On the Inference of Convergence, Selection, and Coevolution of Rapidly Evolving Populations

by

Nicolas Strauli

DISSERTATION

Submitted in partial satisfaction of the requirements for degree of

DOCTOR OF PHILOSOPHY

in

Biomedical Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:

Katherine Pollard — Katherine Pollard

9B0DDB91029D4F1...

Chair

DocuSigned by:

Ryan Hernandez — Ryan Hernandez

D7401DE9E4D8...

DocuSigned by:

Satish Pillai — Satish Pillai

B1C0A7BF8A451...

DocuSigned by:

Michael Wilson — Michael Wilson

354F13DACE76493...

Committee Members

To my father, Peter Strauli.

I never quite made it to male model,

but at least there’s a doctor in the family.

Shneeky leeky pops.

Acknowledgements

Personal

I’d like to thank my late mother, Joan Rushton, who died when I was 10. Over the years, my visual memory of her has faded, yet I suspect that I am a lot like her. The warm thought that she would be proud of me has been my prime source of comfort and guidance throughout my life.

I’d also like to thank my late father, Peter Strauli. who died three years ago. My father raised me as a single parent and gave me exactly the kind of support I needed: unconditional love, with little questions. The antithesis of a helicopter-parent, his only concerns were that I was happy, did my homework, and went to school with clean fingernails. The rest was up to me.

Like most, my older brother, Christian Strauli, first served as a role model, and then at some point morphed into my best friend. He probably still views himself as a role model, which I quietly view as hilarious. Our father would frequently say that Christian was the brains and I was the brawns, which (like any good fatherly statement) seemed perfectly calibrated to piss us both off. In many ways we have both lived out our lives to prove this statement wrong: Christian pursued the idyllic life of surfing and gardening on the central coast of California, and I pursued a life of science and research in metropolitan San Francisco. However, we both secretly know that our dad was right: I am clearly the better surfer, and Christian is way smarter than me. It’s hard to overstate the impact that Christian has had on my life. In our early years we clung to each other like a mutual life-raft through a tumultuous childhood. During adolescence Christian was a moral compass, providing me with an excellent example of how not to be a jerk. Now, Christian, and his family, provide a foundation of love and support, without which, I would be quite lost.

The other half of this familial foundation comes from Laura Hunkler. Christian and Laura have set down roots in central California and made a family. Like any tree with strong roots, this provides the basis for an ecosystem of friends and family. Laura has enabled me to be a part of this ecosystem, which has been a great stabilizing force in my life.

Usually, when it comes time for a scientist to thank their significant other, they will say something like ‘there’s no words that can describe blah-blah-blah…’. This is clearly a copout. Scientists aren’t poets so we resort to idioms. However, I am going to try (and likely fail) to properly thank Manon Eckhardt. Here goes:

If you want to see all that is good, kind, and pure in this world, you look into Manon’s eyes when she is smiling; if you want to see all that is cruel, unjust, and painful, you look into her eyes when she is sad. Manon is an endless well spring of life. Her curiosity will intrigue you, educate you, and tire you out. Her happiness will infect you, inspire you, and tire you out. She’s as quick to laugh at her own tears, as a well-timed fart. Early in our relationship she told me that my soul is “hella old”. Her soul is hella young. Manon is a talented scientist, yet she does not help me with my science. She helps me to feel better when I am sad, and to celebrate when I am happy. Apparently, Manon gets something from me out of our relationship, although I can’t say what that is. I feel to be the great beneficiary in this transaction, yet she would probably say that she feels the same. A life without Manon Eckhardt would be worth living, but it wouldn't be worth celebrating. Manon is a celebration of life.

Professional

I would like to thank the anonymous patients that made most of the research in this thesis possible. When performing my work on human samples, I would often look at the small vials of blood and marvel at how kind and generous these people must have been to donate their time, and person to enable researchers like me to do our work. This generosity has both grounded and inspired me to make my contributions to biomedical research justify these individuals' generosity.

I often feel as if I did half of a PhD prior to starting my real PhD. This preliminary PhD was, in actuality, a research technician position in Dr. Ophir Klein's lab. Ophir gave me my start in science. Were it not for his early mentorship and support, I am certain that I would not be in graduate school today. Ophir taught me what it means to be a good mentor, and how to advocate for those working beneath you. While in the Klein Lab, I primarily preformed research under the tutelage of Dr. Brian Biehs. Brian was an incredibly talented post doc who had an interesting mixture of a salty and supportive attitude about research. This turned out to be the perfect mentor for someone who was considering starting a career as a researcher. His saltyness kept me honest, and his support kept me motivated. Instead of blind encouragement, he openly expressed skepticism about enrolling in a PhD program, which is the most important thing that one can do for a young mind considering this path. I thank Dr. Biehs for his early support and his honesty.

Although I have never officially met him, I would like to thank Dr. Mike McCune. Dr. McCune was a well-established professor at UCSF when I emailed him about an idea for a project that I had concerning coevolving HIV and antibodies. Dr. McCune took the time to put me in touch with Mike Busch at the Blood Systems Research Institute (BSRI), which led to my collaboration with Satish Pillai's lab. Dr. McCune showed me what it means to foster collaborations and to be a good steward of science.

Through the connections of Dr. McCune, I was able to get in contact with Dr. Chris Pilcher, who directed the OPTIONS cohort of HIV+ human specimens. Dr. Pilcher allowed us full access to their repository, which made Chapter 4 of this thesis possible. Selecting samples from a repository is not an easy process, and Shelley Facente graciously gave her time and effort to make this possible.

After being put in contact with the BSRI, I quickly became a collaborator with Satish Pillai and Mohamed Abdel-Mohsen. Satish and Mohamed are like the "Click and Clack the Tappet brothers" of science. If you approach them with a scientific question they will engage in a long humorous discussion, filled with casual insults, and will eventually expertly diagnose the problem and figure out a solution. Satish and Mohamed taught me what it means to be good collaborators, and also taught me a little about foosball. I would also like to thank the entire Pillai Lab at the BSRI, they taught me how to handle human specimens, amplify HIV, and make sequencing libraries. They were my surrogate lab for many months, and were incredibly welcoming.

While attending a group meeting during the early days of my graduate career, a post-doc in Dr. Katie Pollard's lab raised her hand to ask a question and then simply said, "I don't get it". This was one of the most influential moments during my PhD. There are few things more demoralizing than sitting through a presentation and being totally lost. However, prior to this moment, I never would have had the courage to so blatantly acknowledge my confusion as this post doc did with her statement, “I don’t get it”. Over the years of group meetings with the Pollard Lab, I have realized that Katie is rarely lost, yet she takes great pains to foster an environment of mutual respect and openness, which allows those involved to feel comfortable enough to admit ignorance and ask questions. This environment allowed me to maximize the amount that I learned from my peers, and also gave me the resolve and confidence to never again accept being lost during a presentation. I thank Katie for fostering this open environment, and this excellent example of leadership.

When I first started graduate school, I became aware of a super chill, goofy dude, with curly hair, who was really into music and human genetics. This dude’s name was Raul Torres, and I quickly became friends with him. Raul has the misfortune of being a Rangers fan, but he is from Texas, so he was basically born into it. I pity this aspect of his personality, but he partially redeemed himself by embracing the Golden State Warriors. Raul and I started graduate school together, joined the same lab, and ended graduate school at nearly the same time (he beat me by one quarter, but in my defense, I had to make my own data). Raul and I also sat next to each other throughout our time in the Hernandez Lab. We have quietly battled through setbacks, milestones, failed projects, manuscript acceptances, wasted days, encouraging results, rejections, stressful talks, boredom, exuberance, terrible plots, beautiful plots, scientific discussions, and the rest of the PhD gamut, all while sitting three feet from one another. At any given time, I may have spent more time with Raul, than my girlfriend. During this time Raul had gone from a colleague to a good friend. I thank Dr. Torres for his generous intellectual input over the years, and also for being a supportive friend.

I was awarded one fellowship during my PhD, and the only reason I got it was because of the thoughtful comments I got from Dr. Lawrence Uricchio. This exemplifies the type of person that Lawrence is: incredibly smart, and incredibly thoughtful. Like Raul, Lawrence transitioned from colleague to friend over the course of our respective PhDs. He is one of the most talented scientists I have interacted with, and also is a great person to have a beer with. I thank Dr. Uricchio for his intellectual input into my science, and also his friendship.

I also would like to thank the entire Hernandez Lab. While the Hernandez Lab's tenure at UCSF has ended, it has always been a supportive environment for excellent science.

Lastly, I would like to thank my advisor Dr. Ryan Hernandez. I feel incredibly lucky to have had the opportunity to work with, and learn from Ryan. Ryan was unwaveringly supportive of both my graduate education and research efforts. By taking me into his lab, he gave me the opportunity to learn an exciting new field. I learned early on not to think of Ryan as my 'boss', but as my 'advisor'. His role was one of advice and support. He gave me room to form my own ideas, design my own projects, and forge my own path; all while having his door perpetually open to offer help when needed. This taught me the most important thing that I have learned in graduate school. How to be an independent investigator.

Acknowledgement of Previously Published Materials and Research Contributions

All of the material in Chapter 2 of this dissertation was previously published in Strauli *et al*. 2016 [1]. Ryan Hernandez advised the research that forms the basis of this chapter, and all analyses therein were designed by a close collaboration between Nicolas Strauli and Ryan Hernandez. The manuscript was written, and figures were created by Nicolas Strauli.

The Methods section of Chapter 3 has been previously published as an accompaniment to Fernandez *et al.* 2016 [2] and Fedewa *et al.* 2018 [3]. These methods were from the work of Nicolas Strauli, and supervised by Ryan Hernandez. Additionally, the part of the Results section of Chapter 3 that has to do with Ebola Virus was the result of a collaboration between Nicolas Strauli and Melissa Spear. Figures 3.8 and 3.9 were created by and used with the permission of Melissa Spear.

The content of Chapter 4 will be submitted to a peer reviewed journal. This research was supervised by Ryan Hernandez and Satish Pillai. The sequencing approach for human immunodeficiency virus was developed by Mohamed Abdel-Mohsen. Specimen selection for this project was aided by Shelley Facente, and Chris Pilcher directed the repository from which these specimens originated. All text and figures were created by Nicolas Strauli.

References

1. Strauli NB, Hernandez RD. Statistical inference of a convergent antibody repertoire response to influenza vaccine. Genome Med [Internet]. BioMed Central; 2016 [cited 2018 Nov 15];8:60. Available from: http://genomemedicine.biomedcentral.com/articles/10.1186/s13073-016-0314-z

2. Fernandes JD, Faust TB, Strauli NB, Smith C, Crosby DC, Nakamura RL, et al. Functional Segregation of Overlapping Genes in HIV. Cell [Internet]. 2016 [cited 2019 Feb 12];167:1762–1773.e12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27984726

3. Fedewa G, Radoshitzky SR, Chī X, Dǒng L, Zeng X, Spear M, et al. Ebola virus, but not Marburg virus, replicates efficiently and without required adaptation in snake cells. Virus Evol [Internet]. Oxford University Press; 2018 [cited 2019 Feb 12];4. Available from: https://academic.oup.com/ve/article/doi/10.1093/ve/vey034/5214739

ABSTRACT

On the Inference of Convergence, Selection, and Coevolution of Rapidly Evolving Populations

Nicolas Strauli

Genetics is currently undergoing a ‘big-data’ revolution, where the advent of deep-sequencing has enabled researchers to routinely create massive datasets, yet the statistical analyses of these data remain challenging. Deep-sequencing has proved particularly useful in the field of evolution, where researchers can sequence rapidly adapting populations to observe evolution taking place in real-time. In this dissertation, I design and implement statistical tools to answer relatively simple evolutionary questions about complex, large, and rapidly adapting populations. Specifically, I develop methods to infer convergence in the antibodies of *in vivo* human B cells, selection in *in vitro* viral populations, and coevolution between *in vivo* populations of human antibodies and autologous human immunodeficiency virus (HIV). I show that the antibodies of human B cells will converge towards similar genetic sequence characteristics when presented with identical influenza vaccines, and that HIV and antibodies can exhibit genetic signatures of coevolution, although this is quite rare. I also show that genetic drift plays a significant role in an experimental population of HIV, and that this needs to be taken into account when inferring selection. Together, I hope that the analyses described in this work prove useful to other investigators with similar evolutionary questions.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Biology as a discipline of science is currently undergoing a revolution as to how it is practiced. Historically, a biologist could conduct an experiment and visually look at the data or use a simple statistical test to determine if her hypothesis was confirmed or rejected. Today, a variety of technological advances have enabled each experiment to generate massive amounts of biological data. While these high-throughput technologies—collectively referred to as 'omics'—have allowed us to answer questions in biology that have previously been unapproachable [1], they have also made the analysis of such data orders of magnitude more complicated. The collective biologist has bid a fond farewell to the days of simple bar-charts and t-tests, and has instead embraced complex visualizations and statistical analyses in order to 'see' what the data is telling her. This thesis is broadly a story about how to design custom statistical tests when one is using large genetic datasets in order to answer relatively simple questions in evolutionary biology.

In no discipline of biology has the 'omics' revolution been more prevalent than in genetics. The advent of 'deep-sequencing' [2]—where billions of short DNA fragments can be simultaneously sequenced—has resulted in genetic data to be available at a scale, speed, and affordability that is an orders of magnitude improvement over previous technologies. This has allowed for studies to publish 1,000's of human genomes at once [3], the ability to characterize entire transcriptomes of different tissues [4], and the ability to survey entire populations of microorganisms [5–7] (to name a few).

One small, yet growing field that has been made possible by the advent of deep-sequencing technologies is the study of immune repertoires. In 1987 Susumu Tonegawa won the Nobel Prize in physiology or medicine for his discovery of how antibody (Ab) diversity within an organism is generated via somatic mutational processes in the genomes of B cells [8]—the cells that express antibodies. However, it was not until the advent of deep-sequencing that we could actually witness the vast complexity and diversity of antibodies that is generated through this process [9]. This population of antibodies is collectively referred to as the antibody repertoire (AbR), and its vast diversity plays a large role in how humans generate (and maintain) humoral immunity. Interestingly, the way in which Abs develop the ability to bind to a specific antigenic target—which is to say, immunity—is strikingly similar (if not identical) to evolution by natural selection, as laid out by Charles Darwin [10], and the modern synthesis, as compiled by Julian Huxley [11]. Through the processes of V(D)J recombination and somatic hypermutation, B cells will impart a smattering of mutations upon their genetic locus that encodes for Abs. This in turn creates a massive amount of phenotypic diversity in the organism's AbR. Then, the B cells that succeed in tightly binding an antigenic target are instructed to proliferate, while those that don't are instructed to die [12]. Thus, completing the evolutionary process.

This process of mutagenesis followed by selection during Ab development is often described in rather dry biomolecular terms. However, being primarily interested in evolution, it is important to point out the higher-level profundity of this evolutionary process. It is essentially a type of 'meta-evolution', where we have evolved the ability to evolve a subset of our tissues. This ability is unique to adaptive immune systems, and is conserved across all jawed vertebrates [13]. Indeed, it seems to be an essential component of metazoan success. A fundamental dilemma for metazoans is how to defend against parasites that have orders of magnitude shorter generation times and thus can evolve infection strategies vastly quicker than metazoans can evolve defenses to them. It’s possible that one answer to this dilemma is that metazoans have evolved an adaptive immune system that itself can evolve at a rate on par with many pathogens.

The ability to survey the AbR using deep-sequencing based methods has allowed researchers for the first time to be able to see this evolutionary process take place at a whole-population scale, which has in turn enabled us to learn a great deal about how adaptive immunity is acquired following an immune insult [14]. One persistent question about immunity (particularly pertaining to B cell mediated immunity) is how idiosyncratic each individual’s immune response is. Each person has a unique AbR through the collective action of random mutation events in the Ab locus of developing B cells. However, if people are given a similar antigenic stimulus, are Abs with similar sequence characteristics selected to proliferate across independent individuals? This is essentially a question of convergent evolution taking place across independent AbRs, and we test for this in Chapter 2 by looking at the AbRs of a handful of individuals who all were given the same influenza vaccine. In this chapter we show that establishing convergence across AbRs is more complicated than previously appreciated, and we develop a proper statistical framework that uses time-series AbR sequence data to answer it.

One particularly gratifying result of the massive influx of genetic data has been to catapult the previously data-light, theory-heavy field of population genetics into the realm of empirical science. For example, in the past, population geneticists used mathematical models of hypothetical populations to predict how a given positively selected allele might logistically rise in frequency; whereas now, one can use genetic data generated from deep sequencers to actually observe this in the real world—or frequently, in controlled experimental conditions. This has opened up the field of experimental evolution, where researchers can observe the evolution of populations of microorganisms in real-time, and under different conditions. When experimental evolution is used in concert with deep-sequencing, one can simultaneously track the relative frequencies of thousands of alleles overtime [15]. In Chapter 3, we develop a method that combines time-series genetic data from experimental evolution studies with population genetics theory in order to estimate the level of selection (selection coefficient), and to statistically test if an allele is under significant selection. This allows researchers to i) identify alleles/loci of importance in their study, and ii) build fitness landscapes, where the relative fitness of each allele in their data can be mapped.

The main drawback of the experimental evolution approach is that the studies typically occur in highly controlled, and artificial laboratory settings. While this gives researchers the unique ability to control parameters such as initial genetic diversity, population size, and demography of a given experiment, it also means that researchers cannot be sure that any adaptations they observe are not the result of organisms adapting to a strange laboratory environment that would not otherwise occur in the natural world. An obvious remedy to this is to use deep sequencing to study natural populations [16,17], however, this approach comes with the drawback that, in the infinitely complicated natural world, it is difficult to identify which of the many known (and unknown) environmental parameters are responsible for *causing* a given adaptation. It seems that there is a tradeoff with studying experimental vs. natural populations. Experimental populations afford one with exquisite control over experimental parameters, allowing researchers to more easily assign causality, yet lose applicability to the natural world. Whereas findings from natural populations are pertinent to the real, natural world, yet occur in such complicated environments that assigning causality is often challenging. Because of this seemingly insurmountable tradeoff, it is important to perform both types of studies. Chapter 3 deals with two experimental evolution studies, but the methods outlined therein could be applied to time-series studies of either experimental or natural populations.

An important feature of memory B cells—a particular type of B cell that tends to be long lived—is that they can undergo iterative rounds of selection. Which is to say that if a memory B cell that has undergone previous rounds of selection, encounters a new antigen that is similar to what it had previously bound, it can undergo further rounds of selection to once again bind the new antigen. The possibility for iterative rounds of selection occurring on the same B cell lineage, raises the possibility for antagonistic coevolution between Abs and a chronic infection. The idea here is that upon infection, neutralizing Abs gradually develop, then the pathogen evolves escape mutations, then the previously neutralizing Abs evolve innovations that allow them once again to bind to their antigen, and then the pathogen once again escapes, and the process repeats until either the infection is cleared, or the host succumbs to the disease. There have been many accounts of this coevolutionary 'arms-race' occurring between HIV and HIV-neutralizing Ab lineages [18]. However, the overall AbR response to chronic infections remains not well characterized. For example, it is not known if these long-term putatively coevolutionary interactions between Ab lineages and HIV are the exception or the rule when it comes to chronic infections. Further, the examples of Ab/HIV coevolution to date have been qualitative in nature. In Chapter 4, we describe a study where we deeply sequence both the HIV population and the AbR in several chronically infected patients over 10-20 time-points per patient. We use this data to characterize the overall genetic interaction between the HIV and the AbR populations, and then search for long-term coevolutionary interactions between HIV and Ab lineages.

The current age of biological research is presented with a unique problem, where it is far easier to create information rich data, than it is to meaningfully examine it. In this thesis I present three approaches for using large genetic datasets to arrive at statistically informed conclusions for relatively simple evolutionary questions: i) How to test for convergence across immune repertoires. ii) How to test for selection in experimental populations. And iii) how to test for coevolution between Abs and chronic infections. I hope that these examples prove useful to other researchers with similar questions, and more importantly, that this provides motivation for others to create their own innovative approaches to answering their unique questions with big data.

References

1. Mathé E, Hays J, Stover D, Chen J. The Omics Revolution Continues: The Maturation of High-Throughput Biological Data Sources. Yearb Med Inform [Internet]. Georg Thieme Verlag KG; 2018 [cited 2019 Jan 29];27:211–22. Available from: http://www.thieme-connect.de/DOI/DOI?10.1055/s-0038-1667085

2. Kulski JK. Next-Generation Sequencing — An Overview of the History, Tools, and “Omic” Applications. Next Gener Seq - Adv Appl Challenges [Internet]. InTech; 2016 [cited 2019 Feb 1]. Available from: http://www.intechopen.com/books/next-generation-sequencing-advances-applications-and-challenges/next-generation-sequencing-an-overview-of-the-history-tools-and-omic-applications

3. Gibbs RA, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, et al. A global reference for human genetic variation. Nature [Internet]. Nature Publishing Group; 2015 [cited 2019 Feb 1];526:68–74. Available from: http://www.nature.com/articles/nature15393

4. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet [Internet]. 2013 [cited 2019 Feb 1];45:580–5. Available from: http://www.nature.com/articles/ng.2653

5. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature [Internet]. 2012 [cited 2019 Feb 1];486:207–14. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22699609

6. Human Microbiome Project Consortium. A framework for human microbiome research. Nature [Internet]. 2012 [cited 2019 Feb 1];486:215–21. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22699610

7. Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, Albert J, et al. Population genomics of intrapatient HIV-1 evolution. Elife [Internet]. 2015 [cited 2019 Feb 1];4. Available from: https://elifesciences.org/articles/11282

8. Tonegawa S. Somatic generation of antibody diversity. Nature [Internet]. Nature Publishing Group; 1983 [cited 2019 Feb 1];302:575–81. Available from: http://www.nature.com/articles/302575a0

9. Jiang N, Weinstein JA, Penland L, White RA, Fisher DS, Quake SR. Determinism and stochasticity during maturation of the zebrafish antibody repertoire. Proc Natl Acad Sci U S A [Internet]. National Academy of Sciences; 2011 [cited 2019 Feb 1];108:5348–53. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21393572

10. Darwin C, Darwin C. On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life / [Internet]. London : John Murray,; 1859 [cited 2019 Feb 4]. Available from: https://www.biodiversitylibrary.org/item/135954

11. Huxley J. Evolution : the modern synthesis. MIT Press; 2010.

12. Murphy K. Janeway’s immunobiology. Garland Science; 2014.

13. Flajnik MF, Kasahara M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. Nat Rev Genet [Internet]. NIH Public Access; 2010 [cited 2019 Feb 4];11:47–59. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19997068

14. VanDuijn MM, Dekker LJ, van IJcken WFJ, Sillevis Smitt PAE, Luider TM. Immune Repertoire after Immunization As Seen by Next-Generation Sequencing and Proteomics. Front Immunol [Internet]. Frontiers Media SA; 2017 [cited 2019 Feb

1];8:1286. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29085363

15. Schlötterer C, Kofler R, Versace E, Tobler R, Franssen SU. Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation. Heredity (Edinb) [Internet]. Nature Publishing Group; 2015 [cited 2019 Feb 4];114:431–40. Available from: http://www.nature.com/articles/hdy201486

16. Garud NR, Good BH, Hallatschek O, Pollard KS. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. Gordo I, editor. PLOS Biol [Internet]. 2019 [cited 2019 Mar 13];17:e3000102. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30673701

17. Faust K, Lahti L, Gonze D, de Vos WM, Raes J. Metagenomics meets time series analysis: unraveling microbial community dynamics. Curr Opin Microbiol [Internet]. 2015 [cited 2019 Mar 13];25:56–66. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26005845

18. Bonsignori M, Liao H-X, Gao F, Williams WB, Alam SM, Montefiori DC, et al. Antibody-virus co-evolution in HIV infection: paths for HIV vaccine development. Immunol Rev [Internet]. 2017 [cited 2019 Jan 14];275:145–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28133802

Chapter 2: Statistical Inference of a convergent antibody repertoire response to influenza vaccine

Introduction

Since the administration of the first designed vaccine by Edward Jenner in 1796 [1], vaccines have proven indispensable for both medicine and medical research. Jenner's work on vaccines are among the rare achievements of science that have fundamentally changed modern life. Perhaps less well known, vaccines also provide a standardized, safe, and ethical way to directly study human adaptive immunity [2]. Most vaccines confer resistance to a given pathogen by stimulating the patient's population of B cells to produce antibodies (Abs) against the inoculated antigens. Each clonal lineage is composed of B cells that are related by a single common naïve B cell ancestor, and the conglomerate of B cells within an individual make up their antibody repertoire (AbR).

Interestingly, the process by which Abs are adapted to more specifically target an insulting antigen is an example of evolution by natural selection. To wit, during B cell development a vast amount of genetic diversity is generated by a series of somatic mutagenic steps, after which, variants that are able to bind an antigen strongly will be positively selected to proliferate [3,4]. The first diversity-generating step in B cell development is a process of somatic recombination that takes place in the bone marrow. The mature Ab protein is composed of two identical light chains and two identical heavy chains. A light chain can be of either the lambda (IGL) or kappa (IGK) variety, whereas the heavy chain has only one possibility (IGH), and the loci encoding these three chains reside in distinct regions of the genome. Here, the Variable (V), Diversity (D), and Joining (J) gene segments in the IGH locus, and V and J gene segments in the light chain loci will recombine [5–7]. Diversity is generated both by selecting one combination out of all the possible combinations of V, (D) and J genes, as well as by the random insertion and deletion of genetic information at the junctions of these gene segments [8]. Further, once a mature B cell binds an antigen, it will be recruited to a lymph follicle and enter a structure known as the germinal center where a process of somatic hyper-mutation (SHM) takes place [3,4]. Random point mutations are smattered onto the variable region of the Ab locus—the area that is responsible for binding antigen—and if these mutations result in high binding affinity, that B cell clone will receive signals to proliferate. This process generates lineages of B cells specific for a given antigen. These mutagenic steps together result in a high concentration of mutations occurring in a region of the Ab called the complementary determining region 3 (CDR3), which is also the region of the Ab that tends to physically interact with antigen. Because of this, the sequence encoding the CDR3 is often used for clonal analysis of Abs, where Abs with the same CDR3 sequence are assumed to be clones. The net effect of this evolutionary process produces extreme temporal dynamism within the AbR, as different lineages grow and shrink in response to different antigenic stimuli [9].

Advancements in next-generation sequencing (NGS) methods have led to recent work in characterizing the AbR's response to a variety of stimuli [9–21] (see Galson *et al* [2] for a review). However, most of this work has focused on methods development, and there has been comparatively little work focusing on what can actually be learned from these data. Contrary to this trend, Greiff *et al.* [22] recently employed a machine learning approach to classify patients' immune status using their AbR sequence data. Much work remains to be done in this relatively new area of research. For example, the overall changes in a patient's AbR could be used to quantitatively assess the response to vaccination. Of particular interest is the ability to use changes in the frequency of individual Abs over time to identify which specific monoclonal Abs (mAbs) respond to a given antigen [23]. For example, if a particular Ab mRNA sequence exhibits a spike in expression in a time-series RNA-seq dataset from peripheral blood then this could be indicative of a vaccine response for that Ab. To address this gap in knowledge, we here seek to leverage time-series information of five patients' AbRs in order to infer the elements that are responding to a trivalent influenza vaccine (TIV).

A particularly useful and intuitive way to model time-series data is to use methods within the greater discipline of functional data analysis (FDA) [24,25]. As opposed to multivariate data analysis (MDA)—which treats each datum as a finite dimensional vector of observations—FDA treats each datum as a continuous function over some dimension, which is often (as in our case) time. FDA-based methods have a rich history of being used for identifying differentially expressed genes over time [26–29], and have the advantage of easily incorporating uneven time-point sampling, and measurement error into each gene's functional model. FDA is also an intuitive way to model gene expression, as each gene's expression level in a tissue is indeed continuously fluctuating over continuous time. Here, we use an FDA-based method presented by Wu and Wu [28], and apply it to time-series AbR data [30] to identify the components of patients' AbR that are responding to a standard TIV.

There is a plethora of time-series gene expression data that have been used to identify genes involved in pathogen defense [31], autoimmunity [31], and vaccine response [30,32]. The longitudinal and cross-sectional nature of these studies allowed the authors to identify the genes that were consistently differentially expressed in response to the given antigenic stimulus across patients. One could perform a similar analysis using a time-series AbR dataset to help identify the determinants of immunity. However, with the exception of Liao *et al.* [33], Laserson *et al.* [9], and more recently Hoehn *et al.* [34], few detailed time-series datasets on the AbR exist. If RNA-seq were performed on an antibody expressing tissue (for example peripheral blood mononuclear cells, or PBMCs), theoretically, many of the RNA transcripts in the data would originate from Ab loci. Should this be the case, there will exist much AbR information within the data that simply need to be bioinformatically mined out. This approach has been used in the context of cancer research to identify the Ab sequence of the cancerous B cell lineage in Chronic Lymphocytic Leukemia patients [35], and to characterize both the AbR and T cell receptor diversity in solid tumor samples [36,37]. In this study, we developed and implemented such a pipeline on the Henn *et al.* 2013 transcriptomic dataset [30] in order to probe the AbR's response to a standard TIV.

There have been several reports of convergent evolutionary signals between independent AbRs that were exposed to a similar antigenic stimulus (recently reviewed by [38]). While there exist relatively precise definitions for convergence in evolutionary biology [39,40], we define convergent AbRs more loosely as those that develop similar characteristics as a response to similar antigens. These characteristics could include similar Ab DNA sequences, similar sets of Ab genes, or similar structural characteristics, among others. In this manuscript we focus on convergence by way of independent AbRs utilizing similar sets of Ab genes, and similar sequences of CDR3s to target the same vaccine. AbR convergence has been shown in a variety of contexts, including dengue virus infections [21], broadly neutralizing Abs against Human

Immunodeficiency Virus [15,18], and influenza vaccination [23,41,42]. With the exception of Parameswaran *et al.* [21] and Cortina-Ceballos et al. [23], these studies relied largely on qualitative evidence for convergence, where Ab sequences from independent patients either cluster closely together on a dendogram [15], or have strikingly similar sequence and/or structural characteristics [18,41,42]. While these examples of AbR convergence may be intuitively convincing, few methods have been developed to statistically test for a convergent AbR response across patients. The importance of statistical analyses can be illustrated by the high correlation of Ab gene expression in different individuals [43,44]. That is, if an Ab gene is expressed highly in one individual, it will tend to also be highly expressed in another individual. In order to soundly establish a convergent signal between patients' AbRs, this correlation in background gene expression must be taken into account. Indeed, Childs *et al.* [45] have used a computational modeling approach to show that a large determinant of AbR diversity post inoculation is its diversity state prior to inoculation. To resolve this, we developed and implemented a statistical methodology that incorporates the baseline similarity between individual AbRs when testing for a convergent signal.

In this study, we first present a bioinformatic pipeline for extracting AbR information from RNA-seq data. We then go on to use FDA-based methods to characterize the Ab response of several patients to a standard TIV. Finally, we present and implement statistical tests for a convergent Ab response between patients to the same TIV. We find that a detailed time-series dataset can be used to identify Abs that are putatively targeting a vaccine, and that—after controlling for background AbR similarities—these vaccine responding Abs can exhibit similar sequence characteristics across patients.

Methods

Data creation:

The RNA-seq dataset for this study was generated by Henn *et al.* [30] [GEO:GSE45764],[46]. The experimental design was as follows: 5 patients were vaccinated with the 2010 seasonal TIV, and peripheral blood was drawn from each patient for 11 days, from day 0 (the day of the vaccination) to day 10 post vaccination. Each patient/time-point sample was divided into 2 sample-types: PBMCs, and sorted B cells. RNA-seq was performed on both the PBMC and B cell sample-types from each time point for all patients. Importantly, the two different sample-types from each sample provide relatively independent technical replicates to gauge the accuracy of our bioinformatic pipeline, described below.

For a detailed description of sample processing and RNA sequencing see [30]. Briefly, PBMCs were isolated using a discontinuous Ficoll gradient centrifugation, and B cells were enriched from heparinized whole blood with RosetteSep Immunodensity separation (Stemcell Technologies, Vancouver, BC, Canada). RNA was extracted with the Qiagen RNeasy micro kit. Barcodes were assigned to each patient/time-point/sample-type, and sequencing libraries were prepared with Illumina TruSeq RNA kits as recommended by Illumina, using 100 ng total RNA as input. The read length was 65 bases, and the mean read depth across patient/time-point/sample-types was 13,724,354.04 reads, with a range of 8,262,317-17,777,695 reads.

Computational pipeline:

State-of-the-art tools for aligning RNA-seq reads to a reference genome, such as TopHat2 [47], were not designed, and are ill-equipped, to handle the various eccentricities of Ab RNA (such as VDJ gene segment recombination, as well as the high number of mutations expected from both VDJ recombination and SHM). Similar to others [35,36], we therefore developed a bioinformatic pipeline that will harvest the Ab transcripts buried in the multitude of reads from an RNA-seq dataset. Conceptually, the pipeline consists of a negative selection step to weed out all non-Ab encoding transcripts, followed by a positive selection step to identify Ab encoding reads (Fig. 2.1A). For the negative selection step, we first created a whole genome reference sequence where all Ab encoding loci in the genome were masked out. We then used TopHat2 to map all reads to this masked-reference genome. Reads that successfully mapped to the masked genome were discarded. We hypothesized that some fraction of the unmapped reads are true Ab sequences. To identify them, we used IgBLAST [48] to positively select for Ab encoding transcripts. We used a stringent threshold (e-value ≤ 10^{-20}) to select the best aligning germline Ab gene (including V, D, and J genes). We also selected the CDR3 sequence in the alignment (if present) using a less stringent e-value threshold of 10^{-6} in order to retrieve a sufficient number of CDR3 sequences.

Overall Ab expression and V gene expression:

We would like to measure the overall level of mRNA expression of Abs, as well as the expression level of individual V genes and CDR3 sequences in each sample. By ‘expression’ we mean a quantitation of the number of mRNA sequences that map to a given Ab locus in this peripheral blood RNA-seq dataset. We first would like to estimate

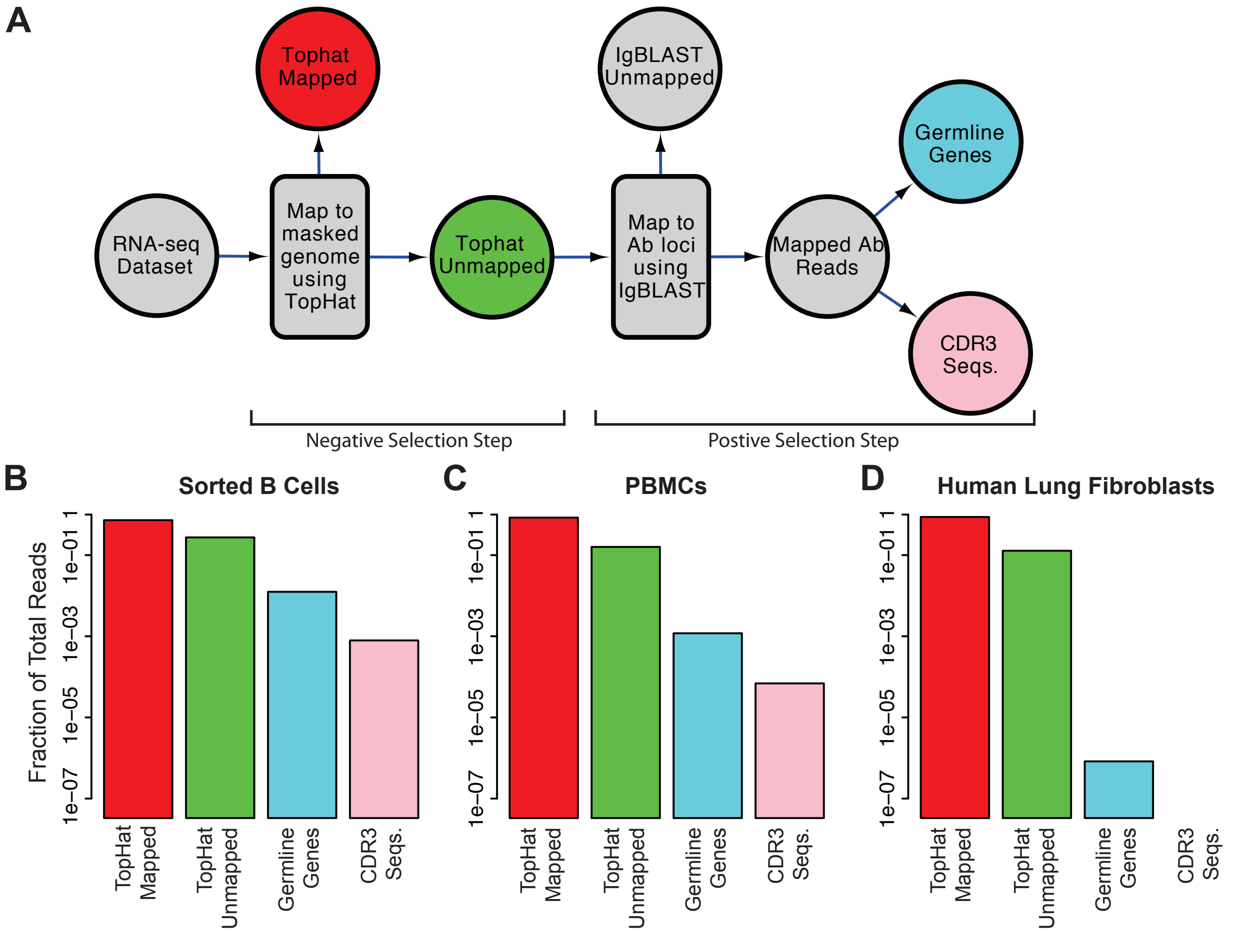


Figure 2.1. Computational pipeline.
(A) Flow diagram of the steps in our bioinformatic pipeline for harvesting Ab reads from a RNA-seq dataset. The pipeline consists of a negative selection step using TopHat2 [47] where non-Ab reads are mapped to a masked reference genome, followed by a positive selection step using IgBLAST [48] where Ab reads are mapped to reference germline Ab sequences. (B) Fraction of reads retrieved for certain steps in the pipeline, in 3 different tissues, out of the number of TopHat mapped reads (red). The colors of the bars correspond to the colors of the steps in (A).

the overall expression of Abs in each sample. To do this we counted the number of mRNA reads that mapped to any gene (V, D, or J) in the variable regions of any Ab loci (heavy, lambda, or kappa chains), and then normalized this by the number of reads that map to anything else in a given sample. We will henceforth refer to this statistic as ‘overall Ab expression’, and it was calculated as follows. Let T be the total number of days in the study, with $t \in [0, T]$, and P be the total number of patients, with $i \in [1, P]$.

For a given patient i, and time-point t, if $M_{i,t}$ is the total number of reads that map to a V, D, or J gene with an e-value ≤ 10^{-20}, and $N_{i,t}$ is the total number of reads that map to anything in the Ab-masked genome (the red circle in Fig. 1A), then overall Ab expression, $A_{i,t}$, for that patient/time-point can be calculated as,

$$A_{i,t} = \frac{M_{i,t}}{N_{i,t}+M_{i,t}}.$$

Because $M_{i,t}$ was very small relative to $N_{i,t}$, we approximated overall Ab expression as,

$$A_{i,t} = M_{i,t}/N_{i,t}.$$

It is important to note that we do not attempt to map reads to any of the constant regions of the Ab loci (IgA, IgE, IgM, IgG, etc.), so our expression level estimates are agnostic to this information. As such, overall Ab expression is a measure of the cumulative mRNA expression of all isotypes in a sample.

Next, we would like to estimate the mRNA expression level for each individual Ab gene. We achieved this by counting the total number of reads that mapped to a given Ab gene, then normalized by both the number of reads that mapped to anything else (as was done for overall Ab expression), as well as by the length of the Ab gene. We will hereto refer to this statistic as ‘gene expression’ and it was calculated as follows. Let V be the total number of unique genes that we detected belonging to a given Ab gene class (e.g., for IGHV, V=68; excluding alleles). For $v \in [1, V]$, let L_v be the length of gene v. If $m_{v,i,t}$ is the total number of reads that map to Ab gene v with an e-value ≤ 10^{-20}, then the gene expression level, $E_{v,i,t}$, of Ab gene v, in patient i at time-point t was calculated as,

$$E_{v,i,t} = \frac{m_{v,i,t}}{N_{i,t}L_v} \cdot 1000.$$

Lastly, we would like to estimate the mRNA expression level of a given CDR3 sequence. This statistic is referred to as 'CDR3 expression', and was calculated by counting the number of times the CDR3 sequence was observed in patient i at time-point t, normalized by $N_{i,t}$.

Ab diversity:

We used the CDR3 sequences in our dataset to estimate AbR diversity. We calculated the mean pairwise genetic distance (commonly referred to as π in population genetics) as our diversity statistic. However, there were different numbers of total reads sequenced for each patient/time-point, and comparing diversity estimates across differing sample sizes is problematic, as the variance of the estimate can change dramatically. To account for this we down-sampled our data until the number of reads for each patient/time-point was equal to the time-point with the least reads. We then calculated diversity from this down-sampled data. To account for possible stochastic effects of down-sampling, we analyzed the mean of 10 independently down-sampled diversity estimates.

Let $C_{i,t}$ be the total number of unique CDR3 sequences found in patient i at time-point t, with $c \in [1, C_{i,t}]$. Let $d_{i,t,c}$ be the number of times the CDR3 sequence c was found in patient i at time-point t, with $U_{i,t} = \sum_{c=1}^{C_{i,t}} d_{i,t,c}$ being the total number of CDR3 sequences detected. Additionally, let $s_{i,t}$ be a list of inferred CDR3 sequences. Antibody diversity, $\pi_{i,t}$, for patient i at time-point t was estimated as,

$$\pi_{i,t} = \frac{\sum_{j=1}^{C_{i,t}-1} \sum_{k=j+1}^{C_{i,t}} d_{i,t,j} \cdot d_{i,t,k} \cdot G(s_{i,t,j}, s_{i,t,k})}{\binom{U_{i,t}}{2}}.$$

Where $G(x, y)$ gives the genetic distance between the two CDR3 sequences, x and y. This was accomplished using the Needleman-Wunsch algorithm encoded by 'needle' in the EMBOSS package to globally align sequences x and y. We then calculated 'genetic distance' by finding the percent of mismatches in this alignment, including gaps.

In words, π can be thought of as the genetic distance that would be expected if one were to randomly pull two CDR3 sequences from a population. Thus, if there are many unique CDR3 sequences, yet only a small subset of these sequences have a high frequency, then π will be relatively low; conversely, if there are the same set of unique CDR3 sequences but their frequencies are evenly distributed, π will be relatively high.

