Quantifying how the mutational tolerance of HIV's envelope protein shapes its evolution

Hugh Haddox

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Jesse D. Bloom, Chair

Julie Overbaugh

Michael Emerman

Program Authorized to Offer Degree:
Molecular and Cellular Biology

University of Washington

Abstract

Quantifying how the mutational tolerance of HIV's envelope protein shapes its evolution

Hugh Haddox

Chair of the Supervisory Committee:
Associate Member Jesse D. Bloom
Fred Hutchinson Cancer Research Center

HIV's most rapidly evolving proteins is its envelope protein (Env). This rapid evolution is driven by continuous selection to evade immunity within HIV-infected hosts. However, as Env evolves, it is also under functional constraint to perform essential functions in the viral lifecycle, including receptor binding and membrane fusion. Since both of the above forces strongly shape Env's evolution, their effects have been difficult to disentangle from one another. As a result, our understanding of these forces is far from complete. A central goal of my graduate research has been to experimentally measure the functional constraint on Env in the lab in the *absence* of external immune selection. There are $\approx$10,000 single amino-acid mutations to a protein of Env's length ($\approx 19 \times 850$). Using a high-throughput technique called deep mutational scanning, I measured the effects of each of these mutations to Env in context of viral replication in cell culture. The results provide an in-depth profile of Env's ability to tolerate each of the 20 amino acids at each site in the protein. Using these data, I examined Env's mutational tolerance variable loops, which rapidly evolve to evade antibodies. It is possible that these loops have a high tolerance for mutations, and that that is one reason they so readily evolve. However, I did not find statistical support that these loops are more tolerant of mutations than other parts of the protein, suggesting that their variability in nature may mainly be due to high levels of diversifying pressure from antibodies. I also examined epitopes of broadly neutralizing antibodies targeting the

CD4 binding site. These epitopes that are highly conserved in nature and are targets in vaccine design. A common assumption is that this conservation is due to high functional constraint at these sites. Indeed, I found that they were less tolerant of mutations than other parts of Env, providing rigorous support for a long-standing hypothesis, and suggesting that these epitopes may have a diminished evolutionary capacity to evade antibodies relative to other sites in the protein, which would make them more vulnerable to immune targeting. Another central goal of my thesis has been to compare Env's mutational tolerance among divergent strains. The same mutation (e.g., A12N) can have different effects in two related proteins due to epistasis (e.g., A12N may only be tolerated in one homolog, but not the other). However, the extent that mutational effects to Env differ between divergent HIV strains is largely unknown. To address this knowledge gap, I repeated the deep mutational-scanning experiment of two Env homologs that have 85% amino-acid identity. The results allowed me to compare each homolog's ability to tolerate each of the 20 amino acids at 616 homologous sites. I found that at a small fraction of sites, the amino acids tolerated in one homolog were largely distinct from the amino acids tolerated in the other homolog. However, only a few sites showed such extreme differences; most sites had changes in mutational tolerance that were only small-to-intermediate in effect size. Thus, these results indicate that Env's mutational tolerance is still substantially conserved between homologs. Overall, my graduate research has increased our knowledge of how Env's underlying mutational tolerance shapes the evolution of antibody epitopes, providing experimental support for the assumption that conserved epitopes targeted in vaccine design are indeed less tolerant of mutations than the rest of the protein, and may thus be less likely to evade an immune response. This work also provides a comprehensive measure of differences in mutational effects across Env, finding that mutational effects are largely conserved between divergent homologs. More broadly, this work was also the first time that deep mutational scanning had been used to comprehensively measure

mutational effects to an HIV protein in context of viral replication. In the future, this technique could be adapted to study any phenotype that is selectable in the lab (e.g., antibody escape).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Graduate school has been an exhilarating experience, largely due to the people involved. First, I would like to thank my mentor, Jesse. Jesse was a large reason why I decided to come to UW and the Fred Hutch for graduate school, not only because I found his research to be extremely interesting, but because he reached out to me during the application process. At the time, my undergraduate mentor, Shelley Copley, had suggested that I could also consider Jesse's lab for a postdoc if I decided on a different school. Reasonable though this suggestion was, I'm extremely happy I decided *not* to kick the can down the road! The Bloom lab has been an excellent place to conduct graduate research, and Jesse has been a large part of that. I was one of his first graduate students. When brainstorming projects, I was extremely fascinated by one he had recently devised on the evolvability of HIV. This project was certain to be challenging, since the lab primarily works on influenza. But, I had no idea how much troubleshooting this project would entail, or the twists and turns it would take. I have many people to thank for its ultimate success, but Jesse's role was the most instrumental: close mentorship at every step, even helping me with difficult or time-consuming experiments on multiple occasions, and providing a constant source of encouragement when the experiments failed, failed even worse than the first time, and then finally succeeded. I leave having learned a huge amount from Jesse, both computationally and experimentally, and in general about how to conduct careful and creative research. Unfortunately, I didn't find out until recently that the only thing I *actually* needed to learn from Jesse in order to graduate was how to beat him in a 5K, which for some reason he has never offered to advise me on how to do.

Next, I would like to thank Adam Dingens. Adam – my little sib during MCB recruitment

– joined the lab a year after me with a fire in his eyes (or if not a fire at least a smoldering glow). His ambition? To use deep mutational scanning to identify antibody-escape mutations to Env. Adam was cautious at first – he didn't want to interfere with my plans if I wanted to head in that direction. But, as I had my eye on Env's disulfide bonds, he started in on the antibody-escape project. And I am extremely glad he did. Working closely with Adam has been one of the best parts of graduate school. He has been invaluable in helping me relate my findings to the broader field of HIV, which he has a very impressive grasp of. And it has been amazing to watch him accomplish the very goals he set at the start of graduate school. Being part of this has allowed me to see results for an ever-increasing number of broadly neutralizing antibodies, all without ever having done a single neutralization assay (Adam cannot believe I got away without doing at least one). His friendly and energetic personality, and love for bad puns, has made sitting in the HIV corner of the lab a wonderful experience that I will greatly miss.

While on the topic of the HIV corner, I would also like to thank an honorary member of the corner – Sarah Hilton. First, I would like to thank her for bearing with me when, on one of the first days of her rotation, when I was helping her gel purify some DNA, I mindlessly threw the gel in the trash after taking a picture of it, but before cutting anything out! I would also like to thank her for all the pointers she has given me in using `phydms` to study Env's evolution, which has lead to some fascinating findings, and Alexandria Wilson for expertly spearheading this line of work this summer, revealing even more interesting (if somewhat confusing!) results.

I would also like to thank the other members of the Bloom lab, both past and current. The lab has been an amazing environment in which to do work, both intellectually and socially. Getting to speculate and problem solve with other people is one of my main joys in science, and this happens on a daily basis in the lab. In particular, I would like to thank Mike Doud and Orr Ashenberg for helpful discussions on comparing amino-acid

preferences between homologs, and both Alistair Russell and Orr for general advice and guidance. A special thanks to Orr for using his special powers as a postdoc to eyeball the DNA concentration of my samples to make sure that the nanodrop was well calibrated.

I have also greatly enjoyed the collaborative environment at the Hutch. In particular, I would like to thank both Julie Overbaugh and Michael Emerman for extremely helpful advise on my project and career. I was the first person in the Bloom lab to work on HIV, so that when it came to specific experimental techniques, as well as relating my work to broader questions in the field, I felt a bit like a fish out of water. The progress I have made would have been impossible without their assistance. I have used Julie as a sounding board on many occasions, and have greatly enjoyed being included in the discussions she organizes surrounding the grant on mother-to-child HIV transmission. She has also been a great person to brainstorm ideas with – including using deep mutational scanning to map antibodies with unknown specificities and how to adapt the technique to map non-neutralizing antibodies. Michael has also been a great sounding board, and organizes the Virology Group Meetings, which have been one of the highlights of graduate school. Almost every time I present, he has come up with insightful comments, at one point suggesting I consider miniprepping unintegrated HIV from cells (a crazy-sounding idea that has worked extremely well!), and at another point helping me realize the importance of my findings on the mutational tolerance at variable loops and epitopes of broadly neutralizing antibodies. Michael and Julie have also been generous in providing plasmids and cells for my experiments.

Members of the Overbaugh and Emerman labs have also been extremely helpful. In particular, I am very grateful to Lily Wu for both training me to work in the HIV room, and for answering any number of questions about basic experimental techniques. I am also very grateful to Stephanie Rainwater, who has likewise answered a countless number of questions on experimental techniques.

Next, I would like to thank the other members of my committee – Jay Shendure and Kelly Lee. My rotation in Jay's lab my first year was a fascinating glimpse into an area of research I had never experienced until then. Despite the size of the lab, Jay made time to regularly meet with me. After I joined the Bloom lab, he has continued to make himself readily available whenever I have wanted to chat about research or careers. I always enjoy interacting with Jay, not only for the advise and insightful comments he has, but also because his high level of enthusiasm for science has an R-nought that is through the roof! Kelly has also been a great committee member, providing excellent insight in committee meetings. I particularly enjoy the ideas of potentially interesting ways to use deep mutational scanning he has spontaneously proposed over emails on multiple occasions.

This section would not be complete without thanking Shelley Copley and Juhan Kim. I conducted four years of research in Shelley's lab as an undergraduate at the University of Colorado. This experience was transformative, nucleating the deep passion for scientific research I have today. Shelley was a fantastic mentor, who was extremely generous with her time. Aside from providing close and insightful mentorship throughout my time in the lab, she also took extra time to regularly read and discuss papers from the field with me one-on-one. This contributed much to the knowledge base I have today, and gave me a first introduction into Jesse's work, she having picked multiple papers of his to discuss. Juhan Kim, a postdoc in the Copley lab, also played a central role in my research experience. He devoted an extraordinary amount of time to helping mentor me at the bench, and was and continues to be a fantastic source of encouragement and friendship. Overall, the opportunities and mentorship Shelley and Juhan gave me were remarkable, and were foundational to my education.

The Hutch and UW have a rich scientific environment. Two groups I would like to particularly acknowledge are the Virology Group and the Genomics Salon. Virology Group has

provided an amazing platform not only to hear about a wide variety of research from other labs, but also a chance to present to and get feedback from the rich virology community in Seattle. I have also greatly enjoyed being part of the The Genomics Salon, which braids together issues in science and society, and has challenged me to take a step back from lab and think about scientific issues in a broader context, which I have come to realize is extremely important. Helping to organize this group, spearheaded by Katherine Xue, with Jolie Carlisle, and Orlando de Lange has been a very fun and rewarding experience.

Nearing the end of this section (almost through, diligent readers who have not already skipped to Chapter 1), I would like to warmly thank the many friends I have had outside of lab who have made this period of my life so enjoyable. This includes a huge number of friends from soccer, climbing, and book clubs. People with whom I have sung King George's part in Hamilton at the top of my lungs while driving across the Australian countryside, and who have sung to me Happy Birthday in both Australian and then Texan accents depending on where we were at the time. People who share my love for brunch and board games. Family friends who regularly invite me and Betsy over for parties and meals, expertly reviving my grandmother's recipe for beef stroganoff. Cousins who have treated me to high tea on the shores of Lake Louise. And my sincerest grati-hugh-tude goes out to Katherine X-hugh and Anne Clark for their constant s-hugh-pply of Hugh puns over the years, and an abundance of f-anne-tastic birthday cakes.

Lastly, I would like to thank my family. My parents, Charley and Lisa, and my sister, Betsy, have been a constant source of happiness and support over my entire life. There have been many moments in graduate school where they have propped me up. They are always among the people most interested to hear about the successes. And they have enriched my life in countless other ways. Go to them for humorous books by PG Wodehouse, captivating musicals about our founding fathers, videos from Saturday Night Live, or a good game of hearts. In particular, getting to once again live in the same city as

Betsy has been immensely enjoyable! I cannot thank them enough.

Chapter 1

INTRODUCTION

A variety of forces shape the evolution of HIV's envelope protein. HIV continues to be a large public-health burden. The WHO estimates that in 2016, there were $>$35 million people living with HIV, with 1.8 million new infections that year (`http://www.who.int/hiv/data/en/`). A characteristic feature of HIV is its ability to rapidly evolve. This evolution has devastating consequences, as it is a major means by which HIV evades both human immunity and anti-viral drugs [134].

One of HIV's most rapidly evolving proteins is its envelope protein (Env) [85]. As with any protein, Env's evolution is shaped by a combination of mutation and selection. Many of these evolutionary forces have been well studied. On the side of mutation, Env's evolution is fueled by HIV's high mutation rate. Laboratory experiments have estimated the mutation rate of HIV's reverse transcriptase to be $\sim$1-3$\times 10^{-5}$ mutations per base pair per replication cycle [103, 3], corresponding to approximately 0.1-0.3 mutations per viral genome per replication cycle. HIV also has a high rate of recombination, which can facilitate adaptation by consolidating beneficial mutations. HIV virions package two RNA genomes. During reverse transcription, the reverse transcriptase can switch between these genomes, leading to recombination [71, 76, 190]. The frequency of this switching has been estimated to occur at a high rate of $\sim$2-3 times per genome [76, 190]. These processes play out on remarkably fast time scales: it is estimated that HIV's generation time is $\sim$2-3 days, and that the number of new virions generated in HIV-infected humans is $\sim 10^9$ per day [128], allowing the virus to rapidly explore its immediate evolutionary space.

On the side of selection, one of the main drivers of Env's evolution is the adaptive race between Env and the immune system of HIV-infected hosts. Env is the only HIV-encoded protein on the surface of viral particles, and is the only known target of antibodies that neutralize the virus. Within months of being infected with HIV, most humans generate anti-Env neutralizing antibodies [6, 174, 139]. However, Env readily evades this initial response through sequence evolution driven by the process described above [6, 174, 139]. Host immunity can then respond by adapting to target new viral variants [174, 139]. Indeed, the immune system has considerable adaptive power: it can elicit multiple antibodies that target different parts of an antigen and optimize binding of these antibodies through somatic hypermutation. However, despite this adaptive potential, immune responses are unable to successfully pin down Env, which readily evades new responses each time they arise [174, 139].

Because HIV causes chronic infections that last years, a large amount of evolution in Env can occur in just a single individual. It has been estimated that, at the DNA level, Env diverges at a rate of $\sim$1% per year of infection, such that after a decade, Env variants can differ from the infecting virus at $\sim$10% of sites [148]. At the same time, the expansion of multiple independent lineages can lead to a high level of standing diversity. This diversity was also found to reach $\sim$10% divergence between Env variants in a single patient after a decade [148]. To put this into perspective, previous work has estimated that Env's diversity in a single individual can reach the same level of diversity as hemagglutinin sequences from all influenza viruses circulating across the entire globe in a given year [85].

The diversity in Env generated within hosts has translated to an even higher level of global diversity in the human population. There have been multiple introductions of HIV into the human population, one of which (HIV-1 group M) gave rise to the current global pandemic [85]. The introduction of group M likely occurred in the first half of the 20th century [86, 47]. Since then, several phylogenetically distinct group-M subtypes have simultaneously spread across the globe, where Env variants from different subtypes tend to differ by 20-35% amino-acid divergence [85]. A large number of sequences have also

radiated from the base of each subtype, where even with a subtype, variants typically differ by 15-20% amino-acid divergence [85]. The spread of such a large number of diverse strains is thought to be facilitated by the fact that much of the human population lacks pre-existing immunity to HIV [61]. In stark contrast, the global diversity of influenza virus experiences frequent bottlenecks, such that only single seasonal influenza lineages (e.g., H3N2) persist over long periods of time. This bottlenecking is thought to partially be due to the fact that much of the human population has pre-existing anti-influenza immunity, which imposes strong selection for immune evasion [61]. Thus, Env's rapid evolution within and between hosts is driven by a combination of forces, which include HIV's high rate of mutation and recombination, diversifying selection to evade immunity, and the lack of pre-existing immunity in much of the human population.

Although Env is able to rapidly evolve, its evolutionary capacity is still limited. That is because Env is also continually under purifying selection to retain its ability to perform essential functions in the viral lifecycle. Specifically, Env functions to bind to cell-surface receptors and then fuse the viral and host membranes, allowing the virus to enter the cell. These functions involve a series of elaborate conformational changes [181]. The first step in this series is for Env to bind HIV's primary receptor, CD4. This binding event induces a dramatic conformational change that exposes Env's co-receptor binding site. Upon binding a co-receptor (typically either CCR5 or CXCR4), Env then undergoes a second large conformational change leading to the insertion of Env's fusion peptide in the cell membrane, and subsequent fusion of the viral and cell membranes. Membrane fusion allows viral proteins and genetic material to enter the cell and complete the viral lifecycle, using cellular resources to ultimately generate viral progeny.

Thus, Env's evolution involves a balance of selective forces, including external selective pressures to evade immunity, which drive Env's evolution, and inherent selective pressures for Env to retain its function, which constrain Env's evolution. Knowledge of these forces is important for a basic understanding of how Env evolves. However, since both of these forces strongly shape Env's evolution, their effects have been difficult to

disentangle from one another, leaving large gaps in our knowledge. The next section describes two such gaps with important implications for understanding Env's ability to readily evade antibodies and which regions of Env may be most vulnerable to antibody targeting.

What is the inherent mutational tolerance of Env's antibody epitopes? A major goal in the field of HIV has been to determine how Env so readily evades the immune system. A parallel goal is to find ways to overcome Env's immune defenses in order to design effective and long-lasting medical interventions. Research in this area has revealed that Env's defenses are manifold. One defense mechanism is dense glycosylation of Env's surface. Structural analysis of Env's glycans, which typically make up an astounding $\sim$50% of Env's mass [94], indicate that they physically shield a large fraction of Env's underlying polypeptide chain from antibodies [160]. Moreover, this shield can evolve to add or remove glycans to thwart antibodies that overcame a previous glycan arrangement [174]. Another of Env's defenses is conformational masking of conserved, functionally important regions [44, 60]. For instance, Env is highly conformationally dynamic, only exposing the CD4 binding site in a subset of conformers. Biochemical evidence suggests that antibodies targeting this site incur a high entropic cost upon binding and ordering Env, decreasing the ability of such antibodies to make energetically favorable interactions [44]. In turn, the co-receptor binding site is conformationally masked in the sense that it is not stably exposed until Env binds CD4 [167, 89, 28]. A third defense is the structurally recessed nature of the CD4 binding site, which may constrain the ability of antibodies to target this region [180, 189]. In comparison, Env's surface-exposed variable loops are commonly targeted by humoral immunity [114], which they readily evade through rapid evolution. Yet another potential defense mechanism is the production of non-functional forms of Env (e.g., via shedding of Env's gp120 subunit) that expose epitopes that are inaccessible in context of intact Env. As a result, these non-functional forms can elicit large numbers of non-neutralizing antibodies, potentially diverting the immune system to epitopes that may be less useful for controlling viral spread [113], although even non-neutralizing antibodies

can have anti-viral effects [122]. Whether or not these “defenses” were evolved for this purpose is open for debate; however, each seem to make it difficult for the immune system to efficiently neutralize the virus.

Another important, but less-studied determinant of antibody escape is Env’s mutational tolerance, i.e., Env’s ability to tolerate mutations while still retaining its ability to fold and perform its basic functions. If a mutation completely disrupts Env’s ability to fold or function, it would not be expected to be beneficial, even if it abrogated antibody binding. Thus, the ability of a site to evade an antibody is intricately tied to its underlying mutational tolerance. It is intriguing to consider that some of Env’s epitopes are more tolerant of mutations than others, and thus have a higher evolutionary capacity to evade the immune system. For instance, Env has several surface-exposed loops with high sequence variability in nature [156, 111]. These loops are commonly targeted by antibodies during HIV infection [114], which they evade via sequence evolution. It is possible that this evolution is facilitated by a high underlying tolerance for mutations. Indeed, a recent study found that immunodominant regions of influenza’s hemagglutinin protein tended to be more tolerant of mutations than other sites in the protein, suggesting that these sites are inherently more evolutionarily pliable, helping to explain hemagglutinin’s ability to evade immunity [164]. Although the same phenomenon may be true for Env’s variable loops, this hypothesis has never been rigorously tested. An alternative hypothesis is that these loops evolve rapidly merely because they are under a high level of diversifying selection as a result of frequent immune targeting. Thus, even if these loops have an average mutational tolerance, they might still be expected to be highly variable. Differentiating between these hypotheses requires disentangling the role of inherent vs. external pressures shaping Env’s evolution.

In contrast to the variable loops, there is a large amount of interest in identifying conserved epitopes that may be more susceptible to immune targeting. A massive effort has identified many “broadly neutralizing antibodies” (bNAbs) that are effective at neutralizing a high fraction of diverse viruses [153, 193, 14, 46, 72, 146]. These antibodies target several conserved regions spanning the length of Env. Currently, there are efforts to elicit

such bNAbs using vaccination or to passively administer bNAbs to humans for prophylactic or therapeutic purposes [88, 104]. Aside from their breadth, another appealing aspect of these antibodies is that they may be more difficult for Env to evade. Not only are their epitopes conserved, but they often overlap with regions of known functional constraint (e.g., the CD4 receptor binding site), such that mutations in these regions may frequently disrupt Env's ability to fold and function. However, this hypothesis has been difficult to test. Once again, an alternative hypothesis to one involving Env's underlying mutational tolerance is that these patterns of sequence conservation may be primarily due to external selective pressures. The epitopes of bNAbs may be under weaker immune selection than other sites in the protein. Only 20% of HIV-infected individuals develop broadly neutralizing antibodies, and only after multiple years of infection. Perhaps these epitopes would evolve much more rapidly if they were under increased selection to do so. Testing these hypotheses requires disentangling inherent vs. external selective pressures.

One way to study inherent functional constrains is to measure the effects of mutations on Env's ability to support viral replication in the lab in the absence of external selective pressures. A large number of studies have done so [119, 34, 10, 56, 98, 74]. These studies have measured the effects of mutations on a wide variety of properties that influence Env's ability to both fold and function, including Env's ability to bind its receptor and co-receptor, to fuse viral and host membranes, to be cleaved by proteases into gp120 and gp41 subunits, and for gp120 and gp41 to remain non-covalently associated following this cleavage event. However, these studies have only been able to examine the effects of a small fraction of mutations to Env, which is no surprise, given that there are $\sim$16,000 single amino-acid mutations to a protein of Env's length ($\approx 19 \times 850$). Thus, we still lack a comprehensive understanding of the ability of each site in Env to tolerate mutations when the only selection is for Env to perform its basic functions in the absence of immune selection.

How much has Env's mutational tolerance shifted over evolutionary time? In general, protein evolution is thought to follow a stepwise process, involving the incremental accumulation of single mutations, each tolerated in the background in which it occurs. The immediate evolutionary space accessible to any protein is thus largely defined by the effects of single point mutations to the protein on organismal fitness. A large amount of work has been devoted to characterizing the effects of mutations on Env's ability to replicate in cell culture in both the presence and absence of immune selection [119, 34, 10, 56, 98, 74, 192, 125, 97, 100]. For practical purposes, most of these studies only characterized mutations in one or a few homologs. However, the effects of mutations to a protein can change over evolutionary time due to epistasis.

One form of epistasis is intra-protein epistasis, i.e., epistasis between two or more residues in the same protein, where the effect of a mutation at one site is influenced by which amino acids are present at other sites. This phenomenon has been well documented and can dramatically alter a protein's adaptive landscape [175, 20, 13, 121, 99, 58, 116, 66, 129]. For illustrative purposes, I will briefly describe one example of this involving the adaptation of vertebrate steroid receptors [121]. Vertebrates have multiple steroid-receptor homologs with differing specificities. These homologs likely evolved from an ancestral version with a single specificity. Through phylogenetic analysis, this study identified several historical substitutions that conferred a switch in receptor specificity along one lineage. However, when these substitutions were introduced by themselves into an ancient version of the receptor, they led to a non-functional protein. This study went on to identify another set of historical substitutions that had a neutral effect when introduced into the ancient receptor by themselves. Strikingly, when the two sets of substitutions were combined in the ancient background, the protein was both highly functional and had switched in its specificity. Thus, the latter group of mutations were permissive in the sense that they enabled the function-switching mutations to be tolerated, but did not confer this switch alone. This example shows that the effects of mutations can "shift" over time (e.g., the function-switching mutations shifted from not being tolerated to being

tolerated).

By analogy, it is possible that the many mutations that have naturally accumulated during Env's evolution have caused the effects of mutations to shift among divergent Env homologs. It seems especially important to consider this possibility given that Env is so diverse in nature. As described above, Env homologs often differ by 20-35% amino-acid divergence, which translates into $>$100 amino-acid differences (Env is $\sim$850 amino acids in length). In principle, even single mutations can substantially shift a protein's mutational tolerance [175, 20, 121, 58, 66]. However, the prevalence of epistasis in long-term protein evolution is also poorly understood, not just for Env, but also in general for any protein.

