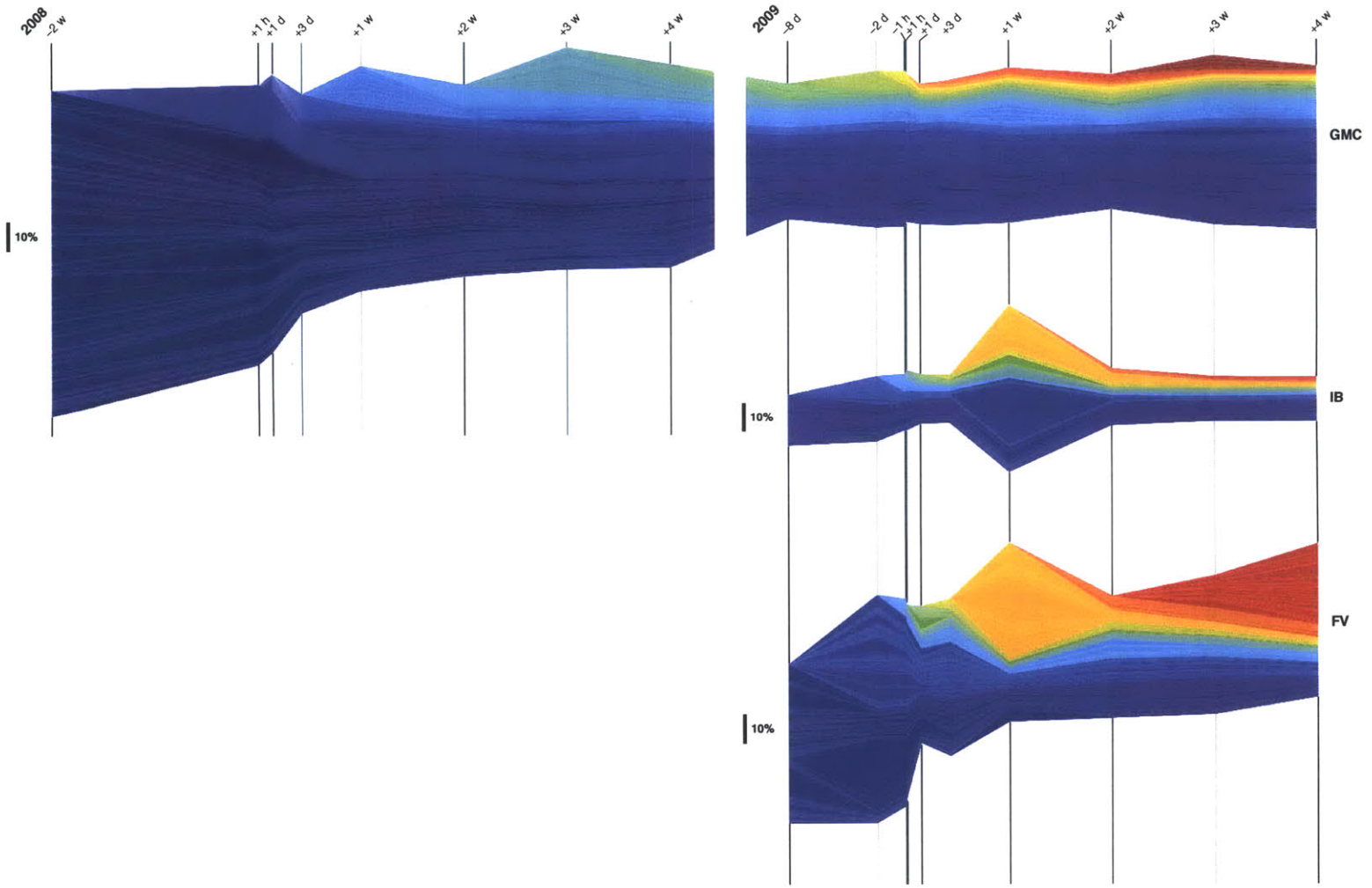


Figure 2.14: Vaccination clone dynamics colored by onset time. Same as Figure 2.13, except clones are colored by onset time. Onset times are ordered spectrally, so that all clones seen in the first time points are blue, followed by cyan, etc.

responding to some other immune challenge prior to vaccination but still responded with large clones 7 days post-vaccination; GMC was also responding to something prior to his first vaccination, with no strong response afterwards, while his second vaccination appears to have produced no significant responses.

We verified that samples that are closer in time share more unique clones. We computed the number of shared CDR3 sequences between all 703 possible pairs of samples across all 38 time points, and observed that closer time points within an individual indeed share a larger number of unique CDR3s (Figure 2.15). Consistent with this, inter-individual comparisons between time points showed very little CDR3 overlap.

We also quantified the range of dynamic behavior of the clones, finding that clones generally fluctuate wildly (Figures 2.16). Interestingly, each individual had a number of clones that were present at every time point sampled, including the samples separated by over a year (257 clones total; Figure 2.17). It is possible that these clones are chronically responding to antigens (foreign or auto) that are always present; indeed, these clones include sequences that are highly mutated (Figure 2.17b).

2.2.5 Clone analysis

It is a commonly accepted that expanding clone populations should arise from an immune challenge about 7 days after flu vaccination [22]. Therefore, we picked a subset of the largest clones from multiple time points before and after vaccination (-2 day, +7 days, +21 days), and synthesized, expressed and panned them by phage display. We were surprised to find very few strong binders against the vaccine hemagglutinin antigens.

Interestingly, even though GMC showed no significant response in 2009, the strongest binder (GMC J-065) was found in his day 7 response of that year. We then applied the Immunitree algorithm on clone GMC J-065 to infer the most likely evolutionary pathway [23]. The tree was also overlayed with selection values estimated using the BASELINe algorithm [18] as well as mutation levels (Figure 2.18). As expected, most nodes in the tree displayed significant negative selection in the FWRs, while some of the nodes show significant positive selection in the CDRs. We are currently in the process of analyzing clones of these trees that are more evolved and show signs of greater selection pressure.

2.3 Discussion

In this study, we generated the first high-throughput profile of the short-timescale dynamics of the antibody heavy chain repertoire. For proper function, the immune system requires the ability to rapidly expand and contract, and such highly dynamic behavior is consistent with

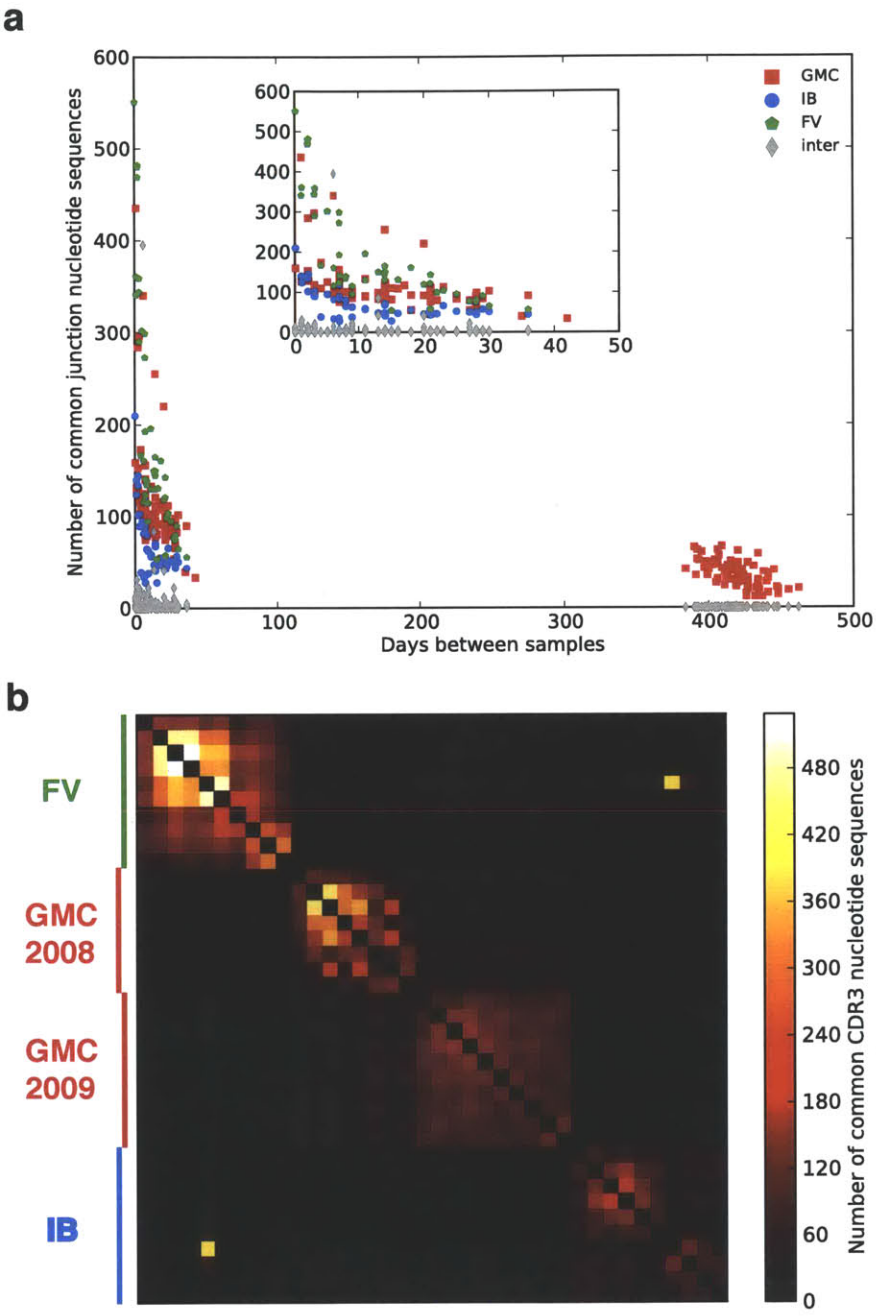


Figure 2.15: Inter-sample CDR3 overlaps. (a) Subsampled CDR3s from each sample are compared for common sequences. Comparisons of time points that are closer in time show higher levels of overlap. Inter-individual comparisons show very little overlap, as expected. (b) Overlap between each sample is plotted showing the three individual blocks. The strong overlap between an IB and an FV sample is likely due to some sample cross-contamination.

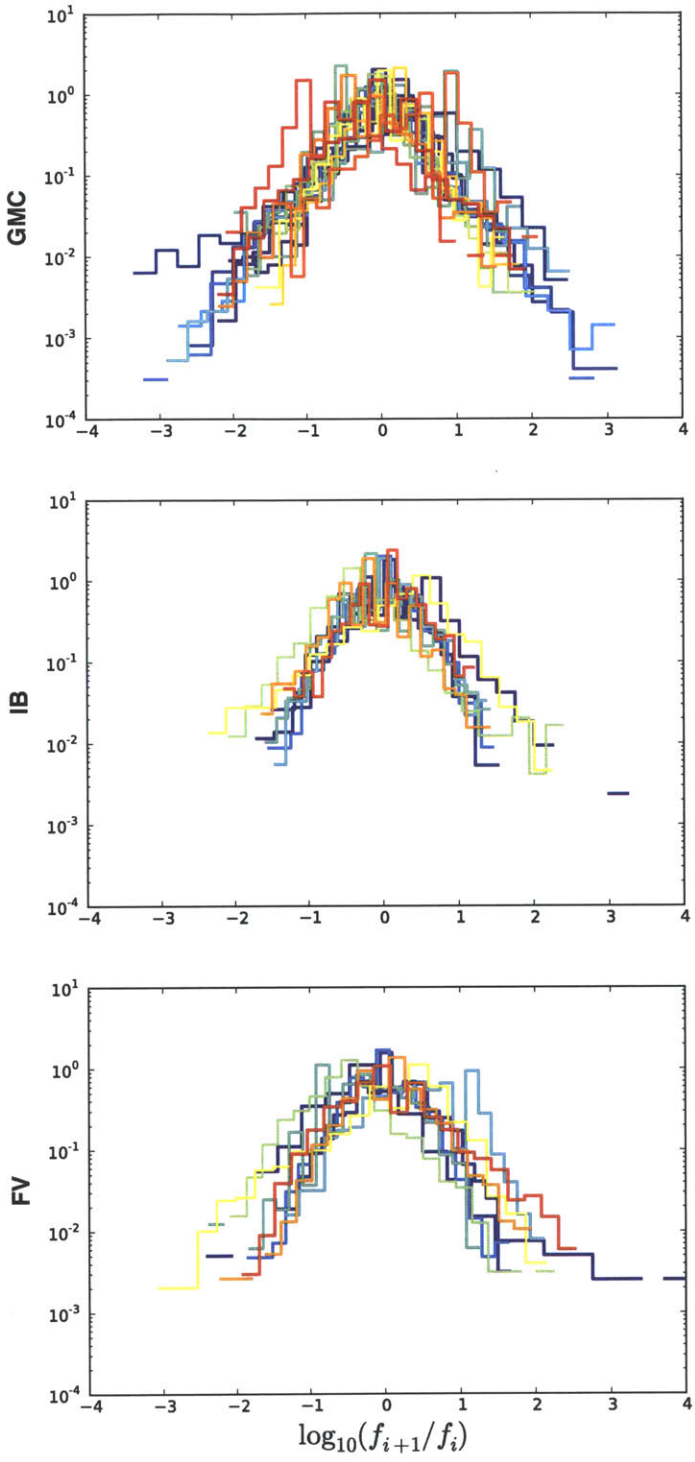


Figure 2.16: Distribution of frequency changes. For each adjacent time point, the $\log_1 0$ ratio of frequencies (f_i) for each clone is histogrammed, when finite. Time points are plotted with different colors, arranged chronologically and spectrally (blue to red).

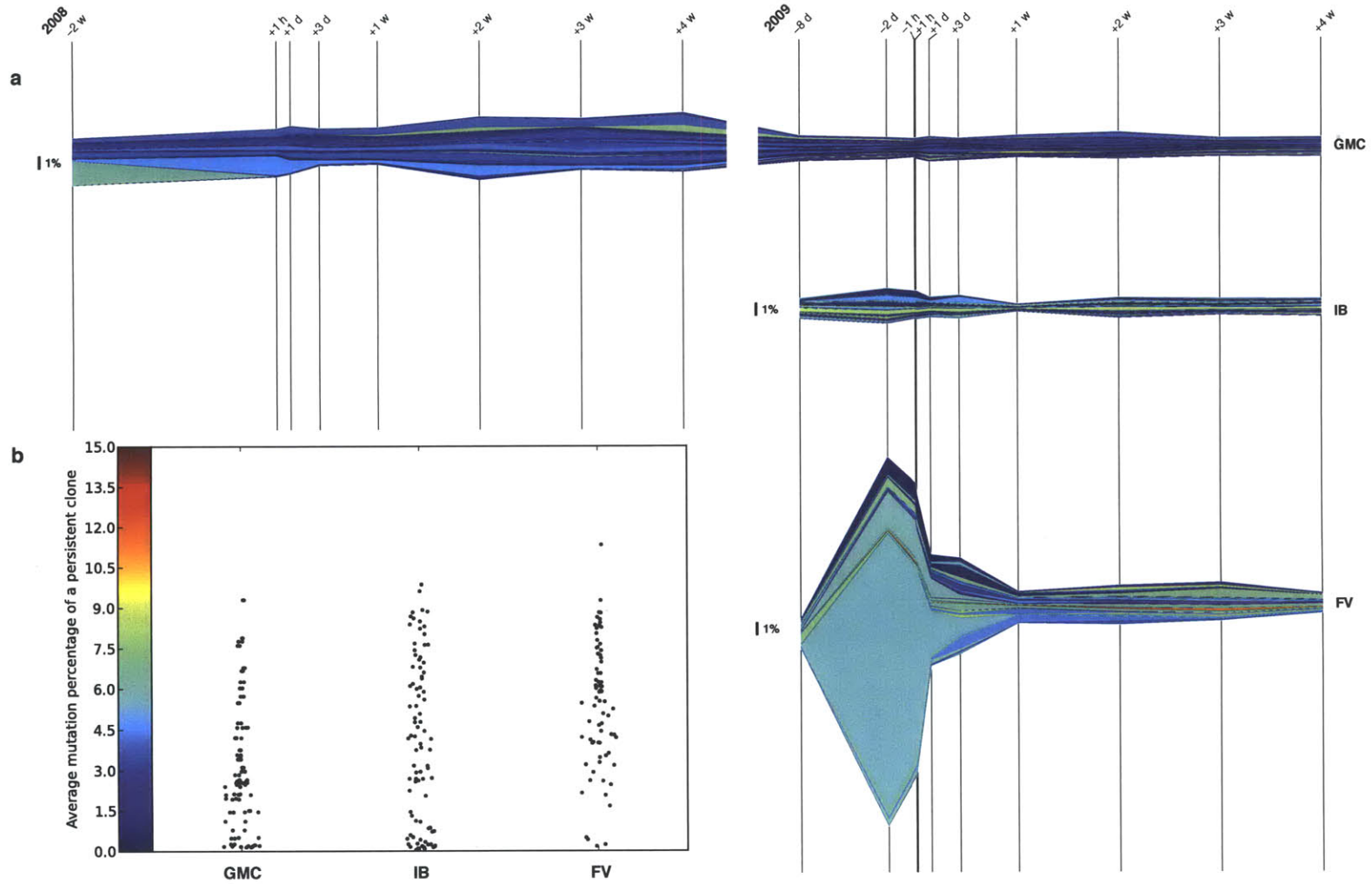


Figure 2.17: Dynamics of persistent clones. (a) Streamgraphs only of clones that are observed in every single time point for a given individual. They are colored as in Figure 2.13. (b) Distribution of average mutation level of the persistent clones.

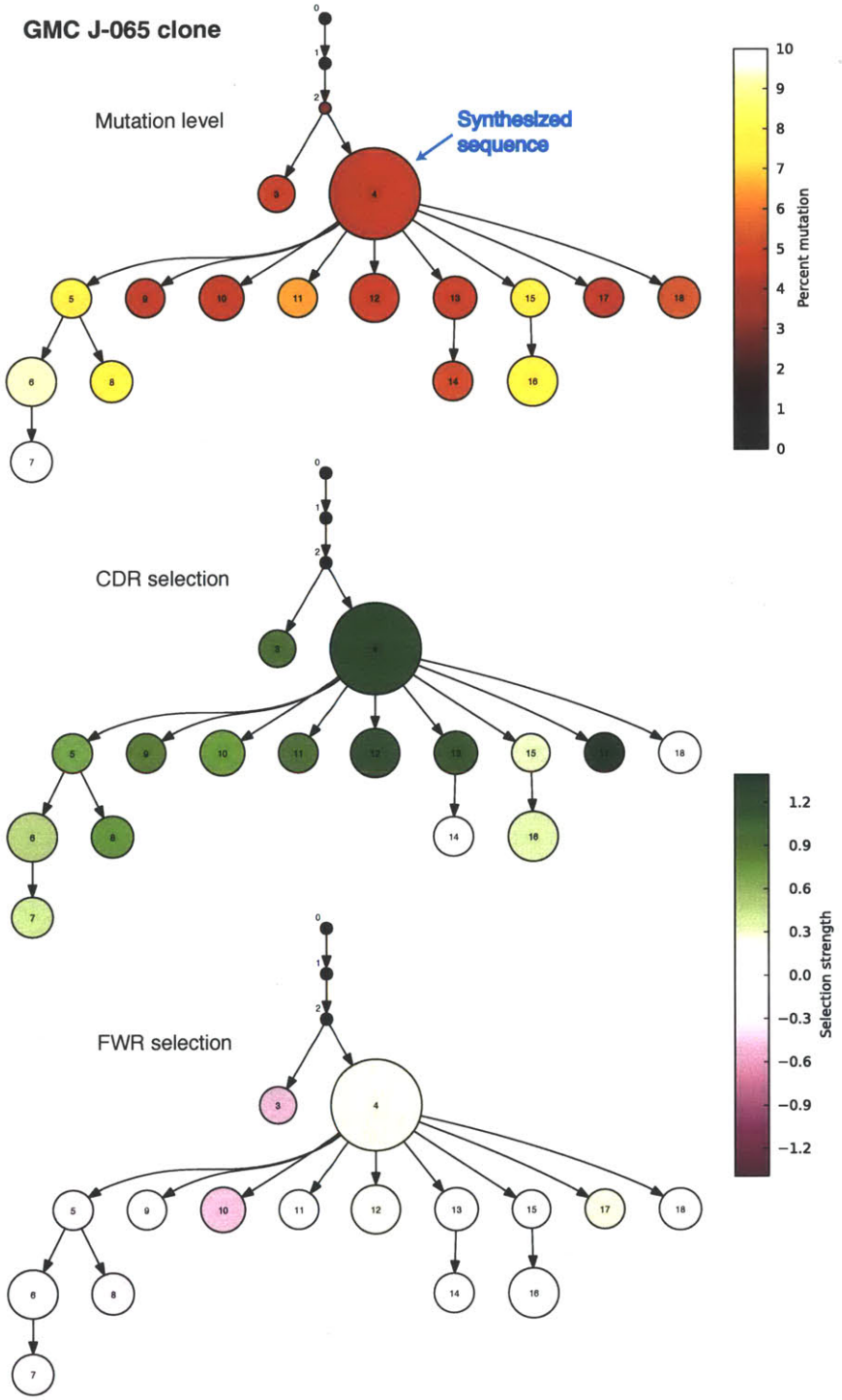


Figure 2.18: GMC J-065 clonal phylogeny. A phylogenetic tree for GMC clone J-065 was constructed with Immunitree and overlayed with sequence mutation data and CDR/FWR selection estimates.

our observations. We also found evidence that even the adaptive immune system (antibody repertoire) functions on an innate-adaptive spectrum, where usage of the germline antibody VDJ library is simultaneously shaped by population selection and somatic selection pressures. Indeed, it is apparent that utilization of the germline library is strongly stereotyped between individuals, but particular clones are highly dynamic.

While we were able to glean significant insights into the immune system from genetics alone, it appears that using the information for predictive purposes still requires a significantly greater amount of data [24]. Analogous to the dichotomy between supervised and unsupervised learning in statistics, we have attempted to understand the dynamics of the immune response using exclusively genetic information (high-throughput) without the limitation of functionally labeling our data (low-throughput). We hope that such approach will eventually enable the analysis of immune functions and also mining the "fossil record" [4] of individual antigen exposures.

While we have thus far not been able to realize this vision, we believe that this study represents a necessary milestone in a collective effort for the development of new tools to harness the full potential of the immune system. To that extent, we are focusing on developing methodologies for high-throughput capture of paired heavy and light chain sequences from single cells. Coupled with significant advances in DNA synthesis technology [8, 25], we should soon be able to assay a large immune repertoire against large, synthetic library of antigens (autoantigens, allergens, infectious agents, etc.) [26–28]. Doing so will further the development of immune repertoire profiling and facilitate our progress towards the next-generation of diagnostics, vaccines, and personalized therapeutic discovery.

2.4 Materials and Methods

2.4.1 Sample collection

Blood samples were collected under the approval of the Personal Genome Project [29]. Sample collection was coordinated with clinically indicated vaccinations for each individual. Total RNA was immediately extracted from each blood sample and stored at -80 until use.

2.4.2 Primer design

All oligonucleotides where ordered from Integrated DNA Technologies (IDT, Coralville, IA). For the design of the upstream variable-region oligonucleotides (IGHV-PCR), we extracted the L-PART1 and L-PART2 sequences from all IMGT/GENE-DB [30] reference segments annotated as "functional" or "ORF". These two segments are spliced together in vivo to form the leader sequence. We positioned our primer sequence to cross the exon-exon boundary to ensure

amplification from cDNA rather than gDNA. For the design of the downstream constant-region oligonucleotides (IGHC-RT and IGHC-PCR), the first 100 nucleotides of the CH1 exon were extracted from the IMGT/GENE-DB. Oligonucleotides were then selected as close as possible to the 5' end of the C-region to take advantage of sequence conservation between different variants, and to ensure that isotypes would be distinguishable.

2.4.3 Sequencing library preparation

We reverse-transcribed the immunoglobulin heavy chain mRNA using a pool of 6 primers specific to the Ig constant regions and amplified the cDNA using 16 cycles of PCR with a pool of 46 V region-specific primers and 6 nested constant region primers. Following ligation of 454-compatible sequencing adapters, we purified the expected VH fragment using PAGE. Each sample derived from a given time-point was uniquely bar-coded during the ligation process, allowing subsequent mixing of all the time points into one common reaction sample (performed independently for each replicate run). Emulsion PCR and 454 GS FLX sequencing were performed directly at the 454 Life Sciences facility according to the manufacturer's standard protocols.

2.4.4 Data processing overview

Following data generation, the resulting reads were processed through an in-house software pipeline. The sequencing reads were filtered for quality, proper fragment size, and presence of a sample identity barcode. The reads were aligned to the reference IMGT database to identify the V, D, and J regions. We then partitioned the reads by VJ usage and hierarchically clustered them using their CDR3 junction to define unique clones. This data was finally used for subsequent time series and statistical analyses, including selection estimation with BASELINe [18] and phylogeny inference with Immunitree.

2.4.5 VDJ alignment process

For each segment we performed a semiglobal dynamic programming alignment against each reference sequence, choosing the best match. To maximize the number of distinguishing nucleotides, we performed our alignment in order of decreasing segment length (V then J then D), and subsequently prune off successfully aligned V or J regions before attempting alignment of the next segment. Since we know that the V and J segments must reside at the ends of the reads, we used a method that is similar to the Needleman-Wunsch algorithm [31]. In contrast to the canonical algorithm, we used zero initial conditions to allow the start of the alignment to occur anywhere without penalty. The alignment is then reconstructed and scored by starting

at the maximum value of the score matrix along the last row or last column, and backtracing. Finally, the identified V or J segments are removed before proceeding to the J or D alignment, respectively. For the D region alignment, we used the canonical Smith-Waterman local alignment algorithm [31], as we have no prior information as to where the D segment should reside. Finally, we compared the performance of our aligner against IMGT/V-QUEST [32] and generate ROC curves (Figure 2.19).

2.4.6 Sequence clustering

We performed sequence clustering in order to group our sequences (reads) into unique clones. This process is primarily used to associate sequences that originated from the same cell/clone, while allowing minor variations attributable to sequencing errors. For most of our work, we chose to use single-linkage agglomerative hierarchical clustering with Levenshtein edit distance as the metric. To make the clustering process more tractable, we partitioned our reads based on VJ identity. Within each partition, we then performed sequence clustering using only the CDR3 junction nucleotide sequence. To account for sequencing errors, we examined the distribution of cophenetic distances observed in the linkage tree, and determined the optimal distance to clip the tree at 4-5 edits (Figure 2.20).

2.4.7 Mutation analysis pipeline

After removing the primers from both ends of each raw read, High V-Quest [33] was used to assign a V and J genes and to align the sequences through the IMGT unique numbering scheme. In this step most of the insertions/deletions were identified and corrected by either removing any insertion or adding "N" to replace any deletion.

Following this step, sequences that potentially had artificial mutations due to incorrect germline assignments were excluded. This was done by: 1) excluding nonfunctional sequences (due to the occurrence of a stop codon and/or due to a shift in the reading frame between the V and the J gene), 2) excluding sequences with more than 14% mutations, 3) excluding sequences with more than 7 mutations in any 12 nucleotide window. This final step was taken in order to account for the possibility of an insertion following a deletion event which can be wrongly viewed as several dense point mutations.

Clonality was determined using a two-step approach. First, the sequences were divided into groups based on equivalence of their V-gene assignment, J-gene assignment, and the number of nucleotides in their junction. Following this step, clones were then defined within each of these groups as a collection of sequences with junction regions that differ from one sequence to any of the others by no more than three point mutations. The threshold of three was determined

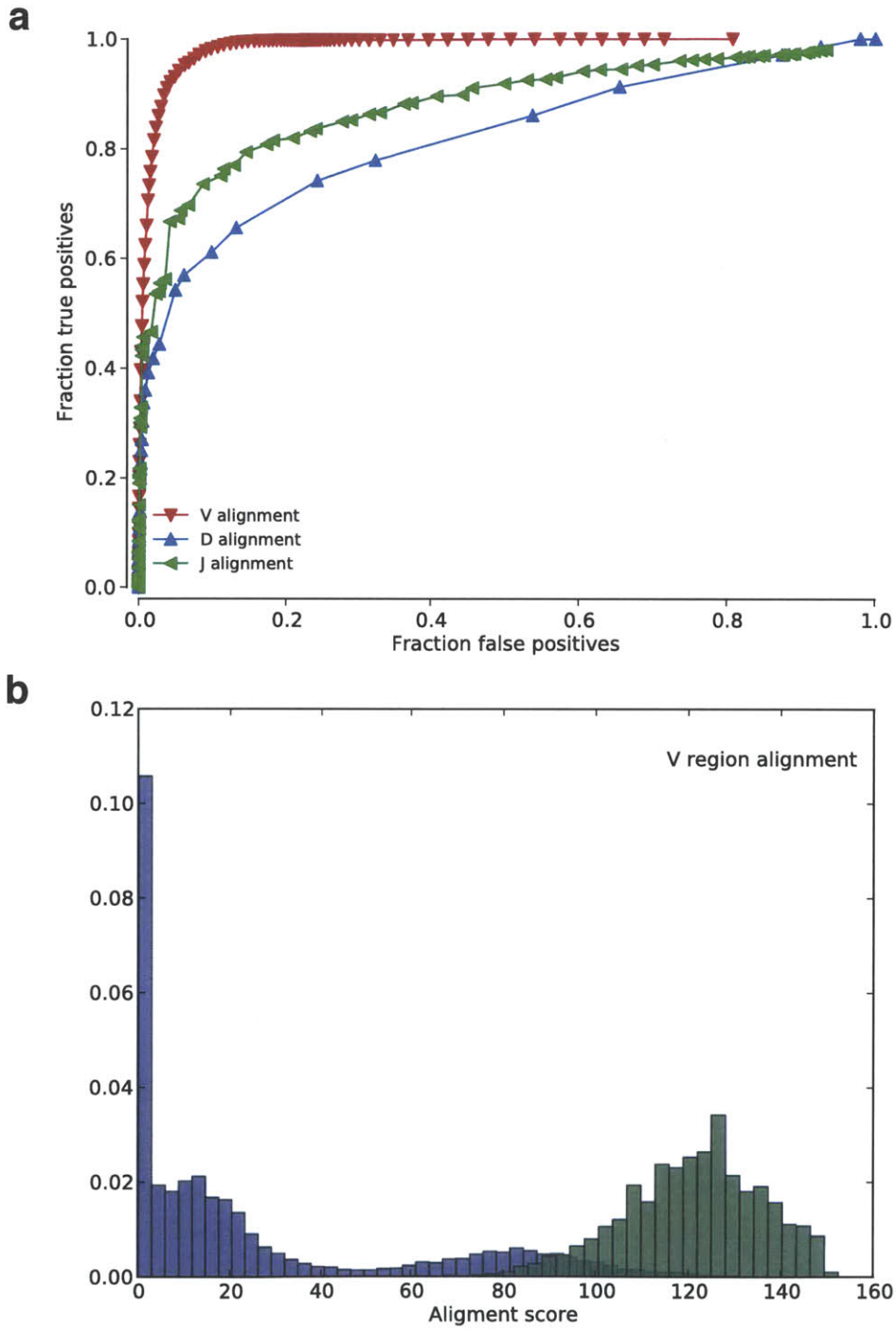


Figure 2.19: VDJ aligner calibration. (a) ROC curves comparing our VDJ aligner to IMGT/V-QUEST as gold-standard. (b) V alignment scores for "correct" alignment versus incorrect alignments.

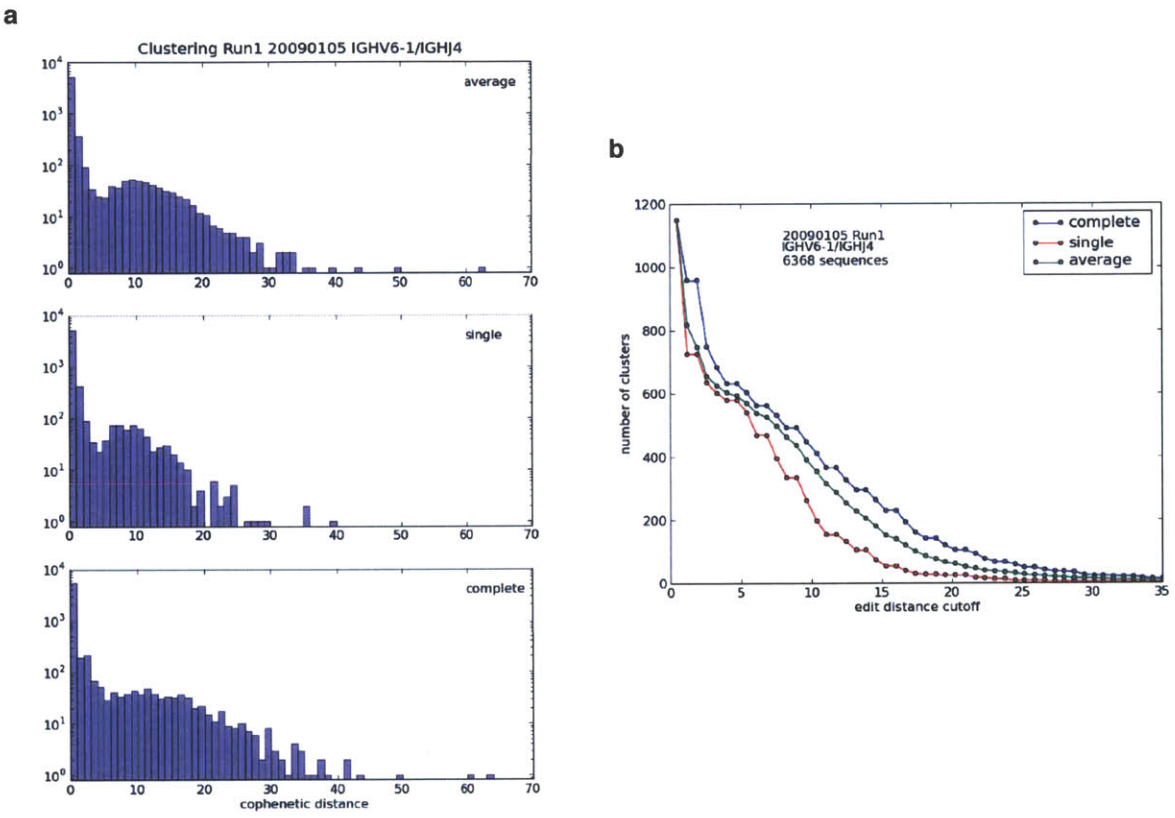


Figure 2.20: Clustering calibration. (a) Distribution of cophenetic distances for single, complete, and average linkage clustering. (b) Number of clusters as a function of clipping threshold.

after manual inspection of the mutation patterns in resulting clones identified through building phylogenetic trees.

2.4.8 Analysis of selection pressures

Selection pressure analysis was carried out using BASELINe (Bayesian estimation of Antigen-driven SELectIoN) [18] based on the local test formalism (see [34]). The output of BASELINe is a full posterior probability distribution function for each sequence and for a collection of sequences. Here, we used the mean selection estimation for each sequence for the tree analysis. For Figure 2.9, we have calculated a combined selection score (and 95% confidence intervals) for each combination of individual, time point and isotype.

2.4.9 Clone phylogeny inference

To determine the most likely phylogeny of a clone of reads, we use the Immunitree algorithm. Immunitree uses a probabilistic generative model that assigns a probability to each possible phylogeny. We apply MCMC to sample from this probability distribution of possible phylogenies, subject to the constraint that the phylogeny must generate the observed empirical data. MCMC generates an entire chain of samples of possible phylogenetic trees. Per MCMC iteration, we perform block gibbs on each of the parameters: phylogenetic tree structure, birth and death times of individual subclones, birth and death rates, mutation rates, read error rates, subclone consensus sequences, and assignment of reads to subclones. Finally, we perform a brief optimization on each of the sampled trees, and select the best such optimized sample as the final output.

2.4.10 V-usage clustering

After assigning the sequences to clones, each clone is associated with one V-gene. A V-gene usage vector for each individual-isotype combination was created. Using a Euclidean distance metric for these vectors, a neighbor joining tree was created in Figure 2.6.

2.4.11 Clone synthesis/affinity

We tested whether we could find antigen-specific clones by choosing the most highly expressed clones at the +7 day time points. We picked a subset of the largest clones from multiple time points before and after vaccination (-2 day, +7 days, +21 days) and synthesized them chemically. Because high-throughput technology to pair heavy and light chains from single cells are yet to be available, we cloned the full light chain repertoires from the corresponding time points.

The constructs were then paired in an scFv format and panned using phage display against the influenza antigens present in the vaccines. After three rounds of selection against hemagglutinin, we found only a single clone at days +7 from GMC-2009 that displayed significant affinity.

2.4.12 Software tools

Processing of raw data was performed by python packages and is available here:

- https://github.com/laserson/vdj
- https://github.com/laserson/pytools

Figures were produced with matplotlib, R, and graphviz. Scripts for figure preparation are available upon request.

2.5 Author contributions

Uri Laserson originated the experiment. Francois Vigneault designed experimental procedures and performed the majority of the experiments, including library preparation, and heavy chain synthesis. Uri Laserson wrote software for analysis, performed the majority of the data analysis, and generated most of the figures. Daniel Gadala-Maria, Gur Yaari, Mohamed Uduman, Jason Kasvin-Felton, and William Kelton performed data analysis, especially around mutation and selection analysis. William Kelton and Sang Taek Jung performed protein expression and characterization on synthesized clones. Jonathan Laserson and Yi Pei Liu generated the Immunitree algorithm and applied it to our data. Rajagopal Chari performed germline analysis. Jehyuk Lee performed phlebotomy. Ido Bachelet performed some expression analysis. Brendan Hickey and Erez Lieberman-Aiden contributed software. Bozena Hanczaruk, Birgitte Simen, and Michael Egholm contributed 454 sequencing services. Daphne Koller, George Georgiou, Steven Kleinstein, and George Church supervised research.

Chapter 3

Broadly Neutralizing HIV-1 Antibodies With Low Levels of Somatic Hypermutation Isolated by Deep Sequencing Analysis

3.1 Introduction

HIV-1 comprises numerous clades and subtypes with recombinant forms constantly emerging and circulating worldwide [35]. This diversity presents an unprecedented challenge to the humoral immune response. For most infections, the immune system readily adapts and eliminates the pathogen, but the high mutation frequency of HIV-1 invariably produces an escape variant, which ultimately leads to chronic infection [36]. Accordingly, a protective vaccine against HIV-1 would likely require the elicitation of broadly neutralizing antibodies (bNAbs), which are capable of neutralizing an extensive cross-clade panel of virus strains and thereby prevent acquisition of the virus rather than clearing it after infection.

Recent work has suggested that approximately 5–20% of chronically infected individuals develop bNAbs to some degree, but the details of how these antibodies emerge and mature remain unclear [37–41]. A common observation among bNAbs, however, is their unusually high mutation rate. While conventional antibodies diverge 5–15% from germline in the affinity maturation process, the CD4 binding site bNAb VRC01 is 40% divergent from germline in its variable heavy chain sequence [42, 43]. To a lesser extent, the quaternary epitope-specific bNAb PG9 is 18% divergent from heavy chain germline, but also has an unusually elongated CDRH3 that is 30 amino acids in length [38]. Finally, the recently described PGT antibodies,

which bind to proteoglycan epitopes involving the glycan at position 332 of Env, are 20–25% divergent from heavy chain germline and demonstrate the highest observed potency against a broad panel of HIV-1 isolates [37].

It remains to be determined if antibodies with high levels of SHM can be elicited by vaccination. In natural infection, isolated anti-gp120 binding antibodies exhibit a mutational level of approximately 20–25% from germline [44]. This high level of mutation stands in stark contrast to the mutation levels induced by vaccination for HIV-1, which range between 4–14% [45–47]. This gap between natural infection and vaccine-induced mAbs is not as wide in other viruses such as influenza, where natural infection yielded antibodies with mutation levels ranging between 15–30% compared to 13% for the broadly neutralizing influenza antibody CR6261, which was isolated via phage display from a healthy vaccinated individual [48, 49]. Given the low level of SHM for HIV-1 mAbs generated by vaccination, it is critical to understand the extent to which antibody maturation is necessary for neutralization breadth and potency.

Characterizing the evolutionary pathway of such highly-mutated bNAbs requires the use of high-throughput DNA sequencing to observe thousands of related antibody variants in a sea of unrelated antibodies (the needles in the haystack). These variants are subsequently used to build an evolutionary tree including the affinity matured bNAb. Recent work using 454 pyrosequencing attempted to uncover a roadmap from germline to affinity-matured mAb using such an approach [43]. They were hampered, however, by phylogeny methods that were ill-suited to analyzing antibody somatic hypermutation (SHM). In this work, we have addressed these issues by using a new phylogeny method, Immunitree, specifically designed for high-throughput analysis of SHM [23]. Because the method is probabilistic, it naturally allows for the incorporation of the state-of-the-art in mutation and sequencing error models. Furthermore, the process of SHM does not guarantee that the leaf nodes in a phylogenetic tree are the most fit. To address this, Immunitree allows for the observation of intermediate clones, which can be further characterized in functional assays. Our approach is also distinct from previous work as the focus is not on the evolution toward an affinity-matured bNAb, but rather on the evolution from the germline. While the former is critically dependent on correct heavy and light chain pairing, the latter assumes that all combinations are possible and focuses instead on the degree of deviation from germline and its effect on neutralization breadth and potency.

We applied this new approach to PBMC samples from the elite neutralizer donor 17 from whom the bNAbs PGT121-123 were isolated [37]. This set of antibodies are among the most potent of bNAbs described to date and have recently been shown to protect against SHIV challenge in passive transfer macaque studies. After 454 sequencing and phylogeny with Immunitree, we identified a predicted antibody precursor that is 90% similar to germline but demonstrated high neutralization potency and moderate breadth. This less-mutated clone appears

to preferentially bind trimeric Env over monomeric gp120, suggesting that Env trimer may be more effective for initiating affinity maturation than monomeric gp120. After estimating selection pressures on the bNAbs, we find they have experienced considerable negative selection. This work suggests that while somatic hypermutation does play a role in increasing breadth and potency, our successful observation of low-mutation neutralizers demonstrates that extensive SHM is not as important for conferring function as previously thought. The results presented here are highly important for vaccine design as it will likely be more feasible to re-elict bNAbs that are less mutated from germline than highly affinity matured antibodies.

3.2 Results

3.2.1 Global repertoire of chronically-infected donor 17 is slightly perturbed relative to health donors

Total RNA was extracted from sorted IgG memory B cells from donor 17 PBMCs. The RNA was reverse-transcribed and used for full-repertoire sequencing. Reads were aligned to the IMGT germline reference database and similar CDR3 sequences were clustered together as clones.

The full-repertoire sequencing did not successfully achieve the read depth necessary to find variants of the PGT antibodies (only 70k and 90k reads for heavy and light loci). However, it did allow us to compare characteristics of the global repertoire of the chronically infected donor 17 with healthy donors (HDs). In contrast to the donor 17 sequencing, which was performed on sorted IgG memory cells, the HD sequencing was performed on PBMCs. Accordingly, to help ensure a fair comparison, we only counted HD reads that were genetically determined to derive from IgG cells (though this approach may have included non-memory IgG cells). Interestingly, the global repertoires of the HD donors were qualitatively similar to the donor 17 repertoire despite the chronic infection (Figure 3.1). Overall SHM levels were similar between all four individuals, while CDR3 length tended to be slightly shorter for donor 17. V-usage tends to be similar between all individuals (>0.8 Spearman correlation between all comparisons), but the HD repertoires cluster closer to each other than the chronically infected patient. However, there could be other confounding variables (e.g., all HD patients are western Caucasian males, while donor 17 is African; donor 17 libraries were prepared with different primer sets).