Comparing B cell and PBMC CDR3 populations:

We used a random sampling approach to test whether or not the CDR3 sequences from B cell and PBMC sample-types were random samples from the same population. Specifically, for a given patient we randomly chose a time-point, then within this time-point, we randomly selected one CDR3 sequence from the B cell dataset and one from the PBMC dataset, where the relative frequency of the CDR3 sequences determined the probability of selection. We then calculated the genetic distance between these two sequences using $G(x, y)$, as was done in the diversity calculation. This process was done 1,000 times to create a distribution of genetic distance values. To create null distributions we repeated this workflow, except sampled pairs of CDR3 sequences from the same population. We used the Mann-Whitney U test to determine if the B cell/PBMC distribution of genetic distances was significantly different from either of the nulls. This process was done for each of the patients.

Test for identifying TIV-responding Abs:

The following method was used to identify both TIV-responding V genes and TIV-responding CDR3 sequences, so we shall henceforth use the notation ‘Ab-element’ to refer to either V gene or CDR3 sequence. For a detailed description of this FPCA based test see Wu and Wu [28], and associated R code [49]. Briefly, the test functions by first converting each of the Ab-element’s expression trajectories into a continuous function over the time-course, t, which we will call an ‘expression function’. This is accomplished by finding the linear combination of the naïve basis functions that best fit the observed Ab-element’s expression data. These expression functions, $X(t)$, can be expressed as,

$$X(t) = \mu(t) \sum_{l=1}^{b} \alpha_l \lambda_l(t).$$

Where $\mu(t)$ is a constant function that is equal to the mean Ab-element expression over the time-course, α_l is the weight given to basis function $\lambda_l(t)$, and b is the number of basis functions in the model.

FPCA is then performed on this set of expression functions. We then identified the first set of eigenfunctions that explain at least 90% of the variance in the data. Once this is done, $X'(t)$ can be re-expressed as a linear combination of this set of eigenfunctions that best fits the observed data.

$$X'(t) = \mu(t) \sum_{l=1}^{c} \xi_l \phi_l(t),$$

Where ξ_l is the weight for each eigenfunction, $\phi_l(t)$, which is often referred to as the functional principal component score, and c is the number of eigenfunctions that form the set of eigenfunctions that together explain at least 90% of the variance in the data (such that their eigenvalues are non-increasing).

Once this is done, the task is then to determine if $X'(t)$ is a better fit to the data than the null hypothesis. The null in this case is that the Ab-element's true expression function is $\mu(t)$ (where the observed deviation around the mean is due to random error). Thus, the null hypothesis is $X^0(t) = \mu(t)$. It is then determined which of the two hypotheses better fit the data by measuring the residual sum of squares (RSS) for the two models, RSS' and RSS^0. The test statistic is given by,

$$F = \frac{RSS^0 - RSS'}{RSS' + \delta}.$$

Where δ is a small constant that is meant to stabilize the variance of F, and is set to equal the variance of the Ab-elements' observed expression values around its estimated expression function. Finally, in order to produce a null distribution of the test statistic, a permutation-based approach is used. The time-points are shuffled and this process is repeated. The Ab-elements whose F statistics were significant relative to the null distribution were deemed TIV-responding. A Benjamini Hochberg correction for multiple tests was used on the p-values within a patient/gene class.

Generation of literature-curated dataset of flu-targeting Abs:

In order to characterize the diversity of Abs that have been reported to physically bind influenza, we scanned the literature and recorded the germline gene identity of all influenza-binding Abs that we found. The generation of this literature-curated dataset qualifies as a meta-analysis, so we created a separate Preferred Reporting Items for

Systematic Reviews and Meta-Analyses (PRISMA) [50] statement that explicitly addresses each item in the PRISMA checklist in order to clearly outline the criteria used to select the studies that contributed to this meta-analysis. See the PRISMA statement and PRISMA Flow Diagram in the supplemental information for details on this meta-analysis.

Test for a globally convergent V gene response:

To determine if patients tend to use similar sets of genes to target TIV, we developed a statistic, which we refer to as 'sum of gene significances' (SGS), and is defined as the number of patients in which a given gene was found to be significant. Because we have 5 patients in our data, SGS is bound between 0 and 5. We computed the SGS value for each gene, and then compared the observed SGS distribution to its null. Our task was then to generate a proper null distribution that takes into account the baseline frequencies at which the different V genes are expressed in a given patient, prior to vaccination. We chose to use a simulation-based null model, where we use day 0 gene frequencies to simulate artificial sets of TIV-responding genes.

These null simulations are best explained by example. Say the number of TIV-responding V genes for patients 1 through 5 were: 3, 6, 4, 7, and 4, respectively. The first step is to sample, without replacement, 3 genes from patient 1's day 0 distribution of gene frequencies. Here, the probability of sampling a given gene for patient 1 is equal to that gene's relative frequency at day 0. We then complete the same process in the other patients by sampling 6 genes from patient 2's day 0 gene frequency distribution; 4 genes from that of patient 3; 7 genes for patient 4; and 4 genes for patient 5. We now have 'null sets' of V genes from each patient, where the composition of these sets only reflect the gene expression levels prior to vaccination. We can then calculate SGS values for each V gene by counting the number of times each gene is present in a ‘null set’. For example if IGHV3-23 was sampled in all patients, then it would have an SGS value of 5, and if IGHV4-59 was sampled in patient 1 and patient 4 then it would have an SGS value of 2. We store these SGS values as a long list of integers. We then repeat the sampling process from each patient’s day 0 gene frequency distribution 1,000 times, and after each trial we append the resulting SGS values for each gene to our long list of integers. Once this is done we can convert this long list of SGS values into a distribution, where this distribution serves as our null, and reflects the SGS values that one might expect to get if they were to randomly sample genes from each patient prior to vaccination. We can then use a multinomial G test to compare our observed SGS distribution to the null.

To generate the ‘naïve’ null distribution we treated each patient independently and then simulated SGS statistics under this model. We did this by first estimating the probability that a gene will be significant (i.e. deemed TIV-responding) in each of the patients. This was done by dividing the number of V genes found to be TIV-responding in a patient by the total number of V genes found in that patient. Once the probability of significance was estimated in all patients, we simulated SGS values based upon these probabilities. This was accomplished by walking through each patient and randomly assigning them a ‘1’ or a ‘0’ (i.e. significant or non-significant), where the probability of getting a ‘1’ is equal to the probability of significance that was previously estimated for that patient. For example, if 3 out of 10 V genes were found to be TIV-responding in patient 1, then this patient would have a probability of 0.3 (3/10) of being assigned a ‘1’.

This assignment of either ‘0’s or ‘1’s was completed for each patient, and by taking the sum across patients we get a simulated SGS value. We then repeat this process 10,000 times to arrive at the distribution of SGS values that one might expect if the probability that a gene is significant in one patient is independent of all the other patients.

Test for convergent response in individual V genes:

This test is similar in spirit to the global test for V gene usage convergence (above), where the day 0 V gene usage is used to generate the null distribution. However, instead of a simulation-based approach to generating this null distribution we develop a closed form solution. P is again the total number of patients in the study (5 in our case), and p_i is the relative proportion of a given V gene at day 0 in the ith patient (where $i \in [1, P]$). S is the set of identifiers for each patient, so $S = \{1,2, \ldots, P\}$, and S_k is the set of all subsets of S that are of size k, so $S_k = \{x | x \subset S, |x| = k\}$, which represents all the different ways to choose k patients from S. If X is the random variable that describes the number of patients in which a given V gene is significant, then the probability of X under the null hypothesis is given by,

$$\Pr(X = x) = \sum_{y \in S_x} \left[\prod_{i \in y} Y(p_i, g_i) \prod_{j \notin y | j \in S} 1 - Y(p_j, g_j) \right].$$

Where $Y(a, b)$ is a function that gives the probability that a gene will be found to be TIV-responding in a single patient, given that that patient has a day 0 gene frequency of a, and b V genes were observed to be TIV-responding in this individual. $Y(a, b)$ is given as,

$$Y(a, b) = 1 - (1 - a)^b.$$

Essentially this can be thought of as a traditional urn problem in probability, where each patient is an urn that contains a given proportion of red balls. The probability of selecting a red ball from an urn is the probability of selecting a given V gene from a patient at day 0. The null distribution is modeled as follows: if g_i is the number of draws made from each urn i (the number of TIV-responding genes found for patient i), and p_i gives the probability of drawing a red ball from urn i (the relative frequency of the Ab gene in question at day 0), and X describes the number of urns from which red balls are drawn (the number of patients in which a particular V gene is identified as TIV-responding); then the probability of X is the null distribution for SGS.

Power simulations for global V gene convergence test:

In order to assess the statistical power of our SGS based tests for convergence, we ran simulations of the data over different parameter values to see how often the simulated data were different than the corresponding null distribution. More specifically, we simulated SGS values for each V gene, and our simulations had two parameters that were varied over a range of possibilities. These parameters were: number of truly convergent genes, and number of patients in the study. These simulations are best illustrated by example.

Say we wish to run simulations where there are 7 patients, and 2 truly convergent genes. The first step is to create ‘simulated’ patients. Here, since we already have 5 observed patients, we will only need to create 2 additional ‘simulated’ patients. For the purposes of the global V gene response convergence test, each patient needs two qualities: a distribution of day 0 gene frequencies, and a number of genes that were found to be TIV-responding for that individual. Both of these values are found by randomly selecting from the 5 existing observed patients. That is, each gene’s day 0 frequency for the simulated patient is found by randomly selecting from the day 0 frequencies for that gene of the 5 observed patients. All of the randomly selected day 0 gene frequencies in the simulated patient are then re-normalized by their sum to make them relative proportions. The number of TIV-responding genes is also randomly selected from the existing values of the observed patients. This is done independently for each simulated patient. The next step is to simulate convergent genes. Two V genes are randomly selected (regardless of their day 0 frequencies) to be truly ‘convergent’. This means that they are significant in all patients (i.e. their SGS value is set to equal 7). For each patient, the remainder of V genes are then randomly selected to be TIV-responding based upon their day 0 frequencies, until the number of genes selected for a given patient equals the total number of TIV-responding genes for that patient. For example, if patient 1 had 5 genes that were found to be TIV-responding, then 2 of these genes are set to be truly convergent (i.e. significant in all patients), and the remaining 3 are randomly drawn from patient 1’s day 0 distribution of gene frequencies, just as was done for our null distribution. Once this is completed for each patient, we have simulated SGS values for each gene, and thus can arrive at a simulated distribution of SGS values. We then compare this simulated distribution to a null distribution, which is generated the same way as described above, except no ‘truly convergent’ genes are assigned and genes are instead solely sampled based upon their day 0 frequencies. This entire process is then run 10,000 times, and power is calculated as the proportion of simulations that yield SGS distributions that are significantly different from the null distribution.

Power calculations for individual V gene convergence test:

We calculated power over a range of parameter values for the convergence test for individual genes. The parameters that we varied for this test were: number of patients in the study, and day 0 gene frequency. Because we have a closed form solution for the null distribution of this test, it is not necessary to run simulations, and we can instead calculate power directly from our equation, albeit with a few simplifying assumptions. For this test, each patient needs two qualities: a day 0 gene frequency, and number of genes found to be TIV-responding. We assume the day 0 frequency for a gene to be the same across all patients, and we set the number of significant genes for each additional patient, beyond the 5 observed, to be the nearest integer to the mean of the 5 observed values. We then plug these values into our equation, and find the probability that a gene would be found to be significant in all the patients, given a starting frequency and a given number of patients. This provides the probability of the null hypothesis, and we calculate statistical power by subtracting this value from 1.

Test for convergent CDR3 response:

To test if two patients have sets of TIV-responding CDR3s that are more similar to each other than would be expected by chance, we again utilized a methodology that hinges on sampling from the day 0 distribution. First, we calculate π (the mean pairwise genetic distance) between the two patients' observed set of TIV-responding CDR3s. If X is the set of TIV-responding CDR3 sequences in patient x, and Y is that of patient y, then $\pi_{x,y}$ between patient x and y was calculated as,

$$\pi_{x,y} = \frac{\sum_{i \in X} \sum_{j \in Y} G(X_i, Y_j)}{|X| \cdot |Y|}.$$

We then generate the null distribution for $\pi_{x,y}$ by randomly sampling (without replacement) from the population of CDR3 sequences at day 0 for both patients x and y, where the frequency of each CDR3 sequence determines the probability that it will be sampled. The number of sequences that are sampled for each patient are equal to the number of CDR3 sequences that were found to be TIV-responding for that patient. These sets of CDR3 sequences form a null set, and are solely informed by the baseline CDR3 expression level of the sequences prior to vaccination. We then calculate $\pi_{x,y}$ between the two null sets from patients x and y, and repeat this sampling process 1,000 times to get a distribution of null $\pi_{x,y}$ values. We can then assess significance of an observed $\pi_{x,y}$ value between two patients by comparing it to the respective null distribution.

Data and software availability:

Data for the immunological assays performed by [30] are available at the ImmPort repository [ImmPort:SDY224],[51]. RNA-seq data generated by [30] are available at the GEO repository [GEO:GSE45764],[46]. The anonymous patients in this study have different naming schemes in different contexts. In this study, patient 1, patient 2, patient 3, patient 4, and patient 5 equates to samples T12, T13, T14, T15, and T16 in the GEO repository; as well as equates to patient IDs S04, S06, S02, S03, and S05 in the Henn et al. study, respectively. All software associated with the analyses herein are available on the GitHub repository [https://github.com/nbstrauli/influenza_vaccination_project],[52].

Results

In this study, we implemented a pipeline to extract Ab sequences from RNA-seq data in order to take advantage of a unique densely sampled time-series dataset comprising RNA-seq data from PBMCs and sorted B cells of 5 patients vaccinated with the 2010 seasonal TIV over a time-course of 11 days [30] (Fig. 2.2).

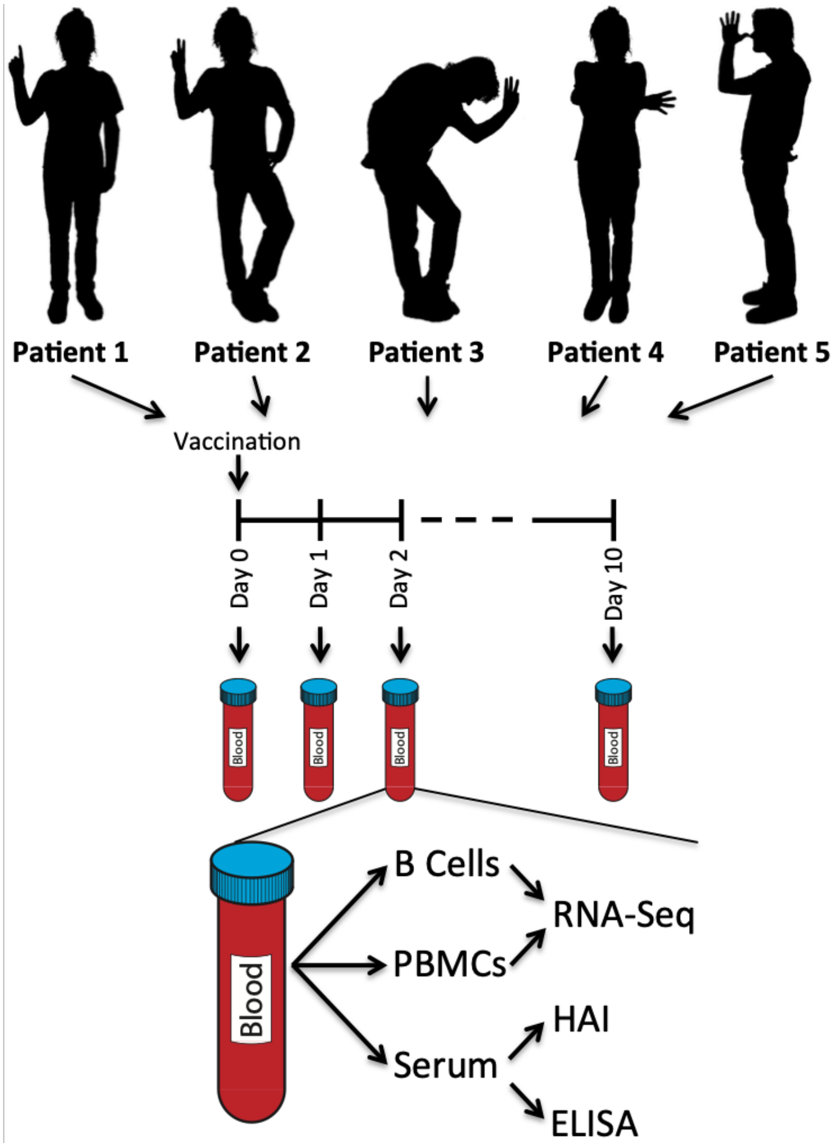


Figure 2.2. Study design.
Schematic representation of the vaccination study. There are 5 patients and each patient is given the same vaccine. Whole blood is drawn immediately prior to vaccination, and each day for 10 days post vaccination (11 time-points total). Each patient/time-point sample is dived into three different sample types: B cells, PBMCs, and serum. The B cells and PBMCs are used for RNA-seq. The serum is used for immunological assays. This figure shows the same information as Figure S1 from Henn *et al.* [29].

*Patient avatars in no way reflect actual patients.

We use the high-resolution temporal information in these data in order to infer the elements of the AbR that are putatively responding to TIV. We then go on to test if the patients in this dataset exhibit more similar responses to TIV than would be expected by chance. That is, we test if these distinct AbRs exhibit convergence in response to the same vaccine.

Quality control of bioinformatic pipeline:

First, we validated that our bioinformatic pipeline (Fig. 2.1A, see methods for a detailed description) extracts meaningful AbR information from RNA-seq data. We hypothesized that the proportion of Ab encoding reads detected should correlate with the expected number of B cells in a given sample-type. We arbitrarily chose the day 7 time-point from patient 1, and applied our pipeline to the RNA-seq data from isolated B cells and PBMCs for this patient/time-point. As a negative control, we also applied our pipeline to RNA-seq data from human tissue cultured lung fibroblasts [53]. Our expectation was that the number of Ab sequences would decrease from B cells to PBMCs, and cultured lung fibroblasts would serve as a negative control with essentially no Ab sequences. Consistent with our expectation, we found that 1.25% of all reads from isolated B cells encode Ab (206,797 of 1.7×10^{7} total reads), PBMCs yielded 0.12% Ab encoding reads (16,214 of 1.4×10^{7} total reads), and cultured lung fibroblasts produced <0.001% Ab encoding reads (25 of 3.0×10^{7} total reads) (Fig. 2.1B).

Broad AbR characteristics:

We next sought to characterize how the AbR broadly behaves in response to TIV. To this end we measured overall Ab expression as the number of Ab mapped reads normalized by the number of non-Ab mapped reads, see methods (Fig. 2.3A).

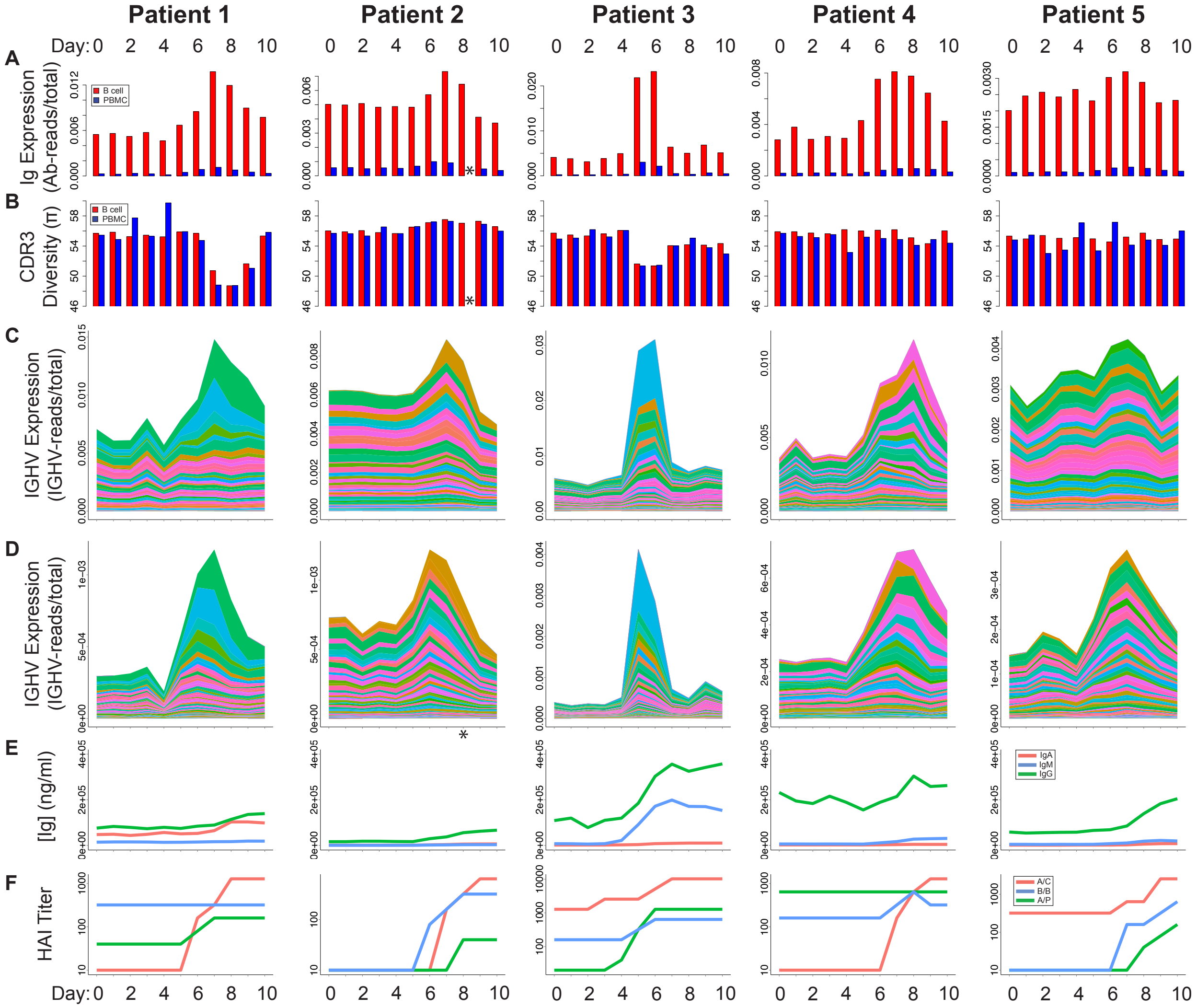


Figure 2.3. AbR response to TIV across patients.
Different metrics were measured for each patient and at each time-point. Metrics are delineated by row, and patients are delineated by column. (A) Overall Ab expression for each patient/time-point. (B) CDR3 diversity for each patient/time-point. B cells and PBMCs are shown in red and blue, respectively. (C and D) Stacked area charts showing the gene expression level for each IGHV gene for each patient/time-point. Colors, corresponding to IGHV genes, are comparable between patients and sample-types, and were sorted by absolute range (max – min). (C) B cell and (D) PBMC data. Complete definitions for the y-axis units in (A,B,C, and D) can be found in methods. (E) ELISA results giving the concentration of Abs that bind TIV for the A, M, and G Ab isotypes (red, blue, and green, respectively). (F) Hemagglutinin inhibition assay results for the three different virus stains in the administered TIV, A/C: A/California/7/2009; B/B: B/Brisbane/60/2008; A/P: A/Perth/16/2009. Data for (E) and (F) were generated by [30], and downloaded from ImmPort [51], [ImmPort:SDY224].

Ab diversity was measured as mean pairwise CDR3 genetic distance (see methods) in each of the patients over the time-course (Fig. 2.3B). We found that each patient had a characteristic peak in overall Ab expression around day 7, although the timing and severity of this peak varied dramatically across patients. Patient 3 had the most dramatic response, which had entirely subsided by day 7, while the response for patient 5 was much more gradual and less pronounced. We note that patient 3 was the only patient to have received the seasonal influenza vaccine for each of the prior 3 years, and received an additional monovalent vaccine the year prior to the study [30]. The monovalent, and seasonal 2009 vaccines had epitopes from two of the strains that were included in the TIV used in this study (Table 2.1).

Table 2.1. Patient History.
Top portion of table lists the demographics and vaccination history for each of the patients. If a patient has an 'X' under one of the vaccinations it means that that individual received the vaccine; a blank means that they did not. The lower portion of the table lists the strains used in each of the vaccines. Entries with thick borders highlight the strains that were used in the study TIV (Seasonal 2010). This table shows the same information as Table S1 from Henn *et al.* [30].

Patient Vaccine Histroy:							
Patient ID	**Sex**	**Age**	**Seasonal 2007**	**Seasonal 2008**	**Seasonal 2009**	**Monovalent 2009**	**Seasonal 2010**
Patient 1	F	47		X	X		
Patient 2	F	36					
Patient 3	F	20	X	X	X	X	
Patient 4	M	27		X	X		
Patient 5	M	23					
		Strains Used in Vaccines:					
		A/H1N1	A/Solomon Isl. /3/2006	A/Brisbane/59 /2007	A/Brisbane/59 /2007	A/California/7 /2009	A/California/7 /2009
		A/H3N2	A/Wisconsin/67 /2005	A/Brisbane/10 /2007	A/Brisbane/10 /2007		A/Victoria/210 /2009
		B	B/Malaysia/2506 /2004	B/Florida/4 /2006	B/Brisbane/60 /2008		B/Brisbane/60 /2008

These results are consistent across both the B cell and the PBMC RNA-seq sample-types. Indeed, overall Ab expression and diversity levels for each of the patients and time-points are highly correlated between the two sample-types (overall Ab expression: Kendall's tau = 0.639, p = 8.715e-12, Ab diversity: Kendall's tau = 0.366, p

= 8.083e-05; Fig. 2.4), suggesting that the overall signal represents the underlying AbR diversity and expression patterns.

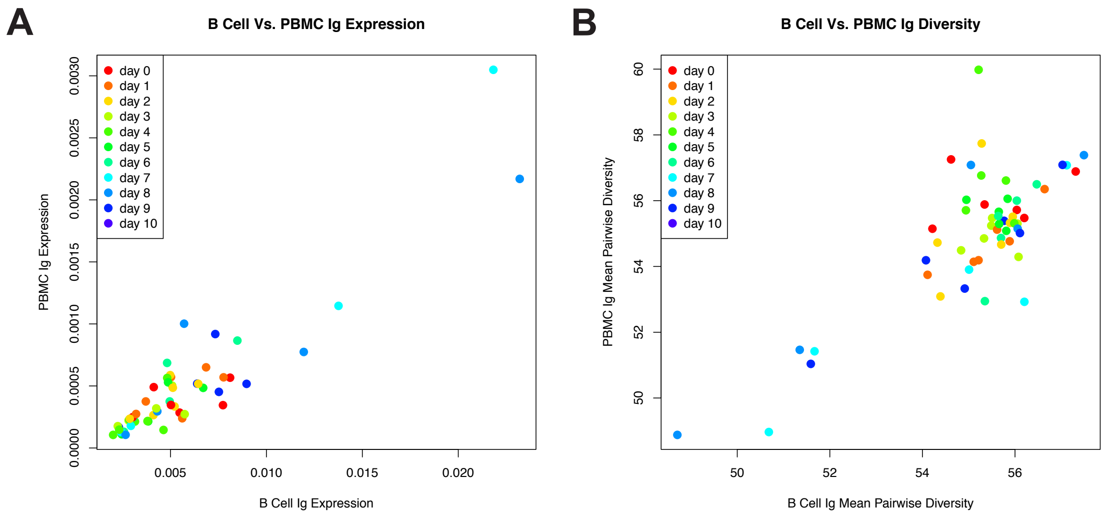


Figure 2.4. Correlation between B cell and PBMC AbR data.
(A) Scatter plot showing the correlation between overall Ab expression in the B cell data to that of the PBMC data for each patient/time-point. Points are colored by time-point. (B) Same as (A) but showing correlation of mean pairwise diversity of CDR3 sequences between the B cell and PBMC data. See methods for how the diversity statistic was calculated.

<u>Comparing B cell and PBMC CDR3 populations:</u>

Because B cells are a subset of PBMCs, we can expect that the RNA-seq data from these two sample-types should yield similar Ab sequences. By checking to see if this is indeed the case in our data, we have another means to check the accuracy of our pipeline. In order to quantify the similarities between the B cell and PBMC sample-types, we focused on CDR3 sequence sets. Specifically, we statistically tested whether the CDR3 sequences from the B cell and PBMC datasets are drawn from the same population. We did this by finding the distribution of genetic distances between CDR3 sequences from different sample-types, and compared this to the same distribution from CDR3s in the same sample type (see methods). We found that none of these three distributions are significantly different in any of the patients (p > 0.07, see Fig. 2.5). We thus conclude that PBMC and B cell datasets can reliably be used to extract Ab sequences from RNA-seq data.

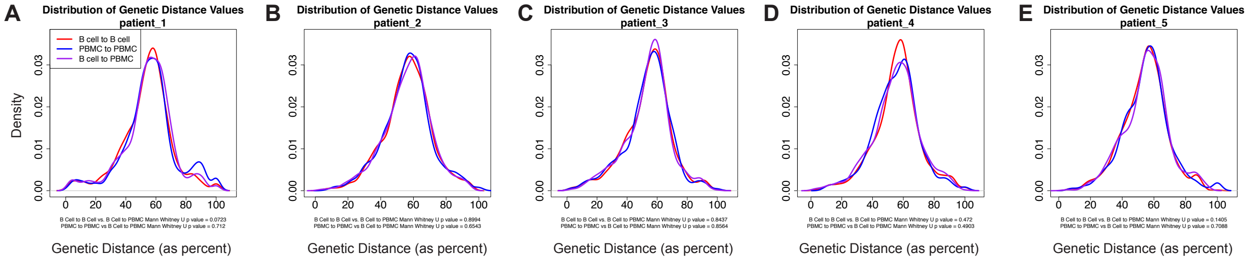


Figure 2.5. Comparing the B cell AbR to the PBMC AbR.
Density plots showing the distribution of genetic distance values for randomly selected CDR3 sequences. Randomly selected sequences were matched by time-point. Red lines show the resulting distribution when comparing CDR3 sequences within the B cell data, blue lines show the same when using the PBMC data, and purple lines show the genetic distance distribution when comparing between datasets. Subtitle below the plots list the Mann Whitney U p-values when comparing the red and the blue distributions to the purple. (A) Patient 1. (B) Patient 2. (C) Patient 3. (D) Patient 4. (E) Patient 5.

V gene and CDR3 usage analysis:

We next sought to analyze how each Ab gene is expressed over the time-course after vaccine administration. We calculated the mRNA expression level of each gene (as number of reads mapped to a given Ab gene normalized by the number of non-Ab mapped reads, see methods) in each of the patients, and at each time-point. We analyzed each class of Ab gene that could produce reliable alignments: V gene heavy chain (IGHV), V gene lambda light chain (IGLV), and V gene kappa light chain (IGKV). We were unable to detect an appreciable number of reads aligning to D or J genes with high confidence, which is likely due to their short lengths. We then generated stacked area charts to observe how the cumulative and individual V gene expression changes over time (Figs. 2.3C, and 2.3D for IGHV; Fig. 2.6 for IGLV and IGKV). We find that the patients with the most dramatic Ab response (patients 1 and 3) also seem to have the largest gene expression increases in very few V genes, and that the increase in these few genes seem to explain a large portion of their rise in overall Ab expression. This is particularly well illustrated in patient 1, where the peak in overall Ab expression is

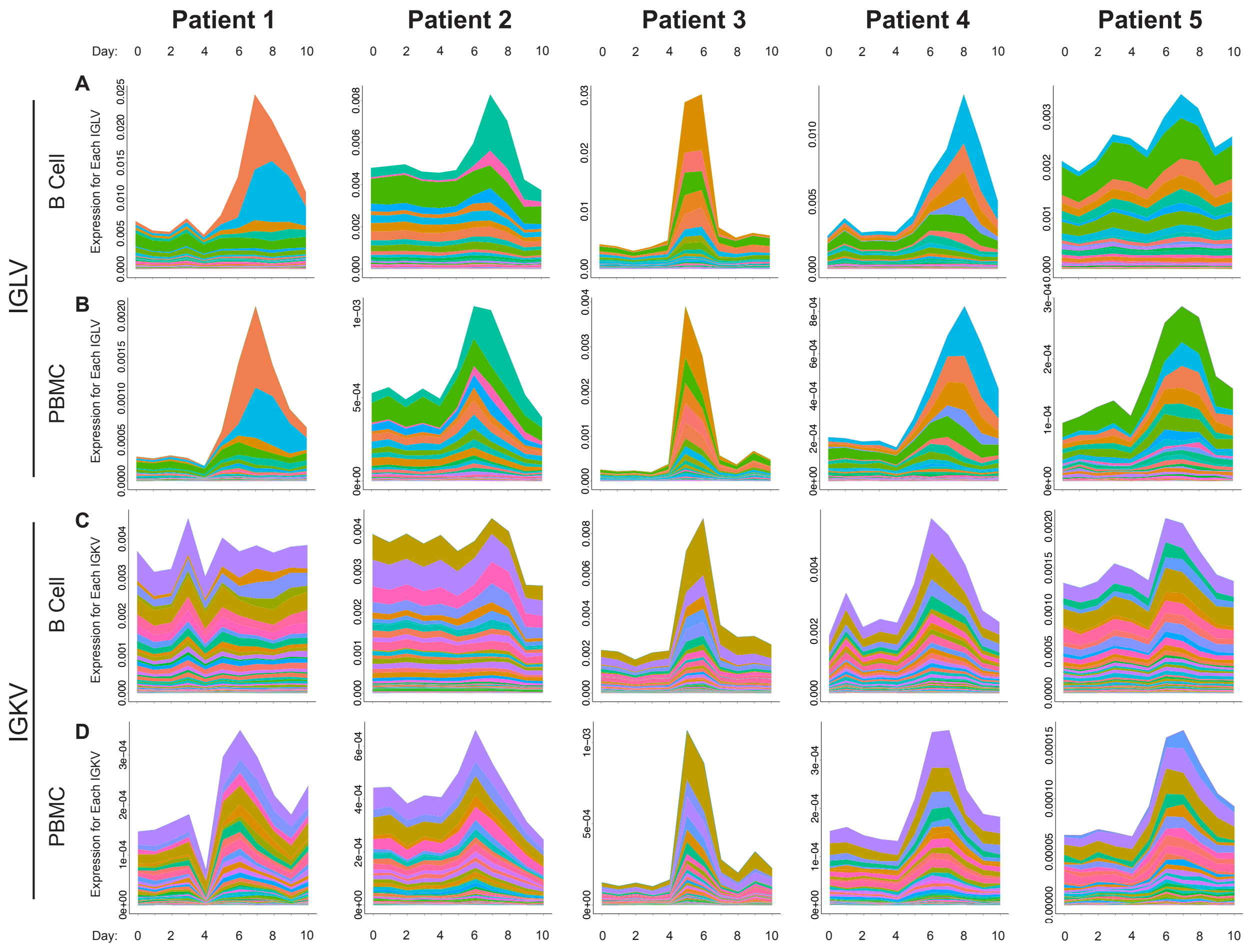


Figure 2.6. Lambda and kappa Ab gene's expression over time.
Stacked area charts showing the cumulative as well as individual Ab gene expression over time for both the lambda and kappa chains. (A and B) Stacked area charts for IGLV genes from the B cell and PBMC data, respectively. (C and D) Same a (A and B) but for the IGKV genes. All distinct colors for plots of IGLV genes correspond to the same genes (i.e. colors are comparable across patients and sample-types). The same is true for plots of IGKV genes.

entirely explained by an increase in gene expression of 2-3 V genes. Moreover, this expression increase coincides with a dip in CDR3 diversity. Together, this suggests that patients 1 and 3 had largely a monoclonal response to TIV. We also note that the other patients showed signs of a predominantly polyclonal Ab response that did not substantially affect diversity. Though we cannot draw strong conclusions about the causes of a polyclonal or monoclonal response in this small sample size, future studies of larger cohorts could elucidate the causes behind this heterogeneity.

We performed an analogous analysis using CDR3 sequence data. We gathered all unique CDR3 sequences for each patient/time-point sample and calculated their mRNA expression level (see methods). We again generated stacked area charts to observe how the predominant sequences change in CDR3 expression over time (Fig. 2.7). We found that these data largely recapitulate the gene expression data, where CDR3 expression expansions tend to occur around the same time as the increases in V gene expression. Patient 1 again shows a dramatic expansion in the expression of a single CDR3 sequence, providing further support for a largely monoclonal response.

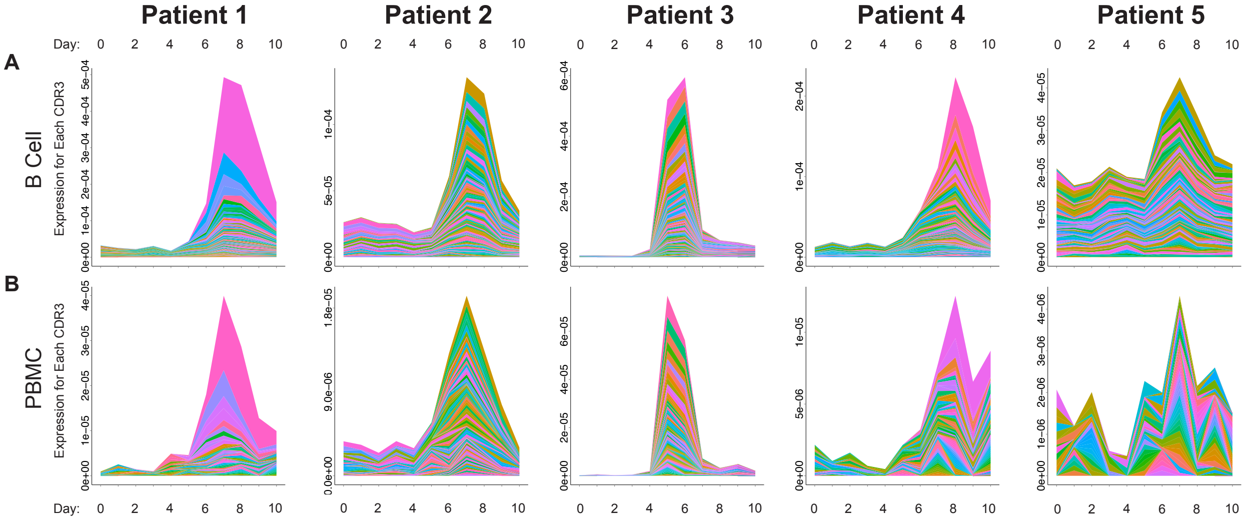


Figure 2.7. CDR3 expression over time.
Stacked area charts showing the cumulative as well as individual expression level for the 100 most frequent CDR3 sequences in the data. (A) B cell data. (B) PBMC data.

There are two factors that contribute to an Ab’s mRNA expression level: the number of B cells harboring the Ab, and the rate of Ab expression for each of these B cells. Because RNA-seq was performed on a heterogeneous population of B cells in the peripheral blood, we cannot distinguish between the two. Further, these two factors are highly dynamic over time, where B cells are constantly migrating in and out of the peripheral blood, in addition to dramatically varying their rate of Ab transcription. Thus, the population of B cells that we sample on day 7 is likely very different from that of day

0. However, whether due to a clonal expansion or an increase in transcription rate, if an Ab gene or CDR3 sequence increases in expression level, it is largely indicative that at least a subset of the B cells harboring this gene or CDR3 are responding to some antigenic stimulus.

Immunological assays:

Given the robust signal in our V gene and CDR3 usage analyses, we sought to validate that the expansions we observed in our data were indeed in response to TIV. Henn *et al.* [30] performed a variety of immunological assays using the sera from each patient/time-point sample. We downloaded these data from [51], [ImmPort:SDY224] to determine if the patients gain immunological reactivity against influenza around the same time as the V gene and CDR3 expression level expansions occur in our data. The results show that vaccine-binding immunoglobin tended to increase around the same time as V gene and CDR3 expansions (Fig. 2.3E). We next sought to establish that the V gene and CDR3 response conferred protectivity against influenza virus. Data from hemagglutinin inhibition (HAI) analyses using the three strains of influenza virus included in the TIV showed that protection to at least one of the strains was gained around the same time as the spike in V gene and CDR3 expression levels (Fig. 2.3F). Together these data suggest that the V gene and CDR3 expression level expansions we observe in our data are direct immunological responses to TIV.

Identifying TIV-responding V genes:

Given the robust signal that we saw in the V gene expression time-course data, we next established a methodology to systematically identify the V genes that appear to be responding to TIV. We utilized a method based on functional principal component analysis (FPCA), which was designed to identify differentially expressed genes over a time-course [28] (see methods for description). We found that it was often the case that the 1st eigenfunction explained over 90% of the variance in the data (Fig. 2.8A and 2.8C).