Some lines of evidence indicate that epistasis may be common in long-term protein evolution. For example, computational simulations suggest that upon an amino-acid change, a protein will rapidly evolve to increase its preference for the derived amino acid over the ancestral state [130]. Thus, proteins that differ by many amino acids might be expected to have radically shifted amino-acid preferences. Another study indicated that epistasis has been prevalent in the long-term evolution a bacterial enzyme involved in leucine biosynthesis [99]. This study compared two homologs of this enzyme with $\sim$50% identity at the amino-acid level. Specifically, they studied the effects of individually switching wildtype amino acids between homologs at each site that differed in wildtype sequence. In the absence of epistasis, these mutations might be expected to be reasonably well tolerated since both homologs are functional enzymes. In the single homolog in which they measured these effects, although most mutations were neutral, approximately a third of mutations decreased enzyme activity, with a fraction leading to severe decreases. Summing these individual effects, which assumes no epistasis, leads to a total decrease that is over 10 times the magnitude of the enzyme's starting catalytic efficiency. The fact that the full combination of mutations actually gives rise to a functional protein (i.e., the other homolog) suggests the effects of the mutations tested were substantially altered by epistasis in this enzyme's natural evolution. While this result is striking, one shortcoming of this study is that it only tests one mutation per site at a subset of sites in the protein. Per-

haps epistasis is less prevalent among sites that are conserved in sequence, or for other amino-acid mutations at the variable sites.

Other lines of evidence point that the effects of mutations on protein stability and function are often conserved in long-term evolution. For instance, a high-throughput study measured the effects of all single amino-acid mutations to two homologs of influenza nucleoprotein with 94% amino-acid identity [41]. Of all 9,443 (=497$\times$19) mutations tested, the effects of these mutations were mostly conserved between homologs, with only a small fraction of mutations having variable effects. A different, lower-throughput study of nucleoprotein measured the effects of multiple mutations that dramatically impacted nucleoprotein stability [8]. This study found that the stability effects were largely conserved among several homologs that were related by 94%-72% amino-acid identity. A third line of evidence comes from another deep mutational-scanning study that compared the effects of mutations to three TIM-barrel homologs related by 30-40% amino-acid identity [26]. This study measured the effects of all amino-acid mutations on protein function across 80 positions in the protein. Despite the high sequence divergence, the effects of mutations were correlated between homologs at many sites. Overall, these experiments suggest that although epistasis shifts effects of mutations over time, these effects remain at least somewhat conserved for many sites. This result may have a structural basis. Protein structures can be well conserved over evolutionary time, even for highly diverged proteins [31]. Since the effects of mutations are highly dependent on the specific structural context in which they occur [32, 138], it might be predicted that these effects would also be conserved. Of note, Env's structure is largely conserved between sequences from different subtypes [160].

In summary, due to these different lines of evidence and the small number of experiments that have addressed this issue, the extent to which the effects of mutations shift during long-term protein evolution is still somewhat unclear. It would be interesting to quantify these shifts for Env since it would provide insight into this broader question. The results would also provide insight into the narrower question of how Env's adaptive po-

tential has changed among highly divergent lineages, and whether effects of mutations measured in the lab in one strain tend to be generalizable to other strains.

Recently developed high-throughput experiments make it possible to quantify the effects of all single amino-acid mutations to a protein of interest. A recently developed technique called deep mutational scanning can be used to measure the effects of thousands of mutations to a protein in a single high-throughput experiment [53, 54]. The basic technique, which is schematized in Fig 1 for Env, is as follows. The first step is to make a library of a gene of interest with random mutations. The next step is to perform a bulk selection for functional variants in the library. Deep sequencing of the libraries before and after selection can then be used to quantify the enrichment or depletion of each mutation. Finally, the deep sequencing data is used to infer the effect of each mutation based on its change in frequency upon selection. The results provide an unprecedented view into how changes in a protein's sequence give rise to changes in its function.

Deep mutational scanning has provided insight into a diverse number of biological questions. Below are a handful of examples illustrating this. As applied to BRCA1, it helped unveil allosteric regulatory mechanisms of this enzyme [157], and to evaluate potential disease-causing mutations among a set of clinically relevant variants with unknown significance [158]. In another study, it was used to create a detailed map of the evolutionary landscape of a pair of co-evolving toxin-antitoxin proteins, suggesting that co-evolution of these proteins proceeds via promiscuous intermediates [1]. This technique has also been used to efficiently engineer enzymes with desirable properties, such as increased thermal stability [141]. A related application is optimization of *de novo* computationally designed proteins, as was used to dramatically improve binding of an anti-influenza protein design [176].

Relevant to my thesis, this technique has also been used to gain insight into viral evolution. Even more relevant to my thesis, many of these studies have analyzed the evolutionary role of underlying functional constraints. For instance, deep mutational scanning

has been used to measure the effects of nearly all single-nucleotide [179] or single amino-acid [164] mutations to influenza's hemagglutinin protein, providing an in-depth profile of this protein's mutational tolerance. As I described above, the latter study found that commonly targeted antigenic sites in this protein tend to have an above-average tolerance for mutations, which may help explain their propensity for evading antibodies [164]. A variation of this technique has also be used to measure the effects of thousands of single-nucleotide mutations scattered across the entire HIV genome [4]. This study was less comprehensive than the above ones in that it only assayed a fraction of all possible single mutations, but it was more comprehensive in the sense that it probed HIV's mutational tolerance at a genome-wide level. The resulting data was used to analyze which mutations were tolerated in the binding pocket of an anti-viral drug targeting HIV's capsid protein, which could be used to redesign inhibitors that specifically tolerate sites with the lowest tolerance for mutations. Another study used deep mutational scanning to investigate the dual constraint imposed by the overlapping coding sequences of HIV's *tat* and *rev* genes, which are important for transcription of the viral genome and export of viral RNAs from the nucleus, respectively [50]. Using artificial constructs that relieve this dual constraint, this study measured the effects of mutations on the functions of *tat* and *rev* independently of one another. Their findings indicate that highly constrained sites in one gene tend not to overlap with highly constrained sites in the other gene, resulting in a division of roles that may have been selected for evolutionarily. This study lays the groundwork for investigating the mutational tolerance at overlapping genes in other viruses [11], which could help elucidate whether the pattern of functional segregation observed for *tat* and *rev* is common. My graduate research builds on this body of work by being the first study to use deep mutational scanning to comprehensively measure the effects of *all* single amino-acid mutations to an HIV protein in context of viral replication. As described in the above two sections, the results greatly expand our knowledge of how underlying functional constraints shape Env's evolution.

Other studies have used deep mutational scanning to investigate the effects of mu-

tations on immune escape. This approach typically involves passaging of mutant viral libraries in cell culture in the presence and absence of immune selection, where mutations enriched in the presence of immune selection are inferred to confer immune escape. For instance, this approach has been used to comprehensively identify all single amino-acid mutations that allow influenza hemagglutinin to escape one of several different antibodies [43]. Similarly, I have collaborated with another graduate student in the Bloom lab to comprehensively identify mutations that allow Env to evade a broadly neutralizing antibody [40]. In both cases, these studies helped to precisely delineate the boundaries and evolutionary potential of these epitopes, and, in the case of Env, suggested a clear biochemical mechanism of antibody escape for some sites. This approach has also been used to study interactions with innate immunity. For instance, a recent study measured the effects of all amino-acid mutations on the ability of influenza's nucleoprotein to escape MxA [9]. The unbiased and comprehensive nature of the approach led to the identification of previously unknown sites that influence MxA sensitivity. In a fascinating conceptual inversion, this technique has also been used to find mutations that *increase* antigen affinity to a target antibody, which was applied to Env immunogens to increase their affinity to germ-line precursors of broadly neutralizing antibodies [159, 75]. Overall, the above studies lay a solid groundwork for using deep mutational scanning to study viral evolution. In the future, this technique could be extended to study a wide variety of viruses and interacting host proteins.

Layout of dissertation. In Chapter 2, I use deep mutational scanning to measure the effects of all single amino-acid changes to Env in context of viral replication in cell culture. Since I conducted this experiment in the absence of external immune selection, the results reveal the underlying ability of each site to tolerate mutations when Env is just being selected for its ability to fold and perform its basic functions. I used these data to test the above hypotheses relating to the mutational tolerance of both the variable loops and epitopes of broadly neutralizing antibodies. I did not find statistical support for the hy-

pothesis that Env's variable loops have an especially high mutational tolerance, indicating that their variability may primarily arise from high levels of diversifying selection. In contrast, I did find support for the hypothesis that epitopes of broadly neutralizing antibodies targeting the CD4 binding site indeed have an especially low mutational tolerance. This second finding rigorously validates a long-standing assumption in the field, and further motivates efforts to target this region, as it may be less evolutionarily pliable, and thus more susceptible to medical interventions.

In Chapter 3, I repeated the deep mutational-scanning experiment on two homologs of Env that are 85% identical at the amino-acid level. The resulting data allowed me to compare the effects of mutations on viral replication in cell culture between these two Envs. I found that most mutations shifted in effect by a small-to-intermediate amount. In contrast, I only identified a handful of mutations with very large shifts on the order of differences expected between non-homologous sites. Thus, although epistasis has led to shifts in Env's mutational tolerance, few sites have been completely remodeled in terms of which amino acids they tolerate.

Chapter 2

EXPERIMENTAL ESTIMATION OF THE EFFECTS OF ALL AMINO-ACID MUTATIONS TO HIV'S ENVELOPE PROTEIN ON VIRAL REPLICATION IN CELL CULTURE

A version of this chapter has been previously published as:

Hugh K Haddox, Adam S Dingens, and Jesse D. Bloom. Experimental estimation of the effects of all amino-acid mutations to HIV's envelope protein on viral replication in cell culture. *PLoS Pathogens*, 12(12): e1006114.

2.1 Abstract

HIV is notorious for its capacity to evade immunity and anti-viral drugs through rapid sequence evolution. Knowledge of the functional effects of mutations to HIV is critical for understanding this evolution. HIV's most rapidly evolving protein is its envelope (Env). Here we use deep mutational scanning to experimentally estimate the effects of all amino-acid mutations to Env on viral replication in cell culture. Most mutations are under purifying selection in our experiments, although a few sites experience strong selection for mutations that enhance HIV's replication in cell culture. We compare our experimental measurements of each site's preference for each amino acid to the actual frequencies of these amino acids in naturally occurring HIV sequences. Our measured amino-acid preferences correlate with amino-acid frequencies in natural sequences for most sites. However, our measured preferences are less concordant with natural amino-acid frequencies at surface-exposed sites that are subject to pressures absent from our experiments such as antibody selection. Our data enable us to quantify the inherent mutational tolerance of each site in Env. We show that the epitopes of broadly neutralizing antibodies have a significantly reduced inherent capacity to tolerate mutations, rigorously validating a pervasive idea in the field. Overall, our results help disentangle the role of inherent functional constraints and external selection pressures in shaping Env's evolution.

2.2 Introduction

HIV evolves rapidly: the envelope (Env) proteins of two viral strains within a single infected host diverge as much in a year as the typical human and chimpanzee ortholog has diverged over $\sim$5-million years [184, 30, 109, 91]. This rapid evolution is central to HIV's biology. Most humans infected with HIV generate antibodies against Env that effectively neutralize viruses from early in the infection [6, 174, 139]. However, Env evolves so rapidly that HIV is able to stay ahead of this antibody response, with new viral variants escaping from antibodies that neutralized their predecessors just months before [6, 174, 139]. Env's

exceptional evolutionary capacity is therefore essential for the maintenance of HIV in the human population.

A protein's evolutionary capacity depends on its ability to tolerate point mutations. Detailed knowledge of how mutations affect Env is therefore key to understanding its evolution. Many studies have estimated the effects of mutations to Env. One strategy is experimental: numerous studies have used site-directed mutagenesis or alanine scanning to measure how specific mutations affect various aspects of Env's function [119, 34, 10, 56, 98, 74, 192, 125, 97, 100]. However, these experiments have examined only a small fraction of the many possible mutations to Env. Another strategy is computational: under certain assumptions, the fitness effects of mutations can be estimated from their frequencies in global or intra-patient HIV sequences [38, 48**?** , 185, 67]. However, these computational strategies are of uncertain accuracy and cannot separate the contributions of inherent functional constraints from those of external selection pressures such as antibodies. Therefore, a more complete and direct delineation of how every mutation affects Env's function would be of great value.

It is now possible to make massively parallel experimental measurements of the effects of protein mutations using deep mutational scanning [53, 54, 22]. These experiments involve creating large libraries of mutants of a gene, subjecting them to bulk functional selections, and quantifying the effect of each mutation by using deep sequencing to assess its frequency pre- and post-selection. Over the last few years, deep mutational scanning has been used to estimate the effects of *all* single amino-acid mutations to a variety of proteins or protein domains [106, 142, 51, 120, 107, 15, 133, 164, 161, 41, 82, 110, 42, 105], as well as to estimate the effects of a fraction of the amino-acid mutations to many additional proteins (e.g., [179, 158, 178]). When these experiments examine all amino-acid mutations, they can be used to compute the mutational tolerance of each protein site, thereby shedding light on a protein's inherent evolutionary capacity. Recently, deep mutational scanning has been used to examine the effects of amino-acid mutations on the binding of antibodies to Env protein displayed on mammalian or yeast cells [159, 75], or

the effects of single-nucleotide mutations scattered across the HIV genome on viral replication in cell culture [5]. However, none of these studies comprehensively measure the effects of all Env amino-acid mutations on viral replication. Therefore, we currently lack comprehensive measurements of the site-specific mutational tolerance of Env.

Here we use deep mutational scanning to experimentally estimate how all amino-acid mutations to the ectodomain and transmembrane domain of Env affect viral replication in cell culture. At most sites, our measurements correlate with the frequencies of amino acids in natural HIV sequences. However, there are large deviations at sites where natural evolution is strongly shaped by factors (e.g., antibodies) that are absent from our experiments. Our results also show that site-to-site variation in Env's inherent capacity to tolerate mutations helps explain why epitopes of broadly neutralizing antibodies are highly conserved in natural isolates. Overall, our work helps elucidate how inherent functional constraints and external selective pressures combine to shape Env's evolution, and demonstrates a powerful experimental approach for comprehensively mapping how mutations affect HIV phenotypes that can be selected for in the lab.

2.3 Results

Deep mutational scanning of Env

We used the deep mutational scanning approach in Fig 2A to estimate the effects of all single amino-acid mutations to Env. We applied this approach to Env from the LAI strain of HIV [127]. LAI is a CXCR4-tropic subtype B virus isolated from a chronically infected individual and then passaged in human T-lymphocytes. We chose this strain because LAI and the closely related HXB2 strain have been widely used to study Env's structure and function [89, 123, 119, 34, 10, 56, 168], providing extensive biochemical data with which to benchmark our results. LAI's Env is 861 amino acids in length. We mutagenized amino acids 31-702 (throughout this paper, we use the HXB2 numbering scheme [84]). We excluded the N-terminal signal peptide and the C-terminal cytoplasmic tail, since mu-

tations in these regions can alter Env expression in ways that affect viral infectivity in cell culture [25, 183, 96]. The region of Env that we mutagenized spanned 677 residues, meaning that there are $677 \times 63 = 42,651$ possible codon mutations, corresponding to $677 \times 19 = 12,863$ possible amino-acid mutations.

To create plasmid libraries containing all these mutations, we used a previously described PCR mutagenesis technique [15] that creates multi-nucleotide (e.g, $\mathtt{gca}\rightarrow\mathtt{CAT}$) as well as single-nucleotide (e.g, $\mathtt{gca}\rightarrow\mathtt{gAa}$) codon mutations. We created three independent plasmid libraries, and carried each library through all subsequent steps independently, meaning that all our measurements were made in true biological triplicate (Fig 2B). We Sanger sequenced 26 clones to estimate the frequency of mutations in the plasmid mutant libraries (Fig 3). There were an average of 1.4 codon mutations per clone, with the number of mutations per clone roughly following a Poisson distribution. The deep sequencing described in the next section found that at least 79% of the $\approx 10^4$ possible amino-acid mutations were observed at least three times in each of the triplicate libraries, and that 98% of mutations were observed at least three times across all three libraries combined. The plasmid libraries therefore sampled most amino-acid mutations to Env.

We produced virus libraries by transfecting each plasmid library into 293T cells. The viruses in the resulting transfection supernatant lack a genotype-phenotype link, since each cell is transfected by many plasmids. We therefore passaged the transfection supernatants twice in SupT1 cells at an MOI of 0.005 to create a genotype-phenotype link and select for functional variants. Importantly, neither 293T nor SupT1 cells express detectable levels of APOBEC3G [150, 137], which can hypermutate HIV genomes [70, 36]. This is a crucial point: although HIV encodes a protein that counteracts APOBEC3G, a fraction of viruses will lack a functional version of this protein and so have their genomes hypermutated in APOBEC3G-expressing cells. For each library, we passaged 5×10^5 infectious particles in order to maintain library diversity. We used Illumina deep sequencing to quantify the frequency of each mutation before and after passaging. In order to increase the sequencing accuracy, we attached unique molecular barcodes or “Primer

IDs" to each PCR amplicon [68, 73, 81, 187]. We sequenced the plasmids to assess the initial mutation frequencies, and sequenced non-integrated viral DNA [152] from infected SupT1 cells to assess the mutation frequencies in the viruses. A concern is that errors from sequencing and viral replication (e.g., from viral reverse transcriptase) would introduce bias. To address this concern, we paired each mutant library with a control in which we generated wildtype virus from unmutated plasmid. Sequencing the control plasmids and viruses enabled us to estimate and statistically correct for the rates of these errors (Fig 4). Overall, these procedures allowed us to implement the deep mutational scanning workflow in Fig 2.

Most mutations are under purifying selection, but a few sites experience selection for cell-culture adaptation mutations

Our deep mutational scanning experiments require that selection purge the virus libraries of non-functional variants. As an initial gene-wide measure of selection, we analyzed how different types of codon mutations (nonsynonymous, synonymous, and stop-codon mutations) changed in frequency after selection. In these analyses, we corrected for background errors from PCR, sequencing, and viral replication by subtracting the mutation frequencies measured in our wildtype controls from those measured in the mutant libraries (Fig 4).

Stop-codon mutations are expected to be uniformly deleterious. Indeed, after correcting for background errors, stop codons were purged to $<$1% of their initial frequency in the twice-passaged viruses for each replicate, indicating strong purifying selection (see the data for "all sites" in Fig 5A). The second viral passage is important for complete selection, as stop codons remain at about $\approx$16% of their initial frequency in viruses that were only been passaged once (Fig 6).

Interpreting the frequencies of nonsynonymous mutations is more nuanced, as different amino-acid mutations have different functional effects. However, a large fraction of

amino-acid mutations are deleterious to any protein [62, 147, 21]. Therefore, one might expect that the frequency of nonsynonymous mutations would decrease substantially in the twice-passaged mutant viruses. But surprisingly, even after correcting for background errors, the average frequency of nonsynonymous mutations in the passaged viruses is $\approx$90% of its value in the mutant plasmids (see the data for "all sites" in Fig 5A). However, the average masks two disparate trends. In each library, a few sites exhibit large increases in the frequency of nonsynonymous mutations, whereas this frequency decreases by nearly two-fold for all other sites (see the data for the subgroups of sites in Fig 5A).

An obvious hypothesis is that at a few sites, amino-acid mutations are favored because they are adaptive for viral replication in cell culture. Consistent with this hypothesis, the sites that experienced large increases in mutation frequencies are similar among the three replicates (Fig 5B), suggestive of reproducible selection for mutations at these sites. Moreover, these sites are spatially clustered in Env's crystal structure in regions where mutations are likely to enhance viral replication in cell culture (Fig 9 and Table 1). One cluster of mutations disrupts potential glycosylation sites at the trimer apex (Fig 9A). This result suggests that some of the glycans that help shield Env from antibodies in nature [77, 174] actually decrease viral fitness in the absence of immune selection. This idea is consistent with previous studies showing that that loss of glycosylation sites can enhance viral infectivity in cell culture [118, 132, 171]. A second cluster overlaps sites where mutations influence Env's conformational dynamics, which are commonly altered by cell-culture passage [112, 163]. It has been hypothesized that neutralization-resistant Envs primarily assume conformations that mask conserved antibody epitopes, while lab-adapted variants more efficiently sample different conformations associated with CD4 binding [63]. Thus, the adaptive mutations we observe may enable Env to more efficiently use CD4 in cell culture, but would not be selected in nature because they expose conserved epitopes. A third cluster is at the co-receptor binding interface (Fig 9B), where mutations may enhance viral entry in cell culture. Therefore, while most of Env is under purifying selection against

changes to the protein sequence, a few sites are under selection for cell-culture adapting amino-acid mutations.

If our experiments are indeed identifying mutations to LAI that are beneficial in cell culture, then one expectation is that some of these mutations might fix after prolonged passage of LAI in cell culture. Interestingly, almost exactly such an experiment was performed in the early study of HIV. The LAI strain used in our study was initially isolated from a chronically infected individual and then passaged in cell culture for a short period of time before cloning [169, 127]. HXB2, another common lab strain, is derived from a variant of LAI that was repeatedly passaged in a variety of cell lines, initially as a contaminant of other viral stocks [149, 27]. There are 23 amino-acid differences between the Env proteins of LAI and HXB2. Although the predecessor for HXB2 was not passaged in the same SupT1 cell line that we used, if its passage in other cell lines led to mutations that were generally adaptive to cell culture, then we would expect them to introduce amino acids in HXB2 that are also selected in our deep mutational scan of LAI. Indeed, we found that most differences between LAI and HXB2 introduced mutations to amino acids that our experiments suggest are more preferred in cell culture than the wildtype LAI amino acid (Table 2). Thus, our results are consistent with the expectation that HXB2 is more adapted to cell culture than LAI.

The average error-corrected frequency of synonymous mutations changes little after selection (an average decrease to 96% of the original frequency; see the data for "all sites" in Fig 5A). This overall trend is consistent with the fact that synonymous mutations usually have smaller functional effects than nonsynonymous mutations. However, synonymous mutations can sometimes have substantial effects [126, 35, 162, 185], particularly in viruses like HIV that are under strong selection for RNA secondary structure and codon usage [64, 172]. To assess selection on synonymous mutations on a more site-specific level, we examined the change in frequency of multi-nucleotide codon mutations across *env*'s primary sequence (Fig 10). The rationale behind examining only multi-nucleotide codon mutations is that they are not appreciably confounded by errors from PCR, deep

sequencing, or *de novo* mutations from viral replication (Fig 4, Fig 7). In a region roughly spanning codons 500 to 600, selection strongly purged both synonymous and nonsynonymous multi-nucleotide codon mutations (Fig 10). This region contains *env*'s Rev-response element (RRE) [49], a highly structured region of RNA that is bound by the Rev protein to control the temporal export of unspliced HIV transcripts from the nucleus [102, 45]. The finding of strong selection on the nucleotide as well as the amino-acid sequence of the RRE region of Env therefore agrees with our biological expectations.

The preference for each amino acid at each site in Env

The previous section examined broad trends in selection averaged across many sites. But our data also enable much more fine-grained estimates of the preference for every amino-acid at every position in Env. We define a site's preference for an amino acid to be proportional to the enrichment or depletion of that amino acid after selection (correcting for the error rates determined using the wildtype controls), normalizing the preferences for each site so that they sum to one. We denote the preference of site r for amino acid a as $\pi_{r,a}$, and compute the preferences from the deep-sequencing data as described in [17]. Since we mutagenized 677 residues in Env, there are $677 \times 20 = 13,540$ preferences. If selection in our experiments exactly parallels selection in nature and there are no shifts in mutational effects as Env evolves, then these preferences are the expected frequencies of each amino acid at each site in an alignment of Env sequences that have reached evolutionary equilibrium under a mutation process that introduces each amino acid with equal probability [15, 16].

Fig 11 shows Env's site-specific amino-acid preferences after averaging across replicates and re-scaling to account for the stringency of selection in our experiments (details of this re-scaling are in the next section). As is immediately obvious from Fig 11, sites vary dramatically in their tolerance for mutations. Some sites strongly prefer a single amino acid, while other sites can tolerate many amino acids. For instance, site 457, an

important receptor-binding residue [119], has a strong preference for aspartic acid. However, this site is adjacent to a variable loop (sites 460-469) where most sites tolerate many amino acids. Another general observation is that when sites tolerate multiple amino acids, they often prefer ones with similar chemical properties. For instance, sites 225 and 226 prefer hydrophobic amino acids, while sites 162 to 164 prefer positively charged amino acids.