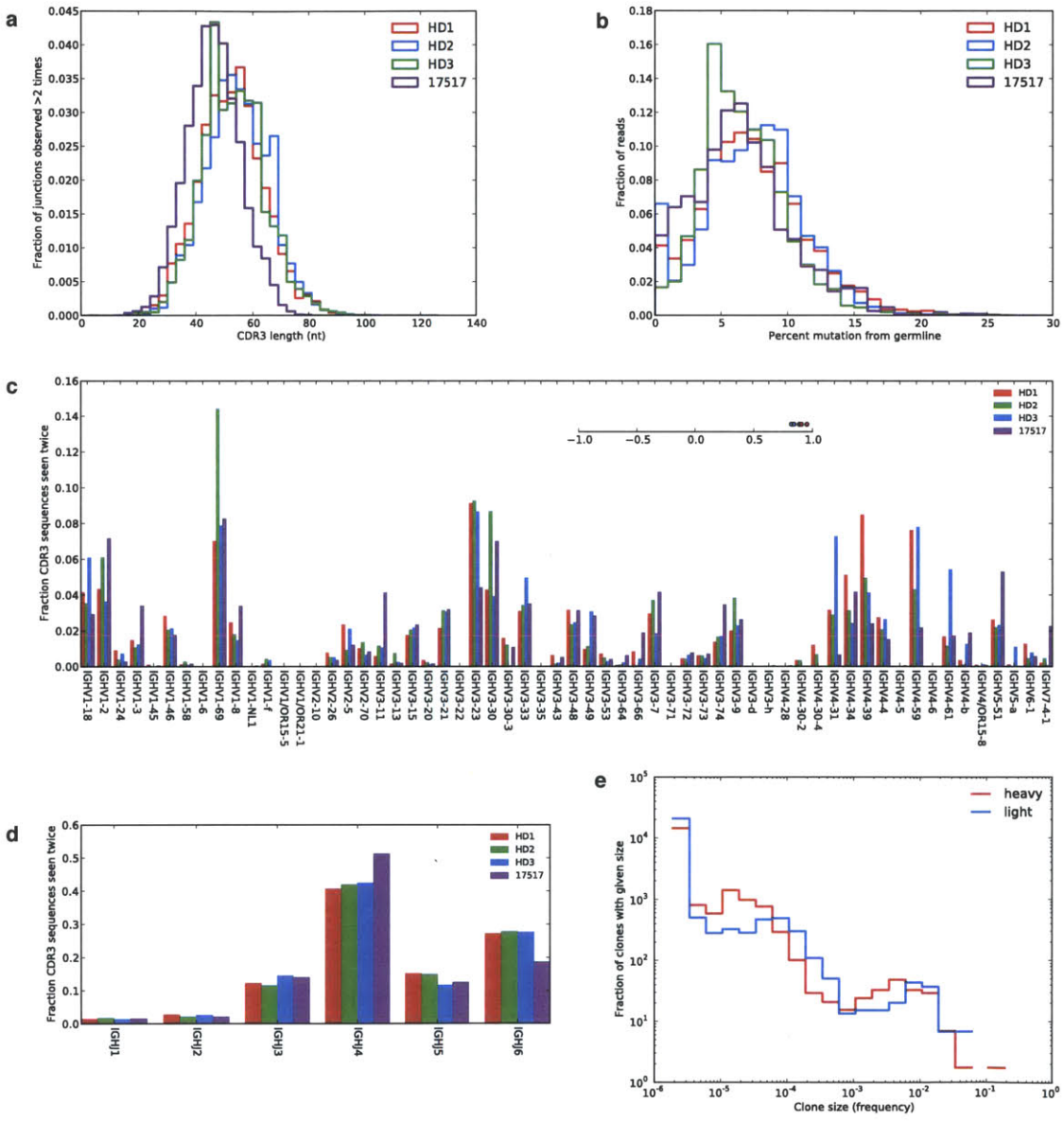


Figure 3.1: Donor 17 global repertoire. (a) CDRH3 length distribution (nucleotides) (b) SHM distribution (c) V-usage (d) J-usage (e) Clone size distribution of donor 17 alone. HD is health donor.

	Full-repertoire				Family-specific			
	Heavy		Light		Heavy		Light	
Total reads	70 063	⌝	92 627	⌝	376 114	⌝	530 197	⌝
IGHV4-59 or IGLV3-21	2 695	4%	5 923	6%	129 410	34%	505 359	95%
Enrichment					*9 fold*		*15 fold*	

Table 3.1: Donor 17 sequencing summary.

3.2.2 Family-specific deep sequencing allows discovery of PGT antibody variants and prediction of low-mutation precursors

Our full-repertoire data set was not sampled deeply enough to capture any PGT variants. Indeed, only 5% of reads were of the target V-gene family and only a small fraction of these would be PGT variants. To address this issue, we designed primers specific for the PGT V-gene families (IGHV4/IGHG for heavy, IGLV3-21/IGLC for light) and sequenced more deeply (Table 3.1). The resulting amplicons were sequenced on the 454 platform, resulting in 376114 heavy chain reads and 530197 light chain reads that are identifiable as immunoglobulins. The V and J gene for each read is determined, along with its percent mutation from the corresponding germline sequence. The IMGT-defined CDR3 sequences from all the reads are then clustered at 90% identity with USEARCH to define clones. For the heavy and light chain loci, we achieved 9- and 15-fold enrichment of the target V-gene family, with much higher overall coverage.

In order to identify variants of the PGT antibody sequences, we first computed divergence-mutation information: each read is scored on its sequence identity to one of the PGT antibodies (divergence) and also for its mutation level compared to the germline V-gene (mutation) (Figure 3.2). Small clusters of reads with above-background identity to the PGT antibodies were easily identifiable. All reads from the high-identity clones were manually extracted and carried forward for phylogeny inference with the Immunitree algorithm.

The Immunitree algorithm performs Bayesian phylogeny inference on antibody sequences [23]. It naturally encodes known models of somatic hypermutation as well as sequencing error rates. And critically, in contrast to traditional phylogeny methods, Immunitree allows for the observation of intermediate nodes in the resulting tree. High-identity reads from heavy and light chain data were separately run through the Immunitree algorithm, including the PGT antibody sequences (Figure 3.3). The PGT antibodies were observed at relatively low levels on the trees. Furthermore, we estimated selection pressures on the reads using the BASELINe algorithm (Figure 3.4) [18]. In an ad hoc manner, we chose both precursor clones as well as clones that were more highly evolved and chemically synthesized them for expression and characterization.

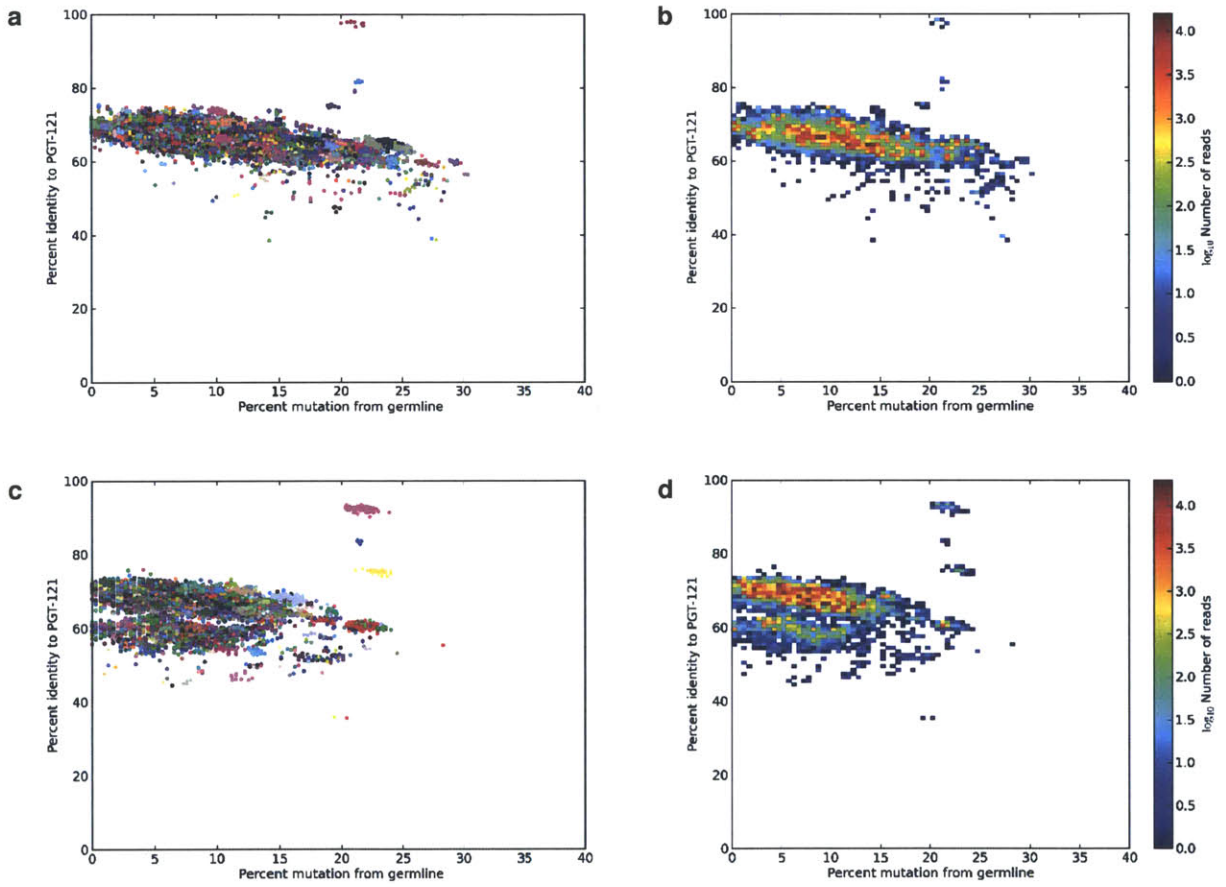


Figure 3.2: Divergence-mutation plots for PGT121. Top: PGT121 heavy chain, bottom: PGT121 light chain. Left: scatter plots; each point is a read. The points are colored according to cluster membership. Right: histogram of reads.

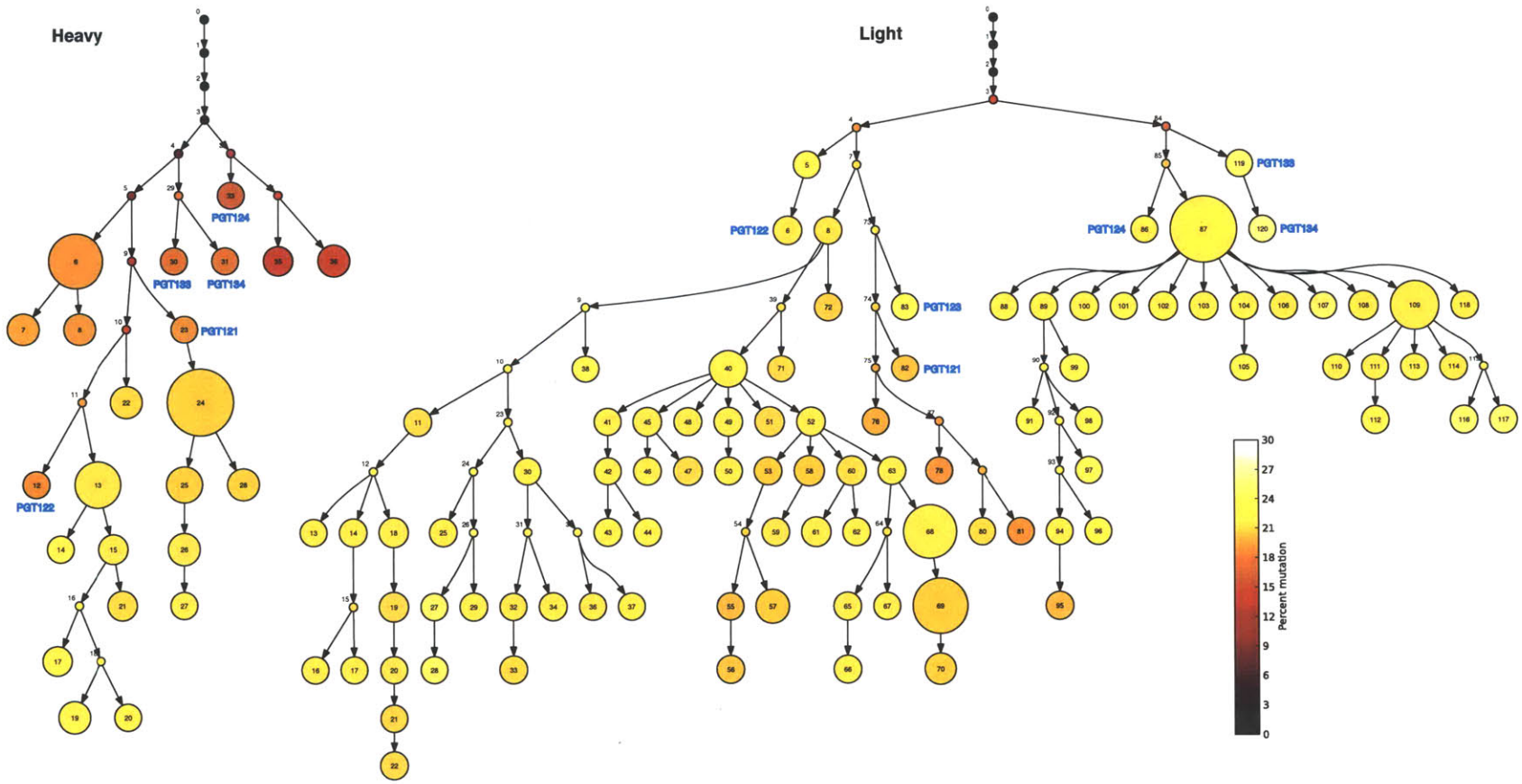


Figure 3.3: Donor 17 antibody phylogeny. Computed with the Immunitree algorithm. Filles, small nodes were not observed, while colored larger nodes had reads that were observed. The nodes are colored based on the percent mutation of the nucleotide sequence of the underlying model sequence. The PGT antibodies are annotated according to their assignments to the nodes.

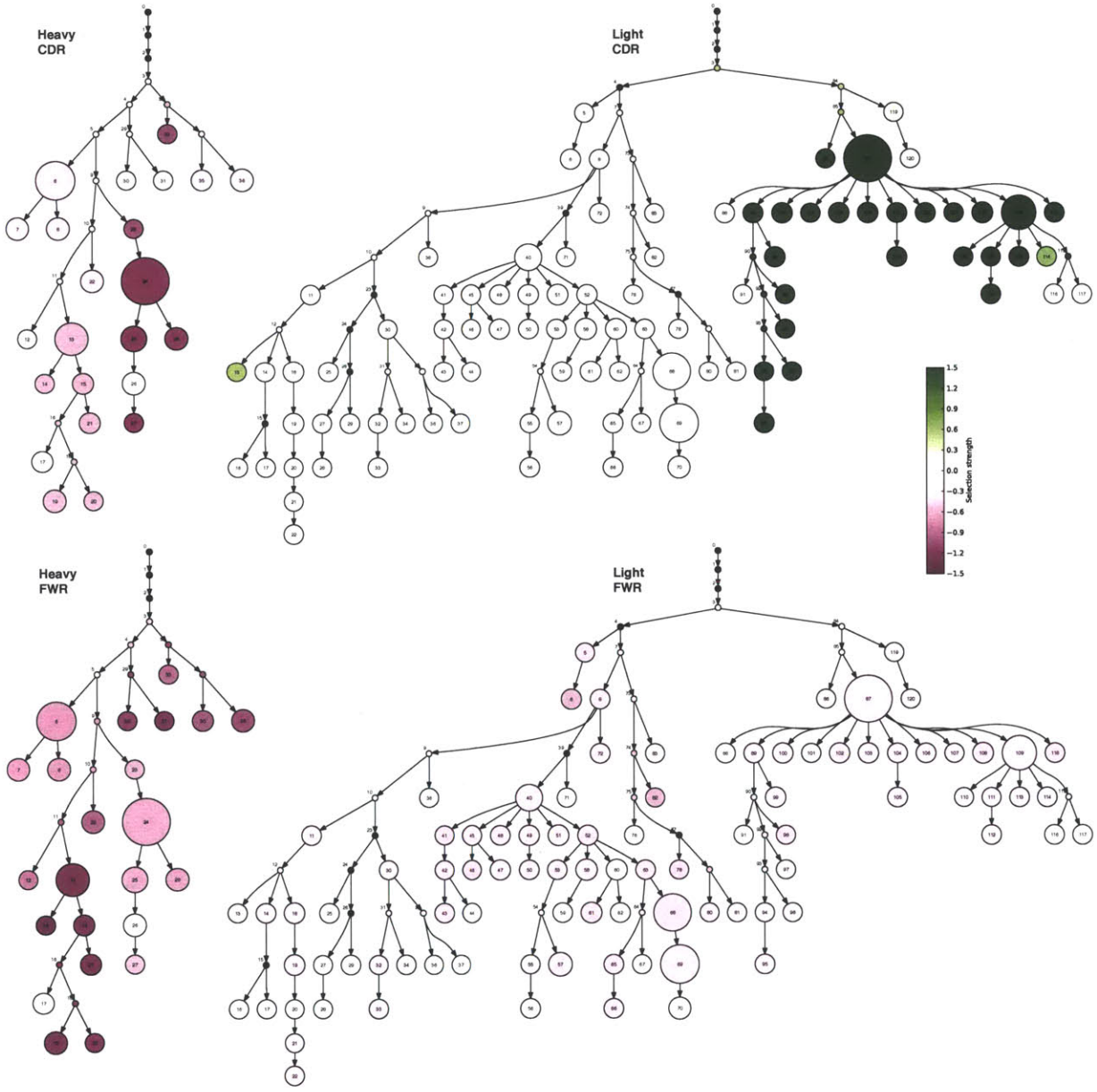


Figure 3.4: Donor 17 antibody phylogeny by selection pressure

3.2.3 Heavy chain and light chain nodes are capable of functionally complementing each other and demonstrate neutralizing activity

While 454 pyrosequencing has the capacity to generate hundreds of thousands of heavy and light chain variable reads, it is unable to preserve the original pairing from the same memory B cell clone. To determine if selected nodes are capable of generating mAb clones with neutralizing activity, selected heavy and light chain nodes were paired and tested for neutralizing activity on a six-virus panel previously determined to be representative of breadth on a larger virus panel (Figure 3.5). All combinations successfully produced recombinant antibody as measured by anti-Fc ELISA. Some combinations, however, were unable to demonstrate neutralizing activity, which could be due to incorrect pairing or missing heavy/light chain clones from the broader repertoire in our analysis.

Interestingly, antibody clones comprising heavy and/or light chain sequences that are more divergent from germline consistently demonstrated greater neutralization potency and/or wider breadth on the six-virus panel (Figure 3.5b). This correlation can also be seen visually on the tree with highly broad and potent clones occupying nodes furthest from germline, which is located at the root of the tree (Figure 3.6). Based on the data, it appears that somatic hypermutation does directly influence the breadth and potency of the antibody. Notably, the predicted early clones 3H+3L and 32H+3L were still capable of neutralizing three out of the six viruses on the panel with high potency. To better understand the contribution of SHM to neutralization breadth and potency the following pairs were expressed, purified, and further characterized: 3H+3L, 32H+3L, and 3H+87L.

3.2.4 Characterization of intermediate nodes demonstrates correlation between level of SHM and breadth and potency

The paired clones 3H+3L, 32H+3L and 3H+87L were expressed in mammalian suspension cells and purified with a protein A column before testing for neutralization breadth and potency on a 38 cross-clade pseudovirus panel. The 3H sequence is 10% (amino acid) mutated from germline, 32H is 16% mutated, 3L is 13% mutated and the 87L sequence is 30% mutated. This pairing allows a direct comparison between the effects of mutations among heavy chain sequences and among light chain sequences. The paired clones were tested for neutralization activity compared to PGT121 (Figure 3.7).

The results indicate that both neutralization breadth and potency increases with divergence from germline. The lowest divergent pair, 3H+3L, demonstrated the lowest breadth at 17/38

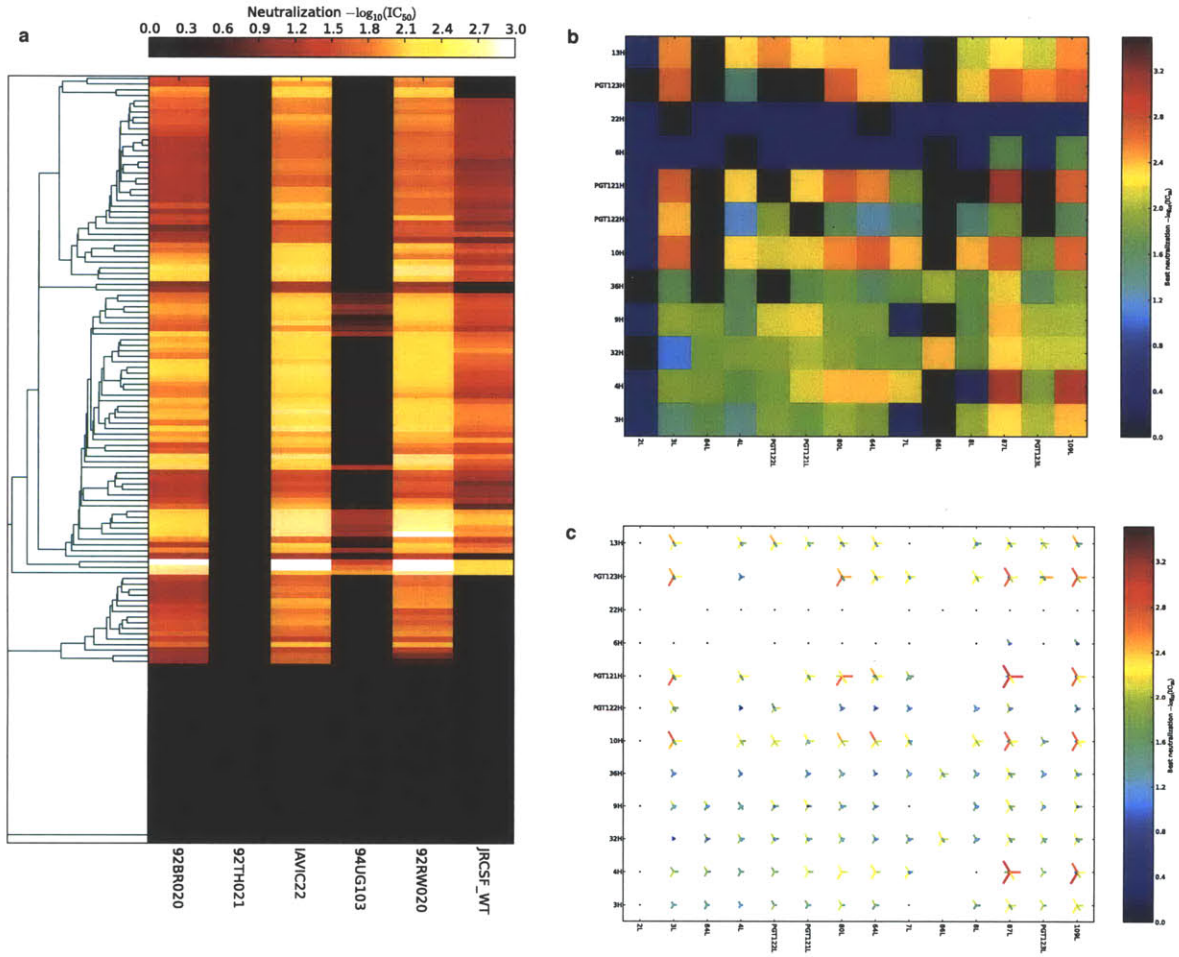


Figure 3.5: Six-virus neutralization panel. (a) Each column is a different virus strain. Each row represents a particular heavy/light chain combination from the tree. The rows are clustered according to their neutralization profiles. (b) Heatmap showing the best performance of a particular heavy/light combination on the 6 viruses. The chains are arranged from bottom-left to top-right as least-mutated to most-mutated (according to nucleotide homology). (c) Each star represents a heavy/light chain combination. Each line in the star represents a virus straing. The length of the line and the color both represent the neutralization potency.

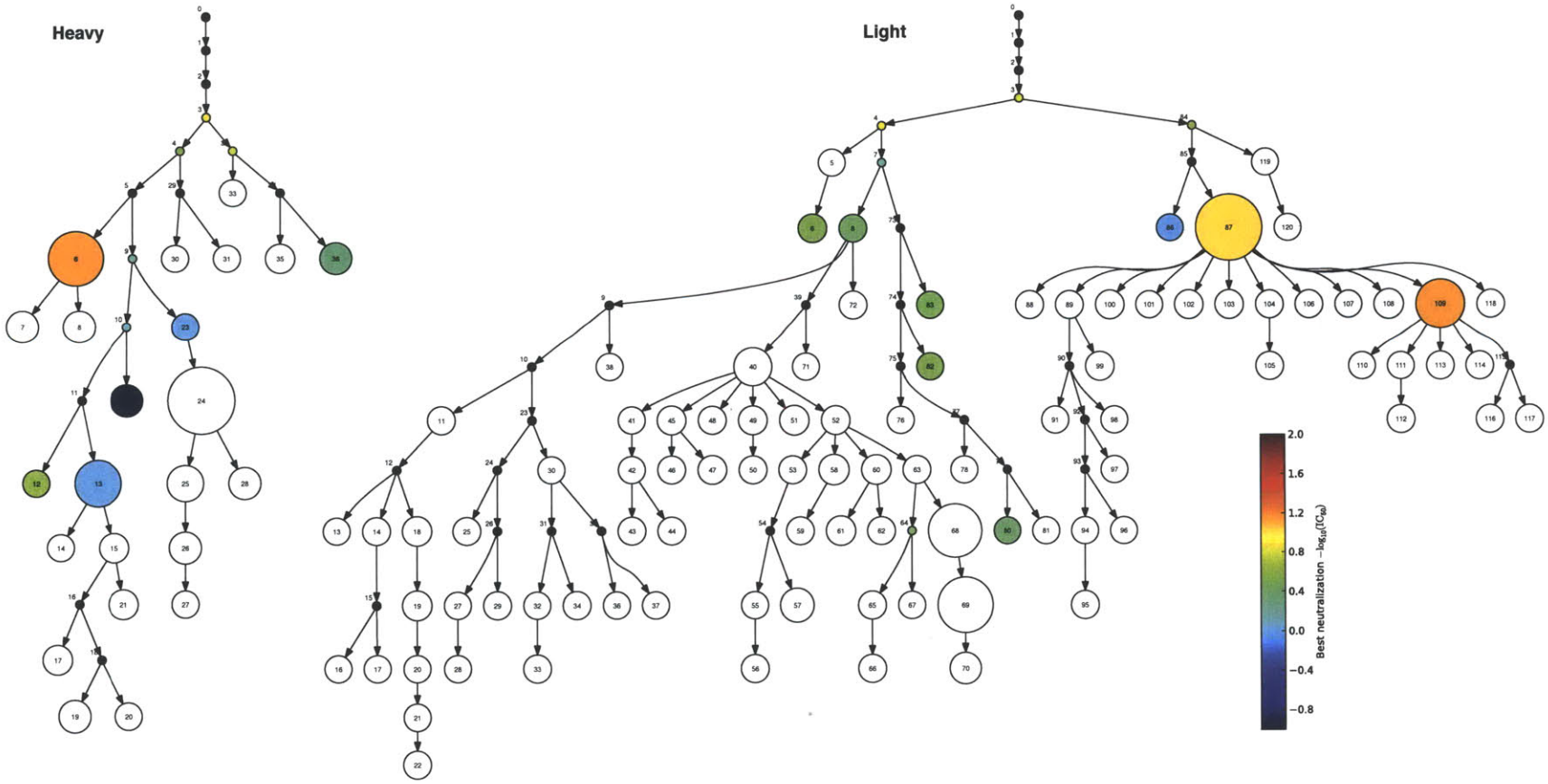


Figure 3.6: Donor 17 antibody phylogeny by neutralization. Nodes that were synthesized are colored; nodes not synthesized are white or black. Nodes are colored according to their best performance in terms of neutralizatino IC50.

CLADE	STRAIN	3H + 3L	32H + 3L	3H + 87L	PGT121
A	92RW020	0.010	0.007	0.004	0.002
	Q23ENV17	0.015	0.007	0.005	0.004
	94UG103	>10	>10	0.268	0.540
B	6535.3	0.010	0.008	0.010	0.003
	92BR020	0.029	0.024	0.008	0.004
	CAAN5342.A2	0.121	0.018	0.015	0.006
	TRO.11	0.207	0.019	0.012	0.006
	JR-FL	0.081	0.051	0.015	0.011
	YU-2	1.725	0.454	0.136	0.028
	RHPA4259.7	>10	0.935	0.033	0.004
	JR-CSF	>10	0.062	0.043	0.020
	AC10.0.29	>10	0.098	0.025	0.015
	PVO.4	>10	1.074	0.171	0.052
	TRJO4551.58	>10	1.103	0.025	0.198
	QH0692.42	>10	>10	0.076	0.039
	SC422661.8	>10	>10	0.109	0.033
	WITO4160.33	>10	>10	1.530	0.117
BC	CNE20	0.011	0.003	0.004	0.002
	CNE53	0.061	0.019	0.011	0.014
C	IAVI C22	0.012	0.008	0.004	0.002
	DU156.12	0.073	0.019	0.020	0.004
	ZM233	0.017	0.574	0.077	0.020
	ZM214M.PL15	0.517	0.408	0.323	0.056
	DU422.1	>10	0.157	0.087	0.009
	DU172.17	>10	0.939	0.116	0.010
	ZM53M.PB12	>10	>10	>10	0.003
G	P1981_C5_3	0.023	0.012	0.009	0.005
	X2131_C1_B5	1.677	0.020	0.012	0.004
	P0402_c2_11	>10	0.035	0.028	0.012
	X1254_c3	>10	0.443	0.284	0.053
	X2088_c9	>10	0.358	0.013	0.008
	X1193_c1	>10	0.285	0.141	0.024
CRF02_AG	T250-4	0.318	0.007	0.005	0.004
	235-47	>10	3.881	0.028	0.033
	263-8	>10	>10	>10	0.160
	% Homology to GL	89%	86%	80%	78%
	% Virus Neutralized	45%	84%	95%	100%
	Mean IC_{50}	0.289	0.380	0.111	0.043
	Median IC_{50}	0.061	0.035	0.025	0.013

Figure 3.7: Broad virus panel neutralization. Values are IC50s in μg/mL.

viruses. The overall mutation level of both heavy and light chains for this clone yields a total homology of 89% to germline in heavy and light chain combined. Interestingly, the clone was still able to maintain a high level of potency with a mean IC50 of 0.29 μg/mL and a median IC50 of 0.061 μg/mL. With a few additional mutations in the heavy chain (an overall heavy and light chain germline homology of 86%), the clone 32H+3L was able to increase breadth and potency, neutralizing 29/38 viruses in the panel with a mean IC50 of 0.35 μg/mL and a median IC50 of 0.035 μg/mL. A higher amount of mutations in the light chain (3H+87L) yields an even broader and more potent antibody capable of neutralizing 35/38 viruses in the panel at a mean IC50 of 0.11 μg/mL and a median IC50 of 0.025 μg/mL. These values suggest that this specific VDJ recombination junction and heavy/light chain pairing might be capable of producing an antibody that is highly potent from the onset and/or the epitope to which it binds to enables a high level of potency. The capacity for breadth, however, appears to be a secondary development and is a direct product of SHM. These findings have valuable implications for immunogen design as it suggests that certain antibodies and/or epitopes may be more capable of eliciting physiologically relevant serum responses than others.

3.2.5 Genomic sequencing and paratope mapping indicate residues that have arisen from SHM and are important for neutralizing breadth and potency

In order to better determine if key residues in the variable region are a result of SHM or are possibly due to polymorphisms in the donor, gDNA from the donor was extracted and V and J genes from both heavy and the light chains were amplified and sequenced (Figure 3.8). The results indicate that the deletion in the 5' end and the insertion in FW3 of the light chain were features that developed following recombination.

In order to the determine the minimal amount of mutations necessary for neutralizing activity, the least mutated clones were aligned to the germline sequence and residues resulting from SHM were identified. Single amino acid changes in 3H and 3L sequences were individually reverted, paired with corresponding light or heavy chain clones, and then tested on the 6-virus panel (Figure 3.9). For the 3H heavy chain, only residues in the CDRH3 demonstrated a significant change in neutralization IC50 and this observation was consistent for multiple light chain pairs. Given these results, we completely reverted the heavy chain to the germline sequence and still found similar neutralization activity, suggesting that most of the residues mediating neutralization are in the CDRH3.

For the 3L light chain, the effects of single amino acid reversions depended on the degree of divergence in the heavy chain. When paired with a less mutated heavy chain (3H), residues in

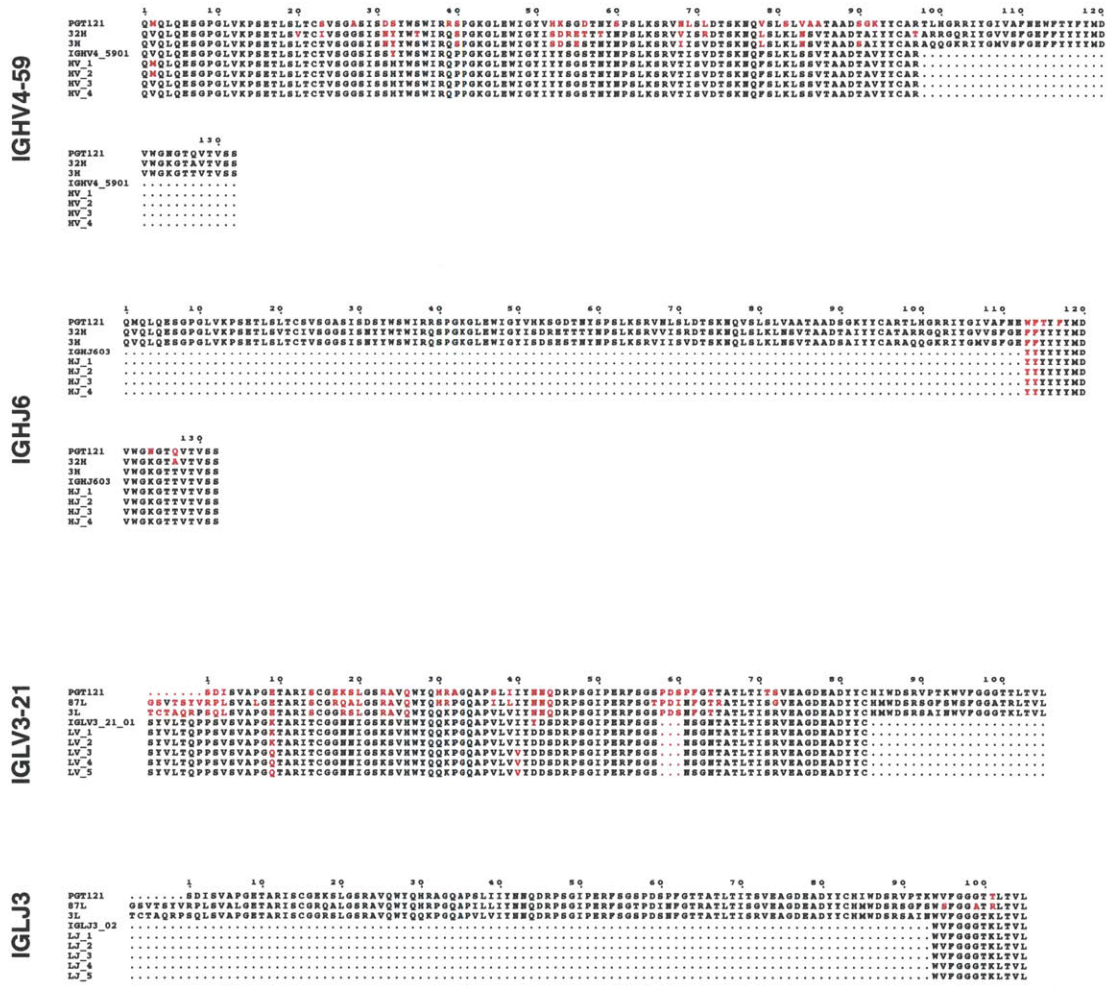


Figure 3.8: Germline sequencing for donor 17. The target V and J genes were amplified from genomic DNA from unrecombined antibody loci. PGT121 is the known antibody sequence. 3H, 32H, 3L, and 87L refer to the antibody variants. IGH* refers to the IMGT reference sequence. HV, HJ, LV, LJ refer to Sanger reads from genomic DNA.

Mutating 3H

Mutating 3H
Paired to 3L

		92BR020	IAVIC22	92RW020
CDRH1	N31S	1.1	0.7	1.1
FR2	S40P	0.8	0.4	0.5
CDRH2	S52Y	1.6	1.8	2.0
	D53Y	1.1	1.4	1.9
	E55G	1.8	1.2	1.5
FR3	I68T	1.3	1.3	1.7
	L78F	1.5	1.0	1.7
	S90T	1.0	0.7	0.9
	I92V	1.0	0.5	0.5
CDRH3	Q99A	1.1	0.8	1.3
	Q100A	1.2	1.0	1.3
	K102A	0.9	1.0	1.2
	Y105A	>100	>100	>100
	M107A	0.5	8.7	45.0
	S109A	0.7	0.5	0.6
	F114A	2.5	1.8	2.6
	Y118A	1.6	1.1	1.9
	3H + 3L	1.0	1.0	1.0

Mutating 3H
Paired to 87L

		92BR020	IAVIC22	94UG103	92RW020	JRCSF WT
CDRH1	N31S	2.5	1.3	0.9	1.3	1.1
FR2	S40P	3.9	1.5	0.1	1.4	1.7
CDRH2	S52Y	4.8	2.0	3.5	1.5	1.1
	D53Y	3.5	1.4	3.5	1.7	2.2
	E55G	3.2	1.5	2.9	1.6	0.9
FR3	I68T	4.2	1.6	3.9	2.0	3.0
	L78F	4.1	2.1	2.9	1.8	1.6
	S90T	1.0	1.2	0.4	1.2	3.1
	I92V	4.9	1.9	0.4	1.3	3.9
CDRH3	Q99A	1.3	1.1	0.9	1.1	2.4
	Q100A	1.9	0.9	0.6	0.8	1.2
	K102A	1.1	0.8	0.5	0.6	0.7
	Y105A	21.9	33.6	>200	120.2	>200
	M107A	2.1	1.2	5.5	1.3	2.6
	S109A	1.2	0.8	0.2	0.7	0.9
	F114A	2.6	1.5	0.4	1.6	2.2
	Y118A	3.6	2.8	2.6	2.1	1.5
	87L + 3L	1.0	1.0	1.0	1.0	1.0

Mutating 3L

Mutating 3L
Paired to 3H

		92BR020	IAVIC22	92RW020
CDRL1	L20I	2.0	2.9	3.4
	R23K	1.5	1.8	1.3
	A24S	2.1	5.5	5.5
FR2	Q26H	13.2	61.0	135.0
	I40V	0.9	0.8	1.0
CDRL2	N42D	2.2	3.2	3.1
	N43D	>100	>100	>100
	Q44S	3.2	11.1	19.0
FR3 Ins	P58A	1.1	0.9	1.2
	D59A	5.7	38.5	181.1
	S60A	1.7	1.7	1.7
	N61A	1.0	1.5	1.3
	F62A	1.4	1.7	1.6
	G63A	1.2	1.6	1.4
	T64A	1.0	0.7	0.8
CDRL3	H84Q	1.5	2.1	1.9
	R89S	>100	>100	>100
	3H + 3L	1.0	1.0	1.0

Mutating 3L
Paired to 32H

		92BR020	IAVIC22	92RW020	JRCSF WT
CDRL1	L20I	4.6	1.3	2.0	2.1
	R23K	5.6	1.3	1.6	2.3
	A24S	5.7	1.2	1.8	2.6
FR2	Q26H	4.3	1.1	3.1	1.6
	I40V	4.5	1.0	1.8	1.4
CDRL2	N42D	7.0	2.0	2.2	2.8
	N43D	>100	>100	>100	>100
	Q44S	2.6	0.8	1.7	2.8
FR3 Ins	P58A	4.3	1.0	1.4	2.5
	D59A	7.0	1.6	4.5	13.0
	S60A	4.6	1.2	1.9	2.1
	N61A	0.8	0.2	0.7	1.1
	F62A	1.2	0.3	0.4	0.5
	G63A	4.4	1.1	1.6	1.7
	T64A	4.8	1.1	1.7	2.0
CDRL3	H84Q	4.3	0.9	1.5	2.8
	R89S	32.3	2.4	18.0	>100
	32H + 3L	1.0	1.0	1.0	1.0

Mutating 3L
Paired to PGT121H

		92BR020	IAVIC22	92RW020	JRCSF WT
CDRL1	L20I	1.1	1.0	1.1	1.4
	R23K	0.8	0.8	1.2	1.2
	A24S	0.5	0.5	0.5	1.1
FR2	Q26H	0.9	0.7	1.0	1.6
	I40V	0.8	0.8	0.7	0.9
CDRL2	N42D	0.7	1.0	1.0	1.2
	N43D	>100	>100	>100	>100
	Q44S	3.8	5.0	4.4	10.7
FR3 Ins	P58A	2.7	3.6	2.4	3.0
	D59A	1.7	1.9	2.5	4.0
	S60A	1.0	1.1	1.1	1.2
	N61A	1.3	1.5	1.4	1.5
	F62A	0.7	0.8	0.9	1.1
	G63A	1.3	1.8	2.0	1.5
	T64A	1.0	1.0	0.9	1.0
CDRL3	H84Q	1.2	0.9	0.6	1.5
	R89S	2.5	2.5	2.4	26.6
	121H + 3L	1.0	1.0	1.0	1.0

Figure 3.9: Paratope mapping. Values are fold difference in IC50 of the unmutated version to the mutated version.

FR2, CDRL2, and the CDRL3 were found to be critical for neutralization activity. Additionally, alanine scanning of the insertion in FR3 abrogated neutralization. When 3L was paired with the highly matured PGT121 heavy chain, however, only residues in CDRL2 and CDRL3 were found to be critical for neutralization. These findings suggest that initially the light chain likely makes more significant contacts on HIV-1 Env than the heavy chain, but that this dependency on light chain binding decreases with increased affinity maturation on the heavy chain. Having identified the residues that are necessary for neutralizaton activity in the light chain, the sequence was fully reverted to germline and might continue to neutralize, but only with the most matured heavy chain, and with reduced potency (these last results are preliminary).

3.2.6 Comparison of affinity-matured sequences identified gain-of-function mutations that enable neutralization breadth and potency

Although the clone 32H is only 3% more mutated than 3H, the 32H+3L pair is capable of neutralizing 12 additional viruses on the 38 virus panel compared to 3H+3L. Similarly, 87L is 18% more mutated than 3L, but this higher affinity-maturation increases the neutralization breadth from 17/38 to 35/38 viruses between 3H+3L and 3H+87L. In order to decipher which residues enables this increase in neutralization breadth, single amino acids in 32H were reverted to amino acids in 3H and then tested against viruses that could be neutralized by 32H+3L but not 3H+3L. A similar approach was perfomed for 87L, in which individual amino acids were mutated to corresponding residues in 3L. As seen in Figure **??**, single amino acid changes did not show demonstrable effects on neutralization IC50 for the 32H reversion (need to do doubles). Results are being repeated for 87L.

3.2.7 The predicted precursor binds preferentially to cell surface Env than monomeric gp120 and may require crosslinking for neutralization

It remains to be determined what form the HIV-1 Env epitope is presented to B cells to trigger activation and begin the affinity maturation process. Many groups have attempted to create germline derivatives of known bNAbs, but these constructs have not been able to bind recombinant gp120 or to neutralize virus strains. It is possible that the host germline antibody responded to an Env clone specific to the infecting virus and that breadth and potency developed gradually as the antibody matured.