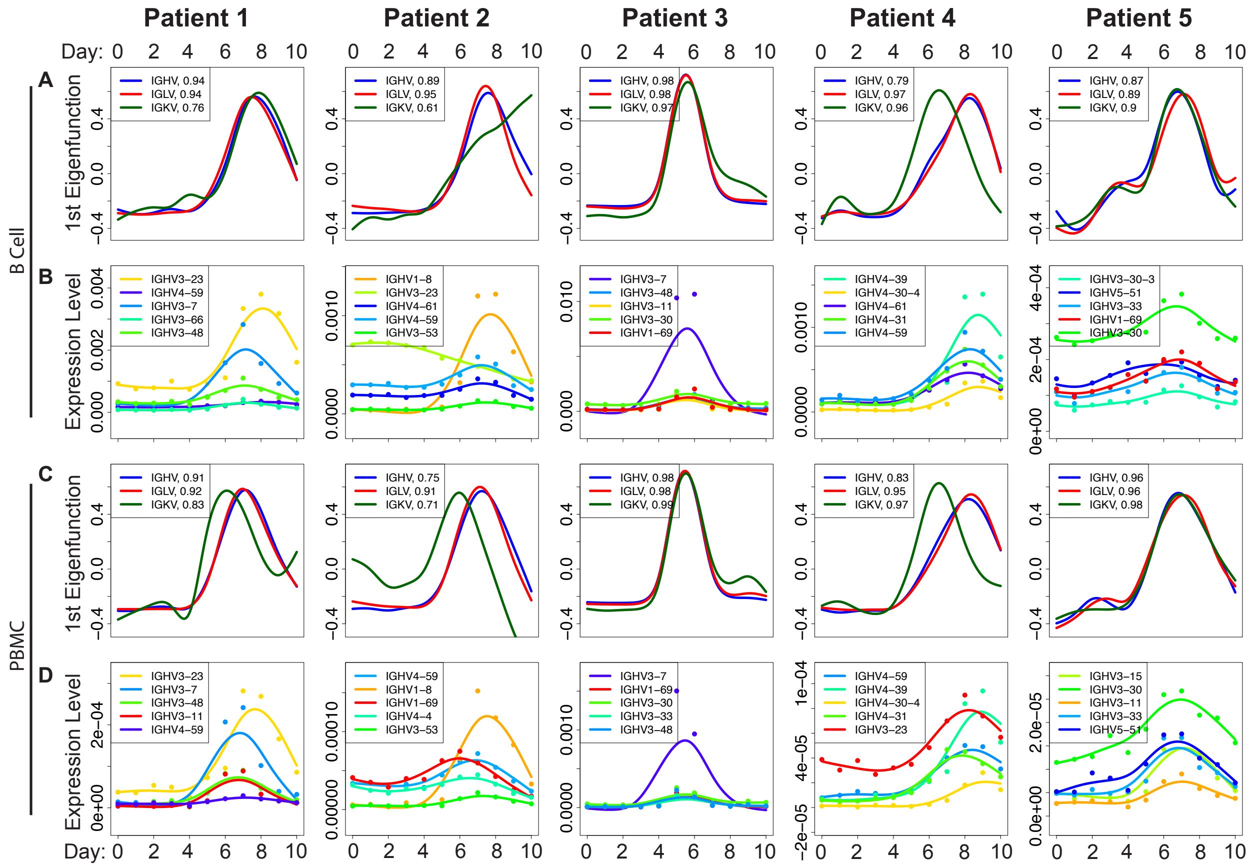


Figure 2.8. Identifying putative TIV-responding V genes
(A) First eigenfunction in the B cell data for each patient and each V gene class. The proportion of the total variance explained by the first eigenfunction is listed in the legend after each respective class of V gene. (B) The top five scoring IGHV genes from the FPCA based test to identify TIV-responding V genes; in the B cell data. The points show the observed data, and the solid lines show the best fitting gene expression function over time. V genes in legend are ordered by p-value, with lowest on top. P-values are based on a permutation test, see Wu and Wu [28] for details. (C and D) Same as (A and B) but from the PBMC data. Colors corresponding to IGHV genes in (B and D) are comparable within a patient. Eigenfunction plots (A and C) were generated using the “eigens” output from the FPCA-based test. IGHV gene expression functions (B and D) were plotted using the “fda” package for the R programming language, and using a smoothing parameter, “lambda” = 0.66.

From this method we were able to identify the genes that seem to be most dramatically responding to TIV (Fig. 2.8B and 2.8D; Table 1). In almost all patients, the top genes identified in the B cell dataset (Fig. 2.8B) are replicated in the PBMC data set (Fig.

2.8D). We deem the V genes identified by this test to be ‘TIV-responding’. It is important to note that while the results of this test provide evidence that these genes are ‘responding’ to TIV, functional validation is required to establish that they actually target TIV. We then assessed whether or not these sets of TIV-responding genes, are more similar across patients then would be expected by chance.

Table 2.2. Top 10 TIV-responding heavy chain V genes.
Lists the top ten scoring IGHV genes in the FPCA-based test for the B cell data. “Lit. Ab Freq.” lists the frequency for each of the V genes in the literature-curated dataset. “Combined” lists the p-values for each of the V genes after using Fisher’s method to combine the p-values from the FPCA-based test across all the patients. Patient 1-5 lists the p-values for each of the individual patients from the FPCA-based test. Genes are sorted by combined p-value.

Gene Name	Lit Ab Freq.	Combined	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5
IGHV3-23	0.1034483	1.08E-14	1.54E-05	4.55E-05	0.0052969	0.000108	0.000154
IGHV1-69	0.237069	5.04E-14	0.0004328	0.002909	0.0002813	6.15E-05	1.54E-05
IGHV3-30	0.0387931	7.19E-13	0.0015373	0.001545	0.00025	0.000692	1.54E-05
IGHV3-7	0.0431034	1.15E-12	0.000209	0.007924	1.54E-05	0.000108	0.003892
IGHV4-61	0.0021552	1.16E-12	0.0013284	7.58E-05	0.0283438	3.08E-05	0.000123
IGHV4-31	0.0043103	1.78E-12	0.0011045	0.004242	0.0015625	3.08E-05	7.69E-05
IGHV3-48	0.0193966	2.36E-12	0.0003284	0.000409	0.0001875	0.000969	0.000969
IGHV4-59	0.1530172	2.77E-12	0.0001194	0.000136	0.0192969	4.62E-05	0.001954
IGHV3-11	0.0043103	3.14E-12	0.0003731	0.000742	0.00025	0.000154	0.003062
IGHV3-30-3	0	4.71E-11	0.002194	0.011303	0.0015313	0.001154	1.54E-05

Testing for a convergent V gene response:

Recently there have been several studies that have reported independent AbRs showing signals of sequence convergence when challenged with a similar antigenic stimulus [15,21,41,42]. This suggests that independent patients may use the same V genes to target similar antigens. Consistent with this, we found that V genes tend to have similar FPCA-based test scores across patients (Fig. 2.9). This suggests that the patients in our data are indeed using similar V genes to target TIV.

However, the sets of TIV-responding V genes were statistically inferred using time-series data, and were not shown to physically bind TIV. To validate these findings, we searched the literature for Abs that have been experimentally shown to bind either

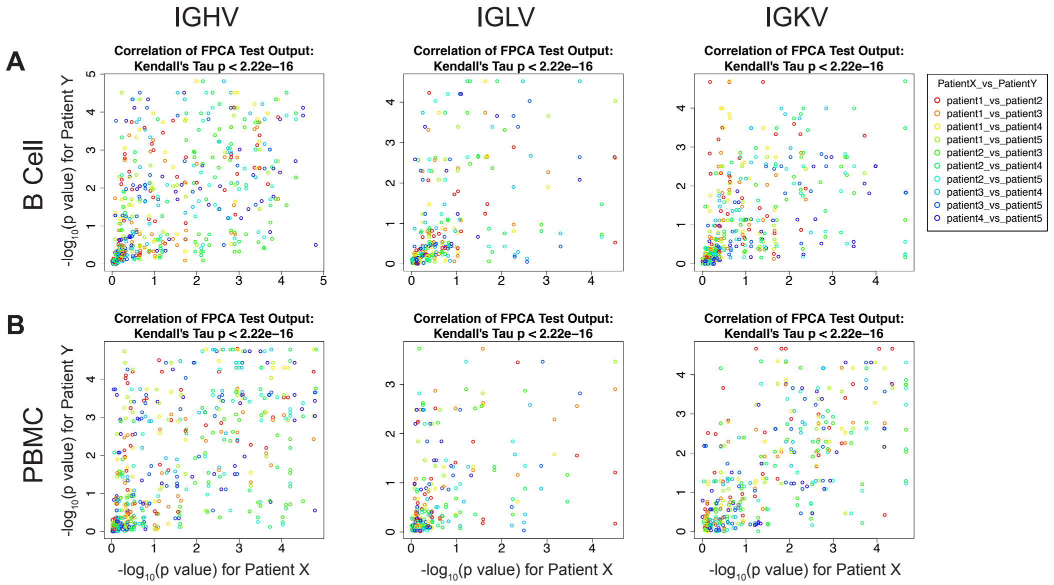


Figure 2.9. Correlation of TIV-responding V genes across patients.
Comparison of FPCA based test p values between all pairwise patients, for each V gene. For each pairwise patient comparison, these plots show the correlation of the p values from the FPCA based test for each of the genes. Points are colored by patient comparison. Correlation p value (Kendall's Tau) is listed in the title of each plot. (A) Scatter plots for IGHV, IGKV, and IGLV genes' p values for the B cell data. (B) Same as (A) but for the PBMC data.

an influenza vaccine or the influenza virus itself. Since most publications do not provide sequence information for the Abs they find, our analysis is limited to the germline genes from which the Abs originated. Our search resulted in 464 Abs that have been shown to bind influenza vaccine or influenza virus (Tables 2.3 and S3).

Table 2.3. Contributions of each publication to the literature curated flu-targeting Ab dataset.
Lists all the publications that contributed to the literature-curated dataset. An 'X' under a given class of Ab genes means that that gene class was reported in the publication. If an entry is blank then the gene class was absent. "Number Abs Retrieved" lists the total number of monoclonal Abs that each publication contributed to the literature-curated dataset.

Publication	IGHV	IGHD	IGHJ	IGLV	IGLJ	IGKV	IGKJ	Number Abs Retrieved
Wrammert, 2011	X	X	X	X	X	X	X	46
Krause, 2011	X	X	X	X	X	X	X	5
Moody, 2011	X		X	X	X	X	X	160
Corti, 2010	X	X	X	X	X	X	X	20
Krause, 2012	X	X	X	X	X	X	X	5
Yu, 2008	X	X	X	X	X	X	X	5
Throsby, 2008	X	X	X					223

We then compared the TIV-responding V genes identified by our FPCA based test to the frequency of each V gene from our literature-curated dataset. Specifically, since each patient is approximately independent, we used Fisher's method to combine FPCA-based p-values across patients. This results in a single p-value for each gene, where significance is increased if a gene is inferred to target TIV in multiple patients. Conversely, significance is diminished if a gene is heterogeneous across patients (Table S4).

We found that these combined p-values are correlated with IGHV gene frequency in the literature-curated dataset (Kendall's Tau, B cell p = 3.115e-5, PBMC p = 2.502e-5). Moreover, we find that ~60% of all Abs in our literature-curated dataset were composed of V genes that were inferred to be TIV-responding across all patients in our analysis (Fig. 2.10A).

In particular, we find that the genes IGHV1-69 and IGHV3-7, which have been shown to consistently target influenza epitopes in several independent studies [41,54–58] have the 2nd and 4th lowest p-values in the B cell data (Table 2.2), and 1st and 2nd lowest p-values in the PBMC data (Table S4), respectively. One of the publications that contributed to our literature-curated dataset used a combinatorial phage display library to select for influenza-targeting Abs (Throsby *et al.* [54]). This is different from the in-vivo selection process that occurs in humans, and thus could introduce unknown bias in the Abs from this study. We removed the data from this study and saw no qualitative difference in the outcome (Fig. 2.11). Together, these data show that (i) the V genes that we identify as TIV-responding with our pipeline are consistent with previous findings in the literature, and (ii) that the patients from the Henn *et al.* dataset, as well as those from several other studies, tend to use similar V genes to target the influenza vaccine.

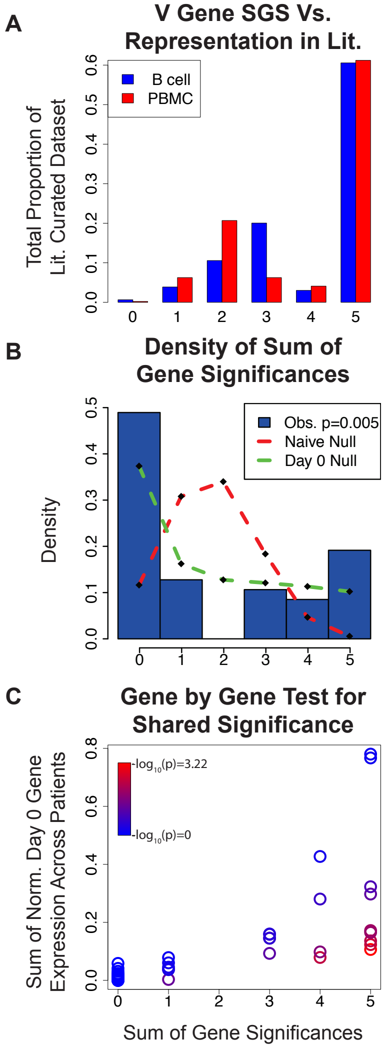


Figure 2.10. Identifying a convergent V gene usage signal across patients.
The x-axis for all plots is the sum of gene significances (SGS) statistic, which is the number of patients for which a given Ab gene was found to be significant. (A) Comparing our results for IGHV to the literature. For each SGS bin, this shows the proportion of the Abs in the literature-curated data that have V genes belonging to this bin. ~60% of the influenza binding Abs in the literature-curated dataset were composed of V genes that had an SGS value of 5. (B) Comparing observed SGS to the null distribution. Blue bars are a histogram showing the observed proportion of IGKV genes from the PBMC data belonging to each SGS bin. Red dashed line shows the 'naïve' null distribution of SGS if each patient were independent from one another (see methods). Green dashed line shows the null distribution of SGS if the baseline similarity in gene expression at day 0 is taken into account. The p-value in the legend shows the result of using a multinomial G test to compare the observed SGS distribution to that of the day 0 null. (C) Comparing the SGS value for each IGKV gene from the PBMC data to that of their respective null expectations. Color indicates the probability of the observed SGS under the null model (p-value, see methods for explanation of null model).

There are two reasonable explanations for this observation. The first is that some V genes have properties that make them naturally better at targeting TIV than others and are thus more likely to show a response across patients. The second is that

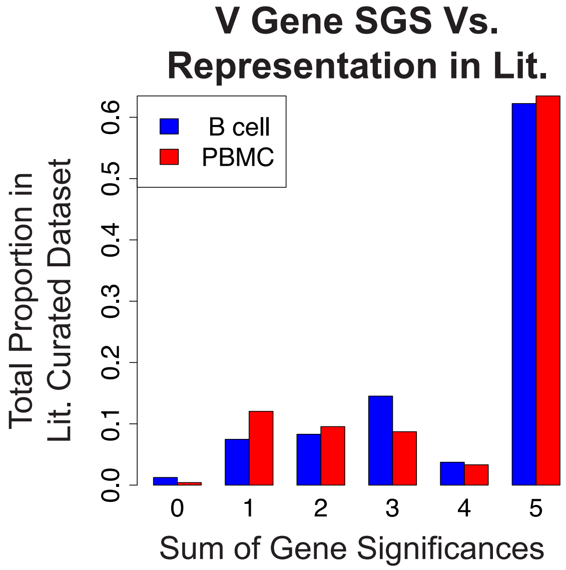


Figure 2.11. Comparison of SGS to literature-curated dataset, excluding Throsby *et al.*
For each SGS bin, this shows the proportion of the Abs in the literature-curated data that have V genes belonging to this bin. The genes that were significant in all 5 patients represented the largest proportion of the genes shown to be influenza binding in the literature.

patients tend to have similar V gene expression levels prior to vaccination, such that the Abs that are selected to respond to TIV tend to have similar V genes across patients simply due to this prior baseline similarity. We argue that this latter explanation has been underappreciated, and thus merits further scrutiny.

Suppose V gene expression levels are correlated across patients, independent of any antigenic stimulus. If Ab lineages were randomly selected to respond to an antigenic stimulus (the null expectation), then we would expect to see similar V genes responding to said antigenic stimulus across patients purely due to the underlying correlation of V gene expression prior to inoculation. We tested for correlations in V gene expression levels prior to vaccination (day 0), and found that they are highly correlated across patients (Fig. 2.12). We therefore developed a statistical test that will take into account the underlying similarity in V gene expression prior to vaccination when determining if the patients in our data tend to use more similar sets of V genes to respond to TIV than would be expected by chance (see methods). For each gene we

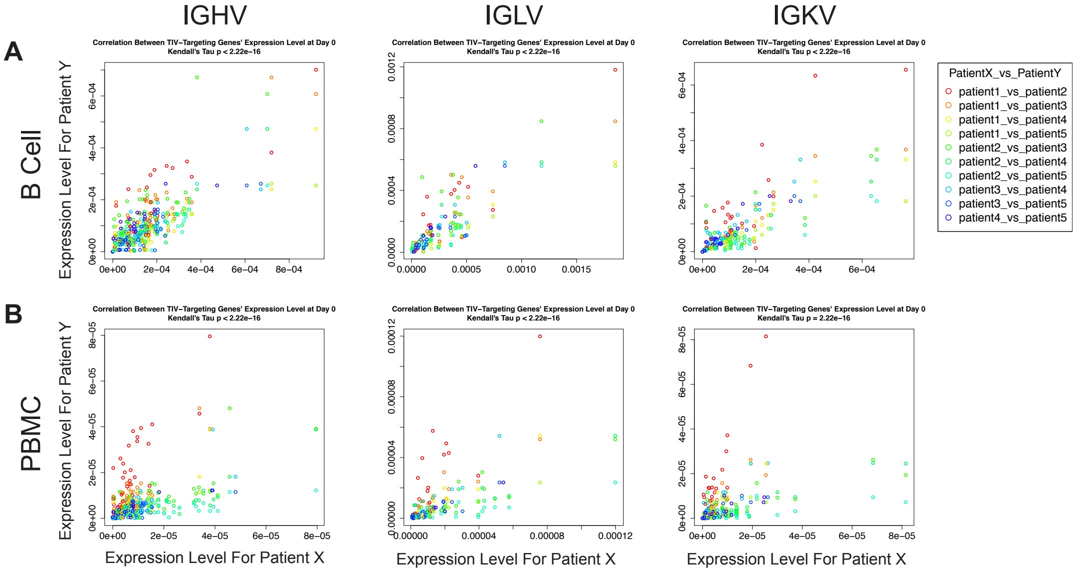


Figure 2.12. Baseline correlation of Ab gene expression at day 0.
Comparison of day 0 expression level between all pairwise patients, for each TIV-responding V gene. For each pairwise patient comparison, these plots show the correlation of the day 0 expression level for each of the TIV-responding V genes. (A) Scatter plots for IGHV, IGKV, and IGLV TIV-responding genes' day 0 expression level for the B cell data. (B) Same as (A) but for the PBMC data.

find the number of patients in which the gene is found to be significant by our FPCA test (referred to as Sum of Gene Significances, or SGS, Tables S5 and S6). We then compare the observed SGS distribution to a null. We found that the observed SGS distribution was significantly different than the null for IGKV from the PBMC dataset (multinomial G-test, p=0.005; Fig. 2.10B; dashed green line vs. histogram) and we saw no evidence for convergent gene usage for other classes of V genes (Fig. 2.13). We then assessed the possibility that this convergent signal was driven by outlier genes that were deemed significant by the FPCA-based test, but do not have expression trajectories representative of a vaccine response (e.g. IGHV3-23 in patient 2, Fig. 2.8B). We performed a rather extensive outlier removal analysis to address this, where we removed these outliers in a variety of different ways (see Appendix for a detailed description). In short, our convergent signal for IGKV was robust to all outlier removal approaches.

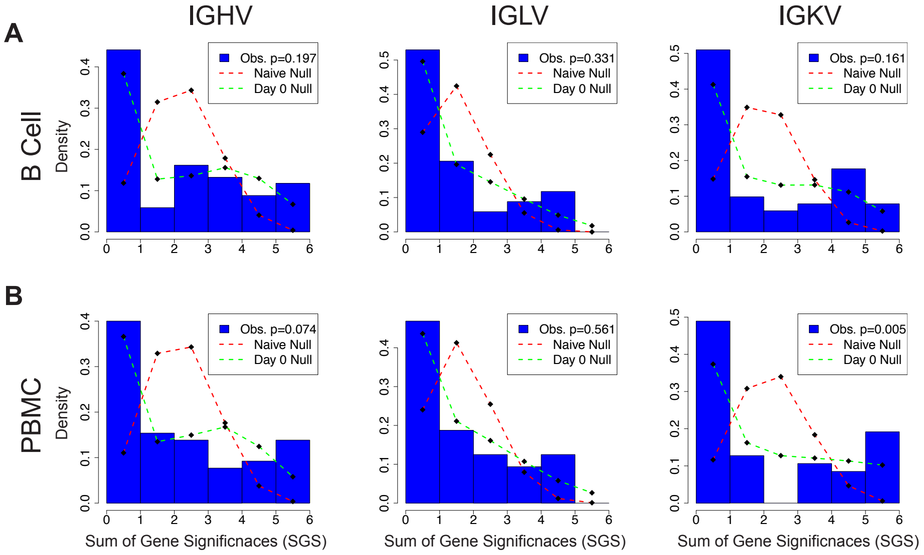


Figure 2.13. Testing for a global gene usage convergent signal.
Comparing observed SGS to the null distribution. Blue bars are histograms showing the observed proportion of V genes belonging to each SGS bin. Red dashed lines show the null distribution of SGS if each patient were independent from one another. Green dashed lines show the null distribution of SGS if the baseline similarity in gene expression at day 0 is taken into account. The p values in the legends show the result of using a multinomial G test to compare the observed SGS distributions to that of the day 0 nulls. (A) Histograms of the IGHV, IGKV, and IGLV genes, for the B cell data. (B) Same as (A) but for the PBMC data.

Given the mixed evidence for a global convergent signal in V gene response to TIV, we investigated each V gene individually (i.e., we test whether a given V gene was found to be TIV-responding in more patients than expected by chance). Similar to our global V gene analysis, we used the gene frequencies at day 0 to construct our null distribution (the null was solved in closed-form, as opposed to simulating; see methods). We found that two V genes showed a significant convergent signal after Bonferroni correction for multiple testing. These were IGHV3-66 on the heavy chain and IGKV3-NL1 on the kappa light chain, using the PBMC data (Tables S7 and S8). In general, these significant V genes had a characteristic expression level trajectory of low expression prior to vaccination, and then increasing in expression post-vaccination. This character of trajectory made it unlikely that the V genes would be selected to respond to the vaccine simply because they were abundant (or highly expressed) prior to vaccination, yet their increase in expression level after vaccination makes them likely candidates for responding to the vaccine.

To our knowledge, neither IGHV3-66 or IGKV3-NL1 have been reported to have shown a convergent response to TIV before, and are absent from our literature-curated dataset. Conversely, the V genes IGHV1-69 and IGHV3-7 — which have been reported in the past as showing a convergent signal when targeting TIV — are not significant in our test. This means that we cannot reject the possibility that these genes were found to be consistently targeting TIV simply due to their tendency to be highly expressed prior to vaccination. While it's possible that the reason for this initial high gene expression is because of prior convergences due to a similar antigenic history, it is also possible that these V genes are highly expressed independent of any antigenic history. We cannot differentiate between these two possibilities, so this baseline correlation must be corrected for.

Together, the results from our tests for a convergent signal in V gene usage show that some patients tend to use similar sets of V genes for particular gene classes, and that a couple of these V genes stand out. While only a subset of our tests yielded a significant convergent signal, we found it notable that there was any convergent signal at all, given the strong baseline correlations across patients prior to vaccination.

Testing for a convergent CDR3 response:

We hypothesized that if the patients within this dataset are capable of using similar sets of V genes to target the same vaccine, then they may use similar sets of CDR3 sequences to target TIV as well. To answer this, we began by again using the FPCA-based test on our CDR3 expression data to identify the putative TIV-responding

CDR3 sequences. We were then left with a list of CDR3 sequences for each patient that appear to be responding to TIV. Our task was then to determine if these lists of TIV-responding CDR3s were more similar between patients than would be expected by chance (see methods). We found that patients 1 and 4 seem to have converged on similar CDR3 sequences to target TIV, whereas patients 2 and 3, and patients 1 and 3 seem to have diverged (Figs. 2.14 and 2.15).

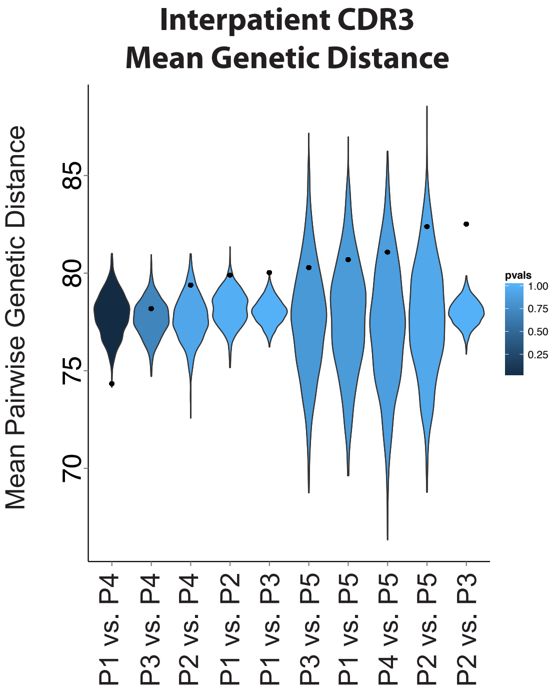


Figure 2.14. Testing for convergent CDR3 sequences across patients.
Black points indicate the observed mean genetic distance between each pair of patients for the TIV-responding CDR3 sequences in the B cell data. Violin plots show the null distribution of mean pairwise distance values for each patient comparison (see methods for how null distribution was created). A point below the null distribution indicates convergent TIV-responding CDR3 sequences, and above indicates divergent TIV-responding CDR3 sequences. Patient comparisons are sorted by observed mean pairwise genetic distance, and distributions are colored by their empirical p-value. P1 vs. P4 p-value = 0.001.

Power calculations:

In order to assess statistical power for our SGS based tests for convergence, we calculated power over a range of parameter values for both our global gene usage convergence test as well as our individual gene convergence test. See methods for a detailed description of how this was done

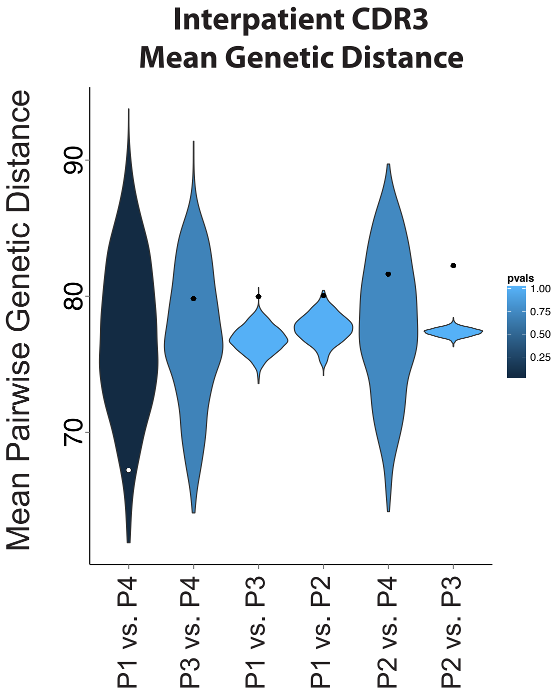


Figure 2.15. Testing for convergent CDR3 sequences across patients.
Violin plots showing the null distribution of mean pairwise distance values for each pairwise patient comparison (see methods for how null distribution was created), for the PBMC data. Points indicate the observed mean genetic distance for the TIV-responding CDR3 sequences between the patient comparison. A point below the null distribution indicates convergent TIV-responding CDR3 sequences, and above indicates divergent TIV-responding CDR3 sequences. Distributions are colored by p value with respect to the observed value. Patient 5 is absent because he had no statistically significant TIV-responding CDR3 sequences in the PBMC data.

To calculate power for our global gene usage test, we designed simulations where we can simulate a given number of truly convergent genes, as well as simulate additional patients. From this we were able to determine how many truly convergent genes, and how many patients are necessary to give sufficient power to differentiate from the null distribution. For example, one can imagine that if there were only 5 V genes that were truly convergent then it might be difficult for the resulting distribution of SGS values to be statistically different from the null. However, if there were 50 patients in the dataset, then it would be unlikely for all 50 of these patients to ‘choose’ the 5 convergent genes by chance, and would allow for a statistical difference from the null. We ran these power simulations over a range of ‘number of patients’, and ‘number of convergent genes' parameter values (Fig. 2.16A). We found that in order to have strong power to reject the null, if there are 5 patients (as in our observed data), there must be greater than 9 convergent genes.

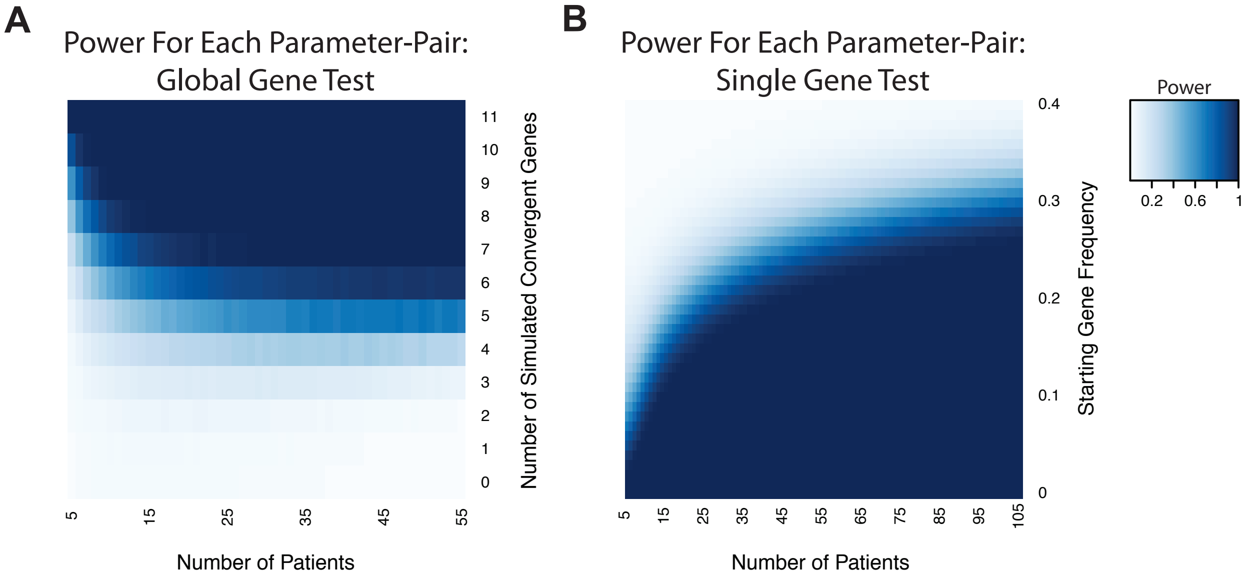


Figure 2.16. Power calculations.
Illustrates the statistical power over a range of parameter values for each of the gene usage convergence tests. (A) Power over a range of simulated patients, and simulated convergent genes for the global gene usage convergence test. Up to 50 convergent genes were simulated but all sets of simulations with greater than 10 convergent genes yielded a power of 1. (B) Power over a range of patients, and day 0 gene frequencies for the individual gene convergence test. Day 0 gene frequency was set to be equal across all patients for each power calculation. Tests with a starting gene frequency of up to 1.0 were run, however, every test with a starting gene frequency greater than 0.4 had a power of 0.0. IGHV expression data from the B cell dataset was used for the simulations/calculations in both (A) and (B).

We also calculated power for our individual gene convergence test over a range of parameter values. Here, the parameters that we varied were 'number of patients' and 'starting gene frequency' (frequency at day 0). In this case, if a gene were highly expressed at day 0 then it would be difficult for this gene to be statistically different from the null hypothesis, as it might be relatively easy for many patients to 'choose' this gene to respond to TIV by chance. However, if there were 100 patients in the study, then it may be unlikely for this gene to be selected in all patients. We found that if there are 5 patients in the study, a gene must have a day 0 frequency lower than 0.06 in each of the patients in order to reliably reject the null hypothesis (Fig. 2.16B).

Discussion

We have mined and characterized the global AbR response to TIV in 5 individuals from RNA-seq data. We find that individuals exhibit a heterogeneous response to TIV. Some of the patients showed characteristics of a monoclonal response, while others responded with much more of a polyclonal character. Interestingly, patient 1, who demonstrated characteristics of the most dramatic monoclonal response, was also the oldest patient (Table 2.1). This is in line with previous work showing that older humans tend to have larger clonal expansions in their AbRs [12]. While all the individuals' overall Ab expression increased markedly post-vaccination, the timing and amplitude of this spike was variable. It is important to note that the patient with arguably the most dramatic Ab response to TIV also had a relatively early spike in overall Ab expression, which had almost completely subsided by day 7. This is the time-point that immunologists typically collect samples for vaccine response studies (see Galson *et al.* [2] Table 1 for examples), and in this individual's case the dramatic signal would have been all but lost if the traditional study design of pre- and post-vaccination time-points were used. This is consistent with the findings of Henn et al. 2013 [30] and further exemplifies the utility of study designs that emphasize dense, longitudinal sampling rather than cross-sectional sampling, as much of the signal would have been missed were there sparser sampling in the time-course. Further, as one decreases the number of time-points, it may become increasingly difficult to distinguish the signal from the noise, which would decrease the power to identify the elements responding to the stimulus.

While targeted sequencing of the Ab locus is unarguably the best way to illustrate the AbR, we, and others [35–37], have shown that a relatively simple bioinformatic pipeline can be implemented to characterize the AbR from RNA-seq data. This will hopefully provide investigators with the ability to leverage their RNA-seq data even further. Sequencing costs continue to plummet each year, however they still remain prohibitive for performing both targeted sequencing, and RNA-seq for the average project budget. If one were interested in overall, population level statistics of the AbR, such as abundance or diversity, or if one were interested in finding/observing the Abs that are highly expressed in the AbR, we would argue that RNA-seq data is more than sufficient for these purposes. However, if one were interested in identifying rare Abs in the population, or needed full Ab sequences, then targeted sequencing of the Ab locus would be advised. In addition to prohibitive sequencing costs, targeted, deep-sequencing of the AbR remains a highly skilled method that involves a great deal of optimization, whereas RNA-seq has well vetted and broadly used protocols. In short, we hope that our method opens up the field of AbR analysis to a broader array of researchers.

The unique, densely sampled time-series dataset from Henn *et al.* [30] provided us with the ability to use functional data analysis methods to statistically identify putative TIV-responding V genes. We found V genes that were commonly TIV-responding across all patients in our dataset, and that these commonly used V genes were also prevalent in influenza targeting Abs collected from the literature. This finding suggests that we have identified V genes that indeed function to target TIV. This also raises the intriguing possibility that some V genes are inherently better than others at targeting TIV, as independent patients seem to be selecting the same V genes to target the vaccine. If this were true it would have interesting implications for the natural design and function of the diversity of genes in the AbR. For example, it could imply that instead of the different V genes providing the basis for an otherwise random exploration of sequence space when optimizing Abs, they could perhaps have evolved as 'specialists' for particular classes of antigens, such that when an Ab is comprised of a particular V gene it is pushed in a particular direction of antigenic space.

As interesting as a convergent signal may be, one must exercise great caution when searching for one. If correlations between individuals exist prior to the selection event, then these correlations must be controlled for in any convergence test. For example, consider a V gene that is highly expressed in many individuals prior to vaccination, and imagine that this V gene was found to be TIV-responding in many patients. As we have pointed out, one does not know if the reason that this V gene was found to be TIV-responding across patients is because it actually has a greater propensity to target TIV than other V genes, or because it was selected randomly due to its high prevalence in the individuals. It is certainly possible that the highly expressed V genes have a greater propensity to target TIV. Indeed, it is possible that the reason they are highly expressed is because of prior vaccinations/antigenic exposure. However, we argue that it is equally possible that some Ab genes have a high endogenous expression level independent of any antigenic stimulus. Because of this, we do not have the statistical ability to de-convolute these two possibilities. Increasing the number of patients in these types of studies would help ameliorate this problem. However, as we show with our power calculations, one experiences diminishing returns in statistical power with adding patients to the study (Fig. 2.16). Alternatively, a synthetic AbR could be created that has a relatively even distribution of Ab elements, and tested for activity against TIV (or other antigens as well).

Despite the strong correlations across patients in V gene expression levels prior to vaccination, we found statistically significant convergent signals in a subset of our tests. We observed global convergence for the IGKV genes, as well as convergence in the individual V genes, IGHV3-66 and IGKV3-NL1. As Dunand and Wilson [38] point out in their review, the V genes IGHV1-69 and IGHV3-7 have been implicated in convergent signals in a huge variety of contexts, including *chronic lymphocytic leukemia [59,60],* Sjögren's syndrome [61], and influenza [41,55–58,62–64] (for both IGHV1-69, and IGHV3-7), as well as human immunodeficiency virus [65–67], and hepatitus C virus [68] (for IGHV1-69 alone). Given that these genes were not significant in our convergence tests, and given the vast array of disparate antigens that these genes have been shown to ‘converge’ towards, it seems that perhaps the simpler explanation may in fact be that these genes have high endogenous expression independent of any antigenic stimulus, and are simply found to consistently respond to a diverse array of antigens by chance. This is a hypothesis that we feel deserves further scrutiny in future studies.

Our method for testing for a convergent signal in the AbR could be easily extendable to other systems. For example, this approach could be applied to meta-genomic microbiome data in order to identify taxa that are consistently responding to some stimulus. It could also be applied to infections in order to see which sequence characteristics of a given pathogenic population are consistently responding to (or resisting) a drug.

References

1. Jenner E. An inquiry into the causes and effects of the variolae vaccinae, a disease discovered in some of the western counties of England, particularly Gloucestershire, and known by the name of the cow pox. printed for the author, by D.N. Shury; 1801.

2. Galson JD, Pollard AJ, Trück J, Kelly DF. Studying the antibody repertoire after vaccination: practical applications. Trends Immunol. 2014;35:319–31.

3. MacLennan IC. Germinal centers. Annu. Rev. Immunol. 1994;12:117–39.

4. Liu YJ, Joshua DE, Williams GT, Smith CA, Gordon J, MacLennan IC. Mechanism of antigen-driven selection in germinal centres. Nature. 1989;342:929–31.

5. Hozumi N, Tonegawa S. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. Proc. Natl. Acad. Sci. 1976;73:3628–32.

6. Brack C, Hirama M, Lenhard-Schuller R, Tonegawa S. A complete immunoglobulin gene is created by somatic recombination. Cell. 1978;15:1–14.

7. Tonegawa S. Somatic generation of antibody diversity. Nature. 1983;302:575–81.

8. Charles A. Janeway, Travers P, Walport M, Shlomchik MJ, Jr CAJ, Travers P, et al. Immunobiology. 5th ed. Garland Science; 2001.

9. Laserson U, Vigneault F, Gadala-Maria D, Yaari G, Uduman M, Heiden JAV, et al. High-resolution antibody dynamics of vaccine-induced immune responses. Proc. Natl. Acad. Sci. 2014;111:4928–33.

10. Dekosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. Nat. Biotechnol. 2013;31:166–9.

11. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. Proc. Natl. Acad. Sci. 2013;110:13463–8.

12. Jiang N, He J, Weinstein JA, Penland L, Sasaki S, He X-S, et al. Lineage Structure of the Human Antibody Repertoire in Response to Influenza Vaccination. Sci. Transl. Med. 2013;5:171ra19–171ra19.

13. Lavinder JJ, Wine Y, Giesecke C, Ippolito GC, Horton AP, Lungu OI, et al. Identification and characterization of the constituent human serum antibodies elicited by vaccination. Proc. Natl. Acad. Sci. 2014;111:2259–64.

14. Wu Y-CB, Kipling D, Dunn-Walters DK. Age-Related Changes in Human Peripheral Blood IGH Repertoire Following Vaccination. Front. Immunol. 2012;3:193.

15. Wu X, Zhou T, Zhu J, Zhang B, Georgiev I, Wang C, et al. Focused Evolution of HIV-1 Neutralizing Antibodies Revealed by Structures and Deep Sequencing. Science. 2011;333:1593–602.

16. Scheid JF, Mouquet H, Feldhahn N, Seaman MS, Velinzon K, Pietzsch J, et al. Broad diversity of neutralizing antibodies isolated from memory B cells in HIV-infected individuals. Nature. 2009;458:636–40.

17. Bonsignori M, Hwang K-K, Chen X, Tsao C-Y, Morris L, Gray E, et al. Analysis of a Clonal Lineage of HIV-1 Envelope V2/V3 Conformational Epitope-Specific

Broadly Neutralizing Antibodies and Their Inferred Unmutated Common Ancestors. J. Virol. 2011;85:9998–10009.

18. Scheid JF, Mouquet H, Ueberheide B, Diskin R, Klein F, Oliveira TYK, et al. Sequence and Structural Convergence of Broad and Potent HIV Antibodies That Mimic CD4 Binding. Science. 2011;333:1633–7.

19. Lu DR, Tan Y-C, Kongpachith S, Cai X, Stein EA, Lindstrom TM, et al. Identifying functional anti-Staphylococcus aureus antibodies by sequencing antibody repertoires of patient plasmablasts. Clin. Immunol. 2014;152:77–89.

20. Zhu J, Ofek G, Yang Y, Zhang B, Louder MK, Lu G, et al. Mining the antibodyome for HIV-1–neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. Proc. Natl. Acad. Sci. 2013;110:6470–5.

21. Parameswaran P, Liu Y, Roskin KM, Jackson KKL, Dixit VP, Lee J-Y, et al. Convergent Antibody Signatures in Human Dengue. Cell Host Microbe. 2013;13:691–700.

22. Greiff V, Bhat P, Cook SC, Menzel U, Kang W, Reddy ST. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. Genome Med. 2015;7:49.

23. Cortina-Ceballos B, Godoy-Lozano EE, Téllez-Sosa J, Ovilla-Muñoz M, Sámano-Sánchez H, Aguilar-Salgado A, et al. Longitudinal analysis of the peripheral B cell repertoire reveals unique effects of immunization with a new influenza virus strain. Genome Med. 2015;7:124.

24. Ramsay J, Silverman BW. Functional Data Analysis. 2nd edition. New York: Springer; 2005.

25. Coffey N, Hinde J. Analyzing time-course microarray data using functional data analysis - a review. 2011 [cited 2015 Jul 4]; Available from: http://aran.library.nuigalway.ie/xmlui/handle/10379/1903

26. Yao F, Müller H-G, Wang J-L. Functional Data Analysis for Sparse Longitudinal Data. J. Am. Stat. Assoc. 2005;100:577–90.

27. Liu X, Yang MCK. Identifying temporally differentially expressed genes through functional principal components analysis. Biostat. Oxf. Engl. 2009;10:667–79.

28. Wu S, Wu H. More powerful significant testing for time course gene expression data using functional principal component analysis approaches. BMC Bioinformatics. 2013;14:6.

29. Storey JD, Xiao W, Leek JT, Tompkins RG, Davis RW. Significance analysis of time course microarray experiments. Proc. Natl. Acad. Sci. U. S. A. 2005;102:12837–42.

30. Henn AD, Wu S, Qiu X, Ruda M, Stover M, Yang H, et al. High-Resolution Temporal Response Patterns to Influenza Vaccine Reveal a Distinct Human Plasma Cell Gene Signature. Sci. Rep. [Internet]. 2013 [cited 2014 Mar 16];3. Available from: http://www.nature.com/srep/2013/130731/srep02327/full/srep02327.html

31. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HYK, Chen R, et al. Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes. Cell. 2012;148:1293–307.