To confirm that our experiments captured known constraints on Env's function, we examined mutations that have been characterized to affect key functions of Env. Table 3 lists mutations known to disrupt an essential disulfide bond, binding to receptor or co-receptor, or protease cleavage. In almost all cases, the deleterious mutation introduces an amino-acid that our experiments report as having a markedly lower preference than the wildtype amino acid. Therefore, our measurements largely concord with existing knowledge about mutations that affect key aspects of Env's function.

A crucial aspect of any high-throughput experiment is assessing the reproducibility of independent replicates. Fig 11 shows the *average* of the preferences measured in each replicate. Fig 12A shows the correlations among the 13,540 site-specific amino-acid preferences estimated from each of the three replicates. The correlations are modest, indicating substantial replicate-to-replicate noise. In principle, this noise could arise from differences in the initial plasmid mutant libraries, bottlenecks during the generation of viruses by transfection, bottlenecks during viral passaging, or bottlenecks during the sequencing of proviral DNA from infected cells. Analysis of technical replicates of the first or second round of viral passaging indicates that most of the noise arises from bottlenecks during the viral passaging or sequencing steps. Specifically, measurements from replicate 3 are no more correlated to those from replicates 3b-1 or 3b-2 (which are repeated passages of the same transfection supernatant, Fig 2B) than they are to those from totally independent replicates (compare Fig 12 and Fig 13). However, replicates 3b-1 and 3b-2 (which shared the first of the two viral passages, Fig 2) do yield more correlated measurements than independent replicates (Fig 13). The existence of bottlenecks during viral

passage is also suggested by the data in Fig 7 and Fig 8. Therefore, the experimental reproducibility could probably be increased by passaging more infectious viruses at each step.

If bottlenecks cause each replicate to sample slightly different mutations, then perhaps the total number of tolerated mutations per site will be similar between replicates, even if the exact mutations differ. To test this hypothesis, we computed the effective number of amino acids tolerated at each site as the exponential of the Shannon entropy of the site's amino-acid preferences. Fig 12B shows that the effective number of amino acids tolerated at each site is more correlated between replicates than the preferences themselves. We further reasoned that even if bottlenecking causes slight variations in the preferred amino acids between replicates, each site would still tend to prefer amino acids with similar chemical characteristics. To test this hypothesis, we quantified the extent that each site preferred hydrophobic or hydrophilic amino acids by computing a site-specific hydrophobicity score from the amino-acid preferences. Fig 12C shows that these preference-weighted hydrophobicities are more correlated between replicates than the preferences. Therefore, even though there is replicate-to-replicate noise in the exact amino acids preferred at a site, the effective number of tolerated amino acids and the chemical properties of these amino acids are similar among replicates.

The amino-acid preferences correlate with amino-acid frequencies in HIV sequence alignments at most sites, but deviate at positions subject to selection pressures absent from our experiments

In the previous section, we showed that our experimentally measured amino-acid preferences captured the constraints on Env's biological function for sites with known mutational effects (Table 3). If this is true across the entire protein, then our measurements should correlate with the frequencies of amino acids in natural HIV sequences. Table 4 shows that there is a modest correlation (Pearson's R ranging from 0.29 to 0.36) between the

preferences from each experimental replicate and the frequencies in an alignment of HIV-1 group-M sequences (a phylogenetic tree of these sequences is in Fig 14A; sites in Env variable loops that can not be reliably aligned are excluded as described in the Methods). Since each replicate suffers from noise due to partial bottlenecking of the viral diversity, we hypothesized that averaging the preferences across replicates should make them more accurate. Indeed, averaging the replicates increased the correlation to $R = 0.4$ (Table 4).

The concordance between deep mutational scanning measurements and natural sequence variation is improved by accounting for differences in the stringency of selection in the experiments compared to natural selection [16, 18]. Specifically, if the measured preference is $\pi_{r,a}$ and the stringency parameter is β, then the re-scaled preference is $(\pi_{r,a})^{\beta} \, / \left[\sum_{a'} (\pi_{r,a'})^{\beta}\right]$. A stringency parameter of $\beta > 1$ means that natural evolution favors the same amino acids as the experiments, but with greater stringency. Table 4 shows that for all replicates, the stringency parameter that maximizes the correlation is > 1. Therefore, natural selection prefers the same amino acids as our experiments, but with greater stringency.

After averaging across replicates and re-scaling by the optimal stringency parameter, the Pearson correlation is 0.44 between our experimentally measured preferences and amino-acid frequencies in the alignment of naturally occurring HIV sequences (Fig 14B). Is this a good correlation? At first glance, a correlation of 0.44 seems unimpressive. But we do not expect a perfect correlation even if the experiments perfectly concord with selection on Env in nature. There are several factors that are expected to reduce the correlation between the experimentally measured preferences and amino-acid frequencies in natural sequences. First, our experiments examine the effects of mutations to Env from the LAI strain. However, it is well known that epistasis can cause the effects of mutations to differ among homologs of the same protein [175, 121], and many examples of this phenomenon have been documented in HIV Env [55, 170, 37, 57]. Therefore, our measurements for the LAI Env are probably not completely generalizable to all other strains. In addition, natural HIV sequences are drawn from a phylogeny (Fig 14A), not an ideal ensemble

of all possible Env sequences. The frequencies of amino acids in this phylogeny reflect evolutionary history as well as natural selection. For instance, if several amino acids are equally preferred at a site, one is likely to be more frequent in the alignment due to historical contingency. Additionally, natural evolution is influenced by the genetic code and mutation biases: a mutation from the tryptophan codon `TGG` to the valine codon `GTT` is extremely unlikely even if valine is more preferred than tryptophan. Mutation biases inherent in reverse transcription [3] or APOBEC3G-induced hypermutation [150] could also bias some evolutionary outcomes over others. Therefore, the correlation will be imperfect even if the preferences completely concord with natural selection – the question is how the actual correlation compares to what is expected given the phylogenetic history and mutation biases.

To determine the expected correlation if the experimentally measured amino-acid preferences reflect conserved constraints in Env, we simulated evolution along the phylogenetic tree in Fig 14A under the assumption that the experimentally measured preferences exactly match natural selection. Specifically, we used `pyvolve` [154] to simulate evolution using the experimentally informed site-specific codon substitution models described in [18], which define mutation-fixation probabilities in terms of the amino-acid preferences. In addition to the preferences and the stringency parameter $\beta = 2.1$ from Table 4, the substitution models in [18] require specification of parameters reflecting biases in the mutation process. We estimated nucleotide mutation bias parameters of $\phi_A = 0.55$, $\phi_C = 0.15$, $\phi_G = 0.11$, and $\phi_T = 0.18$ from the frequencies at the third-nucleotide codon position in sequences in the group-M alignment for sites where the most common amino acid had 4-fold codon degeneracy. We used the transition-transversion ratio of $\kappa = 4.4$ estimated in [117]. For these simulations, we scaled the branch lengths so that the average pairwise protein divergence was the same in the actual and simulated alignments.

The correlation between the preferences and amino-acid frequencies in a representative simulated alignment is shown in Fig 14C. As this plot illustrates, the expected correlation is only about 0.46 if the experimentally measured preferences exactly describe

natural selection on Env under our model. The simulated frequencies in Fig 14C show the same pattern of bi-modality (most values near zero or one) as the actual frequencies in Fig 14B despite the fact that the preferences used in the simulations allow multiple amino acids at most sites (see Fig 11). This fact illustrates that bi-modality in the amino-acid frequencies can arise from the historical contingency inherent in a phylogenetic tree even if multiple amino acids are tolerated at most sites. As a control, we also simulated evolution using substitution models in which the preferences have been randomized among sites (Fig 14D); as should be the case, there is no correlation in these control simulations. So the actual correlation is nearly as high as expected if natural selection concords with the preferences measured in our experiment.

We next investigated if there are parts of Env for which there is an especially low correlation between our experimentally measured preferences and natural amino-acid frequencies. For instance, antibodies exert selection on the surface of Env in nature [174, 139, 148, 122]. We therefore examined the actual and simulated correlations between the preferences and frequencies as a function of solvent accessibility (Fig 14E,F). For all sites (right side of Fig 14E, left side of Fig 14F), the actual correlation is only slightly lower than the range of correlations in 100 simulations. For more buried sites, both the simulated and actual correlations increase (Fig 14E), presumably because sites in the core of Env tend to have stronger preferences for specific amino acids. But as sites become more surface-exposed, the actual correlation drops below the value expected from the simulations (Fig 14F). Therefore, our experiments provide a relatively worse description of natural selection on Env's surface than its core – probably because the evolution of the protein's core is shaped mostly by inherent functional constraints that are effectively captured by our experiments, whereas the surface is subject to selection pressures (e.g., antibodies) that are not modeled in our experiments.

Comparing disulfide-bonded cysteines and glycosylation sites vividly illustrates this dichotomy between inherent functional constraints and external selection pressures. Env has 10 highly conserved disulfide bonds, most of which are essential for the protein's

inherent function [168]. Env also has numerous N-linked glycosylation sites, many of which are also highly conserved in nature, where they help shield the protein from antibodies [77, 174]. In contrast to the disulfides, only some glycosylation sites are important for Env's function in the absence of immune selection [118, 171]. Fig 15 shows that our experimentally measured preferences are highly correlated with natural amino-acid frequencies at the sites of the disulfides, but not at the glycosylation sites. This result can easily be rationalized: the disulfides are inherently necessary for Env's function, whereas many glycosylation sites are important largely because of the external selection imposed by antibodies. Our experiments therefore accurately reflect the natural constraints on the former but not the latter.

The fact that we found well-tolerated mutations at all of Env's glycosylation sites (Fig 16A) might seem surprising given that other studies have shown that some glycosylation sites are important for Env's function in certain HIV strains [118, 171]. However, these studies were all performed in HIV strains substantially diverged from LAI. A study in HXB2 (which is closely related to LAI) found that individual mutations are at least partially tolerated at all glycosylation sites in Env's gp120 subunit when assaying for viral infectivity in cell culture [93]. Therefore, glycosylation sites may be especially expendable in the LAI strain used in our study.

Env has a low mutational tolerance in broadly neutralizing antibody epitopes

Different sites in Env evolve at different rates in natural HIV sequences. For instance, sites on the apical surface of Env evolve especially rapidly [108]. These differences in evolutionary rate arise from two factors. First, some sites are inherently better at tolerating mutations without disrupting Env's essential functions. Second, some sites are under stronger immune selection for rapid sequence change. However, since Env in nature is under selection both to maintain its function and escape immunity, it is difficult to deconvolve these factors.

Our experiments estimate each site's inherent tolerance for mutations under selection purely for Env's function in cell culture, without the confounding effects of immune selection (for the remainder of this section, we define a site's mutational tolerance as the Shannon entropy of its amino-acid preferences shown in Fig 11). We can therefore assess whether regions of Env that evolve rapidly or slowly in nature also have unusually high or low inherent tolerance to mutations.

We focused on two regions of Env. First, we analyzed portions of the protein classified as "variable loops" due to extensive variation in nature [156, 111]. These loops are frequently targeted by antibodies that drive rapid sequence evolution [114, 122]. Because these loops evolve rapidly, we hypothesized they would have a high inherent mutational tolerance. But an alternative hypothesis is that their rapid evolution more attributable to strong selection from antibodies than an unusually high mutational tolerance. Second, we focused on epitopes of antibodies that broadly neutralize many HIV strains. Because these epitopes are highly conserved in nature and often overlap with regions of known functional constraint [153, 193, 14, 46, 72, 146], we hypothesized they would have a low mutational tolerance. However, an alternative hypothesis is that these epitopes evolve slowly not because they are mutationally intolerant but simply because they are under weaker immune selection. Indeed, broad immune responses targeting these epitopes only develop in 20% of infected individuals and generally only after multiple years of infection [88].

In testing these hypotheses, it is important to control for other properties known to affect mutational tolerance. This can be done by using multiple linear regression to simultaneously analyze how several independent variables affect the dependent variable of mutational tolerance. Relative solvent accessibility (RSA) is the strongest determinant of mutational tolerance in proteins [135], so we included RSA as a variable in the regression. The region of *env* that contains the RRE is under strong nucleotide-level constraint [49, 102, 45, Fig 10], so we also included being in the RRE as a binary variable in the regression. We defined the variable loops as indicated in Fig 11, and included be-

ing in one of these loops as a binary variable in the regression. Finally, we used crystal structures to delineate broadly neutralizing antibody epitopes. We focused on broadly neutralizing antibodies targeting the CD4 binding site, since most other broadly neutralizing antibodies target either glycans (which are subject to pressures that are not well-modeled in our experiments; Fig 15A) or a membrane-proximal region of gp41 that is not fully resolved in crystal structures of trimeric Env making it impossible to correct for RSA. Specifically, we analyzed the three antibodies with the greatest breadth from [189]: VRC01 (PDB 3NGB [188]), 12A21 (PDB 4JPW [83]), and 3BNC117 (PDB 4JPV [83]). We defined a site as part of an epitope if it was within a $4\mathring{A}$ inter-atomic distance of the antibody, and included the number of epitopes in which a site is found as a discrete variable in the regression.

The results of the multiple linear regression are in Table 5. As expected, increased solvent accessibility is strongly associated with increased mutational tolerance, whereas presence in the RRE is strongly associated with decreased mutational tolerance. After correcting for these effects, sites in broadly neutralizing epitopes have significantly reduced mutational tolerance. In contrast, sites in the variable loops have higher mutational tolerance, but this effect is not statistically significant. Some of the loops are more variable in nature than others [191]. However, even when the loops are considered independently, none of these regions has a statistically significant association with mutational tolerance (Table 6). Overall, this analysis provides statistical confirmation of something that is widely assumed in the study of HIV: broadly neutralizing antibodies are unique because they target regions of Env that are inherently intolerant of mutations. However, we fail to find strong statistical support for the hypothesis that variable loops are especially tolerant of mutations. Thus, the rapid evolution of these loops in nature is probably more attributable to strong immune selection than exceptionally high inherent mutational tolerance.

2.4 Discussion

We have used deep mutational scanning to experimentally estimate the effects of all amino-acid mutations to most of HIV Env. Our experiments select for Env variants that enable HIV to undergo multi-cycle replication in a T-cell line. The broad trends in our data are consistent with what is expected from general considerations of how gene sequence maps to protein function: stop codons are efficiently purged by selection, many but not all nonsynonymous mutations are selected against, and synonymous mutations are less affected by selection except at regions where the nucleotide sequence itself is known to be biologically important. We also find a few sites where nonsynonymous mutations are strongly favored by selection in our experiments, probably because they adapt the virus to cell culture by affecting Env's conformational dynamics, co-receptor binding, and glycosylation.

We use our experimental data to estimate the preference of each site in Env for each amino acid. We show that these preferences correlate with amino-acid frequencies in natural HIV sequences nearly as well as would be expected if the experimentally measured preferences capture the true selection on Env in nature. The strongest deviations between our measurements and amino-acid frequencies in HIV sequences occur at sites on the surface of the virus that in nature are targeted by pressures (such as antibodies) that are not present in our experiments.

The ability to identify deviations between our measurements and amino-acid frequencies in nature points to a powerful aspect of our approach: it can de-convolve the role of inherent functional constraints and external selection pressures in shaping Env's evolution. For instance, it is known that some regions of Env are conserved in nature and thus are susceptible to broadly neutralizing antibodies. But other regions of Env such as the variable loops exhibit extensive variability and are generally targeted by more strain-specific antibodies. To what extent are these patterns of conservation shaped by Env's inherent capacity to evolve versus the fact that immune selection tends to target the vari-

able loops more readily than the broadly neutralizing antibody epitopes? By measuring Env's mutational tolerance at each site under functional selection alone, we show that the epitopes of broadly neutralizing antibodies indeed have a reduced capacity to tolerate mutations irrespective of the action of immune selection. However, we do not find strong statistical support for the hypothesis that the variable loops are especially tolerant of mutations compared to the rest of the protein. Thus, the rapid evolution of these loops probably results more from strong immune selection than exceptionally high inherent mutational tolerance. In the future, our measurements could also be used to examine the role of Env's mutational tolerance in shaping the evolution of epitopes targeted by cellular immunity [182].

More generally, our experiments provide high-throughput experimental data that can augment computational efforts to infer features of HIV's fitness landscape [38, 48**?** , 87, 67]. Such data will aid in efforts to understand viral evolutionary dynamics both within and between patients. Our study examined the replication of the CXCR4-tropic LAI strain isolated from a chronically infected individual, and used a T-cell line that expresses high levels of receptor relative to many primary cells [92, 79]. This experimental setting is obviously a simplified representation of the actual environment in which HIV replicates. However, we anticipate that our approach could be extended to examine the effects of Env mutations in more complex experimental settings that may better mimic the selection on viruses in humans. For instance, comparing our measurements to those made on transmitted-founder viruses should help elucidate how selective constraints differ among HIV strains. Examining viral replication in cells with different receptor and co-receptor distributions should make it possible to isolate the role of cell-type specific selection in shaping HIV evolution [95, 12]. Adding factors such as antibodies should enable the comprehensive identification of how mutations affect HIV immune escape. Such experiments will augment the results described here with maps of how mutational effects shift under various biologically relevant scenarios, thereby further enhancing our ability to understand the internal and external forces driving HIV evolution.

2.5 Methods

Data and computer code

The computer code to analyze the sequencing data and generate the figures is provided in a series of IPython notebooks in S3 File. Illumina sequencing data are available from the Sequence Read Archive (`http://www.ncbi.nlm.nih.gov/sra`) under the accession numbers in S10 File.

Sequence numbering

We use the HXB2 numbering system [84] unless otherwise noted. The "variable loop" definitions were taken from `http://www.hiv.lanl.gov/`, not including the flanking disulfide-bonded cysteines as part of the loops.

Codon mutant libraries

We created the codon mutant libraries in the context of the pro-viral genomic plasmid pLAI, which encodes the LAI strain of HIV [127]. This plasmid was obtained from the lab of Michael Emerman. The plasmid sequence is in S4 File.

We created codon mutant libraries of *env* using the PCR mutagenesis technique described in [15] (see also [164, 42]) except that we performed two total rounds of mutagenesis rather than the three rounds in [15]. The codon tiling mutagenic primers are in S5 File. The end primers were: 5'-ttggaatttctggcccagaccgtctcatgagagtgaaggagaaatatcagcacttg-3' and 5'-catctgctgctggctcagc-3'. We created three replicate libraries by performing all the steps independently for each replicate starting with independent plasmid preps.

We cloned the PCR mutagenized *env* amplicons into the LAI plasmid with high efficiency to create plasmid mutant libraries. To seamlessly clone the PCR products into the proviral plasmid, we created a recipient version of the plasmid that had *env* replaced by GFP flanked by restriction sites for BsmBI, which cleaves outside its recognition se-

quence. We named this recipient plasmid pLAI-δ*env*-BsmBI; its sequence is in S6 File. We digested both this recipient plasmid and the gel-purified PCR amplicons with BsmBI (there are BsmBI sites at either end of the PCR amplicon), gel purified the digested PCR products, and ligated them into the plasmid using a T4 DNA ligase. We column purified the ligation products, electroporated them into competent cells (Invitrogen,12033-015), and plated the transformed cells on LB plates supplemented with 100 μg/mL ampicillin. For each of the three replicate libraries, we performed enough transformations to yield $>$1.4 million unique colonies as estimated by plating dilutions of each transformation on separate plates. Control ligations lacking an insert yielded at least 10-fold fewer colonies. The transformed cells were scraped from the plates, grown in liquid LB-ampicillin at 37°C for $\sim$4 hours, and mini-prepped to obtain the plasmid mutant libraries. For the wildtype controls, we prepped three independent cultures of the wildtype LAI proviral plasmid.

Generation and passaging of viruses

We generated the mutant virus libraries by transfecting the mutant plasmid libraries into 293T cells obtained from the American Type Culture Collection (ATCC). For each replicate, we transfected two 12-well tissue-culture plates to increase the diversity of the generated viruses. Specifically, we plated 293T cells at 2.4×10^5 cells/well in D10 media (DMEM supplemented with 10% FBS, 1% 200 mM L-glutamine, and 1% of a solution of 10,000 units/mL penicillin and 10,000 μg/mL streptomycin). The next day, we transfected each well with 1 μg plasmid using BioT (Bioland Scientific LLC, B01-01). For the three wildtype controls we used the same process but with only a single 12-well plate per replicate. At one day post-transfection, we aspirated the old media, replacing it with fresh D10. At $\sim$60 hours post-transfection, we filtered the transfection supernatants through 0.4 μm filters. To remove residual plasmid DNA from the transfection, we then treated the filtrate with DNase-I (Roche, 4716728001) at a final concentration of 100 U/mL in the presence of 10 mM magnesium chloride (Sigma, M8266) at 37°C for 20-30 minutes. We froze aliquots of

the DNase-treated supernatant at -80°C. Aliquots were thawed and titered by TZM-bl and TCID-50 assays as described below.

We passaged the transfection supernatants in SupT1 cells obtained from the NIH AIDS Reagent Program [2]. SupT1 cells were maintained in a media identical to the D10 described above except that the DMEM was replaced with RPMI-1640 (GE Healthcare Life Sciences, SH30255.01). Before infecting cells, for replicates 1, 2, and 3 (but not replicate 3b), we first filtered thawed transfection supernatants through a 0.2 μm filter in an effort to remove any large viral aggregates. We then infected 10^8 SupT1 cells with 5×10^5 TZM-bl units of the mutant library transfection supernatant in a final volume of 100 mL SupT1 culture medium in a vented tissue-culture flask (Fisher Scientific, 14-826-80). In parallel, we passaged 10^5 TZM-bl units of transfection supernatant for each wildtype control in 20 million SupT1 cells in a final volume of 20 mL. At one day post-infection, we pelleted cells at 300$\times$g for 4 minutes and resuspended in fresh media to the same volume as before. At two days post-infection, we added fresh media equal to the volume already in the flask to dilute the cells and provide fresh media. We harvested virus at three days post-infection (for replicates 1, 2, and 3) or four days post-infection (for replicate 3b) by pelleting cell debri at 300$\times$g for 4 minutes and then collecting the viral supernatant for storage at -80°C. To remove residual culture media and plasmid DNA from the cell pellets, we washed pellets two times in PBS. The washed cells were resuspended in PBS to a final concentration of 10^7 cells/mL, and aliquots were frozen at -80°C for DNA purification.

We conducted a second passage by infecting new cells with the passage-1 viral supernatants. The second passage differed from the first passage in the following ways: Before infecting cells, we filtered passage-1 supernatant of replicate 3b-2 through a 0.2 μm filter but did not filter any of the other replicates. We also had to modify the passaging conditions for some replicates due to low titers of the passage-1 supernatants. For viruses in which the passage-1 supernatant was at too low a concentration to infect at an MOI of 0.005 in the volumes indicated above, we added additional passage-1 supernatant, and then reduced the volume to that indicated above during the day-one media change. As

stated in the Results section, passaging more than 5×10^5 TZM-bl units of the mutant library at each step would probably help increase reproducibility between experimental replicates.

Virus titering by TCID$_{50}$ and TZM-bl assays

We measured viral titers using TZM-bl reporter cells obtained from the NIH AIDS Reagent Program [173]. Specifically, we added 2×10^4 cells in 0.5 mL D10 to each well of a 12-well plate. We made dilutions of viral inoculum and infected cells with 100 uL of each dilution. At 2 days post-infection, we fixed cells in a solution of 1% formaldehyde and 0.2% glutaraldehyde in PBS for 5 minutes at room temperature, washed with PBS to remove the fixing solution, and stained for beta-galactosidase activity with a solution of 4 mM potassium ferrocyanide, 4 mM potassium ferricyanide, and 0.4 mg/mL X-gal in PBS at 37°C for 50 minutes. After washing cells with PBS to remove the staining solution, we used a microscope to count the number of blue cells per well, computing the viral titer as the number of blue cells per mL of viral inoculum.

We were concerned that the infectious titer in SupT1 cells might differ from the TZM-bl titers. We therefore also performed TCID$_{50}$ assay to directly measure infectious titers in SupT1 cells. To do this, we made dilutions of viral transfection supernatant in a 96-well tissue-culture plate and added SupT1 cells at a final concentration of 2.5×10^5 cells/mL in a final volume of 180 μL/well. At 4 and 8 days post-infection, we passaged supernatant 1:10 into fresh media to prevent cells from becoming over confluent. At 12 days post-infection, we measured the titer of culture supernatants using the TZM-bl assay to determine which SupT1 infections had led to the production of virus. Based on binary scoring from these TZM-bl assays, we calculated titers using the Reed-Muench formula [136] as implemented at `https://github.com/jbloomlab/reedmuenchcalculator`. At least for the LAI strain used in our experiments, the SupT1 TCID$_{50}$ titers were approximately equal to the TZM-bl titers. Therefore, we used only the less time-consuming TZM-bl assay for all subsequent

titering.