To measure differences in binding affinity between the antibody pairs in this study, ELISAs

32H to 3H

	JR-CSF	YU-2
V20L	1.9	1.5
I23T	3.6	1.2
T35S	1.5	1.1
R54S	2.2	1.1
T56S	0.9	0.7
T58N	0.7	0.5
V68I	1.9	4.3
R71V	1.2	1.1

87L to 3L

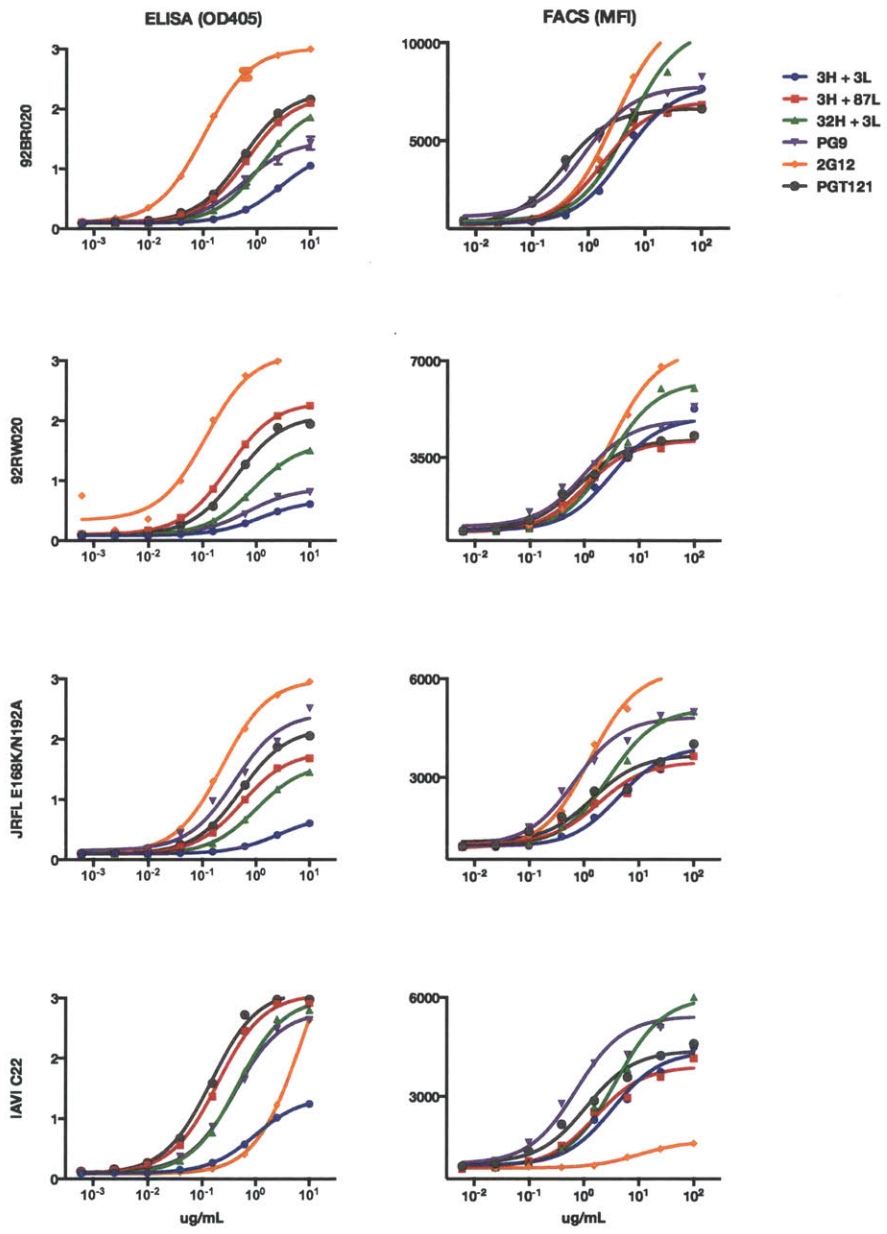
	JR-CSF	YU-2	94UG103
R17G Q18R A19S	2.1	1.5	5.5
H30Q R31K	0.8	0.7	2.2
I37V	7.1	> 8.0	> 8.0
L39V	6.2	> 8.0	> 8.0
T57S	1.2	1.2	2.3
I60S	3.0	2.3	4.2
R65T	3.0	1.9	2.8
G72R	1.1	0.9	2.1
G91A F92I S93N	1.9	1.1	5.9
A99G	1.2	0.8	1.5
R101K	1.2	0.9	1.9

Figure 3.10: Targeted reversion to detect neutralization. Values are fold differences in IC50 of pre-reverted to post-reverted.

were perfomed using both recombinant and lysed virus supernatants. The results show that the least mutated pair (3H+3L) had very little affinity for recombinantly produced monomeric gp120 compared to the more matured pairs (3H+87L and 32H+3L) (Figure 3.11a). This difference in affinity is mirrored by ELISA binding to gp120 protein extracted from lysed virus supernatents. Interestingly, binding to cell surface Env trimer measured by flow cytometry did not show a marked difference between 3H+3L and 3H+87L (figure). These results suggest that the least mutated clone prefers binding to cell surface trimer more than monomeric gp120 and that despite this difference in affinity, the antibody is still capable of neutralization activity. Next, we wanted to determine if the affinity for trimer involved crosslinking between gp120 protomers. To test this possibility, the antibody IgGs were digested into Fab fragments and tested in neutralization assays. The results demonstrate that the least mutated pair 3H+3L loses neutralizing activity when tested as Fabs, while the more mutated pairs neutralize with similar potency (Figure 3.11b). Accordingly, it is possible that less mutated precursors are able to crosslink between protomers within a trimer and that this increased valency served as a means to increase avidity and thereby positive selection.

Finally, we wanted to determine if this preference for trimer was exclusive to cell surface expression or if binding to recombinant versions of Env trimer was possible. To test this, an ELISA was performed using YU2 foldon gp140 trimer. Unlike the quarternary epitope-specific antibody PG9, 3H+3L was able to have a measurable affinity for the gp140 foldon. This finding suggests that the preference for trimer is distinct from the trimer epitope that is unique to PG9.

a



b

CLADE	STRAIN	3H + 3L	3H + 87L	PGT121
A	92RW020	23	2	3
	92BR020	7	3	5
B	YU2 WT	17	2	3
	JR-FL WT	43	3	6
C	ZM214M.PL15	58	6	14
	DU156.12	8	2	5
	IAVIC22 WT	25	1	6

Figure 3.11: gp120 trimer versus monomer. (a) ELISA and FACS affinity assays. (b) Fold difference of affinity of IgG-formatted antibodes versus Fab fragments to gp120 trimer.

More importantly, this finding also suggests that trimeric Env would serve as a better immuogen than monomeric gp120 in binding to candidate germline antibodies.

3.3 Discussion

A number of bNAbs against HIV-1 Env have been identified and the epitopes to which they bind have been structurally and/or biochemically defined [37, 38, 42, 50–54]. The concept of structure-based reverse vaccinology is to utilize this information to guide the design of immunogens, which would be capable of re-eliciting bNAbs following vaccination. The ideal vaccine would likely attempt to re-elicit a cocktail of antibodies to target various sites of vulnerability on HIV-1 Env, but it remains to be determined if some epitopes are more readibly elicitable than others. Currently, most immunogen design focuses on pushing the immune response towards a fully affinity matured antibody. All known bNAbs, however, are mutated to the point that vaccination is unlikely to reproduce them. An alternative approach would be to find less-mutated versions of the bNAbs that still maintain considerable neutralization breadth and potency. To do so, we must delineate the landscape that defines bNAbs. In other words, understanding the degree to which the level of deviation from germline is necessary for a bNAb's neutralization breadth/potency would prioritize different targets and directly inform immunogen design approaches.

The work presented here attempts to execute exactly this program, as the focus is primarily on the role of SHM on the development of antibodies' neutralization breadth and potency. Using 454 deep sequencing and a novel approach to phylogeny, we were able to predict and identify clones with neutralizing activity that are closer in homology to germline antibody. As our focus is on the evolution from germline and its effect on neutralization breadth and potency, we assume that all heavy and light chain combinations are possible. This approach differs from previous studies, which have attempted to map the evolution of highly mutated bNAbs and is therefore dependent on accurate heavy and light chain pairing. Furthermore, we are primarily interested in demonstrating the existence of low-mutation antibodies that could plausibly be elicited by vaccination. While our work found a positive correlation between the degree of SHM and neutralization breadth, pairing the least divergent heavy and light chain clones (3H+3L) showed that clones with relatively lower levels of SHM still demonstrated a high level of potency and moderate breadth. It is possible that this specific family of antibodies and/or specific VDJ recombination was able to generate a clone that is highly potent against viruses harboring the epitope and that wider breadth developed subsequently through SHM. This would stand in contrast to the MPER antibody 4E10, however, which demonstrates the highest breadth among all bNAbs, but lacks in potency. Interestingly, 4E10 is relatively less divergent from germline and

it is possible that the potency for these types of antibodies could increase through SHM.

Notwithstanding the target, there are two important obstacles in re-eliciting a bNAb of interest. The first is priming naïve B cells with an immunogen capable of binding germline precursors. The second is driving this B cell response towards a specific affinity maturation pathway. The data presented here suggests that the binding epitope on HIV-1 can evolve over time. For this specific antibody family, it appears that the affinity begins first on trimeric forms of Env and subsequently matures to gain tighter avidity on monomeric forms of gp120. This process could likely be due to a valency effect, where oligmeric forms could enhance low binding affinities. In terms of application to immunogen design, a possible strategy would first involve priming with trimeric forms of Env and subsequently boosting with monomeric immunogens. Indeed, it has been shown by a number of studies that trimeric forms of Env do provide a stronger immune response than monomer alone.

The ideal vaccine antibody response that would be protective against HIV-1 will likely possess three properties: wide breadth to protect against HIV-1 genetic diversity, high potency to produce physiologically viable serum titers, and low divergence from germline to increase the likelihood of re-elicitation. We chose to focus on the bNAbs PGT121-123 because they were recently described to be the most potent antibodies identified to date and are still capable of neutralizing >70% of HIV-1 isolates. These antibodies are ideal vaccine targets because their high potency suggests that even a modest response is capable of providing protection. The caveat to these antibodies, however, is their high level of mutation, with only 75% homology to germline for heavy and light chain combined. To date, antibodies elicited through vaccination have not demonstrated this high level of SHM. Accordingly, identifying antibodies that are capable of breadth and potency but have lower levels of SHM would redefine bNAb re-elicitation targets. To better understand the contribution of SHM to neutralization breadth and potency, we used deep sequencing analysis to identify and predict clones with varying levels of divergence from germline. These results discovered a predicted antibody that is 10% mutated, but still capable of neutralizing 17/38 viruses in a cross clade panel with a potent mean IC50 of 0.289 μg/mL. This analysis can be extended to other bNAbs that target different epitopes to similarly determine if less mutated antibodies can be identified. These results have important implications for vaccine design as it suggests that bNAbs do not require unconventionally high levels of SHM in order to have breadth and potency.

3.4 Materials and Methods

3.4.1 Human specimens

Peripheral blood monoculear cells (PBMCs) were obtained from donor 17, a HIV-1 infected donor who is a part of the previously described IAVI Protocol G cohort. All human samples were collected with informed consent under clinical protocols approved by the appropriate institutional review board.

3.4.2 Cell sorting and RNA extraction

Frozen vials of 10×10^6 PBMCs were thawed and washed before staining with Pacific Blue labeled anti-CD3 (UCHT1), Pacific Blue labeled anti-CD14 (M5E2), FITC labeled anti-CD19 (HIB19), PE-Cy5 labeled anti-CD10 (HI10a), PE labeled anti-CD27 (M-T271), and APC labeled anti-CD21 (B-ly4), all from BD Biosciences. Sorts were done on a high speed BD FACSAria into miRVana lysis buffer (Ambion). All sorted populations were CD3-, CD14-, and CD19+. Immature B cells were CD10+, exhausted tissue-like memory were CD10-, CD21-, CD27-, activated mature B cells and resting memory B cells were combined in the CD10-, CD27+ gate and short-lived peripheral plasmablasts were CD10-, CD27++, CD21low [55]. Approximate cell yields per patient were: Immature, $1–2\times10^4$, Memory $0.2–1\times10^6$, Exhausted $5–10\times10^4$, Plasmablasts $2–5\times10^3$. Total RNA from the cells was then extracted using the mirVana RNA extraction kit (Ambion) according to manufacturer's instructions. RNA was quantitated on a 2100 Bioanalyzer (Agilent).

3.4.3 Full-repertoire sequencing library preparation

Total RNA was reverse transcribed as follows. Each sample was used for two heavy chain and two light chain RT reactions. For each reaction, 9.5 µL of total RNA was combined with 2.5 µL gene-specific primer (2 µM each), and 1 µL of 10 mM dNTP. The mixture was incubated at 65°C for 5 min, then 1 min on ice. Meanwhile, for each reaction the following mix was prepared: 4 µL of 5x FS buffer, 1 µL 0.1 M DTT, 40 U RNase inhibitor (Enzymatics), and 1 µL SuperScript III RT (200 U/µL; Invitrogen). The mix was added to the reaction and incubated at 55°C for 60 min, followed by inactivation at 70°C for 15 min. The two cDNA reactions for each sample-locus combination were combined prior to PCR, and the RNA/DNA hybrid was removed with 2 µL RNase H (Enzymatics) incubated at 37°C for 20 min.

For each sample-locus combination, four 50 µL PCR reactions were performed from the corresponding cDNA samples. Each reaction was comprised as follows: 27.5 µL water, 10 µL cDNA, 10 µL 5x HF buffer (detergent-free), 1 µL 10 mM dNTP, 0.5 µL each primer (from 2.5

μM stock), and 0.5 μL Phusion II Hot Start polymerase. Note that the J-side primers included sample barcodes, and all primers included adaptor sequences for 454 emPCR. The reactions were cycled as follows.

Heavy chain:

1. 98°C, 30 s
2. 98°C, 10 s
3. 58°C, 30 s
4. 72°C, 40 s
5. Goto 2, 20x
6. 72°C, 5 min
7. 10°C, forever

Light chain:

1. 98°C, 30 s
2. 98°C, 10 s
3. 65°C, 30 s
4. 72°C, 40 s
5. Goto 2, 19x
6. 72°C, 5 min
7. 10°C, forever

The reactions were cleaned and concentrated with AMPure XP beads (Agencourt) used at the standard 1:1.8 ratio and eluted in 30 μL Tris buffer. The desired bands were finally purified using a Pippin prep 1.5% gel, gated from 400–480 bp (heavy) and 370–450 bp (light). The eluates were cleaned on a QIAquick column (Qiagen), eluted in 30 μL, and quantitated on a 2100 Bioanalyzer. The samples were finally mixed to produce equimolar ratios, and sent to 454 Life Sciences for sequencing. The sequencing was performed according to manufacturer's protocol.

3.4.4 Family-specific sequencing library preparation

Reverse transcription was performed as follows. 10 µL total RNA was combined with 2 µL RT primer mix (50 µM oligo-dT and 25 µM random hexamer). The mixture was heated at 95°C for 1 min, 65°C for 5 min, then cooled on ice for 1 min. Then the following were added: 4 µL 5x FS buffer, 1 µL 10 mM dNTP mix, 1 µL 0.1 M DTT, 1 µL RNase inhibitor (Enzymatics), 1 µL SuperScript III RT (Invitrogen), and incubated as follows: 25°C for 10 min, 35°C for 5 min, 55°C for 45 min, 85°C for 5 min. RNA/DNA hybrid was removed by adding 4 µL E. coli RNase H (Enzymatics).

PCR reactions were assembled as follows: 13.75 µL water, 5 µL cDNA, 5 µL 5x HF buffer, 0.5 µL 10 mM dNTP, 0.25 µL of each 100 µM primer stock, 0.25 µL Phusion Hot Start. Cycle:

1. 98°C, 60 s
2. 98°C, 10 s
3. 62°C, 20 s
4. 72°C, 20 s
5. Goto 2, 24x
6. 72°C, 5 min
7. 4°C, forever

Samples were purified on a QIAquick column, and run on a 2% agarose gel. The desired bands were cut out, cleaned on Qiagen MinElute gel extraction kit, eluted twice with 10 µL EB buffer, and quantitated on a 2100 Bioanalyzer. Samples were sent to SeqWright for 454 sequencing, which was performed per manufacturer's instructions.

3.4.5 Raw data processing: VDJ alignment and clone definition

Raw sequencing data were processed using in-house tools written in python. Reads were split into barcodes (if necessary), size-filtered, and aligned to IMGT's germline VDJ reference database. The alignment was accomplished by performing semi-global dynamic programming alignment of each read against all possible germline sequences, choosing the best match if it met a minimum threshold score. The scores were kept low, as we were interested in sequences that were very highly mutated. The V region is aligned first, then removed, followed by J, then removed, followed by D. The IMGT-defined CDR3 sequence of each read was then extracted. The

CDR3 sequences were sorted by abundance and clustered with USEARCH5.1 with the options "`-minlen 0 -global -usersort -iddef 1 -id 0.9`". Finally, each CDR3 sequence was aligned to the "target" antibody sequences of PGT121-123 to determine a "divergence" value from these antibodies

3.4.6 Antibody variant identification and analysis

The divergence-mutation plots are used a a tool to "fish" for reads that are similar to the known PGT121–123 antibodies. High-identity clusters of sequences and clusters that are above "background" (i.e., the large mass of reads) are manually identified and used as input for a phylogeny inference algorithm specifically designed for SHM, Immunitree.

Immunitree implements a Bayesian model of somatic hypermutation of clones, including probabilistic models of SHM and sequencing error, and also allows for the observation of intermediate nodes in the tree [23]. It performs Markov chain Monte Carlo over the tree structure, birth/death times of the subclones, birth/death, mutation, and sequencing error rates, subclone consensus sequences, and assignment of reads to nodes.

Nodes were were chosen on an ad hoc basis for synthesis. The underlying node consensus sequence was analyzed for possible out-of-frame indels and manually corrected. (This was sometimes necessary for precursor nodes for which there was a dearth of data.)

The tree structure is also used for multiple computations and to overlay different information. Significantly, we estimate the selection pressure that a given node has experienced using the BASELINe algorithm [18]. It performs a Bayesian estimation of selection pressure by comparing the observed number of replacement/silent mutations in the CDRs/FWRs of the node consensus sequence.

3.4.7 Software tools

Raw data analysis tools are available here:

- https://github.com/laserson/vdj
- https://github.com/laserson/pytools

Figures were generated primarily with matplotlib, and trees visualized with Graphviz. Immunitree is implemented in MATLAB, and BASELINe is implemented in R.

3.4.8 Antibody and protein expression and purification

Antibody sequences were synthesized by GeneWiz and cloned into heavy and light chain Nuss Vectors. The plasmids were then co-transfected at a 1:1 ratio in either HEK 293T or 293

FreeStyle cells using Fugene 6 (Promega) or 293fection (Invitrogen), respectively. Transfections were performed according to the manufacturer's protocol and antibody supernatents were harvested four days following transfection. Antibodies produced in 293T cells were quantified by ELISA as described below and used directly in neutralization assays. Antibodies produced in 293 freestyle cells were further purified over a protein A column and dialyzed against phosphate-buffered saline. Mutations were introduced by site-directed mutagenesis using the QuikChange site-directed mutagenesis kit (Stratagene) and mutants were verified by Sanger DNA sequencing.

3.4.9 Pseudovirus production and neutralization assays

To produce pseudoviruses, plasmids encoding Env were co-transfected with an Env-deficient genomic backbone plasmid (pSG3ΔEnv) in a 1:2 ratio with the transfection reagent Fugene 6 (Promega). Pseudoviruses were harvested 72 hours post transfection for use in neutralization assays. Neutralizing activity was assessed using a single round of replication pseudovirus assay and TZM-bl target cells, as described previously. Briefly, TZM-bl cells containing the luceriface reporter gene were seeded in a 96-well flat bottom plate at a concentration of 20000 cells/well. The serially diluted virus/antibody mixture, which was pre-incubated for 1 hr, was then added to the cells and luminescence was quantified 72 hrs following infection via lysis and addition of Bright-GloTM Luciferase substrate (Promega). To determine IC50 values, serial dilutions of mAbs were incubated with virus and the dose-response curves were fitted using nonlinear regression.

3.4.10 ELISA assays

Ninety-six-well ELISA plates were coated overnight at 4°C with 50 uL PBS containing 100 ng of goat anti-human IgG Fc (Pierce) or 100 ng of gp120 per well. The wells were washed four times with PBS containing 0.05% Tween 20 and blocked with 3% BSA at room temperature for 1 h. Serial dilutions or mAb were then added to the wells, and the plates were incubated at room temperature for 1 hour. After washing four times, goat anti-human IgG F(ab')2 conjugated to alkaline phosphatase (Pierce) was diluted 1:1000 in PBS containing 1% BSA and 0.025% Tween 20 and added to the wells. The plate was incubated at room temperature for 1 h, washed four times, and the plate was developed by adding 50 uL of alkaline phosphatase substrate (Sigma) to 5 mL alkaline phosphatase staining buffer (pH 9.8), according to the manufacturer's instructions. The optical density at 405 nm was read on a microplate reader (Molecular Devices). Antibody concentration was calculated by linear regression using a standard concentration curve of purified IgG protein.

3.4.11 Cell surface binding assays

To produce cell surface Env trimer, plasmids encoding Env were co-transfected with an Env-deficient genomic backbone plasmid (pSG3ΔEnv) in a 1:2 ratio with the transfection reagent Fugene 6 (Promega). The cells are then harvested 48 hrs following transfection and the supernatent discarded. Titrating amounts of mAbs were added to the transfected cells and incubated for 1h at 4°C in 1x PBS. The cells were washed three times with 1x PBS and fixed with 2% PFA (vendor) for 20 min at RT. Following three washes with 1x PBS, the cells were then stained with a 1:200 dilution of goat anti-human IgG F(ab')2 conjugated to phycoerythrin (Jackson) for 1h at RT. Binding was analyzed using flow cytometry, and binding curves were generated by plotting the mean fluorescence intensity of antigen binding as a function of antibody concentration. FlowJo software was used for data interpretation.

3.5 Author contributions

Uri Laserson performed sequencing experiments and data analysis. Francois Vigneault performed experiments. Raj Chari performed germline analysis. Devin Sok performed antibody characterization experiments and data analysis. Laura Walker, Khoa Le, Karen Saye, and Alejandra Ramos performed experiments. David Smith, Caroline Ignacio, Birgitte Simen, and Elizabeth St. John performed sequencing services. Jonathan Laserson and Yi Pei Liu performed antibody phylogeny analysis. Alison Mahan performed cell sorts. Gur Yaari performed selection analysis. George Church, Daphne Koller, Galit Alter, Steven Kleinstein, Pascal Poignard, and Dennis Burton supervised research.

Chapter 4

Autoantigen Discovery With a Synthetic Human Peptidome[1]

4.1 Abstract

In this study, we improve on current autoantigen discovery approaches by creating a synthetic representation of the complete human proteome, the T7 "peptidome" phage display library (T7-Pep), and use it to profile the autoantibody repertoires of individual patients. We provide methods for 1) designing and cloning large libraries of DNA microarray-derived oligonucleotides encoding peptides for display on bacteriophage, and 2) analyzing the peptide libraries using high throughput DNA sequencing. We applied phage immunoprecipitation sequencing (PhIP-Seq) to identify both known and novel autoantibodies contained in the spinal fluid of three patients with paraneoplastic neurological syndromes. We also show how our approach can be used more generally to identify peptide-protein interactions and point toward ways in which this technology will be further developed in the future. We envision that PhIP-Seq can become an important new tool in autoantibody analysis, as well as proteomic research in general.

4.2 Introduction

Vertebrate immune systems have evolved sophisticated genetic mechanisms to generate antibody repertoires, which are combinatorial libraries of affinity molecules capable of distinguishing between self and non-self. Recent data highlight the delicate balance in higher mammals between energy utilization, robust immune defense against pathogens, and autoimmunity[56]. In humans, loss of tolerance to self-antigens results in a number of diseases including type I

[1] Published as [26]: Larman, et al., Autoantigen Discovery With a Synthetic Human Peptidome, *Nature Biotechnology*, 29 (6): 535–541, 2011.

diabetes, multiple sclerosis, and rheumatoid arthritis. Knowledge of the self antigens involved in autoimmune processes is not only important for understanding the disease etiology, but can also be used to develop accurate diagnostic tests. In addition, physicians may someday utilize antigen-specific therapies to target auto-reactive immune cells for destruction or quiescence.

Traditional approaches to identification of autoantibody targets largely rely on expression of fragmented cDNA libraries. Important technical limitations of this method include the small fraction of clones expressing in-frame coding sequences (with a lower bound of 6%)[57], and the highly skewed representation of differentially expressed cDNAs. Nevertheless, expression cloning has led to the discovery of many important autoantigens[58–60]. Strides have been made to improve peptide display systems[61, 62], but there remains an important unmet need for better display libraries and methods to analyze binding interactions.

Here, we have constructed the first synthetic representation of the complete human proteome, which we have engineered for display as peptides on the surface of T7 phage. This T7 "peptidome" library (T7-Pep) was extensively characterized and found to be both faithful to its in silico design and uniform in its representation. We combined our T7-Pep library with high-throughput DNA sequencing to identify autoantibody-peptide interactions, a method we call phage immunoprecipitation sequencing (PhIP-Seq). This approach provides several advantages over traditional methods, including comprehensive and unbiased proteome representation, peptide enrichment quantification, and a streamlined, multiplexed protocol requiring just one round of enrichment. We have applied PhIP-Seq to interrogate the autoantibody repertoire in the spinal fluid of patients with neurological autoimmunity and identified both known and novel autoantigens. We further demonstrate how PhIP-Seq can also be used more generally to identify peptide-protein interactions.

4.3 Results

4.3.1 Construction and characterization of the T7-Pep library

We sought to create a synthetic representation of the human proteome. We began by extracting all open reading frame (ORF) sequences available from build 35.1 of the human genome (24239; 23% of which had "predicted" status). When there were multiple isoforms of the same protein, we randomly selected one representative ORF. We modified the codon usage by eliminating restriction sites used for cloning and by substituting very low abundance codons in E. coli with more abundant synonymous codons. We parsed this database into sequences of 108 nucleotides encoding 36 amino acid tiles with an overlap of seven residues between consecutive peptides (Figure 4.1), the estimated size of a linear epitope. Finally, the stop codon of each ORF was

removed so that all peptides could be cloned in-frame with a C-terminal FLAG tag.

The final library design includes 413611 peptides spanning the entire coding region of the human genome. The peptide-coding sequences were synthesized as 140-mer oligonucleotides with primer sequences on releasable DNA microarrays in 19 pools of 22000 oligos each, PCR-amplified and cloned into a derivative of the T7Select 10-3b phage display vector (Novagen; Figure 4.1b i and Supplementary Methods). We also generated two additional libraries comprising the N-terminal and C-terminal peptidomes (T7-NPep, T7-CPep), which encode only the first and last 24 codons from each ORF.

The extent of vector re-ligation, multiple insertions, mutations, and correct in-frame phage-displayed peptides was determined by plaque PCR analysis (Table 4.1), clone sequencing (Figure 4.1c), and FLAG expression (Table 4.2) of randomly sampled phage from all subpools. Sequencing revealed that 83% of the inserts lacked frameshifting mutations. These data indicate that a much greater fraction of in-frame, ORF-derived peptides is expressed by our synthetic libraries compared to those constructed from cDNA (Table 4.3).

After combining 5 x 108 phage from each subpool and amplifying the final library, Illumina sequencing was performed at a median depth of 45-fold coverage (Figure 4.1d) and detected 91.2% of the expected clones. Chao1 analysis was performed to estimate the actual library complexity (assuming infinite sampling), which predicted that >91.8% of the library was represented (Figure 4.2) [63]. In addition, T7-Pep is highly uniform, with 78% of the library members within 10-fold abundance (having been sequenced between 10 and 100 times). These data suggest that our library encodes a much more complete and uniform representation of the human proteome than can otherwise be achieved with existing technologies (Table 4.3).

We next optimized a phage immunoprecipitation protocol for detecting antibody-peptide interactions within complex mixtures (Figure 4.3). By combining this protocol with T7-Pep and deep sequencing DNA analysis, we have developed a new method to quantitatively profile autoantibody repertoires in patients (Figure 4.1b).

4.3.2 Analysis of a PND patient with NOVA autoantibodies

Cancers often elicit cellular and humoral immune responses against tumor antigens which may limit disease progression[64]. In rare cases, tumor immunity can recognize central nervous system (CNS) antigens, triggering a devastating autoimmune process called paraneoplastic neurological disorder (PND). Clinical presentations of PND are heterogeneous and correlate with the CNS autoantigens involved. PND has served as a model for CNS autoimmunity, and the application of phage display to PND autoantigen discovery has met with much success [58, 65].

To assess the performance of PhIP-Seq for autoantigen discovery, we examined a sample of cerebrospinal fluid (CSF) from a 63-year-old female (Patient A) with non-small cell lung cancer

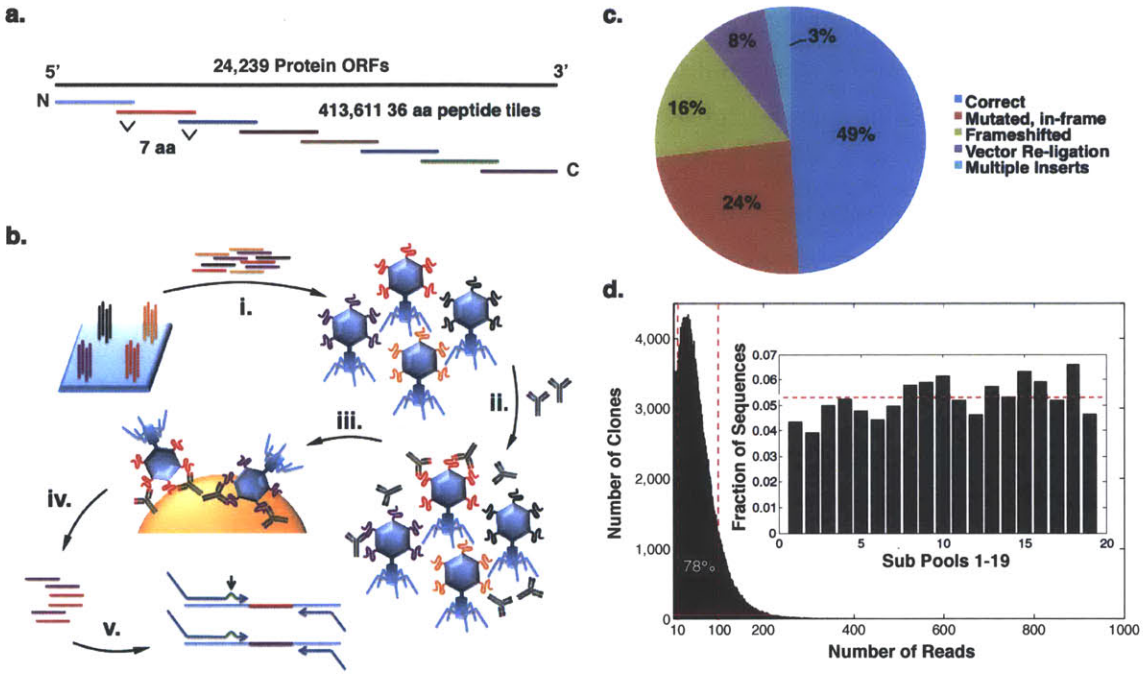


Figure 4.1: Construction and characterization of T7-Pep and the PhIP-Seq methodology. (a) The T7-Pep library is made from 413611 DNA sequences encoding 36 amino acid peptide tiles that span 24239 unique ORFs from build 35.1 of the human genome. Each tile overlaps its neighbors by seven amino acids on each side. (b) The DNA sequences from (a) were printed as 140-mer oligos on releasable DNA microarrays. (i) After oligo release, the DNA was PCR-amplified and cloned into a FLAG-expressing derivative of the T7Select 10-3b mid copy phage display system. (ii) The T7-Pep library is mixed with patient samples containing autoantibodies. (iii) Antibodies and bound phage are captured on magnetic protein A/G coated beads. (iv) DNA from the immunoprecipitated phage is recovered and (v) library inserts are PCR-amplified with sequencing adapters. A single nucleotide change (arrow) is introduced for multiplex analysis. (c) Pie chart showing results of plaque sequencing of 71 phage from T7-Pep Pool 1 and T7-CPep Pool 1. (d) Histogram plot showing results from Illumina sequencing of T7-Pep. 78% of the total area lies between the vertical red lines at 10 and 100 reads, demonstrating the relative uniformity of the library. Representation of each subpool in T7-Pep (inset) compared to expected (horizontal red line).

Pool	Plaques analyzed	Plaques with multiple inserts	% Multiple inserts	Plaques with vector religation	% Vector Religation
T7-Pep pool 1	45	1	2.2	1	2.2
T7-Pep pool 2	39	3	7.7	0	0.0
T7-Pep pool 3	39	1	2.6	0	0.0
T7-Pep pool 4	38	3	7.9	0	0.0
T7-Pep pool 5	38	2	5.3	0	0.0
T7-Pep pool 6	39	0	0.0	0	0.0
T7-Pep pool 7	31	1	3.2	0	0.0
T7-Pep pool 8	62	3	4.8	1	1.6
T7-Pep pool 9	54	0	0.0	0	0.0
T7-Pep pool 10	31	1	3.2	0	0.0
T7-Pep pool 11	62	3	4.8	1	1.6
T7-Pep pool 12	69	1	1.4	4	5.8
T7-Pep pool 13	31	0	0.0	0	0.0
T7-Pep pool 14	31	1	3.2	0	0.0
T7-Pep pool 15	31	1	3.2	1	3.2
T7-Pep pool 16	31	0	0.0	1	3.2
T7-Pep pool 17	30	1	3.3	0	0.0
T7-Pep pool 18	30	1	3.3	0	0.0
T7-Pep pool 19	31	1	3.2	0	0.0
T7-NPep pool 1	46	3	6.5	1	2.2
T7-CPep pool 1	47	2	4.3	0	0.0
T7-NPep pool 2	48	0	0.0	3	6.3
T7-CPep pool 2	44	1	2.3	1	2.3
Total	**947**	**30**	**3.2**	**14**	**1.5**

Table 4.1: Subpool analysis of multiple insertions and vector re-ligation after cloning of the T7-Pep, T7-NPep, and T7-CPep libraries. Phage plaques from each subpool were randomly selected and PCR analyzed to examine the frequency of multiple insertions and vector religations present within each pool.

Pool	FLAG-positive plaques	T7 tail fiber positive plaques	% in-frame phage
T7-Pep pool 2	44	69	64%
T7-Pep pool 3	61	94	65%
T7-Pep pool 4	43	64	67%
T7-Pep pool 5	48	70	69%
Total	196	297	66%

Table 4.2: Subpool analysis of FLAG expression after cloning of T7-Pep. Plaque lifts from four subpools were analyzed by immunoblotting using FLAG and T7 tail fiber antibodies to measure in-frame and total plaques, respectively. Plaques staining positive were counted and a percentage of in-frame, FLAG-expressing phage was determined. The vast majority of frameshifting mutations present in the phage inserts is due to errors in DNA chemical synthesis on the releasable DNA microarrays. In parallel oligonucleotide synthesis, sequence integrity can be compromised by depurination side reactions, inefficient nucleoside coupling, and reversible 5'-hydroxyl deprotection reactions, leading to mutations of the desired oligonucleotide.

Feature	Classic cDNA Phage Display	Protein Array	T7-Pep + PhIP-Seq
Proteome representation	• Incomplete • Highly skewed distribution	• Small fraction • Uniform distribution	• Nearly complete • Uniform distribution
Fraction of clones expressing an ORF peptide in frame	As low as 6%	Up to 100%	~83%
Size of displayed peptides	Up to full-length proteins	Up to full-length proteins	36 amino acid overlapping tiles
Rounds of selection	Requires multiple selection rounds, which favor more abundant and faster growing clones[79]	No selection	Single selection, which eliminates clone growth bias and population bottleneck
Analysis	Individual clone sequencing: • Initial abundance unknown • Requires population bottleneck	Microarray scanning: • Quantitative • Statistical analysis of antibody binding	Deep sequencing of library: • Quantify population before and after a single round of selection • Statistical analysis of enrichments
Determination of antibody polyclonality	Difficult	Not possible	Often straightforward for antigens of known crystal structure
Epitope mapping	Difficult	Not possible	Often straightforward
Effort	Labor intensive	Minimal	Minimal
Sample throughput	Low	Medium	Adaptable to 96 well format
Multiplexing capability	No	No	Yes
Cost	Low	Moderate to high	Moderate

Table 4.3: Comparison between T7-Pep + PhIP-Seq and current proteomic methods for autoantigen discovery

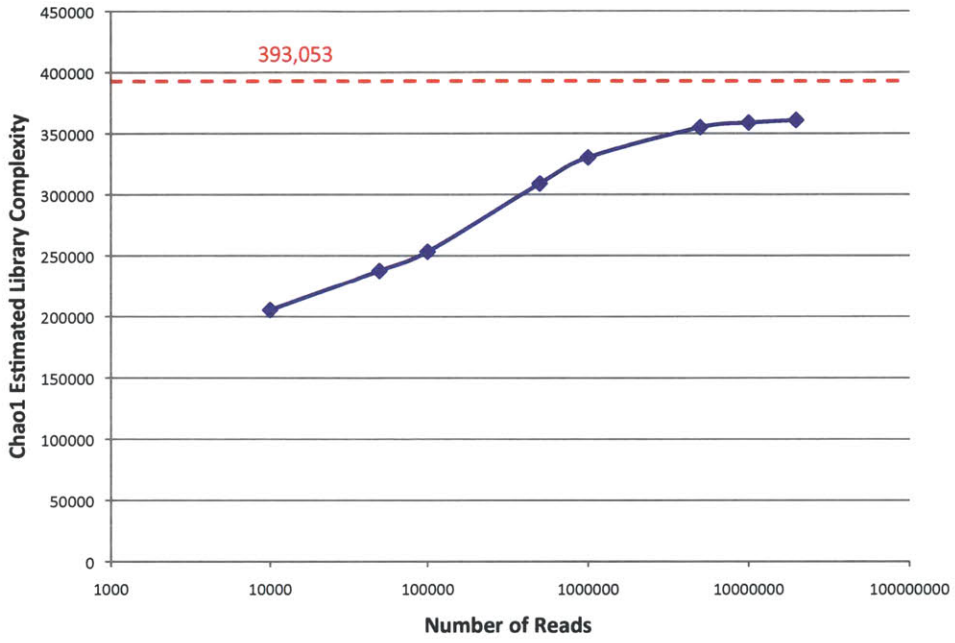


Figure 4.2: The effect of sequencing depth on estimated library complexity. Chao1 estimates of library complexity given by $S_{\text{Chao1}} = S_{\text{obs}} + n_1^2/2n_2$ are shown as a function of simulated T7-Pep library sampling. S_{Chao1} is the estimated complexity, where S_{obs} is the observed library complexity, n_1 is the number of library members observed once, and n_2 is the number of library members observed twice. For the data points shown, S_{obs}, n_1, and n_2 were simulated by randomly sampling the actual sequencing data "Number of Reads" times without replacement. S_{Chao1} was then calculated as above. The sequencing depth actually achieved, 20 million reads, appears to be near saturating with respect to Chao1 estimate of the library complexity, at 361070 library members (or 91.8% of the 393053 resolvable clones).

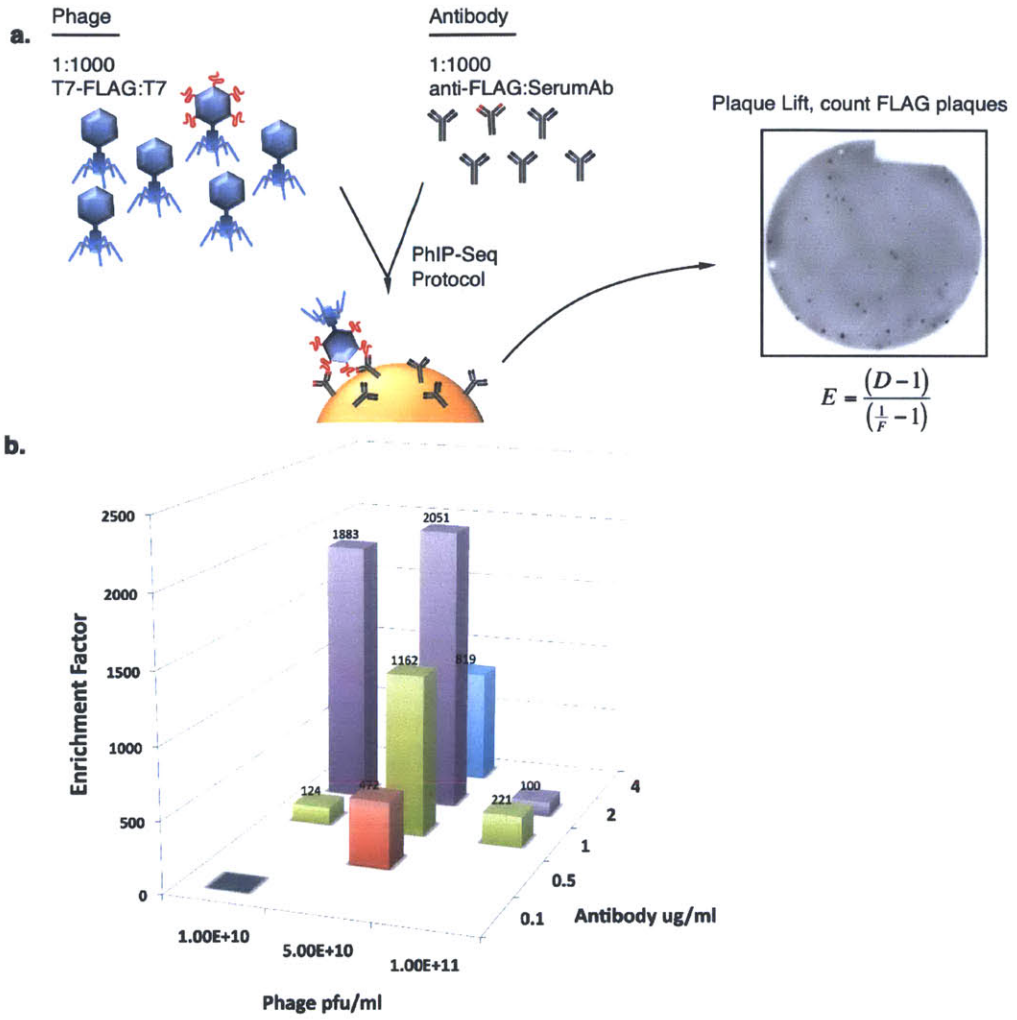


Figure 4.3: Optimization of PhIP-Seq target enrichment. A. A FLAG-expressing T7 phage (depicted with red peptide) was diluted at 1:1,000 into native, non-FLAG-expressing T7 phage to mimic a target peptide within the T7-Pep library. An anti-FLAG monoclonal antibody (M2, Sigma-Aldrich; shown with red variable region) was diluted 1:1,000 into human serum antibodies (shown with black variable region) to mimic a rare autoantibody within a patient's antibody repertoire. After performing the IP, plaque lift analysis for FLAG expression was performed to determine enrichment using the equation shown (E = enrichment; D = dilution factor = 1,000; F = fraction of FLAG expressing clones on plaque lift). Enrichment was optimized with respect to type of beads, number of washes, order of antibody-phage/antibody-bead complex formation, and relative concentrations of phage and antibody. B. Enrichment factor was found to depend on the relative concentrations of phage and antibody during complex formation. We thus varied these parameters independently and found an optimum at about 5x1010 pfu/ml phage and 2 mg/ml total antibody.

(NSCLC) who presented with a PND syndrome and was found to have anti-NOVA autoantibodies [66]. The NOVA autoantigen (neuro-oncological ventral antigen, or "Ri") is commonly targeted in PND triggered by lung or gynecological cancers, and results in ataxia with or without opsoclonus/myoclonus. A concentration of 2 µg/ml of CSF antibody was spiked with 2 ng/ml of an antibody specific to SAPK4 (positive control) to monitor enrichment of the targeted peptide on protein A/G beads. Despite extensive washing, 298667 unique clones (83% of the input library) were found in the immunoprecipitate. A significant correlation was observed between the abundance of input clones and immunoprecipitated clones (Figure 4.4a), likely due to weak nonspecific interactions with the beads.