32. Franco LM, Bucasas KL, Wells JM, Niño D, Wang X, Zapata GE, et al. Integrative genomic analysis of the human immune response to influenza vaccination.

eLife [Internet]. 2013 [cited 2014 Mar 16];2. Available from: http://elife.elifesciences.org/content/2/e00299

33. Liao H-X, Lynch R, Zhou T, Gao F, Alam M, Boyd S, et al. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. Nature [Internet]. 2013 [cited 2013 Apr 4]; Available from: http://dx.doi.org/10.1038/nature12053

34. Hoehn KB, Gall A, Bashford-Rogers R, Fidler SJ, Kaye S, Weber JN, et al. Dynamics of immunoglobulin sequence diversity in HIV-1 infected individuals. Phil Trans R Soc B. 2015;370:20140241.

35. Blachly JS, Ruppert AS, Zhao W, Long S, Flynn J, Flinn I, et al. Immunoglobulin transcript sequence and somatic hypermutation computation from unselected RNA-seq reads in chronic lymphocytic leukemia. Proc. Natl. Acad. Sci. 2015;112:4322–7.

36. Iglesia MD, Vincent BG, Parker JS, Hoadley KA, Carey LA, Perou CM, et al. Prognostic B-cell Signatures Using mRNA-Seq in Patients with Subtype-Specific Breast and Ovarian Cancer. Clin. Cancer Res. 2014;20:3818–29.

37. Brown SD, Raeburn LA, Holt RA. Profiling tissue-resident T cell repertoires by RNA sequencing. Genome Med. 2015;7:125.

38. Dunand CJH, Wilson PC. Restricted, canonical, stereotyped and convergent immunoglobulin responses. Phil Trans R Soc B. 2015;370:20140238.

39. Parker J, Tsagkogeorga G, Cotton JA, Liu Y, Provero P, Stupka E, et al. Genome-wide signatures of convergent evolution in echolocating mammals. Nature [Internet]. 2013 [cited 2013 Sep 6];advance online publication. Available from:

http://www.nature.com/nature/journal/vaop/ncurrent/full/nature12511.html?WT.ec_id=NATURE-20130905

40. Doolittle RF. Convergent evolution: the need to be explicit. Trends Biochem. Sci. 1994;19:15–8.

41. Jackson KJL, Liu Y, Roskin KM, Glanville J, Hoh RA, Seo K, et al. Human Responses to Influenza Vaccination Show Seroconversion Signatures and Convergent Antibody Rearrangements. Cell Host Microbe. 2014;16:105–14.

42. Krause JC, Tsibane T, Tumpey TM, Huffman CJ, Briney BS, Smith SA, et al. Epitope-Specific Human Influenza Antibody Repertoires Diversify by B Cell Intraclonal Sequence Divergence and Interclonal Convergence. J. Immunol. 2011;187:3704–11.

43. Arnaout R, Lee W, Cahill P, Honan T, Sparrow T, Weiand M, et al. High-Resolution Description of Antibody Heavy-Chain Repertoires in Humans. PLoS ONE. 2011;6:e22365.

44. Boyd SD, Gaëta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD, et al. Individual Variation in the Germline Ig Gene Repertoire Inferred from Variable Region Gene Rearrangements. J. Immunol. 2010;184:6986–92.

45. Childs LM, Baskerville EB, Cobey S. Trade-offs in antibody repertoires to complex antigens. Phil Trans R Soc B. 2015;370:20140245.

46. GEO Accession viewer [Internet]. [cited 2016 Mar 21]. Available from: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45764

47. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36.

48. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res. 2013;41:W34–40.

49. R code for significant testing for time course gene expression data using functional principal component analysis approaches — Immune Modeling Community [Internet]. [cited 2015 Jun 24]. Available from: http://www.imcportal.org/repository/software/r-code-for-significant-testing-for-time-course-gene-expression-data-using-functional-principal-component-analysis-approaches

50. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med. 2009;6:e1000097.

51. ImmPort: Immunology Database and Analysis Portal - Study Detail [Internet]. [cited 2015 Jul 29]. Available from: https://immport.niaid.nih.gov/immportWeb/clinical/study/displayStudyDetails.do?itemList=SDY224

52. nbstrauli/influenza_vaccination_project [Internet]. GitHub. [cited 2015 Dec 9]. Available from: https://github.com/nbstrauli/influenza_vaccination_project

53. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol. 2013;31:46–53.

54. Throsby M, van den Brink E, Jongeneelen M, Poon LLM, Alard P, Cornelissen L, et al. Heterosubtypic Neutralizing Monoclonal Antibodies Cross-

Protective against H5N1 and H1N1 Recovered from Human IgM+ Memory B Cells. PLoS ONE. 2008;3:e3942.

55. Ekiert DC, Bhabha G, Elsliger M-A, Friesen RHE, Jongeneelen M, Throsby M, et al. Antibody Recognition of a Highly Conserved Influenza Virus Epitope. Science. 2009;324:246–51.

56. Sui J, Hwang WC, Perez S, Wei G, Aird D, Chen L, et al. Structural and Functional Bases for Broad-Spectrum Neutralization of Avian and Human Influenza A Viruses. Nat. Struct. Mol. Biol. 2009;16:265–73.

57. Corti D, Suguitan AL, Pinna D, Silacci C, Fernandez-Rodriguez BM, Vanzetta F, et al. Heterosubtypic neutralizing antibodies are produced by individuals immunized with a seasonal influenza vaccine. J. Clin. Invest. 2010;120:1663–73.

58. Wrammert J, Koutsonanos D, Li G-M, Edupuganti S, Sui J, Morrissey M, et al. Broadly cross-reactive antibodies dominate the human B cell response against 2009 pandemic H1N1 influenza virus infection. J. Exp. Med. 2011;208:181–93.

59. Fais F, Ghiotto F, Hashimoto S, Sellars B, Valetto A, Allen SL, et al. Chronic lymphocytic leukemia B cells express restricted sets of mutated and unmutated antigen receptors. J. Clin. Invest. 1998;102:1515–25.

60. Messmer BT, Albesiano E, Efremov DG, Ghiotto F, Allen SL, Kolitz J, et al. Multiple Distinct Sets of Stereotyped Antigen Receptors Indicate a Role for Antigen in Promoting Chronic Lymphocytic Leukemia. J. Exp. Med. 2004;200:519–25.

61. De Re V, De Vita S, Gasparotto D, Marzotto A, Carbone A, Ferraccioli G, et al. Salivary gland B cell lymphoproliferative disorders in Sjögren’s syndrome present a

restricted use of antigen receptor gene segments similar to those used by hepatitis C virus-associated non-Hodgkins's lymphomas. Eur. J. Immunol. 2002;32:903–10.

62. Dreyfus C, Laursen NS, Kwaks T, Zuijdgeest D, Khayat R, Ekiert DC, et al. Highly conserved protective epitopes on influenza B viruses. Science. 2012;337:1343–8.

63. Avnir Y, Tallarico AS, Zhu Q, Bennett AS, Connelly G, Sheehan J, et al. Molecular Signatures of Hemagglutinin Stem-Directed Heterosubtypic Human Neutralizing Antibodies against Influenza A Viruses. PLoS Pathog. [Internet]. 2014 [cited 2016 Mar 8];10. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4006906/

64. Thomson CA, Wang Y, Jackson LM, Olson M, Wang W, Liavonchanka A, et al. Pandemic H1N1 Influenza Infection and Vaccination in Humans Induces Cross-Protective Antibodies that Target the Hemagglutinin Stem. Front. Immunol. [Internet]. 2012 [cited 2016 Mar 8];3. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3347682/

65. Sabin C, Corti D, Buzon V, Seaman MS, Hulsik DL, Hinz A, et al. Crystal Structure and Size-Dependent Neutralization Properties of HK20, a Human Monoclonal Antibody Binding to the Highly Conserved Heptad Repeat 1 of gp41. PLOS Pathog. 2010;6:e1001195.

66. Luftig MA, Mattu M, Di Giovine P, Geleziunas R, Hrin R, Barbato G, et al. Structural basis for HIV-1 neutralization by a gp41 fusion intermediate-directed antibody. Nat. Struct. Mol. Biol. 2006;13:740–7.

67. Gustchina E, Li M, Louis JM, Anderson DE, Lloyd J, Frisch C, et al. Structural basis of HIV-1 neutralization by affinity matured Fabs directed against the internal trimeric coiled-coil of gp41. PLoS Pathog. 2010;6:e1001182.

68. Ivanovski M, Silvestri F, Pozzato G, Anand S, Mazzaro C, Burrone OR, et al. Somatic hypermutation, clonal diversity, and preferential expression of the VH 51p1/VL kv325 immunoglobulin gene combination in hepatitis C virus-associated immunocytomas. Blood. 1998;91:2433–42.

Chapter 3: Estimating and testing for selection in experimentally evolving populations

Introduction

The advent of deep sequencing technologies has allowed researchers to observe evolution taking place in controlled laboratory settings. This is done by simultaneously tracking thousands of allele frequencies within a large (typically microbial) population over time. A fundamental question when performing these evolution experiments is: which of these alleles are under selection? The first step to answering this question is to generate some quantitative statistic that reflects the fitness of each allele. This is calculated by comparing the frequency of a given allele at the beginning of the experiment, to its frequency at the end. In which case, an increase or decrease in the allele's frequency would indicate positive or negative selection, respectively. Commonly used statistics for measuring this frequency-change are based upon taking the ratio, or log ratio, of the ending allele frequency over the starting allele frequency [1,2]. While these methods are statistically robust, they have two drawbacks: i) they do not directly incorporate time (or number of generations) that elapsed during the experiment, and ii) they are difficult to incorporate into established population genetics theory.

The importance of time is well illustrated by the following example. An allele that doubles in frequency in 50 generations is under much higher selection than an allele that doubles in frequency in 100 generations. This is typically not an issue in evolution experiments because populations from the same study are generally left to evolve for the same amount of time. However, if this were not the case—such as when comparing results from different studies—then time would need to be taken into account when comparing fitness values.

Being able to integrate a fitness statistic into established theory is important because it can provide consistency and comparability across studies, and more importantly, gives one the ability to leverage the theoretical work of others to generate expectations. For example, if an allele rises in frequency from 1/8 to 1/4, then one could use population genetics theory to arrive at the range of fitness values that could reasonably give rise to such an observation. Moreover, one could design a test to see if an allele with no fitness advantage (i.e. a fitness of 0) could give rise to such an observation due to neutral drift, which is to say, the probability that the observation was due to chance.

In this study we implement an easy, computationally fast method for estimating the selection coefficient [3] (i.e. fitness, or s), and also develop a statistical framework to test if alleles are under significant selection. Our methods require a number of population genetic parameters that are usually not known when studying natural populations, yet are typically readily available in the context of laboratory experiments. These parameters include: the number of generations elapsed during the experiment, the population size across time-points, and the starting and ending allele frequencies. We also incorporate experimental/sequencing error into our framework.

We then go on to apply this method to two datasets of deeply sequenced evolving populations: the first consists of human immunodeficiency virus (HIV), and the second consists of Ebola virus (EBOV). The HIV dataset is from a study where the authors mutated every amino acid of the virus's *tat* and *rev* genes to every other possible amino acid, and then combined each of these alleles at roughly equal frequencies into a highly diverse artificial population [4]. The authors then let this population evolve in tissue culture by letting the diversity of alleles compete, and consequently rise or fall in frequency as a function of their relative fitnesses. This was done for a total of two replicates. These so called ‘deep mutational scanning’ experiments have recently been employed in a variety of contexts, and have proven quite useful to elucidate the fitness landscape for numerous proteins [5]. For this particular study, the authors used this approach to determine the fitness advantage of having overlapping reading frames in the *tat* and *rev* genes of the HIV genome. The EBOV dataset consists of a more homogeneous wildtype population of EBOV that was serially passaged using either a snake cell line (JK cells) or a human cell line (HeLa cells). In this study, the authors sought to observe how much adaptation, if any, is necessary for EBOV to stably infect reptilian cells, while using the human cell line as a negative control. This was performed for a total of three replicates for each cell line.

Together, we have developed a framework for easily estimating fitness for each allele in an experimentally evolving population, and then testing which of those alleles are under statistically significant selection. We then use this approach to help identify alleles of interest in two experimental evolution studies.

Methods

Point estimate of s:

Assuming a large Wright-Fisher population, if an allele has a selection coefficient of s, then (by the definition) it is expected to increase in frequency in the next generation by a factor of $1 + s$. So, if allele A has a frequency of p_i at generation i, and a selection coefficient of s, then the expectation of its frequency in the following generation is,

$$p_{i+1} = \frac{p_i(1+s)}{q_i+p_i(1+s)} .$$

Here, the numerator reflects the expected amount that the allele will increase in one generation, and the denominator serves to normalize this new value by the total frequency in the next generation. This value is known as the *mean fitness*. The reason p_{i+1} is normalized by the mean fitness is so that it remains a true frequency that has a range of $[0,1]$.

Similarly, if allele a has a frequency of $q_i = 1 - p_i$ at generation i, and is not under selection ($s = 0$), then its expected frequency in the following generation is given by,

$$q_{i+1} = \frac{q_i}{q_i + p_i(1+s)} = 1 - p_{i+1} \ .$$

Note that the count of allele a is not expected to change (as it is not selected), however, its relative frequency will go down because of the increase in allele A, which is reflected in the denominator.

One can also easily generate an expression for the ratio of the two allele's frequencies in the following generation, which we will call an allele ratio.

$$\frac{p_{i+1}}{q_{i+1}} = \frac{p_i(1+s)}{q_i} \ .$$

Notice that the denominators in the previous equations cancel out, simplifying the equation for the allele ratio after one generation.

We now wish to determine the expected allele ratio after t generations. After 2 generations (assuming the selection coefficient remains constant) the allele ratio is,

$$\frac{p_{i+2}}{q_{i+2}} = \frac{p_{i+1}}{q_{i+1}}(1+s) = \frac{p_i(1+s)}{q_i}(1+s) = \frac{p_i}{q_i}(1+s)^2 \ ,$$

and after 3 generations the allele ratio is,

$$\frac{p_{i+3}}{q_{i+3}} = \frac{p_{i+2}}{q_{i+2}}(1+s) = \frac{p_i}{q_i}(1+s)^2(1+s) = \frac{p_i}{q_i}(1+s)^3 \ .$$

It is then easy to see that after t generations the expected allele ratio will be,

$$\frac{p_{i+t}}{q_{i+t}} = \frac{p_i}{q_i}(1+s)^t \ .$$

We can then rearrange this equation to solve for s and arrive at,

$$s = e^{\frac{\ln(p_{i+t}/q_{i+t}) - \ln\,(p_i/qi)}{t}} \ .$$

Thus, we have an expression for the expected selection coefficient when provided with a change in allele frequency over some period of time. Please note that this material was largely adapted from Joseph Felsenstein's book, Theoretical Evolutionary Genetics [6].

Estimating experimental error in EBOV data:

A PhiX control was sequenced in the EBOV data, and we were able to use this information to estimate the sequencing error rate for each position. In the case of PhiX sequences, the expectation is that all reads will have the reference allele, and any reads that deviate from this expectation are the result of sequencing error. We estimate the error rate by counting the number of instances that we observed a mutation 'away' from a given nucleotide, and the number of times we observe a mutation 'to' that nucleotide. We then normalized these counts by the total number of reads observed for that nucleotide to get mutation rates. Thus, each nucleotide has an estimated 'away' mutation rate and an estimated 'to' mutation rate associated with it. We represent the rate at which nucleotide Y mutates to any other nucleotide as $R_{Y,\cdot}$, and the rate at which any other nucleotide mutates to Y as $R_{\cdot,Y}$.

Estimating experimental error in HIV data:

Because HIV wildtype (non-randomized) positions were also sequenced in the data, we were able to use this information to estimate the overall error rate for each amino acid type. Similar to PhiX in the EBOV data, the expectation is that all sequenced wildtype positions will have the reference allele, and any instance that deviates from this expectation are the result of some sort of error during the course of our experimental protocol. The advantage of using wildtype positions relative to PhiX in the EBOV data is that this error includes random mutations during the experiment, PCR error, and sequencing error. We then estimate $R_{Y,.}$ and $R_{.,Y}$ the same as was done in the EBOV data.

Estimating population growth in EBOV data:

Estimation of the EBOV population size at each of the sequencing time points was achieved by two-step reverse transcription droplet digital PCR (RT-ddPCR) [7,8]. These figures gave us estimates of the EBOV census population size directly at the time of sequencing, and just before the population was bottlenecked due to passaging. The severity of each bottleneck was estimated by first reducing the population by the volume used to start the next passage (1/40), and then incorporating an estimate of the proportion of viable virions in the supernatant (0.00243678, see [7] for details). We then modeled an exponential growth curve between each bottleneck and the following observed population size at the time of sequencing (Fig. 3.1).

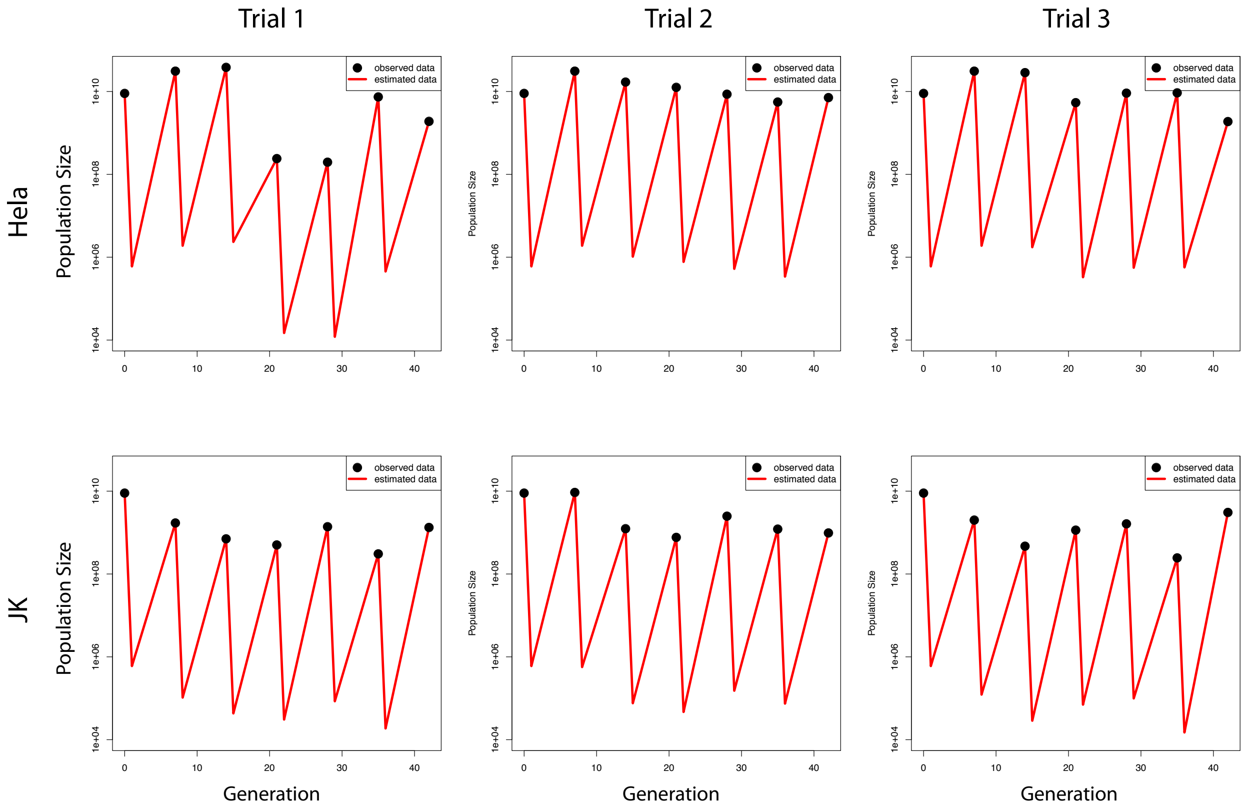


Figure 3.1. Estimated population size trajectories for EBOV populations.
Shows the observed and estimated population sizes for each of the EBOV populations over time. The black points give the observed population size for a given time using RT-ddPCR. The red lines show the estimated population size following passaging bottlenecks, and recovery. Columns of panels correspond to experimental trials 1-3 (from left to right), and rows of panels correspond to HeLa and JK cells (top to bottom).

The form of the exponential curve we used was

$$X_t = X_0(1+r)^t.$$

Here, X_0 is set to be the population size directly after a given bottleneck. X_t is set to equal the census population size at the following sequencing time-point, and t is always 6, as there are about 6 generations between each bottleneck and the following sequencing time-point. We then simply solve for r to get an exponential curve that connects the bottleneck population size to the following sequencing population size, over 6 generations. This gave us estimates of the population size at each of the 42 generations over the course of the experiments. While these estimates were quite rough, they provided us with the general shape of the EBOV demography over the course of the evolution experiments, which we were able to incorporate into our simulations.

Estimating population growth in HIV data:

We used data on the growth of a wildtype population of HIV under the same experimental protocol as in the HIV evolution experiments, to estimate the overall growth of the HIV population in each experiment. Specifically, the p24 concentration ([p24] in pg/uL) was measured for several time-points during the course of the experiment, and this was done a total of three times (see [4] for details). We set the overall population size of HIV to be equal to the mean [p24] value at each time-point. We then fit a logistic curve to these observed values to get a population growth function over the time-course of the experiments (Fig. 3.2).

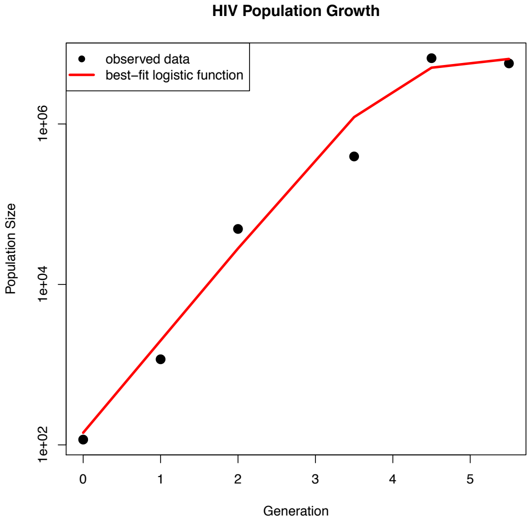


Figure 3.2. Estimated population size trajectory for HIV populations.
Shows the best parametric fit of a logistic population growth function (red line) to the observed data (black points).

Here is the function that resulted from that fit:

$$N_t = f(t) = \frac{6604652.673}{1+e^{10.7974-2.646389t}} \text{ ,}$$

where N_t is the population size at generation t.

Neutral simulations:

We sought to simulate, *in silico*, our evolution experiments in order estimate the range that a neutral allele's frequency could feasibly change during the experiments. This simulated range served as our null distribution, and we generated a unique null distribution for each allele in our observed data (done separately for both EBOV and HIV). The neutral simulations had six parameters: the overall population growth function, the number of generations, the starting allele frequency, the nucleotide identity of the allele, and the starting and ending read depth for the experiment. Each of these parameters will be explained below.

A set of 10,000 neutral simulations were run for each allele in the data, and each simulation was run as follows. We model the neutral allele in question as allele A, with frequency p and collapsed all the other alleles in the population to be allele a, with frequency $q = 1 - p$. If the population size and frequency for allele A at generation t is N_t and p_t, respectively, then the count of A is $P_t = N_t p_t$, and the count of a is $Q_t = N_t - P_t$. The first task is to initialize a starting allele frequency, p_0 for the simulation. There exists an observed starting frequency, p_0' for each allele, however, this observation has experimental error associated with it due to sampling the population and sequencing. Thus, we derived a probability distribution for the true allele frequency, given the observed allele frequency and sequencing depth. We model our prior expectation of the true allele frequency using the beta distribution with shape parameters α and β

$$\Pr(p_0|\alpha,\beta) = \frac{p_0^{\alpha-1} q_0^{\beta-1}}{B(\alpha,\beta)},$$

where $B(\alpha,\beta)$ is the beta function and serves as a normalizing constant. α and β were estimated using the method of moments (MOM) approach, where p_0' is treated as the result of a series of Bernoulli trials, and we use the mean and variance of these trials to estimate α and β. We then model the probability of observing a given allele count, P'_0, given a true starting allele frequency, p_0 and the sample size N'_0 using the binomial distribution,

$$\Pr(P'_0|N'_0, p_0) = \binom{N'_0}{P'_0} p_0^{P'_0} q_0^{N'_0 - P'_0} \text{ ,}$$

where N'_0 is equal to the read depth for the allele's position at time 0. We then use a Bayesian framework to model the probability of the true allele frequency, given our prior and observations,

$$\Pr(p_0|P'_0, N'_0, \alpha, \beta) = \frac{\Pr\left(P'_0\middle|N'_0, p_0\right)\Pr(p_0|\alpha, \beta)}{\sum_{A=0}^{A=N_0} \Pr\left(P'_0\middle|N'_0, A/N_0\right)\Pr(A/N_0|\alpha, \beta)} \text{ .}$$

After substituting in the above equations, this simplifies to,

$$\Pr(p_0|P'_0, N'_0, \alpha, \beta) = \frac{p_0^{\alpha+P'_0-1} q_0^{\beta+(N'_0-P'_0)-1}}{B(\alpha,\beta)} \text{ ,}$$

which is itself a beta distribution. We then randomly sample from this distribution to get a value for the starting allele frequency p_0 for the simulation.

Once p_0 is set, we ran a Wright-Fisher simulation [9] of neutral drift over the number of generations that elapsed for the duration of the evolution experiment. This was 42 generations for EBOV [7] and 6 generations for HIV [4], and will be referred to as T for simplicity. Under this framework, the probability of an allele count in the next generation (P_{t+1}) follows the binomial distribution,

$$\Pr(P_{t+1}) = \binom{f(t+1)}{P_{t+1}} p_t^{P_{t+1}} q_t^{Q_{t+1}} \text{ .}$$

Thus, to get a random value for P_{t+1} in the simulation we simply randomly sample from the above distribution. Once the Wright-Fisher simulation is complete, we again use the binomial distribution to simulate sub-sampling of the population. We do this because the true population size at the end of the experiment is quite large, and when we sequence this population, we are only observing a subset. The size of this subsample is equal to the number of reads that mapped to the allele's position in the genome (i.e. read depth). If N_T is the final population size, let the subsample size be N'_T. Similarly, if P_T is the count of A at the final generation, then P'_T is the count of A resulting from subsampling. We again use the binomial to model the probability of P'_T,

$$\Pr(P'_T) = \binom{N'_T}{P'_T} p_T^{P'_T} q_T^{N'_T - P'_T} \ ,$$

and randomly sample from this distribution to simulate a value for P'_T. In order to simulate random additions to and subtractions from allele A that occur due mutations from experimental error, we again use a binomial sampling approach. If the identity of allele A is Y then the rate at which Y is mutated to anything else is $R_{Y,\cdot}$, and the rate at which anything else mutates to Y is $R_{\cdot,Y}$. Let the number of units that is added to P'_T due to experimental error be represented as $X_{\cdot,Y}$, and the number of units subtracted be $X_{Y,\cdot}$. The probability of these two values are again found using the binomial,

$$\Pr\left(X_{\cdot,Y}\right) = \binom{N'_T - P'_T}{X_{\cdot,Y}} R_{\cdot,Y}^{X_{\cdot,Y}} (1 - R_{\cdot,Y})^{N'_T - P'_T - X_{\cdot,Y}}; \text{ and}$$

$$\Pr\left(X_{Y,\cdot}\right) = \binom{P'_T}{X_{Y,\cdot}} R_{Y,\cdot}^{X_{Y,\cdot}} (1 - R_{Y,\cdot})^{P'_T - X_{Y,\cdot}} \ .$$

We randomly sample from these two distributions to get discrete values for $X_{\cdot,Y}$ and $X_{Y,\cdot}$. Finally, Let the count of allele A, at generation T, after subsampling, and after introducing random experimental error be P''_T. P''_T is then given by,

$$P''_T = \ P'_T - X_{Y,\cdot} + X_{\cdot,Y} \ .$$

P_T'' is the final allele count that we use in the simulation. This process is then repeated 10,000 times to arrive at a simulated distribution for P_T''. This distribution represents the range of ending allele frequencies that one could expect for a neutral allele, if this allele had the same starting frequency, read depth, and nucleotide identity as a given observed allele in our data.

Simulations with selection:

We ran the same simulations of our evolution experiments as described above, but included a selection parameter in order to test the accuracy of our selection coefficient estimates of observed alleles. The only component of the framework described above that changed in this case was the Wright-Fisher simulations. Here, the expected frequency of allele A with a selection coefficient of s at generation $t+1$ is given by,

$$E[p_{t+1}] = \frac{p_t(1+s)}{q_t+p_t(1+s)} \text{ ,}$$

and the probability of P_{t+1} with selection is given by,

$$\Pr(P_{t+1}) = \binom{f(t+1)}{P_{t+1}} E[p_{t+1}]^{\,P_{t+1}} (1-E[p_{t+1}])^{Q_{t+1}} \text{ .}$$

Results

Estimating fitness for each allele in HIV data:

We used a closed-form solution of the selection coefficient to get a point estimate of the fitness for each allele in the HIV data. This equation uses the observed beginning and ending allele frequencies to calculate the most probable selection coefficient, and is accurate in large populations [3,6] (Fig. 3.3).

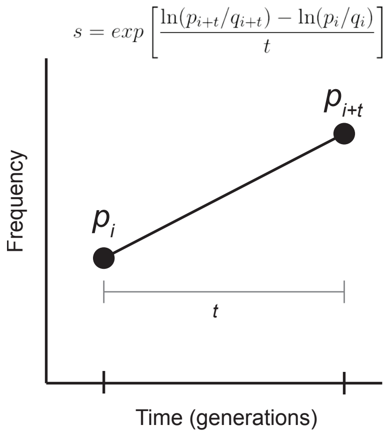

$s = exp\left[\frac{\ln(p_{i+t}/q_{i+t}) - \ln(p_i/q_i)}{t}\right]$
p_{i+t}
p_i
t
Frequency
Time (generations)

Figure 3.3. Point estimate of *s*.
Illustrates the parameters that contribute to the closed form estimation of the selection coefficient (equation at the top of the plot).

We compared our calculation of s to the ratio of the ending over starting allele frequencies, which is the basis for many commonly used statistics to gauge fitness [10], and found that they track well with one another (Fig. 3.4).

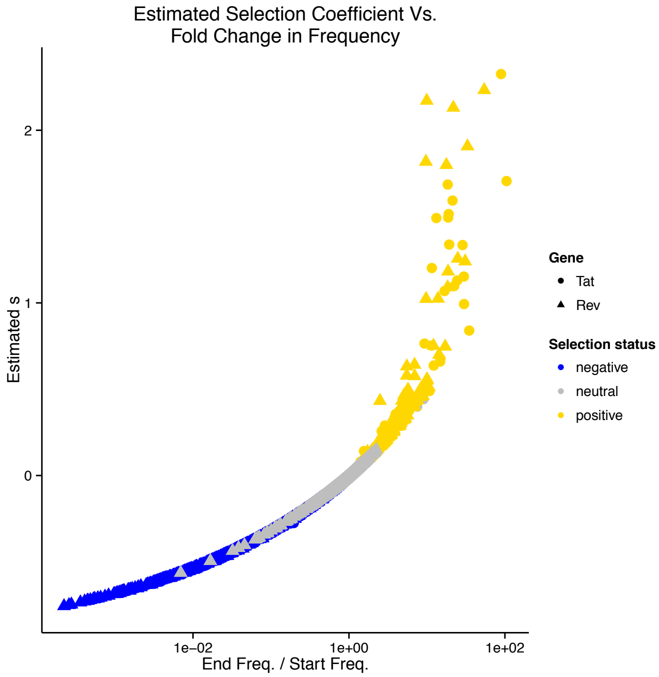


Figure 3.4. Relationship between fold-change in frequency to our fitness estimates.
Each point is one allele in the HIV dataset. Color indicates negative (blue), neutral (grey), or positive (gold) selection, as determined by our simulation-based statistical test. Shape indicates the gene (*tat* or *rev*).

However, we found estimating the selection coefficient to be more interpretable than statistics that take the ratio (or log-ratio) of the start and end allele frequencies, as the selection coefficient is rooted in a mathematical model of a changing population over time.

Checking the accuracy of fitness estimates in HIV data:

Because s is rooted in a population genetics framework, We were able to run Wright-Fisher simulations [9] with selection (see Methods) to gauge the accuracy of our fitness estimates. Specifically, we ran 1,000 simulations with a randomly selected value of s, between -2 and 2. For each simulation, we then calculated the empirical s, using our closed form equation. We found that the closed-form equation for the selection coefficient was very accurate relative to the true value in the simulations (Fig. 3.5).

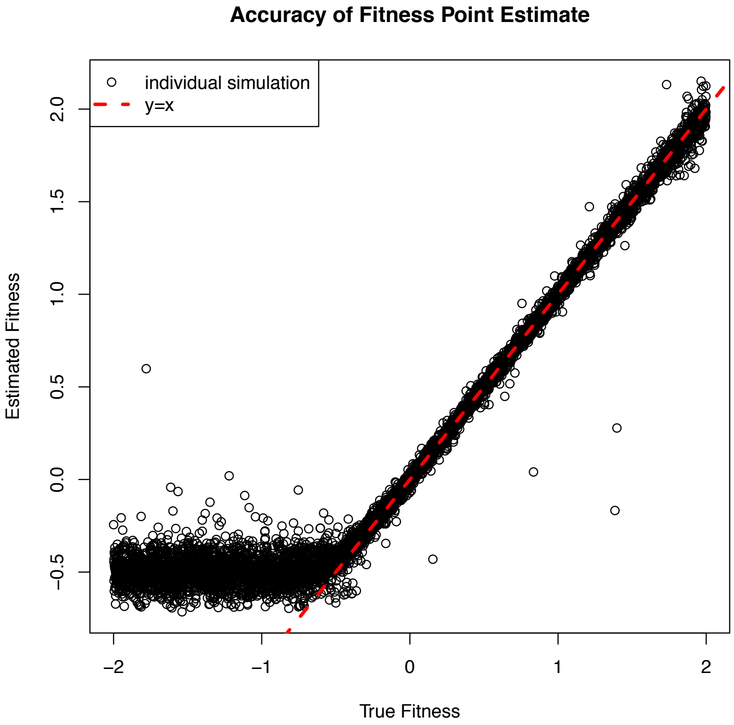


Figure 3.5. Correlation between the estimated fitness and the true fitness under our simulation framework.
Each point corresponds to one simulation. The red dashed line shows y=x.

However, we note that when the ‘true’ selection coefficient is quite low (<-0.5) then one’s ability to estimate it using observed data decreases. This is because a moderately negatively selected allele will fall to a frequency of 0 just the same as an extremely negatively selected allele, if the starting frequency is low. This conceptual limit to one’s ability to infer negative selection with low starting frequency is well illustrated by the plateauing of points on the left-hand side of Figure 3.5.

Testing for significantly selected alleles in HIV data:

There are two factors that can cause an allele to change in frequency over time: selection, and neutral drift. Our point estimates of s work under the asymptotic assumption that a population size is approaching ∞. When this happens, populations begin to behave deterministically, and the effects of drift become negligible. While our populations are large, they are not quite large enough to ignore the effects of drift. Because of this, we designed a statistical test to identify the alleles whose change in frequency cannot be accounted for by drift alone. Which is to say, to identify alleles under significant positive or negative selection (Fig. 3.6).

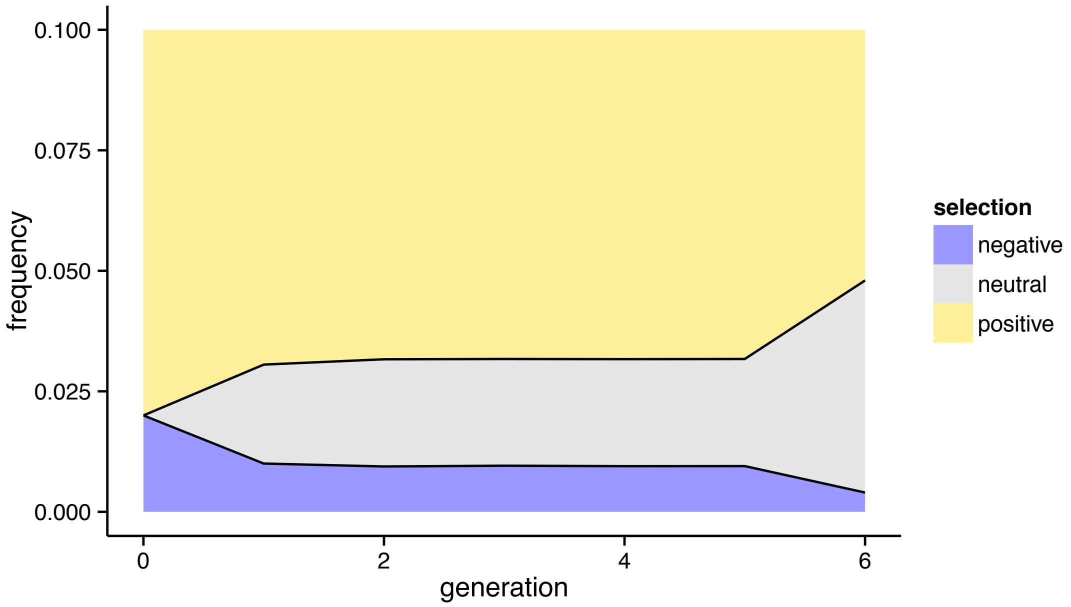


Figure 3.6. An illustration of the neutral simulations for a hypothetical allele. The allele has a starting frequency of 0.02, an ending read depth of 500 reads, and an amino acid identity of Arginine. The grey area depicts the range of trajectories that this allele could take if it were neutral. If an ending allele frequency were observed to be above or below this neutral expectation, it would be deemed positively or negatively selected, respectively. The black dots indicate the upper and lower bounds for the ending allele frequency that would still be considered neutral. These upper and lower bounds correspond to relative fitness values of 0.380 and -0.699, respectively, which means neutrality cannot be rejected for any observed fitness value that resides between this interval.

Of the alleles that failed our test for neutrality, those that had positive point estimates for s were deemed under positive selection, and those that had negative s estimates were deemed under negative selection. We found that generally (and perhaps not surprisingly) alleles with large $|s|$ were under significant selection, and alleles with small $|s|$ were not. However, we also found some notable exceptions to this, where some alleles that had large values of $|s|$ (i.e. large changes in allele frequency over the time-course) were not under significant selection, and also that alleles with small $|s|$ actually were under significant selection (Fig. 3.7).

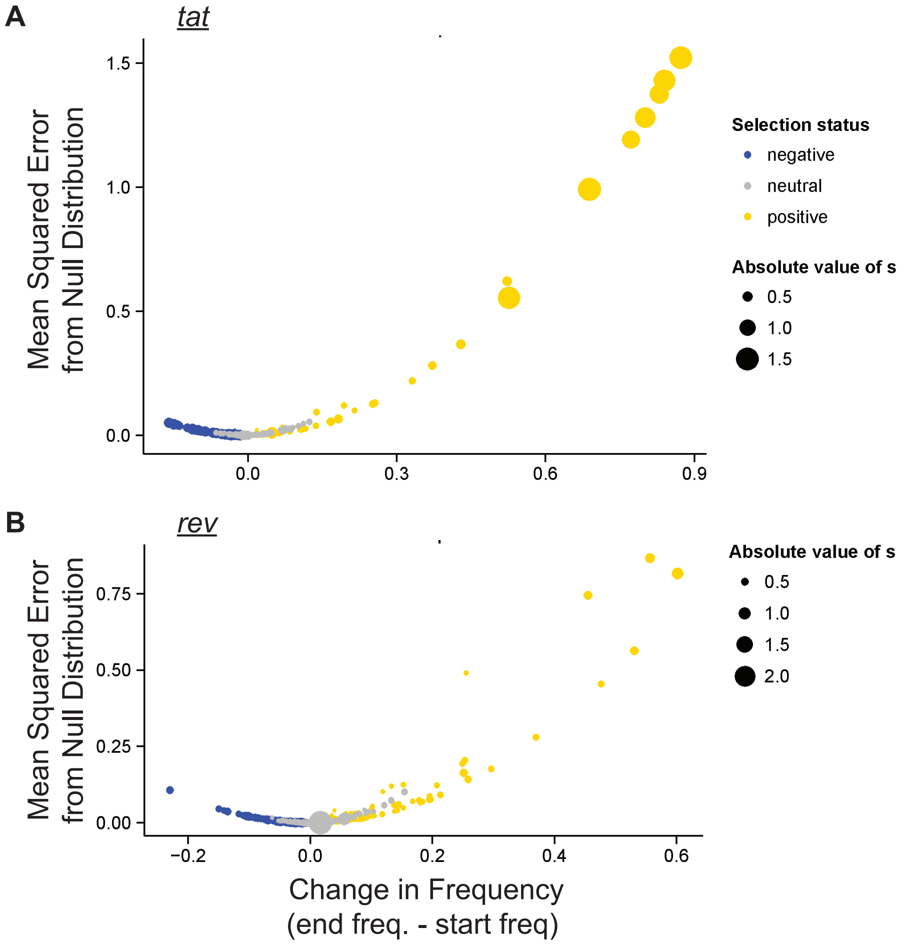


Figure 3.7. Results of simulation-based test for significant selection.
Shows the relationship between the observed change in allele frequency to the distance from the null distribution. The x-axis shows the observed absolute change in allele frequency for a given allele, and the y-axis reflects the distance (as squared error) that a given allele's ending allele frequency is to its null distribution. (A) Tat. (B) Rev.

This suggests that if one were to only use point estimates of fitness, and not take into account the effects of neutral drift, then they could encounter a high rate of false positives and false negatives.

<u>Estimating fitness for each allele in EBOV data:</u>

As was done for the HIV data, we estimated s for each allele in the EBOV dataset (Fig. 3.8).