Generation of samples for Illumina sequencing

We purified non-integrated viral DNA from aliquots of frozen SupT1 cells using a mini-prep kit (Qiagen, 27104) with $\sim 10^7$ cells per prep. In some cases, we then concentrated the purified DNA using Agencourt AMPure XP beads (Beckman Coulter, A63880) using a bead-to-sample ratio of 1.0 and eluting with half of the starting sample volume.

We next generated PCR amplicons of *env* to use as templates for Illumina sequencing. We created these amplicons from plasmid or mini-prepped non-integrated viral DNA by PCR using the primers 5'-agcgacgaagacctcctcaag-3' and 5'-acagcactattctttagttcctgactcc-3'. PCRs were performed in 20 μl or 50 μl volumes using KOD Hot Start Master Mix (71842, EMD Millipore) with 0.3 μM of each primer and 3 ng/μl of mini-prepped DNA or 0.3 ng/μl of plasmid as template. The PCR program was:

1. 95 °C, 2 minutes
2. 95 °C, 20 seconds
3. 70 °C, 1 second
4. 64.3 °C, 10 seconds (cooling to this temperature at 0.5 °C/second)
5. 70 °C, 1 minute 48 seconds
6. Go to 2, 27 times
7. hold at 4 °C

For replicate 3b, there were a few modifications: the annealing temperature was 64.9 °C, the extension time was 54 seconds, and we performed only 25 cycles. To quantify the number of unique template molecules amplified in each PCR, we performed standard curves using known amounts of template *env* in pro-viral plasmid, and ran the the bands on an agarose gel alongside our amplicons for visual quantification. We performed a sufficient number of PCR reactions to ensure that amplicons from plasmid were coming from $> 10^6$ unique template molecules, and amplicons from viral DNA were coming from

$\sim 2 \times 10^5$ template molecules. All PCR products were purified with Agencourt beads (using a sample-to-bead ratio of 1.0) and quantified by Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies, P7589).

We deep sequenced these amplicons using the strategy for barcoded-subamplicon sequencing in [42], dividing *env* into six subamplicons (this is a variation of the strategy originally described in [68, 73, 81]). The sequences of the primers used in the two rounds of PCR are in S9 File. Our first-round PCR conditions slightly differed from [42]: our 25 μL PCRs contained 12.5 μL KOD Hot Start Master Mix, 0.3 μM of each primer, and 5 ng of purified amplicon. For replicates 1, 2, and 3, the first-round PCR program was:

1. 95 °C, 2 minutes
2. 95 °C, 20 seconds
3. 70 °C, 1 seconds
4. 60 °C, 10 seconds (cooling to this temperature at 0.5 °C/second)
5. 70 °C, 10 seconds
6. Go to 2, 10 times
7. 95 °C, 1 min
8. hold 4 °C

For replicate 3b, we used the same program, but with 9 PCR cycles instead of 11. Prior to the second round PCR, we bottlenecked each subamplicon by diluting it to a concentration that should have yielded between 3 and 5 $\times 10^5$ unique single-stranded molecules per subamplicon per sample. We purified the second-round PCR products using Agencourt beads, quantified with PicoGreen, pooled in equimolar amounts, and purified by agarose gel electrophoresis, excising DNA corresponding to the expected $\sim$500 base pairs in length. We sequenced the purified DNA using multiple runs of an Illumina MiSeq with 2$\times$275 bp paired-end reads.

Analysis of deep-sequencing data

We used dms_tools (`http://jbloomlab.github.io/dms_tools/`), version 1.1.dev13, to filter and align the deep-sequencing reads, count the number of times each codon mutation was observed both before and after selection, and infer Env's site-specific amino-acid preferences using the algorithm described in [17]. The code that performs this analysis is in S3 File. Figures summarizing the results of the deep sequencing are also in this supplementary file.

Alignment of group-M env sequences

We downloaded the 2014 filtered web alignment of *env* from `http://www.hiv.lanl.gov/`, including all subtypes for HIV-1/SIVcpz. We then curated this alignment in the following ways. First, we removed sequences differed in length from HXB2 (including gap characters) or contained a premature stop codon, ambiguous residue, or frame-shift mutation. Next, we removed columns in the alignment for which we lacked deep mutational scanning data, columns that had $>$5% gap characters, or columns in variable loops that appeared poorly aligned by eye. Finally, we randomly selected 30 sequences per subtype for group-M subtypes A, B, C, D, F, and G, for a total of 180 sequences. The resulting alignment is in S7 File. The phylogenetic tree in Fig 14 was inferred using RAxML [155] with the GTRCAT substitution model.

Computing relative solvent accessibilities

We computed absolute solvent accessibilities based on the PDB structure 4TVP (including all three Env monomers after removing antibody chains) using DSSP [166, 80]. We normalized absolute solvent accessibilities to relative ones using the maximum accessibilities provided in the first table of [165]. The relative solvent accessibilities are listed in S8 File.

Supplemental files

Descriptions of each file are shown below. Please see the publication for the actual files [65].

S1 File. Average of the amino-acid preferences measured in the replicates. Sites are numbered using the HXB2 scheme. The same preferences re-scaled by the optimal stringency parameter are in S2 File.

S2 File. Amino-acid preferences re-scaled by the optimal stringency parameter. The preferences in S1 File re-scaled by the optimal stringency parameter of $\beta = 2.1$. These are the data plotted in Fig. 11. However, preferences for stop codons are listed in this file, but not shown Fig 11.

S3 File. iPython notebooks that perform the data analysis steps described in this paper.

S4 File. A Genbank file with the sequence of the LAI pro-viral plasmid.

S5 File. The codon tiling primers used to construct the mutant libraries.

S6 File. The recipient pro-viral plasmid, which has *env* replaced by partial GFP and beta globin genes flanked by BsmBI sites.

S7 File. The alignment of group M Env sequences.

S8 File. The relative solvent accessibilities of all sites in Env present in the crystal structure.

S9 File. The PCR primers used in the barcoded-subamplicon sequencing.

S10 File. The SRA accession numbers for deep sequencing data. Samples are named as follows: mutDNA-1 denotes the mutant plasmid library for replicate 1; DNA-1 denotes the wildtype plasmid for replicate 1; mutvirus-p2-1 denotes the twice-passaged mutant viral libraries for replicate 1; virus-p2-1 denotes the twice-passaged wildtype virus for replicate 1.

Acknowledgments

Thanks to Michael Emerman for providing the LAI plasmid and numerous helpful suggestions, to Julie Overbaugh for numerous helpful suggestions, and to Lily Wu for HIV training.

Chapter 3

MAPPING SITES OF SHIFTING AND CONSTANT MUTATIONAL EFFECTS ON THE EVOLUTIONARY LANDSCAPE OF HIV ENVELOPE

3.1 Abstract

HIV's envelope protein (Env) evolves rapidly. The immediate evolutionary space accessible to any viral variant is largely determined by the effects of single amino-acid mutations on viral fitness. Due to epistasis, these effects can "shift" over evolutionary time as the virus traverses through sequence space (e.g., the effect of a mutation tolerated in one viral variant may shift such that it is not tolerated in another related variant). In principle, even single amino-acid mutations can substantially alter a protein's mutational tolerance. However, the prevalence of such shifts in long-term protein evolution in nature is largely unknown. Here, we comprehensively quantified the ways in which the effects of mutations to Env have shifted or remained constant among two divergent Env homologs with 85% sequence identity, corresponding to $>$100 amino-acid differences. We did so by experimentally measuring the effects of all single amino-acid mutations to each Env homolog on the ability of that homolog to support viral replication in cell culture. In total, we analyzed these effects at a total of 616 homologous sites. We observed that a small fraction of mutations had largely shifted in their effects, where a given mutation was well tolerated in one homolog, but strongly disfavored in the other. In contrast, we observed that the majority of mutations had shifted in their effects by only small-to-intermediate amounts. Thus, our data suggest that mutational effects are substantially conserved between these divergent Env homologs despite $>$100 underlying amino-acid differences. Overall, our data indicate that although epistasis can cause the effects of mutations to shift over evolutionary time, large shifts have been rare in the long-term evolutionary period separating

the Env homologs we considered.

3.2 Introduction

HIV's envelope protein (Env) rapidly evolves. In just a single HIV-infected individual, the infecting virus can give rise to a pool of Env variants that are as diverse as all globally circulating influenza strains in a given year [85]. Env's ability to rapidly evolve has dire consequences for the immune system. Most HIV-infected individuals generate anti-Env antibodies that neutralize the virus [6, 174, 139]. However, Env readily evades this response through rapid sequence evolution [6, 174, 139]. The immune system can adapt to target new viral variants through through the elicitation of new antibodies and refining the specificity of antibodies through somatic hypermutation. But, Env is able to evade new immune responses each time they arise [6, 174, 139]. Understanding this protein's ability to do so is thus central to understanding HIV's ability to evade immunity.

The immediate evolutionary space surrounding Env is largely defined by the effects of single point mutations on viral replication. To better understand this space, a large number of studies have characterized the effects of mutations on Env's ability to support viral replication in cell culture and evade antibodies [119, 34, 10, 56, 98, 74, 192, 125, 97, 100, 65]. For practical reasons, these studies often measured these effects in just a single or a few genetic backgrounds. However, effects of mutations can change over evolutionary time due to epistasis. One form of epistasis is intra-protein epistasis, where the effect of a mutation at one site in a protein is influenced by which amino acids are present at other sites in the same protein. Thus, as a protein evolves, substitutions at one site may "shift" the effects of mutations at other sites in the protein, such that a mutation that is not tolerated in one background is tolerated in a closely related one. This phenomenon has been documented for a wide variety of proteins [175, 20, 13, 121, 99, 58, 116, 66, 129], including Env [55, 170, 37, 57], and is described in greater detail in the Introduction of this thesis.

Considering the effects of mutations in different genetic backgrounds is important for a protein as diverse as Env. Since HIV's introduction into humans, several phylogenetically distinct subtypes have arisen (Fig 17). The level of sequence divergence in Env between subtypes is typically 20-35% amino-acid divergence [85]. Even within a subtype, Env variants typically differ by 15-20% amino-acid divergence [85]. Thus, given that Env is $\sim$850 amino acids in length, homologs often differ by $>$100 amino acids. In principle, even single amino-acid changes can cause large shifts in the effect of mutations at other sites in a protein [175, 121, 58]. Thus, mutational effects may substantially differ between Env homologs. However, the frequency of large shifts in long-term protein evolution is still largely unknown. Some studies suggest that such shifts may be rare [41, 8].

Here, we quantified shifts in the effects of mutations on Env's ability to support viral replication in cell culture among two divergent Env homologs that share 85% amino-acid identity. A technique called deep mutational scanning can be used to measure the effects of all possible single amino-acid mutations to a protein or protein domain of interest in a single high-throughput experiment [106, 142, 51, 120, 107, 15, 133, 164, 161, 41, 82, 110, 42, 105]. We applied this technique to measure the ability of each homolog to tolerate all single amino-acid amino-acids across 612 homologous sites. A small fraction of mutations had dramatically different effects between homologs, with the mutation being well tolerated in one homolog, but highly deleterious to the other homolog. However, the effects of most mutations had only shifted by small-to-intermediate amounts, suggesting that the effects of most mutations are largely conserved between the divergent homologs despite underlying amino-acid differences at 15% of sites.

3.3 Results

Deep mutational scanning of two Env homologs from transmitted-founder viruses

Previously, we used deep mutational scanning to estimate the effects of all single amino-acid mutations to Env from the LAI strain of HIV, which is a lab-passaged CXCR4-tropic

strain isolated late in infection [127, 65]. Here, we sought to examine the effects of mutations in viruses isolated from earlier in infection, near the time of transmission. We chose two Envs – BG505.W6M.C2 and BF520.W14M.C2 (hereafter referred to as BG505 Env and BF520 Env, respectively) – both from viruses isolated from HIV-infected infants shortly after mother-to-child transmission [59]. Both infants rapidly developed broad plasma responses [59], and an anti-Env broadly neutralizing antibody (bNAb) has been isolated from BF520 [151]. We have previously used deep mutational scanning to comprehensively identify all amino-acid mutations that allow BF520 Env to escape a different bNAb [40]. BG505 Env has been extensively studied from a structural and immunological standpoint [78, 101, 124, 72, 46, 144, 160], and variants of this Env are being tested as a vaccine immunogens [143, 144, 39]. Specifically, we examined the T332N variant of BG505 Env, which has a glycosylation site that is targeted by many bNAbs, but is absent in wild-type BG505 Env [143]. Thus, we considered strains that are relevant to the effort to create antibody-based immunotherapies targeting Env.

The trees in Fig 17 show the relationship between the Envs from BG505, BF520, and LAI in context of other HIV-1 sequences. Fig 17A, which includes group-M sequences from several different subtypes, shows that BG505 and BF520 cluster within subtype-A, while LAI clusters within subtype-B. Fig 17B, which includes an larger number of just subtype-A sequences, shows that BG505 and BF520 are separated by long branches, which is typical of any two sequences in this star-like phylogeny. BG505 and BF520 Env are 85% identical at the amino-acid level, while both are 73% identical to LAI Env. Since each of these Envs are $\sim$850 amino-acids in length, these levels of divergence correspond to > 100 amino-acid differences between each pair of homologs.

We used the deep mutational-scanning approach schematized in Fig 18 to quantify the effects of all single amino-acid changes to both BG505 and BF520 Env. We followed the same technique that we used for LAI [65], with a few modifications. For each homolog, we created a library of *env* genes with random codon mutations. We cloned these libraries into full-length proviral plasmids encoding the Q23 strain of HIV [131], us-

ing high-efficiency cloning to obtain $>$1 million unique plasmid clones per library. Sanger sequencing revealed that for both BG505 and BF520, the codon mutations were evenly distributed across the genes, with an average of 1.5 and 1.1 codon mutations per gene, respectively (see Fig 19 and [40]). We then generated mutant viruses by transfecting the plasmid library into 293T cells. Since transfected cells each receive multiple plasmids, the resulting viruses are unlikely to have a genotype-phenotype link. To establish this link and select for functional variants, we first passaged viral libraries for four days at a low initial multiplicity of infection (MOI) of 0.01 in SupT1 cells expressing CCR5. We then infected SupT1 cells expressing CCR5 with the passaged libraries at a high MOI ($>$1) and harvested reverse-transcribed unintegrated viral DNA 12 hours post-infection, before additional rounds of replication could occur. Essentially, this second infection imposes an additional round of selection for viruses with entry-competent Env variants. Finally, we deep sequenced the plasmid library before selection and the unintegrated proviral DNA after selection to quantify the change in frequency of each mutation. We estimated error rates from PCR and deep sequencing by sequencing wildtype plasmids. We also estimated error rates from viral replication by sequencing wildtype viruses passaged in parallel with the mutant viruses. We conducted the entire experiment in biological triplicate for each homolog (Fig 18B) (except that we only sequenced a single wildtype plasmid for the BF520 experiment).

For the deep mutational-scanning experiments to succeed, the selection step must efficiently purge deleterious mutations from the library. As a first measure of the strength of selection, we examined changes in per-codon mutation frequencies averaged across all sites in Env. Codon-level mutations can be classified as being nonsynonymous, synonymous, or leading to a stop codon. Deep sequencing of the mutant libraries allowed us to quantify the frequency each of these mutation types. In addition, deep sequencing of the wildtype controls allowed us to quantify mutational errors from PCR, deep sequencing, and viral replication. We corrected for these errors by subtracting mutation frequencies in the wildtype plasmids from the those in the mutant plasmids, and the mutation fre-

quencies in the wildtype viruses from those in the mutant viruses. Figure 20A shows the resulting error-corrected per-codon mutation frequencies for each replicate before and after selection. The overall mutation frequency was higher in the BG505 libraries compared to the BF520 libraries, consistent with the higher mutation frequency estimated by Sanger sequencing Fig 19.

The observed changes in per-codon mutation frequencies are consistent with the experiments imposing strong purifying selection. Mutations leading to pre-mature stop codons are expected to be highly deleterious and thus efficiently purged by selection. Indeed, stop codons were purged to between 3-17% their starting frequencies in each replicate (see the red numbers in Figure 20A). This purging was more efficient in the BG505 replicates than in the BF520 replicates, suggesting that the BG505 experiments imposed stronger selection. Nonsynonymous mutations can have more varied effects depending on the site and amino-acid change in question. A variety of studies indicate that many if not most amino-acid mutations tend to be deleterious to any protein [62, 147, 21]. Consistent with this expectation, we found that nonsynonymous mutations were purged to 40-50% their starting frequencies across all replicates. In contrast, the effects of synonymous mutations might be expected to be more neutral than the effects of nonsynonymous mutations. This expectation does not always hold [126, 35, 162, 185], especially in regions where codon mutations have the potential to disrupt important RNA secondary structures, such as HIV's Rev-response element, which overlaps with the *env* codon sequence [49, 65]. However, when averaging across all sites, we found that the synonymous mutations were mostly retained upon selection, only decreasing to 87-95% their starting frequency in each replicate. Overall, these changes are broadly consistent with selection strongly purging deleterious variants in the libraries.

Next, we used the deep-sequencing data to infer each homolog's site-specific amino-acid preferences. For each site (r), we inferred that site's relative preference ($\pi_{r,a}$) for each amino acid (a) as values that are proportional to that amino acid's enrichment or depletion upon selection. These preferences are defined to sum to one at each site

(i.e., $\sum_a \pi_{r,a} = 1$). In making these inferences, we used the deep-sequencing data of the wildtype controls to statistically correct for errors. We mutagenized 670 sites in BG505 and and 662 sites BF520. Thus, we estimated 14,070 (= 670 $\times$ 21) and 13,902 (=662 $\times$ 21) site-specific amino-acid preferences for these homologs, respectively. Fig 22 shows the preferences for BG505 averaged between replicates and rescaled to optimally reflect the strength of selection in nature (as described in the next section). Fig 23 is the same as Fig 22, but shows the preferences for BF520 instead of BG505. Each homolog's preferences are highly heterogeneous among sites. For instance, although some sites strongly prefer just a single amino acid (e.g., site 54), other sites, such as the region of variable loop 4 between sites 396-413, prefer a large number of amino acids roughly equally. Of sites that tolerate multiple amino acids, some sites prefer amino acids with a wide variety of chemical properties (e.g., site 31), while other sites only prefer amino acids with similar properties (e.g., site 35 only prefers aromatic amino acids). Overall, these data quantify each homolog's ability to tolerate each amino acid at each site.

Comparing the preferences between experimental replicates of the same homolog allowed us to quantify the level of experimental noise in our estimates (Figs 20 B and C). Our estimates were largely reproducible between replicates for a give homolog, with Pearson correlation coefficients ranging from 0.59-0.75. The estimates were more reproducible for BG505 than for BF520. This trend might be explained by the fact that stop codons were less efficiently purged in the BF520 experiments (Fig 20A). Thus, a higher fraction of nonfunctional variants may have randomly survived the selection step in the BF520 experiments, which would be expected to result in increased noise. Overall, however, the above results indicate that our estimates were largely reproducible for both homologs.

Re-scaling the experimental measurements to optimally describe HIV evolution in nature

Next, we sought to compare Env's amino-acid preferences between BG505 and BF520. We also sought to compare these homologs to LAI [65]. However, a concern is that

differences we see between homologs may be due to experimental differences, rather than true biological differences. For instance, the LAI experiments involved passaging the viral libraries for a substantially longer amount of time in cell culture, which may have imposed increased selection. Additionally, even though we conducted the BG505 and BF520 experiments using very similar protocols, the experiments seemed to have exerted different levels of purifying selection, as indicated by differential purging of stop codons (Fig 20A). The strength of selection in our experiments may also differ from the strength of selection in nature. Ideally, however, we would like to compare the preferences after rescaling them to reflect the intensity of selection in a natural setting.

To address these concerns, we estimated differences in selection strength between our experiments and nature using a phylogenetic approach. A software package called `phydms` can be used to incorporate site-specific amino-acid preferences into substitution models for maximum-likelihood phylogenetics [69]. These models, known as experimentally informed codon models (ExpCMs), define the probability of a substitution at site r to amino-acid j from amino-acid i as being proportional to the ratio of the site's preferences for these amino acids: $\frac{\pi_{r,j}}{\pi_{r,i}}$. Thus, if j is more preferred than i, the probability of substituting from j to i is greater than substituting from i to j. These models also define a stringency parameter (β) that rescales the preference for each amino acid a at a site to be: $\pi_{r,a}^{\beta}/[\sum_a \pi_{r,a}^{\beta}]$. Given an alignment of natural sequences, ExpCMs can be used to infer the value of this parameter that rescales the preferences to optimally describe the evolution of these sequences. If the inferred stringency parameter is greater than one, the indication is that selection in nature tends to prefer the same amino acids as the experiments, but with greater stringency.

We created ExpCMs with each homolog's averaged preferences. We then used these ExpCMs to analyze the *env* sequences used to make the trees in Fig 17. To do so, we first used `RAxML` to infer the topology of each tree using the GTRCAT model. Fixing this topology, we then optimized the tree's branch lengths using either a standard codon-substitution model (YNGKP M5) or one of the ExpCMs. Table 7 compares the

performance of each model with respect to each tree. For both trees, the ExpCMs describe Env's natural evolution far better than the standard model, as gauged by differences Akaike information criterion (AIC), which quantifies the likelihood of the data given the model while taking into account the number of model parameters, with higher AIC values indicating worse model performance. Notably, BG505 and BF520's preferences describe Env's evolution substantially better than LAI's preferences, though the ExpCM for LAI still outperforms the standard model. As a control, for each homolog, we averaged that homolog's preferences across all sites and made a ExpCM that modeled each site using these averaged preferences. As expected, these site-averaged models perform worse than the site-specific ones, indicating that the preferences recapitulate site-specific constraints in nature. The site-averaged models perform about as well as the standard model, which also lacks site-specific information. Tables 8 and 9 show the results of ExpCMs made using the preferences of individual replicates.

For both alignments, the ExpCMs inferred stringency parameters of >1 for both BG505 and BF520, indicating that selection was weaker in our experiments for these homologs than in nature. The results are largely consistent between the two alignments, indicating that the rescaling is robust to the different levels of sequence diversity we examined. We decided to rescale the preferences for BG505 and BF520 using the stringency parameters from the group-M analysis (Table 7). These resulting rescaled preferences are shown in Figs 22 and 23 for BG505 and BF520, respectively. Since both homologs were rescaled with respect to the same sequences, we expect this rescaling to also normalize the preferences across experiments, helping to correct for differences in selection strength between experiments.

In contrast to BG505 and BF520, for both alignments, the ExpCMs inferred a stringency parameter of $\sim$1 for LAI (Table 7). Thus, both the stringency parameter and the overall model performance was lower for LAI than for the other homologs. One explanation for these differences is biological. Perhaps the preferences for LAI – a lab-adapted virus isolated from chronic infection – are simply less representative of the dominant selective

pressures that shape Env's evolution in nature. An alternative explanation is experimental. It is possible that if we repeated the LAI deep mutational scan using the same method we used for BG505 and BF520, our estimates would change. We are currently repeating these experiments. Thus, for the remainder of this chapter, I will only focus on comparing the preferences for BG505 and BF520.

Quantifying shifts in mutational effects between BG505 and BF520

Having rescaled the preferences of each homolog to optimally describe Env's evolution in nature, we next sought to compare the preferences between homologs to determine how much Env's preferences have shifted over evolutionary time. We estimated preferences for 670 and 662 sites in BG505 and BF520, respectively. We compared the subset of 616 sites that are shared between homologs and are readily alignable in the group-M multiple-sequence alignment. Thus, in total, we compared 12,320 (=616$\times$20) amino-acid preferences between the homologs.

As an initial gene-wide measure of these shifts, we simply quantified the correlation of the preferences between homologs. We expected some of the observed differences to be due to experimental noise rather than true biological differences. We estimated this noise by correlating the preferences between replicates for the same homolog after rescaling the preferences for each replicate using the stringency parameter from Table 7 for the appropriate homolog, as inferred from the group-M alignments. Fig 24A shows this correlation between a single pair of replicates for both BG505 and BF520. These comparisons provide an approximate ceiling for the expected correlation of replicates between homologs if both homologs have identical preferences. Conversely, if Env's preferences have substantially shifted over evolutionary time, the correlation between homologs is expected to be much lower. Fig 24B shows the correlation between homologs for a single pair of replicates. These replicates are nearly as well correlated as two replicates from the same homolog, suggesting that Env's preferences are substantially conserved

between homologs. As a control, we correlated Env's preferences to the preferences of a non-homologous protein – influenza hemagglutinin (HA) – estimated in another study [42] and rescaled with the appropriate stringency parameter from Table 1 of that study. HA's preferences were highly repeatable between the pair of replicates shown in Fig 24A. However, comparing replicates between BG505 and HA across all 480 sites that overlap in sequential numbering yielded a very low correlation, as would be expected when comparing preferences for non-homologous sites. Fig 24 C shows that these trends are consistent across all three replicates for each protein. Thus, the correlations between Env homologs are much higher than the expected correlation between two non-homologous proteins.