To approximate the expected distribution of IP'ed clones' abundances, we employed a two-parameter generalized Poisson (GP) model (as recently demonstrated for RNA-seq data [68]) and found that this distribution family fits the data well at various input abundances (Figure 4.4b). We calculated the GP parameter values for each input abundance level [67] and regressed these parameters to form our null model for the calculation of enrichment significance (p-values) of each clone (Figure 4.4c and online methods). Comparing the two PhIP-Seq replicates revealed that the most significantly enriched clones were the same in both replicas (Figure 4.4d), highlighting the assay's reproducibility. This contrasts dramatically with a comparison of clones enriched by two different patients (Figure 4.5). Performing PhIP-Seq in the absence of patient antibodies identified phage capable of binding to the protein A/G beads. We thus defined Patient A positive clones as those clones with a reproducible Log10 p-value greater than a cutoff (Figure 4.4d, dashed red line), but not significantly enriched on beads alone ($P < 10^{-3}$). Patient A positives included the expected SAPK4-targeted positive control peptide ($P < 10^{-15}$), the expected NOVA1 autoantigen ($P < 10^{-15}$), and six additional candidate autoantigens (Table 4.4).

We tested three of these predictions by expressing full-length TGIF2LX, DBR1 and PCDH1 in 293T cells and immunoblotting with patient CSF. TGIF2LX (TGFB-Induced factor homeobox 2-like, X-linked) was confirmed as a novel autoantigen, as we detected strong immunoreactivity at the expected molecular weight (Figure 4.6a). Full-length DBR1 and PCDH1, while expressed well in 293T cells (not shown), were not detected by CSF antibodies. We observed two bands in the untransfected lysate migrating at approximately 50 and 62 KDa, possibly representing endogenously expressed proteins that correspond either to untested candidates or to false negatives of the PhIP-Seq assay.

Strikingly, the hypothetical protein LOC26080 had seven distinct peptides that were significantly enriched, and they all appeared to share a nine residue repetitive motif. We used MEME software [69] to characterize this motif, which represents the likely epitope recognized by Patient A's autoantibodies (Figure 4.6b).

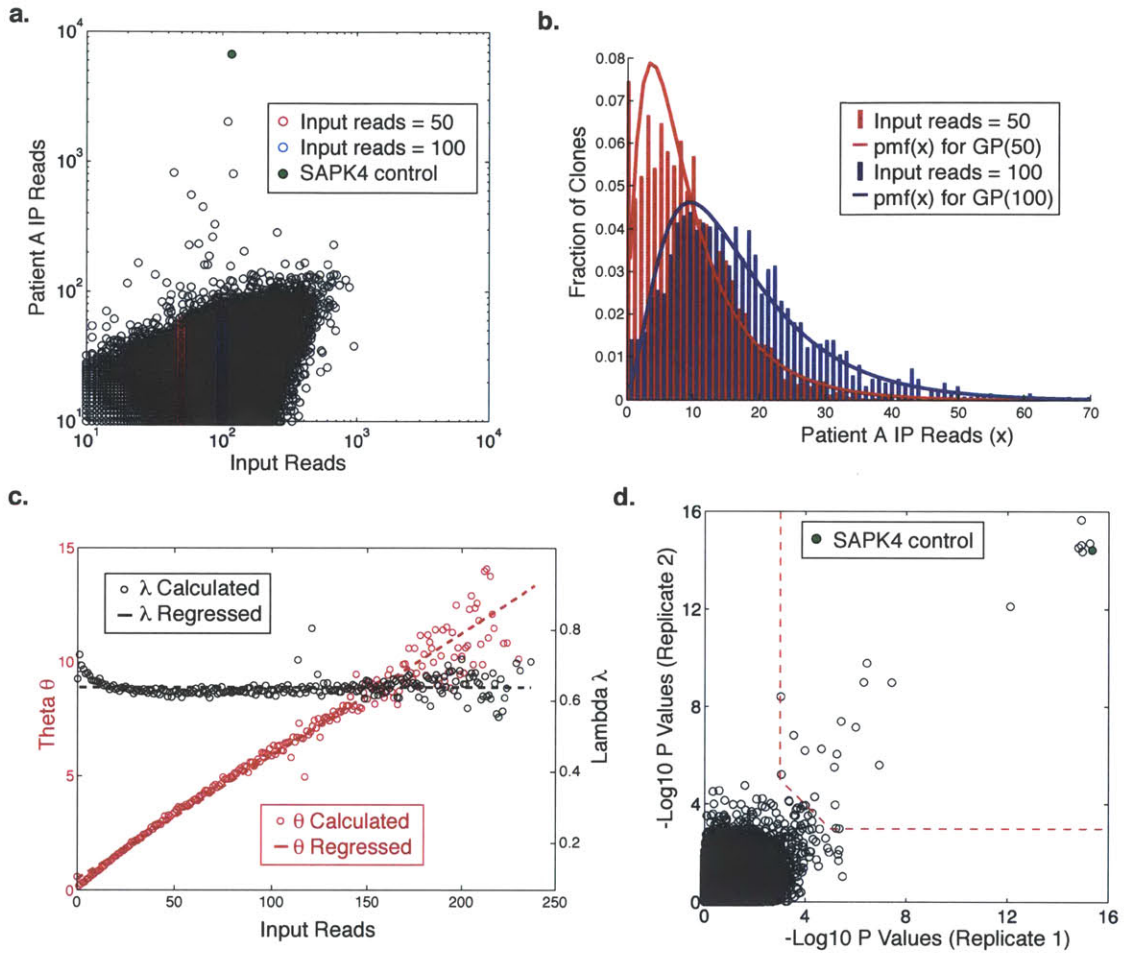


Figure 4.4: Statistical analysis of PhIP-Seq data. (a) Scatter plot comparing sequencing reads from T7-Pep input library and from Patient A immunoprecipitated (IP) phage (Pearson coefficient = 0.435; P = 0). Highlighted are all clones with an input abundance of 50 reads (red), and all clones with an input abundance of 100 reads (blue). The target of the SAPK4 control antibody is highlighted in green. (b) Histogram plot of sequencing reads from the data highlighted in (a) with corresponding colors. The curves are fit with a generalized Poisson (GP) distribution. pmf is the probability mass function of the corresponding GP distribution and x is the number of IP sequencing reads. (c) Plots of lambda and theta for each input abundance, calculated using the method of Consul et al [67]. Lambda is regressed to its average value (black dashed line) and theta is linearly regressed (red dashed curve). (d) Scatter plot comparing clone enrichment significances (as -Log10 p-value) from two independent PhIP-Seq experiments using CSF from Patient A. Red dashed line shows the cutoff for considering a clone to be significantly enriched, and the SAPK4 control antibody target is highlighted in green.

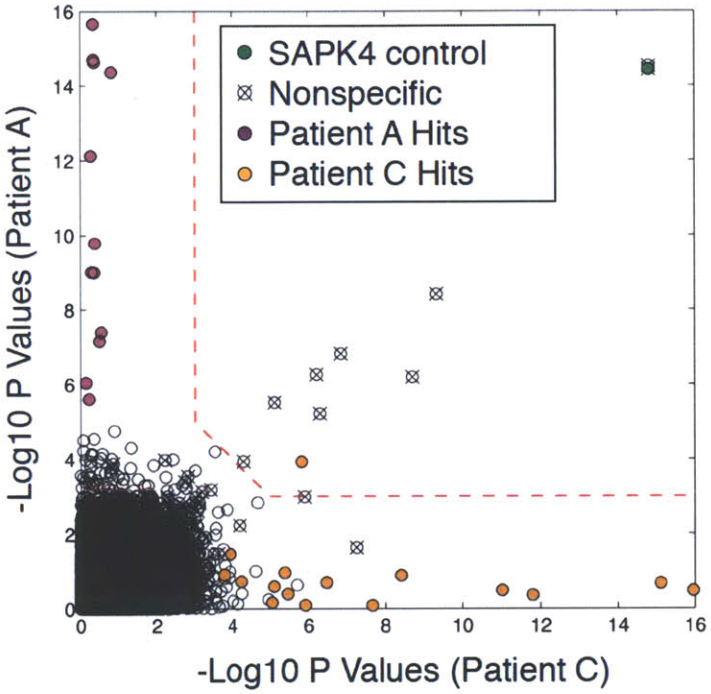


Figure 4.5: Comparison of PhIP-Seq experiments on different patients. Scatter plot as in Figure 4.4d from text, but comparing clone enrichment p-values from two different patients: Patient A (y-axis) versus Patient C (x-axis). Both experiments included the SAPK4 spike-in antibody. X'ed circles were enriched by beads and SAPK4 antibody alone (no patient antibody in IP). Filled purple and orange circles are the Patient A- and Patient C-specific positives given in Table 4.4 from the text.

Patient Info	-Log10 P value	Protein	Peptides	Validation
A: 63 y.o. female with non-small cell lung cancer. Presents with classic cerebellar syndrome. ***CSF positive for anti-NOVA antibodies.***	*15.38*	*NEURO-ONCOLOGICAL VENTRAL ANTIGEN 1 (NOVA1)*	*1*	***WB+***
	14.76	**HYPOTHETICAL PROTEIN LOC26080**	**7**	**DB+**
	14.54	**TGFB-INDUCED FACTOR HOMEOBOX 2-LIKE, X-LINKED (TGIF2LX)**	**1**	**WB+**
	8.00	NEBULIN (NEB)	1	NT
	6.49	**DEBRANCHING ENZYME HOMOLOG 1 (DBR1)**	**1**	**WB-,DB+**
	6.20	**PROTOCADHERIN 1 (PCDH1)**	**1**	**WB-,DB+**
	4.29	INSULIN RECEPTOR (INSR)	1	NT
B: 59 y.o. female with non-small cell lung cancer. Presents with dysarthria, ataxia, head titubation and muscle lock. Paraneoplastic antibody panel is negative.	15.18	SOLUTE CARRIER FAMILY 25 MEMBER 43 (SLC25A43)	1	NT
	13.06	**GLUTAMATE DECARBOXYLASE 2 (GAD65)**	**2**	**RIA+,WB-, IP+**
	12.96	**TESTIS EXPRESSED SEQUENCE 2 (TEX2)**	**1**	**DB+**
	12.11	ATAXIN 7-LIKE 3 ISOFORM B (ATXN7L3)	1	NT
	11.93	ETS-RELATED TRANSCRIPTION FACTOR ELF-1 (ELF1)	1	NT
	11.91	**TGFB-INDUCED FACTOR HOMEOBOX 2-LIKE, X-LINKED (TGIF2LX)**	**1**	**WB+**
	11.34	INSULIN RECEPTOR SUBSTRATE 4 (IRS4)	1	NT
	6.98	HEPATOMA-DERIVED GROWTH FACTOR-RELATED PROTEIN 2 (HDGFRP2)	1	NT
	6.60	TUBULIN, BETA (TUBB)	1	WB-
	6.54	**CANCER/TESTIS ANTIGEN 2 (CTAG2)**	**1**	**WB+**
	6.30	**DENN/MADD DOMAIN CONTAINING 1A (DENDD1A)**	**1**	**WB-,DB+**
	6.09	DOUBLESEX AND MAB-3 RELATED TRANSCRIPTION FACTOR (DMRT2)	1	NT
	5.53	TUDOR AND KH DOMAIN CONTAINING ISOFORM A (TDRKH)	1	NT
C: 59 y.o. female with melanoma. Presents with ataxia, dysarthria, horizontal gaze palsy. Paraneoplastic antibody panel is negative. However, CSF stained brain and cerebellar IHC slides.	**15.72**	**TRIPARTITE MOTIF-CONTAINING 67 (TRIM67)**	**2**	**WB+**
	15.65	**TRIPARTITE MOTIF-CONTAINING 9 (TRIM9)**	**3**	**WB+**
	12.13	**FIBROBLAST GROWTH FACTOR 9 (GLIA-ACTIVATING FACTOR) (FGF9)**	**1**	**WB-,DB+**
	10.18	**DUAL-SPECIFICITY TYROSINE-(Y)-PHOSPHORYLATION REGULATED KINASE 3 (DYRK3)**	**1**	**WB-,DB+**
	6.93	CENTROSOMAL PROTEIN 152KDA (CEP152)	1	NT
	6.57	TITIN (TTN)	1	NT
	6.34	NUCLEOPORIN LIKE 2 (NUPL2)	1	NT
	5.43	**HISTONE DEACETYLASE 1 (HDAC1)**	**1**	**WB-,DB+**
	5.36	**MITOCHONDRIAL RIBOSOMAL PROTEIN L39 (MRPL39)**	**1**	**WB-,DB+**
	5.35	**CHROMOSOME 10 OPEN READING FRAME 82 (C10ORF82)**	**1**	**WB-,DB+**
	5.15	NLR FAMILY, PYRIN DOMAIN CONTAINING 5 (NLRP5)	1	NT
	4.83	TASPASE, THREONINE ASPARTASE, 1 (TASP1)	1	NT
	4.70	KIAA0090	1	NT
	4.55	SERINE (OR CYSTEINE) PROTEINASE INHIBITOR, CLADE A (ALPHA-1 ANTIPROTEINASE, ANTITRYPSIN), MEMBER 9 (SERPINA9)	1	NT
	4.21	**PROTEIN TYROSINE PHOSPHATASE, NON-RECEPTOR TYPE 9 (PTPN9)**	**1**	**WB-,DB+**

Table 4.4: Results of PhIP-Seq for 3 Patients. A previously validated autoantigen is shown in italics. Autoantigens confirmed by any secondary assay are shown in bold. Confirmation of patient antibodies with the full-length protein via western blot is indicated by red type. Average of replicate -Log10 p-values are shown in column 2. If multiple peptides from the same ORF are enriched, the average -Log10 p-value of the most significantly enriched peptide is shown. Secondary validation assay abbreviations: WB = western blot of full-length proteins; IP = immunoprecipitation of full-length proteins followed by western blotting for the fusion tag; RIA = radioimmunoassay; DB = dot blot; NT = not tested. Validation assay is followed by "+" or "-" depending on whether the results were positive or negative.

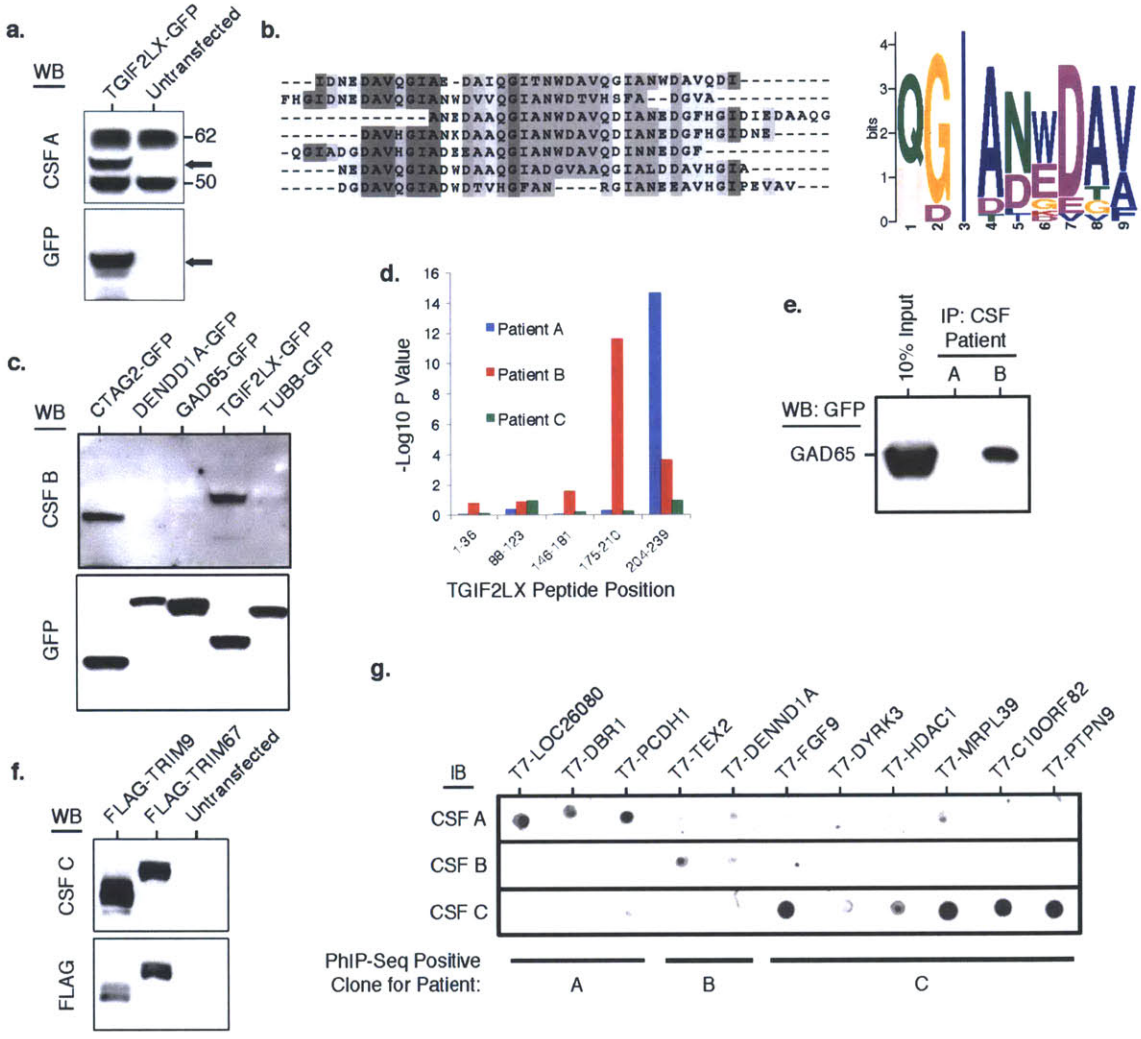


Figure 4.6: Validation of full-length PhIP-Seq candidates. (a) Western blot with CSF from Patient A, staining for full-length TGIF2LX-GFP expressed in 293T cells by transient transfection. Bands corresponding to TGIF2LX-GFP are denoted by an arrow. (b) ClustalW alignment of the seven significantly enriched hypothetical protein LOC26080 peptides, and the nine-element MEME-generated recognition motif. (c) Western blot with CSF from Patient B, staining for indicated full-length proteins expressed in 293T cells by transient transfection. (d) Bar graph of -Log10 p-values of enrichment for the indicated TGIF2LX peptides by the three patients. (e) Immunoprecipitation of the GAD65-GFP from 293T cell transfected lysate by CSF from Patient B (but not Patient A). (f) Western blot with CSF from Patient C, staining for indicated full-length proteinss expressed in 293T cells by transient transfection. (g) Phage lysates from candidate T7 clones were spotted directly onto nitrocellulose membranes, which were subsequently immunoblotted with patient CSF.

4.3.3 Analysis of two PND patients with uncharacterized autoantibodies

Having established that PhIP-Seq could reliably identify known and novel autoantigens, we examined CSF from two additional patients who had suggestive PND presentations but tested negative for a panel of commercially available PND autoantigens. Patient B was a 59-year-old female with NSCLC, who presented with dysarthria, ataxia, head titubation and muscular rigidity. PhIP-Seq analysis yielded three particularly interesting candidate autoantigens: TGIF2LX, CTAG2 (cancer/testis antigen 2), and GAD65 (glutamate decarboxylase 2) (Table 4.4). Both TGIF2LX and CTAG2 were confirmed by immunoblotting (Figure 4.6c). Surprisingly, Patient B, like Patient A, was auto-reactive against TGIF2LX. The enriched peptide was distinct from, but overlapped the peptide enriched by Patient A (Figure 4.6d).

CTAG2 is a member of a family of cancer/testis antigen (CTAG) proteins that are normally germ cell restricted, but frequently expressed in cancers and often elicit anti-tumor immune responses [70]. Several reports have described both humoral and cellular immune responses targeted against CTAG2 [71, 72]. TGIF2LX is also testis restricted [73, 74] and may be a new CTAG family member. As a negative control, we found TGIF2LX reactivity to be absent in the CSF of three patients with non-PND CNS autoimmunity and oligoclonal Ig bands (Figure 4.7). Having confirmed TGIF2LX autoreactivity in two NSCLC patients, we wondered whether it could be a new biomarker for this disease. However, the serum of 15 additional NSCLC patients without PND did not contain TGIF2LX antibodies detectable by immunoblotting (Figure 4.8).

Neither CTAG2 nor TGIF2LX is expressed in the brain, and thus are unlikely to explain the neurological syndrome experienced by Patient B. GAD65, however, is the rate-limiting enzyme in the synthesis of the inhibitory neurotransmitter GABA. GAD65 is also a well-characterized autoantigen targeted in the autoimmune disorder Stiff Person Syndrome (SPS; OMIM ID 184850). Two non-overlapping GAD65 peptides derived from the domain known to be targeted by pathogenic autoantibodies in SPS patients [75, 76] were enriched by Patient B's CSF. A commercial radioimmunoassay (RIA 81596; Mayo Medical Laboratories), confirmed the presence of high titer anti-GAD65 autoantibodies (5.12 nmol/L; >250 fold above the reference range). Surprisingly, direct immunoblotting with Patient B's CSF did not demonstrate reactivity (Figure 4.6c), suggesting that denatured GAD65 epitopes are not recognized by Patient B's antibodies. Successful immunoprecipitation of GAD65 from the same cell lysate with CSF confirmed this hypothesis (Figure 4.6e).

Patient C, a 59-year-old female with PND secondary to melanoma, had an unusual presentation that included horizontal gaze palsy. PhIP-Seq analysis of Patient C's CSF yielded five significantly enriched peptides from two homologous members of the tripartite motif (TRIM)

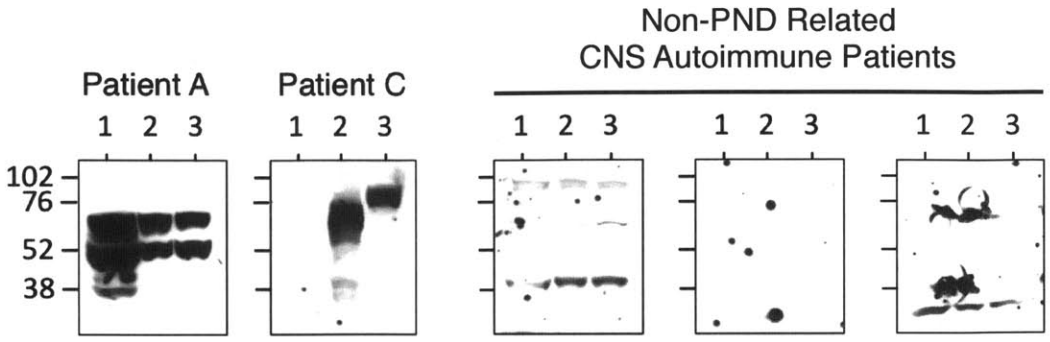


Figure 4.7: TRIM9 and TRIM67 autoreactivity is not present nonspecifically in CSF. Western blotting with CSF from Patients A and C, as well as three patients with non-PND related CNS autoimmune syndromes. In each blot, lanes 1, 2, and 3 were loaded with lysate from 293T cells overexpressing either TGIF2LX-GFP, FLAG-TRIM9, or FLAG-TRIM67, respectively.

family, TRIM9 and TRIM67 (Table 4.4). Both candidate autoantigens were confirmed by immunoblotting lysates from TRIM9- or TRIM67-overexpressing cells (Figure 4.6f). TRIM67 is expressed in some normal tissues (including skin) and is often highly expressed in melanoma [74]. TRIM9 has recently emerged as a brain-specific E3 ubiquitin ligase and has been implicated in neurodegenerative disease processes [77]. Based on their high degree of homology, our data suggest the possibility that tumor immunity targeting TRIM67 might have spread to, or cross-reacted with TRIM9 in the CNS (Figure 4.9). TRIM9 and TRIM67 autoreactivity was not detected in the CSF of three patients with non-PND CNS autoimmunity (Figure 4.7).

In total, 16 of the candidate autoantigens in Table 4.4 were available to us as full-length Gateway Entry clones from the ORFeome collection [78]. Of these, 10 were not confirmed by immunoblotting or immunoprecipitation of the full-length protein. We wondered whether this reflected a high rate of false positive discovery inherent to PhIP-Seq, or rather a requirement that the peptides be presented with intact conformation, as was the case for GAD65. We synthesized 9 of these 10 candidate T7 clones, plus 2 additional high confidence T7 clones, for validation in a dot blot assay. Each of these clones exhibited immunoreactivity above background with the appropriate patients' spinal fluid as predicted by the PhIP-Seq dataset (Figure 4.6g; Figure 4.10). This finding indicates that PhIP-Seq analysis can have a low rate of false positive discovery, and supports the hypothesis that the 36 amino acid peptides retain a significant amount of secondary structure during display on the T7 coat.

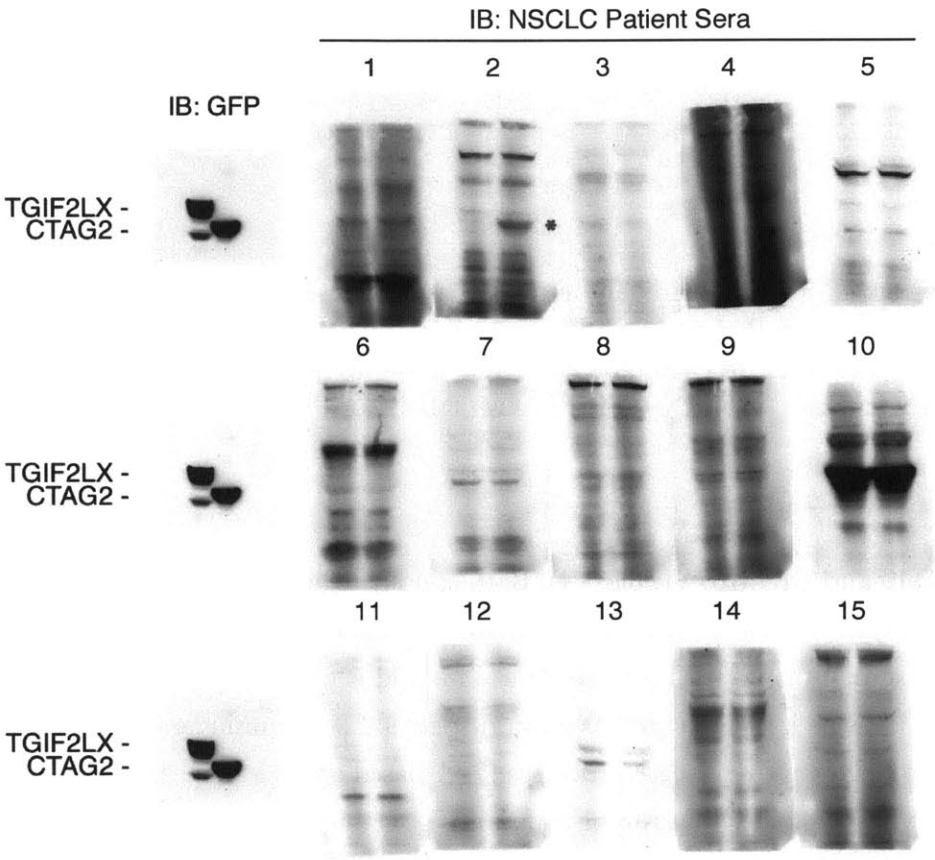


Figure 4.8: Immunoblots for TGIF2LX and CTAG2 reactivity in the serum of NSCLC patients without PND. Sera from fifteen non-small cell lung cancer (NSCLC) patients were used to blot SDS-PAGE separated 293T cell lysate overexpressing either TGIF2LX (left lane) or CTAG2 (right lane), fused with C-terminal GFP. Staining for GFP (left blot) demonstrates overexpression of TGIF2LX and CTAG2 at the expected weights. Only patient 2 was found to have anti-CTAG2 serum antibodies (marked by *). No patients were found to have anti-TGIF2LX serum antibodies.

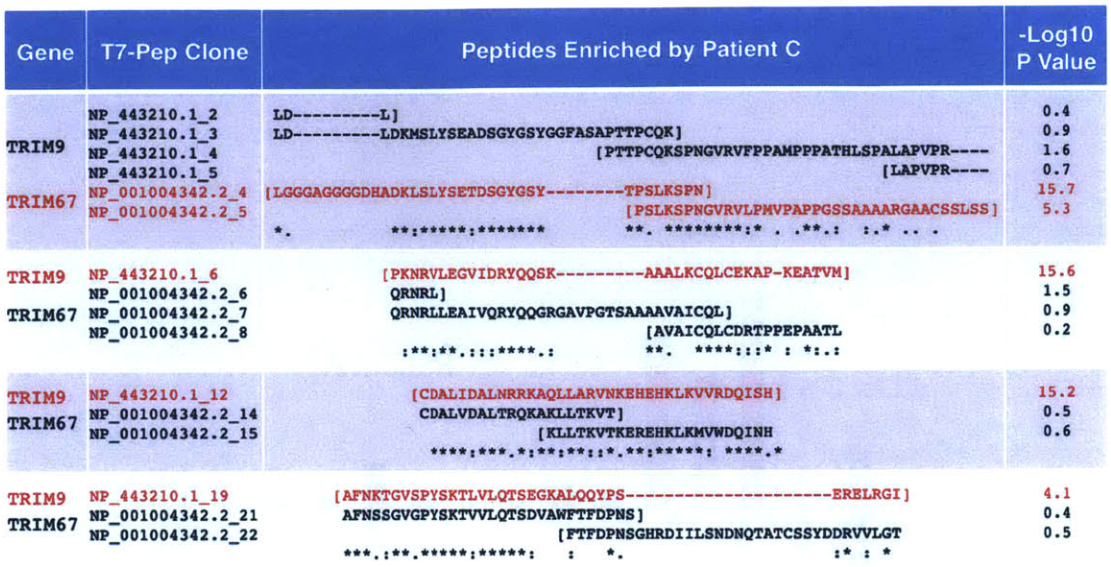


Figure 4.9: Alignment among enriched peptides from TRIM9 and TRIM67. Significantly enriched peptides (in red) from TRIM9 and TRIM67 shown with corresponding ClustalW-aligned peptides from the homologous protein (in black). Boundaries of phage-displayed peptides are denoted with brackets. Peptides are shown next to their −Log10 p-value of enrichment.

4.3.4 PhIP-Seq can identify peptide-protein interactions

The utility of T7-Pep is not limited to autoantigen identification. To explore more general interactions, we have used the library in an in vitro peptide-protein "two-hybrid" interaction experiment with GST-RPA2 (replication protein A2) as bait for T7-Pep. We were again able to utilize the generalized Poisson method for determining the significance of phage clones' enrichment. Whereas GST alone did not significantly enrich any library clones ($P < 10^{-4}$; Figure 4.11), PhIP-Seq with GST-RPA2 robustly identified the N-terminal peptide from the known interactor SMARCAL1 ($P < 10^{-14}$, Figure 4.12), among others (Supplementary Table 3.3). The enriched SMARCAL1 peptide contains a previously identified motif known to bind RPA2 [79, 80]. Peptides from four proteins known to contain this motif (UNG2, TIPIN, XPA and RAD52) were significantly disrupted by the positions of the breaks between peptides (Table 4.6). One peptide from UNG2 retained most of the motif and that peptide was correspondingly enriched ($P < 10^{-5}$), demonstrating the power of this approach to identify linear interaction motifs.

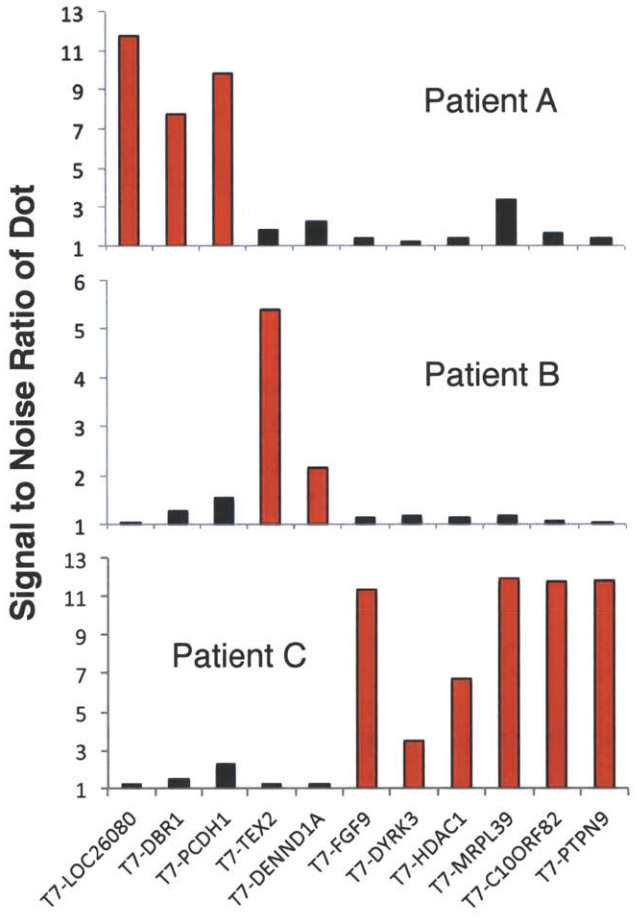


Figure 4.10: Quantification of T7 Candidate Dot Blots. The dot blots in Figure 4.6g were analyzed to determine the signal-to-noise ratio arising from each T7 candidate clone immunoblotted with each of the patients' spinal fluid. The data from the candidates expected to react with a given patient's antibodies are shown in red, whereas that data from the candidates that are expected not to react with a given patient's antibodies are shown in black.

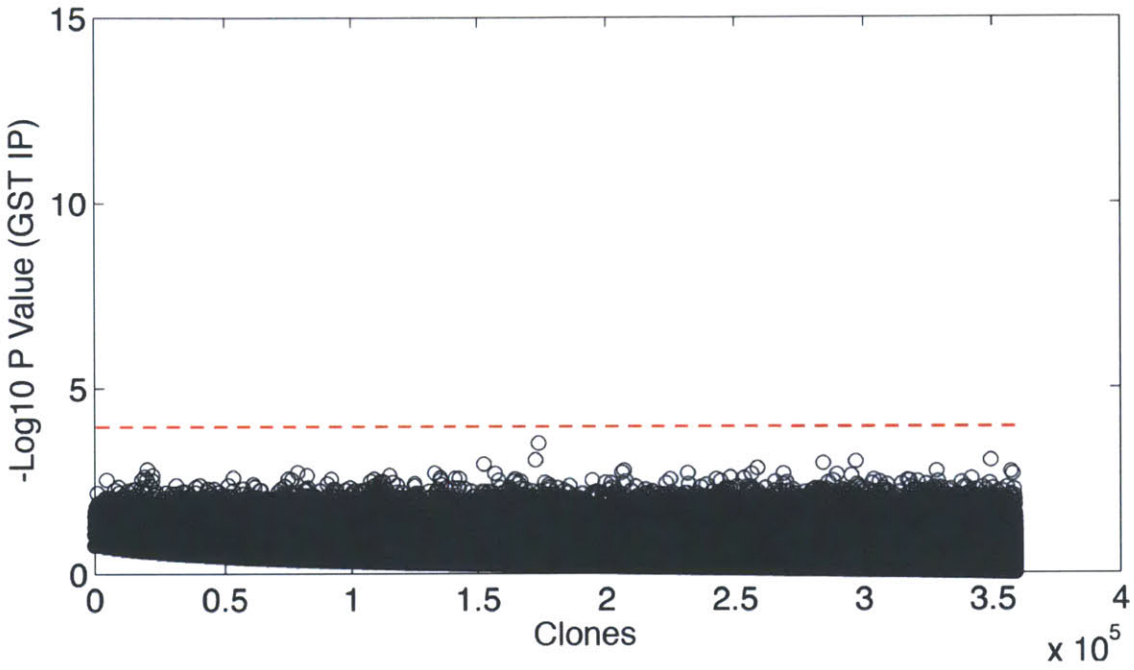


Figure 4.11: PhIP-Seq −Log10 p-values for T7-Pep enrichment by GST alone. GST coated glutathione magnetic beads were used to precipitate phage from the T7-Pep library. Illumina sequencing data was analyzed using the generalized Poisson method. No library members were significantly enriched by GST alone ($P < 10^{-4}$).

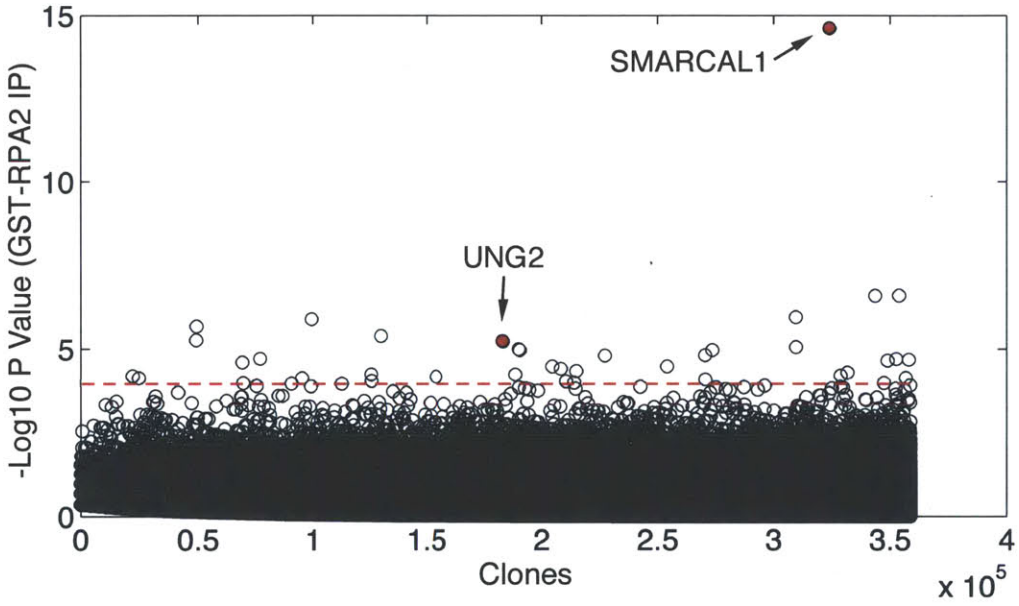


Figure 4.12: PhIP-Seq can identify protein-protein interactions. GST-RPA2 was used to precipitate phage from the T7-Pep library on magnetic glutathione beads. -Log10 p-values of enrichment were calculated using the generalized Poisson method. Clones are arranged in increasing input abundance from left to right. The experiment identified two of the known RPA2 binding partners SMARCAL1 ($P < 10^{-14}$) and UNG2 ($P < 10^{-5}$), highlighted in red

Rank	T7-Pep Clone	Peptide	Log P GST	Log P GST-RPA2	Gene Symbol
1	NP_054859.2_1	MSLPLTEEQRKKIEENRQKALARRAEKLLAEQHQRT	0.29	14.61	SMARCAL1
2	NP_055877.3_31	TPPSMSAALPFPAGGLGMPPSLPPPPLQPPSLPLSM	1.09	6.60	PPRC1
3	NP_006360.3_18	TLSYNGLGSNIFRLLDSLRALSGQAGCRLRALHLSD	2.23	6.59	LRRC41
4	NP_060903.2_28	AVLQQNPSVLEPAAVGGEAASKPAGSMKPACPASTS	0.07	5.95	KDM3A
5	XP_372311.2_13	LTLYDGPNVSSPSYGPYCRGDTSIAPFVASSNQVFI	1.65	5.90	LOC389958
6	NP_060876.2_13	LTPVTTSTVLSSPSGFNPSGTVSQETFPSGETTISS	1.87	5.68	MUC4
7	XP_372592.2_4	AALIHVPPLSRGLPASLLGRALRVIIQEMLEEVGKP	0.28	5.39	PGPEP1L
8	NP_057131.1_2	ITAEEMYDIFGKYGPIRQIRVGNTPETRGTAYVVYE	0.47	5.26	SF3B14
9	NP_003353.1_3	AEQLDRIQRNKAAALLRLAARNVPVGFGESWKKHLS	0.20	5.23	UNG2
10	NP_443728.2_10	IRPMDDDLLKLLLPLMLQYSDEFVQSAYLSRRLAYF	1.32	5.06	MED12L
11	NP_004981.2_2	ISTVGPEDCVVPFLTRPKVPVLQLDSGNYLFSTSAI	1.60	5.00	MARS
12	NP_078997.2_120	TTSTSQSAASSNNTYPHLSCFSMKSWPNILFQASAR	0.22	4.96	ZFHX4
13	NP_004697.2_27	ITETAGSLKVPAPASRPKPRPSPSSTREPLLSSSEN	2.14	4.96	ARHGEF1
14	NP_003425.2_23	SHLSRHRKTTSVHHRLPVQPDPEPCAGQPSDSLYSL	1.28	4.81	ZNF133
15	NP_996882.1_34	LDRFKNRLKDYPQYCQHLASISHFMQFPHHLQEYIE	0.54	4.80	CNOT1
16	NP_783324.1_3	PHPSALSSVPIQANALDVSELPTQPVYSSPRRLNCA	2.11	4.71	RAB3IP
17	NP_000341.1_5	PESQHLGRIWTELHILSQFMDTLRTHPERIAGRGIR	0.02	4.70	ABCA4
18	NP_689896.1_1	MNRKWEAKLKQIEERASHYERKPLSSVYRPRLSKPE	0.78	4.68	CCDC111
19	XP_095991.7_15	IKTRDICNQLQQPGFPVTVTVESPSSSEVEEVDDSS	1.04	4.65	CEP78
20	NP_741996.1_43	ITNGLAMKNNEISVIQNGGIPQLPVSLGGSALPPLG	1.60	4.60	SALL3
21	NP_997191.1_49	SVYGWATLVSERSKNGMQRILIPFIPAFYINQSELV	0.86	4.48	NUP210L
22	NP_775902.2_9	IHSGERPYECSECGKLFMWSSTLITHQRVHTGKRPY	1.31	4.41	ZNF547
23	XP_496363.1_6	PVRRGYWGNKIGKPHTVPCKVTGRCGSALVHLIPVP	1.66	4.34	RNF11
24	NP_001026.1_92	LSRKLFWGIFDALSQKKYEQELFKLALPCLSAVAGA	0.05	4.29	RYR2
25	NP_006359.3_11	THQWLDGSDCVLQAPGNTSCLLHYMPQAPSAEPPLE	0.78	4.24	CREB3
26	NP_002146.2_6	ITVPAYFNDSQRQATKDAGAIAGLNVLRIINEPTAA	1.37	4.23	HSPA6
27	NP_065987.1_4	ITPTRELAIQIDEVLSHFTKHFPEFSQILWIGGRNP	2.50	4.18	DDX55
28	NP_079279.2_13	SHHDTAVLITRYDICSSKEKCNMLGLSYLGTICDPL	0.24	4.17	ADAMTS20
29	NP_055987.1_47	PKGEPTRRGRGGTFRRGGRDPGGRPSRPSTLRRPAY	1.62	4.13	BAT2L2
30	NP_005712.1_2	IIPSCIAIKESAKVGDQAQRRVMKGVDDLDFFIGDE	1.82	4.13	ACTR3
31	NP_079165.3_2	SPPSQLFSSVTSWKKRFFILSKAGEKSFSLSYYKDH	1.75	4.13	C10ORF81
32	NP_006609.2_26	PPYKYKLRYRYTLDDLYPMMNALKLRAESYNEWALN	0.63	4.08	LOC100133760
33	NP_149163.2_41	INLTIRGHEVVGIVGRTGSGKSSLGMALFRLVEPMA	1.01	4.04	ABCC11
34	NP_001004750.1_11	IHFLFPPFMNPFIYSIKTKQIQSGILRLFSLPHSRA	0.79	4.03	OR51B6

Table 4.5: Candidate RPA2 interacting proteins. PhIP-Seq was performed using GST-RPA2 as bait, and enrichment scores (−Log10 p-values estimated by the generalized Poisson method) were compared to enrichment on GST alone.