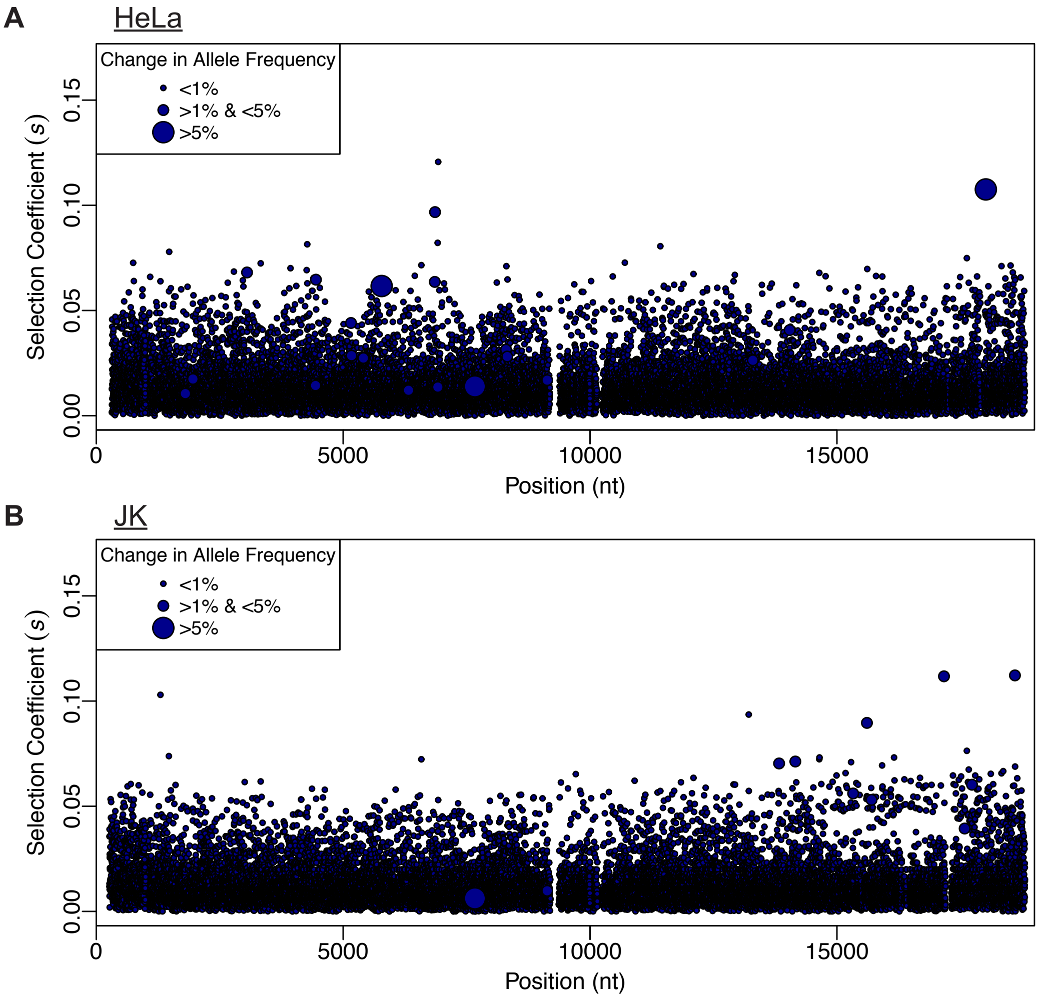


Figure 3.8. Selection coefficients for each position in the EBOV genome.
Shows the maximum estimated s (y-axis) across all alleles at a given site, and across the three trials for each position in the EBOV genome (x-axis). There is one circle shown for each position, and the diameter of each circle reflects the absolute change in frequency over the time-course. (A) JK. (B) HeLa.

We found that alleles overall had comparably lower values of s than the HIV dataset, which is perhaps not surprising considering the EBOV population began as wildtype, and was then left to adapt (if need be) to either a human or snake cellular environment. These selection coefficients are on the order of what has been estimated in natural human populations, whereas s values in experimental evolution studies are typically much higher.

Identifying positively selected alleles in EBOV data:

Despite the low selection regime in the EBOV data. We sought to identify the alleles that were under the highest selection in the EBOV populations. To do this we again used our simulation-based test to identify the significantly non-neutral alleles, and added extra criteria where alleles must be non-synonymous variants that were present in all three biological replicates. We found only 5 and 3 alleles that met these criteria in the HeLa and JK cell lines, respectively (Fig. 3.9).

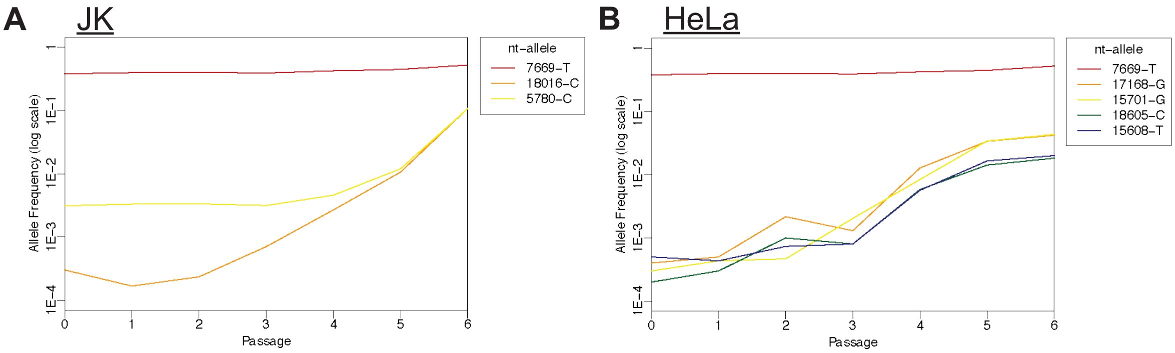


Figure 3.9. Most positively selected alleles in EBOV data.
Shows the frequency trajectories of the positively selected alleles that met our criteria. Each color corresponds to an allele, and the number followed by a letter in the legends gives the nucleotide position in the reference genome, followed by the nucleotide identity of each allele, respectively. (A) JK. (B) HeLa.

Discussion

In this study we have presented a fitness statistic, s, that may provide some benefit over more commonly used statistics in experimental evolution studies. These benefits include: incorporating time (or number of generations in the study) to provide comparability across studies, and providing one with the ability to simulate populations *in-silico* to check the accuracy of fitness estimates. The drawback of this approach is that one needs to know the generation time of the populations in a study. However, as we have demonstrated in our two applications of this approach, generation time of organisms that are typically used in experimental evolution studies is generally known [4,7].

As Jewett *et al.* have shown [3], deterministic estimates of s are accurate in large populations, however, when populations become sufficiently small, these estimates can become unstable. Because experimental evolution studies will often involve very small founder populations (as in our HIV data, Fig. 3.2), or complicated population size histories due to serial passaging (as in our EBOV data, Fig. 3.1), we found it unwise to disregard the possible effects of neutral drift. In addition to this, every study will have varying amounts of uncertainty introduced from experimental error (i.e. mutations, and sequencing error), and random subsampling of the population when sequencing. Thus, we developed a statistical framework that will take into account these sources of uncertainty when determining whether or not a given allele is under significant selection. Many methods have been developed to estimate s given time-series data [11–13], however these methods typically rely on a diffusion approximation of a Wright-Fisher process [14]. The diffusion approximation involves (among other things) a rescaling of the time dimension to become continuous. This is inappropriate in studies that span relatively few generations, such as our HIV and EBOV datasets. Thus we used a discrete-time Wright-Fisher process in our simulations, which allowed us to estimate the exact probability distribution of the ending allele frequencies of a neutral allele.

An important caveat to our Wright-Fisher simulation-based approach is that it does not capture the dynamics of clonal interference, where different alleles do not act in isolation, but instead compete with one another to reach fixation (frequency of one) in the population [15]. One approach to ameliorate this would to be to simulate all alleles together and jointly infer their selection coefficients. While this would be more computationally expensive (because a range of s values would need to be considered), it could potentially incorporate clonal interference into a simulation framework.

We were able to see that alleles that were deemed neutral by our test were actually capable of exhibiting large changes in allele frequency over the time-course (Fig. 3.10). This exemplifies the importance of using a statistical approach when assigning selection to alleles in experimental evolution studies. For example, in the HIV dataset, an allele in the *rev* gene that had one of the highest selection coefficients in the data was also deemed to be neutral by our test (Fig. 3.7B).

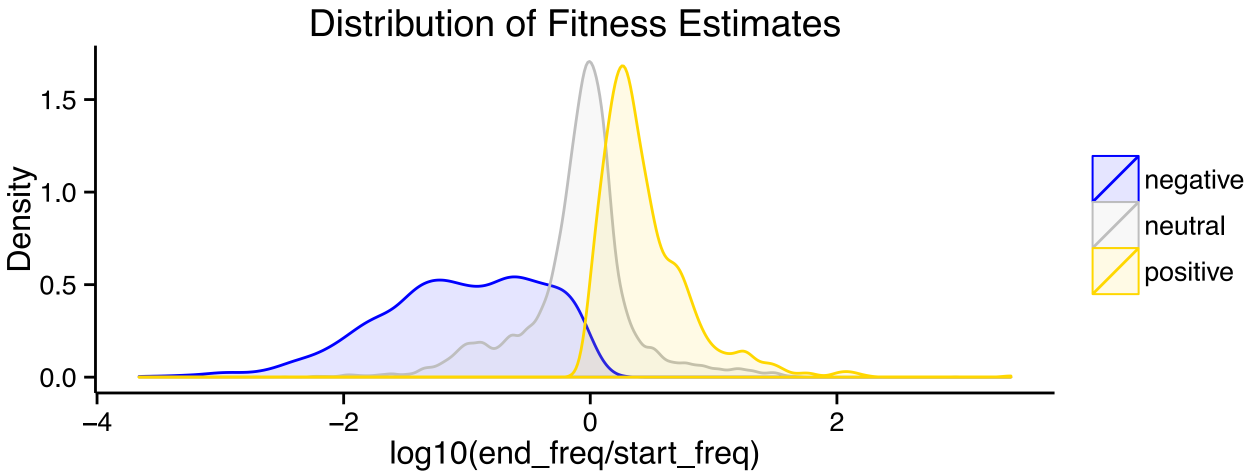


Figure 3.10. Distribution of fold change values in HIV data.
Shows the observed distribution of fold change values for alleles found to be under negative (blue), neutral (grey), or positive (gold) selection, as determined by our simulation-based test. Neutral alleles were sometimes found to have relatively extreme changes in frequency (left and right tails of grey distribution). Likewise, alleles under significant positive or negative selection were sometimes found to have relatively small changes in frequency (right tail of blue distribution, and left tail of gold distribution). Any values that appear above or below 0 in the blue or gold distributions, respectively, are due to the density smoothing function and were not seen in our data.

Together, we have put together a statistical framework for answering questions about selection in allele frequency data from experimental evolution studies. We hope that this will help future investigators who wish to identify sets of alleles in their data that are important (i.e. under selection). While implementing a simulation-based statistical test can be burdensome, we have shown that it can help identify type I and type II error.

References

1. McLaughlin Jr RN, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R. The spatial architecture of protein function and adaptation. Nature [Internet]. Nature Publishing Group; 2012 [cited 2019 Feb 12];491:138–42. Available from: http://www.nature.com/articles/nature11500

2. Bloom JD. Software for the analysis and visualization of deep mutational scanning data. BMC Bioinformatics [Internet]. BioMed Central; 2015 [cited 2019 Feb 12];16:168. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0590-4

3. Jewett EM, Steinrücken M, Song YS. The Effects of Population Size Histories on Estimates of Selection Coefficients from Time-Series Genetic Data. Mol Biol Evol [Internet]. Oxford University Press; 2016 [cited 2019 Feb 13];33:3002–27. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msw173

4. Fernandes JD, Faust TB, Strauli NB, Smith C, Crosby DC, Nakamura RL, et al. Functional Segregation of Overlapping Genes in HIV. Cell [Internet]. 2016 [cited 2019 Feb 12];167:1762–1773.e12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27984726

5. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods [Internet]. Nature Publishing Group; 2014 [cited 2019 Feb 13];11:801–7. Available from: http://www.nature.com/articles/nmeth.3027

6. Felsenstein J. Theoretical Evolutionary Genetics [Internet]. Seattle; 2016 [cited 2019 Feb 12]. Available from: http://evolution.genetics.washington.edu/pgbook/pgbook.html

7. Fedewa G, Radoshitzky SR, Chī X, Dŏng L, Zeng X, Spear M, et al. Ebola virus, but not Marburg virus, replicates efficiently and without required adaptation in snake cells. Virus Evol [Internet]. Oxford University Press; 2018 [cited 2019 Feb 12];4. Available from: https://academic.oup.com/ve/article/doi/10.1093/ve/vey034/5214739

8. Hindson BJ, Ness KD, Masquelier DA, Belgrader P, Heredia NJ, Makarewicz AJ, et al. High-Throughput Droplet Digital PCR System for Absolute Quantitation of DNA Copy Number. Anal Chem [Internet]. 2011 [cited 2019 Feb 13];83:8604–10. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22035192

9. Hartl DL, Clark AG. Principles of population genetics. Sinauer Associates; 2007.

10. Wiser MJ, Lenski RE. A Comparison of Methods to Measure Fitness in Escherichia coli. Blanchard JL, editor. PLoS One [Internet]. Public Library of Science; 2015 [cited 2019 Feb 14];10:e0126210. Available from: http://dx.plos.org/10.1371/journal.pone.0126210

11. Schraiber JG, Evans SN, Slatkin M. Bayesian Inference of Natural Selection from Allele Frequency Time Series. Genetics [Internet]. Genetics; 2016 [cited 2019 Feb 18];203:493–511. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27010022

12. Malaspinas A-S, Malaspinas O, Evans SN, Slatkin M. Estimating Allele Age and Selection Coefficient from Time-Serial Data. Genetics [Internet]. 2012 [cited 2019 Feb 18];192:599–607. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22851647

13. Steinrücken M, Bhaskar A, Song YS. A NOVEL SPECTRAL METHOD FOR INFERRING GENERAL DIPLOID SELECTION FROM TIME SERIES GENETIC DATA. Ann Appl Stat [Internet]. NIH Public Access; 2014 [cited 2019 Feb 18];8:2203–22.

Available from: http://www.ncbi.nlm.nih.gov/pubmed/25598858

14. Kimura M. Diffusion models in population genetics. J Appl Probab [Internet]. 1964 [cited 2019 Feb 18];1:177–232. Available from: https://www.cambridge.org/core/product/identifier/S0021900200108368/type/journal_article

15. Park S-C, Krug J. Clonal interference in large populations. Proc Natl Acad Sci [Internet]. 2007 [cited 2019 Mar 13];104:18135–40. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17984061

Chapter 4: The genetic interaction between HIV and the antibody repertoire

Introduction

Since the beginning of the modern pandemic in 1981 [1,2], human immunodeficiency virus (HIV)—the virus that causes acquired immunodeficiency syndrome (AIDS)—has been the source of incredible scientific scrutiny. While there has been great progress in the development of antiretroviral therapeutics (ARTs), which can now manage the disease indefinitely (albeit only for those who can afford them), a cure remains elusive [3], and little progress has been made toward a preventative vaccine [4]. Prevention efforts have recently made significant headway by implementing preexposure prophylaxis (PrEP) to high risk individuals. However, this strategy has its downsides, such as a reliance on daily self-administration, significant financial burden, and health side effects [5]. Thus, cure and vaccine strategies remain the elusive goal for HIV research. Together, the distinct gains in AIDS treatment, yet relative lack of gains in HIV prevention, has resulted in a stalemate of sorts, where instead of HIV being triumphantly eradicated by modern science, it has settled into a persistent, yet treatable reality of human life.

The fervent hope for progress is particularly palpable in HIV vaccine research. This fervor is mainly fueled by the fact that effective HIV immunity is entirely possible and well documented, as it occurs naturally in 10-20% of those chronically infected [6–8]. If a vaccine (or vaccine regimen) were to be designed that could somehow recapitulate whatever happens during the humoral immune response of these 10-20% (almost) immune individuals, then triumph could be within reach! While a good deal of progress has been made in this avenue of research, it has not yet resulted in an effective vaccine. Among the promising discoveries are broadly neutralizing antibodies (bnAbs), which are monoclonal antibodies (Abs) that can single-handedly neutralize up to 90% of heterologous HIV strains [9,10]. To study the development of these bnAbs, researchers have utilized a post-hoc deep-sequencing approach by which a set of primers are developed that will preferentially amplify a subset of the antibody repertoire (AbR)—the population of antibodies in an organism—that is known to contain a bnAb lineage [10–12]. The advantage of this approach is that one can cut through the incredible noise and complexity of the AbR to focus on a particular lineage of importance. At the same time, such approaches miss the diversity of Ab lineages interacting with HIV that may have important effects on immunity outcomes. To our knowledge, only [13] have deeply sequenced the AbR in an unbiased fashion in the context of HIV infection, but they did not collect paired HIV sequence data to directly study genetic interactions.

The common narrative of bnAb development is that they are in a coevolutionary arms race [14] with the autologous HIV population [15,16]. There is good reason to suspect that this coevolutionary hypothesis is true: bnAbs tend to be quite derived relative to their inferred naïve ancestors [11], they tend to take a long time to develop (years), and there tends to be a time dependence on neutralization capabilities (i.e. HIV-neutralizing Abs are more likely to neutralize autologous virus from the past, and less likely to neutralize contemporaneous or future autologous virus) [12,17]. However, there is also evidence contrary to the arms race hypothesis: bnAbs can arise relatively quickly, and with few mutations [18–21], and superinfections—multiple HIV infections in the same individual—don’t necessarily drive further evolution in existing HIV-neutralizing Ab lineages, but rather promote the development of *de novo* HIV-targeting

Ab lineages [22,23]. Similarly, in the context of malaria infection, repeated immunizations with a complex malaria antigen tends to promote the activation of *de novo* naïve Ab lineages, rather than the evolution of already existing malaria-targeting Abs [24].

A better understanding of the interaction taking place between HIV and the AbR over time could potentially shed a good deal of light on how HIV-immunity is achieved. However, longitudinal sequence studies of HIV infections tend to either focus on HIV [25–29] or the AbR [13,30,31], but not both. To our knowledge, only three studies have deeply sequenced the AbR along with the autologous HIV population, but each of these studies consisted of a single individual with relatively few time-points [11,12,32]. In this study, we hope to ameliorate this dearth of data, and to also shed light on the genetic interaction between these putatively coevolving populations.

Results

Sequencing HIV *env* and IGH

We collected a total of 119 cryopreserved peripheral blood samples from the OPTIONS cohort at the University of California, San Francisco (UCSF). The samples all originated from men aged 25-48 years old at the estimated time of infection in San Francisco, and each patient had 10-20 longitudinal samples (Table 4.1, and Figure 4.1).

Table 4.1. Patient demographics.
MSM – men having sex with men. * at estimated time of infection

ID	Age*	Date*	Gender	Ethnicity	Exposure	Num. Samples
1	30	6/7/98	Male	White/European American	MSM	20
2	25	2/17/99	Male	Asian	MSM	17
3	32	7/4/01	Male	Hispanic/Latino	MSM	10
4	44	8/8/01	Male	White/European American	MSM	10
5	34	7/20/03	Male	Hispanic/Latino	Unknown	11
6	25	1/18/05	Male	White/European American	MSM	10
7	35	6/6/05	Male	White/European American	MSM	10
8	48	9/3/08	Male	White/European American	MSM	10
9	33	2/2/09	Male	Asian	MSM	10
10	31	4/3/09	Male	White/European American	MSM	10

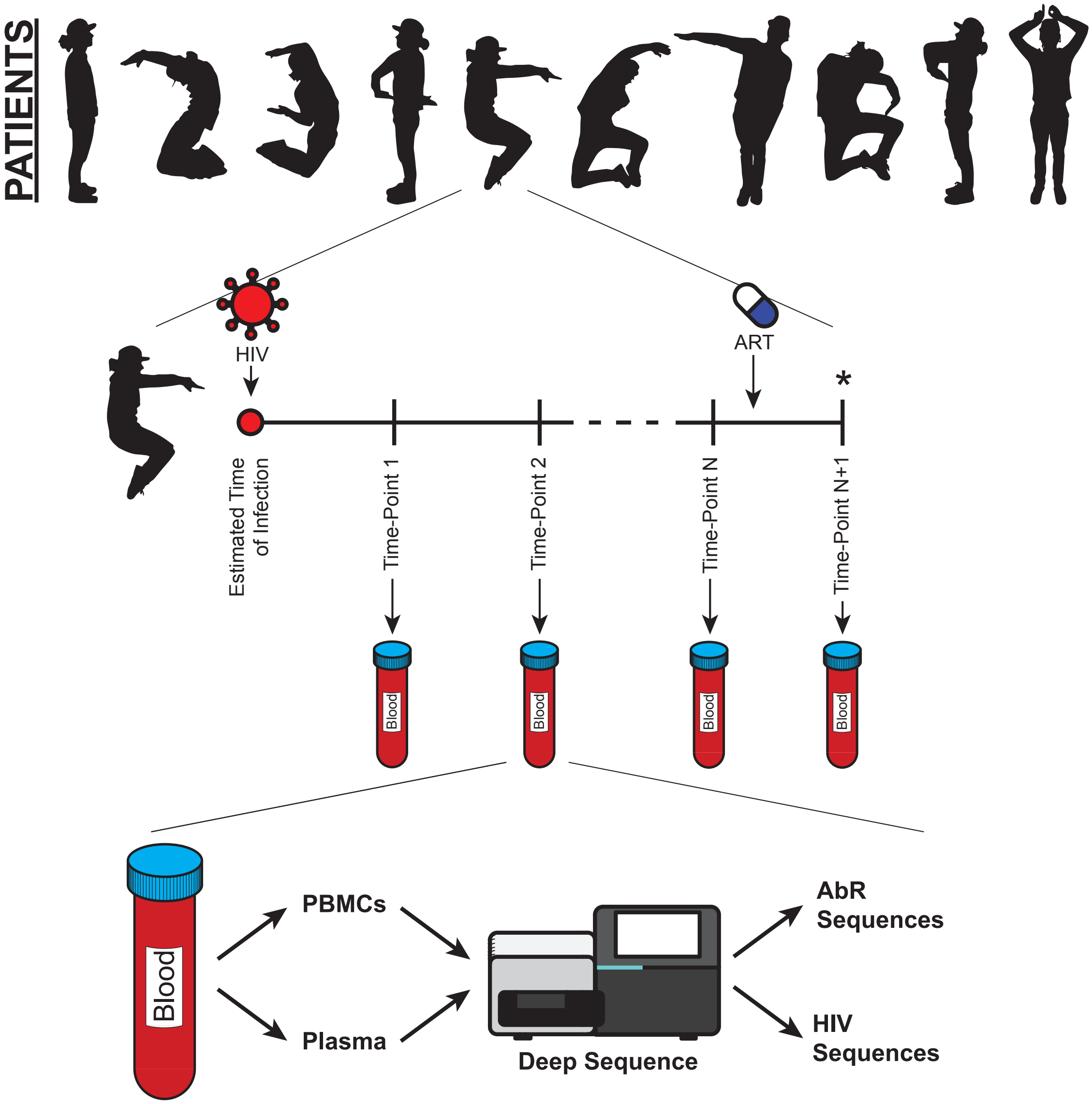


Figure 4.1. Schematic of study.
There were 10 patients in our study. Peripheral blood was drawn from each patient, post HIV infection, for 10-20 longitudinal time-points. Each blood sample was divided into PBMCs and plasma. Ab sequences were derived from the PBMCs and HIV sequences were derived from plasma. Note that patient avatars in no way reflect actual patients and instead reflect one lonely scientist.
*Only patients 1, 2, and 5 had the final time-point with ART exposure.

They were also all collected prior to administration of ART, with the exception of the last time-point of patients 1, 2, and 5. We chose to deeply sequence the C2-V3 region of the *env* gene because of its rich history of interactions with Abs, as evidenced by the HIV epitope map from Los Alamos National Labs (LANL) [33–35]. Of the ART naive samples, we were unable to successfully amplify C2-V3 from 12. In 11 of these cases, low viral load was the presumed cause for lack of amplification, but the 9th time-point of patient 6 was unsuccessful despite high viral load (Figure 4.2, 4.3). Interestingly, 8 of

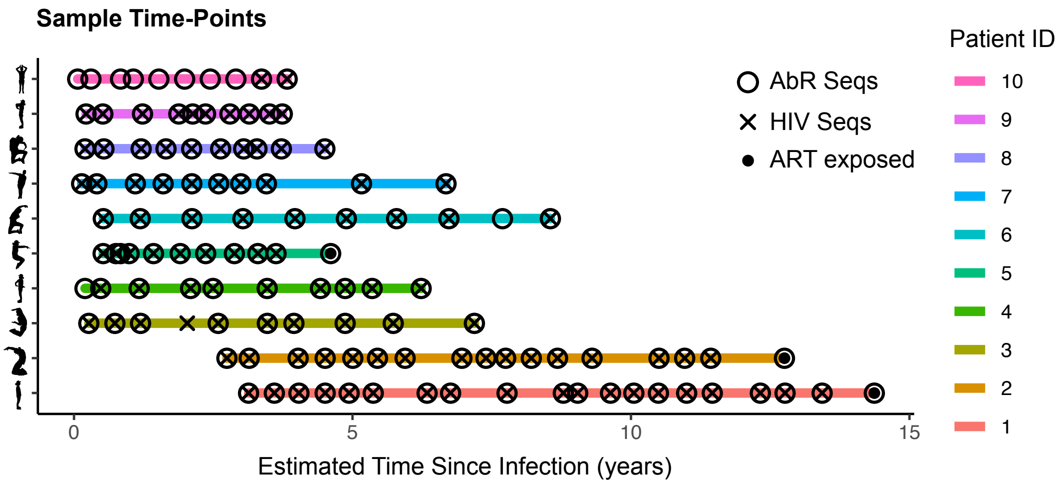


Figure 4.2. Sampled time-points.
Depicts the time since infection for each sample, in each of the patients. Open circles indicate that the AbR was successfully sequenced, crosses indicate that HIV was successfully sequenced, and solid circles indicate that the patient was on ART at this time.

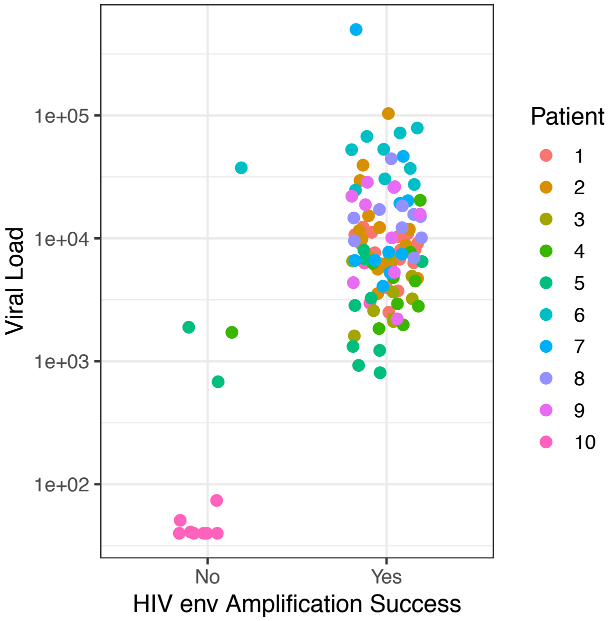


Figure 4.3. HIV amplification success and viral load.
Depicts the relationship between HIV viral load and our ability to amplify HIV *env* prior to sequencing. All samples in which *env* could not be amplified had low to exceptionally low viral load, with the exception of one sample from patient 6.

the 11 unsuccessfully amplified and low viral load samples were the first eight time-points of patient 10. We were also unable to amplify C2-V3 from any of the ART-exposed samples. While viral load measurements were not available for these samples, the presumed cause for our inability to amplify C2-V3 was again low viral load, given their ART status. The initial *env* sequencing depth ranged from 3,771-101,831 reads per sample, and after cleaning the data with several quality control (QC) steps, this ranged from 2,276-56,914 reads per sample (Figure 4.4).

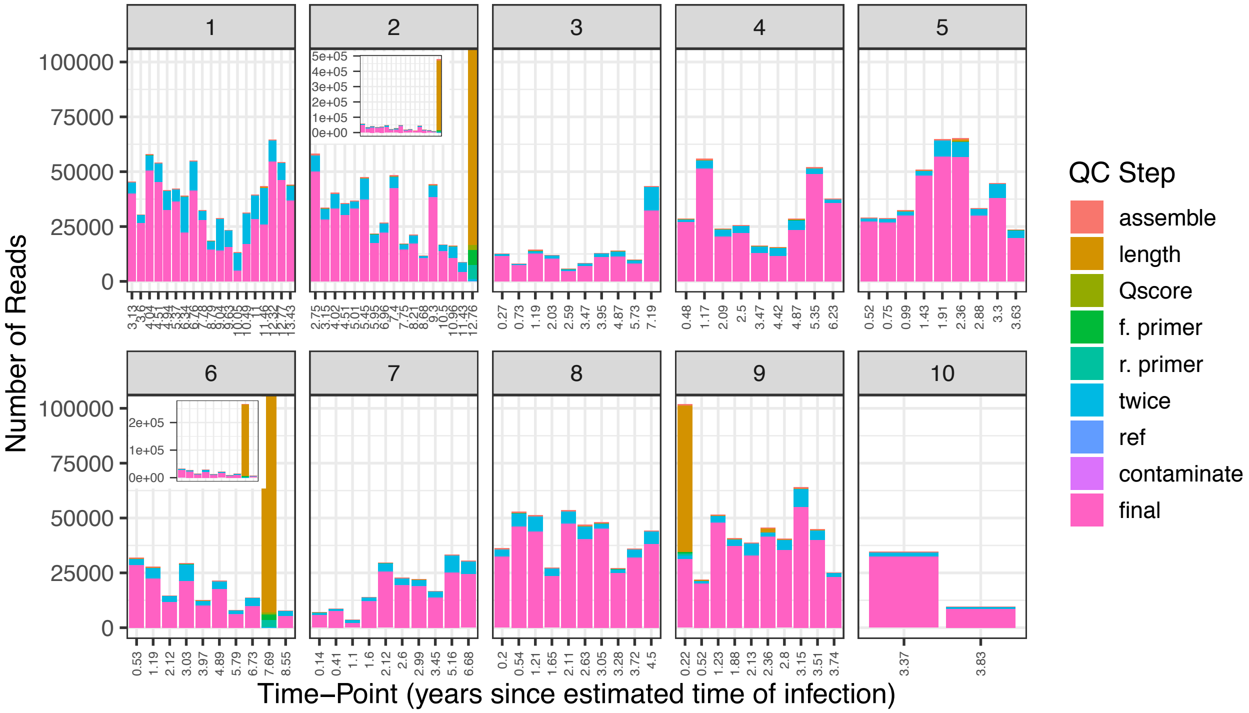


Figure 4.4. Read depth and QC filtering for HIV data.
Stacked bar charts depicting the total read depth and proportion of reads that were filtered after each QC step. The height of each bar shows initial read depth (directly after sequencing) for each HIV sample. The relative lengths of each color within a bar show how many reads were filtered out from a given QC step. The legend maps color to QC step, and QC steps are shown (from top to bottom) in the order that they were performed: “assemble” = could not assemble read pairs; “length” = too short; “Qscore” = mean Q score too low; “f. primer” = could not align forward primer; “r. primer” = could not align reverse primer; “twice” = sequence did not occur at least twice; “ref” = did not align to an HIV reference sequence; “contaminate” = sequence was found in other studies; “final” = the final sequences that passed all QC. Numeric titles for each panel give the patient ID. Insets in patients 2 and 6 show the full breadth of the y-axes, depicting the data from two samples that did not successfully amplify *env*, but were included as a negative control. They did not pass QC.

We also deeply sequenced the variable region of the immunoglobulin heavy chain locus (IGH), the product of which we refer to as the antibody repertoire (AbR). The AbR was successfully sequenced in all samples, with the exception of the fourth time-point of patient 3. Sequence data was generated for this sample, but it exhibited an abrupt clonal expansion of a magnitude that was a clear outlier for patient 3, and not seen in any other sample, so was discarded (Figure 4.5, and 4.2). Initial AbR

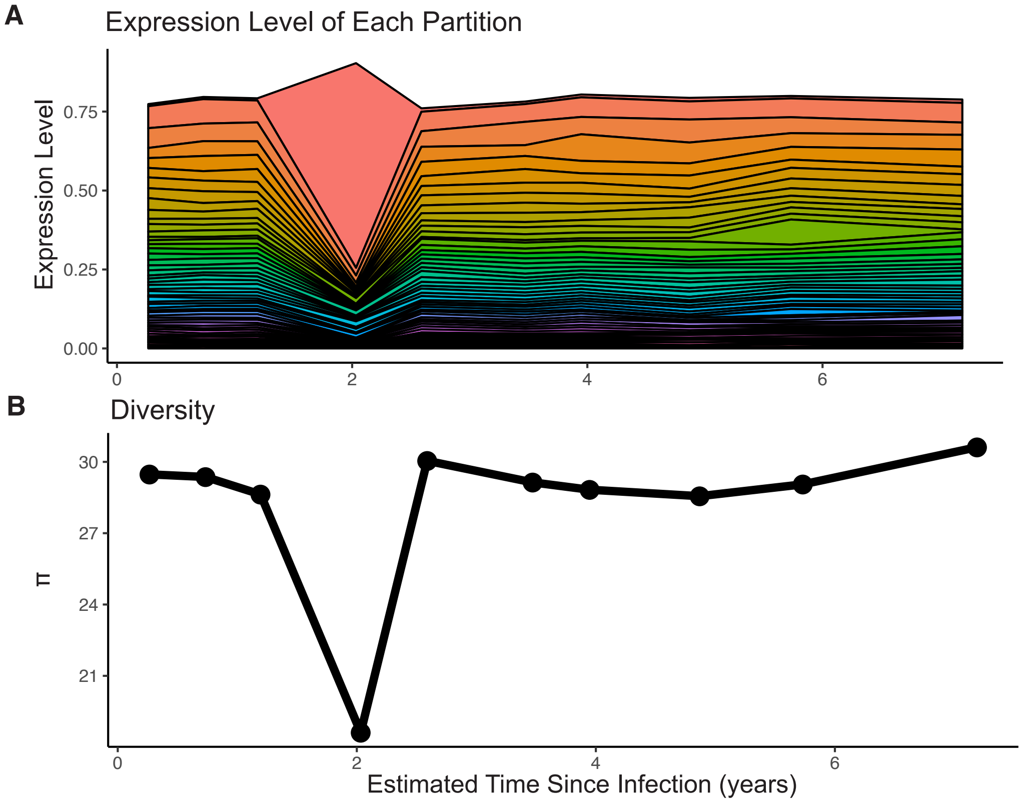


Figure 4.5. Anomalous time-point in patient 3.
(A) Stacked area chart shows the unrealistic clonal expansion occurring at the 4th time-point of patient 3. Each unique color represents a unique AbR partition, and the width of a given color represents the relative frequency of that AbR partition at that time-point. (B) A line plot showing the corresponding dramatic drop in AbR diversity due to this anomaly.

sequencing depth ranged from 669,331-2,669,662 reads per sample, and after QC this ranged from 160,291-552,479 reads per sample (Figure 4.6).

Characterizing the HIV population and AbR over time

In order to quantify the broad attributes of the AbR and HIV populations over time, we calculated a variety of summary statistics that characterized the genetic diversity, divergence, selection, and abundance for each of the populations. As others have reported [29,36], we found that diversity, divergence, and selection all tend to increase with time since estimated date of HIV infection, with large perturbations over

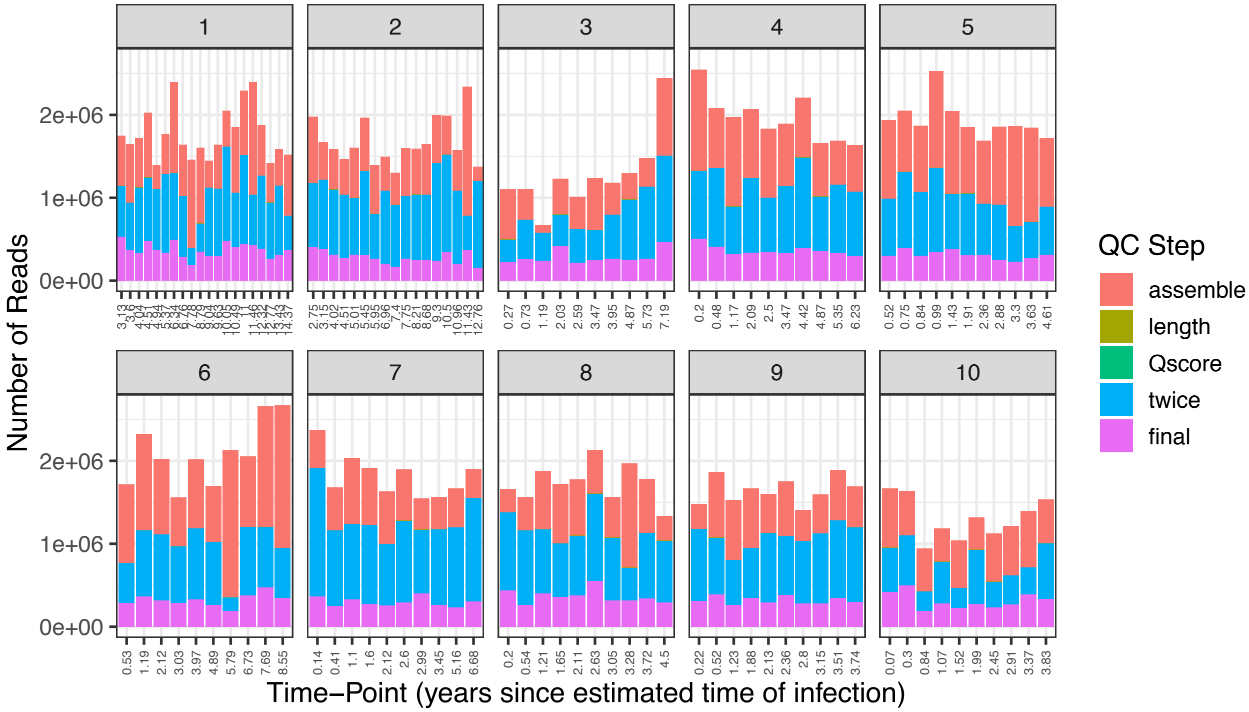


Figure 4.6. Read depth and QC filtering for AbR data.
Stacked bar charts depicting the total read depth and proportion of reads that were filtered after each QC step. The height of each bar shows initial read depth (directly after sequencing) for each AbR sample. The relative lengths of each color within a bar show how many reads were filtered out from a given QC step. The legend maps a color to a QC step, and QC steps are shown (from top to bottom) in the order that they were performed: "assemble" = could not assemble read pairs; "length" = too short; "Qscore" = mean Q score too low; "twice" = sequence did not occur at least twice; "final" = the final sequences that passed all QC. Numeric titles for each panel give the patient ID.

smaller time-scales (Figures 4.7, 4.8, and 4.9). Of note, the high viral load of the first time-point of patient 7 suggests that the acute viremia phase of early HIV infection was captured, and, as noted previously, patient 10 had very low viral load for the first 8 time-points, which explains why amplification of HIV *env* was unsuccessful for these samples. Patient 6 exhibited strong evidence for a super-infection occurring between the 1st and 2nd time-points (Figure 4.10). Super infections are not uncommon with HIV [37], however they will cause a sudden injection of 'artificial' genetic divergence relative to the initial infecting virus. Thus, we accounted for this superinfection when calculating divergence summary statistics for patient 6 (see Methods).

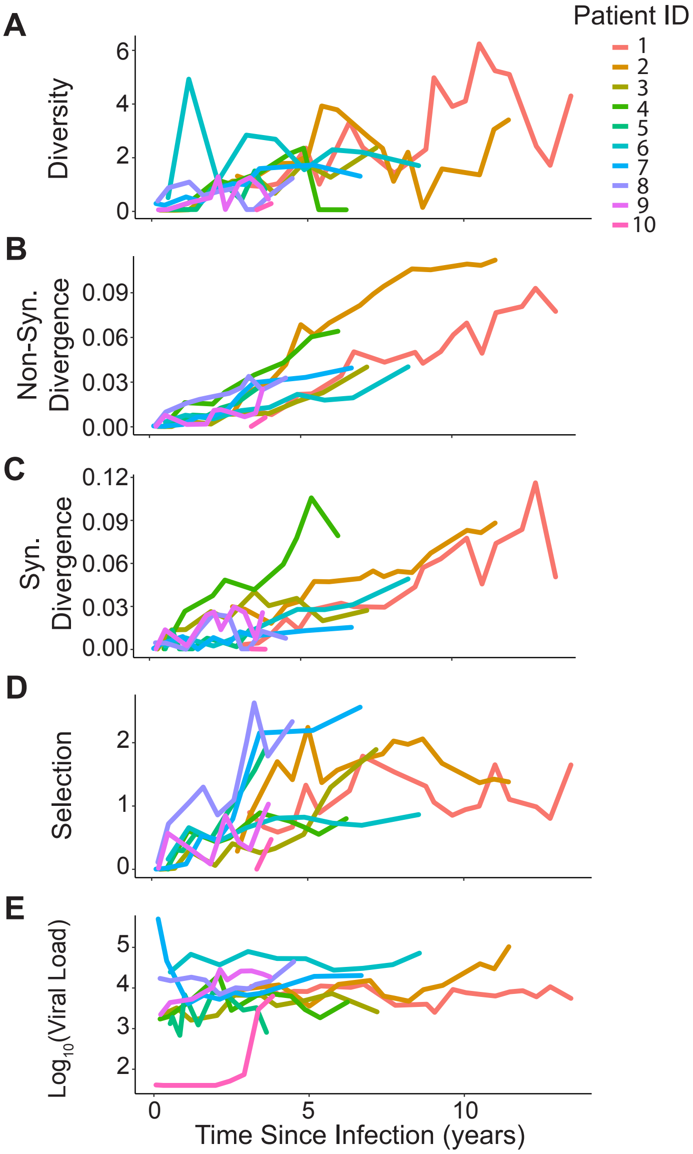


Figure 4.7. HIV summary statistic trajectories foreach patient.
Each line shows the trajectory of a given summary statistic, in a given patient over time. Each patient has a unique color. (A) Diversity. (B) nonsynonymous divergence. (C) synonymous divergence. (D) Selection. (E) Viral load.