Next, we sought to quantify shifts in Env's preferences at a site-specific level. Correlating preferences between replicates of the same homolog showed that some sites were substantially influenced by experimental noise Fig 24. Thus, we quantified shifts using an approach that corrects for noise, similar to the one used in a previous study that deep mutational-scanning data between homologs of influenza's nucleoprotein [41]. Fig 25A shows how we quantified shifts for a few example sites. For a given site r, we define the distance in preferences between an arbitrary pair of replicates i and j as: $D_r^{i,j} = \frac{1}{2}\sum_a |\pi_{r,a}^i - \pi_{r,a}^j|$. We then use this metric to measure the distance between all pairwise combinations of replicates from both homologs. We quantified the un-corrected distance between homologs as the root mean square of all replicate-replicate distances between homologs ($RMSD_{between}$). Since $D_r^{i,j}$ can range between 0-1, $RMSD_{between}$ also ranges between 0-1, with 0 indicating no differences and 1 indicating extreme differences between the homologs. Next, we quantified experimental noise as the root mean square of all replicate-replicate distances where both replicates came from the *same* homolog ($RMSD_{within}$). Finally, we computed the noise-corrected distance between homologs ($RMSD_{corrected}$) by subtracting $RMSD_{between}$ by $RMSD_{within}$.

Sites where we repeatedly measured large differences between homologs are expected to have $RMSD_{corrected}$ values much greater than zero (e.g., sites 512 and 309;

Fig 25A). In contrast, sites where we repeatedly measured small differences between homologs are expected to have $RMSD_{corrected}$ values close to zero (e.g., site 296). Sites where the noise completely overwhelmed the biological signal are also expected to have $RMSD_{corrected}$ values close to zero (e.g., site 607). Thus, a $RMSD_{corrected}$ value near zero could indicate that a site's preferences are conserved between homologs, or that the noise at a site is high.

The blue histogram in Fig 25B shows the distribution of site-specific $RMSD_{corrected}$ values between BG505 and BF520 for all 616 sites being compared (Fig 24). Most sites have distances greater than zero. Yet, only a small fraction of sites have large distances as extreme as sites 512 and 309 ($RMSD_{corrected}$ = 0.30-0.47; Fig 25A), as indicated by the small tail at the positive end of the distribution. Thus, it appears that few sites have dramatically shifted in preference. However, since $RMSD_{corrected}$ is influenced by noise, and since this noise is heterogeneous between sites, it was unclear exactly what the distribution would look like if Env's preferences had actually shifted at a large number of sites. We estimated the expected shape of this distribution by comparing BG505's preferences with the preferences from influenza HA, comparing all 480 sites that overlap in primary sequence. Since these proteins are non-homologous, we expected large site-specific differences in their preferences, mimicking large shifts during evolution. The green histogram in Fig 25B shows the resulting distribution of $RMSD_{corrected}$ values between Env and HA. As expected, the values in the Env-HA comparison tend to be much larger than the values in the Env-Env comparison. Thus, most differences between Env homologs are small-to-intermediate in effect size relative to the typical differences we observe for non-homologous sites.

To examine the preference shifts in greater detail, we made a logo plot that shows the estimated shift for each amino acid at each of the 616 sites being compared (Fig 26). As described in the figure legend, this logo plot shows the results of subtracting the averaged preferences between homologs and then adjusting the total height of the letters at each site to correspond to that site's $RMSD_{corrected}$ value from Fig 25B. Most sites

have small-to-intermediate stack heights, reflecting the fact that most site have relatively small-to-intermediate $RMSD_{corrected}$ values in Fig 25B. Sites where we detect large shifts between homologs are distributed throughout Env's primary sequence. The observed shifts follow a variety of patterns. At some sites, the preferences are strongly shifted for just a few amino acids (e.g., sites 309 and 288). At other sites, the preferences are strongly shifted in terms of the overall number of tolerated mutations (e.g., sites 512, 516, 599). Many of the largest shifts involved shifting from one hyrdrophobic amino acid to another (e.g., sites 165, 288, 307, 309). Overall, the shifts in the preferences are highly heterogeneous between sites. We are currently validating our findings for a subset of mutations by independently cloning each of these mutations into proviral plasmids, and then testing the ability of mutant viruses to replicate in cell culture.

Next, we sought to explore the evolutionary basis of these shifts. One prediction is that the shifts would be concentrated at sites that differ between homologs. For instance, at site 309, Env's preferences shift with wildtype amino acid, where the BG505 wildtype amino acid (isoleucine) is more preferred in BG505 and the BF520 wildtype amino acid (leucine) is more preferred in BF520. This trend was observed in another study that compared shifts in mutational effects between homologs of influenza's nucleoprotein [41]. However, when we considered all sites, we did not find evidence that shifts at variable sites were stronger than shifts at conserved sites (Fig 27). In the future, I plan to examine other evolutionary explanations for these shifts. I also plan to examine these shifts in context of Env's structure, which might lend insight into their functional basis.

3.4 Discussion

We experimentally estimated the effects of all single amino-acid changes to two Env homologs with 85% amino-acid identity. We did so using a high-throughput technique called deep mutational scanning. For each homolog, this approach involved making libraries of *env* with random codon mutations, selecting for mutations that supported viral replication

in cell culture, and then deep sequencing the starting and ending libraries to quantify the effect of each mutation. These experiments imposed strong purifying selection, as indicated by the near-complete depletion of mutations leading to premature stop codons. For each homolog, the results allowed us to estimate each site's preference for each of the 20 amino acids. We found that these estimates were largely repeatable between experimental replicates of the same homolog.

We then compared site-specific amino-acid preferences between homologs. In an initial gene-wide comparison, we simply correlated the preferences between homologs across all sites. We expected that the differences we observed would be due to a combination of true biological differences and experimental noise. We quantified experimental noise by correlating the preferences between experimental replicates of the same homolog. When we then correlated the preferences between replicates from different homologs, we found that the correlation was nearly as high as the correlation between replicates from the same homolog, suggesting that the true preferences of each homolog are fairly well conserved. As a control, we compared Env's preferences with the preferences from a non-homologous protein – influenza HA – among sites that overlap in primary sequence of these proteins. As expected, the correlation in preferences between these non-homologous proteins was very low. This finding provided the first indication that although Env's preferences differ to some extent between homologs, they are still much more conserved than expected for two non-homologous proteins.

Next, we quantified shifts in Env's amino-acid preferences at a more site-specific level. We computed shifts using a metric that takes into account experimental noise. At each site, this metric estimates the expected shift due to the noise, as captured by experimental replicates, and then corrects for this noise when measuring the shifts between homologs. We found that for most sites, the shifts between homologs were small-to-intermediate in effect size, with only a few sites having large shifts. We used the same metric to compare Env and influenza HA at sites that overlap in primary sequence. As expected for two non-homologous proteins, we found that most sites had large shifts in mutational effects. Thus,

this site-specific analysis provided additional support that the shifts observed between Env homologs were not nearly as large as shifts between non-homologous proteins.

Overall, these results help elucidate the strength of epistasis in Env's long-term evolution. Although large shifts in mutational effects can occur during protein evolution, our results indicate that such events were rare for the Env homologs we analyzed. Deep mutational scanning has also been used to analyze mutational shifts between other homologs. One study compared two homologs of influenza's nucleoprotein with 94% amino-acid identity [41]. This study found that most mutational effects were conserved. A lower-throughput study of nucleoprotein mutations with a variety of stability effects found that these effects were also largely conserved among homologs that ranged from 94%-72% amino-acid identity [8]. Another deep mutational-scanning study compared three homologous TIM-barrel proteins, each with 30-40% amino-acid identity to one another [26]. Even for these considerably divergent homologs, mutational effects were still found to be correlated between homologs at many sites. Thus, our findings with Env are qualitatively similar to the above studies in suggesting that although epistasis can cause mutational effects to diverge over time, it does not usually completely erase site-specific preferences.

Our results also shed light on the evolutionary basis of shifts in mutational effects. The deep mutational scan comparing nucleoprotein homologs found that shifts were enriched at sites that differed in wildtype amino-acid sequence between homologs [41]. This trend might be expected if proteins quickly evolve to accommodate historical amino-acid changes, as suggested from computational modeling [130]. However, the median shift in Env's preferences was similar between sites that were the same vs. different in wildtype sequence. Additional comparisons will be required to determine which pattern is more typical.

Future work could expand upon our findings to analyze other Env homologs and the effects of mutations on other Env phenotypes. In this study, we compared two homologs from subtype A. It is possible that mutational effects are more shifted between more divergent Envs. Deep mutational scanning of Env from a different subtype could address this

question (our previous results for LAI are difficult to compare with BG505 and BF520, for reasons described above). The specific phenotype we selected for in our study was the ability of Env to support viral replication in cell culture in the *absence* of immune selection (e.g., antibodies). However, it is of great interest to characterize the effects of mutations to Env on antibody escape, especially for antibodies being used in immunotherapies or as templates in vaccine design. We have previously used deep mutational scanning to comprehensively identify single amino-acid mutations that allow BF520 Env to escape a bNAb [40]. This technique, when applied to two different Env homologs, could be used to comprehensively determine the extent to which antibody-escape mutations have shifted during Env evolution.

3.5 Methods

Sequence numbering

We use the HXB2 numbering system [84] for each Env homolog unless otherwise noted. We defined Env's "variable loops" according to `http://www.hiv.lanl.gov/`, but did not consider the disulfide-bonded cysteines as part of the loops.

Codon-mutant libraries

For BF520 *env*, we used the codon-mutant libraries generated from a previous study [40]. For BG505 *env*, we generated codon-mutant libraries using essentially the same methods as the previous study, but with a few modifications. We computationally generated the sequences of the codon-tiling primers for the PCR-mutagenesis step using the same approach as in [40]. The sequences of these primers will be made available as a supplemental file upon publication of this work. The end primers for this mutagenesis step were: 5'-tgaaggcaaaactactggtccgtctcgagcagaagacagtggcaatgaga-3' and 5'-gctacaaatgcatataacagcgtctcattctttccctaacctcaggcca-3'. As with BF520, we cloned the BG505 *env* libraries into the *env* locus of the full-length proviral genome of HIV strain

Q23 [131] – another subtype-A transmitted/founder virus, using the same high-efficiency cloning vector used previously [40]. To do so, we first digested the cloning vector with BsmBI. We then used PCR to elongate the amplicons to include on either end 30 base-pairs that are identical in sequence to the ends of the BsmBI-digested vector. The primers we used for this PCR were: 5'-agataggttaattgagagaataagagaaagagcagaagacagtggcaatgagagtgatgg-3' and 5'-ctcctggtgctgctggaggggcacgtctcattctttccctaacctcaggccatcc-3'. Next, we used NEBuilder HiFi DNA Assembly (NEB, E2621S) to clone the *env* amplicons into the BsmBI-digested plasmids. We purified the assembled products using Agencourt AMPure XP beads (Beckman Coulter, A63880) using a bead-to-sample ratio of 1.5, and then transformed the purified products into Stellar electrocompetent cells (Takara, 636765). The transformations yielded between 1.5-3.6 million unique clones for each of the three replicate libraries, as estimated by plating 1:2,000 dilutions of the transformations. As before, we scraped the plated colonies and maxiprepped the plasmid DNA, but this time we did not include the 4-hour outgrowth step after the scraping step. For the wildtype controls, we maxiprepped three independent cultures of wildtype BG505 *env* in the *env* locus of the full-length Q23 proviral plasmid.

Generation and passaging of viruses

For BF520, we analyzed the viruses generated and passaged in our previous study [40]. For BG505, we generated and passaged viral libraries using essentially the same methods. First, for each replicate, we generated mutant viruses by transfecting 293T cells in three 6-well plates with a mixture of 2 ug mutant plasmid DNA and 6 uL FuGENE 6 Transfection Reagent (Promega, E269A) and 100 uL DMEM per well. 293T cells were seeded in D10 media (DMEM supplemented with 10% FBS, 1% 200 mM L-glutamine, and 1% of a solution of 10,000 units/mL penicillin and 10,000 μg/mL streptomycin) the day before transfection with 0.5 million cells per well, such that they were approximately 50% confluent the next day. In parallel, we generated wildtype viruses by transfecting one 6-well

plate of 293T cells with wildtype plasmid, using the same amount of DNA and transfection reagent as above. At 2days post-transfection, we harvested the transfection supernatant, passed it through a 0.2μm filter, treated the supernatant with DNAse to digest residual plasmid DNA (as in [65]), and froze aliquots at -80°C. We thawed and titered aliquots using the TZM-bl assay in the presence of 10 μg/mL DEAE-dextran as described in [40].

Next, we conducted the initial viral passage in SupT1.CCR5 cells (obtained from Dr. James Hoxie [23]). During this passage, cells were maintained in R10 media, which has the same composition as D10 (described above), but has RPMI-1640 (GE Healthcare Life Sciences, SH30255.01) in the place of DMEM, with 10 μg/mL DEAE-dextran to enhance viral infection. We infected cells with 4 million (for replicate 1) or 5 million (for replicates 2 and 3) TZM-bl infectious units of mutant virus at an MOI of 0.01, with cells at a starting concentration of 1 million cells/mL in vented tissue-culture flasks (Fisher Scientific, 14-826-80). At 1 day post-infection, we pelleted cells, aspirated the supernatant, and resuspended cell pellets in fresh media including DEAE-dextran. At 2 days post-infection, we doubled the volume of each culture with fresh media including DEAE-dextran. At 4 days post-infection, we pelleted cells, passed the virus-containing supernatant through a 0.2 μm filter, concentrated the virus $\sim$30 fold using ultracentrifugation as described in [40], and then froze aliquots at -80°C. In parallel, for each replicate, we also passaged 0.2 million (for replicate 1) or 0.5 million (for replicates 2 and 3) infectious units of wildtype virus with and using the same conditions. We thawed and titered these aliquots using the TZM-bl assay in the presence of 10 μg/mL DEAE-dextran.

We conducted a second, much shorter viral passage by infecting cells with the passage-1 viruses. For each virus, we infected 1 million TMZ-bl infectious units into 1 million SupT1.CCR5 cells in the presence of 100 μg/mL DEAE-dextran (a 10-fold higher concentration than to concentration we used in the TZM-bl assay to titer the virus, meaning the actual number of infectious units in this passage was probably higher). Three hours post-infection, we pelleted the cells and resuspended them in fresh media without any DEAE-dextran. At 12 hours post-infection, we pelleted cells, washed them once with

PBS, and then used a miniprep kit to harvest reverse-transcribed unintegrated viral DNA.

Generation of samples for Illumina sequencing

We deeply sequenced each plasmid library before selection and each viral library after the two passages, as well as the wildtype plasmids and wildtype viruses that served as controls. First, we generated PCR amplicons of *env* using the same approach as [40] with the primers: 5'-GAAGACAGTGGCAATGAGAGTGATGG-3' and 5'-TTCCCTAACCTCAGGCCATCC-3'. Next, we sequenced these amplicons using a barcoded sub-amplicon sequencing strategy to reduce the sequencing error rate, as previously described [42, 65], dividing *env* into seven amplicons. The primer pairs used to generate these subamplicons are as follows, where N characters represent randomized sites in the barcode region of the primer:

- 5'-CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNGATCTTGGGGATGATAATAATCTGTAGTGC-3' and 5'-GGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNGGTGACATTGGTACACTGTAGAGTAAC-3'
- 5'-CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNCCATGTGTAAAGTTAACCCCTCTCTGC-3' and 5'-GGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNGAACTTCTTATCCTTACACTTTAGGATCGC-3'
- 5'-CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNCATACATTATTGTGCCCCAGCTGG-3' and 5'-GGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNCAATGTGCTTGTCTTATATCCCCTATTATGTC-3'
- 5'-CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNGGACAAGCATTCTATGCAACAGGG-3' and 5'-GGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNTGCTTTATTCTGCATGGGAGAGTTATAC-3'
- 5'-CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNGTCAAATAGCACGGGGTCAAATGAC-3' and 5'-GGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNCCAAGGAAGACAGCTCCTATTCCAAC-3'

- 5'-CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNGTGGTGGGGAGAGAAAAAAGAGC-3' and 5'-GGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNGTTTCTATTACTCCAACTAGAGTTCCAGGG-3'

- 5'-CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNGGAAAACTCATCTGCACCACTAATGTG-3' and 5'-GGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNTGAGTATCCCTGCCTAACTCTATGTATTACAG-3'

The resulting deep-sequencing data will be made available on the Sequence Read Archive upon the publication of this work.

Analysis of deep-sequencing data

We analyzed the deep-sequencing data using the `dms_tools` software package [17]. The specific code we used will be made available upon the publication of this work. A summary of our workflow is as follows, and involves multiple programs encoded in `dms_tools`: First, we aligned sequencing reads to *env* and computed site-specific codon mutation frequencies using `dms_barcodedsubamplicons`, using `dms_summarizealignments` to make plots summarizing these frequencies. Next, we inferred Env's preferences from the codon counts using `dms_inferprefs`. We then used a custom script to compare these preferences between homologs.

Alignments and phylogenetic analyses of HIV-1 env sequences

We made the trees in Fig **??** using multiple-sequence alignments downloaded from the HIV sequence database (`http://www.hiv.lanl.gov/`) that we then manually curated in the following ways: First, we removed sequences that differed in length from the HXB2 reference sequence (including gap characters), or which contained a premature stop codon, ambiguous residue, or frame-shift mutation. Next, we removed columns in the alignment for which we lack deep mutational scanning data or are not present in HXB2, columns

that have $>$5% gap characters, or columns in variable loops that looked poorly aligned by eye. The sequences used to make the trees in Fig **??** were randomly selected from the larger alignment, with a defined number of sequences per clade, as described in the figure. The topologies of these trees were inferred by `RAxML` using the GTRCAT model. In the `phydms` [69] analyses, we fixed the topologies of these trees, and then compared the ability of the different models to optimize the branch lengths. The trees shown in Fig **??** are the results of optimizing the branch lengths using the YNGKP M5 model. The code we used to generate these alignments, build the trees, and conduct the `phydms` analyses will be made available upon publication of this work.

Chapter 4

CONCLUSION

Env is shaped by a wide variety of evolutionary forces. Two such forces are inherent selection for Env to fold and function and external selection for Env to evade the immune system. Since both of these forces strongly influence Env's evolution, their individual effects have been difficult to disentangle from one another. In my graduate research, I used deep mutational scanning to quantify the effects of all amino-acid mutations to Env in context of three homologs. Since I performed these experiments in the lab in the absence of immune selection, the results provide an in-depth view of the inherent selection on Env in the absence of external selection. In Chapter 2, I used these data to compare the mutational tolerance of different sites in Env. I found that conserved antibody epitopes overlapping with CD4 binding site tended to be less tolerant of mutations than other sites in the protein, after correcting for other variables influencing mutational tolerance. This finding provided rigorous support for a long-standing hypothesis in the field, and suggests that since residues in these epitopes are less tolerant of mutations in the absence of immune selection, that they may have a reduced evolutionary capacity to evade an immune response – antibody-escape mutations must still preserve Env's ability to function or they will not be propagated. In contrast, my data did not support the hypothesis that Env's variable loops are more tolerant of mutations than other sites in Env. Thus, the high variability of these loops in nature may primarily be driven by strong diversifying selection for these sites to evade host immunity. In Chapter 3, I proceeded to compare Env's mutational tolerance among two divergent Env homologs with 85% amino-acid identity. I found that a small fraction of mutations had dramatically different effects in the homologs. However, most mutations differed by small-to-intermediate effect sizes, suggesting that

Env's mutational tolerance is largely conserved among the homologs I examined, despite many ($>$100) underlying amino-acid sequence differences. Relevant to the field of HIV, this finding suggests that measurements of the effects of mutations on viral replication for one Env homolog tend to be roughly generalizable to other homologs. Relevant to the broader field of protein evolution, this finding helps elucidate the extent that the effects of mutations shift over long-term protein evolution, suggesting that these shifts are mostly small-to-intermediate, at least for the level of sequence divergence that I considered.

Below, I describe additional ways that my data could be used to gain insight into HIV evolution. My research is also valuable from a technological standpoint. Other groups had used deep mutational scanning to measure the ability of cell surface-displayed Env to bind antibodies, or to measure the effects of thousands of mutations scattered across the genome on viral replication. However, I was the first to use this approach to comprehensively measure the effects of all amino-acid mutations to an HIV protein in the context of viral replication. Below, I also describe a variety of exciting ways that this method could be adapted to study other aspects of Env.

Using evolutionary models to detect sites in Env that evolve differently in nature than expected based Env's amino-acid preferences the lab. In the previous chapters, I discuss how inherent selection shapes the evolution of several regions in Env: glycosylation sites, disulfide bonds, variable loops, and epitopes of broadly neutralizing antibodies. Each of these regions was pre-selected for playing a known role in immune escape or, in the case of the disulfide bonds, structural integrity. In the future, it would be interesting to take a more agnostic approach in analyzing how inherent vs. external selection shape Env's evolution across all sites in the protein. Indeed, the data I generated are extremely comprehensive in that I measured the effects of all amino-acid mutations at nearly all sites in Env, for three different homologs. In theory, this data provides a null model for how we would expect Env's evolution to proceed if it were just shaped by inherent selection.

Sites that deviate from this null model may be under additional external selective pressures in nature, and of biological interest for that reason. For instance, I would expect some sites to evolve much faster in nature than expected based on this null model. These sites might be under strong diversifying selection to evade immunity. In contrast, I would expect other sites to evolve much slower in nature than expected based on the null. These sites might also be under immune selection, but selection that is purifying instead of diversifying. Env's glycosylation sites are an excellent example of this phenomenon. Many of these sites are highly conserved in nature, despite being highly tolerant of mutations in the lab. This high conservation is likely because glycans are important for shielding Env from antibodies, and are thus strongly selected for in nature, but not the lab. My data provides a sensitive way to detect this phenomenon across all other sites.

This approach is conceptually similar to standard evolutionary models for detecting diversifying or purifying selection. These models can be used to infer the relative rate of non-synonymous to synonymous substitutions in a gene. Sites with high rates (>1) are inferred to be under diversifying selection, whereas sites with low rates (<1) are inferred to be under purifying selection. The Bloom lab has developed similar evolutionary models that also take into account a protein's amino-acid preferences, as measured in the lab [69]. Another study found that these models can be much more sensitive than standard models in identifying sites under diversifying selection [19]. They also provide a more informative null model for detecting sites that evolve especially slowly [19]. I have applied these models to Env, and preliminarily, there are many sites that I detect as being under a significant amount of either diversifying or purifying selection. In the future, it would be very interesting to further characterize these sites. Is there always a clear explanation for why these sites behave differently in nature vs. the experiments (e.g., immune selection constraining glycosylation sites) or do we identify sites that evolve differently for some unknown reason?

Using deep mutational scanning to measure the effects of mutations on other phenotypes in the lab. My research is exciting from a technological standpoint since it is the first time that deep mutational scanning was used to measure the effects of all single amino-acid changes to any HIV protein in the context of viral replication. In theory, this technique could be extended to measure the effects of mutations on any Env phenotype that is measurable in the lab. Thus, I envision it could be used to address a wide variety of questions.

For instance, I have collaborated with another graduate student in the lab to use this method to comprehensively identify mutations that allow Env to evade a broadly neutralizing antibody [40]. To do so, we simply selected the mutant viral libraries for their ability to enter cells both in the presence and absence of the antibody, and then determined which mutations were enriched in the presence of the antibody. In theory, this approach could be applied to study any neutralizing antibody. For antibodies with unknown epitopes, the results could help precisely delineate where that antibody binds, as escape mutations would be expected to be enriched in the antibody epitope. For antibodies that are being used as templates in vaccine design, or which are being passively administered as a therapeutic or prophylactic, the results could help identify ways that the virus could evolve to escape the antibody. By analogy, the selection could also be conducted with anti-viral drugs that target Env, such as T20 (i.e., DP178) [177]. This knowledge could help inform how medical interventions are designed, as well as evaluating the results of such interventions.