Gene	T7-Pep Clone	Aligned Peptide	-Log10 P Value
SMARCAL1	**NP_054859.2_1**	**MSLPLTEEQRK-KIEENRQK--ALARRAEKLLAEQHQRT**	**14.6**
UNG2	NP_003353.1_2	...PSSPLSAEQLD-RI--	0.1
	NP_003353.1_3	**AEQLD-RI--QRNKAAAL----LRLAARNVPV...**	**5.2**
TIPIN	NP_060328.1_7	...LSRSLTEEQQR-RIE--RNKQLA	1.1
	NP_060328.1_8	E--RNKQLALERRQAKLLSNSQTL...	0.4
XPA	NP_000371.1_1	...QPAELPASVRA-SIERKRQRAL	0.3
	NP_000371.1_2	RKRQRALML--RQARLAARPYSA...	0.1
RAD52	NP_002870.2_9	...SLSSSAVESEATHQRKLRQKQLQQQF	1
	NP_002870.2_10	KQLQQQFR-ERMEKQQVRV...	0.1

Table 4.6: Dependence of peptide-RPA2 interaction on integrity of RPA2 binding motif. Aligned phage peptides containing the RPA2-binding motif (underlined) are shown next to their -Log10 p-value of enrichment. Significantly enriched peptides are shown in bold.

4.4 Discussion

We have developed a new proteomic technology called Phage Immunoprecipitation Sequencing (PhIP-Seq), which is based on a synthetic phage library (T7-Pep) made to uniformly express the complete human peptidome on the coat of T7 phage particles. Combining T7-Pep with high throughput DNA sequencing enables a variety of innovative proteomic investigations. In addition to applications in autoimmune disease, PhIP-Seq can be utilized to identify peptide-protein interactions and can be a viable alternative to two-hybrid analyses. From a methodological perspective, the robust single-round enrichment signals and the ability to adapt the assay to 96-well format suggests the feasibility of performing automated PhIP-Seq screens on large sets of samples.

Antibodies bind protein antigens by a variety of mechanisms and several studies have uncovered some general themes underlying these interactions. For instance, antibody combining surfaces on natively folded proteins tend to be dominated by "discontinuous" epitopes, which are patches of 4–14 amino acid side chains formed by two or more noncontiguous peptides brought into proximity during protein folding [81, 82]. If the protein is divided into its constituent peptides, antibody affinity is expected to decrease due to 1) the loss of contacts contributed by noncontiguous residues, and 2) the increased entropic costs of binding a free peptide as opposed to the natively constrained peptide. The degree to which individual peptides are still able to interact with a given antibody is difficult to predict, and is expected to vary widely. While our study demonstrates the utility of 36 amino acid tiles, further work will be required to define the true false negative discovery rate inherent to the use of T7-Pep. Autoantibodies that target

normally inaccessible epitopes have also been reported, such as those that recognize proteolytic cleavage products [83, 84], misfolded proteins or protein aggregates [85, 86]. Antigen discovery with full-length, folded proteins may thus be less sensitive than tiled peptides in some such circumstances.

In our study, performing PhIP-Seq with CSF from a well characterized PND patient (Patient A), identified a known (NOVA1) and a novel, testis-restricted [74] autoantigen (TGIF2LX). Since we also found anti-TGIF2LX antibodies in the spinal fluid of a second PND patient with NSCLC, this protein may represent a new cancer-testis antigen family member, and should be further investigated as a biomarker for PND. PhIP-Seq analysis of CSF from two PND patients with uncharacterized antibodies (Patients B and C) uncovered likely neuronal targets of their autoimmune syndromes. In Patient B, high titer anti-GAD65 antibodies bound two distinct peptides from the region of the protein associated with Stiff Person Syndrome (SPS). Interestingly, GAD65 targeting in SPS occurs more often in patients without cancer, raising the possibility that at least part of this neurological syndrome may have been unrelated to the patient's cancer. This finding highlights the utility of unbiased antibody profiling to distinguish between deceptively similar disease states [87]. In Patient C, we identified TRIM9 as a likely neuronal autoantigen and suggest the possibility of epitope spreading from tumor-derived TRIM67 as a potential mechanism. It should be noted that demonstration of a protein's autoreactivity is not evidence for its role in disease pathogenesis, since the autoantibodies might be incidental in nature, arise due to epitope spreading, or might simply exhibit non-cognate cross-reactivity.

Several interesting features of the T7-Pep + PhIP-Seq platform emerged during this proof-of-concept study. We found that patient antibodies targeting GAD65 robustly recognized two 36 amino acid peptides, but not the corresponding denatured full-length proteins, indicating that an important degree of conformational information is retained in the peptide library. Second, for proteins with known crystal structures, using tiled peptides can facilitate determination of the antibody clonality, as well as the location of the targeted epitope. Finally, the simultaneous quantification of a large number of peptide enrichments permits the discovery of epitope motifs. Autoantibodies from Patient A targeted seven peptides from a repetitive hypothetical protein, and we were thus able to calculate a motif that most likely represented the antigenic epitope, a task less easily performed with alternative technologies.

T7-Pep could be improved in several ways. The generation of longer oligos will decrease the complexity of the library, thereby increasing the sampling depth and making it possible to generate domain libraries that capture more protein-folding units. In addition, PhIP-Seq with libraries of peptides from human pathogens could permit rapid analysis of antibodies to infectious agents, thus aiding vaccine research and the diagnosis of infectious diseases.

We have taken a synthetic biological approach to develop a proteomic resource useful in

translational medicine. When combined with high throughput DNA sequencing, our methodology permits unbiased and quantitative analysis of autoantibody repertoires in human patients. PhIP-Seq thus complements existing proteomic technologies in the study of autoimmune processes for which the relevant autoantigens remain unknown.

4.5 Methods

4.5.1 Design of T7-Pep, T7-CPep and T7-NPep ORF sequences

We first downloaded all human protein and cDNA sequences available from the RefSeq database at build 35.1 of the human genome. Accession numbers between a protein and its cDNA were matched, and the paired sequences were used to construct the library. All the ATG start codons in the cDNAs were compared to the corresponding protein sequences until the correct ORF sequence was found. Seventy-two nucleotide (nt) fragments were then separated and overlapped with adjoining sequences by 21 nt (7 amino acids). Each DNA fragment was then scanned for the eight relatively rare codons in E. coli (CTA, ATA, CCC, CGA, CGG, AGA, AGG, GGA), and they were replaced by more abundant, synonymous codons (selected randomly if there was more than one replacement available). After that, each DNA fragment was rescanned for the four restriction sites (EcoRI, XhoI, BseRI, MmeI), and they were eliminated by replacement of one codon with a different, abundant, synonymous codon. Sequences were scanned iteratively to ensure the final ORF fragments were free of both rare codons and restriction sites. Finally, common primer sequences were added.

4.5.2 Cloning of T7-Pep

The proteome-wide library (19 pools of 22,000 synthetic oligos per pool) and N/C-terminal libraries (two pools each of 18,000 synthetic oligos per pool) were PCR-amplified as 23 independent pools with common primer sequences using the following conditions: 250 mM dNTPs, 2.5 mM MgCl2, 0.5 µM each primer, 1 µl Taq polymerase and 350 ng oligo DNA per 50 µl reaction. The thermal profile was 1. 95 °C 30 s, 2. 94 °C 35 s, 3. 50 °C 35 s, 4. 72 °C 30 s, 5. Go to step 2 3x, 6. 72 °C 5 min, 7. 95 °C 30 s, 8. 94 °C 35 s, 9. 70 °C 35 s, 10. 72 °C 30 s, 11. Go to step 8 29x, 12. 72 °C 5 min The PCR product was then digested and cloned into the EcoRI/SalI sites of the T7FNS2 vector with an average representation of at least 100 copies of each peptide maintained during each cloning step. The T7FNS2 vector is a derivative of the T7Select 10-3b vector (Novagen), which is a lytic, mid-copy phage display system, and displays 5-15 copies as C-terminal fusions with the T7 capsid protein. We modified the T7Select 10-3b vector to generate T7FNS2 by inserting a sequence encoding a FLAG epitope in the NotI and

XhoI sites to generate an in-frame FLAG C-terminal fusion with the inserted peptide. Cloning of the synthetic peptide libraries into the T7FNS2 vector results in a C-terminal fusion of the ORF fragments with the T7 10B capsid protein, followed by a C-terminal FLAG epitope tag and stop codon (except for those in T7-CPep, which retain the native stop codons).

4.5.3 Patient samples

Collection and usage of human specimens from consenting patients were approved by the Brigham and Women's Hospital Institutional Review Board (protocol no. 2003-P-000655). Cerebrospinal fluid was aliquoted and kept at −80 °C until used, and freeze-thawing was avoided as much as possible after that. Neurological evaluations were performed by a board-certified neurologist. Serum samples from patients with confirmed NSCLC were from Bioserve.

4.5.4 Detailed PhIP-Seq protocol

The following were the multiplex barcode-introducing forward primers. The common P5 sequence for Illumina sequencing is in bold. The underlined segment was where the sequencing primer annealed. The 3-nt barcode is in italics.

HsORF-FL-mmBC1-F

AATGATACGGCGACCACCGAAGGTGTGATGCTCGGGGATCCAGGAATTCCACTGCGC

HsORF-FL-mmBC2-F

AATGATACGGCGACCACCGAAGGTGTGATGCTCGGGGATCCAGGAATTCCGCCGCGC

HsORF-FL-mmBC3-F

AATGATACGGCGACCACCGAAGGTGTGATGCTCGGGGATCCAGGAATTCCCCTGCGC

HsORF-FL-mmBC4-F

AATGATACGGCGACCACCGAAGGTGTGATGCTCGGGGATCCAGGAATTCCTCTGCGC

HsORF-FL-mmBC5-F

AATGATACGGCGACCACCGAAGGTGTGATGCTCGGGGATCCAGGAATTCCGATGCGC

HsORF-FL-mmBC6-F

AATGATACGGCGACCACCGAAGGTGTGATGCTCGGGGATCCAGGAATTCCGGTGCGC

HsORF-FL-mmBC7-F

AATGATACGGCGACCACCGAAGGTGTGATGCTCGGGGATCCAGGAATTCCGTTGCGC

HsORF-FL-mmBC8-F

AATGATACGGCGACCACCGAAGGTGTGATGCTCGGGGATCCAGGAATTCCGCGGCGC

P7-T7Down (this is the common reverse primer):

CAAGCAGAAGACGGCATACGAC ACTG AACCCCTCAAGACCCGTTTA

mmBC-FL_seq_prim (for sequencing the barcode and the library insert at P5 in forward direc-

tion):

AGGTGTGATGCTCGGGGATCCAGGAATTCC

Immunoprecipitation wash buffer consisted of 150 mM NaCl, 50 mM Tris-HCl, 0.1% NP-40 (pH 7.5).

Procedure: 1.5 ml tubes were blocked (including under cap) with 3% fraction V bovine serum albumin (BSA) in tris-buffered saline with 0.5% tween-20 (TBST) overnight at 4 °C rotating. Positive control SAPK4 C-19 antibody (Santa Cruz, sc-7585) was added (2 ng/ml final concentration; 1/1,000 of patient antibody) to phage stock (5 x 1010 pfu T7-Pep/ml final concentration) and mixed before being added to patient antibody (2 µg/ml final concentration). Each IP reaction was brought to a final volume of 1 ml using M9LB (Novagen).

Note: replicas were independent after this point (that is, there were two IP reactions as above for each sample).

Tubes were rotated at 4 °C for 24 h. 40 µl of 1:1 mix of Protein A and Protein G coated magnetic Dynabeads (Invitrogen, 100.02D and 100.04.D) slurry was added to each tube. Tubes were rotated for 4 more hours at 4 °C. Beads were washed 6 times in 500 µl IP wash buffer by pipetting up and down eight times per wash. Tubes were changed after every second wash. As much wash buffer as possible was removed and beads were resuspended in 30 µl H2O. IP was then heated at 90 °C for 10 min to denature phage and release DNA. 50 µl PCR reactions were prepared with TaKaRa HS Ex polymerase (TAKARA BIO), using the entire 30 µl of IP: 9.5 µl H2O, 5 µl 10x TaKaRa buffer, 4 µl dNTP (2.5 mM each), 0.5 µl P7-BC-T7Down (200 µM), 0.5 µl P5-mmBCn-F (100 µM), 0.5 µl TaKaRa HS Ex enzyme mix, 30 µl phage IP. The thermal profile was 1. 98 °C 10 s, 2. 56 °C 15 s, 3. 72 °C 25 s, 4. Go to step 1 39x, 5. 72 °C 7 min The number of cycles can optionally be increased to 45. PCR products were gel purified individually. Concentration was measured and then 500 ng of each barcoded sample was mixed together and Illumina sequencing was then performed on final material, using mmBC-FL_seq_prim as sequencing primer.

The first seven nt calls arose from the DNA barcode, and were used to parse the data by sample. Remaining sequence was aligned against the reference file. The reference sequences were truncated to the length of the reads and alignment was constrained to the appropriate strand.

4.5.5 RPA2-peptide interaction screen

Full-length, sequence-verified RPA2 was recombined from an available entry vector into pDEST-15 for inducible expression in E. coli as an N-terminal GST-fusion protein. A pDEST-15 clone expressing GST alone was used as a negative control. Protein expression was induced with 0.1 µM IPTG for 5 hours at 30°C. Protein lysate from 50 ml of bacterial culture was prepared in

1.5 ml of lysis buffer (50 mM tris pH 7.5, 500 mM NaCl, 10% glycerol, 1% triton, 10 mg/ml lysozyme) and sonicated before removing insoluble material by centrifugation. 40 µl of MagneGST Glutathione beads (Promega, V8611) were incubated in 1 ml of undiluted bacterial lysate for 2 hours. Beads were then washed 3 times with PBS. 1 ml of M9LB containing 5x1010 pfu of T7-Pep was then used to resuspend the beads (now coated with GST or GST-RPA2). The mix was rotated 24 hours at 4°C. At this point the beads were washed 6 times in 500 µl IP wash buffer, and the remaining protocol for PhIP-Seq given above was followed precisely.

4.5.6 Estimation of general Poisson model parameters and regressions

We assessed several distribution families for their ability to appropriately model the PhIP-Seq enrichment data, and found the two-parameter generalized Poisson distribution to be the best:

$$\text{pmf}(x) = \theta(\theta + x\lambda)^{x-1}e^{-\theta-x\lambda}/x!$$

For each value of input read number that had at least 50 corresponding clones, we used the following maximum likelihood estimators to calculate the values of lambda (λ) and theta (θ) for the corresponding distribution of n IP reads (xi) [67].

$$\sum_{i=1}^{n} \frac{x_i(1-x_i)}{X+(x_i-X)\lambda} - nX = 0$$

where

$$X = \sum_{i=1}^{n} \frac{x_i}{n}$$

and

$$\theta = X(1-\lambda)$$

Upon calculation of λ across all the input read numbers, we found it to be approximately constant. For each experiment, we thus regressed this parameter to be equal to the mean of all calculated λ's (Figure 4.4c). Calculation of θ's for all input values revealed the near linearity of this parameter, and so we linearly regressed this parameter prior to calculating the p-values.

4.5.7 Western blot validation of candidate autoantigens

We utilized the ORFeome collection of full-length proteins, which was generated by PCR and Gateway recombinational cloning [88], as a source for testing autoantigen candidates by im-

munoblot. Entry vectors were recombined into the appropriate mammalian expression vector (CMV promoter driving ORF expression with either C-terminal GFP fusion or N-terminal FLAG epitope tag) and miniprepped for transient transfection.

293T cells were plated 24 hours before transfection at a density of 0.8 million cells per well of a 6-well plate and grown in DMEM containing 10% FBS. TransIT-293T transfection reagent (Mirus, MIR 2700) was mixed with 2 µg expression plasmid per well, and added to the cells. After 24 hours, cells were harvested in 200 µl standard 1x RIPA-based laemmli/DTT sample buffer with Complete protease inhibitor cocktail (Roche) and sonicated for 30 seconds. Insoluble material was removed by centrifugation. 2-20 µl of lysate was run on 4-20 Bis-Tris polyacrylamide gels and transferred onto nitrocellulose using the iBlot system (Invitrogen). Membranes were blocked 1 hr in 5% milk and then stained with either patient CSF (1:250 to 1:1,000) or the appropriate primary anti-GFP (JL-8 monoclonal antibody; Clontech, 632381) or anti-FLAG (M2 monoclonal antibody; Sigma-Aldrich, F9291) antibody in 2.5% milk, TBST. Human antibody from CSF was detected with 1:3,000 peroxidase-conjugated goat affinity purified anti-Human IGG (whole molecule) secondary antibody (MP Biomedicals, 55252) in 2.5% milk, TBST.

For IP-western blotting, cell lysate was harvested in standard RIPA buffer with Complete protease inhibitor cocktail and sonicated for 30 seconds. Insoluble material was removed by centrifugation. 150 µl of lysate was mixed with 1 µg of patient antibodies and rotated overnight at 4°C. A 40 µl slurry of 1:1 mix of Protein A coated magnetic Dynabeads and Protein G coated magnetic Dynabeads was added to each tube. Tubes were rotated 4 hours at 4°C. Beads were washed 3 times in 500 µl RIPA buffer, and then harvested in 25 µl of laemmli/DTT sample buffer. The IP'ed protein and 10% of the input lysate were subject to SDS-PAGE analysis as above, and protein was detected by staining for the protein tag (e.g. GFP).

4.5.8 Dot blot validation of candidate autoantigens

Individual clones were made by synthesizing the peptide-encoding insert as a single, long DNA oligo (IDT, Ultramer™) that was PCR amplified and then cloned into T7FNS2 in the same way as described for the library. Clones were sequence verified and titered. 2 µl of each clone, after normalizing for titer, was spotted directly onto a nitrocellulose membrane and allowed to dry for 30 minutes. Membranes were blocked with 5% milk, TBST for 1 hour at room temperature, and then stained overnight at 4°C with 1 µg/ml of CSF antibody diluted in a solution containing a 1:1 mix of 5% milk, TBST and T7 10-3b-FLAG phage lysate. Human antibody from CSF was then detected with 1:3,000 peroxidase-conjugated goat affinity purified anti-Human IGG (whole molecule) secondary antibody (MP Biomedicals, 55252) in 2.5% milk, TBST. Quantification was performed by scanning developed films and analyzing the .tiff file

with ImageJ software.

4.6 Author contributions

S.J.E. conceived the project, which was supervised by N.L.S. and S.J.E. Z.Z. designed the DNA sequences for synthesis. Oligo libraries were constructed by E.M.L. Cloning was performed by M.Z.L., M.A.M.G., and N.L.S. The T7-Pep, T7-NPep, and T7-CPep phage libraries were constructed by N.L.S. and characterized by N.L.S. and H.B.L. The PhIP-Seq protocol was developed and implemented by H.B.L. Clinical evaluations and patient sample acquisitions were performed by S.K. Statistical analysis of PhIP-Seq data was conceived by U.L. under the supervision of G.M.C. and implemented by H.B.L. PhIP-Seq candidates were confirmed by H.B.L. The RPA2 experiment was performed by A.C. The manuscript was prepared by H.B.L. and edited by N.L.S. and S.J.E.

Chapter 5

High Throughput PhIP-Seq Definition of Autoantibody Repertoires in Health and Disease[1]

5.1 Abstract

Autoimmune disease results from a loss of tolerance to self-antigens in genetically susceptible individuals. Understanding this process requires knowledge of the target molecules, and thus a number of techniques have been developed to determine immune receptor specificities. We have previously reported the construction of a T7 phage-displayed synthetic human peptidome and its application to autoantigen discovery using cerebrospinal fluid from 3 patients with paraneoplastic neurological disorder. Here we present data from the first large-scale phage immunoprecipitation sequencing ("PhIP-Seq") screen of 298 independent antibody repertoires, including those from 73 healthy sera. The resulting database of peptide enrichments characterizes each individual's unique "autoantibodyome", and includes specificities found to occur frequently in the general population or associated with disease. Sera from 39 type 1 diabetes (T1D) patients were screened, revealing a prematurely polyautoreactive phenotype compared to their matched controls. Screening a collection of cerebrospinal fluids and sera from 63 multiple sclerosis patients uncovered novel, as well as previously reported specificities. Finally, a screen of synovial fluids and sera from 64 rheumatoid arthritis patients revealed novel recurrent autoantibody specificities that were independent of seropositivity status. In sum, this work demonstrates the utility of performing PhIP-Seq screens on large numbers of individuals and is a step toward defining the full complement of autoimmunoreactivities in health and disease.

[1] Submitted for publication as: Larman, et al., High Throughput PhIP-Seq Definition of Autoantibody Repertoires in Health and Disease, 2012.

5.2 Introduction

Our understanding of autoimmunity has been limited by available technologies, which cannot capture the molecular complexity of intact immune systems in large numbers of individuals. To address these limitations, we have recently developed an unbiased proteomic technology, phage immunoprecipitation sequencing (PhIP-Seq), with the capacity to quantitatively measure interactions between an individual's antibody repertoire and each of over 400000 overlapping 36 mer peptides that together span the open reading frames of the human genome [26]. In this work, we have improved the PhIP-Seq method in two ways. First, sample processing was made 96-well plate compatible and implemented on a Biomek FX liquid handling robot, resulting in a more reproducible, high throughput protocol. Second, we developed a method to perform 96-plex analysis of individual PhIP-Seq libraries using just 2-3 lanes of an Illumina HiSeq 2000 flow cell [89]. This degree of multiplexing reduces the cost of each screen to about $25 per sample, thereby enabling cohort-scale repertoire screening projects.

There are several autoimmune diseases of relatively high incidence for which the role of antibody-mediated autoimmunity is appreciated but not understood. Of these, we selected type 1 diabetes (T1D), multiple sclerosis (MS) and rheumatoid arthritis (RA) for autoantibody repertoire analysis by high-throughput PhIP-Seq screening. Strong genetic linkage to class II HLA alleles in each of these diseases supports the view that there is an important role for antigen presentation and subsequent activation of CD4+ helper T cells with self-specificity [90]. The role of B cells in these diseases is less clear, but several lines of evidence indicate that analysis of secreted antibodies may provide insight into disease pathogenesis. For example, pancreatic beta cell destruction in T1D is thought to be largely a consequence of CD8+ T cell activity, yet autoantibodies targeting islet-associated antigens are routinely used for diagnosis and risk stratification [91]. In MS, secondary lymphoid tissue with germinal center activity often forms in the meninges of patients with advanced disease [92] and oligoclonal IgG bands of unknown specificity are frequently found in cerebrospinal fluid (CSF) [93]. Patients with RA are classified as seropositive or seronegative depending on the presence of rheumatoid factor (antibodies against the Fc portion of IgG) and/or antibodies that recognize the citrulline post translational protein modification. Beneficial clinical response to CD20+ B cell depletion therapy in RA has prompted the adoption of rituximab as a second line therapy for patients with high disease activity and features of a poor prognosis [94, 95]. In the treatment of MS and T1D, several studies have demonstrated a benefit after B cell depletion, but with perhaps more elusive optimal dosing regimens [96, 97].

Here we report the first high throughput PhIP-Seq analysis of autoantibody repertoires from a large number of T1D, RA, and MS patients, for comparison to each other and to a set of 73

healthy controls. Our findings describe both known and novel properties of immunoenriched peptides, and sets the stage for additional large scale PhIP-Seq investigations.

5.3 Results

5.3.1 Polyautoreactivity and screen sensitivity

We used PhIP-Seq to analyze 298 antibody repertoires. This collection of samples included 39 sera obtained from newly diagnosed T1D patients, 44 synovial fluids and 20 sera from RA patients, and 28 CSF samples and 35 sera from MS patients (including 6 matching CSF/serum sets). Additionally, 73 sera from healthy donors, including a set of 41 age/sex-matched controls for the T1D cohort, were analyzed. To control for differences in fluid composition, we screened synovial fluid samples from 19 individuals with gout or osteoarthritis, as well as CSF from 10 patients with non-MS associated meningitis, subacute sclerosing panencephalitis, or paraneoplastic neurological disorder. Finally, we had previously screened a collection of 29 sera from patients with estrogen and progesterone receptor positive breast cancer (BC), and while analysis of the BC dataset is not presented here, it was utilized to increase power of the antigen-disease specificity tests. Table 5.1 provides a summary of these samples. A more detailed description can be found in Table 5.2.

After immunoprecipitation of the T7-Pep phage library with patient antibodies, peptide enrichments were quantified using massively parallel DNA sequencing. We considered peptides with a P-value $< 10^{-4}$ (-log10 P-value greater than 4) as scoring positively above background (see Methods, Figure 5.1) [26]. To exclude peptides that immunoprecipitated nonspecifically, we ignored 1404 peptides that displayed enrichment with -log10 P-values equal to 3 or greater in two or more out of 8 negative control (no patient sample) IPs.

We first turned our attention to the data from the 73 healthy donors. In sum, 14604 unique peptides were enriched by at least one healthy donor. An overwhelming majority (12727) of these autoreactivities were "personal" in the sense that they were observed to occur in only one individual (Figure 5.2A). At the other extreme, we observed a smaller number of peptides that were more frequently enriched by healthy individuals. For example, we found that serum from 40% of individuals significantly enriched a single peptide from the activin receptor type IIB (ACVR2B), and serum from 44% of individuals had reactivity against a peptide from melanoma antigen family E, 1 (MAGEE1). Notably, these two autoreactivities were not significantly correlated, suggesting that they arise independently of each other. As it is not immediately clear whether these common autoantibodies are antigen driven or simply cross-reactive, we looked for evidence of multi-epitope targeting within the database. Whereas we did find convincing

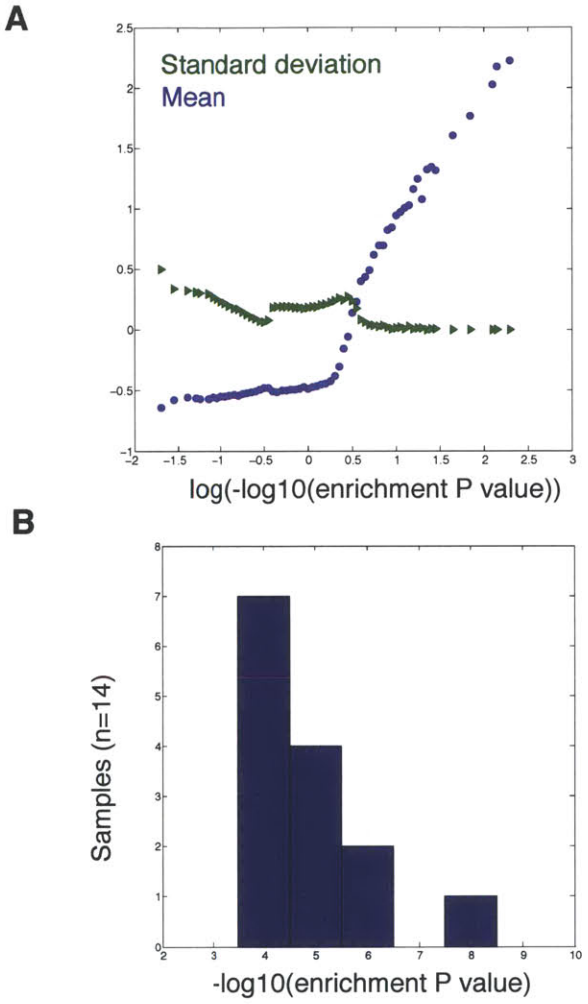


Figure 5.1: Dataset reproducibility threshold. P value threshold for reproducibility was established using the data from each sample duplicate pair. Scatter plots of duplicate 1 versus duplicate 2 were used to perform a signal-to-noise analyses. A. Typical behavior of a duplicate scatterplot. As -log10 P values increase, the average mean (signal) increases while the standard deviation (noise) decreases. The point at which they cross is considered the reproducibility threshold. B. Histogram plot showing where the mean passed above the variance in all screen duplicates. Based on this analysis, we chose -log10 P value = 4 as a low stringency cutoff for reproducibility.

Class	Subclass	Fluid	Total
Type 1 Diabetes		serum	39
Multiple Sclerosis	(RRMS/SPMS/PPMS)	serum	35
		CSF	28
Rheumatoid Arthritis	Seropositive	serum	10
		synovial fluid	22
	Seronegative	serum	10
		synovial fluid	22
Healthy Controls		serum	73
Non MS CSF Controls	SSPE, PND, Meningitis	CSF	10
Non RA synovial fluid controls	Gout, OA	synovial fluid	20
Breast Cancer	ER+/PR+	serum	29
		Total	**298**

Table 5.1: Summary of the samples screened by high throughput PhIP-Seq. RR, relapse remitting MS; SP, secondary progressive MS; PP, primary progressive MS; SSPE, subacute sclerosing panencephalitis; PND, paraneoplastic neurological disorder; OA, osteoarthritis; ER+, estrogen receptor positive; PR+, progesterone receptor positive. Six sets of MS CSF/serum samples are patient matched.

examples of antigen-driven responses (e.g. the scleroderma antigen CENPC1, Figure 5.2B), this was not true for ACVR2B or MAGEE1. We therefore conclude that these recurrent anti-peptide antibodies are most likely cross-reactive and because they occur frequently in the serum of healthy individuals are unlikely to have a pathological consequence.

Patterns of disease-associated autoreactivity may only become apparent in the context of aggregated peptide enrichments, since different individuals may produce antibodies that recognize distinct epitopes of the same protein. We therefore collapsed the peptide enrichment matrix onto an ORF enrichment matrix by taking the most significant value from the set of peptides corresponding to each ORF. Again, if this -log10 P-value was greater than 4, the ORF was considered enriched by the individual. Analysis of ORF enrichments by healthy individuals resulted in a distribution similar to the peptide enrichments, with the majority of significantly enriched ORFs (58%) arising in just one person (Figure 5.2C). This analysis is biased toward larger proteins being commonly enriched, and indeed significant reactivity against at least one peptide from titin (TTN, the largest ORF in our library) was observed in 45 of the 73 healthy individuals (Supplementary Discussion).

We screened a collection of serum samples obtained from 39 newly diagnosed T1D patients. As controls for comparison, we screened sera from 41 healthy donors (matched for age and gender) in the same automated PhIP-Seq run. Titers of clinically utilized autoantibody biomarkers

Experiment 1						
Class	Subclass	Male	Female	Age	Fluid	Total
Breast Cancer	ER+/PR+	0	29	52.3 (7.0)	serum	29
Multiple Sclerosis	(RRMS/SPMS)	0	29	52.8 (6.7)	serum	29
Healthy Controls		0	30	48.1 (9.5)	serum	30

Experiment 2						
Class	Subclass	Male	Female	Age	Fluid	Total
Multiple Sclerosis						
	RRMS (in remission)	0	6	39.9 (7.6)	serum	6
	RRMS (in remission)	5	12	45.2 (8.9)	CSF	17
	SPMS	10	1	44.4 (7.6)	CSF	11
Controls						
	Meningitis	4	0	?	CSF	4
	PND	0	2	61 (2)	CSF	2
	SSPE	3	1	17.5 (2.6)	CSF	4

Experiment 3						
Class	Subclass	Male	Female	Age	Fluid	Total
Type 1 Diabetes		21	18	17.4 (9.4)	serum	39
Healthy Controls		23	20	20.1 (10.4)	serum	43

Experiment 4						
Class	Subclass	Male	Female	Age	Fluid	Total
Rheumatoid Arthritis						
	Seropositive	?	?	?	serum	10
	Seronegative	?	?	?	serum	10
	Seropositive	4	17	60.7 (18.3)	synovial	22
	Seronegative	9	13	54.3 (16.6)	synovial	22
Controls						
	Gout	8	2	55.2 (14.3)	synovial	10
	Osteoarthritis	2	8	65.4 (11.1)	synovial	10

Table 5.2: Detailed composition of patient cohorts. Each experiment represents a different 96 well plate of samples, whose positions were randomized across the plate. ER+, estrogen receptor positive; PR+, progesterone receptor positive; RR, relapse remitting MS; SP, secondary progressive MS; PP, primary progressive MS; ?, unknown status; PND, paraneoplastic neurological disorder; SSPE, subacute sclerosing panencephalitis. Average patient age is given alongside standard deviation of ages in parentheses.

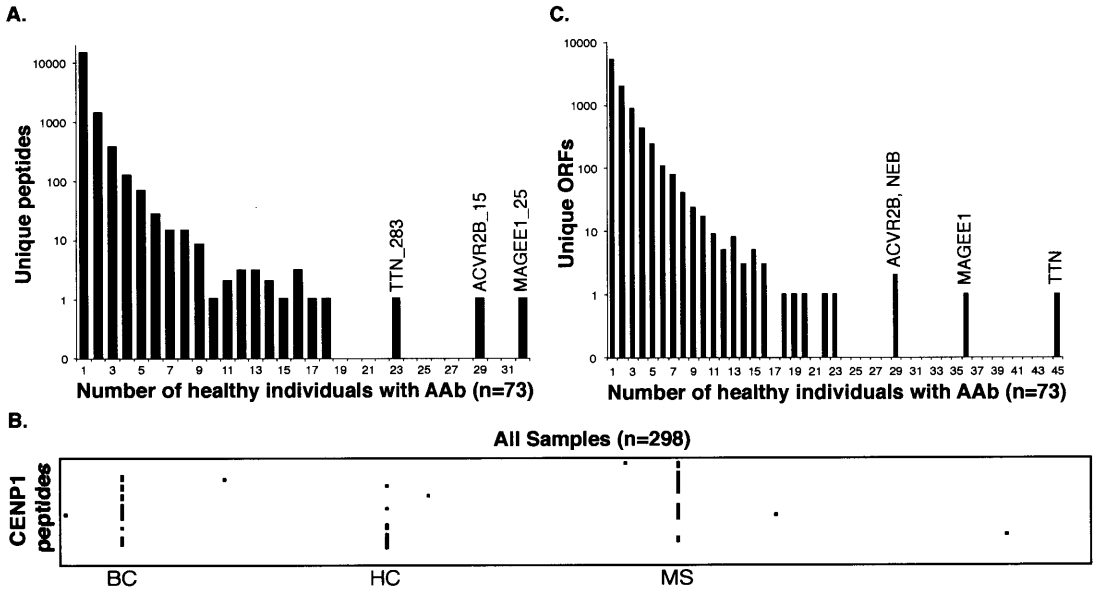


Figure 5.2: Enrichment recurrence and multi-epitope targeting. A. Frequency distribution of the 14604 unique peptides enriched by greater than -log10 P value of 4 in healthy individuals. Peptides enriched above a threshold of -log10 P value of 3 or greater in 2 or more negative controls were considered nonspecific and removed from the analysis. AAb, autoantibody; TTN, titin; ACVR2B, activin receptor type-2B; MAGEE1, melanoma antigen family E, 1. Number following underscore denotes which 36-residue tile was enriched (order is N- to C-terminus). B. Multi-epitope targeting of the CENPC1 protein. Peptides are organized from top to bottom. Peptide enrichment -log10 P values greater than 4 are colored black and less than 4 are colored white. Three individuals exhibit evidence of multi-epitope responses (BC, breast cancer; HC, healthy control; MS, multiple sclerosis). C. Frequency distribution of the 7619 unique ORFs enriched by greater than -log10 P value of 5 in healthy individuals. NEB, nebulin.

(islet cell cytoplasmic antibody, "ICA"; insulin autoantibody, "IAA"; glutamic acid decarboxylase 2 antibodies, "GADA"; protein tyrosine phosphatase, receptor type, N antibodies, "PTPRNA" or "IA2A"; zinc transporter, member 8 antibodies, "ZnT8A") were also measured for each of the T1D patients and controls. In order to determine the false negative rate (sensitivity), we compared radioimmunoassay (RIA) measurements for each biomarker in each individual with the corresponding PhIP-Seq ORF enrichment scores. No PhIP-Seq enrichment was observed in any of the patients for insulin or ZnT8A, whereas GAD2 and PTPRN enrichment was observed in some of the T1D patients who had the highest RIA titers for those antigens (Figures 5.3A and 5.4).

We reasoned that if the amount of antibody-self peptide cross-reactivity in any way reflected the complexity of the antibody repertoire, then serum of older individuals should bind more unique peptides compared to their younger counterpart. Comparing ages 12 and under ("young") with those 18 and older ("adult"), we observed a significant difference in the number of enriched peptides between young and adult healthy controls ($P = 0.03$; Student's t test, 1 tail; Figure 5.3B). However, when we performed the same analysis of the T1D cohort, we found young T1D patients to be significantly precocious in their development of autoreactive antibodies compared to their age-matched healthy counterpart ($P = 0.01$). There was no difference in the number of enriched peptides between healthy and T1D adults, or between young and adult T1D patients.

5.3.2 Disease-specific autoantibodies

We next identified peptide and ORF autoreactivities specifically associated with each autoimmune disease under investigation. For this analysis, each disease group was compared to all other samples, in the form of a Fisher's exact test to determine significance of association. This analysis was performed for each peptide in the library, and so a distribution of >400,000 Fisher's P values was obtained. To account for multiple hypothesis testing, we created a null distribution of "expected" Fisher's P values by randomly permuting the sample labels 1000 times (Methods). We compared the distribution of expected significance values to that which was actually observed, and then set a threshold for 10% false discovery rate (FDR). All peptide/ORF autoreactivities that exhibited disease association with this level of confidence are reported in Table 5.3 (see Table 5.4 for peptide sequences).

We first examined peptide/ORF autoreactivities specifically associated with RA (Table 5.3, Figure 5.5). Of the 16 peptides with an FDR <10%, 11 assorted with patients nonrandomly as two peptide clusters, "RA1" and "RA2", composed of 3 and 8 peptides, respectively. Interestingly, none of the RA-associated enrichments appeared to correlate with seropositivity (i.e. reactivity against rheumatoid factor and/or citrullinated peptide; Figure 5.5B). Despite attempts

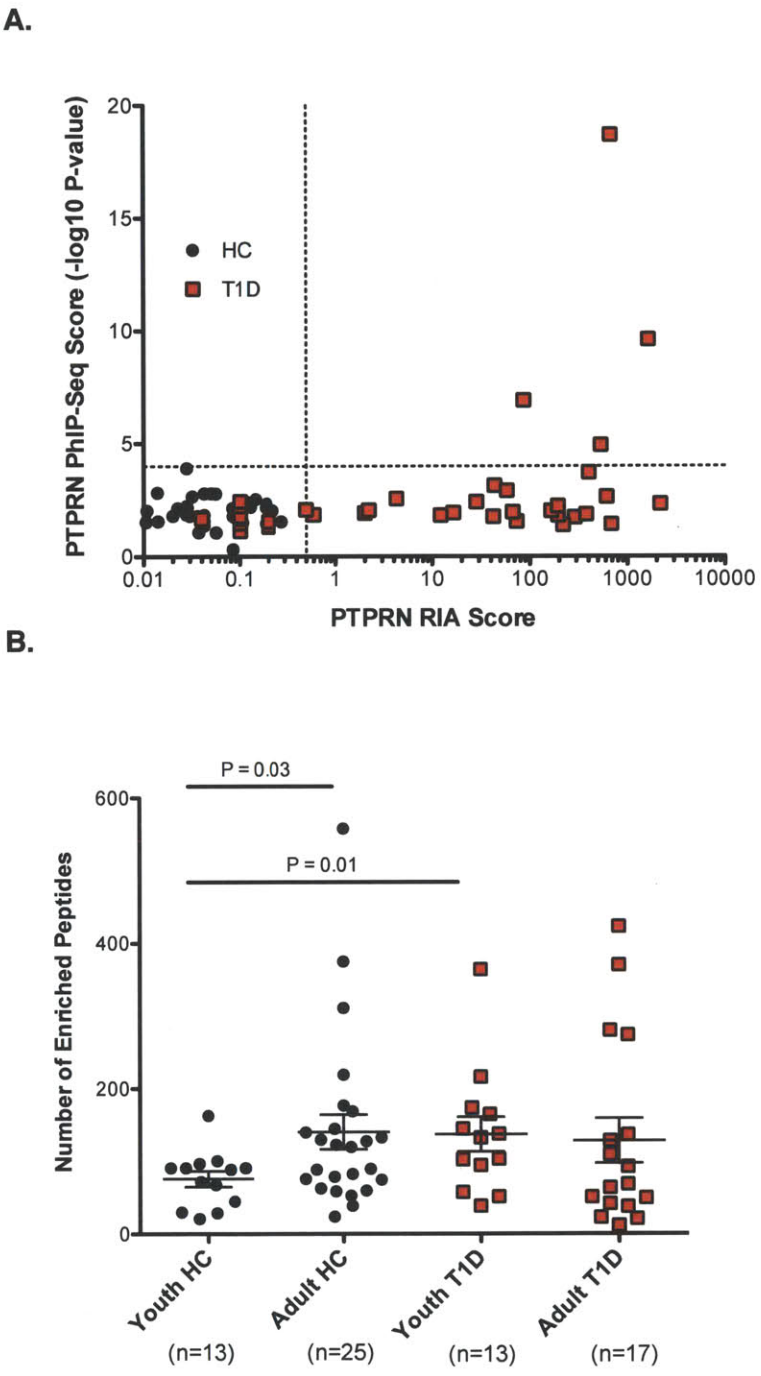


Figure 5.3: Analysis of T1D and healthy control sera. A. Sensitivity of PTPRN (IA2) autoantibody detection by PhIP-Seq compared to RIA in T1D patients (red boxes) and healthy controls (circles). PhIP-Seq values correspond to the most enriched peptide from the PTPRN ORF. A value of >0.5 is considered positive for RIA. B. Comparison of the total number of unique peptides enriched by individuals of different age groups and disease status. "Youth" individuals are 12 years old or younger; "Adult" individuals are 18 years old or older. Statistical comparisons of the means were performed using the Student's t test, with one tail.

Dz	Gene Symbol	Gene name associated with peptide or ORF	-log10 Fisher P val	Cluster	Summary of positives							Extra-cellular
					T1D (39)	RA (64)	MS (57)	HC (73)	BC (29)	OA/Gout (20)	CSF Ctrl (10)	
RA	BCOR	BCL6 corepressor	5.9	RA1	0	11	0	1	0	1	0	N
	LOC645453	ring finger protein, LIM domain interacting; similar to ring finger protein (C3H2C3 type) 6	5.3	RA1	0	9	1	0	0	0	0	N
	ATAD5	ATPase family, AAA domain containing 5	4.1	RA1	0	6	0	0	0	0	0	N
	Hcn3	hyperpolarization activated cyclic nucleotide-gated potassium channel 3	4.0		0	7	0	0	0	1	0	P
	FAM135A	*family with sequence similarity 135, member A*	*3.4*		*0*	*7*	*1*	*0*	*0*	*1*	*0*	?
	HRNR	hornerin	3.4		0	7	0	0	1	0	0	N
	ADAM33	ADAM metallopeptidase domain 33	3.4	RA2	0	7	0	0	0	1	0	L
	PTK2	PTK2 protein tyrosine kinase 2	3.4	RA2	0	5	0	0	0	0	0	N
	SNRPB	small nuclear ribonucleoprotein polypeptides B and B1	3.1	RA2	0	8	1	1	1	1	0	N
	KRT33B	keratin 33B	2.7	RA2	0	5	0	0	0	1	0	N
	ATXN2	ataxin 2	2.7	RA2	0	5	0	0	0	1	0	N
	S100A11	S100 calcium binding protein A11; S100 calcium binding protein A11 pseudogene	2.7		0	5	0	0	0	1	0	N
	Lrba	LPS-responsive vesicle trafficking, beach and anchor containing	2.7	RA2	0	5	0	0	0	1	0	N
	CREB3L1	cAMP responsive element binding protein 3-like 1	2.7	RA2	0	5	0	0	0	1	0	N
	SEPT8	septin 8	2.7	RA2	0	5	0	0	0	1	0	N
MS	Krt75	keratin 75	6.7	MS1	0	0	9	0	0	0	0	N
	TRIO	triple functional domain (PTPRF interacting)	5.9	MS1	0	0	8	0	0	0	0	N
	Sox17	SRY (sex determining region Y)-box 17	5.4		0	1	13	5	1	0	0	N
	LOC388182	LOC388182	5.1	MS1	0	0	7	0	0	0	0	?
	METTL23	methyltransferase like 23	5.1	MS1	0	0	7	0	0	0	0	?
	DENND4C	DENN/MADD domain containing 4C	5.1	MS1	0	1	9	0	0	0	1	N
	PPARGC1A	*peroxisome proliferator-activated receptor gamma, coactivator 1 alpha*	*5.0*		*0*	*0*	*8*	*0*	*1*	*0*	*0*	N
	SFRS16	splicing factor, arginine/serine-rich 16	4.4	MS1	0	0	6	0	0	0	0	N
	KIAA1045	KIAA1045	4.4	MS1	0	0	6	0	0	0	0	?
	FRMD4B	FERM domain-containing protein 4B	4.4	MS1	0	0	6	0	0	0	0	N
	N/A	N/A	4.4	MS1	0	0	6	0	0	0	0	?
	RIMS2	regulating synaptic membrane exocytosis 2	4.3	MS1	0	1	7	0	0	0	0	Y
	PPP1R10	protein phosphatase 1, regulatory (inhibitor) subunit 10	3.8		1	2	15	9	4	1	0	N
	Baz2a	bromodomain adjacent to zinc finger domain, 2A	3.6	MS1	0	0	5	0	0	0	0	N
	tes	testis derived transcript (3 LIM domains)	3.6		0	0	5	0	0	0	0	P
	USP11	ubiquitin specific peptidase 11	3.1		0	0	6	0	2	0	0	N

Table 5.3: Peptide/ORF enrichments associated with disease. All disease-associated autoantigens with a false discovery rate of 10% are listed. ORF-only associations are shown in italics. If the peptide is among a nonrandomly assorted cluster, the name of that cluster is provided. The summary of enrichments provides the total number of individuals from each group that displayed immunoreactivity against the peptide/ORF. T1D, type 1 diabetes; RA, rheumatoid arthritis; MS, multiple sclerosis; HC, healthy controls; BC, breast cancer; OA, osteoarthritis; CSF, cerebrospinal fluid. In parentheses are the total number of individuals from the group. N/A: no longer associated with an expressed sequence. Last column indicates whether protein is predicted to be localized extracellularly: "N", no; "Y", yes; "P", possibly; "L", likely; "?", unknown.