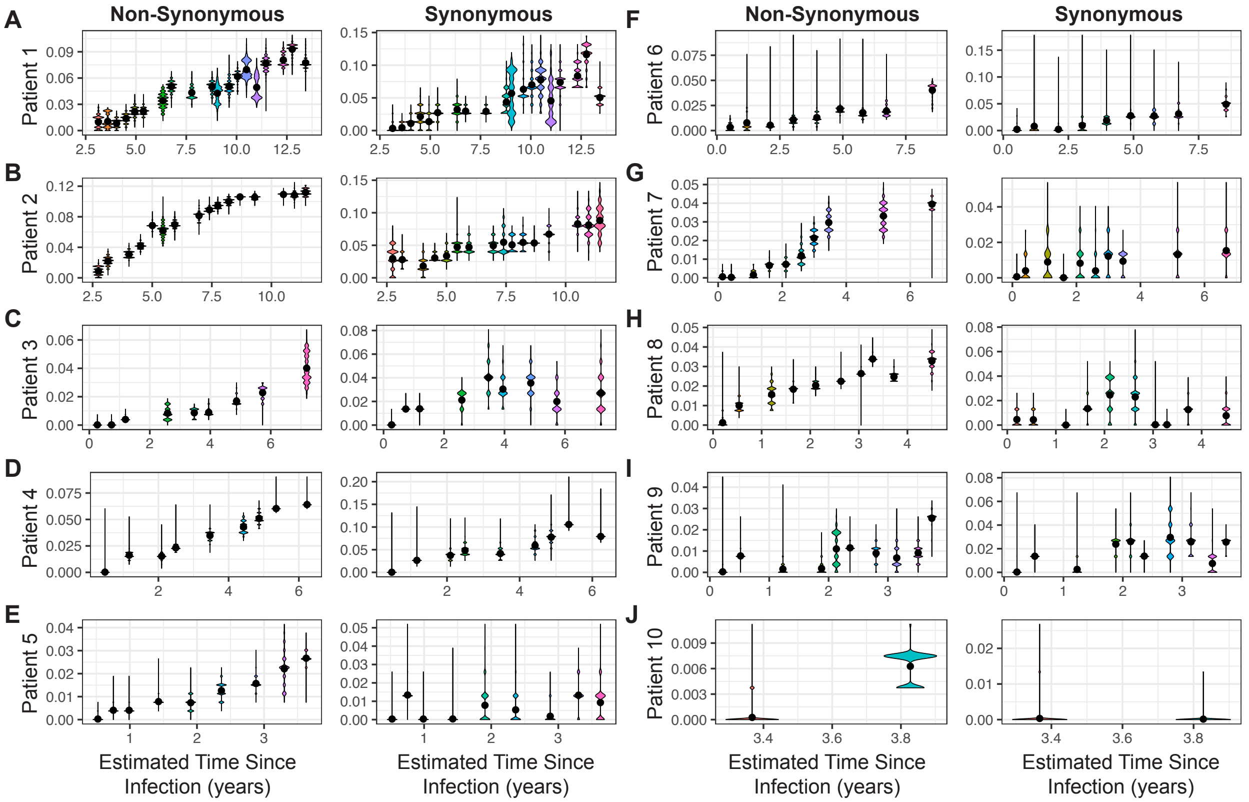


Figure 4.8. Distributions of HIV divergence values.

Each violin plot gives the distribution of HIV divergence values for a given time-point, where black points show the mean of the distribution. Plots on the left side of a panel show non-synonymous divergence, and plots on the right show synonymous divergence. (A) Patient 1. (B) Patient 2. (C) Patient 3. (D) Patient 4. (E) Patient 5. (F) Patient 6. (G) Patient 7. (H) Patient 8. (I) Patient 9. (J) Patient 10.

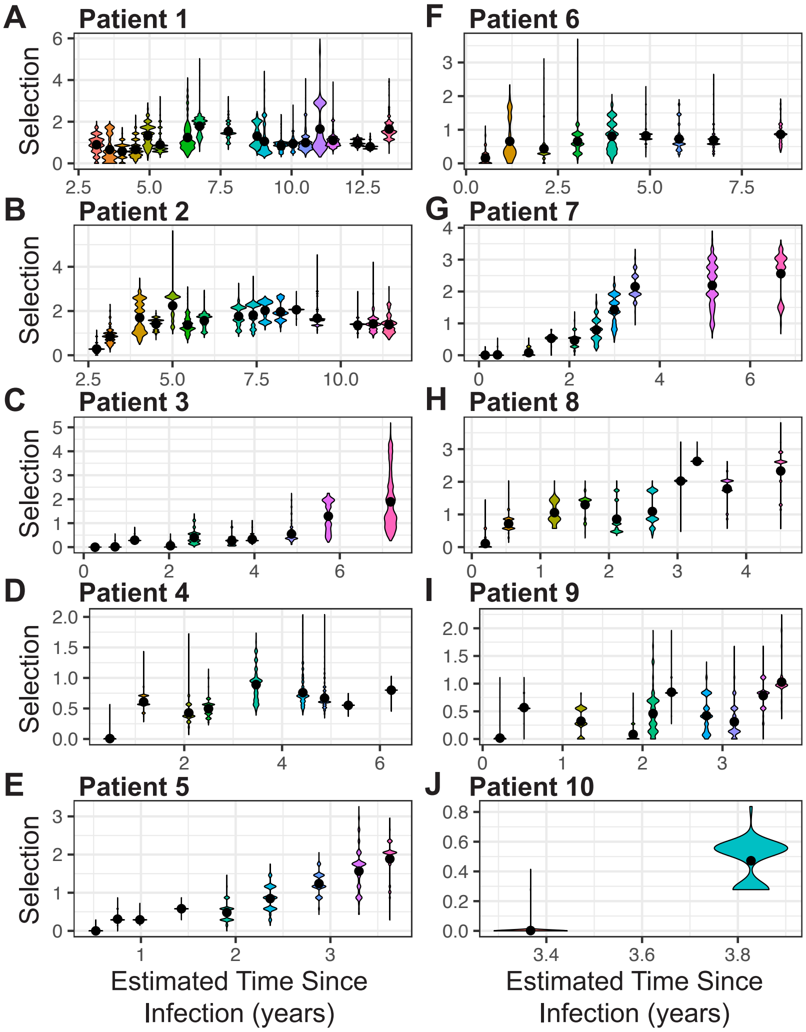


Figure 4.9. Distributions of HIV selection values.

Each violin plot gives the distribution of HIV selection values at a given time-point, where black points show the mean of the distribution. (A) Patient 1. (B) Patient 2. (C) Patient 3. (D) Patient 4. (E) Patient 5. (F) Patient 6. (G) Patient 7. (H) Patient 8. (I) Patient 9. (J) Patient 10.

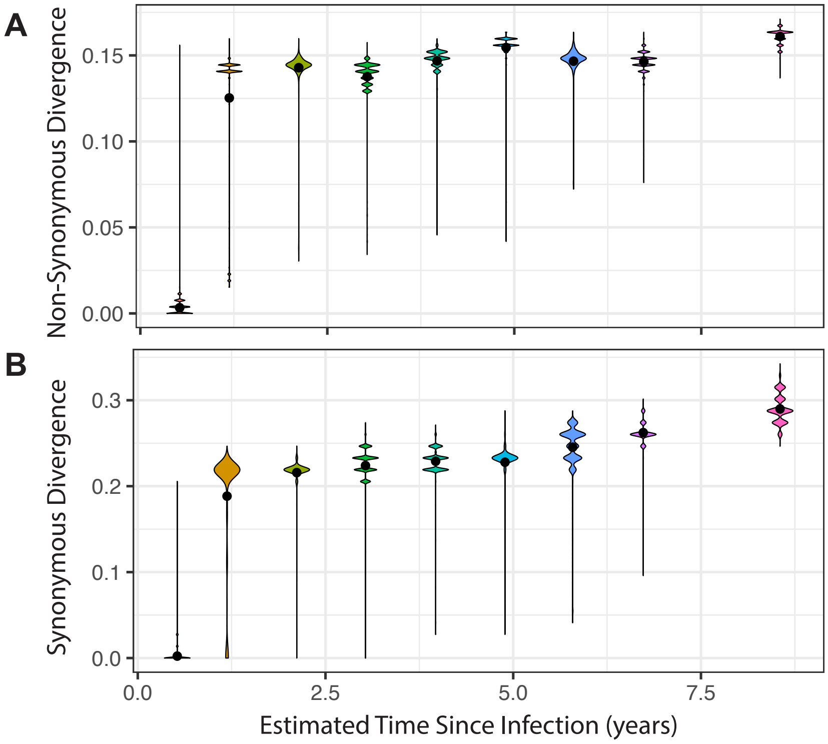


Figure 4.10. Super infection in patient 6.
Each violin plot gives the distribution of HIV divergence values for a given time-point for patient 6, where black points show the mean of the distributions. Divergence was calculated here in the same way as other patients in order to illustrate super-infection. (A) Non-synonymous divergence. (B) Synonymous divergence. The large increases in both synonymous and non-synonymous divergence between the 1st and 2nd time-points are indicative of a super-infection.

The trajectories of the AbR summary statistics did not show any obvious stereotyped pattern across patients (Figures 4.11, 4.12, and 4.13). However, there were a couple data points that suggested interactions with the HIV population: the second time-point of patient 7 (0.41 years post infection) showed a large increase in selection (in both the framework regions, FWR, and complementarity determining regions, CDR), which could be a response to the initial viremia in the prior time-point; and the ninth time-point of patient 10 (3.37 years post infection) also showed a large increase in selection with a concomitant drop in diversity, which could have been a response to the large increase in viral load at that time-point.

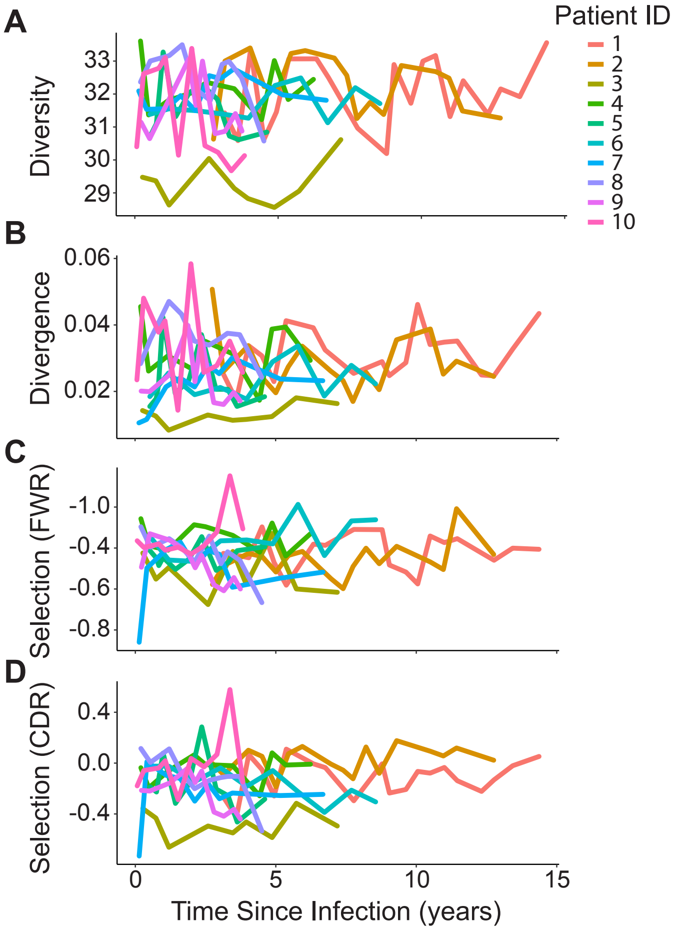


Figure 4.11. AbR summary statistic trajectories foreach patient.
Each line shows the trajectory of a given summary statistic, in a given patient over time. Each patient has a unique color. (A) Diversity. (B) Divergence. (C) Selection in the FWR. (D) Selection in the CDR.

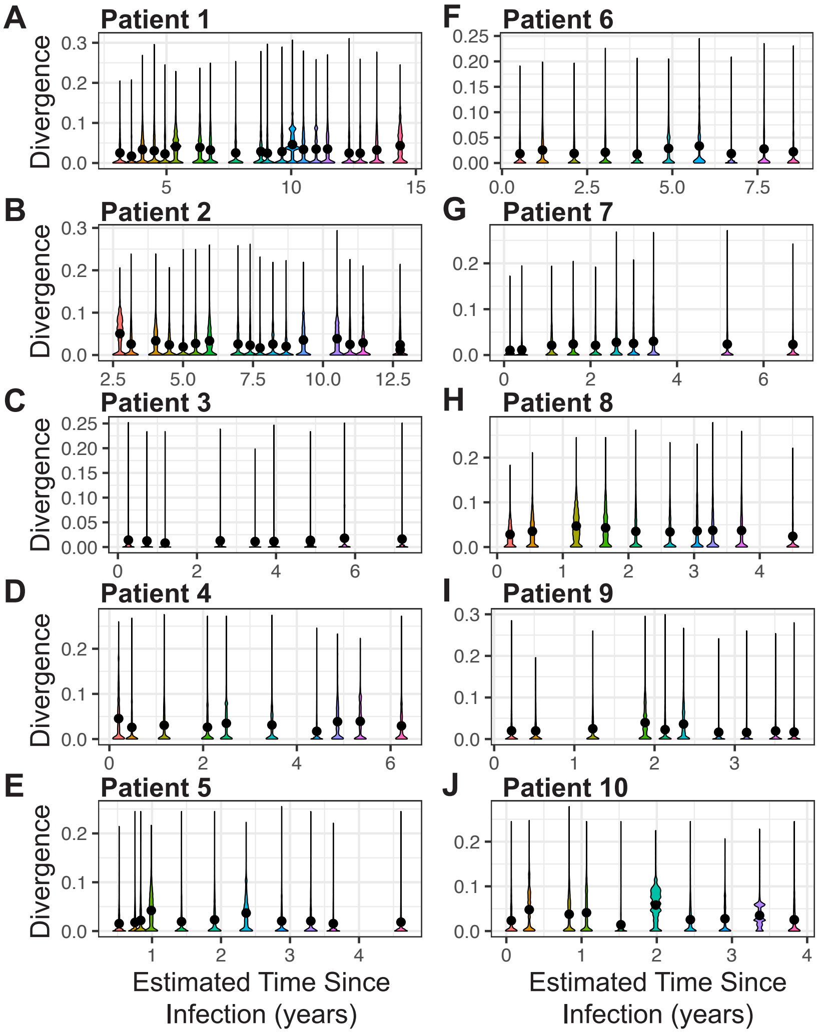


Figure 4.12. Distributions of antibody divergence values.

Each violin plot gives the distribution of antibody divergence values for a given time-point, where black points show the mean of the distribution. (A) Patient 1. (B) Patient 2. (C) Patient 3. (D) Patient 4. (E) Patient 5. (F) Patient 6. (G) Patient 7. (H) Patient 8. (I) Patient 9. (J) Patient 10.

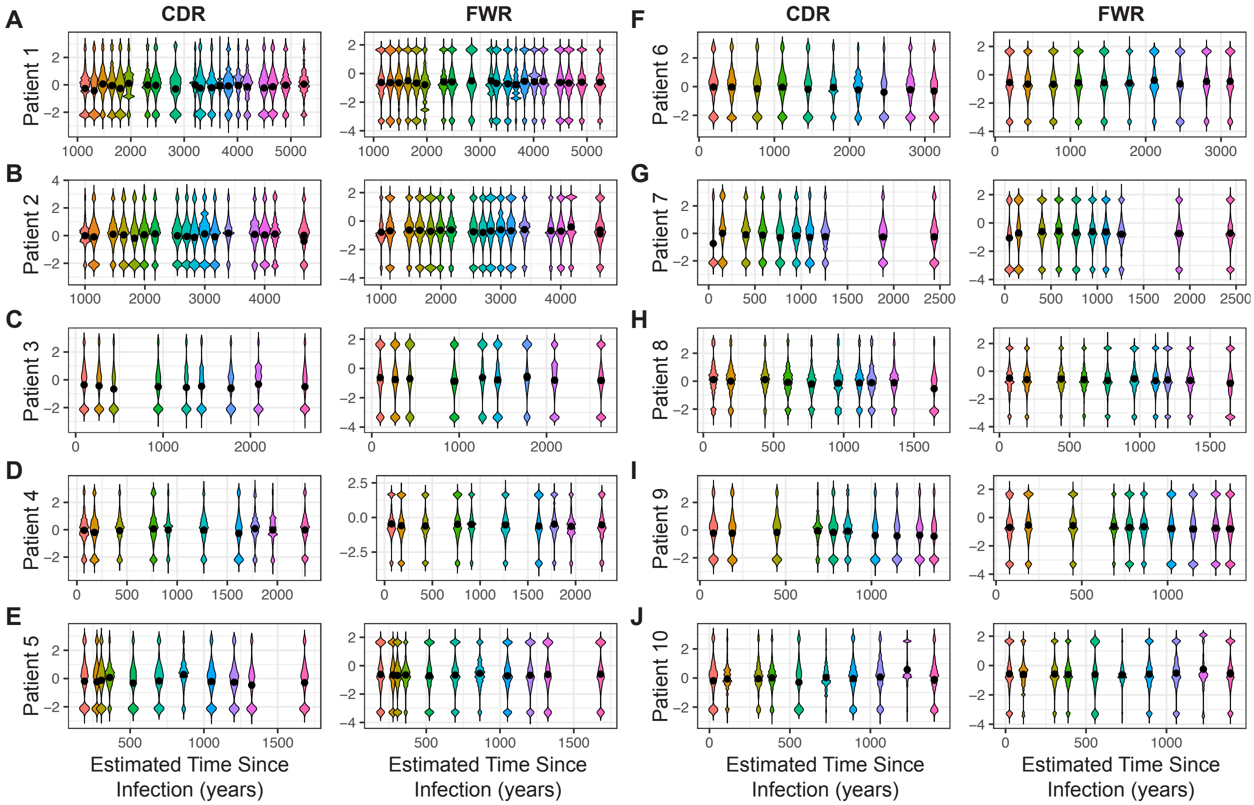


Figure 4.13. Distributions of AbR selection values.
Each violin plot gives the distribution of AbR selection values at a given time-point, where black points show the mean of the distribution. Plots on the left side of a panel show selection in the CDR, and plots on the right show selection in the FWR. (A) Patient 1. (B) Patient 2. (C) Patient 3. (D) Patient 4. (E) Patient 5. (F) Patient 6. (G) Patient 7. (H) Patient 8. (I) Patient 9. (J) Patient 10.

<u>Testing for whole-population level interactions</u>

In order to understand how much of an effect HIV generally has on the AbRs of patients, we first pooled all the data across patients and used a regression framework to test if any of the AbR summary statistics were significantly correlated with that of the HIV population (while controlling for patient-specific effects, see Methods). We performed this test in a pairwise fashion on all AbR summary statistics against all HIV summary statistics. Similar to Hoehn's work [13], we found no significant correlation between AbR diversity and viral load, yet we did find a small correlation between AbR and HIV diversity (p= 0.02, Figure 4.14 A). While this association was small and marginally significant—indeed, it ceases to be significant after controlling for multiple tests—we found any association at the whole-population level surprising, and thus, worth reporting.

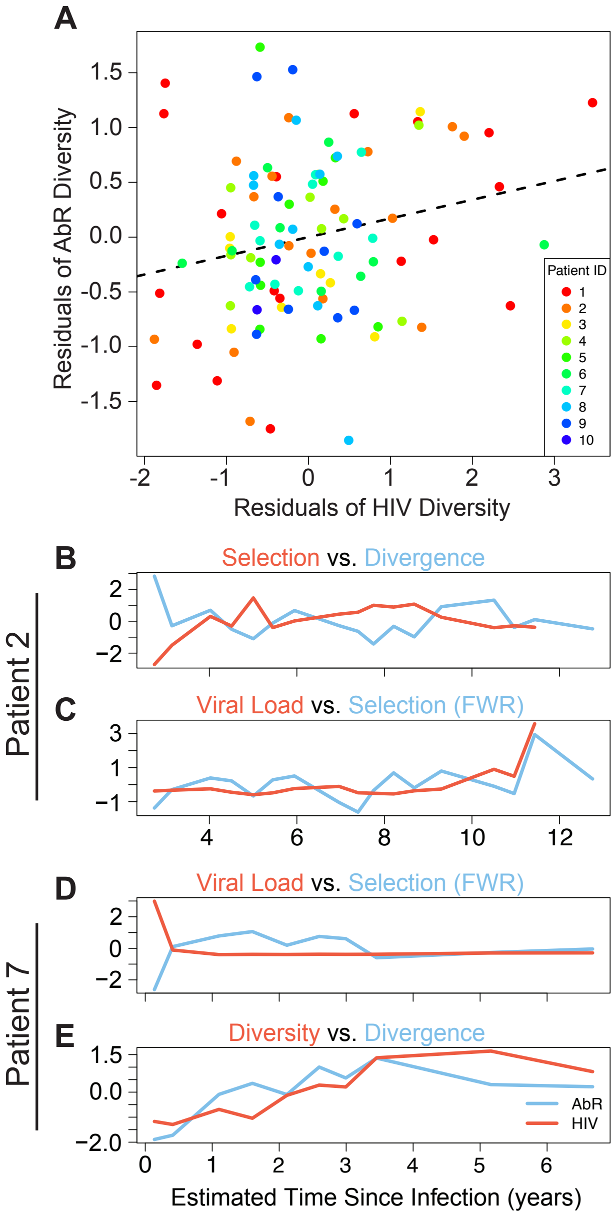


Figure 4.14. Whole population level associations between AbR and HIV summary statistics. (A) Scatter plot showing positive correlation between HIV diversity (x-axis) and AbR diversity (y-axis). Each point represents a sample with diversity values from both the AbR and HIV sequence data. The axes show diversity values after patient specific effects have been regressed out. Dashed line shows positive relationship between AbR and HIV diversity, as given by our linear regression (see Methods). (B-E) Shows associations between summary statistic trajectories in AbR (blue) and HIV (red) at the individual patient level. (B) HIV selection with AbR Divergence, in patient 2. (C) Viral load with AbR FWR-selection, in patient 2. (D) Viral load with AbR FWR-selection, in patient 7. (E) HIV diversity with AbR divergence, in patient 7.

It is possible that some patients' AbRs interact with their autologous HIV population more than others, thus we also tested for interactions between summary statistics on a patient-by-patient approach. Because the number of data-points per patient is relatively low, we opted to use a permutation based test in order to accurately estimate type 1 error [38] (see Methods). In patient 2, We found that selection and viral load in the HIV population were associated with divergence and selection (FWR) in the AbR, respectively (p=0.009 and p=0.0425, Figure 4.14 B, and C). We also found that in patient 7, viral load and diversity in the HIV population were associated with selection (FWR) and divergence in the AbR, respectively (p=0.040 and p=0.040, Figure 4.14 D, and E). We note that all of these associations were positive correlations, with the exception of HIV selection and AbR divergence in patient 2, which was anticorrelated.

Together, these data suggest that a large proportion of the AbR may be responding to the HIV infection, and that this response can be detected in AbR sequence data.

Identifying partitions of the AbR that interact with HIV

The AbR is an exceedingly complex population consisting of a myriad of Ab lineages capable of simultaneously binding and neutralizing a countless number of antigenic targets. In order to reduce this complexity, and to identify specific parts of the AbR that may be interacting with HIV, we first partitioned the AbR across time based on the germline identity of each sequence's V and J gene segments (Figure 4.15), and then tested each AbR partition for evidence of interactions with the autologous HIV population using similar summary statistics as the overall population (see Methods). Using an analogous permutation-based test as was used when comparing the overall

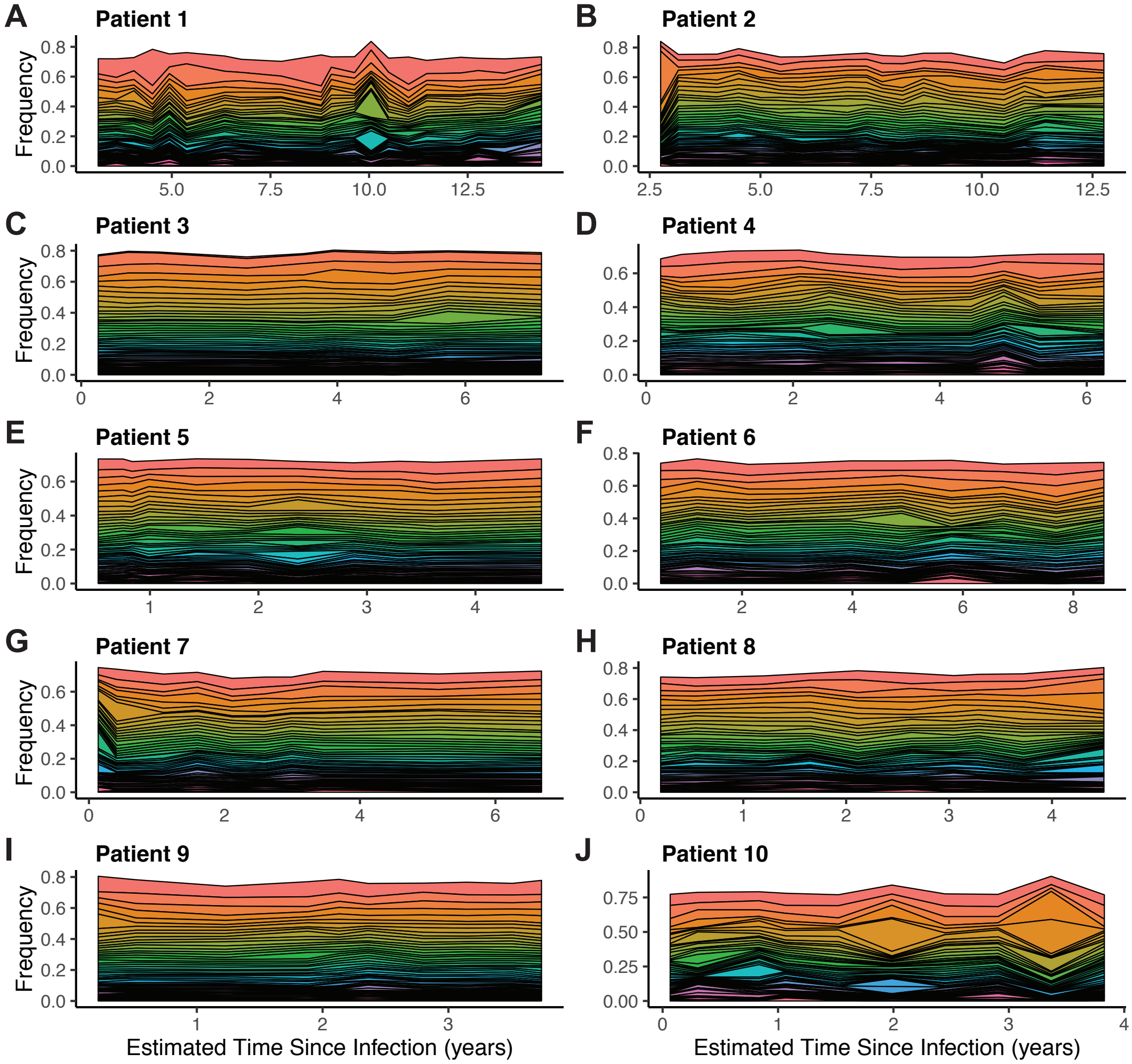


Figure 4.15. AbR partition frequencies over time.
Stacked area charts of AbR partition frequencies over time. Each unique color represents a unique AbR partition, and the width of a given color represents the relative frequency of that AbR partition at that time-point. Only the top 50 AbR partitions are shown, when sorted by the sum of frequencies across all time-points. (A) Patient 1. (B) Patient 2. (C) Patient 3. (D) Patient 4. (E) Patient 5. (F) Patient 6. (G) Patient 7. (H) Patient 8. (I) Patient 9. (J) Patient 10.

populations, we found significant associations between AbR-partition trajectories and HIV trajectories in patients 3, 7, and 8 (Figures 4.16, 4.17, and 4.18). Of these associations, AbR-partition frequency tended to be associated with viral load. For example, patient 7 had a distinct viral load trajectory—presumably due to acute viremia—and the frequency trajectory of IGHV4-31:IGHJ5 AbR partition was positively associated with the unique shape of this trajectory (while the diversity of this AbR partition was negatively associated with viral load). This suggests a clonal expansion occurred in this AbR-partition in response to HIV, which caused an increase in the partition's frequency with a concomitant drop in diversity. Similarly, the frequency trajectory of two AbR partitions with the same V gene segment (IGHV6-1:IGHJ5 and IGHV6-1:IGHJ4) were positively associated with viral load in patient 8, suggesting that the IGHV6-1 gene segment in this patient may have had a predisposition to targeting HIV. In patient 3, the frequency trajectory of the IGHV3-30:IGHJ3 partition was negatively associated with both non-synonymous divergence and selection in the HIV population, suggesting that escape mutations in HIV have caused a drop in frequency of the interacting AbR partition.

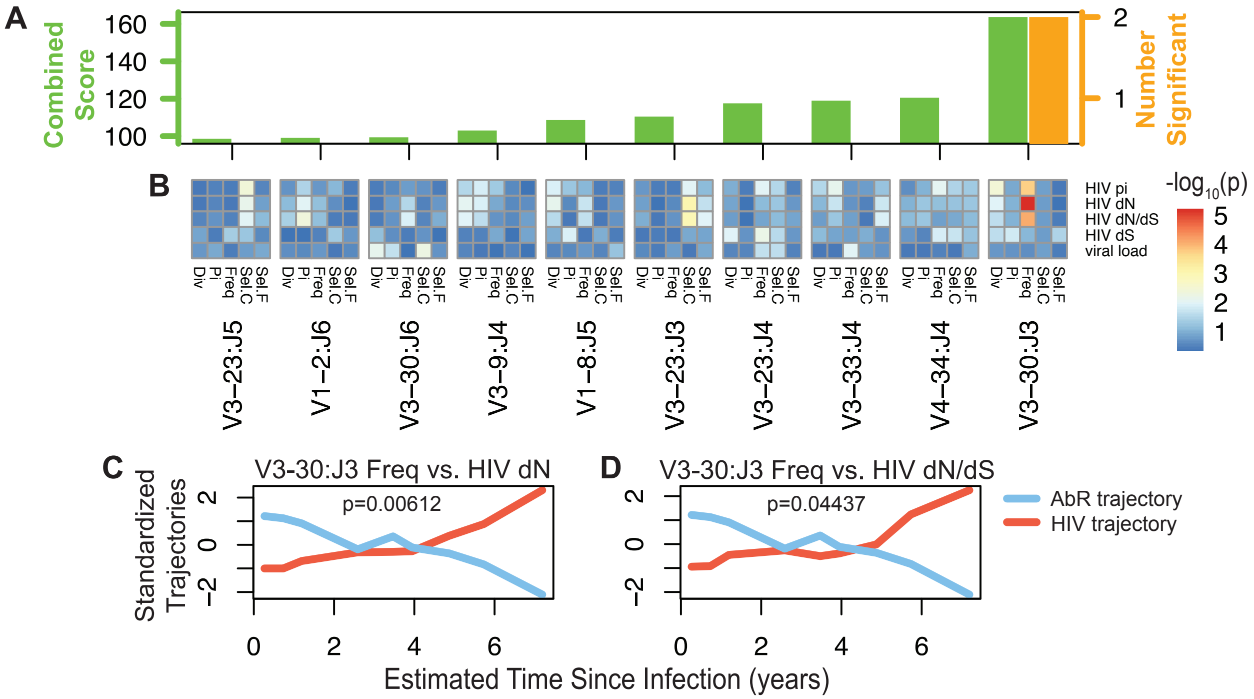


Figure 4.16. Results of permutation-based test to identify HIV-associated AbR partitions in patient 3.
(A) A barplot showing the combined score from the permutation-based test (left axis), and the number of significant associations (right axis) for the top 10 AbR partitions. AbR partitions were sorted first by the number of significant associations, then by their combined score from the permutation-based test (ascending, left-right). (B) Heatmaps depicting the significance (-log10(p-value)) foreach test run within each of the top 10 AbR partitions. Columns correspond to summary statistics of the AbR partitions: Div=divergence, Pi=diversity, Freq=relative frequency, Sel.C=CDR selection, and Sel.F=FWR selection. Rows correspond to summary statistics of the HIV population: HIV pi=diversity, HIV dN=nonsynonymous

divergence, HIV dS=synonymous divergence, HIV dN/dS=selection, and viral load is self-explanatory. The color of each element in the heatmaps shows the significance of the association between a given AbR partition summary statistic trajectory with a given HIV population summary statistic trajectory. (C, D) Shows the AbR partition (blue) and HIV population (red) trajectories that were significantly associated. (C) Frequency trajectory of the IGHV3-30:IGHJ3 AbR partition with the divergence trajectory of the HIV population. (D) Frequency trajectory of the IGHV3-30:IGHJ3 AbR partition with the selection trajectory of the HIV population.

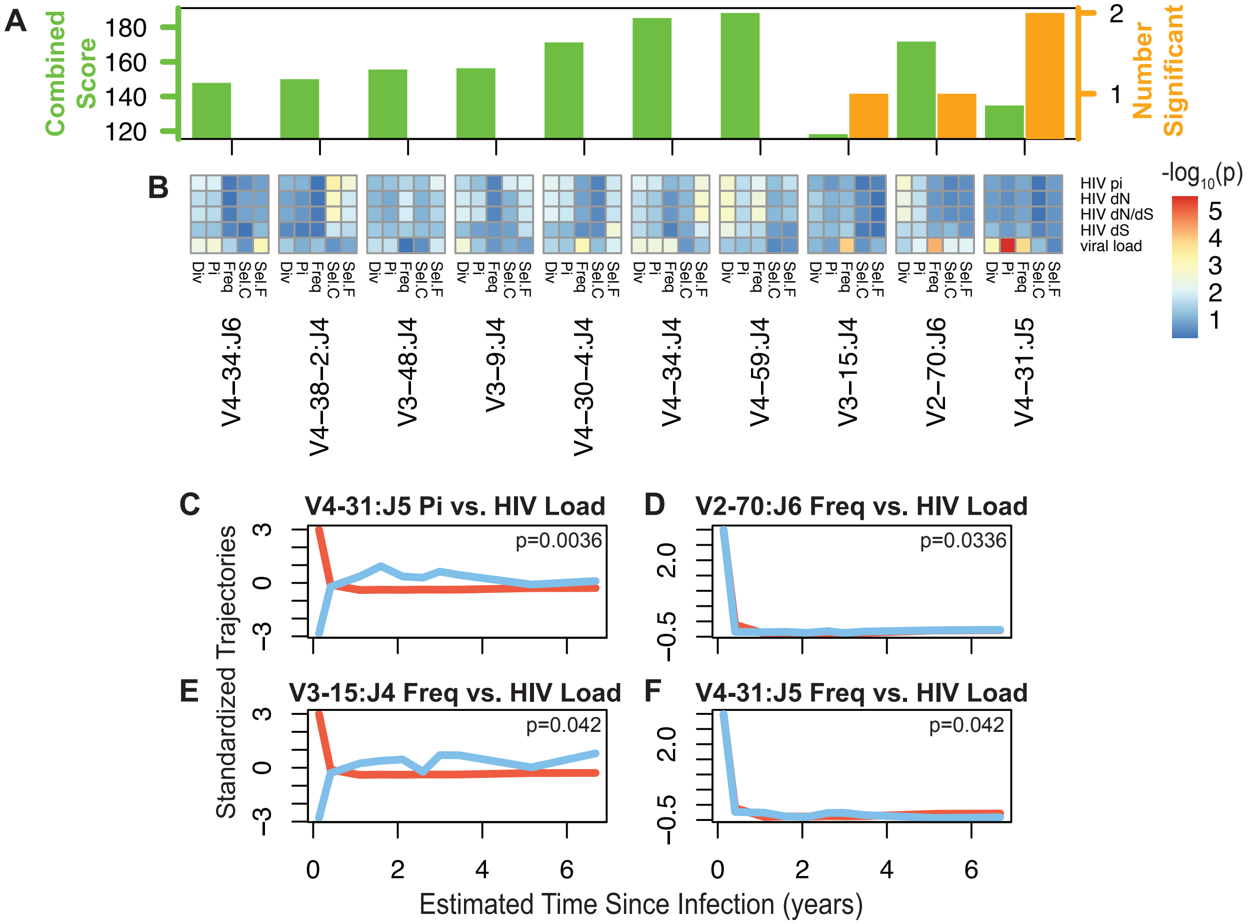


Figure 4.17. Results of permutation-based test to identify HIV-associated AbR partitions in patient 7.

(A) A barplot showing the combined score from the permutation-based test (left axis), and the number of significant associations (right axis) for the top 10 AbR partitions. AbR partitions were sorted first by the number of significant associations, then by their combined score from the permutation-based test (ascending, left-right). (B) Heatmaps depicting the significance (-log$_{10}$(p-value)) foreach test run within each of the top 10 AbR partitions. Columns correspond to summary statistics of the AbR partitions: Div=divergence, Pi=diversity, Freq=relative frequency, Sel.C=CDR selection, and Sel.F=FWR selection. Rows correspond to summary statistics of the HIV population: HIV pi=diversity, HIV dN=nonsynonymous divergence, HIV dS=synonymous divergence, HIV dN/dS=selection, and viral load is self-explanatory. The color of each element in the heatmaps shows the significance of the association between a given AbR partition summary statistic trajectory with a given HIV population summary statistic trajectory. (C-F) Shows the AbR partition (blue) and HIV population (red) trajectories that were significantly associated. (C) Diversity trajectory of the IGHV4-31:IGHJ5 AbR partition with the viral load trajectory of the HIV population. (D) Frequency trajectory of the IGHV2-70:IGHJ6 AbR partition with the viral load trajectory of the HIV population. (E) Frequency trajectory of the IGHV3-15:IGHJ4 AbR partition with the viral load trajectory of the HIV population. (F) Frequency trajectory of the IGHV4-31:IGHJ5 AbR partition with the viral load trajectory of the HIV population.

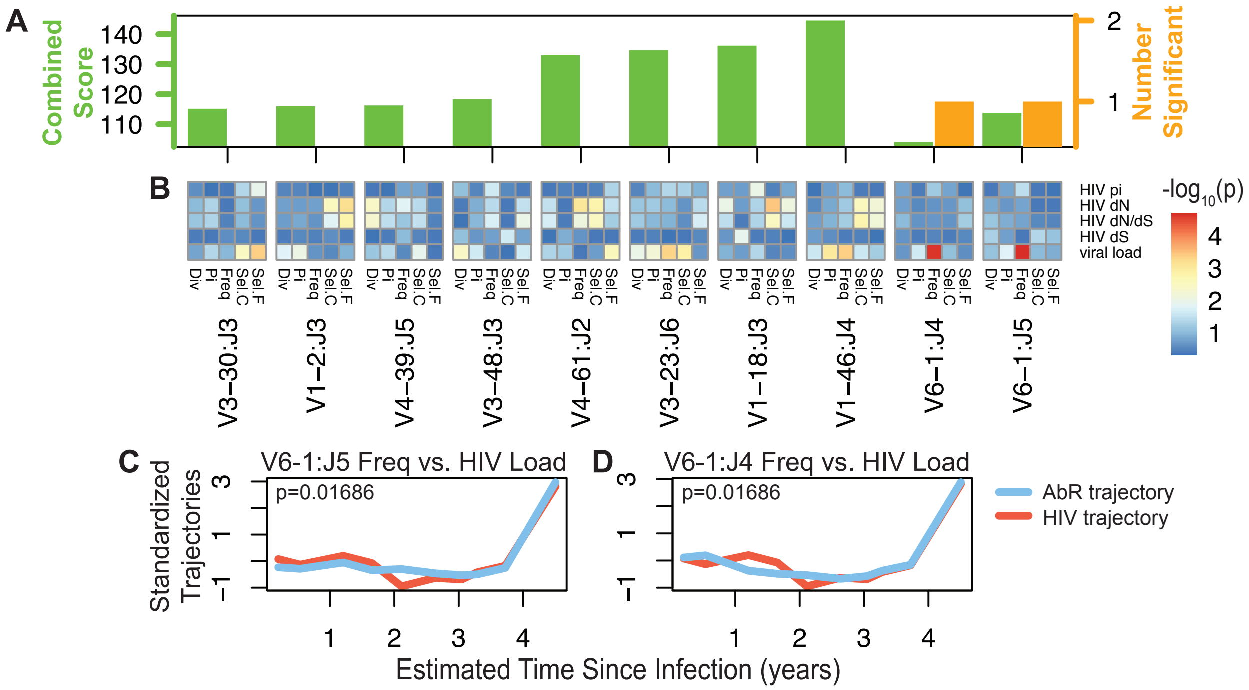


Figure 4.18. Results of permutation-based test to identify HIV-associated AbR partitions in patient 8.

(A) A barplot showing the combined score from the permutation-based test (left axis), and the number of significant associations (right axis) for the top 10 AbR partitions. AbR partitions were sorted first by the number of significant associations, then by their combined score from the permutation-based test (ascending, left-right). (B) Heatmaps depicting the significance (-log$_{10}$(p-value)) foreach permutation test within each of the top 10 AbR partitions. Columns correspond to summary statistics of the AbR partitions: Div=divergence, Pi=diversity, Freq=relative frequency, Sel.C=CDR selection, and Sel.F=FWR selection. Rows correspond to summary statistics of the HIV population: HIV pi=diversity, HIV dN=nonsynonymous divergence, HIV dS=synonymous divergence, HIV dN/dS=selection, and viral load is self-explanatory. The color of each element in the heatmaps shows the significance of the association between a given AbR partition summary statistic trajectory with a given HIV population summary statistic trajectory. (C, D) Shows the AbR partition (blue) and HIV population (red) trajectories that were significantly associated. (C) Frequency trajectory of the IGHV6-1:IGHJ5 AbR partition with the viral load trajectory of the HIV population. (D) Frequency trajectory of the IGHV6-1:IGHJ4 AbR partition with the viral load trajectory of the HIV population.

Validating the HIV-associated AbR partitions

In order to establish that our permutation-based test is in fact identifying AbR partitions that had a biological response to the HIV infection and was not the result of random noise in the data, we sought to compare our results to previous findings in the literature. We first used Fisher's method to compile all the results of our permutation-based test into a single score for each V gene segment (see Methods), and then used the database of HIV bnAbs from bNAber [39] to compare the incidence of known HIV-binding V gene segments in the literature, to how well these V gene segments score in our test (Table 4.2). We found that V gene segments that had been shown to bind HIV

Table 4.2. Literature datasets.