Further down the road, it may be possible to identify antibody-escape mutations not just to single monoclonal antibodies, but to polyclonal serum. An important validation experiment would be to mix two monoclonal antibodies at different proportions to gauge the sensitivity of this approach to different mixtures of antibodies. Once validated, characterizing escape mutations to polyclonal serum could help address a variety of questions, such as: What are the specific epitopes targeted by humoral immunity during a natural infection? How does this response change overtime as the virus evolves? Do viruses evolve

via the same antibody-escape mutations we identify in our experiments? And lastly, which epitopes are targeted upon vaccination with an antigen?

Our current method for identifying antibody-escape mutations requires that antibodies neutralize the virus to prevent it from replicating. However, there are also non-neutralizing antibodies that do not directly block HIV from entering cells, but combat the virus through other mechanisms such as antibody-dependent cellular cytotoxicity [122]. In the future, the experiment could be tweaked to allow characterization of these antibodies as well. A straight-forward modification would be to first incubate the viral libraries with and without antibodies, and then before infecting cells, separate the antibody-bound viruses from the unbound viruses. This separation step could be accomplished using a either a column or magnetic beads with a secondary antibody that binds to the first. Infecting cells with the unbound viruses would then be important for selecting viruses that both escaped antibody binding and were still functional (unfolded proteins would not be expected to bind antibodies).

In addition to being targeted by humoral immunity, Env also interacts with a family of proteins from innate immunity called interferon-induced transmembrane proteins (IFITMs) [52]. Although sequence differences in Env are known to influence IFITM restriction, the details of this interaction are still largely unclear. In the future, it could be interesting to repeat the deep mutational scan in the presence of IFITM to map single amino-acid mutations that enable escape. This might help suggest a mechanism by which these proteins interact.

A very different application of deep mutational scanning than the ones above could be to use it to examine the dual RNA- and protein-level constrains at HIV's Rev-response element (RRE). The RRE encodes a RNA secondary structure that is essential for nuclear export of certain viral RNAs. This region also overlaps with part of *env*'s coding sequence. Thus, the RRE is under constraint at the levels of both RNA and protein. Such dual constraint is common in viruses with small genomes, presumably because the space to encode new functions is very limited. A fascinating question is: to what extent is the

evolution of the Env protein restricted by the RRE RNA? It may be possible to answer this question by repeating the deep mutational scanning experiment with a modified virus in which the RRE is expendable, with its function being replaced by a sequence element from another virus [24]. In this context, mutations would only be selected for their effects on protein function. I would expect a substantial number of sites be have an increased tolerance for mutations in this scenario.

It could also be interesting to repeat the deep mutational scan of Env in a variety of different cells. All of the experiments in my graduate research involved functional selection for Env's ability to support replication in Sup-T1 T cells. However, in nature, HIV replicates in multiple different cell types with varying levels of receptor and co-receptors on the cell surface [33, 92]. Thus, our experiments are probably a vast oversimplification of the variety of selective environments in which Env evolves in natures, even without considering external selection from immunity. HIV can also shift its co-receptor preference during an infection, with earlier viruses typically preferring CCR5 and later viruses sometimes evolving the ability to use CXCR4 [7]. Deep mutational scanning in cells with different receptor compositions and levels could help examine cell-type specific selection as well as the sequence determinants of switching co-receptor specificity.

Finally, it should be possible to adapt this method to study other HIV proteins. Many of the steps that I optimized for Env should be generalizable to other proteins (e.g., generating mutant plasmid libraries with high-efficiency cloning and deep sequencing of mutant viral libraries). However, the selection step may need to be tweaked depending on the protein and where it acts in the viral lifecycle and how many rounds of selection are required to efficiently purge deleterious mutations. Moreover, a concern with oligomeric proteins like Gag is that there may be dominant-negative effects that allow non-functional variants to randomly bottleneck the library when all mutant variants are transfected into cells at once.

Quantifying the prevalence of epistasis among additional Env homologs and between Env and other genes in HIV. In my graduate research, I measured mutational effects to three homologs of Env: LAI, BF520, and BG505. I conducted an in-depth comparison of BF520 and BG505, which are both from subtype A, and have 85% amino-acid identity. For reasons described previously, there are some experimental and biological factors that may confound the comparison between these two homologs and LAI. In the comparison between BF520 and BG505, I observed that only a few sites had undergone large shifts in amino-acid preferences. But maybe Env's preferences have shifted by a much larger amount between more divergent homologs. Env homologs from different subtypes typically range between 65-80% amino-acid identity. In the future, it would be interesting to quantify shifts among more divergent homologs to test the above hypothesis. An even more adventurous experiment would be to use repeat the deep mutational scanning experiment with an SIV Env in a relevant primate cell line. I would expect much larger shifts to have occurred over the much longer amount of evolutionary time that separates HIV Env from different SIV Env proteins, and due to the different selective pressures these Envs face (e.g., usage of human vs. non-human receptors). However, it's also possible that this comparison could sites where Env's preferences have remained conserved. Indeed overall fold of gp120 is conserved between HIV and at least one homolog of SIV gp120 [29].

Lastly, in my research, I only investigated *intra*-protein epistasis between residues in Env. However, it is also possible that there are epistatic interactions between Env and other proteins in the HIV genomes. For instance, Env is known to functionally interact with Gag [115]. A straight forward experiment could be to clone the library of mutant *env* genes into two two different HIV genomes and then select for functional variants in both contexts. Such an experiment would help determine how much Env's amino-acid preferences have shifted over evolutionary time due to *inter*-protein epistasis, which would have similar implications as my research on intra-protein epistasis.

BIBLIOGRAPHY

[1] Christopher D Aakre, Julien Herrou, Tuyen N Phung, Barrett S Perchuk, Sean Crosson, and Michael T Laub. Evolving new protein-protein interaction specificity through promiscuous intermediates. *Cell*, 163(3):594–606, 2015.

[2] DV Ablashi, ZN Berneman, B Kramarsky, J Whitman, Y Asano, and GR Pearson. Human herpesvirus-7 (hhv-7): current status. *Clinical and diagnostic virology*, 4(1):1–13, 1995.

[3] Michael E Abram, Andrea L Ferris, Wei Shao, W Gregory Alvord, and Stephen H Hughes. Nature, position, and frequency of mutations made in a single cycle of hiv-1 replication. *Journal of virology*, 84(19):9864–9878, 2010.

[4] Laith Q Al-Mawsawi, Nicholas C Wu, C Anders Olson, Vivian Cai Shi, Hangfei Qi, Xiaojuan Zheng, Ting-Ting Wu, and Ren Sun. High-throughput profiling of point mutations across the hiv-1 genome. *Retrovirology*, 11(1):124, 2014.

[5] Laith Q Al-Mawsawi, Nicholas C Wu, CA Olson, Vivian C Shi, Hangfei Qi, Xiaojuan Zheng, Ting-Ting Wu, and Ren Sun. High-throughput profiling of point mutations across the hiv-1 genome. *Retrovirology*, 11(1):124, 2014.

[6] Jan Albert, Bengt Abrahamsson, Karoly Nagy, Elisabeth Aurelius, Hans Gaines, Gunnel Nyström, and Eva Maria Fenyö. Rapid development of isolate-specific neutralizing antibodies after primary hiv-1 infection and consequent emergence of virus variants which resist neutralization by autologous sera. *Aids*, 4(2):107–112, 1990.

[7] Ghalib Alkhatib. The biology of ccr5 and cxcr4. *Current Opinion in HIV and AIDS*, 4(2):96, 2009.

[8] Orr Ashenberg, L Ian Gong, and Jesse D Bloom. Mutational effects on stability are largely conserved during protein evolution. *Proceedings of the National Academy of Sciences*, 110(52):21071–21076, 2013.

[9] Orr Ashenberg, Jai Padmakumar, Michael B Doud, and Jesse D Bloom. Deep mutational scanning identifies sites in influenza nucleoprotein that affect viral inhibition by mxa. *PLoS pathogens*, 13(3):e1006288, 2017.

[10] Stéphane Basmaciogullari, Gregory J Babcock, Donald Van Ryk, Woj Wojtowicz, and Joseph Sodroski. Identification of conserved and variable structures in the human immunodeficiency virus gp120 glycoprotein of importance for cxcr4 binding. *Journal of virology*, 76(21):10791–10800, 2002.

[11] Robert Belshaw, Oliver G Pybus, and Andrew Rambaut. The evolution of genome compression and genomic novelty in rna viruses. *Genome research*, 17(10):1496–1504, 2007.

[12] Edward A Berger, Philip M Murphy, and Joshua M Farber. Chemokine receptors as hiv-1 coreceptors: roles in viral entry, tropism, and disease. *Annual review of immunology*, 17(1):657–700, 1999.

[13] Shimon Bershtein, Michal Segal, Roy Bekerman, Nobuhiko Tokuriki, and Dan S Tawfik. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature*, 444(7121):929, 2006.

[14] Claudia Blattner, Jeong Hyun Lee, Kwinten Sliepen, Ronald Derking, Emilia Falkowska, Alba Torrents de la Peña, Albert Cupo, Jean-Philippe Julien, Marit van Gils, Peter S Lee, et al. Structural delineation of a quaternary, cleavage-dependent epitope at the gp41-gp120 interface on intact hiv-1 env trimers. *Immunity*, 40(5):669–680, 2014.

[15] Jesse D Bloom. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Mol. Biol. Evol*, 31(8):1956–1978, 2014.

[16] Jesse D Bloom. An experimentally informed evolutionary model improves phylogenetic fit to divergent lactamase homologs. *Molecular biology and evolution*, 31(10):2753–2769, 2014.

[17] Jesse D Bloom. Software for the analysis and visualization of deep mutational scanning data. *BMC bioinformatics*, 16:168, 2015.

[18] Jesse D. Bloom. Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. *bioRxiv*, page 037689, 2016.

[19] Jesse D Bloom. Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. *Biology direct*, 12(1):1, 2017.

[20] Jesse D Bloom, Sy T Labthavikul, Christopher R Otey, and Frances H Arnold. Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences*, 103(15):5869–5874, 2006.

[21] Jesse D Bloom, Jonathan J Silberg, Claus O Wilke, D Allan Drummond, Christoph Adami, and Frances H Arnold. Thermodynamic prediction of protein neutrality. *Proceedings of the National Academy of Sciences of the United States of America*, 102(3):606–611, 2005.

[22] Jeffrey I Boucher, Pamela Cote, Julia Flynn, Li Jiang, Aneth Laban, Parul Mishra, Benjamin P Roscoe, and Daniel NA Bolon. Viewing protein fitness landscapes through a next-gen lens. *Genetics*, 198(2):461–471, 2014.

[23] David F Boyd, Dylan Peterson, Beth S Haggarty, Andrea PO Jordan, Michael J Hogan, Leslie Goo, James A Hoxie, and Julie Overbaugh. Mutations in hiv-1 envelope that enhance entry with the macaque cd4 receptor alter antibody recognition by

disrupting quaternary interactions within the trimer. *Journal of virology*, 89(2):894–907, 2015.

[24] Molly Bray, Susan Prasad, John W Dubay, Eric Hunter, Kuan-Teh Jeang, David Rekosh, and ML Hammarskjöld. A small element from the mason-pfizer monkey virus genome makes human immunodeficiency virus type 1 expression and replication rev-independent. *Proceedings of the National Academy of Sciences*, 91(4):1256–1260, 1994.

[25] L Chakrabarti, M Emerman, P Tiollais, and P Sonigo. The cytoplasmic domain of simian immunodeficiency virus transmembrane protein modulates infectivity. *Journal of virology*, 63(10):4395–4403, 1989.

[26] Yvonne H Chan, Sergey V Venev, Konstantin B Zeldovich, and C Robert Matthews. Correlation of fitness landscapes from three orthologous tim barrels originates from sequence and structure constraints. *Nature Communications*, 8:14614, 2017.

[27] Sheng-Yung P Chang, Barbara H Bowman, Judith B Weiss, Rebeca E Garcia, and Thomas J White. The origin of hiv-1 isolate htlv-iiib. *Nature*, 363(6428):466–469, 1993.

[28] Bing Chen, Erik M Vogan, Haiyun Gong, John J Skehel, et al. Structure of an unliganded simian immunodeficiency virus gp120 core. *Nature*, 433(7028):834, 2005.

[29] Bing Chen, Erik M Vogan, Haiyun Gong, John J Skehel, Don C Wiley, and Stephen C Harrison. Determining the structure of an unliganded and fully glycosylated siv gp120 envelope glycoprotein. *Structure*, 13(2):197–211, 2005.

[30] F-C Chen, EJ Vallender, H Wang, C-S Tzeng, and W-H Li. Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences. *Journal of Heredity*, 92(6):481–489, 2001.

[31] Cyrus Chothia and Arthur M Lesk. The relation between the divergence of sequence and structure in proteins. *The EMBO journal*, 5(4):823, 1986.

[32] Peter Y Chou and Gerald D Fasman. Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochemistry*, 13(2):211–222, 1974.

[33] Paul R Clapham and Áine McKnight. Hiv-1 receptors and cell tropism. *British medical bulletin*, 58(1):43–59, 2001.

[34] A Cordonnier, L Montagnier, and M Emerman. Single amino-acid changes in hiv envelope affect viral tropism and receptor binding. *Nature*, 340(6234):571, 1989.

[35] José M Cuevas, Pilar Domingo-Calap, and Rafael Sanjuán. The fitness effects of synonymous mutations in dna and rna viruses. *Molecular Biology and Evolution*, 29(1):17–20, 2012.

[36] José M Cuevas, Ron Geller, Raquel Garijo, José López-Aldeguer, and Rafael Sanjuán. Extremely high mutation rate of hiv-1 in vivo. *PLoS Biol*, 13(9):e1002251, 2015.

[37] Jack da Silva, Mia Coetzer, Rebecca Nedellec, Cristina Pastore, and Donald E Mosier. Fitness epistasis and constraints on adaptation in a human immunodeficiency virus type 1 protein region. *Genetics*, 185(1):293–303, 2010.

[38] Vincent Dahirel, Karthik Shekhar, Florencia Pereyra, Toshiyuki Miura, Mikita Artyomov, Shiv Talsania, Todd M Allen, Marcus Altfeld, Mary Carrington, Darrell J Irvine, et al. Coordinate linkage of hiv evolution reveals regions of immunological vulnerability. *Proceedings of the National Academy of Sciences*, 108(28):11530–11535, 2011.

[39] Steven W de Taeye, Gabriel Ozorowski, Alba Torrents de la Peña, Miklos Guttman, Jean-Philippe Julien, Tom LGM van den Kerkhof, Judith A Burger, Laura K

Pritchard, Pavel Pugach, Anila Yasmeen, et al. Immunogenicity of stabilized hiv-1 envelope trimers with reduced exposure of non-neutralizing epitopes. *Cell*, 163(7):1702–1715, 2015.

[40] Adam S Dingens, Hugh K Haddox, Julie Overbaugh, and Jesse D Bloom. Comprehensive mapping of hiv-1 escape from a broadly neutralizing antibody. *Cell Host & Microbe*, 2017.

[41] MB Doud, O Ashenberg, and JD Bloom. Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. *Molecular biology and evolution*, 32(11):2944–2960, 2015.

[42] Michael B Doud and Jesse D Bloom. Accurate measurement of the effects of all amino-acid mutations to influenza hemagglutinin. *Viruses*, 8:155, 2016.

[43] Michael B Doud, Scott E Hensley, and Jesse D Bloom. Complete mapping of viral escape from neutralizing antibodies. *PLoS pathogens*, 13(3):e1006271, 2017.

[44] Michael L Doyle, David J Casper, Claudia Cicala, et al. Hiv-1 evades antibody-mediated neutralization through conformational masking of receptor-binding sites. *Nature*, 420(6916):678, 2002.

[45] Michael Emerman, Rosemay Vazeux, and Keith Peden. The rev gene product of the human immunodeficiency virus affects envelope-specific rna localization. *Cell*, 57(7):1155–1165, 1989.

[46] Emilia Falkowska, Khoa M Le, Alejandra Ramos, Katie J Doores, Jeong Hyun Lee, Claudia Blattner, Alejandro Ramirez, Ronald Derking, Marit J van Gils, Chi-Hui Liang, et al. Broadly neutralizing hiv antibodies define a glycan-dependent epitope on the prefusion conformation of gp41 on cleaved envelope trimers. *Immunity*, 40(5):657–668, 2014.

[47] Nuno R Faria, Andrew Rambaut, Marc A Suchard, Guy Baele, Trevor Bedford, Melissa J Ward, Andrew J Tatem, João D Sousa, Nimalan Arinaminpathy, Jacques Pépin, et al. The early spread and epidemic ignition of hiv-1 in human populations. *Science*, 346(6205):56–61, 2014.

[48] Andrew L Ferguson, Jaclyn K Mann, Saleha Omarjee, Thumbi Ndungu, Bruce D Walker, and Arup K Chakraborty. Translating hiv sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity*, 38(3):606–617, 2013.

[49] Jason Fernandes, Bhargavi Jayaraman, and Alan Frankel. The hiv-1 rev response element: an rna scaffold that directs the cooperative assembly of a homo-oligomeric ribonucleoprotein complex. *RNA biology*, 9(1):6–11, 2012.

[50] Jason D Fernandes, Tyler B Faust, Nicolas B Strauli, Cynthia Smith, David C Crosby, Robert L Nakamura, Ryan D Hernandez, and Alan D Frankel. Functional segregation of overlapping genes in hiv. *Cell*, 167(7):1762–1773, 2016.

[51] Elad Firnberg, Jason W Labonte, Jeffrey J Gray, and Marc Ostermeier. A comprehensive, high-resolution map of a genes fitness landscape. *Molecular biology and evolution*, 31(6):1581–1592, 2014.

[52] Toshana L Foster, Harry Wilson, Shilpa S Iyer, Karen Coss, Katie Doores, Sarah Smith, Paul Kellam, Andrés Finzi, Persephone Borrow, Beatrice H Hahn, et al. Resistance of transmitted founder hiv-1 to ifitm-mediated restriction. *Cell host & microbe*, 20(4):429–442, 2016.

[53] Douglas M Fowler, Carlos L Araya, Sarel J Fleishman, Elizabeth H Kellogg, Jason J Stephany, David Baker, and Stanley Fields. High-resolution mapping of protein sequence-function relationships. *Nature methods*, 7(9):741–746, 2010.

[54] Douglas M Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature methods*, 11(8):801–807, 2014.

[55] EO Freed and MA Martin. Evidence for a functional interaction between the v1/v2 and c4 domains of human immunodeficiency virus type 1 envelope glycoprotein gp120. *Journal of virology*, 68(4):2503–2512, 1994.

[56] EO Freed, DJ Myers, and R Risser. Mutational analysis of the cleavage sequence of the human immunodeficiency virus type 1 envelope glycoprotein precursor gp160. *Journal of virology*, 63(11):4670–4675, 1989.

[57] Romain Gasser, Meriem Hamoudi, Martina Pellicciotta, Zhicheng Zhou, Clara Visdeloup, Philippe Colin, Martine Braibant, Bernard Lagane, and Matteo Negroni. Buffering deleterious polymorphisms in highly constrained parts of hiv-1 envelope by flexible regions. *Retrovirology*, 13(1):50, 2016.

[58] Lizhi Ian Gong, Marc A Suchard, and Jesse D Bloom. Stability-mediated epistasis constrains the evolution of an influenza protein. *Elife*, 2:e00631, 2013.

[59] Leslie Goo, Vrasha Chohan, Ruth Nduati, and Julie Overbaugh. Early development of broadly neutralizing antibodies in hiv-1-infected infants. *Nature medicine*, 20(6):655–658, 2014.

[60] Leslie Goo, Caitlin Milligan, Cassandra A Simonich, Ruth Nduati, and Julie Overbaugh. Neutralizing antibody escape during hiv-1 mother-to-child transmission involves conformational masking of distal epitopes in envelope. *Journal of virology*, 86(18):9566–9582, 2012.

[61] Bryan T Grenfell, Oliver G Pybus, Julia R Gog, James LN Wood, Janet M Daly, Jenny A Mumford, and Edward C Holmes. Unifying the epidemiological and evolutionary dynamics of pathogens. *science*, 303(5656):327–332, 2004.

[62] Haiwei H Guo, Juno Choe, and Lawrence A Loeb. Protein tolerance to random amino acid change. *Proceedings of the National Academy of Sciences of the United States of America*, 101(25):9205–9210, 2004.

[63] Miklos Guttman, Albert Cupo, Jean-Philippe Julien, Rogier W Sanders, Ian A Wilson, John P Moore, and Kelly K Lee. Antibody potency relates to the ability to recognize the closed, pre-fusion form of hiv env. *Nature communications*, 6, 2015.

[64] Jürgen Haas, Eun-Chung Park, and Brian Seed. Codon usage limitation in the expression of hiv-1 envelope glycoprotein. *Current Biology*, 6(3):315–324, 1996.

[65] Hugh K Haddox, Adam S Dingens, and Jesse D Bloom. Experimental estimation of the effects of all amino-acid mutations to hiv?s envelope protein on viral replication in cell culture. *PLoS pathogens*, 12(12):e1006114, 2016.

[66] Michael J Harms and Joseph W Thornton. Historical contingency and its biophysical basis in glucocorticoid receptor evolution. *Nature*, 512(7513):203, 2014.

[67] Marion Hartl, Kristof Theys, Alison Feder, Maoz Gelbart, Adi Stern, and Pleuni S Pennings. Within-patient hiv mutation frequencies reveal fitness costs of cpg dinucleotides, drastic amino acid changes and g$\rightarrow$a mutations. *bioRxiv*, page 057026, 2016.

[68] Joseph B Hiatt, Rupali P Patwardhan, Emily H Turner, Choli Lee, and Jay Shendure. Parallel, tag-directed assembly of locally derived short sequence reads. *Nat. Methods*, 7(2):119–122, 2010.

[69] Sarah K Hilton, Michael B Doud, and Jesse D Bloom. phydms: Software for phylogenetic analyses informed by deep mutational scanning. *bioRxiv*, page 121830, 2017.

[70] Ya-Chi Ho, Liang Shan, Nina N Hosmane, Jeffrey Wang, Sarah B Laskey, Daniel IS Rosenbloom, Jun Lai, Joel N Blankson, Janet D Siliciano, and Robert F Siliciano.

Replication-competent noninduced proviruses in the latent reservoir increase barrier to hiv-1 cure. *Cell*, 155(3):540–551, 2013.

[71] Wei-Shau Hu and Howard M Temin. Genetic consequences of packaging two rna genomes in one retroviral particle: pseudodiploidy and high rate of genetic recombination. *Proceedings of the National Academy of Sciences*, 87(4):1556–1560, 1990.

[72] Jinghe Huang, Byong H Kang, Marie Pancera, Jeong Hyun Lee, Tommy Tong, Yu Feng, Ivelin S Georgiev, Gwo-Yu Chuang, Aliaksandr Druz, Nicole A Doria-Rose, et al. Broad and potent hiv-1 neutralization by a human antibody that binds the gp41-120 interface. *Nature*, 515(7525):138, 2014.

[73] Cassandra B Jabara, Corbin D Jones, Jeffrey Roach, Jeffrey A Anderson, and Ronald Swanstrom. Accurate sampling and deep sequencing of the HIV-1 protease gene using a primer ID. *Proceedings of the National Academy of Sciences*, 108(50):20166–20171, 2011.

[74] Amy Jacobs, Jayita Sen, Lijun Rong, and Michael Caffrey. Alanine scanning mutants of the hiv gp41 loop. *Journal of Biological Chemistry*, 280(29):27284–27288, 2005.

[75] Joseph G Jardine, Daniel W Kulp, Colin Havenar-Daughton, Anita Sarkar, Bryan Briney, Devin Sok, Fabian Sesterhenn, June Ereño-Orbea, Oleksandr Kalyuzhniy, Isaiah Deresa, et al. Hiv-1 broadly neutralizing antibody precursor b cells revealed by germline-targeting immunogen. *Science*, 351(6280):1458–1463, 2016.

[76] Amanda E Jetzt, Hong Yu, George J Klarmann, Yacov Ron, Bradley D Preston, and Joseph P Dougherty. High rate of recombination throughout the human immunodeficiency virus type 1 genome. *Journal of virology*, 74(3):1234–1240, 2000.

[77] Welkin E Johnson and Ronald C Desrosiers. Viral persistence: Hiv's strategies of immune system evasion. *Annual review of medicine*, 53(1):499–518, 2002.

[78] Jean-Philippe Julien, Albert Cupo, Devin Sok, Robyn L Stanfield, Dmitry Lyumkis, Marc C Deller, Per-Johan Klasse, Dennis R Burton, Rogier W Sanders, John P Moore, et al. Crystal structure of a soluble cleaved hiv-1 envelope trimer. *Science*, 342(6165):1477–1483, 2013.