Dz	Gene Symbol	Gene name associated with peptide or ORF	Cluster
RA	BCOR	KPSKLAKRIANSAGYVGDRFKCVTTELYADSSQLSR	RA1
	LOC645453	VLDLQVRRVRPGEYRQRDSIASRTRSRSQTPNNTVT	RA1
	ATAD5	SKNISKAKQLIEKAKALHISRSKVTEEIAIPLRRSS	RA1
	Hcn3	MYFIQHGLLSVLARGARDTRLTDGSYFGEICLLTRG	
	HRNR	STHGQHGSTSGQSSSCGQHGASSGQSSSHGQHGSGS	
	ADAM33	VLQGHIPGQPVTPHWVLDGQPWRTVSLEEPVSKPDM	RA2
	PTK2	RTHAVSVSETDDYAEIIDEEDTYTMPSTRDYEIQRE	RA2
	SNRPB	RPPMGPPMGIPPGRGTPMGMPPPGMRPPPPGMRGLL	RA2
	KRT33B	LECEINTYRSLLESEDCKLPSNPCATTNACEKPIGS	RA2
	ATXN2	ELTANEELEALENDVSNGWDPNDMFRYNEENYGVVS	RA2
	S100A11	MAKISSPTETERCIESLIAVFQKYAGKDGYNYTLSK	
	Lrba	VMDNMVMACGGILPLLSAATSATHELENIEPTQGLS	RA2
	CREB3L1	MDAVLEPFPADRLFPGSSFLDLGDLNESDFLNNAHF	RA2
	SEPT8	LKIRRSLFDYHDTRIHVCLYFITPTGHSLKSLDLVT	RA2
MS	Krt75	MSRQSSITFQSGSRRGFSTTSAIT**PAAGRSR**FSSVS	MS1
	TRIO	SGGPSSCGGA**PSTSRSR**PSRIPQPVRHHPPVLVSSA	MS1
	Sox17	SALHVYYGAMGSPGAGGGRGFQMQPQHQHQHQHQHH	
	LOC388182	KRTPPAPQNPGGSTQAPQRVVGKSHSGIRM**PAKSRN**	MS1
	METTL23	MLGRSRATATW**PAASRSR**SLAARSLPRSPARPGPND	MS1
	DENND4C	LDHGSPAQENPESEKSS**PAVSRSK**TFTGRFKQQTPS	MS1
	SFRS16	TRSRSHSPS**PSQSRSR**SRSRSQSPSPSPAREKLTRP	MS1
	KIAA1045	SAAPEPR**PAPGRSR**AMGVLMSKRQTVEQVQKVSLAV	MS1
	FRMD4B	WPGRTVKDEFGGRLDP**PAASRSR**RREPRRAGRWGRG	MS1
	N/A	NPGCHPSPTPMLASQRHCRDSAAVCVHVT**PSPSRSL**	MS1
	RIMS2	SMPSLMTGRSAPPS**PALSRSH**PRTGSVQTSPSSTPV	MS1
	PPP1R10	HRPHEGPGGGMGAGGGHRPHEGPGGSMGGSGGHRPH	
	Baz2a	AAHASLNPALFSMKMELAGSNTTASS**PARARSR**PLK	MS1
	tes	KRNVMILTNPVAAKKNVSINTVTYEWAPPVQNQALA	
	USP11	ISHSCVGCRRERTAMATVAANPAAAAAAVAAAAAVT	

Table 5.4: Sequences of MS and RA specific peptides. Sequences of peptides associated with disease at a false positive discovery rate of 10%. If the peptide is among a nonrandomly assorted cluster, the name of that cluster is provided. The MS1 motifs are highlighted in bold.

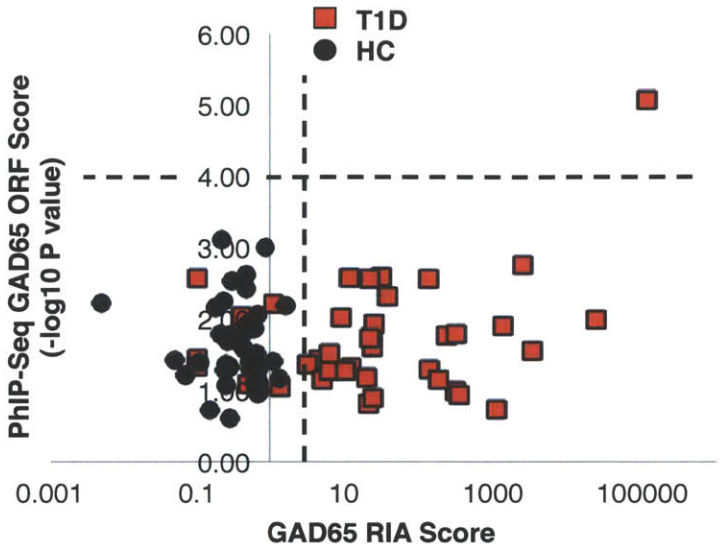


Figure 5.4: PhIP-Seq false negative rate for GAD65 autoantibodies. Sensitivity of GAD65 autoantibody detection by PhIP-Seq compared to RIA in T1D patients and healthy controls. PhIP-Seq values correspond to the most enriched peptide from the ORF. Dashed lines illustrate threshold for positivity.

to uncover a shared sequence motif among RA1- and RA2-clustered peptides using blastp and MEME algorithms, none was identified [69].

MS patients are frequently found to have oligoclonal immunoglobulin in their CSF, which is resolvable by isoelectric focusing. As the presence of these oligoclonal IgGs is the most consistent laboratory abnormality in MS (detectable in about 95% of patients compared with 10%–15% of controls), it has long been assumed that the specificities of intrathecally-produced antibodies harbor clues to the pathogenesis of the disease. We therefore screened 28 CSF samples and 35 serum samples (including 6 CSF-serum pairs) from patients with clinically definite MS. As additional negative controls, we screened 10 CSF samples from individuals with subacute sclerosing panencephalitis (SSPE), paraneoplastic neurological disorder (PND) and meningitis. We examined the set of 15 peptides that were enriched by MS patients with a disease association FDR of <10% (Figure 4A, Table 5.3). Eleven of these peptides assorted non-randomly among a subset of MS patients, and motif discovery revealed a 7 amino acid sequence contained in all of them ("MS1", Figures 5.6B-D). Notably, a motif nearly identical to MS1 was previously identified by Cepok et al. in a similar screen of MS CSF samples [98], and they reported an alignment with the BRRF2 protein of the Epstein-Barr virus, a pathogen repeatedly implicated in MS pathogenesis. We performed an alignment of the MS1 motif against the UniProt database of all proteins from viruses with human tropism, collapsed onto 90% identity clusters (7546 UniRef

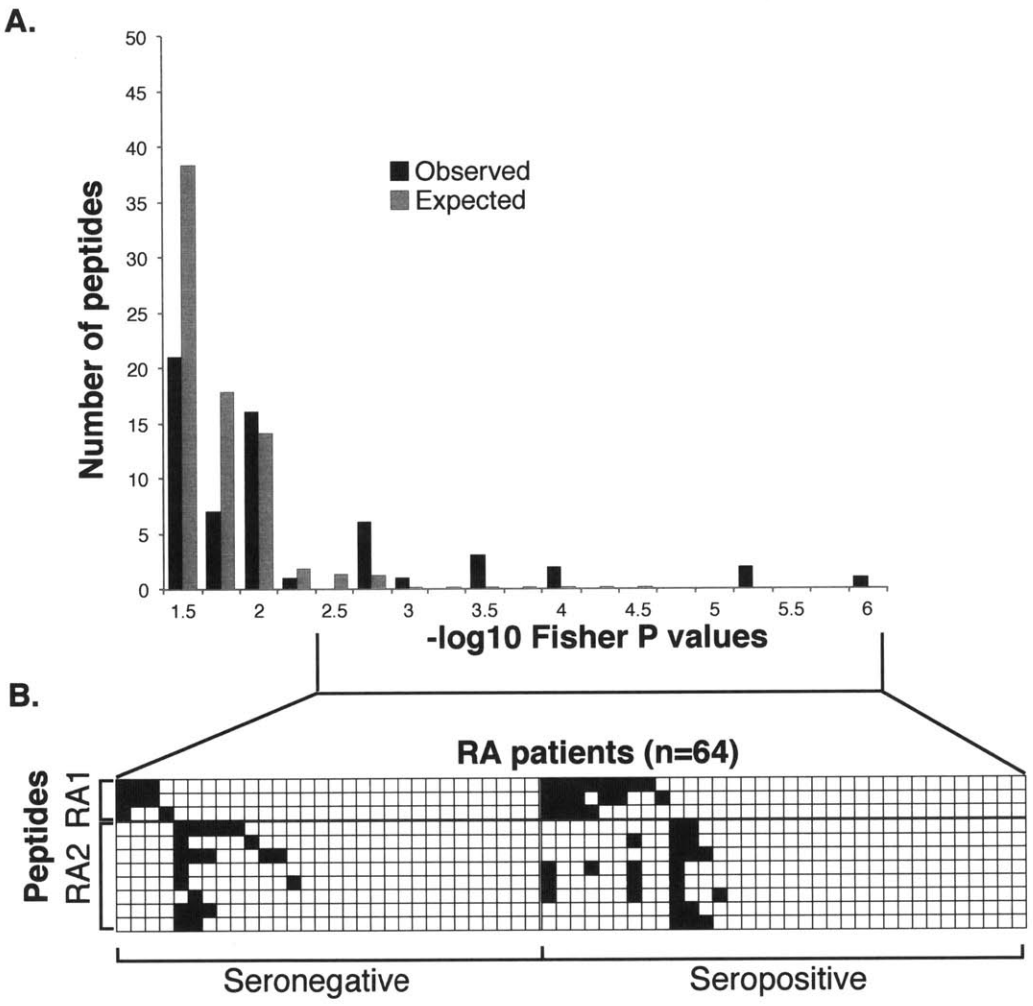


Figure 5.5: RA associated peptides and their clusters. A. Permutation analysis of peptide enrichments associated with RA. "Observed" bars indicate the number of peptides associated with RA at a given P-value by Fisher's exact test. "Expected" bars show the number of peptides expected to have a -log10 Fisher P-value at least that extreme due to chance alone (as determined by permuting sample labels). B. RA1 and RA2 peptide enrichment heat map (as in Figure 5.2) illustrating nonrandom segregation of peptide enrichments (rows) and RA patients (columns). -Log10 P-values less than 4 are white and greater than 4 are black. Patients are organized by their seropositivity, and by their RA1/RA2 status.

sequences; 656 unique taxa), and also found the best alignment to be with the EBV BRRF2 protein (E value = 1.2; sequence: PAASRSK).

We considered the possibility that a peptide containing the MS1 motif might have clinical utility in the form of an ELISA assay. To this end, we immobilized the peptide which performed best in our PhIP-Seq screen, Krt75_1 (9 positives of 57 MS samples, versus 0 positive of 239 non MS samples). Of 25 MS CSF samples tested by ELISA, 3 were positive, compared to 0 of 19 CSF samples from individuals with other inflammatory neurological diseases (Figure 5.7A). Eight of the ELISA-tested MS samples had also been screened using PhIP-Seq, and we found the latter method to have a greater sensitivity (Figure 5.7B).

5.3.3 Analysis of matched MS samples

As part of our collection, we obtained six sets of MS CSF-serum pairs. Each of these samples was screened in duplicate, and we considered only those peptides that were reproducibly enriched with a -log10 P-value greater than 3 in both replicates from either compartment. For each of these MS patient pairs, we plotted the average -log10 P-value for each peptide's CSF enrichment against the average serum enrichment (Figure 5). In all cases we observed a strong correlation in the enrichment profiles between these two fluid compartments. A majority of the enrichments were found in both compartments, with a trend toward stronger enrichment in the serum. In several cases, however, we did find peptides that were more highly enriched in the CSF compartment. For example, CSF from patient 9292 enriched two homologous peptides from interferon alpha 5 and 14 much more significantly than serum from the same patient (Figure 5.8A; Table 5.5). This is unlikely to reflect cross-reactivity of inhibitor antibodies to therapeutic interferon beta, however, as the homologous peptide from interferon beta was not enriched in either compartment.

We systematically examined all the CSF-specifically enriched peptides (enriched by CSF antibodies with -log10 P-value of at least 3 greater than the corresponding serum enrichment) that were identified in the six patients (Table 5.5). Motif discovery was performed on each set of CSF-specific peptides, and one motif was uncovered for patient 10894 (Figure 5.8B and Table 5.5). This motif was searched into the database of human viruses, and a significant alignment was found with the major capsid protein VP1 of the JC polyomavirus (JCV; E value = 0.03; sequence: RRVKNP). Similar to EBV, JCV infection is highly prevalent, infecting 70 to 90 percent of humans. Also of note, JCV can cross the blood-brain barrier into the central nervous system, where it infects oligodendrocytes and astrocytes, possibly through the 5-HT2A serotonin receptor [99].

Some MS patients exhibited little or no CSF-specific autoreactivity, an example of which is shown in Figure 5.8C (patient 8911). This patient, however, did have serum samples drawn on

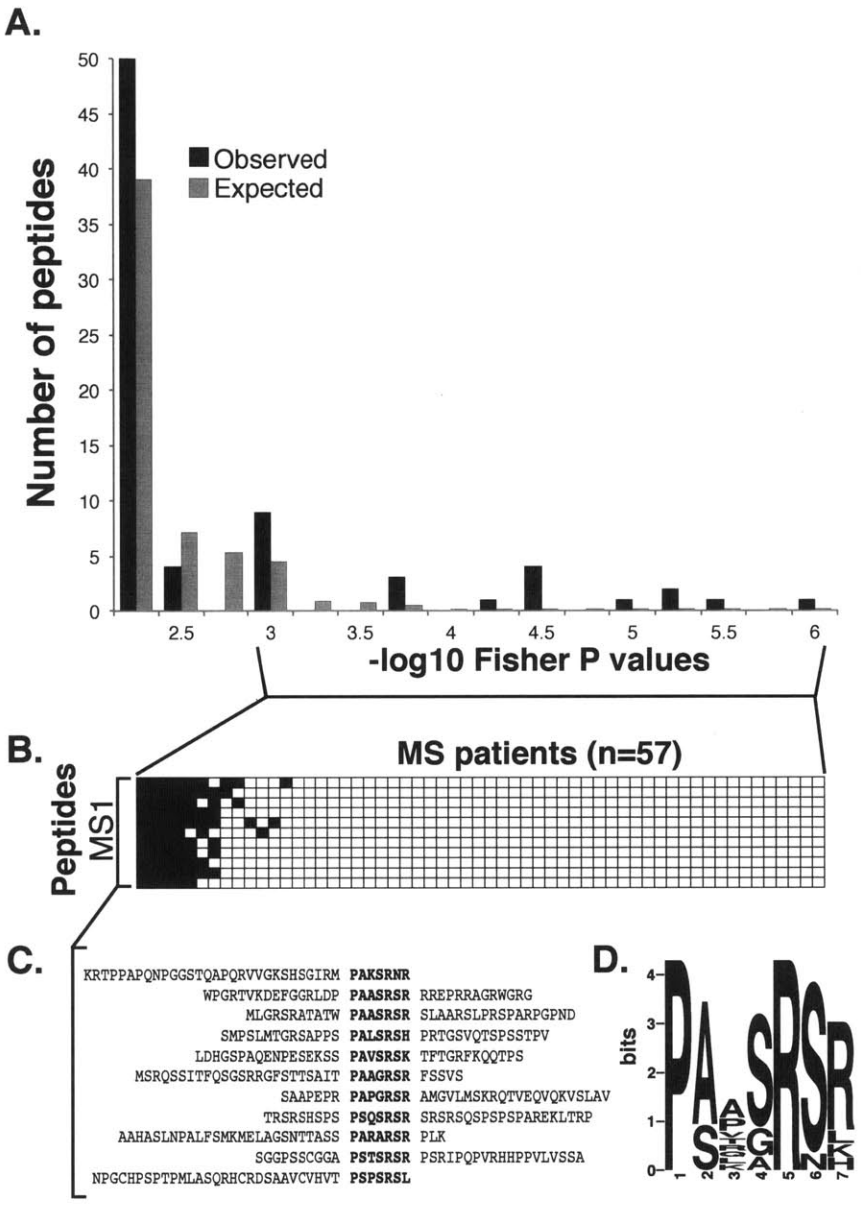


Figure 5.6: MS associated peptides share a sequence motif. A. Permutation analysis of peptide enrichments associated with MS. "Observed" bars indicate the number of peptides associated with MS at a given P-value by Fisher's exact test. "Expected" bars show the number of peptides expected to have a -log10 Fisher P-value at least that extreme due to chance alone (as determined by permuting sample labels). B. Peptide enrichment heat map (as in Figure 5.2) illustrating nonrandom segregation of MS1 peptide enrichments (rows) and MS patients (columns). -Log10 P-values of enrichment less than 4 are white and greater than 4 are black. Patients are organized by their MS1 status. C. Alignment of the co-segregated peptides reveals a shared epitope. D. MS associated epitope (MS1) motif logo, calculated from the peptides in C (MEME software).

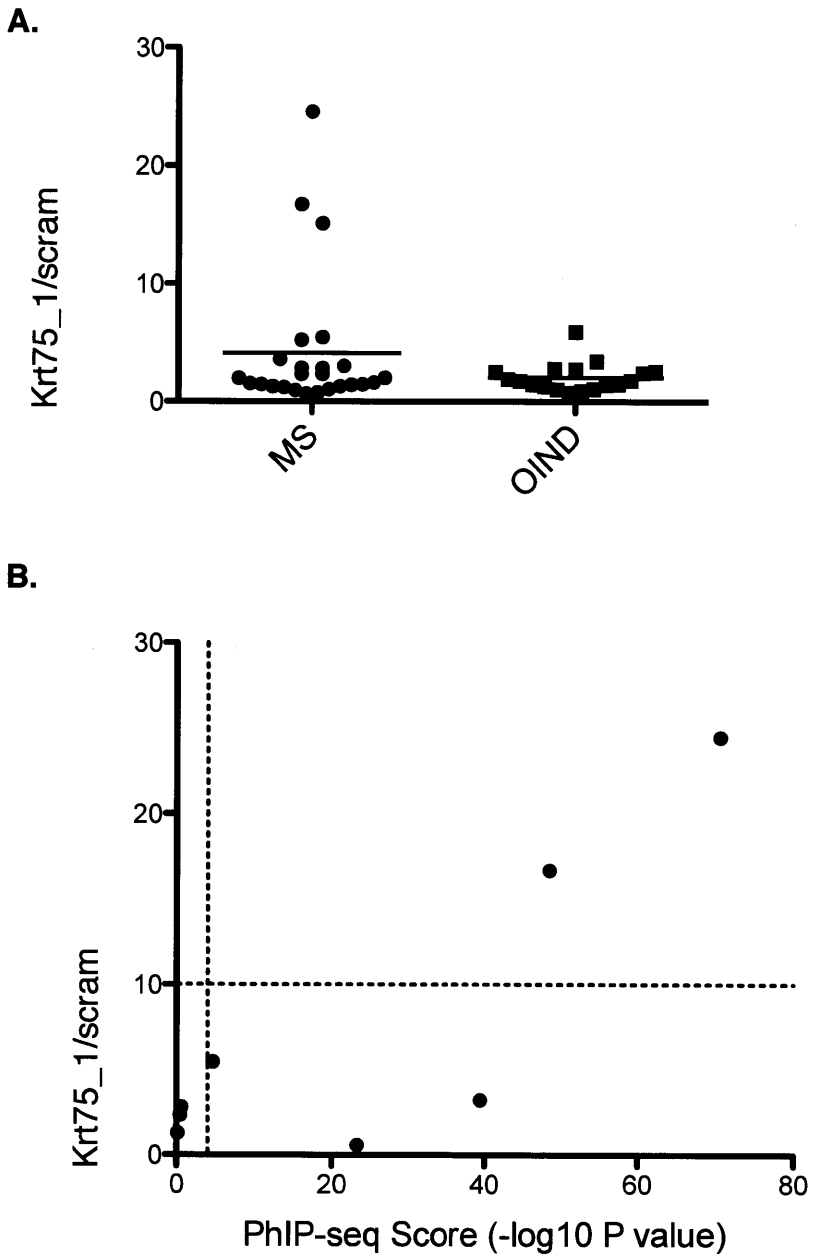


Figure 5.7: ELISA testing of MS peptide Krt75_1. A. Results of ELISA assay comparing ratio of signal from Krt75_1 peptide to signal from scrambled peptide ("scram") between 25 MS patient CSF samples and 19 other inflammatory neurological diseases ("OIND") patient CSF samples. Horizontal bar indicates mean ratio of cohort. B. Comparison between ELISA assay and PhIP-Seq assay. ELISA confirmed 2 of the 5 PhIP-Seq positive CSF samples. 8 CSF samples in total were tested by both methods. Dotted lines indicate threshold of positivity.

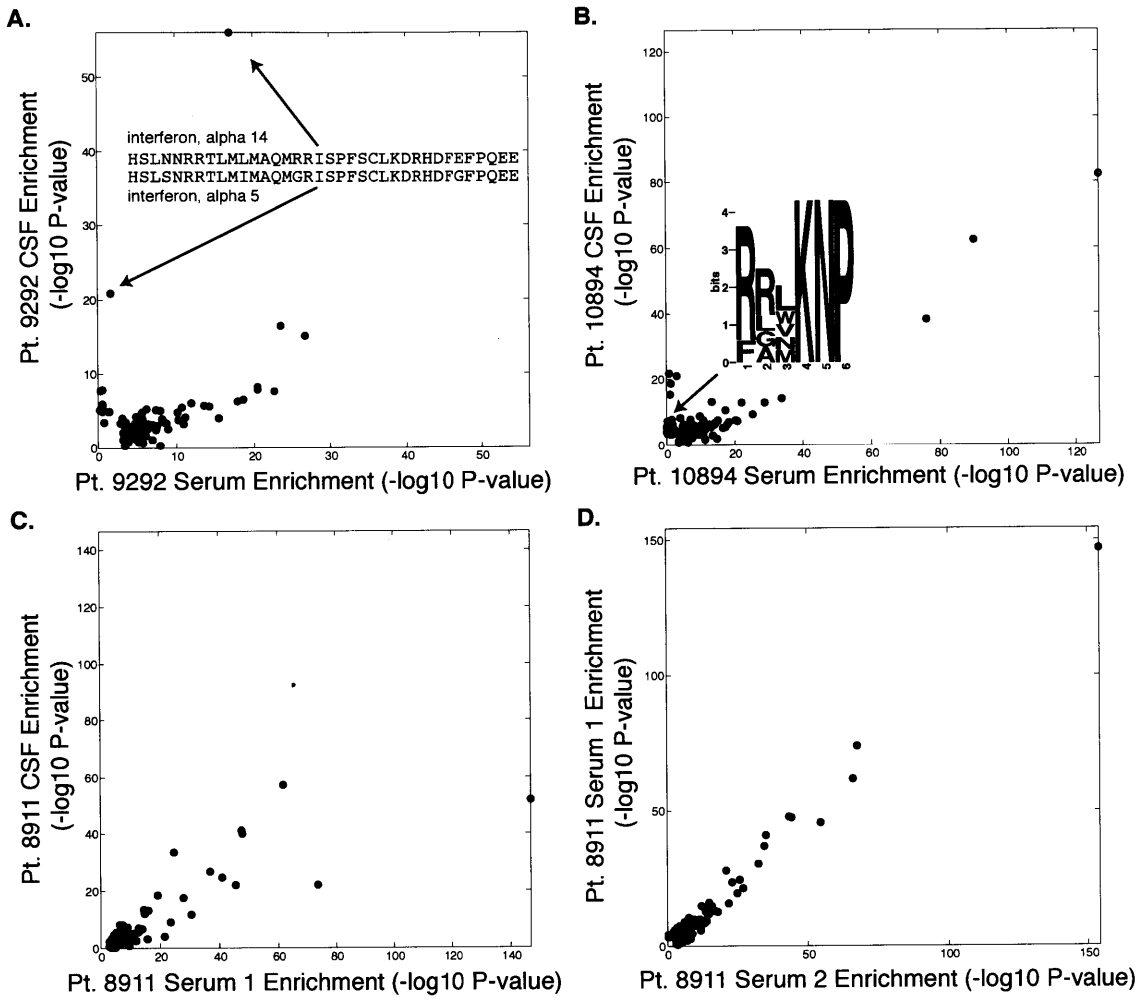


Figure 5.8: Analysis of MS patient CSF/serum pairs. Scatter plots of matched samples from the same individuals. Each sample was analyzed in duplicate and the average -log10 P-value is plotted for peptides enriched by more than -log10 P value of 3 in both duplicates. A. Patient 9292 peptide enrichments in CSF versus serum, and enrichment of nearly identical peptides from IFN-α5/14 specifically in the CSF. B. Patient 10894 peptide enrichments in CSF versus serum, and CSF-specific enrichment of the JCV motif is illustrated. C. Patient 8911 peptide enrichments in CSF versus serum. D. PhIP-Seq serum profile in patient 8911 taken at two time points.

Patient	-log10 P-value CSF	Serum	Peptide sequence	Symbol	Gene name
	56.1	17.0	HSLNNRRTLMLMAQMRRISPFSCLKDRHDFEFPQEE	IFNA14	interferon, alpha 14
	20.8	1.6	HSLSNRRTLMIMAQMGRISPFSCLKDRHDFGFPQEE	IFNA5	interferon, alpha 5
	7.7	0.2	IIANALSSEPACLAEIEEDKARRILELSGSSSEDSE	PRKDC	similar to protein kinase, DNA-activated, catalytic polypeptide
	7.9	0.6	RPLTTQKLILRVESLLEVRPGNTRQKKQEDHSSGSL	LOC283682	hypothetical protein LOC283682
9292	5.7	0.5	AFDVQASPNEGFVNQNITIFYRDRLGLYPRFDSAGR	HYAL2	hyaluronoglucosaminidase 2
	5.2	0.2	QFQLLEQEITKPVENDISKWKPSQSLPTTNSGVSAQ	Nedd9	neural precursor cell expressed, developmentally down-regulated 9
	4.9	0.5	AESLLEAGDMLQFHDVRDAAAEFLEKNLFPSNCLGM	Klhl25	kelch-like 25 (Drosophila)
	5.0	1.4	MVLGKVKSLTISFDCLNDSNVPVYSSGDTVSGRVNL	arrdc3	arrestin domain containing 3
	4.9	1.3	NKVLIAQKLHECARCGKNFSWHSDLILHEQIHSGEK	ZNF311	zinc finger protein 311
	16.5	8.6	LYSAPIFSSNYSSRSGTAAGAVPPPHPVSHPSPGHN	Glis3	GLIS family zinc finger 3
	6.6	1.1	AQEMFTYICNHIKVATNRGNLRSAITVFPQRCPGRG	NOS3	nitric oxide synthase 3 (endothelial cell)
	42.6	38.2	LDNTKFRSHEGETAYIRVKVDGPRSPSYGRSRSRSR	SFRS1	splicing factor, arginine/serine-rich 1
	8.2	3.9	SDWEKSSNGRQWKPQLGFNRDRRPVHLDQAAFRTLG	XRN2	5-3 exoribonuclease 2
9316	5.1	1.1	KLIAGKLRETLNISGPPLKAGKTRNFYGLHQDFPSV	LAP3P2	LAP3P2 pseudogene
	4.1	0.4	CVHTPSPCSTPPVKSLEEDPWPRVNSKDHIPALVRS	PPM1D	protein phosphatase 1D magnesium-dependent, delta isoform
	9.6	5.9	EESDIDSEASSAFFMAKKKTPPKRERKPSGGSSRGN	GTF2F1	general transcription factor IIF, polypeptide 1, 74kDa
	13.7	10.1	SPHEAWNRLHRAPPSFPAPPPWPKSVDAERVSALTN	FBRSL1	fibrosin-like 1
	4.0	0.6	NKKMMERLHKIKICIKESGQKLKKSRSVQSREENEM	HERC1	hect (homologous to the E6-AP (UBE3A) carboxyl terminus) domain and RCC1 (CHC1)-like domain (RLD) 1
	3.5	0.4	PAGALTPGPSTRLRCSVILAPAHPGETGPSPAPLRS	removed	removed
	13.0	0.7	VDEYMLPFEEEIGQHPSLEELQEVVVHKKMRPTIKD	ACVR2B	activin A receptor, type IIB
	30.5	21.7	RKITRPSQRPKTPPTDIIVYTELPNAESRSKVVSCP	KIR2DL1	killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 2
	8.3	0.7	TDTSLTMDIYFDENMKPLEHLNHDSVWNFHVWNDCW	Tgm1	transglutaminase 1
	17.7	11.5	QQVCHAIANISDRRKPSLGKRHPPFRLPQEHRLFER	NCAPD2	non-SMC condensin I complex, subunit D2
	7.4	1.7	QIQVTHGKVDVGKKAEAVATVVAAVDQARVREPREP	TTN	titin
	21.7	16.3	DTNKAHPDNKAEFPSYLLGGRSGALKNFVIPKIKRD	nipbl	Nipped-B homolog (Drosophila)
	5.8	0.4	ALGEFVLVEKDVKISKKGKIYNLNEGNAKYFDRAVT	*removed*	*removed*
	6.3	1.0	KSQLQKVSGVFSSFMTPEKRMVRRIAELSRDKCTYF	RIN2	Ras and Rab interactor 2
	6.3	1.6	PFPSSPPFPSSPPFPSSPPFPSSPPFPSSPPFPSSP	*removed*	*removed*
	6.5	1.9	MMSEKKKKNPTPESVAIGELKGTSKENRNLLFSGSP	ARHGAP19	Rho GTPase activating protein 19
	5.5	0.8	MAELQQLQEFEIPTGREALRGNHSALLRVADYCEDN	ABI3	ABI family, member 3
9358	6.0	1.8	LVNSLKVWGKKRDRKSAIQDIRISPDNRFLAVGSSE	EML6	echinoderm microtubule associated protein like 6
	8.4	4.4	EDNPRDLQLLRHDLPLHPAVVKPHLGHVPDYLVPPA	ddx56	DEAD (Asp-Glu-Ala-Asp) box polypeptide 56
	5.4	1.4	IRMPPLRNVGAGGVSGAIRTPRPMGQEASVTTGLGR	ELFN1	leucine-rich repeat and fibronectin type III domain containing 1
	8.3	4.4	AVWCQNVKTASDCGAVKHCLQTVWNKPTVKSLPCDI	PSAP	prosaposin
	71.7	67.9	KKKSSWKAFFGVVEKEDSQSTPAKVSAQGQRTLEYQ	BCL2L14	BCL2-like 14 (apoptosis facilitator)
	5.1	1.4	IRLPSLYHVLGPTAADAGPESEKGDEEVCEPAVSPP	POPDC2	popeye domain containing 2
	5.6	2.1	IKEIKMEEERNIIPREEKPIEDEIERKENIKPSLGS	ARID4B	AT rich interactive domain 4B (RBP1-like)
	6.0	2.5	SSSLTVHKRTHVGRETIRNGSLPLSMSHPYCGPLAN	ZNF333	zinc finger protein 333
	6.1	2.8	QNIWNINLQLRPSLITGIMKDSGNKPPGLLPRKGLY	gstk1	glutathione S-transferase kappa 1
	4.9	1.8	KKRSLWDTIKKKKISASTSHNRRVSNIQNVNKTFSV	ASPM	asp (abnormal spindle) homolog, microcephaly associated (Drosophila)
	4.9	1.9	GKDRVVSLSEKNFKQVLKKYDLLCLYYHEPVSSDKV	Casq2	calsequestrin 2 (cardiac muscle)
	3.3	0.3	LRSQLILKLRQHYRELCQQREGIEPPRESFNRWMLE	PCIF1	PDX1 C-terminal inhibiting factor 1
9733	4.1	0.4	QSQSAHLWKSPFPDVVPLQPEVSSYRRGRKKRVPYT	HOXC13	homeobox C13
	21.8	0.9	RAACYFTMGLYEKALEDSEKALGLDSESIRALFRKA	ZC3H7B	zinc finger CCCH-type containing 7B
	21.2	3.1	LNKATDPSMSEQDWSAIQNFCEQVNTDPNGPTHAPW	Gga2	golgi associated, gamma adaptin ear containing, ARF binding protein 2
	18.9	1.3	AEDLEDVRAEGTEDVGTEGTEDVGAEDSEDIRAESS	*removed*	*removed*
	15.4	1.1	LRLEAPSPKAIVTRTALRNLSMQKGFNDKFCYGDIT	PDZD8	PDZ domain containing 8
	7.2	0.4	EDGGSEITNYIVDKRETSRPNWAQVSATVPITSCSV	TTN	titin
	7.7	1.6	SLLPEGEDTFLSESDSEEERSSSKRRGRGSQKDTRA	GTF3C1	general transcription factor IIIC, polypeptide 1, alpha 220kDa
	6.6	0.9	KVDEYTDTDLYTGEFLSFADDLLSGLGTSCVAAGRS	Astn2	astrotactin 2
	6.0	0.4	VSDVSRDSVNLTWTEPASDGGSKITNYIVEKCATTA	TTN	titin
	6.1	0.6	RPVPGCVNTTEMDIRKCRRLKNPQKVKKSVIGVTEE	RGS6	regulator of G-protein signaling 6
	5.5	0.1	LLDTQRDGLQNYEALLGLTNLSGRSDKLRQKIFKER	UNC45B	unc-45 homolog B (C
10894	5.6	1.0	GEDGSRRFGYCRRLLPGGKGKRLPEVYCIVSRLGCF	DENND2A	DENN/MADD domain containing 2A
	5.0	0.5	LLPRTKGFTTAVKCLRGTVAAVYDVTLNFRGMKMPS	AGPAT3	1-acylglycerol-3-phosphate O-acyltransferase 3
	6.3	1.8	RSRSKDEYEKSRSRSRSRSPKENGKGDIKSKSRSRS	Sfrs6	splicing factor, arginine/serine-rich 6; similar to arginine/serine-rich splicing factor 6
	5.8	1.3	EQKLKLERLMKNPDKAVPIPEKMSEWAPRPPPEFVR	prkrip1	PRKR interacting protein 1 (IL11 inducible)
	5.0	0.5	SPGEWQQASAGPLHLSVPEPGRAWKNPERGSKSRWS	BCYRN1	brain cytoplasmic RNA 1 (non-protein coding)
	5.1	0.7	PEFEDSEEVRRIWNRAIPLWELPDQEEVQLADTMFG	c10orf2	chromosome 10 open reading frame 2
	8.3	4.2	EKTIEDYIDKKMNDADSTSVEAMYSGASQCRHEKKN	IRAK4	interleukin-1 receptor-associated kinase 4
	3.9	0.1	QKTCQEQELLKQEDISMTNLGSMACPIMEPLHLENT	CCDC168	coiled-coil domain containing 168
	4.2	0.9	SVGKQDKSGLLMKLQNLCTRLDQDESFSQRLPLNIE	Apaf1	apoptotic peptidase activating factor 1
	3.7	0.4	DAVYLDSEEERQEYVLTQQGFIYQGSAKFIKNIPWN	tgm2	transglutaminase 2 (C polypeptide, protein-glutamine-gamma-glutamyltransferase)
	3.6	0.3	AEDPNLNQPVWMKPCRINSSYFRRVKNPNNLDEIKS	CEP350	centrosomal protein 350kDa
	3.7	0.7	SVSASSCVLSWGEPKDGGGTEITNYIVEKRESGTTA	TTN	titin
8911	33.6	24.5	SGSGQSPGHGQRGSGSRQSPSYGRHGSGSGRSSSSG	HRNR	hornerin

Table 5.5: Peptides more enriched in CSF. Average -log10 P-values of duplicate peptide enrichments are reported if they were at least 3 larger after IP with CSF compared to serum from the same patient. The CSF-specific motif in patient 10894 is shown in bold.

two separate occasions within one year, which allowed us to examine the persistence of PhIP-Seq enrichments over this length of time. The scatterplot (Figure 5.8D) reveals minimal time-dependent changes.

We were intrigued by the observation that the ACVR2B_15 peptide, which was very frequently enriched in the sera of healthy individuals (Figure 5.2A), was enriched only in the CSF compartment of MS patient 9358 (-log10 P-value of 17.8 in CSF compared to 0.8 in serum; Table 5.5). In hopes of identifying an epitope within the ACVR2B_15 peptide, we searched for peptide enrichments that were highly correlated with ACVR2B_15. MEME analysis revealed a motif shared by 3 peptides that were only enriched when ACVR2B was also enriched. Interestingly, the most significant viral peptide alignment was again found within the proteome of EBV, but this time with the latent membrane protein-1 (LMP-1; E value = 0.03; sequence: LTEEVANK).

5.4 Discussion

In this study we report the first large scale PhIP-Seq screen of a population of individuals with different autoimmune diseases for direct comparison to healthy controls and to each other. These data provide an unbiased, proteomic-scale assessment of precise autoreactivities found within 298 independent antibody repertoires. The vast majority of autoreactivities were individually unique, lending support to the notion that each person possesses a unique "autoantibodyome", of which the impact on phenotype remains to be explored. It is interesting to note that as our database of enriched peptides grows, so will the number of peptides recurrently enriched by a small fraction of the population - a situation analogous to the ongoing identification of progressively less common alleles in sequenced genomes. Screening large numbers of genotyped individuals will additionally reveal correlations between autoreactivities and HLA haplotypes, antibody variable domain alleles, and other immunogenetic modifiers.

Our unbiased method revealed a large number of novel peptide autoreactivities, but when compared to RIA-determined titers of known autoantibodies, appears to suffer from relatively low sensitivity. We detected no anti-insulin antibodies in the T1D patients, with the important caveat that we did not charcoal-extract insulin from the serum prior to performing PhIP-Seq, which is standard protocol for the RIA assay. It is therefore possible that the anti-insulin antibodies were occupied by endogenous or injected insulin and therefore not accessible for peptide binding. Additionally, ZnT8 RIA titers were obtained using a fusion protein consisting of two allelic variants of the immunodominant epitope, and so the single consensus sequence in T7-Pep (the "CR" variant) may have contributed to the low sensitivity. The most important source of the high false negative rate, however, is most likely the limited amount of conformational

structure inherent to 36 amino acid peptide tiles. The findings presented here thus highlight the need for improved display libraries that include more complex epitopes.

Despite this limitation, we observed a significantly accelerated polyautoreactivity in the sera of younger T1D patients compared with their matched controls. To our knowledge, this finding has not been explicitly reported previously. Several possible factors may confound this finding. Perhaps most obvious is the role that HLA haplotype could play, since T1D genetic risk is tightly linked to MHC class II alleles. It would therefore be interesting to explore the relationship between T1D risk and protection-conferring alleles and PhIP-Seq polyreactivity. Epidemiologically, these data are consistent with the existence of a "risk window", during which increased polyreactivity provides more opportunities to acquire pathogenic autoreactivity.

While we did not observe T1D-specific peptide/ORF enrichments at a frequency above that expected by chance, the immunodominant peptide from PTPRN was enriched by 3 T1D individuals and none of the non-T1D individuals. Increasing the power of this study would thus likely reveal this known, as well as additional novel, T1D-associated autoantigens.

In contrast to T1D, our RA study was sufficiently powered to uncover novel disease-associated anti-peptide antibodies. 13 out of the 64 RA patients (20%) exhibited immunoreactivity against at least one RA1 peptide, compared to 3 of 232 non RA individuals (1.3%). In addition, 16 of the RA patients (25%) exhibited immunoreactivity against at least one RA2 peptide, compared to 6 non-RA individuals (2.6%). Taken together, 26 of the RA patients (41%) exhibited immunoreactivity against at least one RA1 or RA2 peptide, compared to 9 non RA individuals (3.9%; $P = 7.8 \times 10^{-13}$, Fisher's exact test, one tail); 16 RA patient samples enriched at least two peptides from RA1 or RA2. Further work is required to determine the nature of the relationships among these correlated autoreactivities.

Much effort has been invested to identify the specificities of oligoclonal bands in the CSF of MS patients. Cortese et al. used a library of constrained nonamers to find mimitopes for CSF antibodies in 2 MS patients . One of the sequences (KPPNP) is contained within several of our library peptides. Of them, one peptide from XP_499190.1 (SQQWRENPRTQNQSAVERKPPNPEPVSSGEKTPEPR), was enriched by 6 of 57 MS patients and 9 of 235 non-MS individuals, and so was weakly associated with MS (Fisher's P value = 0.05). Perhaps most notable of these studies, however, Rand et al. used a small collection of CSF samples from MS patients to screen a phage library of random hexamers [100]. They uncovered a recurrently enriched sequence (RRPFF) in several individuals, and reported alignment with the heat shock protein αB crystallin and the Epstein-Barr virus nuclear antigen (EBNA-1). In our study, the most commonly enriched peptide by healthy individuals, MAGEE1_25, contains this same sequence (RAFAEGWQALPHFRRPFFEEAAAEVPSPDSEVSSYS; 32 of 73 positive; Figure 5.2A). We also detected MAGEE1_25 immunoreactivity in 10 of the 27 MS CSF

samples and in 1/7 non-MS CSF controls. MAGEE1_25 was enriched with equal frequency in the serum of MS patients compared to healthy controls (17/29 MS and 17/29 HC). Of the six MS patients for which we had matching CSF and serum samples, two had MAGEE1_25 antibodies. Both of them exhibited stronger enrichment in their serum than in their CSF. Taken together, we believe the PhIP-Seq data are consistent with a scenario in which RRPFF antibodies occur with equal frequency in the serum of MS and healthy individuals, and suggest that they are unlikely to be produced specifically within the CNS. In sum, we detected three EBV associated epitopes and one JCV associated epitope in our analysis. The EBNA-1 and LMP-1 epitopes were ubiquitously enriched, whereas the BRRF2 epitope (MS1) was targeted with high specificity by patients with MS. Importantly, the MS1 antibodies exhibited a notable degree of polyspecificity for self peptides (Figure 5.6B).

The majority of autoreactivities observed in MS patients' CSF were also observed in the serum of the same individuals, though usually to a lesser extent. This result is somewhat surprising, given that the total IgG concentration in CSF tends to be dominated by intrathecal production. One explanation is that the majority of these intrathecal antibodies do not bind epitopes contained within the T7-Pep phage library. Our comprehensive scan of minimally conformational autoepitopes would therefore suggest that a universal MS associated intrathecal specificity, if it exists, is likely to be either highly conformational, not human, post translationally modified, or includes a non-protein component. Our results also imply that a blood test (without lumbar puncture) will likely be sufficient to detect MS specific autoantibodies.