Gene Name	V Gene Score	HIV Literature Dataset Count	Flu Literature Dataset Count
IGHV4-59	5.692163182	5	71
IGHV1-18	5.135748975	1	14
IGHV3-23	4.069059968	0	48
IGHV3-30-3	3.996480894	0	0
IGHV4-39	3.931435639	7	37
IGHV3-30	3.826571542	14	18
IGHV4-61	3.825996333	0	1
IGHV1-46	3.708051547	0	0
IGHV1-8	3.705576595	18	0
IGHV3-21	3.301116439	1	40
IGHV3-11	3.233458041	0	2
IGHV3-33	3.020795329	1	3
IGHV3-66	2.99037641	0	0
IGHV3-74	2.801852721	0	0
IGHV3-7	2.769498585	0	19
IGHV4-34	2.036779778	3	7
IGHV1-2	2.029337852	18	7
IGHV1-3	2.018780447	1	0
IGHV6-1	1.922879261	0	0
IGHV1-69	1.809503004	2	108
IGHV2-5	1.723749802	1	8
IGHV4-4	1.672966404	0	10
IGHV3-15	1.532535707	1	2
IGHV3-9	1.532147253	0	5
IGHV3-53	1.446589506	0	1
IGHV1-24	1.298021556	0	0
IGHV7-4-1	1.217541317	0	2
IGHV5-51	1.212203964	0	7
IGHV3-48	1.206518076	0	9
IGHV4-30-2	1.189591763	0	0
IGHV3-13	1.123988038	0	6
IGHV3-64	0.851467542	0	1
IGHV4-30-4	0.790029158	0	3
IGHV4-31	0.708777704	0	2
IGHV3-72	0.298207811	0	0
IGHV2-70D	0.220188607	0	0
IGHV3-43	0.209900779	0	0
IGHV2-70	0.115635873	0	1
IGHV4-38-2	0.103112693	0	0
Total	NA	**73**	**432**

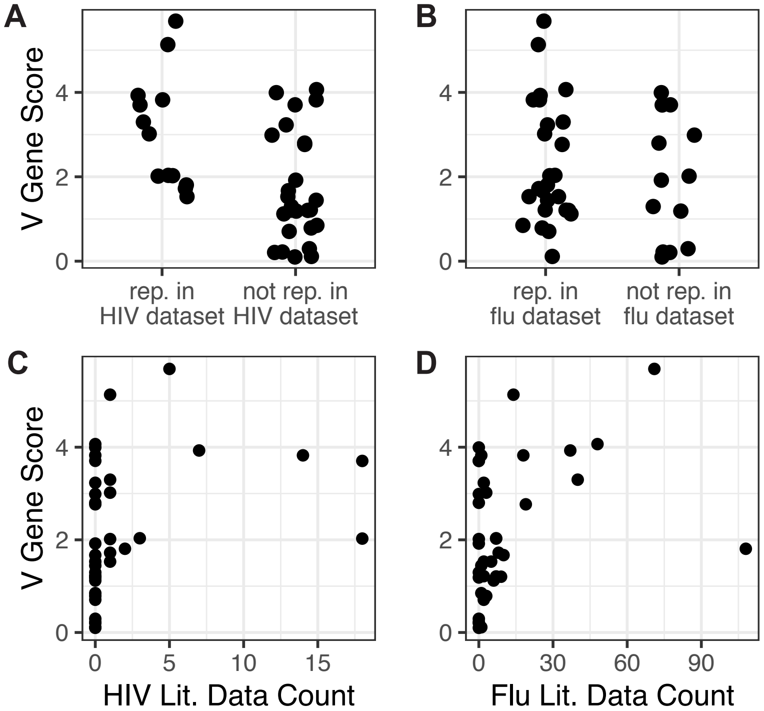


Figure 4.19. Literature validation of HIV-associated AbR partitions.
(A, B) Shows the overall score from our permutation-based test for each V gene segment when partitioned as being represented in a literature dataset (points to the left), vs. not represented in a literature dataset (points to the right). (A) Pertains to the HIV literature dataset, where ‘representation’ is defined as the V gene segment occurring at least once. (B) Pertains to the influenza literature dataset, where ‘representation’ is defined as the V gene segment occurring at least 10 times. (C, D) Scatter plots showing the relationship between the overall score from our permutation-based test for each V gene segment (y-axis) and the count of a given V gene segment in a literature dataset (x-axis). (C) HIV literature dataset. (D) Influenza literature dataset.

in the literature dataset, tended to score higher in our permutation-based test (p=1.52e-3, Mann-Whitney U test, Figure 4.19 A, and C). This suggests that our test is indeed identifying a biological response to HIV. It also suggests that some V gene segments may have a predisposition to bind HIV. However, a simpler—and perhaps less exciting—explanation is that these V gene segments have a predisposition to bind anything (either due to high endogenous expression, being ‘sticky’, or something else). In order to differentiate between these two possibilities we performed a similar test except instead of comparing our results to Abs known to target HIV, we compared our results to a literature dataset that we previously compiled of Abs that have been shown to bind to influenza [40] (Table 4.2). Similar to the HIV literature dataset, we found that

V genes that were well represented in the influenza literature dataset also tended to score highly in our permutation-based test (p=7.28e-4, Mann-Whitney U test, Figure 4.19 B, and D). This suggests that, while we are likely identifying a biological response to HIV, the response may not be specific to HIV.

Testing for coevolution

Coevolution between HIV and a handful of well-known bnAbs antibodies has been extensively reported [12,15] and reviewed [41]. Coevolution provides an intellectually compelling explanation for the development of bnAbs against HIV, however, examples tend to be anecdotal and qualitative (likely due to small sample sizes). While we cannot be sure that bnAbs exist in our data, we sought to test if coevolution is a predominant driver of HIV-targeting Ab development generally. We tested for genetic signals of coevolution in our data by first dividing the AbR data of each patient into time-course lineages of Abs (Figure 4.20, and 4.21). We then use MAFFT to create a multiple sequence alignment (MSA) of each Ab lineage, and compare each of these Ab lineage MSAs with a representative MSA of the HIV population overtime using a mutual information (MI) statistic. Importantly, we reduce the complexity of the amino acid code to a code of 'change' or 'no-change' prior to calculating MI (see Methods) [42–45].

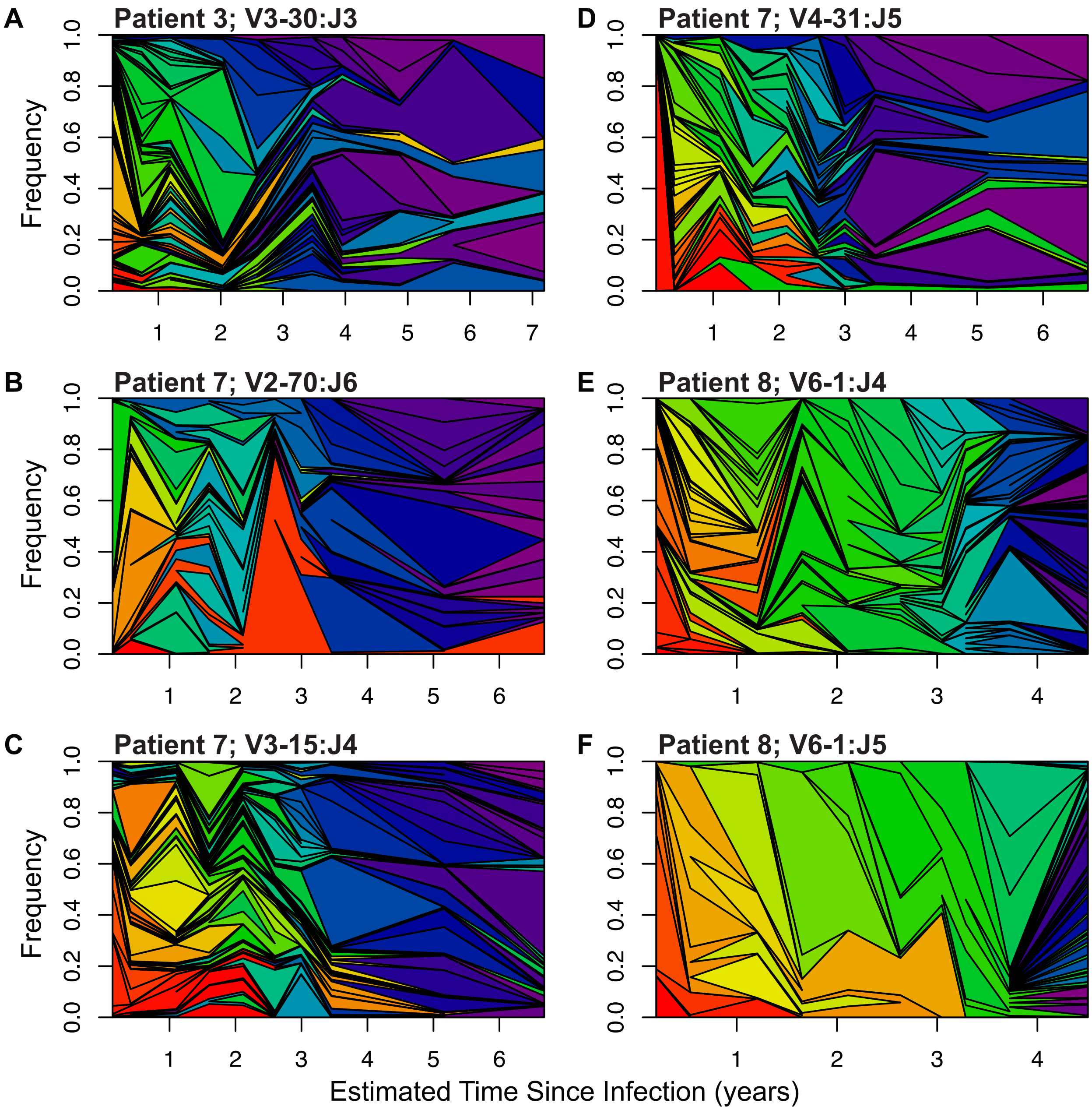
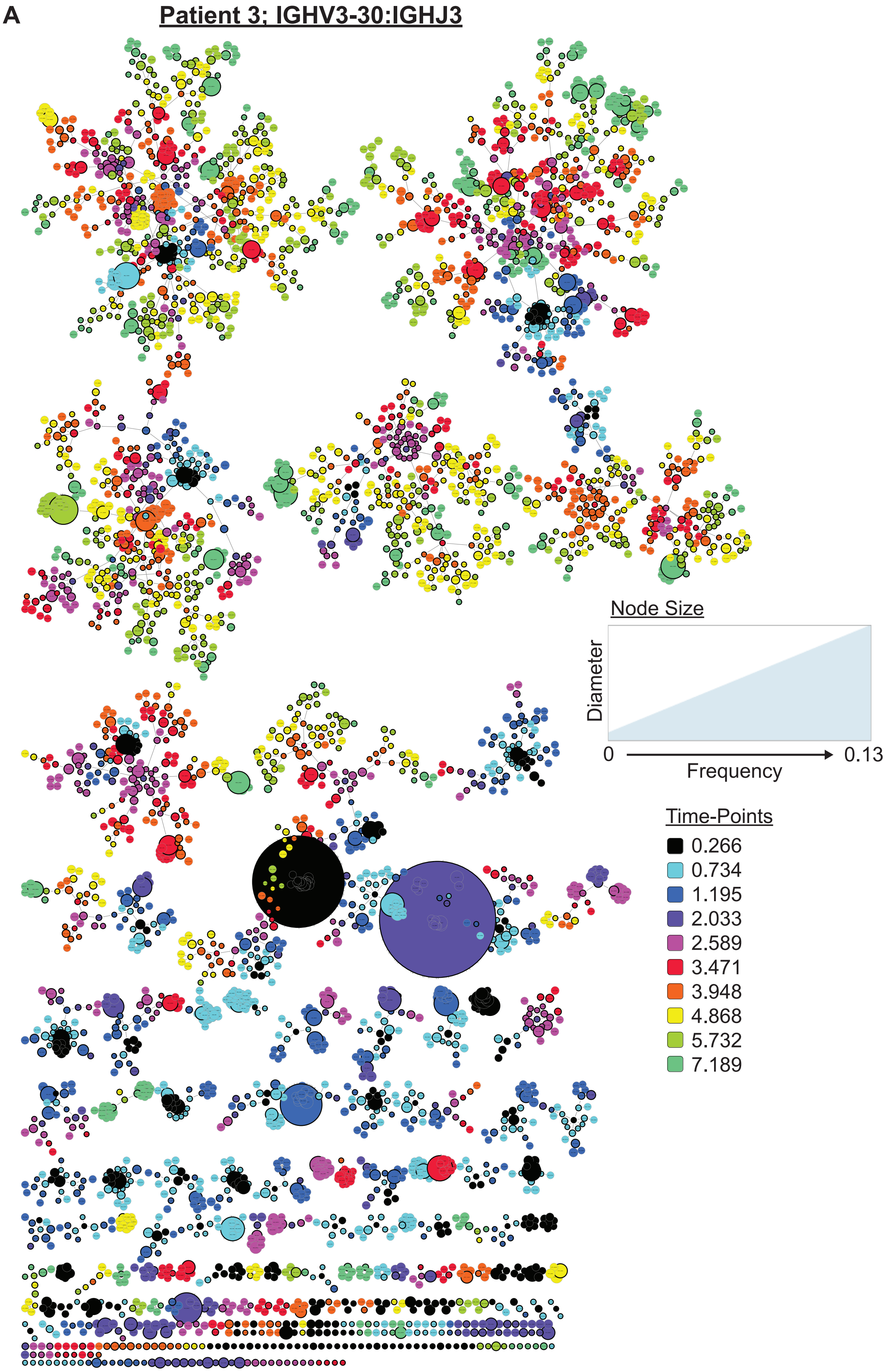
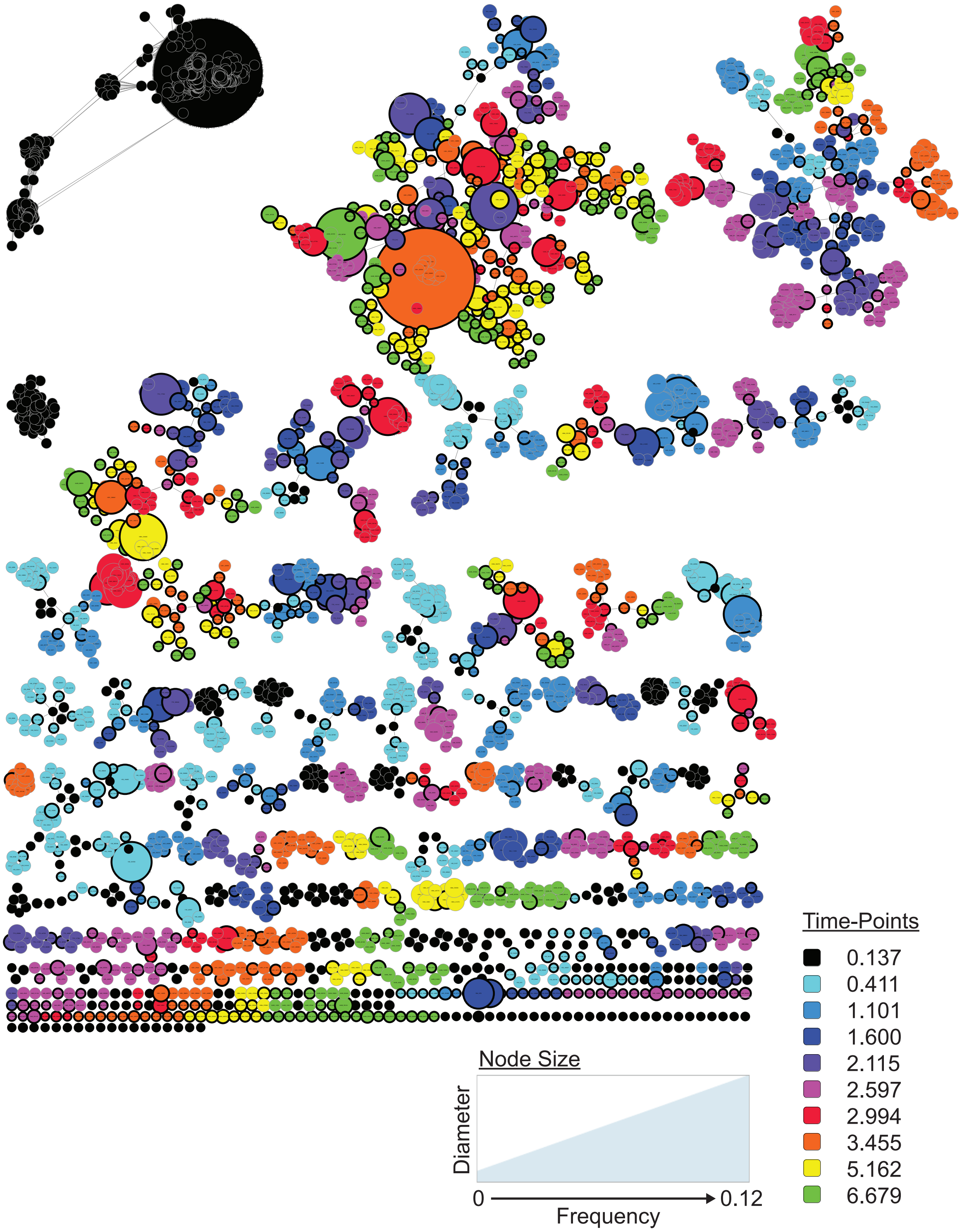


Figure 4.20. Muller plots of Ab lineages.
Depicts the relative frequency of each lineage within a given HIV-associated AbR partition. Each unique color represents a unique lineage. If a new lineage arises within the bounds of a preexisting lineage, then the new lineage is a daughter of the preexisting, parent lineage. Lineages that began earlier in the time-course have colors closer to the red side of the spectrum, while lineages that began later in the time-course have colors closer to violet. Only lineages that exceeded 0.0001 frequency in the larger AbR for at least one time-point are included in the plot. (A) AbR partition IGHV3-30:IGHJ3 in patient 3. (B) AbR partition IGHV2-70:IGHJ6 in patient 7. (C) AbR partition IGHV3-15:IGHJ4 in patient 7. (D) AbR partition IGHV4-31:IGHJ5 in patient 7. (E) AbR partition IGHV6-1:IGHJ4 in patient 8. (F) AbR partition IGHV6-1:IGHJ5 in patient 8.


A
Patient 3; IGHV3-30:IGHJ3
Node Size
Diameter
0
Frequency
0.13
Time-Points
0.266
0.734
1.195
2.033
2.589
3.471
3.948
4.868
5.732
7.189

B

Patient 7; IGHV4-31:IGHJ5



C

Patient 6; IGHV6-1:IGHJ5

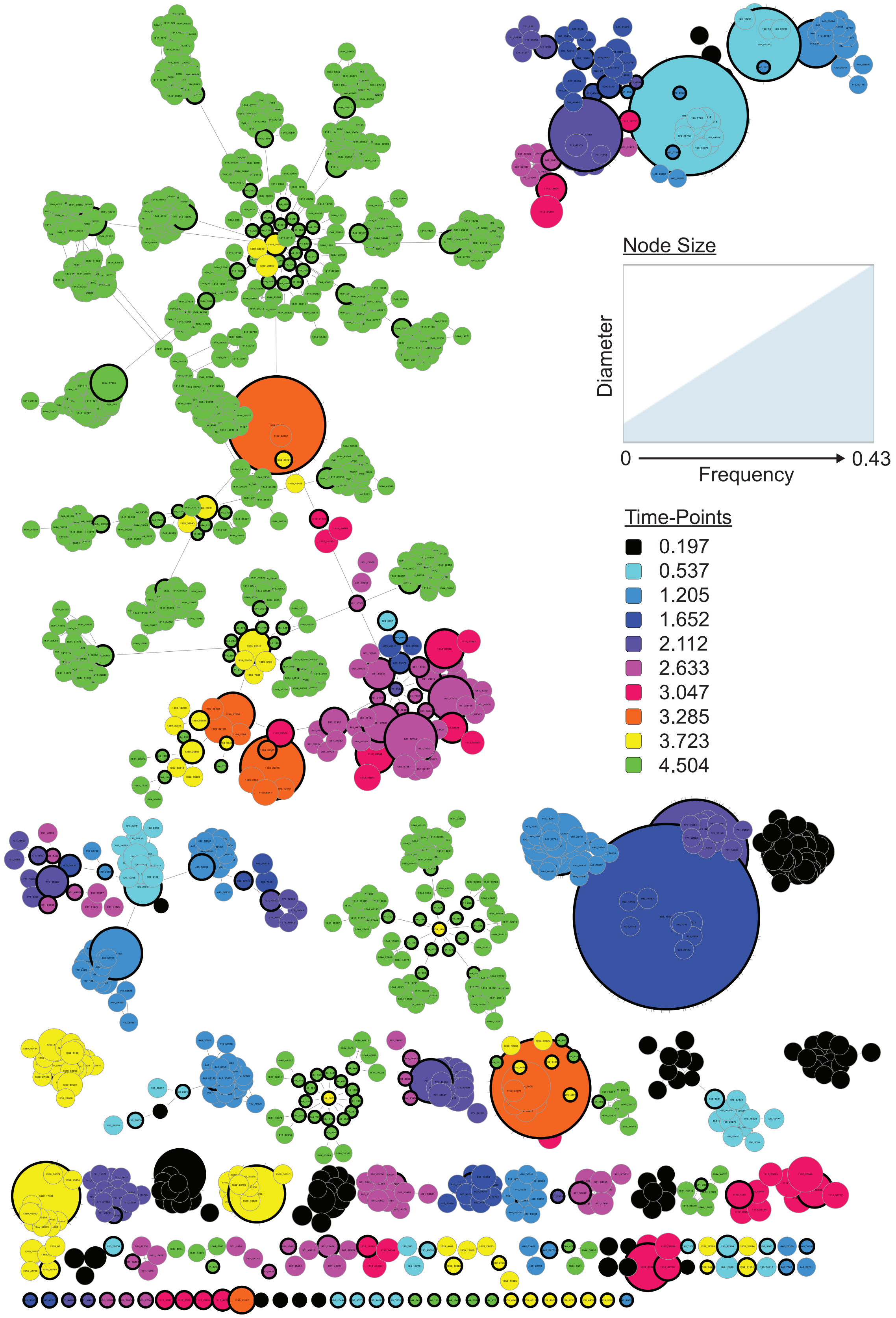


Figure 4.21. Network plots of AbR lineages.
Depicts the structure of each Ab lineage within a given HIV-associated AbR partition. Each node represents a unique sequence, and the size of a given node reflects the frequency of that sequence (within the AbR partition), at a given time-point. The fill color of a node reflects time-point, and if a node has a thicker black border then it was assigned to be a representative sequence of a sequence-cluster (borders are otherwise thinner and gray). Nodes of the same color (i.e. sequences in the same time-point) are linked with an edge if they were in the same cluster. Nodes that have different colors are linked with an edge if they were assigned to the same lineage. This inter-time-point linking only occurs between representative sequences. Taken together, each isolated grouping of nodes shows a family of related lineages. (A) AbR partition IGHV3-30:IGHJ3 in patient 3. (B) AbR partition IGHV4-31:IGHJ5 in patient 7. (C) AbR partition IGHV6-1:IGHJ5 in patient 8.

If coevolution is a common characteristic of HIV-targeting Ab lineages, then we might expect that the HIV-associated AbR partitions that we previously identified will have abnormally high MI values. Thus, we compared the mean MI values of the lineages within the all of the HIV-associated AbR partitions (IGHV3-30:IGHJ3 of patient 3, IGHV2-70:IGHJ6 of patient 7, IGHV3-15:IGHJ4 of patient 7, IGHV4-31:IGHJ5 of patient 7, IGHV6-1:IGHJ4 of patient 8, IGHV6-1:IGHJ5 of patient 8) to the distribution of mean MI values from the rest of the lineages (Figure 4.22 A). In addition to mean MI, we also compared the length of lineages (i.e. the number of time-points for which a lineage is present in the data), because coevolving lineages might be expected to persist in the AbR longer than non-coevolving lineages. We found no evidence of coevolution in the HIV-associated AbR partitions, with the exception of a single lineage in IGHV6-1:IGHJ5 of patient 8, which had a mean MI value that was in the 99.55th percentile relative to all the Ab lineages in non-HIV-associated AbR partitions. We found that this result persisted when comparing to a simulated null distribution that controls for uncertainty when assigning lineages across time-points (Figure 4.22 B).

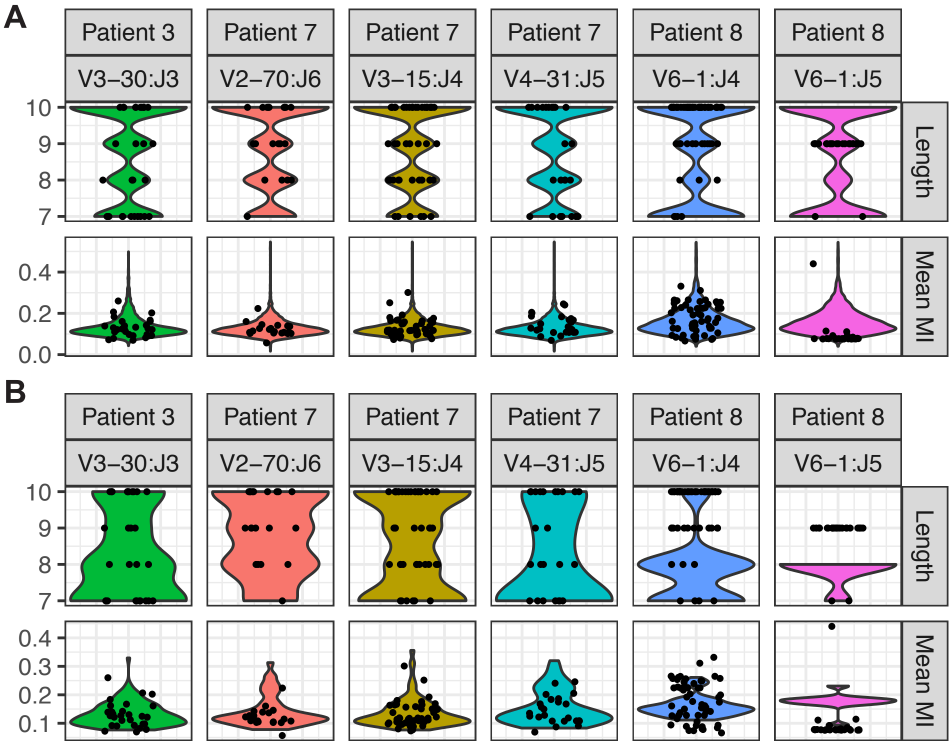


Figure 4.22. Coevolution test for individual lineages.
Each black dot corresponds to a value for a specific lineage within an HIV-associated AbR partition, and are jittered across the x-axis. The identifying information for the specific HIV-associated partition is given by the column labels for the plots (ex. The values for the lineages within the IGHV3-30:IGHJ3 AbR partition of patient 3 are found in the left-most column). The colored violin plots behind the points represent a given null-comparison for the points, where color corresponds to patient. The top row of plots shows the values and null distributions for the length of lineages (i.e. how many time-points they were present). The bottom row shows the values and null distributions for the mean MI of lineages. (A) The null distributions are made up of the values from the lineages in AbR partitions that were not HIV-associated. (B) The null distributions are made up of the values from the null simulation for each given AbR partition.

Lastly, we test for a global coevolutionary signal, agnostic to whether or not a lineage belongs to an HIV-associated partition. To do this we gather the mean MI values across all of the observed lineages within a patient, and then compare this distribution to that of the mean MI values from the simulated null lineages (Figure 4.23). If Ab/HIV coevolution were taking place on a large scale, we would expect to see a shift towards higher MI values in the observed distribution relative to the simulated null. However, we see no evidence of this, and instead see that, if anything, the simulated null lineages tend to have higher MI values. This suggests that, if coevolution is taking place at all, it is either a weak force or exceedingly rare in these patients. Alternatively, we cannot rule out the possibility that our test was underpowered and as a result was unable to detect a coevolutionary signal in our genetic data.

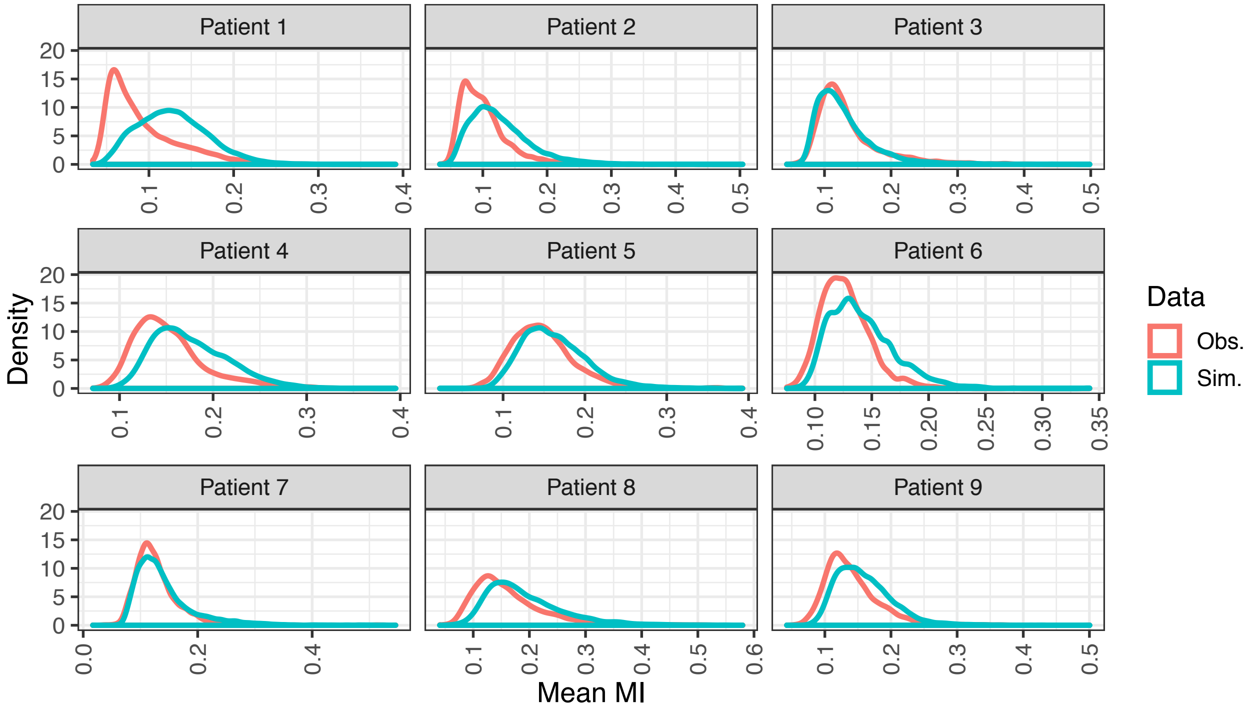


Figure 4.23. Gloabal tests for coevolution.
Each line gives the distribution of mean MI values for a given patient. Salmon colored lines correspond to observed lineages (i.e. lineages inferred from the data), and turquoise colored lines correspond to lineages from the null simulations. Patient 10 is omitted because of limited HIV sequence data.

Discussion

In this study we have created a relatively large dataset of Ab sequences and HIV sequences from 119 longitudinal samples. While more samples would clearly be preferable, this is, to our knowledge, the largest dataset of its kind. The HIV literature currently encompasses an abundance of AbR sequence datasets from HIV+ individuals, however, these datasets primarily originate from the amplification of a particular Ab lineage that was known to contain an HIV-bnAb, prior to deep sequencing. These ‘biased’ AbR datasets are quite useful to home in on the development of a particularly interesting Ab lineage. However, we argue that it is equally important to understand the humoral immune response to HIV infection on a global/systems scale for the following reasons: i) It is just as important to understand why an HIV-targeting Ab lineage failed to develop broad neutralization ability, as it is to understand why a lineage succeeded in developing it. ii) It is quite possible that a significant proportion of humans that develop broad immunity to future HIV infections, do so in a polyclonal manner [23]. Meaning that broad neutralization depth against HIV is achieved via the cooperative action of many Ab lineages, each simultaneously targeting different epitopes, or different versions of the same epitope. iii) The population dynamics of HIV-bnAbs (in addition to HIV-binding Abs in general) are poorly understood. For example, do these Abs persist at high or low frequency in the greater population? How does this frequency change over time? What type of selection drives their development (ex: positive, negative, balancing, etc.)? These types of questions are difficult, if not impossible to answer without understanding the larger population context for which these Ab lineages exist. In this study, we have taken the preliminary steps towards addressing these types of questions. Specifically, we have developed a statistical approach to identify the partitions of the AbR that are likely responding to HIV. Once this has been established, questions like those enumerated above, can be answered. Further, we hope that the sequence datasets we have created here will provide a useful resource for others with similar lines of inquiry.

While [13] found no correlation between AbR diversity and viral load in their data, they did not have the means to address other characteristics of the HIV population, as they did not have HIV sequence data at their disposal. However, they did find that AbR diversity was lower in HIV+ individuals than healthy controls. This suggested that HIV may have a broad effect on the AbR, yet the details of this effect remained unclear. We have presented a small positive correlation between overall AbR and HIV diversity across all our samples. A possible scenario that would explain this observation is one where AbR diversity is decreased due to clonal expansions in Ab lineages that target HIV, which in turn causes a decrease in HIV diversity, due to positive selection for escape mutations. Once the HIV population has escaped, its diversity will return, and diversity in the AbR will also return because the previous clonal expansion will have vanished due to its target having escaped. However, we stress that this correlation had nominal significance and should be treated cautiously. The AbR is an especially complicated population that is capable of simultaneously responding to countless antigens, thus even a fleeting correlation with HIV at the whole-population level may be worthy of further follow up studies.

A key first step towards illuminating the global interaction between the AbR and an HIV infection is to be able to identify the subset of the AbR that is actually responding to HIV. Similar to our previous work in the context of influenza vaccination [40], we leveraged the time-series nature of our dataset to identify partitions of the AbR that seem to be associated with HIV. We purposefully made no prior assumptions about what types of interactions we might find. For example, the common narrative of HIV-targeting Ab lineages is that they are under intense positive selection. Thus, one might have the expectation that Ab selection will be positively correlated with HIV selection. However, it is also possible that an HIV-targeting Ab lineage is under intense negative selection, where there is a preference for amino acids *not* to change so as to not ablate their binding ability, or perhaps to not change a strict structural conformation that is required to access an epitope. In this case, one might expect Ab selection to be negatively correlated with HIV selection. We therefore compared all AbR summary statistics to all HIV summary statistics. This was a double-edged sword, as it gave us the privilege of an unbiased approach, yet greatly increased the number of tests that were performed, and hence the severity of our multiple tests correction. As such, we were only able to identify a handful of AbR partitions that were significantly associated with HIV. This suggests that long-term interactions between Ab lineages and HIV are rare, and that Ab/HIV interactions may be of a more transitory nature, where an antibody binds to HIV, then HIV escapes, and then another unrelated Ab binds to the escape mutant, and so on. Another possibility is that our test was simply underpowered and had many false negatives. One way to ameliorate this would be to first filter AbR partitions based on some statistic (e.g. divergence) and then test for associations using a different, orthogonal statistic (e.g. diversity). Further, there remains a great deal of powerful analyses that could be done with the HIV sequence data. In principle, one could divide the HIV population into lineages and test each HIV lineage against each AbR partition. This would increase the number of tests, but could also illuminate interactions that would be otherwise hidden.

Lastly, we tested for a coevolutionary signal in our data. Coevolution in sequence data is notoriously hard to establish [46], and to our knowledge, reports of HIV/Ab coevolution to date have been universally qualitative, with little or no statistical analyses [11,12,15,18,20,32,47–49]. When it comes to claims of coevolution, there are two sources of uncertainty that we have attempted to account for in this study. i) When both the Ab lineage and the putative HIV epitope are under positive selection, it is very easy for mutations to be correlated by chance rather than by coevolution. ii) There is a huge amount of uncertainty when assigning Abs to lineages, especially when trying to link a given Ab sequence to other Ab sequences that existed months-years prior. AbRs are incredibly dynamic populations with high turnover, and high mutation rates. In a population such as this, where *de novo* lineages are continuously being added, it is important for one to account for the possibility that two similar Ab sequences—even if strikingly similar—may not be of the same lineage. By creating a simulated null dataset from shuffled Ab lineages, we were able to create a null distribution of MI values that took both of these confounders into account. After doing this we found no global signal for coevolution, yet we did find one isolated Ab lineage in patient 8 that showed compelling evidence for it. This suggests that while coevolution between Ab lineages and HIV is possible, it is likely exceedingly rare and/or hard to detect. Given that other sequence datasets of the AbR in the context of HIV infection have about the same or fewer time-points than the patients in our dataset, we suspect that claims of coevolution in these data would be equally hard to make.

Coevolution has been responsible for some of the most remarkable phenotypes known (e.g. the cheetah's speed, a flower's beauty, the strangeness of genitalia [50]), yet it remains unclear as to how much of a role it plays in the development of HIV-targeting Abs. This may seem academic, but it has important implications for vaccine strategies. If coevolution is the predominant force in the development of HIV-bnAbs, then a vaccine regimen that mimics the evolution of the HIV epitope would be desired, as this would recapitulate the coevolutionary process. However, it is also possible that HIV-bnAbs occur as rare, random events, whereby a (typically diverged) Ab lineage 'stumbles' upon broad neutralization breadth by chance. In this case, one might desire a vaccine regimen that has a very diverse array of HIV epitopes so as to maximize the chances that this rare event occurs.

It remains unknown how much of a role each of these scenarios play in the development of HIV neutralization breadth, however, it has been shown that neutralization breadth is positively correlated with viral load and HIV diversity [51–53]. This is suggestive (although far from conclusive) that coevolution may play less of a role, as high viral diversity is not a necessity for an arms race. An interesting future study would be to use simulations or mathematical modeling to gain a better understanding of which evolutionary parameters (e.g. population size, mutation rate, selection strength, population diversity, etc.) in the HIV and AbR populations promote coevolution, and which do not. This was partially done by Nourmohammad et al. [54], but coevolution was more of a feature of their model rather than a variable being tested. For example, it could be that a highly diverse AbR, with a slower mutation rate than HIV would be less likely to coevolve, and more likely to generate neutralizing Abs via chance recombination events. Whereas a less diverse AbR, with a mutation rate on par with HIV, may be more likely to engage in arms races. However, speculation is difficult when concerning two complex populations that engage in a complicated interaction, and simulation frameworks, such as those used by Murugan et al. [24], that include the introduction of novel naïve Ab lineages into the population could be used to gain insight into this line of inquiry. We hope that this study has provided a sound example of how to go about formally testing for coevolution in order to differentiate between these two possibilities.

Methods

Patient selection and sample processing

Samples were obtained from the OPTIONS cohort at UCSF. All patients provided written informed consent, and the study was approved by the UCSF Committee on Human Research. Our sole criterion for selecting patients from this cohort was to find those with the greatest number of samples available prior to the administration of ART. All the patients in this study were men who contracted HIV via sexual transmission, with the exception of patient 5, who became infected by unknown means (Table 1). Each peripheral blood sample was divided into plasma and peripheral blood mononuclear cells (PBMCs) by density gradient centrifugation using Ficoll-hypaque. After separation, PBMCs and plasma were aliquoted in cryopreservation media, and cryo-preserved in a specimen repository. Plasma viral load was measured at each patient visit.

Viral load

In early samples (prior to ~2009), a combination of a branched DNA assay and an ultra-sensitive PCR assay from Roche were used to measure HIV load. In later samples (post ~2009), the Abbott RealTime HIV-1 Viral Load assay was employed to measure load.

Estimated time of infection

The time of initial infection was estimated using the following criteria: i) If a patient first presents with detectable viral load, but negative enzyme immunoassay (EIA) or western blot, and then presents a positive western blot in the following visit, then the estimated time of infection is given by 24 days prior to the first visit. ii) If a patient first presents with an indeterminate western blot, and then a positive western blot following repeat testing, then the estimated time of infection is given by 24 days prior to this first test. iii) If patient first presents with a positive western blot, and has a documented negative HIV test result within at most 180 days prior to first test, then the estimated time of infection is calculated as 24 days prior to the midpoint between the first positive test and prior negative test. We note that first visit here corresponds to the first visit in the OPTIONS study at UCSF, and not the first sample in our study.

C2V3 amplification and ultra-deep sequencing

HIV RNA was isolated from plasma samples using the Maxwell 16 Viral Total Nucleic Acid Purification Kit (Promega). cDNA was synthesized using the SuperScript III First-Strand Synthesis System (Invitrogen) with a gag-specific primer: 5'-GCACTYAAGGCAAGCTTTATTGAGGCTTA-3'. The C2/V3 region (~416bp) of HIV *env* was amplified using a nested PCR approach with Phusion High-Fidelity PCR Master Mix (New England Biolabs). The outer primers were: 5'-ATTACAGTAGAAAAATTCCCCT-3' and 5'-CAAAGGTATCCTTTGAGCCAAT-3'. The inner primers were: 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAACAGGACCAGGATCCAATGTCAGCACAGTACAAT-3' and 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCGTTAAAGCTTCTGGGTCCCCTCCTGAG-3', where the underlined portions indicate the Illumina adapter sequence. A unique barcode was added to each amplicon using the Nextera XT Index Kit (Illumina) and the barcoded amplicons were mixed to generate a sequencing library. Paired-end sequencing (2×300 bp) was performed using the Illumina MiSeq instrument and the MiSeq Reagent Kit v3.

IGH amplification and ultra-deep sequencing

Total RNA was extracted from PBMCs using the Qiagen RNeasy Mini Kit. To reverse transcribe, amplify IGH encoding RNA, and generate sequencing-ready libraries, we used iRepertoire's long read iR-Profile Kit and followed the procedure as described in the accompanying protocol [55]. Paired-end sequencing (2×300 bp) was performed using the Illumina MiSeq instrument and the MiSeq Reagent Kit v3.

HIV sequence data QC

Sequences from different samples were de-multiplexed by barcode using the internal software on the Illumina machine. We used the software package pRESTO [56] to assemble read pairs, remove sequences shorter than 300bp, remove sequences with a mean quality score less than 30, mask the primer sequences, and remove sequences that only occur once in a given sample. We then use an in house implementation of BLAST [57] to check that each sequence has at least 70% identity to at least one HIV subtype *env* reference sequence, which were downloaded from LANL [58]. In order to check for possible contaminations from HIV sequences outside of our study, we again used BLAST to map each of our sequences to every *env* sequence within the LANL database. In this case, any of our sequences that had 99% identity or more to any sequence within the *env* database would be deemed a contaminant. We found that all the samples from patient 8 in our study had a significant amount of identity with sequences derived from patient ID: 9036 in the LANL database. We also found that all the samples from our patient 9 had significant identity with sequences in the LANL database derived from patient ID: 9018. There are two reasonable explanations for this: i) these samples had a large degree of contamination, or ii) that our patients 8 and 9 are the same as patients 9036 and 9018 in the LANL database, respectively. We conclude that the latter is the more likely explanation because of the following rationale. This large degree of ‘contamination’ only occurred in patients 8 and 9, and it occurred in all their samples, however, the samples from these patients were processed in different batches. These patients’ diversity, and divergence trajectories showed a relatively steady increase over time, which would not be consistent with contamination (see results). Lastly, patients 9036 and 9018 from the LANL database both correspond to the study [59] which also recruited patients from San Francisco, CA. None of the other samples in our study had detectable contamination using this method.

To check for cross contamination of sequences across samples in our study, we used a clustering approach. We first reduced the size of the dataset by grouping the sequences within each sample that have an edit distance less than or equal to 4 (see “Clustering sequences with samples” section below). We then choose the sequence within each group (or cluster) that has the highest count to be the ‘representative sequence’ for that cluster. After which we pool all the representative sequences across all samples and cluster these pooled representative sequences using the same clustering algorithm. To identify clusters of sequences that were likely cross contaminants we used the following criteria: the cluster had to i) have a representative sequence that clustered closer with sequences from a different patient than with sequences from the same sample, and ii) have a frequency less than 0.001 within its sample. All sequences within clusters that satisfied these criteria were removed. This effectively identified low frequency sequences that were closer in genetic distance to sequences from another patient. Lastly, we used a phylogenetic approach (see “Making phylogenetic trees” section below) to i) make phylogenetic trees of the representative sequences in each sample, and ii) remove any sequence that is more closely related to a representative sequence that was identified as a cross contaminant (as described above) then to the other representative sequences in the sample (Figure 4.24).