[79] David Kabat, Susan L Kozak, Kathy Wehrly, and Bruce Chesebro. Differences in cd4 dependence for infectivity of laboratory-adapted and primary patient isolates of human immunodeficiency virus type 1. *Journal of virology*, 68(4):2570–2577, 1994.

[80] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[81] Isaac Kinde, Jian Wu, Nick Papadopoulos, Kenneth W Kinzler, and Bert Vogelstein. Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences*, 108(23):9530–9535, 2011.

[82] Jacob O Kitzman, Lea M Starita, Russell S Lo, Stanley Fields, and Jay Shendure. Massively parallel single-amino-acid mutagenesis. *Nature methods*, 12(3):203–206, 2015.

[83] Florian Klein, Ron Diskin, Johannes F Scheid, Christian Gaebler, Hugo Mouquet, Ivelin S Georgiev, Marie Pancera, Tongqing Zhou, Reha-Baris Incesu, Brooks Zhongzheng Fu, et al. Somatic mutations of the immunoglobulin framework are generally required for broad and potent hiv-1 neutralization. *Cell*, 153(1):126–138, 2013.

[84] Bette Korber, Brian T Foley, C Kuiken, Satish K Pillai, Joseph G Sodroski, et al. Numbering positions in hiv relative to hxb2cg. *Human retroviruses and AIDS*, 3:102–111, 1998.

[85] Bette Korber, Brian Gaschen, Karina Yusim, Rama Thakallapally, Can Kesmir, and Vincent Detours. Evolutionary and immunological implications of contemporary hiv-1 variation. *British medical bulletin*, 58(1):19–42, 2001.

[86] Bette Korber, M Muldoon, J Theiler, F Gao, R Gupta, A Lapedes, BH Hahn, S Wolinsky, and T Bhattacharya. Timing the ancestor of the hiv-1 pandemic strains. *science*, 288(5472):1789–1796, 2000.

[87] Roger D Kouyos, Gabriel E Leventhal, Trevor Hinkley, Mojgan Haddad, Jeannette M Whitcomb, Christos J Petropoulos, and Sebastian Bonhoeffer. Exploring the complexity of the hiv-1 fitness landscape. *PLoS Genet*, 8(3):e1002551, 2012.

[88] Peter D Kwong, John R Mascola, and Gary J Nabel. Broadly neutralizing antibodies and the search for an hiv-1 vaccine: the end of the beginning. *Nature Reviews Immunology*, 13(9):693–701, 2013.

[89] Peter D Kwong, Richard Wyatt, James Robinson, Raymond W Sweet, Joseph Sodroski, and Wayne A Hendrickson. Structure of an hiv gp120 envelope glycoprotein in complex with the cd4 receptor and a neutralizing human antibody. *Nature*, 393(6686):648–659, 1998.

[90] Jack Kyte and Russell F Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–132, 1982.

[91] Kevin E Langergraber, Kay Prüfer, Carolyn Rowney, Christophe Boesch, Catherine Crockford, Katie Fawcett, Eiji Inoue, Miho Inoue-Muruyama, John C Mitani, Martin N Muller, et al. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proceedings of the National Academy of Sciences*, 109(39):15716–15721, 2012.

[92] Benhur Lee, Matthew Sharron, Luis J Montaner, Drew Weissman, and Robert W Doms. Quantification of cd4, ccr5, and cxcr4 levels on lymphocyte subsets, den-

dritic cells, and differentially conditioned monocyte-derived macrophages. *Proceedings of the National Academy of Sciences*, 96(9):5215–5220, 1999.

[93] Woan-Ruoh Lee, Wan-Jr Syu, Bin Du, Michiko Matsuda, Shencao Tan, Andrea Wolf, Max Essex, and Tun Hou Lee. Nonrandom distribution of gp120 n-linked glycosylation sites important for infectivity of human immunodeficiency virus type 1. *Proceedings of the National Academy of Sciences*, 89(6):2213–2217, 1992.

[94] Cordelia K Leonard, Michael W Spellman, Lavon Riddle, Reed J Harris, James N Thomas, and TJ Gregory. Assignment of intrachain disulfide bonds and characterization of potential glycosylation sites of the type 1 recombinant human immunodeficiency virus envelope glycoprotein (gp120) expressed in chinese hamster ovary cells. *Journal of Biological Chemistry*, 265(18):10373–10382, 1990.

[95] Jay A Levy. Pathogenesis of human immunodeficiency virus infection. *Microbiological reviews*, 57(1):183–289, 1993.

[96] Yan Li, Lizhong Luo, David Y Thomas, and O Yong Kang. Control of expression, glycosylation, and secretion of hiv-1 gp120 by homologous and heterologous signal sequences. *Virology*, 204(1):266–278, 1994.

[97] Yuxing Li, Sijy O'Dell, Laura M Walker, Xueling Wu, Javier Guenaga, Yu Feng, Stephen D Schmidt, Krisha McKee, Mark K Louder, Julie E Ledgerwood, et al. Mechanism of neutralization by the broadly neutralizing hiv-1 monoclonal antibody vrc01. *Journal of virology*, 85(17):8954–8967, 2011.

[98] Min Lu, Marisa O Stoller, Shilong Wang, Jie Liu, Melinda B Fagan, and Jack H Nunberg. Structural and functional analysis of interhelical interactions in the human immunodeficiency virus type 1 gp41 envelope glycoprotein by alanine-scanning mutagenesis. *Journal of virology*, 75(22):11146–11156, 2001.

[99] Mark Lunzer, G Brian Golding, and Antony M Dean. Pervasive cryptic epistasis in molecular evolution. *PLoS genetics*, 6(10):e1001162, 2010.

[100] Rebecca M Lynch, Patrick Wong, Lillian Tran, Sijy O'Dell, Martha C Nason, Yuxing Li, Xueling Wu, and John R Mascola. Hiv-1 fitness cost associated with escape from the vrc01 class of cd4 binding site neutralizing antibodies. *Journal of virology*, 89(8):4201–4213, 2015.

[101] Dmitry Lyumkis, Jean-Philippe Julien, Natalia de Val, Albert Cupo, Clinton S Potter, Per-Johan Klasse, Dennis R Burton, Rogier W Sanders, John P Moore, Bridget Carragher, et al. Cryo-em structure of a fully glycosylated soluble cleaved hiv-1 envelope trimer. *Science*, 342(6165):1484–1490, 2013.

[102] Michael H Malim, Joachim Hauber, Shu-Yun Le, Jacob V Maizel, and Bryan R Cullen. The hiv-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mrna. *Nature*, 338(6212):254–257, 1989.

[103] Louis M Mansky and Howard M Temin. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *Journal of virology*, 69(8):5087–5094, 1995.

[104] David M Margolis, Richard A Koup, and Guido Ferrari. Hiv antibodies for treatment of hiv infection. *Immunological Reviews*, 275(1):313–323, 2017.

[105] David Mavor, James Fraser, et al. Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. *eLife*, 5:e15802, 2016.

[106] Richard N McLaughlin Jr, Frank J Poelwijk, Arjun Raman, Walraj S Gosal, and Rama Ranganathan. The spatial architecture of protein function and adaptation. *Nature*, 491(7422):138–142, 2012.

[107] Alexandre Melnikov, Peter Rogov, Li Wang, Andreas Gnirke, and Tarjei S Mikkelsen. Comprehensive mutational scanning of a kinase in vivo reveals

substrate-dependent fitness landscapes. *Nucleic Acids Research*, 42(14):e112, 2014.

[108] Austin G Meyer and Claus O Wilke. The utility of protein structure as a predictor of site-wise dn/ds varies widely among hiv-1 proteins. *Journal of The Royal Society Interface*, 12(111):20150579, 2015.

[109] Tarjei Mikkelsen, LaDeana Hillier, Evan Eichler, Michael Zody, David Jaffe, Shiaw-Pyng Yang, Wolfgang Enard, Ines Hellmann, Kerstin Lindblad-Toh, Tasha Altheide, et al. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87, 2005.

[110] Parul Mishra, Julia M Flynn, Tyler N Starr, and Daniel NA Bolon. Systematic mutant analyses elucidate general and client-specific aspects of hsp90 function. *Cell reports*, 15(3):588–598, 2016.

[111] Susanne Modrow, BEATRICE H Hahn, GEORGE M Shaw, ROBERT C Gallo, F Wong-Staal, and H Wolf. Computer-assisted analysis of envelope protein sequences of seven human immunodeficiency virus isolates: prediction of antigenic epitopes in conserved and variable regions. *Journal of virology*, 61(2):570–578, 1987.

[112] John P Moore, Yunzhen Cao, Limo Qing, Quentin J Sattentau, Javashree Pyati, Raju Koduri, James Robinson, CF 3rd Barbas, Dennis R Burton, and David D Ho. Primary isolates of human immunodeficiency virus type 1 are relatively resistant to neutralization by monoclonal antibodies to gp120, and their neutralization is not predicted by studies with monomeric gp120. *Journal of virology*, 69(1):101–109, 1995.

[113] Penny L Moore, Emma T Crooks, Lauren Porter, Ping Zhu, Charmagne S Cayanan, Henry Grise, Paul Corcoran, Michael B Zwick, Michael Franti, Lynn Morris, et al. Na-

ture of nonfunctional envelope proteins on the surface of human immunodeficiency virus type 1. *Journal of virology*, 80(5):2515–2528, 2006.

[114] Penny L Moore, Elin S Gray, and Lynn Morris. Specificity of the autologous neutralizing antibody response. *Current opinion in HIV and AIDS*, 4(5):358, 2009.

[115] Tsutomu Murakami. Roles of the interactions between env and gag proteins in the hiv-1 replication cycle. *Microbiology and immunology*, 52(5):287–295, 2008.

[116] Chandrasekhar Natarajan, Noriko Inoguchi, Roy E Weber, Angela Fago, Hideaki Moriyama, and Jay F Storz. Epistasis among adaptive mutations in deer mouse hemoglobin. *Science*, 340(6138):1324–1327, 2013.

[117] Rasmus Nielsen and Ziheng Yang. Likelihood models for detecting positively selected amino acid sites and applications to the hiv-1 envelope gene. *Genetics*, 148(3):929–936, 1998.

[118] Shinji Ohgimoto, Tatsuo Shioda, Kazuyasu Mori, Emi E Nakayama, Huiling Hu, and Yoshiyuki Nagai. Location-specific, unequal contribution of the n glycans in simian immunodeficiency virus gp120 to viral infectivity and removal of multiple glycans without disturbing infectivity. *Journal of virology*, 72(10):8365–8370, 1998.

[119] U Olshevsky, E Helseth, C Furman, J Li, W Haseltine, and J Sodroski. Identification of individual human immunodeficiency virus type 1 gp120 amino acids important for cd4 receptor binding. *Journal of virology*, 64(12):5701–5707, 1990.

[120] C Anders Olson, Nicholas C Wu, and Ren Sun. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Current Biology*, 24(22):2643–2651, 2014.

[121] Eric A Ortlund, Jamie T Bridgham, Matthew R Redinbo, and Joseph W Thornton. Crystal structure of an ancient protein: evolution by conformational epistasis. *Science*, 317(5844):1544–1548, 2007.

[122] Julie Overbaugh and Lynn Morris. The antibody response against hiv-1. *Cold Spring Harbor perspectives in medicine*, 2(1):a007039, 2012.

[123] Marie Pancera, Shahzad Majeed, Yih-En Andrew Ban, Lei Chen, Chih-chin Huang, Leopold Kong, Young Do Kwon, Jonathan Stuckey, Tongqing Zhou, James E Robinson, et al. Structure of hiv-1 gp120 with gp41-interactive region reveals layered envelope architecture and basis of conformational mobility. *Proceedings of the National Academy of Sciences*, 107(3):1166–1171, 2010.

[124] Marie Pancera, Tongqing Zhou, Aliaksandr Druz, Ivelin S Georgiev, Cinque Soto, Jason Gorman, Jinghe Huang, Priyamvada Acharya, Gwo-Yu Chuang, Gilad Ofek, et al. Structure and immune recognition of trimeric pre-fusion hiv-1 env. *Nature*, 514(7523):455–461, 2014.

[125] Ralph Pantophlet, Erica Ollmann Saphire, Pascal Poignard, Paul WHI Parren, Ian A Wilson, and Dennis R Burton. Fine mapping of the interaction of neutralizing and nonneutralizing monoclonal antibodies with the cd4 binding site of human immunodeficiency virus type 1 gp120. *Journal of virology*, 77(1):642–658, 2003.

[126] Joanna L Parmley, JV Chamary, and Laurence D Hurst. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Molecular biology and evolution*, 23(2):301–309, 2006.

[127] Keith Peden, Michael Emerman, and Luc Montagnier. Changes in growth properties on passage in tissue culture of viruses derived from infectious molecular clones of hiv-1 lai, hiv-1 mal, and hiv-1 eli. *Virology*, 185(2):661–672, 1991.

[128] Alan S Perelson, Avidan U Neumann, Martin Markowitz, John M Leonard, David D Ho, et al. Hiv-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science*, 271(5255):1582–1586, 1996.

[129] Anna I Podgornaia and Michael T Laub. Pervasive degeneracy and epistasis in a protein-protein interface. *Science*, 347(6222):673–677, 2015.

[130] David D Pollock, Grant Thiltgen, and Richard A Goldstein. Amino acid coevolution induces an evolutionary stokes shift. *Proceedings of the National Academy of Sciences*, 109(21):E1352–E1359, 2012.

[131] Mary Poss and Julie Overbaugh. Variants from the diverse virus population identified at seroconversion of a clade a human immunodeficiency virus type 1-infected woman have distinct biological properties. *Journal of virology*, 73(7):5255–5264, 1999.

[132] Pavel Pugach, Shawn E Kuhmann, Joann Taylor, Andre J Marozsan, Amy Snyder, Thomas Ketas, Steven M Wolinsky, Bette T Korber, and John P Moore. The prolonged culture of human immunodeficiency virus type 1 in primary lymphocytes increases its sensitivity to neutralization by soluble cd4. *Virology*, 321(1):8–22, 2004.

[133] Hangfei Qi, C Anders Olson, Nicholas C Wu, Ruian Ke, Claude Loverdo, Virginia Chu, Shawna Truong, Roland Remenyi, Zugen Chen, Yushen Du, et al. A quantitative high-resolution genetic profile rapidly identifies sequence determinants of hepatitis c viral fitness and drug sensitivity. *PLoS Pathog*, 10(4):e1004064, 2014.

[134] Andrew Rambaut, David Posada, Keith A Crandall, and Edward C Holmes. The causes and consequences of hiv evolution. *Nature reviews. Genetics*, 5(1):52, 2004.

[135] D C Ramsey, M P Scherrer, T Zhou, and C O Wilke. The relationship between relative solvent accessibility and evolutionary rate in protein evolution. *Genetics*, 188:479–488, 2011.

[136] Lowell Jacob Reed and Hugo Muench. A simple method of estimating fifty per cent endpoints. *American journal of epidemiology*, 27(3):493–497, 1938.

[137] Eric W Refsland, Mark D Stenglein, Keisuke Shindo, John S Albin, William L Brown, and Reuben S Harris. Quantitative profiling of the full apobec3 mrna repertoire in lymphocytes and tissues: implications for hiv-1 restriction. *Nucleic acids research*, 38(13):4274–4284, 2010.

[138] Jane S Richardson and David C Richardson. Amino acid preferences for specific locations at the ends of alpha helices. *Science*, 240(4859):1648, 1988.

[139] Douglas D Richman, Terri Wrin, Susan J Little, and Christos J Petropoulos. Rapid evolution of the neutralizing antibody response to hiv type 1 infection. *Proceedings of the National Academy of Sciences*, 100(7):4144–4149, 2003.

[140] Carlo D Rizzuto, Richard Wyatt, Nivia Hernández-Ramos, Ying Sun, Peter D Kwong, Wayne A Hendrickson, and Joseph Sodroski. A conserved hiv gp120 glycoprotein structure involved in chemokine receptor binding. *Science*, 280(5371):1949–1953, 1998.

[141] Philip A Romero, Tuan M Tran, and Adam R Abate. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proceedings of the National Academy of Sciences*, 112(23):7159–7164, 2015.

[142] Benjamin P Roscoe, Kelly M Thayer, Konstantin B Zeldovich, David Fushman, and Daniel NA Bolon. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *Journal of molecular biology*, 425(8):1363–1377, 2013.

[143] Rogier W Sanders, Ronald Derking, Albert Cupo, Jean-Philippe Julien, Anila Yasmeen, Natalia de Val, Helen J Kim, Claudia Blattner, Alba Torrents de la Peña, Jacob Korzun, et al. A next-generation cleaved, soluble hiv-1 env trimer, bg505

sosip. 664 gp140, expresses multiple epitopes for broadly neutralizing but not non-neutralizing antibodies. *PLoS pathogens*, 9(9):e1003618, 2013.

[144] Rogier W Sanders, Marit J Van Gils, Ronald Derking, Devin Sok, Thomas J Ketas, Judith A Burger, Gabriel Ozorowski, Albert Cupo, Cassandra Simonich, Leslie Goo, et al. Hiv-1 neutralizing antibodies induced by native-like envelope trimers. *Science*, 349(6244):aac4223, 2015.

[145] Rogier W Sanders, Mika Vesanen, Norbert Schuelke, Aditi Master, Linnea Schiffner, Roopa Kalyanaraman, Maciej Paluch, Ben Berkhout, Paul J Maddon, William C Olson, et al. Stabilization of the soluble, cleaved, trimeric form of the envelope glycoprotein complex of human immunodeficiency virus type 1. *Journal of virology*, 76(17):8875–8889, 2002.

[146] Louise Scharf, Johannes F Scheid, Jeong Hyun Lee, Anthony P West, Courtney Chen, Han Gao, Priyanthi NP Gnanapragasam, René Mares, Michael S Seaman, Andrew B Ward, et al. Antibody 8anc195 reveals a site of broad vulnerability on the hiv-1 envelope spike. *Cell reports*, 7(3):785–795, 2014.

[147] Sasha Shafikhani, RA Siegel, E Ferrari, and Volker Schellenberger. Generation of large libraries of random mutants in bacillus subtilis by pcr-based plasmid multimerization. *Biotechniques*, 23(2):304–311, 1997.

[148] RAJ Shankarappa, Joseph B Margolick, Stephen J Gange, Allen G Rodrigo, David Upchurch, Homayoon Farzadegan, Phalguni Gupta, Charles R Rinaldo, Gerald H Learn, XI He, et al. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *Journal of virology*, 73(12):10489–10502, 1999.

[149] George M Shaw and Beatrice H Hahn Suresh K Arya. Molecular characterization

of human t-cell leukemia (lymphotropic) virus type iii in the acquired immune deficiency syndrome. *Science l984*, 226:l, 1984.

[150] Ann M Sheehy, Nathan C Gaddis, Jonathan D Choi, and Michael H Malim. Isolation of a human gene that inhibits hiv-1 infection and is suppressed by the viral vif protein. *Nature*, 418(6898):646–650, 2002.

[151] Cassandra A Simonich, Katherine L Williams, Hans P Verkerke, James A Williams, Ruth Nduati, Kelly K Lee, and Julie Overbaugh. Hiv-1 neutralizing antibodies with limited hypermutation from an infant. *Cell*, 166(1):77–87, 2016.

[152] Richard D Sloan and Mark A Wainberg. The role of unintegrated dna in hiv infection. *Retrovirology*, 8(1):1, 2011.

[153] Devin Sok, Matthias Pauthner, Bryan Briney, Jeong Hyun Lee, Karen L Saye-Francisco, Jessica Hsueh, Alejandra Ramos, Khoa M Le, Meaghan Jones, Joseph G Jardine, et al. A prominent site of antibody vulnerability on hiv envelope incorporates a motif associated with ccr5 binding and its camouflaging glycans. *Immunity*, 45(1):31–45, 2016.

[154] Stephanie J Spielman and Claus O Wilke. Pyvolve: a flexible python module for simulating sequences along phylogenies. *PloS one*, 10(9):e0139047, 2015.

[155] Alexandros Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, page btu033, 2014.

[156] Bruno R Starcich, Beatrice H Hahn, George M Shaw, Paul D McNeely, Susanne Modrow, Hans Wolf, Elizabeth S Parks, Wade P Parks, Steven F Josephs, Robert C Gallo, et al. Identification and characterization of conserved and variable regions in the envelope gene of htlv-iii/lav, the retrovirus of aids. *Cell*, 45(5):637–648, 1986.

[157] Lea M Starita, Jonathan N Pruneda, Russell S Lo, Douglas M Fowler, Helen J Kim, Joseph B Hiatt, Jay Shendure, Peter S Brzovic, Stanley Fields, and Rachel E Klevit.

Activity-enhancing mutations in an e3 ubiquitin ligase identified by high-throughput mutagenesis. *Proceedings of the National Academy of Sciences*, 110(14):E1263–E1272, 2013.

[158] Lea M Starita, David L Young, Muhtadi Islam, Jacob O Kitzman, Justin Gullingsrud, Ronald J Hause, Douglas M Fowler, Jeffrey D Parvin, Jay Shendure, and Stanley Fields. Massively parallel functional analysis of brca1 ring domain variants. *Genetics*, 200(2):413–422, 2015.

[159] Jon M Steichen, Daniel W Kulp, Talar Tokatlian, Amelia Escolano, Pia Dosenovic, Robyn L Stanfield, Laura E McCoy, Gabriel Ozorowski, Xiaozhen Hu, Oleksandr Kalyuzhniy, et al. Hiv vaccine design to target germline precursors of glycan-dependent broadly neutralizing antibodies. *Immunity*, 45(3):483–496, 2016.

[160] Guillaume BE Stewart-Jones, Cinque Soto, Thomas Lemmin, Gwo-Yu Chuang, Aliaksandr Druz, Rui Kong, Paul V Thomas, Kshitij Wagh, Tongqing Zhou, Anna-Janina Behrens, et al. Trimeric hiv-1-env structures define glycan shields from clades a, b, and g. *Cell*, 165(4):813–826, 2016.

[161] Michael A Stiffler, Doeke R Hekstra, and Rama Ranganathan. Evolvability as a function of purifying selection in tem-1 β-lactamase. *Cell*, 160(5):882–892, 2015.

[162] Arvind R Subramaniam, Aaron DeLoughery, Niels Bradshaw, Yun Chen, Erin OShea, Richard Losick, and Yunrong Chai. A serine sensor for multicellularity in a bacterium. *Elife*, 2:e01501, 2013.

[163] Nancy Sullivan, Ying Sun, John Li, Wolfgang Hofmann, and Joseph Sodroski. Replicative function and neutralization sensitivity of envelope glycoproteins from primary and t-cell line-passaged human immunodeficiency virus type 1 isolates. *Journal of virology*, 69(7):4413–4422, 1995.

[164] Bargavi Thyagarajan and Jesse D Bloom. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *eLife*, 3:e03300, 2014.

[165] Matthew Z Tien, Austin G Meyer, Dariya K Sydykova, Stephanie J Spielman, and Claus O Wilke. Maximum allowed solvent accessibilites of residues in proteins. *PloS one*, 8(11):e80635, 2013.

[166] Wouter G Touw, Coos Baakman, Jon Black, Tim AH te Beek, Elmar Krieger, Robbie P Joosten, and Gert Vriend. A series of pdb-related databanks for everyday needs. *Nucleic acids research*, page gku1028, 2014.

[167] Alexandra Trkola, Tatjana Dragic, James Arthos, James M Binley, et al. Cd4-dependent, antibody-sensitive interactions between hiv-1 and its co-receptor ccr-5. *Nature*, 384(6605):184, 1996.

[168] Eelco van Anken, Rogier W Sanders, I Marije Liscaljet, Aafke Land, Ilja Bontjer, Sonja Tillemans, Alexey A Nabatov, William A Paxton, Ben Berkhout, and Ineke Braakman. Only five of 10 strictly conserved disulfide bonds are essential for folding and eight for function of the hiv-1 envelope glycoprotein. *Molecular biology of the cell*, 19(10):4298–4309, 2008.

[169] Simon Wain-Hobson, Jean-Pierre Vartanian, Michel Henry, Nicole Chenciner, Rémi Cheynier, Sylvie Delassus, L Pedroza Martins, Monica Sala, Marie-Thérèse Nugeyre, Denise Guétard, et al. Lav revisited: origins of the early hiv-1 isolates from institut pasteur. *Science*, 252(5008):961–965, 1991.

[170] Wei-Kung Wang, Max Essex, and Tun-Hou Lee. Single amino acid substitution in constant region 1 or 4 of gp120 causes the phenotype of a human immunodeficiency virus type 1 variant with mutations in hypervariable regions 1 and 2 to revert. *Journal of virology*, 70(1):607–611, 1996.