The findings presented here point to the accumulating value of high throughput, low cost PhIP-Seq screening. As the sample size of our database grows, so will the power to detect rare, yet significantly disease-associated autoantibodies. Quantitative elucidation of these diverse autoreactivities will be particularly important for understanding complex, heterogeneous autoimmune disease pathogeneses. In the future, methods that query linear and conformational epitopes, as well as T cell epitopes, for both human and pathogen antigens will eventually provide us with a comprehensive description of autoimmunity.

5.5 Supplementary Discussion

Titin autoantibodies have long been investigated in association with myasthenia gravis (MG), since they occur in 20-30% of patients with this autoimmune disease [101, 102]. The main immunogenic region of titin was mapped to a 30 KDa fragment spanning amino acids 7025–7311 of the novex-2 isoform (NP_597681.3) [103]. A second, recently discovered, MG-associated immunodominant region was mapped to amino acids 10319–10532.(19) Within our dataset, three individuals (one healthy, one MS patient, and one BC patient), demonstrated reactivity against

a single peptide from the first region (7193–7228), and one BC patient from our study had antibodies targeting a peptide within the second region (10441–10476). It would be interesting to determine whether these peptides are in fact the minimal epitopes of the MG-associated titin antibodies. Strikingly, one titin peptide (8179–8214), which is not derived from either MG-associated region, was enriched by 85 individuals, making it the third most commonly enriched peptide by healthy controls (Figure 5.2A). To our knowledge, autoreactivity toward this peptide has not been previously described, but due to its prevalence in healthy controls, is unlikely to have pathological consequences.

5.6 Methods

5.6.1 Patient samples

Specimens originating from patients were collected after informed written consent was obtained and under a protocol approved by the local governing human research protection committee. Specimens which did not include personally identifiable private information or intervention or interaction with an individual were collected under an exempt protocol approved by the local governing human research protection committee.

5.6.2 T1D patient samples and matched controls

Type 1 diabetic patients (n=39, <40 years at diagnosis, male/female ratio = 1.18, average age 18 ± 2 years, range 3–37 years) were consecutively recruited by a Belgian network of endocrinologists between May 2004 and January 2006. Blood was sampled within 7 days from clinical onset/diagnosis by the Belgian Diabetes Registry (www.bdronline.be). Only diabetic patients with three or more samples during yearly follow-up by the Registry were included in this study. Age/sex-matched healthy control samples (n=41, male/female ratio = 1.18, average age 18 ± 2 years, range 4–37 years) were obtained from patients undergoing elective minor surgery. Controls were verified to be negative for all known type 1 diabetic autoantibodies.

5.6.3 Insulin, GAD65, PTPRN and ZnT8 autoantibody radioimmunoassay

After acid charcoal extraction of the endogenous and/or injected insulin, serum was incubated with radioactive labeled human recombinant insulin (mono-125I-tyrosin-A14-insulin) in the presence and absence of an excess of unlabeled insulin. Immune complexes were precipitated using polyethylene glycol (PEG). After washing (to remove the unbound 125I-insulin), radioac-

tivity of the PEG precipitate was measured. The IAA concentration is expressed as specific 125I-insulin binding capacity of the serum (% tracer bound of the total amount of tracer added). Sera with insulin binding $\geq$ 0.6% were considered IAA positive.

GAD65, PTPRN (amino acids 603–980), and ZnT8 (gene SEC30A8 is a chimeric construct of two peptides, amino acids 268–369), were produced in-house using in vitro transcription/translation of pEX9 (cDNA) using the Promega L4600 TnT-Kit. For ZnT8, the CR variant carries 325Arg while the CW variant carries 325Trp. The chimeric CW-CR construct contains both CR and CW [104]. The diabetes autoantibodies were determined by liquid-phase radiobinding assays as described previously [101].

5.6.4 Islet cell IgG cytoplasmic autoantibodies

Indirect immunofluorescence was performed on non-fixed cryosections of human O+ donor pancreas, calibrated to a Juvenile Diabetes Foundation (JDF)-standard (assigned arbitrarily an ICA titer of 200 JDF-units). Pancreas sections were incubated with a serial dilution of the unknown serum, washed with phosphate buffer, and attached anti-islet IgG visualized by FITC-labeled rabbit anti-human IgG gamma chain antibody. When islet immunoreactivity was detected, the exact ICA titer was determined by further serial dilution (2-fold step), and samples with titers $\geq$ 12 JDF-units are considered ICA+.

5.6.5 MS and encephalitis patient samples

A detailed clinical intake form was collected from outside investigators, summarizing the patient's neurological history, relapse features, neurological examination, MRI and CSF findings. For samples collected at the Brigham and Women's Hospital, the same information was obtained from the MS Center's clinical database. Patients were diagnosed with relapsing-remitting MS according to the McDonald criteria.

Viral encephalitis serum samples were provided by the New York State Department of Health. Sera from patients infected with West Nile virus or St. Louis Encephalitis virus were reactive in ELISA tests and were confirmed by cross species plaque reduction neutralization tests with paired acute and convalescent sera. Sera from patients with enteroviral infection were collected on the same day as spinal fluids for which PCR tests for enteroviruses were positive [102]. Healthy control samples were collected at Brigham and Women's Hospital from subjects self-reported to be free of MS or other autoimmune disease. All serum and CSF samples were stored in aliquots at -80°C.

5.6.6 Patient synovial fluids

Human knee synovial fluids were obtained as discarded material from patients with various arthritides undergoing diagnostic or therapeutic arthrocentesis. Arthritis diagnosis was ascertained by an American Board of Internal Medicine certified Rheumatologist and/or by review of laboratory, radiologic and clinic notes and by applying ACR classification criteria [103].

5.6.7 Breast cancer patient sera

Breast cancer patient serum samples were obtained from the Dana-Farber/ Harvard Cancer Center (DF/HCC) Breast SPORE Blood Bank. These samples were originally collected under Protocol #93-085 at the DF/HCC.

5.6.8 Phage immunoprecipitation

The T7-Pep library was prepared as described previously [26] and stored at -80 °C until used. For all samples, the final amount of Ig added to each 1 ml IP mix was approximately 2 µg. Serum/plasma samples were assumed to have 10 µg/ul of Ig, and so were diluted 10x in PBS before addition of 2 µl to the IP mix. If patient samples were derived from a different fluid compartment, their protein content was measured by Bradford assay and converted to an Ig concentration in the following way. For CSF the Ig fraction was assumed to be 29% of the total protein concentration. For synovial fluid, we used the following conversion: [Ig conc] = $0.154 \times$ [total protein conc] + 0.098. Sample dilutions were performed in a 96 well polystyrene PCR plate that had been blocked overnight with 1% fraction V or agarose purified BSA (Invitrogen) in PBS to minimize the amount of Ig lost to nonspecific binding of the polystyrene plate.

Each 1 ml IP mix contained 5x1010 T7-Pep phage particles and 2 ng of Positive control SAPK4 C-19 antibody (Santa Cruz, sc-7585) diluted in M9LB (for 1L: 46.7 ml 20X M9 salts, 18.7 ml 20% glucose (filtered), 0.93 ml 1 M MgSO4, 934 ml LB) with 100 µg/ml ampicillin. 1 ml IP mixes were placed in each well of a 96 deep well plate (Cole-Parmer, EW-07904-04). At this point, each patient sample or control was randomly assigned to a position on the IP plate and the appropriate volume for 2 µg of Ig was added to each IP. The plate was then carefully sealed with adhesive optical tape (Applied Biosystems) and placed on a rotator for 20 hours, mixing at 4 °C.

The plate was briefly centrifuged to collect volume. 40 µl of 1:1 Protein A / Protein G slurry (Invitrogen, 100-02D, 100-04D) was added to each well. The re-sealed plate was then placed on rotator for 4 hours at 4 °C.

The plate was briefly centrifuged. At this point the beads were subjected to an automated IP protocol, which was carried out on a BioMek FX liquid handling robot. Briefly, IPs were

washed in 440 µl IP Wash Buffer (150 mM NaCl, 50 mM Tris-HCL, 0.1% NP-40, pH 7.5) by pipetting up and down 30 times, for a total of 3 washes. Wash buffer was removed after magnetic separation on a 96 well magnet. Beads were moved to a new, clean plate after the second wash. After the final wash, IPs were resuspended in 40 µl of pure water and transferred to a new polystyrene PCR plate. This plate was heated to 95 °C for 10 minutes and then frozen at -80 °C until next step.

5.6.9 Preparation of immunoprecipitated T7-Pep sequencing libraries

Primers used (underlined sequences anneal with initial template, x's are the index barcode):
PCR1 forward: "IS7_HsORF5_2"
ACACTCTTTCCCTACACGACTCCAGTCAGGTGTGATGCTC
PCR1 reverse: "IS8_HsORF3_2"
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCGAGCTTATCGTCGTCATCC
PCR2 forward: "IS4_HsORF5_2"
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACTCCAGT
PCR2 reverse: "index N" (set of 96)
CAAGCAGAAGACGGCATACGAGATxxxxxxGTGACTGGAGTTCAGACGTGT
P5_Primer:
AATGATACGGCGACCACCGA
P7_Primer_2:
CAAGCAGAAGACGGCATACGA
Internal HsORF3' "TaqMan" FAM Probe:
GCCGCAAGCTTGTCGAGCGATG (modified with 5' 6-FAM-ZEN-3' Iowa Black FQ)
T7-Pep Library Sequencing Primer "T7-Pep_96_SP":
GCTCGGGGATCCAGGAATTCCGCTGCGC
Standard Illumina Multiplex Index Sequencing Primer "Index SP":
GATCGGAAGAGCACACGTCTGAACTCCAGTCAC

We tested the sensitivity of several DNA polymerases to residual NP-40 detergent from the wash buffer. Some of these enzymes performed poorly in the presence of this contaminant. We found the Herculase II Fusion DNA Polymerase (Agilent) to perform the most efficiently under all conditions, and so developed the following PCR protocol to recover IP'ed T7-Pep libraries. For each 50 µl PCR1 reaction, the following components were mixed with 30 µl from each IP: 8.75 µl water, 10 µl 5x Herculase Buffer, 0.5 µl of 100 mM dNTP, 0.125 µl of 100 uM IS7_HsORF5_2 forward primer, 0.125 µl of 100 uM IS8_HsORF3_2 reverse primer, and 0.5 µl of Herculase II enzyme. The reaction was then brought to 95 °C for 2 min, and cycled 30

times with the following thermal profile. 1. 95 °C, 20s 2. 58 °C, 30s 3. 72 °C, 30s and then subjected to a final extension for 3 min at 72 °C.

A set of 96, 7 nucleotide barcode-containing primers for PCR2 were designed using the method of Meyer et al. [89] to 1) be compatible with standard Illumina multiplex sequencing, 2) be base-balanced to maximize Illumina cluster definition, and 3) have no fewer than 3 nucleotide differences between them to minimize misalignment. [89] This set of oligos was purchased from Invitrogen in 10 µl 25 uM aliquots and then diluted to a final concentration of 2.5 uM by adding 90 µl of water.

For each 50 µl PCR2 reaction, the following components were mixed with 5 µl of the appropriate index primer and 1.5 µl of unpurified PCR1 product: 27.9 µl water, 10 µl 5x Herculase Buffer, 0.5 µl of 100 mM dNTP, 0.125 µl of 100 uM IS4_HsORF5_2 forward primer, and 0.5 µl of Herculase II enzyme. The reaction was then brought to 95 °C for 2 min, and then cycled 10 times with the following thermal profile. 1. 95 °C, 20s 2. 58 °C, 30s 3. 72 °C, 30s and then subjected to a final extension for 3 min at 72 °C.

Unpurified PCR2 product was next quantified using real time quantitative PCR on a 7500 Fast PCR-System (Applied Biosystems). Each PCR2 product was serially diluted 100 fold to a final 10,000x dilution in water. 4 µl of this dilution was added to 16 µl of master mix composed of: 4 µl water, 10 µl Universal TaqMan 2X PCR Master Mix (Applied Biosystems, PO4475), and 2 µl of a P5/FAM Probe/P7_2 mix (5 uM P5, 5 uM P7_2, and 2.5 uM FAM Probe). The thermal profile was: 1. 50 °C, 2m 2. 95 °C, 10m 3. 95 °C, 15s 4. 60 °C, 2m and steps 3 and 4 were repeated 35 times. We estimated the DNA concentration (in ng/ul) by $[Conc] = 5000 * 10^{(C_t - 3.0964)/-4.5781}$. 300 ng of each PCR2 product were then combined in a single tube, mixed, and run on a 2% agarose gel. The dominant band at 316 bp was cut out and column purified twice (QIAGEN).

This 96-plex pooled library was sequenced on 2 or 3 lanes of an Illumina HiSeq 2000 using 93+7 single end cycles (93 cycles from the "T7-Pep_96_SP" primer, and 7 cycles from the "Index SP" primer) to obtain between 300 and 450 million reads.

5.6.10 PhIP-Seq informatics pipeline

We developed an informatics pipeline for processing the single end, 100 nucleotide sequencing data generated from high throughput PhIP-Seq experiments. Unless otherwise noted, scripts were written in python, and are available online for download from: https://github.com/laserson/phip-stat This pipeline was implemented on Harvard Medical School's Orchestra Shared Research Cluster. The pipeline assumes that the initial data set is a single .fastq file (not "de-multiplexed") and that the barcode is in the header of each read. If reads have been de-multiplexed one can skip fastq2parts.py and proceed to bowtie_parts_with_LSF.py. Note that these commands are

for dispatch to the Platform LSF job scheduler.

The count data for each IP was then analyzed one sample at a time by comparison to the counts obtained by sequencing the un-enriched T7-Pep library. We used our generalized Poisson significance assignment algorithm [26] to compute -log10 P-values for each peptide/sample pair. Briefly, the IP count distribution for each input count was fitted to a generalized Poisson (GP) distribution. The two GP parameters, λ and θ were then regressed to form a joint distribution between the IP counts and the GP parameters such that each IP count could be evaluated for its likelihood of enrichment.

5.6.11 Analysis of high-throughput PhIP-Seq enrichment data

All computational analysis was performed in MATLAB software (MathWorks). Reproducibility between each replica pair was assessed as follows. Scatter plots of the log10 of the -log10 P-values were generated, and a sliding window of width 0.05 was moved in steps of 0.05 from -2 to 3 across the x-axis. The mean and standard deviation of the values within this window were calculated at each step and plotted as a function of -log10 P-values (see Figure 5.1A for example). For all such plots, at low -log10 P-values the standard deviation is larger than the mean. At high -log10 P-values, however, the reverse is true. For each pair, we determined the -log10 P-value at which the mean was equal to the standard deviation (analogous to the "signal" being equal to the "noise"). A histogram plot of these values are given as Figure 5.1B. Based on this data, we chose a -log10 P-value of 4 to be our cutoff for considering a peptide to be significantly enriched. Within each 96-well plate screened, several samples were run in duplicate so that the reproducibility of each run's automated IPs could be assessed. We found that occasionally, sequences from random clones were amplified dramatically only in one of the replicas. The cause of these potential false positives is under investigation, but they seemed to follow no particular pattern so did not contribute to disease association of enriched clones. They are unlikely to be due purely to spurious PCR amplification, as the same clones were amplified from the same wells with two independent PCR reactions using two different enzymes.

For analyses of peptide/ORF-disease association, we set all -log10 P-values less than 4 equal to 0, and -log10 P-values greater than 4 equal to 1. This allowed us to sum the "hits" for each peptide/ORF in each disease category and then to compute the P value for association using Fisher's exact test. To correct for multiple hypothesis testing, we performed a permutation analysis by randomly permuting the sample names and then calculating the "null" Fisher P-values for each peptide/ORF. This was repeated 1000 times and a final histogram of null Fisher P-values was constructed. Finally, an "expected" Fisher P-value distribution could be calculated for each P-value by summing the null distribution from each P-value to infinity. This expected distribution indicates how many peptide/ORF associations with a P-value at least as extreme,

would be expected by chance alone, given the same dataset with randomly permuted sample names. We corrected for bias due to differences in the total number of hits between samples by requiring that the difference in total number of hits after permutation is less than 3% compared to before permutation. To find the 10% false discovery rate threshold, we compare the expected Fisher P-values at each P-value to the sum of the observed Fisher P-values that are at least that extreme. The P-value at which this ratio is 1:10 is then set as the 10% FDR threshold.

5.6.12 ELISA testing of CSF samples

High binding capacity streptavidin-coated 96-well ELISA plates (Pierce, USA) were coated with biotin-Krt75_1 or biotin-scrambled peptide at 5μg/mL in Tris-buffer saline with 0.05% Tween plus 0.1% bovine serum albumin (TBST-BSA), pH 7.2, for 2 hours at room temperature. After three washes with TBST-BSA, CSF samples were normalized to 5 μg/mL IgG and then incubated in the wells with gentle agitation for 1 hour. Wells were washed three times with TBST-BSA. Secondary goat anti-human HRP (Chemicon, USA) was prepared at 1:20,000 and incubated in the wells for one hour with gentle agitation. After three washes with TBST-BSA, 50 uL of One-Step Ultra TMB ELISA developing reagent (Thermo Scientific, USA) was added to each well and allowed to develop for 5 minutes. The reaction was stopped by addition of 50 μl of 1M sulfuric acid. The optical density of each well was measured at 455nm. Data is reported as fold difference from signal from Krt75_1 versus that from scrambled peptide. □

5.7 Acknowledgements

We would like to thank Stewart Rudnicki at the Institute of Chemistry and Cell Biology (ICCB) at Harvard Medical School for helping to automate PhIP-Seq. Paul I W de Bakker provided valuable statistical direction. We also thank the Dana-Farber/Harvard Cancer Center (DF/HCC) Specialized Programs of Research Excellence (SPORE) in breast cancer for providing valuable breast cancer patient sera and the Human Brain and Spinal Fluid Resource Centre (VA Greater Los Angeles Healthcare System, West Los Angeles Healthcare Center) for providing CSF. This work was supported in part by grants from the United States Department of Defense (W81XWH-10-1-0994 and W81XWH-04-1-0197), and a HITI/Helmsley Trust Pilot Grants in Type 1 Diabetes to S.J.E. N.L.S. is a fellow of the Susan G. Komen for the Cure Foundation. L.Q. receives support from Fondo de Investigaciones Sanitarias (CM 09/00017), Carlos III Institute of Health, Spain. K.C.O. receives support from the Nancy Davis Foundation. P.A.N. was supported in part by the Cogan Family Foundation. S.J.E. is an investigator with the Howard Hughes Medical Institute.

5.8 Author contributions

S.J.E. conceived and supervised the project. N.L.S. constructed the T7-Pep library. H.B.L. performed the screen and together with U.L. analyzed the data. U.L. developed the informatics pipeline that was used to analyze the data. P.A.N. and R.M.P. provided the RA samples. K.C.O. and P.L.D.J. provided MS and control samples. G.A.M. provided T1D samples and matched controls, and assisted with analysis of the T1D dataset. P.I.W.D.B. provided statistical support. This chapter was prepared by H.B.L. and edited by S.J.E.

Chapter 6

Conclusion and Future Directions

We outlined a vision for using high-throughput DNA sequencing as an assay for antibody-antigen interactions. In this thesis, I described several steps in this direction; however, in each case, we only generate data for half the picture: either the antibodies alone, or the antigens alone. In fact, in the case of the antibodies, we have not even generated information on the entire receptor, as we have only obtained a single chain of the antibody heterodimer at a time. Below I address some proposed methods for which we have performed preliminary work to address these limitations.

6.1 Methods for single-cell coupling of heavy and light chains[1]

In Chapter 2, we successfully developed methods to characterize the VH antibody repertoire of an individual human. However, knowledge of the heavy chains alone is not sufficient to truly characterize the repertoire, and more importantly, does not allow the reconstruction of the antibodies of interest. To rigorously confirm that certain clones are involved in immune responses or to discover new antibodies against antigens of interest, it is a requirement to successfully capture both the heavy and light chains of individual antibodies. Because of the lack of methods for capturing paired VH and VL chains in high-throughput, the best available protocols involve sorting single cells into individual wells and performing PCR for the heavy and light chains serially (e.g., [105]). However, even with automated liquid handling robots, typical throughputs are practically limited to 10^6 (at great expense). Another popular solution is to capture heavy and light chain repertoires separately, and associate them randomly with each other in expression vectors [106]. However, heavy and light chain pairing is likely far from randomly distributed,

[1] Adapted from thesis proposal from December 2009.

and so these methods do not provide accurate portrayals of the underlying repertoires. (Indeed, this is the approach we followed, to marginal success.)

The goal of this proposal is to develop a general method for capturing paired heavy and light chains in millions of single cells in a single-reaction format. These methods should allow the simultaneous manipulation of millions of cells in parallel, while keeping them isolated from each other to maintain the natural chain pairing. The overall experimental design can be split into two parts: the chain-linking biochemistry and the cell insulation method. Solutions for each part can mostly be chosen independently.

All of the biochemical methods proposed are ultimately based on PCR. We can choose between amplifying the target chains from the genomic DNA or perform RT-PCR from the expressed mRNAs. The former requires no reverse transcription step but has the risk of amplifying non-functional receptors, while the latter can benefit from higher copy numbers and should only capture functional, expressed receptors. The physical cross-linking can occur through multiple mechanisms. The first is standard splicing-by-overlap-extension PCR (SOE-PCR, or fusion PCR or crossover PCR), whereby two of the PCR primers have complementary sequences so that the two amplicons function as primers and they fuse to each other [107]. The main advantage of this method is that it has been used extensively, and the overlap sequence can be designed so that the fused construct is immediately in a usable scFv format. The next mechanism is similar to the SOE-PCR in that tags are incorporated into the PCR primers. In this case, the tags contain loxP sites, so that fusion will occur upon Cre-mediated recombination [108, 109]. Finally, in the case of emulsion methods (see below), the final option for biochemistry is to amplify both the heavy and light chains onto beads [6, 110]. One advantage is that the beads can be processed immediately for sequencing on bead-based next-generation sequencing systems. However, this is also a disadvantage, as it severely limits the range of options after chain coupling. One alternative bead-based method is to amplify both chains onto beads, and then couple the chains on the beads. This will increase the specificity of the whole process, albeit at increased complexity of the protocol. All these methods are summarized in Figure 6.1.

The cell insulation methods fall into two main categories: in-cell methods and emulsion methods. In-cell methods emulate the earliest attempt to couple two chains together, performed in Greg Winter's group at MRC [111]. In this method, the cells are fixed in formalin and permeabilized to allow the diffusion of biochemical reagents into the cell. The cell membrane functions as the barrier that prevents cross-contamination of heavy and light chains between cells. The advantages of this general approach are the relative simplicity of fixing the cells and also the ability to serially apply reagent sets to all cells in parallel. However, the permeabilization step is a two-edged sword, and potentially increases the chance that Ig chains will leak out of cells and lead to cross-contamination.

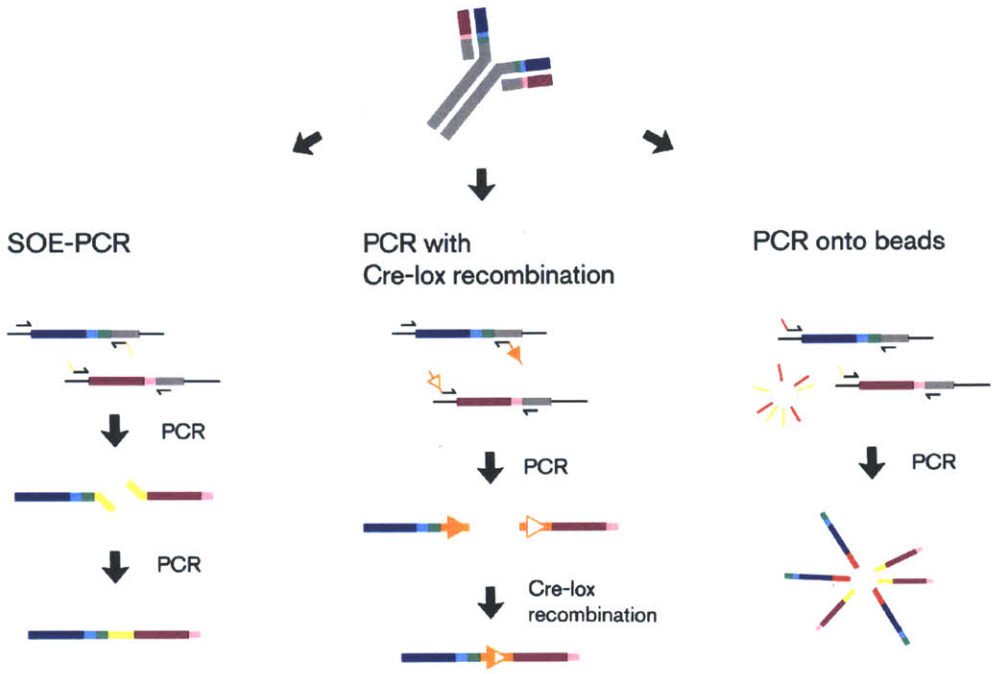


Figure 6.1: Chain coupling methods

For the emulsion-based methods, single cells are placed into individual compartments of a water-in-oil emulsion [112, 113]. The primary advantage of such an approach is that the oil-based separation of compartments should potentially provide nearly absolute insulation from chain cross-contamination. But while the oil-separated compartments should stop any exchange of material between compartments, a common problem of thermal cycling emulsions is that compartments fuse together, leading to non-clonality. Furthermore, it is considerably more difficult to manipulate emulsions. Emulsions are generally formed using physical methods (e.g., vortexing) that depend on Poisson statistics to achieve clonality [114, 115]. However, this tends to lead to a small fraction of non-clonal compartments, and also leads to a large number of unoccupied compartments. However, to combat these problems, other groups, e.g. David Weitz's group, have generated emulsions using microfluidic technology [112]. An additional disadvantage to using emulsion methods is that once an emulsion is formed, it is difficult to exchange additional material with the compartments in a controlled fashion. However, some companies are researching this exact probem, like RainDance technologies, which has developed a sophisticated technology for fusing emulsion droplets in a controlled fashion [116, 117]. Finally, emulsion PCR is often performed in conditions that are far from standardized protocols. In these unique conditions, the performance and fidelities of the relevant enzymes have not been well characterized. The cell insulation methods are summarized in Figure 6.2.

With these general considerations, we have identified six strategies to achieve coupling of

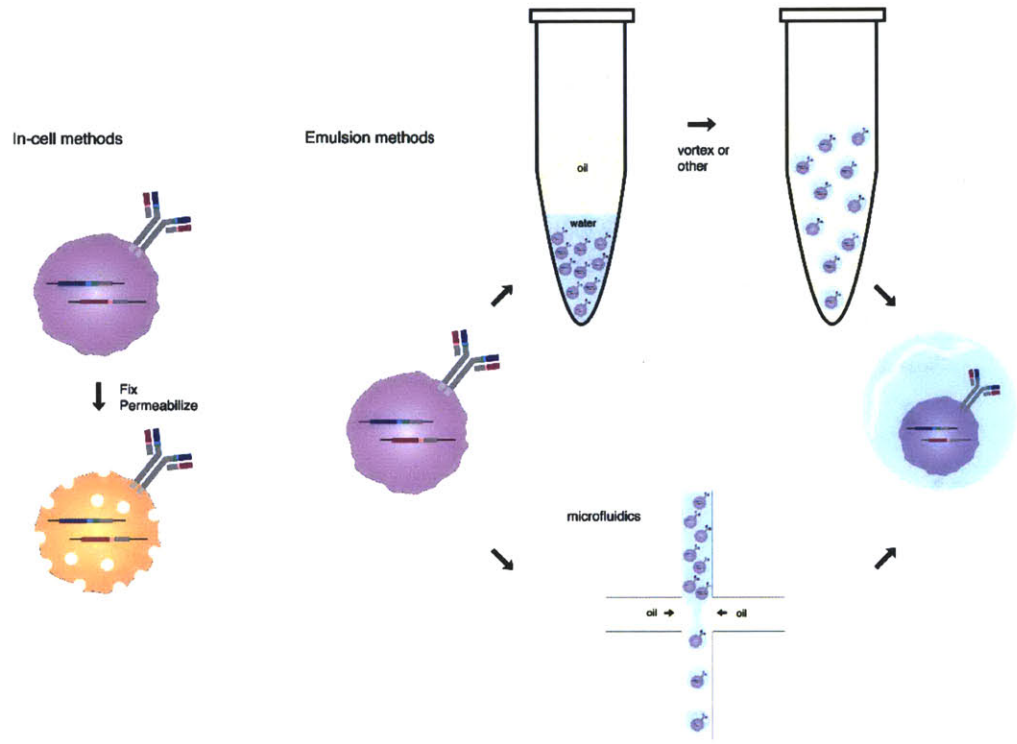


Figure 6.2: Cell insulation methods

heavy and light chains. They are displayed in Figure 6.3:

1. *Emulsion PCR from gDNA.* This is the simplest emulsion-based approach. The cells are placed in the emulsion along with reagents for a traditional PCR reaction. We then perform the SOE-PCR using the gDNA as a template.

2. *In-cell RT and SOE-PCR or Cre-Lox coupling.* Replicate the work of Embleton et al. [109, 111]. This involves fixing cells in formalin and permeabilizing them using one of several methods (e.g., proteinase K). Because all cells are in solution, it allows a traditional RT-PCR reaction by applying the relevant enzymes serially.

3. *Tth-mediated emulsion RT-PCR.* As described above, it is preferable to capture the Ig chains from the mRNA sequence, as this avoids any non-functional receptor rearrangements and also benefits from the potentially higher copy-numbers of expressed cells. However, emulsion PCR only allows us to add biochemical reagents once. This makes Tth polymerase, which is capable of performing both RT and PCR [118], quite attractive for an emulsion context. However, as it is already known that Tth has slightly lower fidelity compared with traditional polymerases such as Taq or Pfu, it remains to be seen whether this unique enzyme is a viable option.

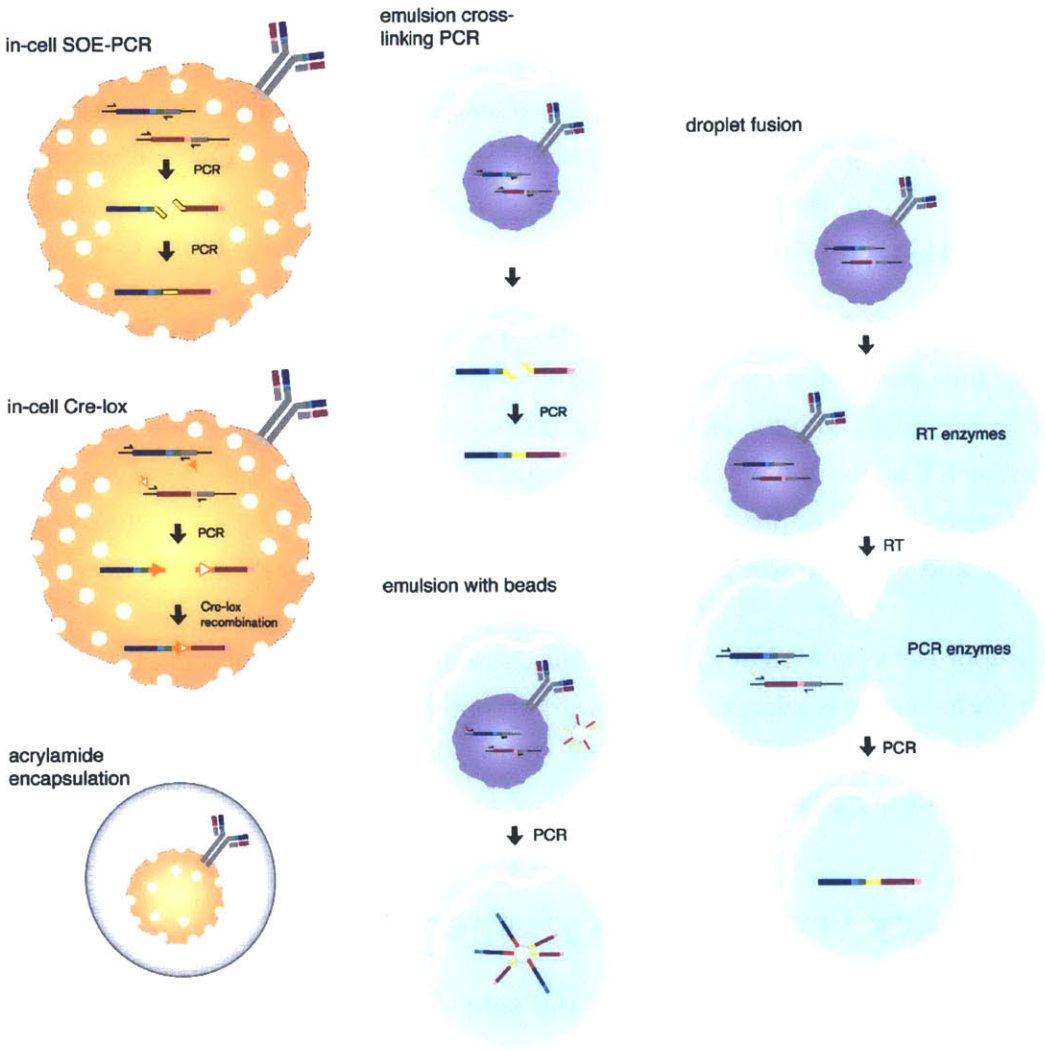


Figure 6.3: Summary of methods for coupling heavy and light chains

4. *Acrylimide-encapsulated in-cell RT-PCR.* This technique is similar to the in-cell RT-PCR, but involves the additional step of encapsulating individual cells in polyacrylimide gels [119]. This will add an additional layer of protection from cross-contamination.

5. *Emulsion PCR onto beads.* As an alternative to SOE-PCR for cross-linking the heavy and light chains, this approach attempts to capture the two chains by conventional PCR onto beads [110]. Similarly to many next-generation library preparation schemes, each emulsion compartment aims to have a single cell and a single bead. The beads are coated with two different primers: one for the heavy chain and one for the light chain. After breaking the emulsion, the beads can be manipulated in a variety of ways. One method to increase the specificity of the technique will be to cross-link the two chains on the beads using a modified Cre-Lox system.

6. *Emulsion RT-PCR with droplet fusion.* This approach attempts to utilize methods that allow us to fuse emulsion droplets in a controlled fashion. In this way, "bags" of enzymes can be serially fused with the emulsion compartments to perform separate biochemical steps, such as RT followed by PCR. However, this technology is still at the forefront of research and represents an approach that is less likely to be fruitful. Still, it would allow us to use robust and well characterized enzymes to separately perform RT and PCR.

These approaches hold significant potential due to the amount of collective experience available with both. There is a large body of work in optimizing RT-PCR reactions *in situ* for use in clinical pathology (e.g., [120]). Regarding emulsions, multiple labs have invested significant resources in performing droplet manipulation wizardry.

6.2 Library versus library experiments

Analogously to methods for linking heavy and light chains, here we would like to leverage a similar approach to characterize antigen-antibody interactions in high-throughput. Current technologies require selecting for new antibodies against a single antigen at a time [121]. A typical experiment involves purifying and immobilizing some antigen of interest, and exposing it to some type of protein display technology encoding a library of candidate antibody sequences (typically in scFv format). After multiple rounds of panning, washing, and amplifying, a small number of clones are sequenced and carried through for further analysis. Alternatively, an animal might be immunized to the antigen of interest to generate a polyclonal response. Antigen-specific lymphocytes must then be harvested (to obtain the polyclonal response) and screened to obtain high-affinity monoclonal antibodies [121]. Our proposed project aims at

developing methods that will allow the selection of new antibodies against multiple antigens in parallel or will allow for a "diagnostic" that can assay many antigens against many antibodies simultaneously.

To achieve these goals, we are proposing to develop similar methods to those proposed for coupling two chains of nucleic acids. Antibodies can be encoded in some protein display format (e.g., phage [122, 123], yeast [124–126], or ribosome display [127]) and antigens can also be packaged with their coding information (e.g., peptide libraries in display format, or whole virus particles). The two libraries are allowed to interact and selected for interacting complexes only [27]. The coding chains of the interacting antigen and antibody are then physically coupled using methods described for heavy/light chain linking. This should allow for the discovery of antigen-specific antibodies for multiple antigens in parallel. Ultimately, this type of approach could be scaled up to test huge antigen libraries against huge antigen libraries (e.g., entire human proteome, or all known viral proteins).

One advantage is that our protocols for capturing full antibodies will have already formatted the captured repertoires in an scFv format, allowing for easy expression using one of the protein display technologies. Compared with many previous studies that generate random antibody libraries, each antibody in our libraries were derived from a functioning, natural immune system. In this way, we hope to capitalize on millions of years of evolution to provide us with efficient antibody libraries that will allow us to quickly discover new functional antibodies.

More concretely, our primary approach for capturing antigen-antibody (Ag-Ab) interactions involves placing single Ag-Ab complexes into individual emulsion compartments. In order to properly display antibodies and antigens, there are several choices for each: phage, yeast, or ribosome display, and whole viral particles (in the case of viral antigens). Ribosome display offers the largest potential libraries and is fully *in vitro*. Phage display is the oldest method and provides large library sizes. Yeast display offers the smallest library sizes, but is particularly appropriate for antigen libraries as it can carry larger payloads and has glycosylation machinery [128]. Use of whole virus particles provides access to the most realistic antigens but has a potentially large genome. It is important to note that because we are cloning naturally expressed repertoires, we expect that the maximum library size of all of these technologies will be sufficient to capture the diversity of sequences in any practical blood samples or antigen sets. Phage-Ab against yeast-Ag has already shown promising results for a general approach to library-against-library selections [27]. The yeast is well suited to larger protein fragments, the two systems can replicate independently, and there are already protocols published. We also believe that ribosome-Ab against ribosome-Ag hold significant potential, as this system is entirely *in vitro*. This allow us to attempt coupling methods that do not depend on emulsions.

Phage-based systems appear to be more geared toward emulsion-based methods for cap-

turing interacting antibodies and antigens. Possibly after several rounds of affinity selection, phage-Ab-yeast-Ag duplexes can be "double-purified" using magnetic sorting or FACS to eliminate non-interacting particles [27, 28]. These duplexes can then be placed into individual emulsion compartments where some type of cross-linking PCR reaction will physically associate the Ag and Ab coding sequence. These cross-linked species can then be prepared for next-generation sequencing and the interactions determined by analyzing the sequencing data. In the case that the Ag used is whole virus particles, the genome sequence can be relatively large. However, if the library of viruses is relatively small, it is sufficient to find a unique "barcode" to identify the specific strain of virus.

For the fully *in vitro* ribosome display system, in addition to the emulsion based protocol, there are possibilities for non-emulsion methods. For example, proximity ligation assay [129] offer an opportunity to capture unique tags on both Ag and Ab libraries. These tags can be engineered to supply enough information to obtain the full corresponding sequences.

6.3 Analyzing HTS fitness experiments: an experiment in crowdsourcing

We realized that the experimental approach used in PhIP-seq was becoming applicable in many other contexts. At its core, the goal is to find a small subset of a large population that exhibits some sort of significant fitness advantage in relation to a particular assay. In the case of PhIP-seq, the library is every protein-coding peptide and the fitness is binding affinity to autoantibodies; in other cases, it may be a small RNA library with the fitness being growth rate or viral resistance. We decided to approach the problem more generally, and to attempt to develop the best statistical methodology that is both accurate, and efficient.

We developed a general model for this type of experimetal design, and encoded it as a probabilistic graphical model (Figure 6.4). We believe this particular type of model is common thanks to developments in HTS, but there do not exist adequate statistical methods for it. In addition to inferring a hidden underlying "fitness" value for each member of the population, the model allows for the observation of multiple time points in the selection experiment.

We implemented a Gibbs sampler to perform inference on this model; however, as this method does not scale well to very large populations, we initiated a collaboration with Harvard Catalyst to solve the problem through TopCoder. This is a platform for running algorithm competitions for client-contributed problems. Using our PGM, we generated many instances of the problem using multiple underlying fitness distributions and sampling depths (Figure 6.5). We developed a scoring model that optimizes for correctly determing the library members with

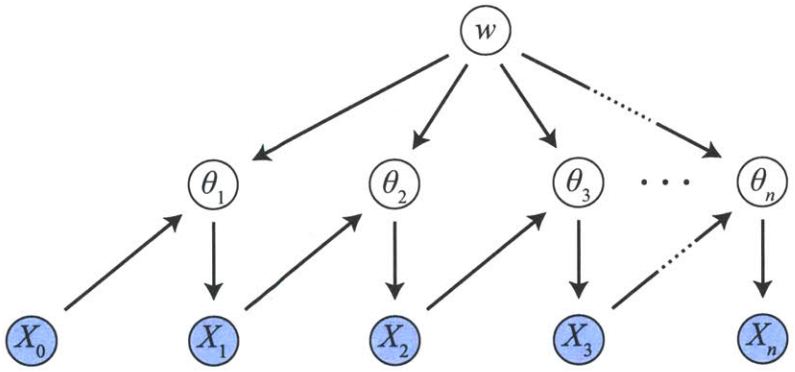


Figure 6.4: Bayesian network for fitness estimation. The only observed variables are the count vectors X_i, which are assumed to be multinomially distributed with probability vector θ_i. These vectors are Dirichlet distributed with information derived from the previous observed counts along with the underlying fitness values w, which are the values we are ultimately interested in.

the most extreme fitness values. (Effectively, this ranks the most promising library members for laboratory follow-up.)

We recently received the results of the competition, and are currently analyzing the different algorithms that were submitted.

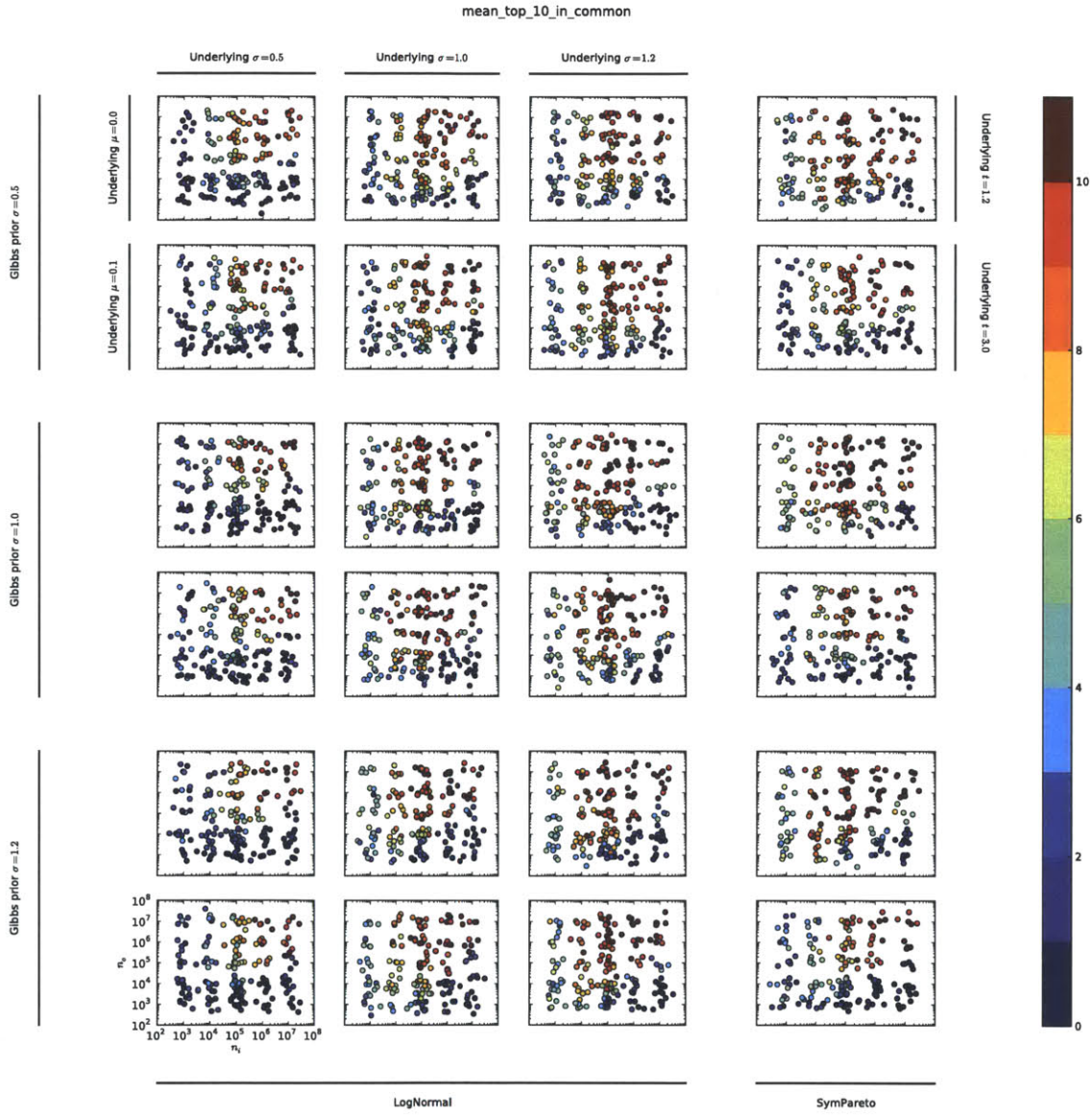


Figure 6.5: PGM model instantiations. Here we instantiate many instances of the PGM using different distributions, and determine the quality of solutions using our Gibbs sampling method.