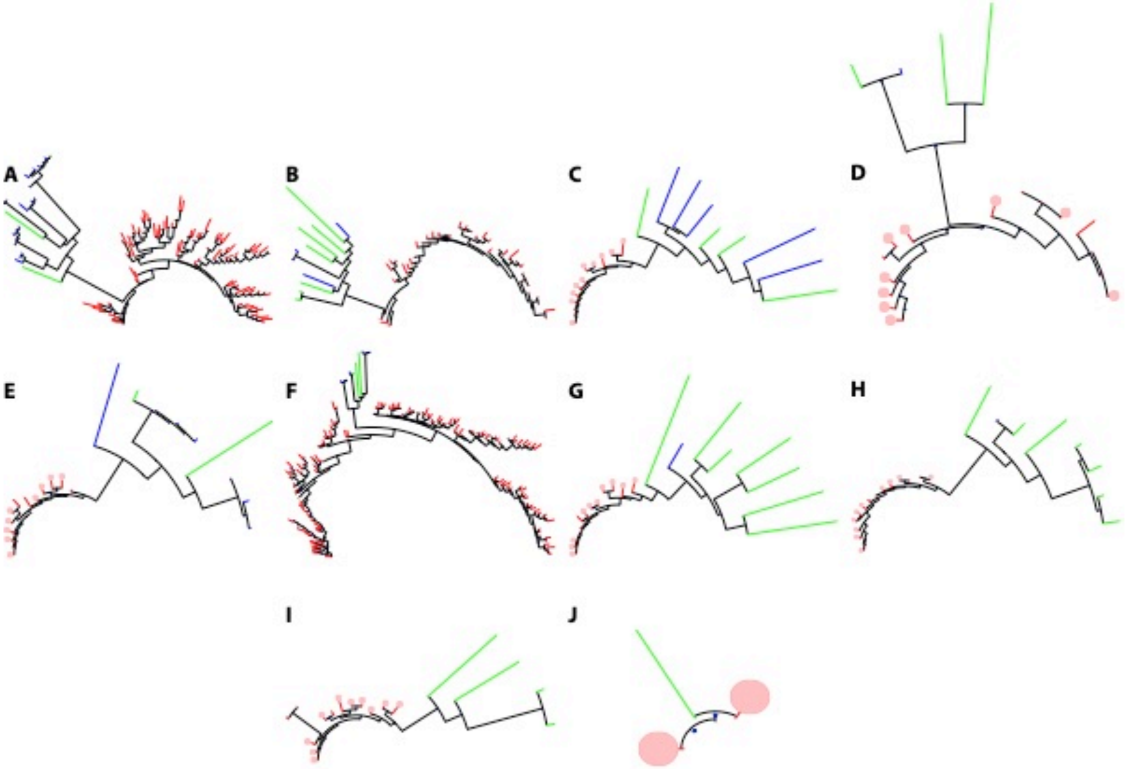


Figure 4.24. Outlier HIV sequences.
Phylogenetic trees depicting the relationship of HIV sequences for each patient. Each leaf on a tree represents a cluster of sequences, and the size of the circle at the terminus of a leaf gives the relative frequency of that cluster at a given time-point. Leaves that are colored blue were identified as outliers because they were found to have a closer genetic distance to clusters from another patient then clusters within their assigned patient (i.e. cross-contaminants). Leaves that are green were identified as outliers because they group closest with leaves identified a cross-contaminants.

Clustering sequences within samples

We use the Needleman-Wunsch algorithm [60] as implemented in the ‘needle’ program from the European Molecular Biology Open Software Suite [61] to globally align each pair of sequences, and calculate the edit distances. Through an in-house algorithm, we then group sequences into a cluster that have an edit distance less than some provided threshold to any other sequence in the cluster. For HIV sequences, this threshold was 4; for AbR sequences, this threshold was 6.

Making phylogenetic trees

To make a phylogenetic tree of a group of sequences, we first make multiple sequence alignments using MAFFT [62], and then construct phylogenetic trees using FastTree [63]. Visualization and analyses of newick formatted files was performed using the ETE toolkit [64].

Calculating HIV divergence

We first assign an HIV reference sequence for each patient by finding the most abundant sequence at the first time-point. Because patient 6 showed extensive evidence of a super-infection occurring at the second time-point, we assigned two reference sequences to patient 6: one from the first time-point and the other from the second. We then translate the reference sequences as well as all other sequences in the data (see “Translating HIV sequences” section below). To find the number of synonymous and non-synonymous changes for a given query HIV sequence, we first codon align it to the patient’s reference. For a given codon, we first calculate the number of expected non-synonymous and synonymous sites as:

$$n = \sum_{i=1}^{3} f_i \text{ , and}$$

$$s = 3 - n \text{ .}$$

Where f_i gives the proportion of all possible nucleotide changes at codon position i of the reference sequence that result in an amino acid change. We denote N and S as the sum of n and s across all codons in a given reference sequence. We then count the number of observed non-synonymous and synonymous changes in a given query HIV sequence as N_o and S_o. If there were multiple mutations, we selected the order of mutations that resulted in the least amount of amino acid changes as the most parsimonious, and thus most likely. The proportion of non-synonymous and synonymous mutations in a given query sequence is then,

$$pN = \frac{N_o}{N} \text{ , and } pS = \frac{S_o}{S} \text{ .}$$

To estimate non-synonymous and synonymous divergence we then use

$$dN = -\frac{3}{4}\ln\left(1 - \frac{4}{3}pN\right) \text{ , and } dS = -\frac{3}{4}\ln\left(1 - \frac{4}{3}pS\right)$$

[65]. The notation above was heavily borrowed from Richard Orton's blog post [66]. Because patient 6 had two reference sequences, the reference sequence that yielded the lower divergence value was used, for a given query sequence.

Calculating HIV selection

We estimated selection in HIV as simply dN/dS.

Translating HIV sequences

In order to translate a given query HIV sequence, we first use needle to globally align it to the reference HXB2 *env* sequence (downloaded from LANL). We then use this alignment to determine the coding frame of the query sequence. Once this is known we translate the query nucleotide sequence using a simple in-house python script.

Calculating diversity

To estimate diversity in both the HIV population and the AbR we calculated the statistic, π [40]. In words, π is the expected genetic distance between two randomly selected sequences from a given sample. Mathematically π can be expressed as

$$\pi = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} c_i \cdot c_j \cdot G(s_i, s_j)}{\binom{M}{2}} \text{ ,}$$

where N gives the total number of unique sequences in the sample, c_i gives the count of sequence i, M gives the total counts of sequences in that sample (i.e. $M = \sum_{i=1}^{N} c_i$), and $G(x, y)$ gives the genetic distance between sequences x and y. We used VSEARCH with the "--allpairs_global" option to globally align all pairs of sequences in a sample [67]. Genetic distance between a pair of sequences was then calculated as the percent of mismatches in the alignment.

AbR sequence data QC

As with the HIV data, we used pRESTO to assemble read pairs, remove reads less than 300bp, remove reads with a mean Q score less than 20, and remove reads that only occur once. We then use IgBLAST [68] to align each sequence to a database of germline immunoglobin genes downloaded from the IMGT website (imgt.org) [69]. After this, we used Change-O to annotate each sequence with its most likely V, D, and J germline gene-segments, identify the FWRs and CDRs, and to construct the likely naïve antibody sequence from the germline gene-segment alignments.

Calculating Ab divergence

Divergence in a given Ab sequence was calculated as the number of changes in the observed sequence relative to the naïve sequence, divided by the length of the sequence. This is excluding the junction region of the sequence, as naïve sequence reconstruction of this region is unreliable.

Estimating Ab Selection

We used the BASELINe program on each individual Ab sequence to estimate selection. For a detailed description of this tool see [70]. Very briefly, BASELINe compares the observed number of mutations in a sequence (relative to its inferred naïve ancestor) to a null distribution of the expected number of mutations under no selection. The program takes local nucleotide motifs into account when calculating mutation probabilities, and returns a sigma value that indicates the distance between the observed number of mutations and the null distribution. A negative sigma indicates fewer mutations than expected (negative selection), and a positive sigma indicates more mutations than expected (positive selection). It does this separately for different regions of the Ab sequence (i.e. the FWR and CDR).

Creating AbR lineages

To cluster the AbR of a given patient, we first divided it into partitions by grouping together all sequences that use the same germline V and J gene-segments (as annotated by Change-O). We then cluster the sequences within a given partition/time-point (see “Clustering sequences within samples” section above), with and edit distance threshold of 6 (Figure 4.25 A). Once clusters are delineated, and similar to the clustering of HIV sequences, we assign the most numerous sequence of each cluster, as the ‘representative sequence’. We then link clusters, within a given partition, across adjacent time-points using the following algorithm. We first find the representative sequence in the previous time-point that has the smallest edit distance to a given representative sequence in a contemporary time-point. If this edit distance is smaller than 30 (Figure 4.25 B), then the two representative sequences (and the clusters they represent) will be linked as being part of the same lineage. This process is carried out independently in each patient, over each representative sequence, and for each adjacent time-point pair (see Figure 4.26). Finally, once all lineages have been assigned in this manner, any lineage that does not rise above 0.0001 frequency, in any of the time-points will be disregarded.

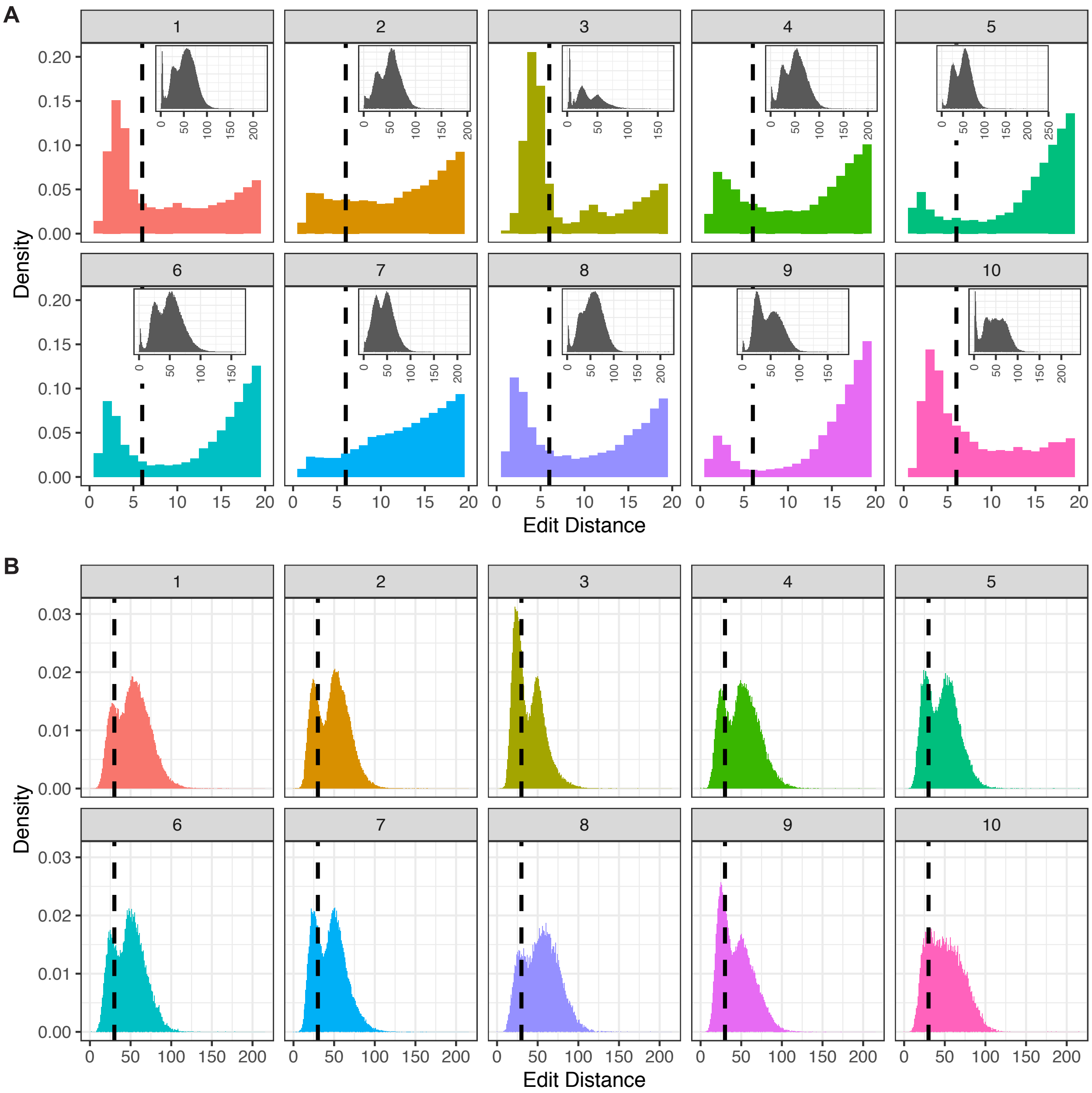


Figure 4.25. Genetic distance distributions for AbR sequences.
Histograms of edit distance values between pairs of antibody sequences in the AbR. Top label foreach panel and color correspond to patient. (A) Histograms of edit distance values for all sequences within a time-point. The within time-point distance values are pooled across all time-points to generate the shown histograms. The main plots give a zoomed-in view for edit distance values 0-20. Insets show the entire histogram. Vertical dashed lines at x=6 show the edit distance cutoff that was used to assign sequences to the same cluster. (B) Histograms of edit distance values between adjacent time-points. This shows the edit distance distributions between pairs of representative cluster sequences across adjacent time-points. Vertical dashed lines at x=30 show the cutoff that was used to assign sequence clusters as being members of the same lineage.

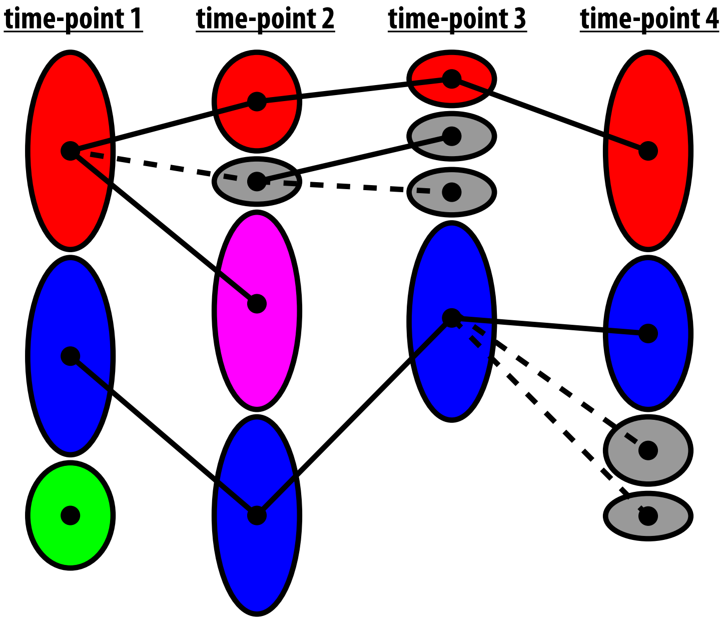


Figure 4.26. Schematic of lineage clustering algorithm.
Each ellipse represents a cluster of sequences, and the dot in the middle is the representative sequence for a given cluster. The lines, or edges, linking clusters across time-points are formed by finding the cluster in the previous time-point that has the representative sequence with the smallest genetic distance to a given query representative sequence. Dashed edges represent lineage connections that fall above the edit distance threshold, while solid edges fall below. Clusters are colored by which lineage they belong to. However, the red cluster in the top left corner is simultaneously a member of the 'red' and 'purple' lineages because there is a bifurcation at time-point 2. In these cases of multifurcations, the parental color is assigned to the daughter cluster that has the lowest edit distance to the parent cluster. However, this is only pertinent for visualizations, such as this figure and Muller plots. Grey denotes lineages that do not meet the minimum frequency threshold.

Simulating null AbR lineages

To simulate random AbR lineages, we began by using our clustered Ab sequences within the partitioned AbR data (see "Creating AbR lineages" section above). We then carry out an identical procedure as was done to create the observed lineages, with the exception that instead of finding the parent cluster in the previous time-point that has the minimum edit distance, a parent cluster is randomly chosen from the previous time-point. Additionally, in order to replicate the aspect of the observed data where each time-point brings a certain number of new lineages to the population, we estimate the probability of a new lineage as

$$r_{i,j} = \frac{l_{i,j}}{C_{i,j}}$$

where $l_{i,j}$ gives the number of new lineages, and $C_{i,j}$ gives the total number of clusters in patient, i, and time-point, j. For the simulated lineages, we then randomly assign each cluster as being a new lineage (i.e. not having any connections with the previous time-point) with probability $r_{i,j}$. In order to have a null dataset that is sufficiently large, we duplicated the observed, clustered data 100 times and then simulated lineages using this 100-fold larger dataset.

Linear modeling of population level interactions

We tested for cross-patient, population-wide interactions of summary statistics using a linear mixed model approach. We model the interaction of a pair of summary statistics as

$$Y_{i,j} = \beta_i + \beta X_{i,j} + \epsilon_{i,j} \ ,$$

where $Y_{i,j}$ gives the value of a given HIV summary statistic for the jth time-point of patient i, $X_{i,j}$ gives the value of a given AbR summary statistic for the same patient/time-point, β_i is a random intercept term to correct for patient specific effects, and $\epsilon_{i,j}$ is a random error term that is assumed to be normally distributed with a mean of 0. This model was implemented in the R programming language using the 'lmer' function of the 'lme4' package [71]. We then use a likelihood ratio test to determine if a model with β≠0 provides a significantly better fit than a model with β=0 ($p \leq 0.05$). If it does, then the given pair of HIV and AbR summary statistics was deemed to be interacting.

Trajectory permutation test

We test for associations between a set of AbR trajectories to one or more HIV trajectories using a permutation-based test. The AbR and HIV trajectories are loaded into memory as matrices where each row is a different trajectory and each column is a time-point in chronological order from left to right. We first standardize each trajectory by subtracting the mean and dividing by the standard deviation. If v is a row vector representing a given trajectory, then v is standardized by

$$v_j' = \frac{v_j - \mu_v}{\sigma_v} ,$$

where μ_v and σ_v give the arithmetic mean and standard deviation of v, respectively. We then calculate the sum of the squared error (SSE) for a given AbR trajectory relative to a given HIV trajectory as

$$SSE(r', v') = \sum_{j \in v'} \left(r_j' - v_j'\right)^2 ,$$

where r' is a standardized HIV trajectory vector. This gives the observed SSE values for each AbR/HIV trajectory pair. If either the HIV or AbR trajectories have missing values, then these time-points are disregarded in the SSE calculation, and a trajectory must have at least 75% of its values defined to be included in the test. We then permute the columns of the AbR trajectory matrix many times and calculate the SSE values for each permuted AbR trajectory after each permutation. This gives our permuted null distribution of SSE values. If an observed AbR/HIV trajectory pair have an SSE value that is significantly outside of this null distribution, then they are deemed to be significantly associated with one another, where significance is appropriately adjusted as based on the number of tests. When conducting this test on whole AbR population trajectories vs. whole HIV population trajectories (Figure 4.14 B-E), we performed 100,000 permutations for each patient. When conducting this test on AbR partition trajectories vs. whole HIV population trajectories (Figures 4.16, 4.17, and 4.18) we performed 1,000 permutations for each patient.

Comparing to literature datasets

To compare the results of our trajectory permutation test to a literature dataset we use a Mann-Whitney U test. However, before this can be done, we first must combine the results of our permutation-based test across patients. When the permutation-based test was employed to identify HIV associated AbR partitions, the structure of the results was as follows: each patient had hundreds of AbR partitions, and each AbR partition had tens of p values associated with it (from comparing each of its summary statistics to each of the HIV population summary statistics). In order to combine p values such that there is one value associated with one V gene segment, we first pool the p values across all AbR partitions that have a given V gene segment, and across all patients. We then use Fisher's method to arrive at an overall V gene score for this pool of p values. If p is a vector of p values associated with a given V gene segment, then we first combine the p values into one overall p value, $p_{overall}$, using Fisher's method:

$$c = -2\sum_{i=1}^{|p|} \ln p_i \text{ , and}$$

$$p_{overall} = \Pr\left(\chi^2_{2|p|} \leq c\right) \text{ .}$$

Where $\chi^2_{2|p|}$ is the chi-squared distribution with $2|p|$ degrees of freedom. Strictly speaking, $p_{overall}$ will tend to be inflated because not all the p values in p are independent (ex. AbR partition trajectories that are associated with HIV selection, will also tend to be associated HIV non-synonymous divergence). However, because this inflation should be the same across V gene segments, and because we are not interested in actual significance but rather need some reasonable method for combining p values into an overall score, Fisher's method should be sufficient. Finally, to avoid $p_{overall}$ being interpreted as a significance level, we take its log transform to arrive at a V gene score

$$V\ gene\ score = -\log_{10} p_{overall}\ .$$

We then used a Mann-Whitney U test to see if V gene segments that were 'well represented' in a literature dataset tended to have significantly different $V\ gene\ score$ values then those that were not 'well represented'. In the case of the dataset of HIV targeting Abs, 'well representation' was defined as presence/absence (i.e. count ≥ 1). The dataset of influenza targeting Abs was relatively large (432 entries), so 'well representation' was defined as a count ≥ 10 (Table S1).

Calculating MI

To measure the amount of association between two sites (columns) in a pair of MSAs we first reduce the complexity of the amino acid code by converting it to a code of 'change' or 'no-change'. In this case, if a site has an amino acid identity that is different than the previous time-point, then it is recorded as a '1' and if it is the same, then it is recorded as a '0' (the first time-point is always '0'). We then use MI to measure the amount of association (coevolution) between two columns in a 'change', 'no-change' alignment. MI is calculated as

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} \Pr(x,y) \log_2 \left(\frac{\Pr(x,y)}{\Pr(x) \cdot \Pr(y)} \right),$$

where X and Y are categorical random variables representing the two different columns being compared. x and y represent particular states of X and Y, respectively (i.e. ‘1’ or ‘0’ for a change/no-change alignment). $\mathrm{Pr}(a)$ is the probability that a given random variable (or MSA column) equals the state, a. If c is a vector that represents a given column of an MSA, then $\mathrm{Pr}(a)$ can be estimated as

$$\mathrm{Pr}(a) = \frac{1}{|c|}\sum_{j \in c} \begin{cases} 1, \; c_j = a \\ 0, \; c_j \neq a \end{cases}.$$

$\mathrm{Pr}(a, b)$ is the joint probability that the random variable representing one MSA column equals a, and simultaneously the random variable representing the other MSA column equals b. If d is a vector that represents a given column of the other MSA, then $\mathrm{Pr}(a, b)$ can be estimated as

$$\mathrm{Pr}(a, b) = \frac{1}{|c|}\sum_{j \in c} \begin{cases} 1, \; c_j = a, d_j = b \\ 0, \; c_j \neq a \mid d_j \neq b \end{cases}.$$

Correcting for multiple tests

Unless otherwise stated, p values from a given statistical test within a patient were corrected for multiple testing using the Benjamini-Hochberg procedure.

Network plots

Networks of lineages were visualized using cytoscape [72].

References

1. Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, et al. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. Science [Internet]. NIH Public Access; 2014 [cited 2018 Nov 28];346:56–61. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25278604

2. Gottlieb M, Schanker H, Fan P, Saxon A, Weisman J, Pozalski I. Pneumocystis pneumonia--Los Angeles. MMWR Morb Mortal Wkly Rep [Internet]. 1981 [cited 2018 Nov 28];30:250–2. Available from: https://www.popline.org/node/421156

3. Stein J, Storcksdieck genannt Bonsmann M, Streeck H. Barriers to HIV Cure. HLA [Internet]. 2016 [cited 2018 Nov 28];88:155–63. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27620852

4. Trovato M, D'Apice L, Prisco A, De Berardinis P. HIV Vaccination: A Roadmap among Advancements and Concerns. Int J Mol Sci [Internet]. 2018 [cited 2018 Nov 28];19:1241. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29671786

5. Weinstein M, Yang OO, Cohen AC. Were we prepared for PrEP? Five years of implementation. AIDS [Internet]. 2017 [cited 2019 Mar 4];31:2303–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28857829

6. Binley JM, Lybarger EA, Crooks ET, Seaman MS, Gray E, Davis KL, et al. Profiling the Specificity of Neutralizing Antibodies in a Large Panel of Plasmas from Patients Chronically Infected with Human Immunodeficiency Virus Type 1 Subtypes B and C. J Virol [Internet]. 2008 [cited 2018 Nov 28];82:11651–68. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18815292

7. Hraber P, Seaman MS, Bailer RT, Mascola JR, Montefiori DC, Korber BT.

Prevalence of broadly neutralizing antibody responses during chronic HIV-1 infection. AIDS [Internet]. 2014 [cited 2018 Nov 28];28:163–9. Available from: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00002030-201401140-00002

8. Sather DN, Armann J, Ching LK, Mavrantoni A, Sellhorn G, Caldwell Z, et al. Factors Associated with the Development of Cross-Reactive Neutralizing Antibodies during Human Immunodeficiency Virus Type 1 Infection. J Virol [Internet]. 2009 [cited 2018 Nov 28];83:757–69. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18987148

9. McCoy LE, Burton DR. Identification and specificity of broadly neutralizing antibodies against HIV. Immunol Rev [Internet]. Wiley/Blackwell (10.1111); 2017 [cited 2018 Nov 28];275:11–20. Available from: http://doi.wiley.com/10.1111/imr.12484

10. Wu X, Zhou T, Zhu J, Zhang B, Georgiev I, Wang C, et al. Focused Evolution of HIV-1 Neutralizing Antibodies Revealed by Structures and Deep Sequencing. Science (80-) [Internet]. 2011 [cited 2018 Nov 28];333:1593–602. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21835983

11. Bhiman JN, Anthony C, Doria-Rose NA, Karimanzira O, Schramm CA, Khoza T, et al. Viral variants that initiate and drive maturation of V1V2-directed HIV-1 broadly neutralizing antibodies. Nat Med [Internet]. 2015 [cited 2018 Nov 28];21:1332–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26457756

12. Liao H-X, Lynch R, Zhou T, Gao F, Alam SM, Boyd SD, et al. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. Nature [Internet]. NIH Public Access; 2013 [cited 2018 Nov 28];496:469–76. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/23552890

13. Hoehn KB, Gall A, Bashford-Rogers R, Fidler SJ, Kaye S, Weber JN, et al. Dynamics of immunoglobulin sequence diversity in HIV-1 infected individuals. Philos Trans R Soc Lond B Biol Sci [Internet]. The Royal Society; 2015 [cited 2018 Nov 28];370. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26194755

14. Dawkins R, Krebs JR. Arms races between and within species. Proc R Soc London Ser B, Biol Sci [Internet]. The Royal Society; 1979 [cited 2018 Nov 28];205:489–511. Available from: http://www.ncbi.nlm.nih.gov/pubmed/42057

15. Rantalainen K, Berndsen ZT, Murrell S, Cao L, Omorodion O, Torres JL, et al. Co-evolution of HIV Envelope and Apex-Targeting Neutralizing Antibody Lineage Provides Benchmarks for Vaccine Design. Cell Rep [Internet]. 2018 [cited 2018 Nov 28];23:3249–61. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29898396

16. Zhou JO, Ton T, Morriss JW, Nguyen D, Fera D. Structural Insights from HIV-Antibody Coevolution and Related Immunization Studies. AIDS Res Hum Retroviruses [Internet]. 2018 [cited 2018 Nov 28];34:760–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29984587

17. Gray ES, Madiga MC, Hermanus T, Moore PL, Wibmer CK, Tumba NL, et al. 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. J Virol [Internet]. American Society for Microbiology (ASM); 2011 [cited 2018 Nov 28];85:4828–40. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21389135

18. Doria-Rose NA, Schramm CA, Gorman J, Moore PL, Bhiman JN, DeKosky BJ, et al. Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies.

Nature [Internet]. 2014 [cited 2018 Nov 28];509:55–62. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24590074

19. Goo L, Chohan V, Nduati R, Overbaugh J. Early development of broadly neutralizing antibodies in HIV-1-infected infants. Nat Med [Internet]. NIH Public Access; 2014 [cited 2019 Mar 4];20:655–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24859529

20. MacLeod DT, Choi NM, Briney B, Garces F, Ver LS, Landais E, et al. Early Antibody Lineage Diversification and Independent Limb Maturation Lead to Broad HIV-1 Neutralization Targeting the Env High-Mannose Patch. Immunity [Internet]. 2016 [cited 2019 Mar 4];44:1215–26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27192579

21. Simonich CA, Williams KL, Verkerke HP, Williams JA, Nduati R, Lee KK, et al. HIV-1 Neutralizing Antibodies with Limited Hypermutation from an Infant. Cell [Internet]. NIH Public Access; 2016 [cited 2018 Nov 28];166:77–87. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27345369

22. Sheward DJ, Marais J, Bekker V, Murrell B, Eren K, Bhiman JN, et al. HIV Superinfection Drives De Novo Antibody Responses and Not Neutralization Breadth. Cell Host Microbe [Internet]. Cell Press; 2018 [cited 2018 Nov 28];24:593–599.e3. Available from: https://www.sciencedirect.com/science/article/pii/S1931312818304852

23. Williams KL, Wang B, Arenz D, Williams JA, Dingens AS, Cortez V, et al. Superinfection Drives HIV Neutralizing Antibody Responses from Several B Cell Lineages that Contribute to a Polyclonal Repertoire. Cell Rep [Internet]. Cell Press; 2018 [cited 2018 Nov 28];23:682–91. Available from:

https://www.sciencedirect.com/science/article/pii/S2211124718304431

24. Murugan R, Buchauer L, Triller G, Kreschel C, Costa G, Pidelaserra Martí G, et al. Clonal selection drives protective memory B cell responses in controlled human malaria infection. Sci Immunol [Internet]. 2018 [cited 2018 Nov 29];3:eaap8029. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29453292

25. Feder AF, Kryazhimskiy S, Plotkin JB. Identifying Signatures of Selection in Genetic Time Series. Genetics [Internet]. 2014 [cited 2019 Feb 18];196:509–22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24318534

26. Fischer W, Ganusov V V., Giorgi EE, Hraber PT, Keele BF, Leitner T, et al. Transmission of Single HIV-1 Genomes and Dynamics of Early Immune Escape Revealed by Ultra-Deep Sequencing. Nixon DF, editor. PLoS One [Internet]. 2010 [cited 2018 Nov 29];5:e12303. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20808830

27. Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, Macalalad AR, et al. Whole Genome Deep Sequencing of HIV-1 Reveals the Impact of Early Minor Variants Upon Immune Recognition During Acute Infection. Walker CM, editor. PLoS Pathog [Internet]. 2012 [cited 2018 Nov 29];8:e1002529. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22412369

28. Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H, et al. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J Virol [Internet]. 1999 [cited 2018 Nov 29];73:10489–502. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10559367

29. Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, Albert J, et al. Population

genomics of intrapatient HIV-1 evolution. Elife [Internet]. 2015 [cited 2018 Nov 29];4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26652000

30. Nourmohammad A, Otwinowski J, Luksza M, Mora T, Walczak AM. Clonal competition in B-cell repertoires during chronic HIV-1 infection. bioRxiv [Internet]. Cold Spring Harbor Laboratory; 2018 [cited 2018 Nov 29];271130. Available from: https://www.biorxiv.org/content/early/2018/02/26/271130

31. Wu X, Zhang Z, Schramm CA, Joyce MG, Do Kwon Y, Zhou T, et al. Maturation and Diversity of the VRC01-Antibody Lineage over 15 Years of Chronic HIV-1 Infection. Cell [Internet]. 2015 [cited 2018 Nov 29];161:470–85. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25865483

32. Landais E, Murrell B, Briney B, Murrell S, Rantalainen K, Berndsen ZT, et al. HIV Envelope Glycoform Heterogeneity and Localized Diversity Govern the Initiation and Maturation of a V2 Apex Broadly Neutralizing Antibody Lineage. Immunity [Internet]. 2017 [cited 2019 Mar 4];47:990–1003.e9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29166592

33. Yusim K, Korber BTM, Brander C, Barouch D, Boer R de, Haynes BF, et al., editors. HIV Molecular Immunology 2016 [Internet]. Los Alamos: Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico; 2016 [cited 2018 Nov 29]. Available from: https://www.hiv.lanl.gov/content/immunology/contents2016.html

34. Hatada M, Yoshimura K, Harada S, Kawanami Y, Shibata J, Matsushita S. Human immunodeficiency virus type 1 evasion of a neutralizing anti-V3 antibody involves acquisition of a potential glycosylation site in V2. J Gen Virol [Internet]. 2010

[cited 2018 Nov 29];91:1335–45. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20032207

35. Ringe R, Das L, Choudhary I, Sharma D, Paranjape R, Chauhan VS, et al. Unique C2V3 sequence in HIV-1 envelope obtained from broadly neutralizing plasma of a slow progressing patient conferred enhanced virus neutralization. PLoS One [Internet]. Public Library of Science; 2012 [cited 2018 Nov 29];7:e46713. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23056416

36. Maldarelli F, Kearney M, Palmer S, Stephens R, Mican J, Polis MA, et al. HIV populations are large and accumulate high genetic diversity in a nonlinear fashion. J Virol [Internet]. American Society for Microbiology Journals; 2013 [cited 2018 Dec 6];87:10313–23. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23678164

37. Redd AD, Quinn TC, Tobian AA. Frequency and implications of HIV superinfection. Lancet Infect Dis [Internet]. 2013 [cited 2019 Mar 4];13:622–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23726798

38. Ernst MD. Permutation Methods: A Basis for Exact Inference. Stat Sci. 2004;19:676–685.

39. Eroshkin AM, LeBlanc A, Weekes D, Post K, Li Z, Rajput A, et al. bNAber: database of broadly neutralizing HIV antibodies. Nucleic Acids Res [Internet]. Oxford University Press; 2014 [cited 2019 Jan 7];42:D1133–9. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt1083

40. Strauli NB, Hernandez RD. Statistical inference of a convergent antibody repertoire response to influenza vaccine. Genome Med [Internet]. BioMed Central; 2016 [cited 2018 Nov 15];8:60. Available from:

http://genomemedicine.biomedcentral.com/articles/10.1186/s13073-016-0314-z

41. Bonsignori M, Liao H-X, Gao F, Williams WB, Alam SM, Montefiori DC, et al. Antibody-virus co-evolution in HIV infection: paths for HIV vaccine development. Immunol Rev [Internet]. 2017 [cited 2019 Feb 4];275:145–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28133802

42. Brandman R, Brandman Y, Pande VS. Sequence Coevolution between RNA and Protein Characterized by Mutual Information between Residue Triplets. Antopolsky M, editor. PLoS One [Internet]. Public Library of Science; 2012 [cited 2019 Jan 14];7:e30022. Available from: https://dx.plos.org/10.1371/journal.pone.0030022

43. Gloor GB, Martin LC, Wahl LM, Dunn SD. Mutual Information in Protein Multiple Sequence Alignments Reveals Two Classes of Coevolving Positions†. American Chemical Society; 2005 [cited 2019 Jan 14]; Available from: https://pubs.acs.org/doi/abs/10.1021/bi050293e

44. Marino Buslje C, Teppa E, Di Doménico T, Delfino JM, Nielsen M. Networks of High Mutual Information Define the Structural Proximity of Catalytic Sites: Implications for Catalytic Residue Identification. Rost B, editor. PLoS Comput Biol [Internet]. Public Library of Science; 2010 [cited 2019 Jan 14];6:e1000978. Available from: https://dx.plos.org/10.1371/journal.pcbi.1000978

45. Simonetti FL, Teppa E, Chernomoretz A, Nielsen M, Marino Buslje C. MISTIC: Mutual information server to infer coevolution. Nucleic Acids Res [Internet]. Oxford University Press; 2013 [cited 2019 Jan 14];41:W8-14. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23716641

46. Avila-Herrera A, Pollard KS. Coevolutionary analyses require

phylogenetically deep alignments and better null models to accurately detect inter-protein contacts within and between species. BMC Bioinformatics [Internet]. 2015 [cited 2019 Mar 4];16:268. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26303588

47. Bonsignori M, Kreider EF, Fera D, Meyerhoff RR, Bradley T, Wiehe K, et al. Staged induction of HIV-1 glycan–dependent broadly neutralizing antibodies. Sci Transl Med [Internet]. 2017 [cited 2019 Mar 4];9:eaai7514. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28298420

48. Bonsignori M, Zhou T, Sheng Z, Chen L, Gao F, Joyce MG, et al. Maturation Pathway from Germline to Broad HIV-1 Neutralizer of a CD4-Mimic Antibody. Cell [Internet]. 2016 [cited 2019 Mar 4];165:449–63. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26949186

49. Gao F, Bonsignori M, Liao H-X, Kumar A, Xia S-M, Lu X, et al. Cooperation of B Cell Lineages in Induction of HIV-1-Broadly Neutralizing Antibodies. Cell [Internet]. 2014 [cited 2019 Mar 4];158:481–91. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25065977

50. Brennan PLR, Prum RO. Mechanisms and Evidence of Genital Coevolution: The Roles of Natural Selection, Mate Choice, and Sexual Conflict. Cold Spring Harb Perspect Biol [Internet]. Cold Spring Harbor Laboratory Press; 2015 [cited 2019 Mar 4];7:a017749. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26134314

51. Landais E, Huang X, Havenar-Daughton C, Murrell B, Price MA, Wickramasinghe L, et al. Broadly Neutralizing Antibody Responses in a Large Longitudinal Sub-Saharan HIV Primary Infection Cohort. Trkola A, editor. PLOS Pathog [Internet]. 2016 [cited 2019 Mar 14];12:e1005369. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/26766578

52. Doria-Rose NA, Klein RM, Daniels MG, O'Dell S, Nason M, Lapedes A, et al. Breadth of Human Immunodeficiency Virus-Specific Neutralizing Activity in Sera: Clustering Analysis and Association with Clinical Variables. J Virol [Internet]. 2010 [cited 2019 Mar 14];84:1631–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19923174

53. Piantadosi A, Panteleeff D, Blish CA, Baeten JM, Jaoko W, McClelland RS, et al. Breadth of Neutralizing Antibody Response to Human Immunodeficiency Virus Type 1 Is Affected by Factors Early in Infection but Does Not Influence Disease Progression. J Virol [Internet]. 2009 [cited 2019 Mar 14];83:10269–74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19640996

54. Nourmohammad A, Otwinowski J, Plotkin JB. Host-Pathogen Coevolution and the Emergence of Broadly Neutralizing Antibodies in Chronic Infections. Cobey S, editor. PLOS Genet [Internet]. 2016 [cited 2019 Mar 14];12:e1006171. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27442127

55. Wang C, Sanders CM, Yang Q, Schroeder HW, Wang E, Babrzadeh F, et al. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. Proc Natl Acad Sci [Internet]. 2010;107:1518–23. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0913939107

56. Vander Heiden JA, Yaari G, Uduman M, Stern JNH, O'connor KC, Hafler DA, et al. PRESTO: A toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. Bioinformatics. 2014;30:1930–2.

57. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment

search tool. J Mol Biol [Internet]. Academic Press; 1990 [cited 2018 Nov 9];215:403–10. Available from: https://www.sciencedirect.com/science/article/pii/S0022283605803602?via%3Dihub

58. Foley B, Leitner T, Apetrei C, Hahn B, Mizrachi I, Mullins J, et al., editors. HIV Sequence Compendium 2017 [Internet]. Los Alamos: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, NM; 2017. Available from: http://www.hiv.lanl.gov/

59. Sturdevant CB, Joseph SB, Schnell G, Price RW, Swanstrom R, Spudich S. Compartmentalized Replication of R5 T Cell-Tropic HIV-1 in the Central Nervous System Early in the Course of Infection. Lifson J, editor. PLOS Pathog [Internet]. Public Library of Science; 2015 [cited 2018 Nov 9];11:e1004720. Available from: https://dx.plos.org/10.1371/journal.ppat.1004720

60. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol [Internet]. Academic Press; 1970 [cited 2018 Nov 11];48:443–53. Available from: https://www.sciencedirect.com/science/article/pii/0022283670900574?via%3Dihub

61. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet [Internet]. Elsevier Current Trends; 2000 [cited 2018 Nov 11];16:276–7. Available from: https://www.sciencedirect.com/science/article/pii/S0168952500020242?via%3Dihub

62. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res [Internet]. Oxford University Press; 2002 [cited 2018 Nov 11];30:3059–66. Available

from: http://www.ncbi.nlm.nih.gov/pubmed/12136088

63. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. Poon AFY, editor. PLoS One [Internet]. Public Library of Science; 2010 [cited 2018 Nov 11];5:e9490. Available from: https://dx.plos.org/10.1371/journal.pone.0009490

64. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol [Internet]. Oxford University Press; 2016 [cited 2018 Nov 26];33:1635–8. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msw046

65. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol [Internet]. Oxford University Press; 1986 [cited 2018 Nov 12];3:418–26. Available from: https://academic.oup.com/mbe/article/3/5/418/988012/Simple-methods-for-estimating-the-numbers-of

66. Orton R. Calculating dNdS for NGS datasets | Bioinformatics I/O [Internet]. 2014 [cited 2019 Mar 4]. Available from: http://bioinformatics.cvr.ac.uk/blog/calculating-dnds-for-ngs-datasets/

67. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ [Internet]. PeerJ Inc.; 2016 [cited 2018 Nov 15];4:e2584. Available from: https://peerj.com/articles/2584

68. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res [Internet]. Oxford University Press; 2013 [cited 2018 Nov 15];41:W34–40. Available from:

http://academic.oup.com/nar/article/41/W1/W34/1097536/IgBLAST-an-immunoglobulin-variable-domain-sequence

69. Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, Chaume D, et al. IMGT/LIGM-DB, the IMGT(R) comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. Nucleic Acids Res [Internet]. Oxford University Press; 2006 [cited 2018 Nov 15];34:D781–4. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkj088

70. Yaari G, Uduman M, Kleinstein SH. Quantifying selection in high-throughput Immunoglobulin sequencing data sets. Nucleic Acids Res [Internet]. 2012 [cited 2018 Nov 15];40:e134–e134. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22641856

71. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using **lme4**. J Stat Softw [Internet]. 2015 [cited 2018 Nov 26];67:1–48. Available from: http://www.jstatsoft.org/v67/i01/

72. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res [Internet]. 2003 [cited 2019 Mar 4];13:2498–504. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14597658

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature

3/26/2019
Date