References

[1] I R Cohen. *Tending Adam's Garden: Evolving the Cognitive Immune Self - Irun R. Cohen - Google Books*. 2000.

[2] George M Church, Yuan Gao, and Sriram Kosuri. Next-Generation Digital Information Storage in DNA. *Science*, pages –, August 2012.

[3] J L JL Xu and M M MM Davis. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity*, 13(1):37–45, July 2000.

[4] R A Lerner, C F Barbas, A S Kang, and D R Burton. On the use of combinatorial antibody libraries to clone the "fossil record" of an individual's immune response. *Proceedings of the National Academy of Sciences of the United States of America*, 88(21):9705–9706, November 1991.

[5] Marcel M Margulies, Michael M Egholm, William E WE Altman, Said S Attiya, Joel S JS Bader, Lisa A LA Bemben, Jan J Berka, Michael S MS Braverman, Yi-Ju YJ Chen, Zhoutao Z Chen, Scott B SB Dewell, Lei L Du, Joseph M JM Fierro, Xavier V XV Gomes, Brian C BC Godwin, Wen W He, Scott S Helgesen, Chun Heen CH Ho, Chun He CH Ho, Gerard P GP Irzyk, Szilveszter C SC Jando, Maria L I ML Alenquer, Thomas P TP Jarvie, Kshama B KB Jirage, Jong-Bum JB Kim, James R JR Knight, Janna R JR Lanza, John H JH Leamon, Steven M SM Lefkowitz, Ming M Lei, Jing J Li, Kenton L KL Lohman, Hong H Lu, Vinod B VB Makhijani, Keith E KE McDade, Michael P MP McKenna, Eugene W EW Myers, Elizabeth E Nickerson, John R JR Nobile, Ramona R Plant, Bernard P BP Puc, Michael T MT Ronan, George T GT Roth, Gary J GJ Sarkis, Jan Fredrik JF Simons, John W JW Simpson, Maithreyan M Srinivasan, Karrie R KR Tartaro, Alexander A Tomasz, Kari A KA Vogt, Greg A GA Volkmer, Shally H SH Wang, Yong Y Wang, Michael P MP Weiner, Pengguang P Yu, Richard F RF Begley, and Jonathan M JM Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, September 2005.

[6] Jay Shendure, Gregory J Porreca, Nikos B Reppas, Xiaoxia Lin, John P McCutcheon, Abraham M Rosenbaum, Michael D Wang, Kun Zhang, Robi D Mitra, and George M Church. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309(5741):1728–1732, September 2005.

[7] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature biotechnology*, 26 (10):1135–1145, October 2008.

[8] Peter A Carr and George M Church. Genome engineering. *Nature biotechnology*, 27(12): 1151–1162, December 2009.

[9] Andrzej Drukier, Katherine Freese, David Spergel, Charles Cantor, George Church, and Takeshi Sano. New Dark Matter Detectors using DNA for Nanometer Tracking. *arXiv.org*, astro-ph.IM, June 2012.

[10] Harlan Robins, Cindy Desmarais, Jessica Matthis, Robert Livingston, Jessica Andriesen, Helena Reijonen, Christopher Carlson, Gerold Nepom, Cassian Yee, and Karen Cerosaletti. Ultra-sensitive detection of rare T cell clones. *Journal of immunological methods*, 375 (1-2):14–19, January 2012.

[11] Joshua A JA Weinstein, Ning N Jiang, Richard A RA White, Daniel S DS Fisher, and Stephen R SR Quake. High-throughput sequencing of the zebrafish antibody repertoire. *Science*, 324(5928):807–810, May 2009.

[12] Aaron C AC Logan, Hong H Gao, Chunlin C Wang, Bita B Sahaf, Carol D CD Jones, Eleanor L EL Marshall, Ismael I Buño, Randall R Armstrong, Andrew Z AZ Fire, Kenneth I KI Weinberg, Michael M Mindrinos, James L JL Zehnder, Scott D SD Boyd, Wenzhong W Xiao, Ronald W RW Davis, and David B DB Miklos. High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. *Proceedings of the National Academy of Sciences of the United States of America*, 108(52):21194–21199, December 2011.

[13] Holden T HT Maecker, Tamsin M TM Lindstrom, William H WH Robinson, Paul J PJ Utz, Matthew M Hale, Scott D SD Boyd, Shai S SS Shen-Orr, and C Garrison CG Fathman. New tools for classification and monitoring of autoimmune diseases. *Nature Reviews: Rheumatology*, 8(6):317–328, January 2012.

[14] Scott D Boyd, Eleanor L Marshall, Jason D Merker, Jay M Maniar, Lyndon N Zhang, Bita Sahaf, Carol D Jones, Birgitte B Simen, Bozena Hanczaruk, Khoa D Nguyen, Kari C Nadeau, Michael Egholm, David B Miklos, James L Zehnder, and Andrew Z Fire. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Science Translational Medicine*, 1(12):12ra23–12ra23, December 2009.

[15] René L Warren, J Douglas Freeman, Thomas Zeng, Gina Choe, Sarah Munro, Richard Moore, John R Webb, and Robert A Holt. 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(5):790–797, May 2011.

[16] Yu-Chang Wu, David Kipling, Hui Sun Leong, Victoria Martin, Alexander A Ademokun, and Deborah K Dunn-Walters. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood*, 116(7):1070–1078, August 2010.

[17] H P Brezinschek, R I Brezinschek, and P E Lipsky. Analysis of the heavy chain repertoire of human peripheral B cells using single-cell polymerase chain reaction. *Journal of immunology (Baltimore, Md. : 1950)*, 155(1):190–202, July 1995.

[18] G Yaari, M Uduman, and S H Kleinstein. Quantifying selection in high-throughput Immunoglobulin sequencing data sets. *Nucleic acids research*, pages –, May 2012.

[19] J Douglas Freeman, René L Warren, John R Webb, Brad H Nelson, and Robert A Holt. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Research*, 19(10):1817–1824, October 2009.

[20] Véronique Giudicelli, Patrice Duroux, Chantal Ginestoux, Géraldine Folch, Joumana Jabado-Michaloud, Denys Chaume, and Marie-Paule Lefranc. IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic acids research*, 34(Database issue):D781–D784, January 2006.

[21] M M MM Davis and P J PJ Bjorkman. T-cell antigen receptor genes and T-cell recognition. *Nature*, 334(6181):395–402, August 1988.

[22] Jens Wrammert, Kenneth Smith, Joe Miller, William A Langley, Kenneth Kokko, Christian Larsen, Nai-Ying Zheng, Israel Mays, Lori Garman, Christina Helms, Judith James, Gillian M Air, J Donald Capra, Rafi Ahmed, and Patrick C Wilson. Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nature*, 453(7195): 667–671, May 2008.

[23] Jonathan Laserson. *Bayesian Assembly of Reads From High Throughput Sequencing*. PhD thesis, Stanford University, October 2011.

[24] Quoc V Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S Corrado, Jeff Dean, and Andrew Y Ng. Building high-level features using large scale unsupervised learning. *arXiv.org*, cs.LG, December 2011.

[25] Sriram Kosuri, Nikolai Eroshenko, Emily M Leproust, Michael Super, Jeffrey Way, Jin Billy Li, and George M Church. Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nature biotechnology*, 28(12):1295–1299, November 2010.

[26] H Benjamin Larman, Zhenming Zhao, Uri Laserson, Mamie Z Li, Alberto Ciccia, M Angelica Martinez Gakidis, George M Church, Santosh Kesari, Emily M Leproust, Nicole L Solimini, and Stephen J Elledge. Autoantigen discovery with a synthetic human peptidome. *Nature biotechnology*, 29(6):535–541, 2011.

[27] D R Bowley, T M Jones, D R Burton, and R A Lerner. Libraries against libraries for combinatorial selection of replicating antigen-antibody pairs. *Proceedings of the National Academy of Sciences of the United States of America*, 106(5):1380–1385, February 2009.

[28] Saurabh R Nirantar and Farid J Ghadessy. Compartmentalized linkage of genes encoding interacting protein pairs. *Proteomics*, 11(7):1335–1339, April 2011.

[29] G M Church. The Personal Genome Project. *Molecular systems biology*, 1(1):–, December 2005.

[30] Véronique V Giudicelli, Denys D Chaume, and Marie-Paule MP Lefranc. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic acids research*, 33(Database issue):D256–D261, January 2005.

[31] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, May 1998.

[32] Xavier Brochet, Marie-Paule Lefranc, and Véronique Giudicelli. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic acids research*, 36(Web Server issue):W503–8, July 2008.

[33] Marie-Paule Lefranc, Véronique Giudicelli, Chantal Ginestoux, Joumana Jabado-Michaloud, Géraldine Folch, Fatena Bellahcene, Yan Wu, Elodie Gemrot, Xavier Brochet, Jérôme Lane, Laetitia Regnier, François Ehrenmann, Gérard Lefranc, and Patrice Duroux. IMGT, the international ImMunoGeneTics information system. *Nucleic acids research*, 37(Database issue):D1006–12, January 2009.

[34] Mohamed Uduman, Gur Yaari, Uri Hershberg, Jacob A Stern, Mark J Shlomchik, and Steven H Kleinstein. Detecting selection in immunoglobulin sequences. *Nucleic acids research*, 39(Web Server issue):W499–504, July 2011.

[35] Barbara S BS Taylor, Magdalena E ME Sobieszczyk, Francine E FE McCutchan, and Scott M SM Hammer. The challenge of HIV-1 subtype diversity. *The New England journal of medicine*, 358(15):1590–1602, April 2008.

[36] Dennis R Burton, Robyn L Stanfield, and Ian A Wilson. Antibody vs. HIV in a clash of evolutionary titans. *Proceedings of the National Academy of Sciences of the United States of America*, 102(42):14943–14948, October 2005.

[37] Laura M Walker, Michael Huber, Katie J Doores, Emilia Falkowska, Robert Pejchal, Jean-Philippe Julien, Sheng-Kai Wang, Alejandra Ramos, Po-Ying Chan-Hui, Matthew Moyle, Jennifer L Mitcham, Phillip W Hammond, Ole A Olsen, Pham Phung, Steven Fling, Chi-Huey Wong, Sanjay Phogat, Terri Wrin, Melissa D Simek, Protocol G Principal Investigators, Wayne C Koff, Ian A Wilson, Dennis R Burton, and Pascal Poignard. Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature*, 477 (7365):466–470, August 2011.

[38] Laura M Walker, Sanjay K Phogat, Po-Ying Chan-Hui, Denise Wagner, Pham Phung, Julie L Goss, Terri Wrin, Melissa D Simek, Steven Fling, Jennifer L Mitcham, Jennifer K Lehrman, Frances H Priddy, Ole A Olsen, Steven M Frey, Phillip W Hammond, Protocol G Principal Investigators, Stephen Kaminsky, Timothy Zamb, Matthew Moyle, Wayne C Koff, Pascal Poignard, and Dennis R Burton. Broad and potent neutralizing antibodies

from an African donor reveal a new HIV-1 vaccine target. *Science*, 326(5950):285–289, October 2009.

[39] Xueling Wu, Zhi-Yong Yang, Yuxing Li, Carl-Magnus Hogerkorp, William R Schief, Michael S Seaman, Tongqing Zhou, Stephen D Schmidt, Lan Wu, Ling Xu, Nancy S Longo, Krisha McKee, Sijy O'Dell, Mark K Louder, Diane L Wycuff, Yu Feng, Martha Nason, Nicole Doria-Rose, Mark Connors, Peter D Kwong, Mario Roederer, Richard T Wyatt, Gary J Nabel, and John R Mascola. Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science*, 329(5993):856–861, August 2010.

[40] Leonidas Stamatatos, Lynn Morris, Dennis R Burton, and John R Mascola. Neutralizing antibodies generated during natural HIV-1 infection: good news for an HIV-1 vaccine? *Nature medicine*, 15(8):866–870, August 2009.

[41] E S Gray, M C Madiga, T Hermanus, P L Moore, C K Wibmer, N L Tumba, L Werner, K Mlisana, S Sibeko, C Williamson, S S Abdool Karim, and L Morris. The Neutralization Breadth of HIV-1 Develops Incrementally over Four Years and Is Associated with CD4+ T Cell Decline and High Viral Load during Acute Infection. *Journal of Virology*, 85(10):4828–4840, April 2011.

[42] Tongqing T Zhou, Ivelin I Georgiev, Xueling X Wu, Zhi-Yong ZY Yang, Kaifan K Dai, Andrés A Finzi, Young Do YD Kwon, Johannes F JF Scheid, Wei W Shi, Ling L Xu, Yongping Y Yang, Jiang J Zhu, Michel C MC Nussenzweig, Joseph J Sodroski, Lawrence L Shapiro, Gary J GJ Nabel, John R JR Mascola, and Peter D PD Kwong. Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. *Science*, 329(5993):811–817, August 2010.

[43] Xueling Wu, Tongqing Zhou, Jiang Zhu, Baoshan Zhang, Ivelin Georgiev, Charlene Wang, Xuejun Chen, Nancy S Longo, Mark Louder, Krisha McKee, Sijy O'Dell, Stephen Perfetto, Stephen D Schmidt, Wei Shi, Lan Wu, Yongping Yang, Zhi-Yong Yang, Zhongjia Yang, Zhenhai Zhang, Mattia Bonsignori, John A Crump, Saidi H Kapiga, Noel E Sam, Barton F Haynes, Melissa Simek, Dennis R Burton, Wayne C Koff, Nicole A Doria-Rose, Mark Connors, NISC Comparative Sequencing Program, James C Mullikin, Gary J Nabel, Mario Roederer, Lawrence Shapiro, Peter D Kwong, and John R Mascola. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science*, 333(6049):1593–1602, September 2011.

[44] Johannes F Scheid, Hugo Mouquet, Niklas Feldhahn, Michael S Seaman, Klara Velinzon, John Pietzsch, Rene G Ott, Robert M Anthony, Henry Zebroski, Arlene Hurley, Adhuna Phogat, Bimal Chakrabarti, Yuxing Li, Mark Connors, Florencia Pereyra, Bruce D Walker, Hedda Wardemann, David Ho, Richard T Wyatt, John R Mascola, Jeffrey V Ravetch, and Michel C Nussenzweig. Broad diversity of neutralizing antibodies isolated from memory B cells in HIV-infected individuals. *Nature*, 458(7238):636–640, April 2009.

[45] Christopher C Sundling, Yuxing Y Li, Nick N Huynh, Christian C Poulsen, Richard R Wilson, Sijy S O'Dell, Yu Y Feng, John R JR Mascola, Richard T RT Wyatt, and Gunilla B GB Karlsson Hedestam. High-Resolution Definition of Vaccine-Elicited B Cell

Responses Against the HIV Primary Receptor Binding Site. *Science Translational Medicine*, 4(142):142ra96–142ra96, July 2012.

[46] Chih-Jen Wei, Jeffrey C Boyington, Patrick M McTamney, Wing-Pui Kong, Melissa B Pearce, Ling Xu, Hanne Andersen, Srinivas Rao, Terrence M Tumpey, Zhi-Yong Yang, and Gary J Nabel. Induction of broadly neutralizing H1N1 influenza antibodies by vaccination. *Science*, 329(5995):1060–1064, August 2010.

[47] M Anthony MA Moody, Nicole L NL Yates, Joshua D JD Amos, Mark S MS Drinker, Joshua A JA Eudailey, Thaddeus C TC Gurley, Dawn J DJ Marshall, John F JF Whitesides, Xi X Chen, Andrew A Foulger, Jae-Sung JS Yu, Ruijun R Zhang, R Ryan RR Meyerhoff, Robert R Parks, Julia Cavanaugh JC Scull, Lu L Wang, Nathan A NA Vandergrift, Joy J Pickeral, Justin J Pollara, Garnett G Kelsoe, S Munir SM Alam, Guido G Ferrari, David C DC Montefiori, Gerald G Voss, Hua-Xin HX Liao, Georgia D GD Tomaras, and Barton F BF Haynes. HIV-1 gp120 vaccine induces affinity maturation in both new and persistent antibody clonal lineages. *Journal of Virology*, 86(14):7496–7507, July 2012.

[48] Xiaocong X Yu, Tshidi T Tsibane, Patricia A PA McGraw, Frances S FS House, Christopher J CJ Keefer, Mark D MD Hicar, Terrence M TM Tumpey, Claudia C Pappas, Lucy A LA Perrone, Osvaldo O Martinez, James J Stevens, Ian A IA Wilson, Patricia V PV Aguilar, Eric L EL Altschuler, Christopher F CF Basler, and James E JE Crowe. Neutralizing antibodies derived from the B cells of 1918 influenza pandemic survivors. *Nature*, 455(7212):532–536, September 2008.

[49] Damian C DC Ekiert, Robert H E RH Friesen, Gira G Bhabha, Ted T Kwaks, Mandy M Jongeneelen, Wenli W Yu, Carla C Ophorst, Freek F Cox, Hans J W M HJ Korse, Boerries B Brandenburg, Ronald R Vogels, Just P J JP Brakenhoff, Ronald R Kompier, Martin H MH Koldijk, Lisette A H M LA Cornelissen, Leo L M LL Poon, Malik M Peiris, Wouter W Koudstaal, Ian A IA Wilson, and Jaap J Goudsmit. A highly conserved neutralizing epitope on group 2 influenza A viruses. *Science*, 333(6044):843–850, August 2011.

[50] C F CF Barbas, D D Hu, N N Dunlop, L L Sawyer, D D Cababa, R M RM Hendry, P L PL Nara, and D R DR Burton. In vitro evolution of a neutralizing human antibody to human immunodeficiency virus type 1 to enhance affinity and broaden strain cross-reactivity. *Proceedings of the National Academy of Sciences of the United States of America*, 91(9): 3809–3813, April 1994.

[51] G Stiegler, R Kunert, M Purtscher, S Wolbank, R Voglauer, F Steindl, and H Katinger. A potent cross-clade neutralizing human monoclonal antibody against a novel epitope on gp41 of human immunodeficiency virus type 1. *AIDS research and human retroviruses*, 17 (18):1757–1765, December 2001.

[52] M B Zwick, A F Labrijn, M Wang, C Spenlehauer, E O Saphire, J M Binley, J P Moore, G Stiegler, H Katinger, D R Burton, and P W Parren. Broadly neutralizing antibodies targeted to the membrane-proximal external region of human immunodeficiency virus type 1 glycoprotein gp41. *Journal of Virology*, 75(22):10892–10905, November 2001.

[53] Daniel A DA Calarese, Christopher N CN Scanlan, Michael B MB Zwick, Songpon S Deechongkit, Yusuke Y Mimura, Renate R Kunert, Ping P Zhu, Mark R MR Wormald, Robyn L RL Stanfield, Kenneth H KH Roux, Jeffery W JW Kelly, Pauline M PM Rudd, Raymond A RA Dwek, Hermann H Katinger, Dennis R DR Burton, and Ian A IA Wilson. Antibody domain exchange is an immunological solution to carbohydrate cluster recognition. *Science*, 300(5628):2065–2071, June 2003.

[54] R Pejchal, K J Doores, L M Walker, R Khayat, P-S Huang, S-K Wang, R L Stanfield, J-P Julien, A Ramos, M Crispin, R Depetris, U Katpally, A Marozsan, A Cupo, S Maloveste, Y Liu, R McBride, Y Ito, R W Sanders, C Ogohara, J C Paulson, T Feizi, C N Scanlan, C-H Wong, J P Moore, W C Olson, A B Ward, P Poignard, W R Schief, D R Burton, and I A Wilson. A Potent and Broad Neutralizing Antibody Recognizes and Penetrates the HIV Glycan Shield. *Science*, 334(6059):1097–1103, October 2011.

[55] Susan Moir and Anthony S Fauci. B cells in HIV infection and disease. *Nature Reviews: Immunology*, 9(4):235–245, April 2009.

[56] Andrea L Graham, Adam D Hayward, Kathryn A Watt, Jill G Pilkington, Josephine M Pemberton, and Daniel H Nussey. Fitness correlates of heritable variation in antibody responsiveness in a wild mammal. *Science*, 330(6004):662–665, October 2010.

[57] Peggy Ho PH Faix, Michael A MA Burg, Michelle M Gonzales, Edward P EP Ravey, Andrew A Baird, and David D Larocca. Phage display of cDNA libraries: enrichment of cDNA expression using open reading frame selection. *BioTechniques*, 36(6):1018–1019, June 2004.

[58] Matthew L Albert and Robert B Darnell. Paraneoplastic neurological degenerations: keys to tumour immunity. *Nature reviews. Cancer*, 4(1):36–44, January 2004.

[59] Xiaoju Wang, Jianjun Yu, Arun Sreekumar, Sooryanarayana Varambally, Ronglai Shen, Donald Giacherio, Rohit Mehra, James E Montie, Kenneth J Pienta, Martin G Sanda, Philip W Kantoff, Mark A Rubin, John T Wei, Debashis Ghosh, and Arul M Chinnaiyan. Autoantibody signatures in prostate cancer. *The New England journal of medicine*, 353(12): 1224–1235, September 2005.

[60] Karen S Anderson, Sahar Sibani, Garrick Wallstrom, Ji Qiu, Eliseo A Mendoza, Jacob Raphael, Eugenie Hainsworth, Wagner R Montor, Jessica Wong, Jin G Park, Naa Lokko, Tanya Logvinenko, Niroshan Ramachandran, Andrew K Godwin, Jeffrey Marks, Paul Engstrom, and Joshua LaBaer. Protein Microarray Signature of Autoantibody Biomarkers for the Early Detection of Breast Cancer. *Journal of Proteome Research*, 10(1):85–96, January 2011.

[61] P Zacchi. Selecting Open Reading Frames From DNA. *Genome Research*, 13(5):980–990, May 2003.

[62] Youngbae Kim, Nora B Caberoy, Gabriela Alvarado, Janet L Davis, William J Feuer, and Wei Li. Identification of Hnrph3 as an autoantigen for acute anterior uveitis. *Clinical Immunology*, pages –, October 2010.

[63] J B Hughes, J J Hellmann, T H Ricketts, and B J Bohannan. Counting the uncountable: statistical approaches to estimating microbial diversity. *Applied and Environmental Microbiology*, 67(10):4399–4406, October 2001.

[64] Jeremy B Swann and Mark J Smyth. Immune surveillance of tumors. *The Journal of clinical investigation*, 117(5):1137–1146, May 2007.

[65] Robert B Darnell and Jerome B Posner. Paraneoplastic syndromes involving the nervous system. *The New England journal of medicine*, 349(16):1543–1554, October 2003.

[66] Kiran K Musunuru and Santosh S Kesari. Paraneoplastic opsoclonus-myoclonus ataxia associated with non-small-cell lung carcinoma. *Journal of Neuro-Oncology*, 90(2):213–216, November 2008.

[67] PC CONSUL and MM SHOUKRI. Maximum-Likelihood Estimation for the Generalized Poisson-Distribution. *Communications in Statistics-Theory and Methods*, 13(12):1533–1547, 1984.

[68] S Srivastava and L Chen. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic acids research*, 38(17):e170–e170, September 2010.

[69] T L TL Bailey and C C Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 2:28–36, January 1994.

[70] L G Almeida, N J Sakabe, A R deOliveira, M C C Silva, A S Mundstein, T Cohen, Y-T Chen, R Chua, S Gurung, S Gnjatic, A A Jungbluth, O L Caballero, A Bairoch, E Kiesler, S L White, A J G Simpson, L J Old, A A Camargo, and A T R Vasconcelos. CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens. *Nucleic acids research*, 37(Database):D816–D819, January 2009.

[71] D Rimoldi, V Rubio-Godoy, V Dutoit, D Lienard, S Salvi, P Guillaume, D Speiser, E Stockert, G Spagnoli, C Servis, J C Cerottini, F Lejeune, P Romero, and D Valmori. Efficient simultaneous presentation of NY-ESO-1/LAGE-1 primary and nonprimary open reading frame-derived CTL epitopes in melanoma. *Journal of immunology (Baltimore, Md. : 1950)*, 165(12):7253–7261, December 2000.

[72] Y-T Chen, A O Güre, S Tsang, E Stockert, E Jäger, A Knuth, and L J Old. Identification of multiple cancer/testis antigens by allogeneic antibody screening of a melanoma cell line library. *Proceedings of the National Academy of Sciences of the United States of America*, 95(12): 6919–6923, June 1998.

[73] Patricia P Blanco-Arias, Carole A CA Sargent, and Nabeel A NA Affara. The human-specific Yp11.2/Xq21.3 homology block encodes a potentially functional testis-specific TGIF-like retroposon. *Mammalian Genome*, 13(8):463–468, August 2002.

[74] Lisa Berglund, Erik Björling, Per Oksvold, Linn Fagerberg, Anna Asplund, Cristina Al-Khalili Szigyarto, Anja Persson, Jenny Ottosson, Henrik Wernérus, Peter Nilsson, Emma Lundberg, Asa Sivertsson, Sanjay Navani, Kenneth Wester, Caroline Kampf, Sophia Hober, Fredrik Pontén, and Mathias Uhlén. A genecentric Human Protein Atlas for expression profiles based on antibodies. *Molecular & cellular proteomics : MCP*, 7(10):2019–2027, October 2008.

[75] L S LI, W A HAGOPIAN, H R BRASHEAR, T DANIELS, and A LERNMARK. Identification of Autoantibody Epitopes of Glutamic-Acid Decarboxylase in Stiff-Man Syndrome Patients. *Journal of immunology (Baltimore, Md. : 1950)*, 152(2):930–934, 1994.

[76] H L Schwartz, J M Chandonia, S F Kash, J Kanaani, E Tunnell, A Domingo, F E Cohen, J P Banga, A M Madec, W Richter, and S Baekkeskov. High-resolution autoreactive epitope mapping and structural modeling of the 65 kDa form of human glutamic acid decarboxylase. *Journal of molecular biology*, 287(5):983–999, April 1999.

[77] Kunikazu Tanji, Tetsu Kamitani, Fumiaki Mori, Akiyoshi Kakita, Hitoshi Takahashi, and Koichi Wakabayashi. TRIM9, a novel brain-specific E3 ubiquitin ligase, is repressed in the brain of Parkinson's disease and dementia with Lewy bodies. *Neurobiology of Disease*, 38(2):210–218, 2010.

[78] Jean-François Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F Berriz, Francis D Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, Niels Klitgord, Christophe Simon, Mike Boxem, Stuart Milstein, Jennifer Rosenberg, Debra S Goldberg, Lan V Zhang, Sharyl L Wong, Giovanni Franklin, Siming Li, Joanna S Albala, Janghoo Lim, Carlene Fraughton, Estelle Llamosas, Sebiha Cevik, Camille Bex, Philippe Lamesch, Robert S Sikorski, Jean Vandenhaute, Huda Y Zoghbi, Alex Smolyar, Stephanie Bosak, Reynaldo Sequerra, Lynn Doucette-Stamm, Michael E Cusick, David E Hill, Frederick P Roth, and Marc Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, October 2005.

[79] A Ciccia, A L Bredemeyer, M E Sowa, M-E Terret, P V Jallepalli, J W Harper, and S J Elledge. The SIOD disorder protein SMARCAL1 is an RPA-interacting protein involved in replication fork restart. *Genes & Development*, 23(20):2415–2425, October 2009.

[80] G G Mer, A A Bochkarev, R R Gupta, E E Bochkareva, L L Frappier, C J CJ Ingles, A M AM Edwards, and W J WJ Chazin. Structural basis for the recognition of DNA repair proteins UNG2, XPA, and RAD52 by replication factor RPA. *Cell*, 103(3):449–456, October 2000.

[81] D J DJ Barlow, M S MS Edwards, and J M JM Thornton. Continuous and discontinuous protein antigenic determinants. *Nature*, 322(6081):747–748, January 1986.

[82] L Jin, B M Fendly, and J A Wells. High resolution functional analysis of antibody-antigen interactions. *Journal of molecular biology*, 226(3):851–865, August 1992.

[83] Katsushi K Miyazaki, Noriaki N Takeda, Naozumi N Ishimaru, Fumie F Omotehara, Rieko R Arakaki, and Yoshio Y Hayashi. Analysis of in vivo role of alpha-fodrin autoantigen in primary Sjogren's syndrome. *The American Journal of Pathology*, 167(4):1051–1059, October 2005.

[84] M Huang, H Ida, M Kamachi, N Iwanaga, Y Izumi, F Tanaka, K Aratake, K Arima, M Tamai, A Hida, H Nakamura, T Origuchi, A Kawakami, N Ogawa, S Sugai, P J Utz, and K Eguchi. Detection of apoptosis-specific autoantibodies directed against granzyme B-induced cleavage fragments of the SS-B (La) autoantigen in sera from patients with primary Sjögren's syndrome. *Clinical and Experimental Immunology*, 142(1):148–154, October 2005.

[85] D C DC Robbins, S M SM Cooper, S E SE Fineberg, and P M PM Mead. Antibodies to covalent aggregates of insulin in blood of insulin-using diabetic patients. *Diabetes*, 36 (7):838–841, July 1987.

[86] Katerina K KK Papachroni, Natalia N Ninkina, Angeliki A Papapanagiotou, Georgios M GM Hadjigeorgiou, Georgia G Xiromerisiou, Alexandros A Papadimitriou, Anastasios A Kalofoutis, and Vladimir L VL Buchman. Autoantibodies to alpha-synuclein in inherited Parkinson's disease. *Journal of Neurochemistry*, 101(3):749–756, May 2007.

[87] M C MC Dalakas, M M Fujii, M M Li, and B B McElroy. The clinical spectrum of anti-GAD antibody-positive patients with stiff-person syndrome. *Neurology*, 55(10):1531–1535, November 2000.

[88] Philippe Lamesch, Ning Li, Stuart Milstein, Changyu Fan, Tong Hao, Gabor Szabo, Zhenjun Hu, Kavitha Venkatesan, Graeme Bethel, Paul Martin, Jane Rogers, Stephanie Lawlor, Stuart McLaren, Amélie Dricot, Heather Borick, Michael E Cusick, Jean Vandenhaute, Ian Dunham, David E Hill, and Marc Vidal. hORFeome v3.1: a resource of human open reading frames representing over 10,000 human genes. *Genomics*, 89(3): 307–315, March 2007.

[89] Matthias Meyer and Martin Kircher. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor protocols*, 2010(6): pdb.prot5448–prot5448, June 2010.

[90] Michelle M A Fernando, Christine R Stevens, Emily C Walsh, Philip L De Jager, Philippe Goyette, Robert M Plenge, Timothy J Vyse, and John D Rioux. Defining the role of the MHC in autoimmunity: a review and pooled analysis. *PLoS genetics*, 4(4):e1000024–e1000024, April 2008.

[91] Li Zhang and George S Eisenbarth. Prediction and prevention of Type 1 diabetes mellitus. *Journal of diabetes*, 3(1):48–57, March 2011.

[92] Barbara Serafini, Barbara Rosicarelli, Roberta Magliozzi, Egidio Stigliano, and Francesca Aloisi. Detection of ectopic B-cell follicles with germinal centers in the meninges of patients with secondary progressive multiple sclerosis. *Brain pathology (Zurich, Switzerland)*, 14(2):164–174, April 2004.

[93] Hans Link and Yu-Min Huang. Oligoclonal bands in multiple sclerosis cerebrospinal fluid: an update on methodology and clinical usefulness. *Journal of neuroimmunology*, 180 (1-2):17–28, November 2006.

[94] Jonathan C W Edwards, Leszek Szczepanski, Jacek Szechinski, Anna Filipowicz-Sosnowska, Paul Emery, David R Close, Randall M Stevens, and Tim Shaw. Efficacy of B-cell-targeted therapy with rituximab in patients with rheumatoid arthritis. *The New England journal of medicine*, 350(25):2572–2581, June 2004.

[95] Kenneth G Saag, Gim Gee Teng, Nivedita M Patkar, Jeremy Anuntiyo, Catherine Finney, Jeffrey R Curtis, Harold E Paulus, Amy Mudano, Maria Pisu, Mary Elkins-Melton, Ryan Outman, Jeroan J Allison, Maria Suarez Almazor, S Louis Jr Bridges, W Winn Chatham, Marc Hochberg, Catherine Maclean, Ted Mikuls, Larry W Moreland, James O'Dell, Anthony M Turkiewicz, and Daniel E Furst. American College of Rheumatology 2008 recommendations for the use of nonbiologic and biologic disease-modifying antirheumatic drugs in rheumatoid arthritis. *Arthritis & Rheumatism-Arthritis Care & Research*, 59(6):762–784, 2008.

[96] Mark D Pescovitz, Carla J Greenbaum, Heidi Krause-Steinrauf, Dorothy J Becker, Stephen E Gitelman, Robin Goland, Peter A Gottlieb, Jennifer B Marks, Paula F McGee, Antoinette M Moran, Philip Raskin, Henry Rodriguez, Desmond A Schatz, Diane Wherrett, Darrell M Wilson, John M Lachin, Jay S Skyler, and Type 1 Diabetes TrialNet Anti-CD20 Study Group. Rituximab, B-lymphocyte depletion, and preservation of beta-cell function. *The New England journal of medicine*, 361(22):2143–2152, November 2009.

[97] Stephen L Hauser, Emmanuelle Waubant, Douglas L Arnold, Timothy Vollmer, Jack Antel, Robert J Fox, Amit Bar-Or, Michael Panzara, Neena Sarkar, Sunil Agarwal, Annette Langer-Gould, Craig H Smith, and HERMES Trial Group. B-cell depletion with rituximab in relapsing-remitting multiple sclerosis. *The New England journal of medicine*, 358 (7):676–688, February 2008.

[98] Sabine S Cepok, Dun D Zhou, Rajneesh R Srivastava, Stefan S Nessler, Susanne S Stei, Konrad K Büssow, Norbert N Sommer, and Bernhard B Hemmer. Identification of Epstein-Barr virus proteins as putative targets of the immune response in multiple sclerosis. *The Journal of clinical investigation*, 115(5):1352–1360, May 2005.

[99] Gwendolyn F Elphick, William Querbes, Joslynn A Jordan, Gretchen V Gee, Sylvia Eash, Kate Manley, Aisling Dugan, Megan Stanifer, Anushree Bhatnagar, Wesley K Kroeze, Bryan L Roth, and Walter J Atwood. The human polyomavirus, JCV, uses serotonin receptors to infect cells. *Science*, 306(5700):1380–1383, November 2004.

[100] K H KH Rand, H H Houck, N D ND Denslow, and K M KM Heilman. Molecular approach to find target(s) for oligoclonal bands in multiple sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 65(1):48–55, July 1998.

[101] J De Grijse, M Asanghanwa, B Nouthe, N Albrecher, P Goubert, I Vermeulen, S Van Der Meeren, K Decochez, I Weets, B Keymeulen, V Lampasona, J Wenzlau, J C Hutton,

D Pipeleers, F K Gorus, and Belgian Diabetes Registry. Predictive power of screening for antibodies against insulinoma-associated protein 2 beta (IA-2beta) and zinc transporter-8 to select first-degree relatives of type 1 diabetic patients with risk of rapid progression to clinical onset of the disease: implications for prevention trials. *Diabetologia*, 53(3):517–524, March 2010.

[102] Kevin C KC O'Connor, Katherine A KA McLaughlin, Philip L PL De Jager, Tanuja T Chitnis, Estelle E Bettelli, Chenqi C Xu, William H WH Robinson, Sunil V SV Cherry, Amit A Bar-Or, Brenda B Banwell, Hikoaki H Fukaura, Toshiyuki T Fukazawa, Silvia S Tenembaum, Susan J SJ Wong, Norma P NP Tavakoli, Zhannat Z Idrissova, Vissia V Viglietta, Kevin K Rostasy, Daniela D Pohl, Russell C RC Dale, Mark M Freedman, Lawrence L Steinman, Guy J GJ Buckle, Vijay K VK Kuchroo, David A DA Hafler, and Kai W KW Wucherpfennig. Self-antigen tetramers discriminate between myelin autoantibodies to native or denatured protein. *Nature medicine*, 13(2):211–217, February 2007.

[103] Eric Boilard, Peter A Nigrovic, Katherine Larabee, Gerald F M Watts, Jonathan S Coblyn, Michael E Weinblatt, Elena M Massarotti, Eileen Remold-O'Donnell, Richard W Farndale, Jerry Ware, and David M Lee. Platelets amplify inflammation in arthritis via collagen-dependent microparticle production. *Science*, 327(5965):580–583, January 2010.

[104] Janet M Wenzlau, Yu Liu, Liping Yu, Ong Moua, Kimberly T Fowler, Sampathkumar Rangasamy, Jay Walters, George S Eisenbarth, Howard W Davidson, and John C Hutton. A common nonsynonymous single nucleotide polymorphism in the SLC30A8 gene determines ZnT8 autoantibody specificity in type 1 diabetes. *Diabetes*, 57(10):2693–2697, October 2008.

[105] Per-Johan Meijer, Peter S Andersen, Margit Haahr Hansen, Lucilla Steinaa, Allan Jensen, Johan Lantto, Martin B Oleksiewicz, Kaja Tengbjerg, Tine R Poulsen, Vincent W Coljee, Søren Bregenholt, John S Haurum, and Lars S Nielsen. Isolation of human antibody repertoires with preservation of the natural heavy and light chain pairing. *Journal of molecular biology*, 358(3):764–772, May 2006.

[106] Graham P Wright, Clare A Notley, Shao-An Xue, Gavin M Bendle, Angelika Holler, Ton N Schumacher, Michael R Ehrenstein, and Hans J Stauss. Adoptive therapy with redirected primary regulatory T cells results in antigen-specific suppression of arthritis. *Proceedings of the National Academy of Sciences*, 106(45):19078–19083, November 2009.

[107] Karin L Heckman and Larry R Pease. Gene splicing and mutagenesis by PCR-driven overlap extension. *Nature protocols*, 2(4):924–932, January 2007.

[108] H H Albert, E C EC Dale, E E Lee, and D W DW Ow. Site-specific integration of DNA into wild-type and mutant lox sites placed in the plant genome. *Plant Journal*, 7(4): 649–659, April 1995.

[109] N Chapal, M Bouanani, M J Embleton, I Navarro-Teulon, M Biard-Piechaczyk, B Pau, and S Peraldi-Roux. In-cell assembly of scFv from human thyroid-infiltrating B cells. *BioTechniques*, 23(3):518–524, September 1997.

[110] Frank Diehl, Meng Li, Yiping He, Kenneth W Kinzler, Bert Vogelstein, and Devin Dressman. BEAMing: single-molecule PCR on microparticles in water-in-oil emulsions. *Nature methods*, 3(7):551–559, July 2006.

[111] M J Embleton, G Gorochov, P T Jones, and G Winter. In-cell PCR from mRNA: amplifying and linking the rearranged immunoglobulin heavy and light chain V-genes within single cells. *Nucleic acids research*, 20(15):3831–3837, August 1992.

[112] Jenifer Clausell-Tormos, Diana Lieber, Jean-Christophe Baret, Abdeslam El-Harrak, Oliver J Miller, Lucas Frenz, Joshua Blouwolff, Katherine J Humphry, Sarah Köster, Honey Duan, Christian Holtze, David A Weitz, Andrew D Griffiths, and Christoph A Merten. Droplet-based microfluidic platforms for the encapsulation and screening of Mammalian cells and multicellular organisms. *Chemistry & Biology*, 15(5):427–437, May 2008.

[113] John H Leamon, Darren R Link, Michael Egholm, and Jonathan M Rothberg. Overview: methods and applications for droplet compartmentalization of biology. *Nature methods*, 3 (7):541–543, July 2006.

[114] Michihiko Nakano, Naohito Nakai, Hirofumi Kurita, Jun Komatsu, Kazunori Takashima, Shinji Katsura, and Akira Mizuno. Single-molecule reverse transcription polymerase chain reaction using water-in-oil emulsion. *Journal of bioscience and bioengineering*, 99(3):293–295, March 2005.

[115] Richard Williams, Sergio G Peisajovich, Oliver J Miller, Shlomo Magdassi, Dan S Tawfik, and Andrew D Griffiths. Amplification of complex gene libraries by emulsion PCR. *Nature methods*, 3(7):545–550, July 2006.

[116] Ryan Tewhey, Jason B Warner, Masakazu Nakano, Brian Libby, Martina Medkova, Patricia H David, Steve K Kotsopoulos, Michael L Samuels, J Brian Hutchison, Jonathan W Larson, Eric J Topol, Michael P Weiner, Olivier Harismendy, Jeff Olson, Darren R Link, and Kelly A Frazer. Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nature biotechnology*, 27(11):1025–1031, November 2009.

[117] RainDance Technologies. URL `http://raindancetech.com/`.

[118] T W TW Myers and D H DH Gelfand. Reverse transcription and DNA amplification by a Thermus thermophilus DNA polymerase. *Biochemistry (Washington)*, 30(31):7661–7666, August 1991.

[119] K Yokoyama, F Saka, T Kai, and E Soeda. Encapsulation of cells in agarose beads for use in the construction of human DNA libraries as yeast artificial chromosomes (YAC). *Jinrui idengaku zasshi. The Japanese journal of human genetics*, 35(2):131–143, June 1990.

[120] Omar Bagasra. Protocols for the in situ PCR-amplification and detection of mRNA and DNA sequences. *Nature protocols*, 2(11):2782–2795, 2007.

[121] Paul J Carter. Potent antibody therapeutics by design. *Nature Reviews: Immunology*, 6(5): 343–357, May 2006.

[122] T T Clackson, H R HR Hoogenboom, A D AD Griffiths, and G G Winter. Making antibody fragments using phage display libraries. *Nature*, 352(6336):624–628, August 1991.

[123] J J McCafferty, A D AD Griffiths, G G Winter, and D J DJ Chiswell. Phage antibodies: filamentous phage displaying antibody variable domains. *Nature*, 348(6301):552–554, December 1990.

[124] E T Boder and K D Wittrup. Yeast surface display for screening combinatorial polypeptide libraries. *Nature biotechnology*, 15(6):553–557, June 1997.

[125] E T ET Boder, K S KS Midelfort, and K D KD Wittrup. Directed evolution of antibody fragments with monovalent femtomolar antigen-binding affinity. *Proceedings of the National Academy of Sciences of the United States of America*, 97(20):10701–10705, September 2000.

[126] Ginger Chao, Wai L Lau, Benjamin J Hackel, Stephen L Sazinsky, Shaun M Lippow, and K Dane Wittrup. Isolating and engineering human antibodies using yeast surface display. *Nature protocols*, 1(2):755–768, 2006.

[127] Christian Zahnd, Patrick Amstutz, and Andreas Plückthun. Ribosome display: selecting and evolving proteins in vitro that specifically bind to a target. *Nature methods*, 4(3):269–279, March 2007.

[128] Hennie R Hoogenboom. Selecting and screening recombinant antibody libraries. *Nature biotechnology*, 23(9):1105–1116, September 2005.

[129] Ola Söderberg, Mats Gullberg, Malin Jarvius, Karin Ridderstråle, Karl-Johan Leuchowius, Jonas Jarvius, Kenneth Wester, Per Hydbring, Fuad Bahram, Lars-Gunnar Larsson, and Ulf Landegren. Direct observation of individual endogenous protein complexes in situ by proximity ligation. *Nature methods*, 3(12):995–1000, December 2006.