[171] Wenbo Wang, Jianhui Nie, Courtney Prochnow, Carolyn Truong, Zheng Jia, Suting Wang, Xiaojiang S Chen, and Youchun Wang. A systematic study of the n-glycosylation sites of hiv-1 envelope protein on infectivity and antibody-mediated neutralization. *Retrovirology*, 10(1):1, 2013.

[172] Joseph M Watts, Kristen K Dang, Robert J Gorelick, Christopher W Leonard, Julian W Bess Jr, Ronald Swanstrom, Christina L Burch, and Kevin M Weeks. Architecture and secondary structure of an entire hiv-1 rna genome. *Nature*, 460(7256):711–716, 2009.

[173] Xiping Wei, Julie M Decker, Hongmei Liu, Zee Zhang, Ramin B Arani, J Michael Kilby, Michael S Saag, Xiaoyun Wu, George M Shaw, and John C Kappes. Emergence of resistant human immunodeficiency virus type 1 in patients receiving fusion inhibitor (t-20) monotherapy. *Antimicrobial agents and chemotherapy*, 46(6):1896–1905, 2002.

[174] Xiping Wei, Julie M Decker, Shuyi Wang, Huxiong Hui, John C Kappes, Xiaoyun Wu, Jesus F Salazar-Gonzalez, Maria G Salazar, J Michael Kilby, Michael S Saag, et al. Antibody neutralization and escape by hiv-1. *Nature*, 422(6929):307–312, 2003.

[175] Daniel M Weinreich, Nigel F Delaney, Mark A DePristo, and Daniel L Hartl. Darwinian evolution can follow only very few mutational paths to fitter proteins. *science*, 312(5770):111–114, 2006.

[176] Timothy A Whitehead, Aaron Chevalier, Yifan Song, Cyrille Dreyfus, Sarel J Fleishman, Cecilia De Mattos, Chris A Myers, Hetunandan Kamisetty, Patrick Blair, Ian A Wilson, et al. Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nature biotechnology*, 30(6):543–548, 2012.

[177] Carl T Wild, Diane C Shugars, Teresa K Greenwell, Charlene B McDanal, and

Thomas J Matthews. Peptides corresponding to a predictive alpha-helical domain of human immunodeficiency virus type 1 gp41 are potent inhibitors of virus infection. *Proceedings of the National Academy of Sciences*, 91(21):9770–9774, 1994.

[178] Nicholas C Wu, C Anders Olson, Yushen Du, Shuai Le, Kevin Tran, Roland Remenyi, Danyang Gong, Laith Q Al-Mawsawi, Hangfei Qi, Ting-Ting Wu, et al. Functional constraint profiling of a viral protein reveals discordance of evolutionary conservation and functionality. *PLoS Genet*, 11(7):e1005310, 2015.

[179] Nicholas C Wu, Arthur P Young, Laith Q Al-Mawsawi, C Anders Olson, Jun Feng, Hangfei Qi, Shu-Hwa Chen, I-Hsuan Lu, Chung-Yen Lin, Robert G Chin, et al. High-throughput profiling of influenza a virus hemagglutinin gene at single-nucleotide resolution. *Scientific reports*, 4:4942, 2014.

[180] Richard Wyatt, Peter D Kwong, Elizabeth Desjardins, Raymond W Sweet, et al. The antigenic structure of the hiv gp120 envelope glycoprotein. *Nature*, 393(6686):705, 1998.

[181] Richard Wyatt and Joseph Sodroski. The hiv-1 envelope glycoproteins: fusogens, antigens, and immunogens. *Science*, 280(5371):1884–1888, 1998.

[182] Karina Yusim, Can Kesmir, Brian Gaschen, Marylyn M Addo, Marcus Altfeld, Søren Brunak, Alexandre Chigaev, Vincent Detours, and Bette T Korber. Clustering patterns of cytotoxic t-lymphocyte epitopes in human immunodeficiency virus type 1 (hiv-1) proteins reveal imprints of immune evasion on hiv-1 global variation. *Journal of virology*, 76(17):8757–8768, 2002.

[183] Eloísa Yuste, Jacqueline D Reeves, Robert W Doms, and Ronald C Desrosiers. Modulation of env content in virions of simian immunodeficiency virus: correlation with cell surface expression and virion infectivity. *Journal of virology*, 78(13):6775–6785, 2004.

[184] Fabio Zanini, Johanna Brodin, Lina Thebo, Christa Lanz, Göran Bratt, Jan Albert, and Richard A Neher. Population genomics of intrapatient hiv-1 evolution. *eLife*, 4:e11282, 2015.

[185] Fabio Zanini and Richard A Neher. Quantifying selection against synonymous mutations in hiv-1 env evolution. *Journal of virology*, 87(21):11843–11850, 2013.

[186] Ming Zhang, Brian Gaschen, Wendy Blay, Brian Foley, Nancy Haigwood, Carla Kuiken, and Bette Korber. Tracking global patterns of n-linked glycosylation site variation in highly variable viral glycoproteins: Hiv, siv, and hcv envelopes and influenza hemagglutinin. *Glycobiology*, 14(12):1229–1246, 2004.

[187] Tian-Hao Zhang, Nicholas C Wu, and Ren Sun. A benchmark study on error-correction by read-pairing and tag-clustering in amplicon-based deep sequencing. *BMC genomics*, 17(1):1, 2016.

[188] Tongqing Zhou, Ivelin Georgiev, Xueling Wu, Zhi-Yong Yang, Kaifan Dai, Andrés Finzi, Young Do Kwon, Johannes F Scheid, Wei Shi, Ling Xu, et al. Structural basis for broad and potent neutralization of hiv-1 by antibody vrc01. *Science*, 329(5993):811–817, 2010.

[189] Tongqing Zhou, Rebecca M Lynch, Lei Chen, Priyamvada Acharya, Xueling Wu, Nicole A Doria-Rose, M Gordon Joyce, Daniel Lingwood, Cinque Soto, Robert T Bailer, et al. Structural repertoire of hiv-1-neutralizing antibodies targeting the cd4 supersite in 14 donors. *Cell*, 161(6):1280–1292, 2015.

[190] Jianling Zhuang, Amanda E Jetzt, Guoli Sun, Hong Yu, George Klarmann, Yacov Ron, Bradley D Preston, and Joseph P Dougherty. Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. *Journal of virology*, 76(22):11273–11282, 2002.

[191] Susan Zolla-Pazner and Timothy Cardozo. Structure–function relationships of hiv-1 envelope sequence-variable regions refocus vaccine design. *Nature reviews Immunology*, 10(7):527–535, 2010.

[192] Michael B Zwick, Richard Jensen, Sarah Church, Meng Wang, Gabriela Stiegler, Renate Kunert, Hermann Katinger, and Dennis R Burton. Anti-human immunodeficiency virus type 1 (hiv-1) antibodies 2f5 and 4e10 require surprisingly few crucial residues in the membrane-proximal external region of glycoprotein gp41 to neutralize hiv-1. *Journal of virology*, 79(2):1252–1261, 2005.

[193] Michael B Zwick, Aran F Labrijn, Meng Wang, Catherine Spenlehauer, Erica Ollmann Saphire, James M Binley, John P Moore, Gabriela Stiegler, Hermann Katinger, Dennis R Burton, et al. Broadly neutralizing antibodies targeted to the membrane-proximal external region of human immunodeficiency virus type 1 glycoprotein gp41. *Journal of virology*, 75(22):10892–10905, 2001.

LIST OF FIGURES

Figure Number Page

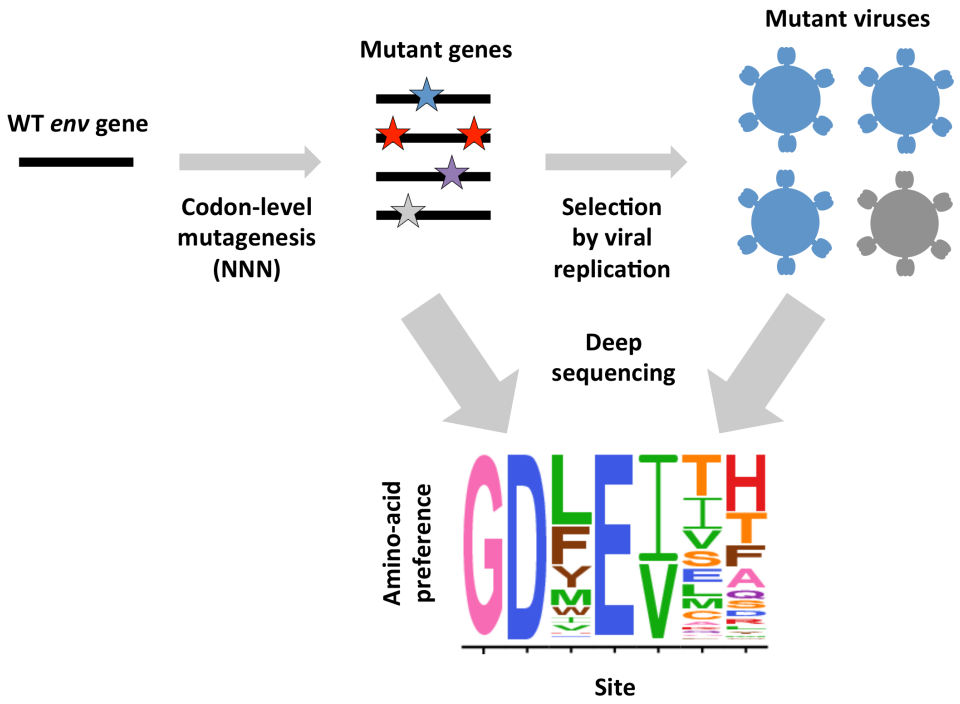

Figure 1: **Deep mutational scanning workflow, as applied to Env.** I used deep mutational scanning to quantify the effects of all single amino-acid mutations to Env. First, I introduced random codon-level mutations into a wildtype *env* gene. Next, I used these genes to generate mutant viruses. I then selected for functional variants by passaging these viruses in cell culture. Since Env is essential for viral replication, I expected this step to enrich for functional variants. Next, I deep sequenced the library before and after selection to quantify the change if frequency of each mutation. Finally, I used these data to infer the preference for each amino acid at each site in the protein. The results provide a comprehensive view of Env's mutational tolerance.

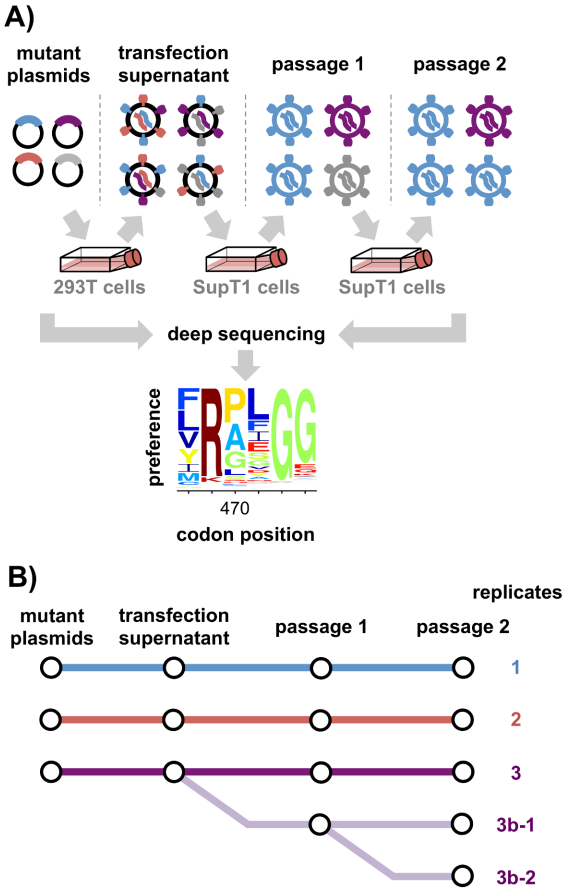

Figure 2: **Deep mutational scanning workflow. (A)** We created libraries of HIV proviral plasmids with random codon mutations in *env*, and generated mutant viruses by transfecting these plasmid libraries into 293T cells. Since cells receive multiple plasmids, there may not be a link between viral genotype and phenotype at this stage. To establish this link and select for functional variants, we passaged the viruses twice at low multiplicity of infection (MOI) in SupT1 cells. We deep sequenced *env* before and after selection to quantify the enrichment or depletion of each mutation, and used these data to estimate the preference of each site for each amino acid. Each mutant library was paired with a control in which cells were transfected with a wildtype HIV proviral plasmid to generate initially wildtype viruses that were passaged in parallel with the mutant viruses. Deep sequencing of these wildtype controls enabled estimation of the rates of apparent mutations arising from deep sequencing and viral replication. **(B)** We performed the entire experiment in triplicate. Additionally, we passaged the replicate-3 transfection supernatant in duplicate (replicate 3b). We also performed the second passage of replicate 3b in duplicate (replicates 3b-1 and 3b-2).

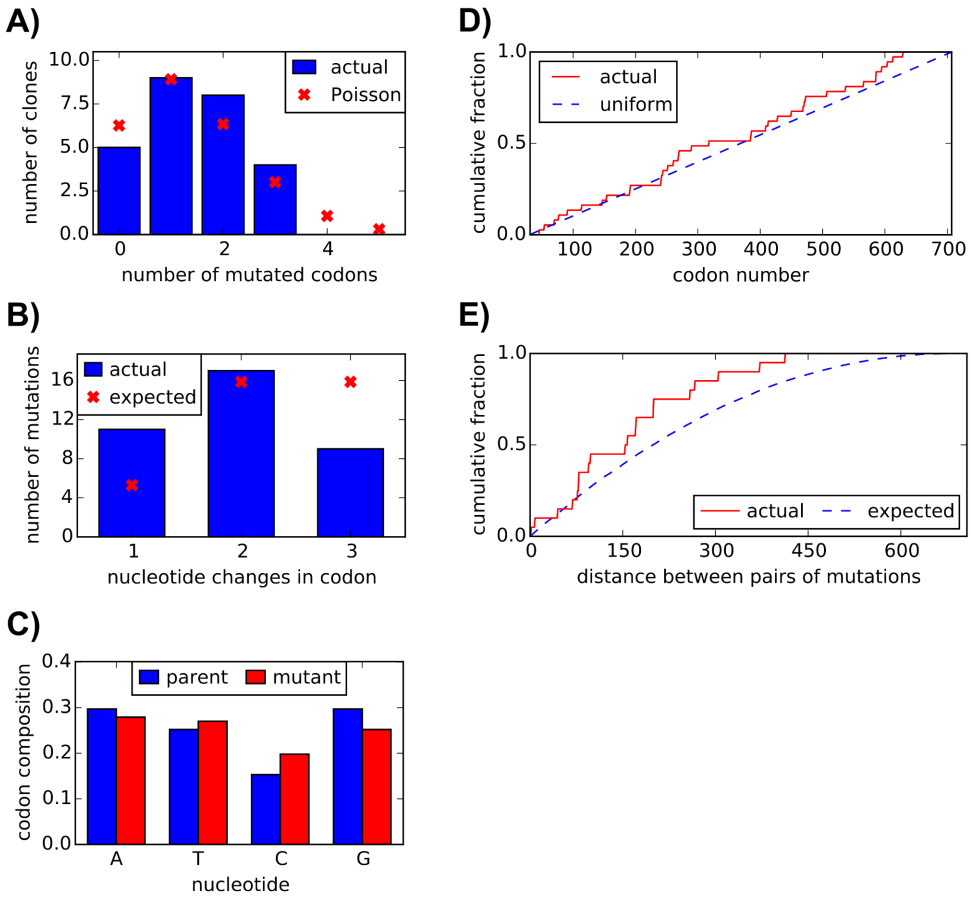

Figure 3: **Sanger sequencing of mutant plasmids shows a roughly uniform distribution of codon mutations, with an average of 1.4 mutations per gene.** We Sanger sequenced 26 clones sampled roughly evenly from the three replicate mutant plasmid libraries prior to any functional selection. **(A)** We observed an average of 1.4 mutant codons per clone. The number of mutant codons per clone closely followed a Poisson distribution. **(B)** Mutant codons had a mix of single-, double-, and triple-nucleotide changes. **(C)** The nucleotide frequencies were fairly uniform in the mutant codons. **(D)** Mutations were distributed roughly evenly along the portion of *env* that we mutagenized (codons 31-707). **(E)** For clones with multiple mutations, we computed pairwise distances between mutations in primary sequence and plotted the cumulative distribution of these distances (red line). For comparison, we simulated the expected distribution of pairwise distances if mutations occurred entirely independently (blue line). The difference between the actual and expected distributions suggests our mutagenesis had a slight bias to introduce mutations closer together than expected by chance.

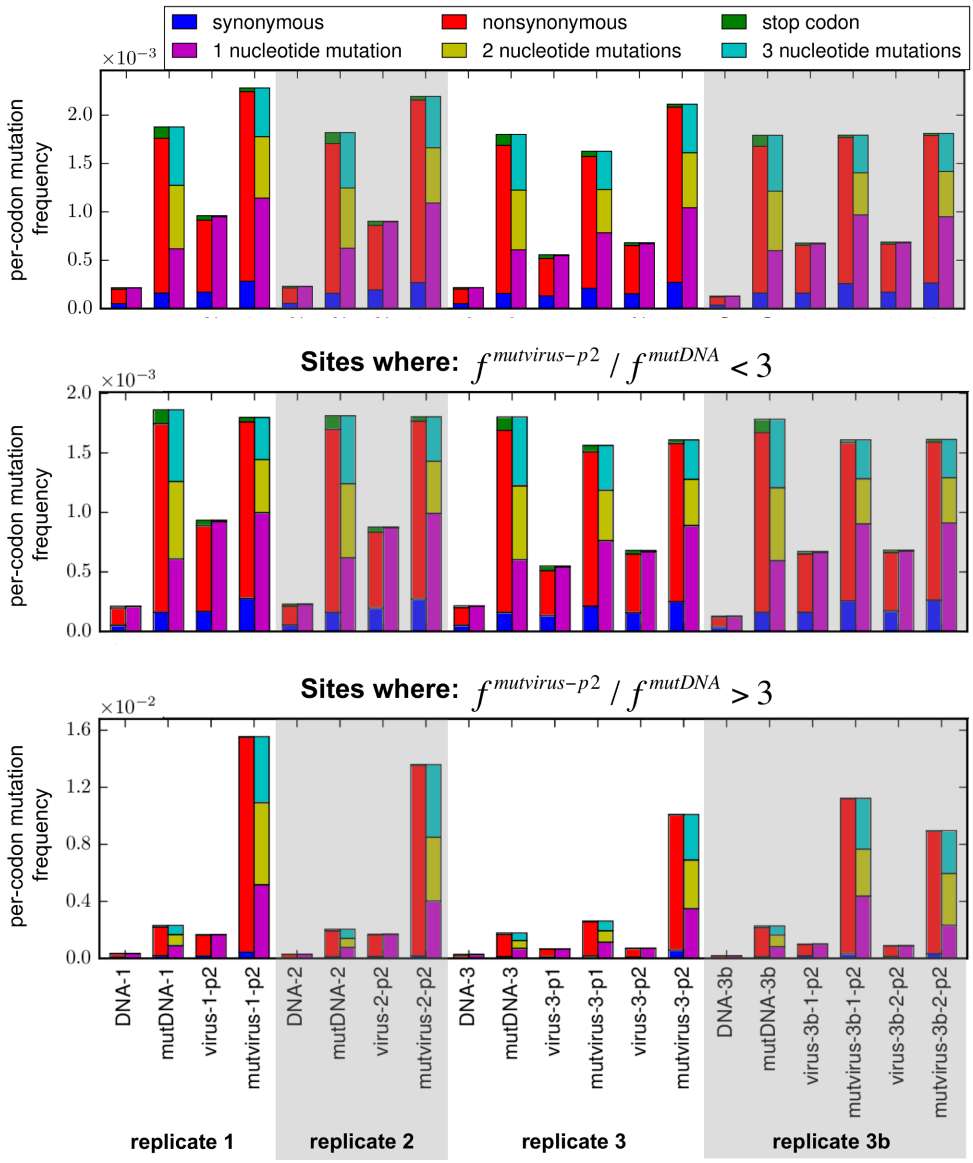

Figure 4: **Codon mutation frequencies of mutant libraries and wildtype controls.** This figure is similar to Fig 5 except that it shows the uncorrected mutation frequencies in the mutant plasmid and mutant virus libraries, and the mutation frequencies in the wildtype controls that were used to correct the mutation frequencies in Fig 5. Codon mutations are classified both by their effect on the protein (synonymous, nonsynonymous, or stop codon) and by the number of nucleotides they change in the codon (one, two, or three). The top panel shows data for all sites, whereas the middle and lower panels show data for the indicated subsets of sites. This this plot, $f^{mutvirus-p2}$ and f^{mutDNA} refer to the nonsynonymous mutation frequency in the twice-passaged mutant viruses and the initial mutant DNA, respectively.

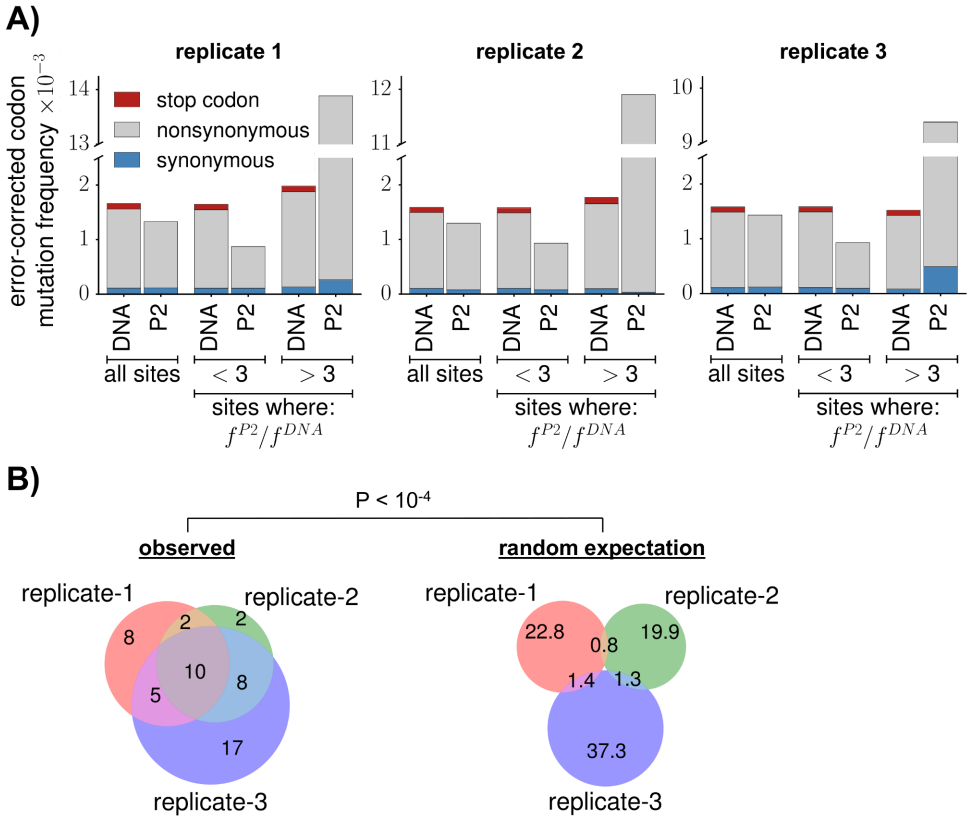

Figure 5: **Selection purged mutations in most of *env*, but favored mutations at a few sites. (A)** For each replicate, we deep sequenced the initial plasmids (DNA) and the viruses after two rounds of passaging (P2). Bars show the per-codon mutation frequency averaged across sites after subtracting error rates determined from the wildtype controls (Fig 4). When mutation frequencies are averaged across all sites, selection purged stop codons to <1% of their frequency in the initial DNA. Selection only slightly reduced the average frequency of nonsynonymous mutations; however, this average results from two distinct trends. For ≈4% of sites, the frequency of nonsynonymous mutations in the twice-passaged viruses (f^{P2}) increased >3-fold relative to the frequency in the initial plasmid DNA (f^{DNA}). For all other sites, the frequency of nonsynonymous mutations decreased substantially after selection. **(B)** The sites at which the error-corrected mutation frequency increased >3-fold are similar between replicates, indicating consistent selection for tissue-culture adaptation at a few positions. The left Venn diagram shows the overlap among replicates in the sites with a >3-fold increase. The right Venn diagram shows the expected overlap if the same number of sites per replicate are randomly drawn from Env's primary sequence. This difference is statistically significant, with $P < 10^{-4}$ when comparing the actual overlap among all three replicates to the random expectation. Another summary view of selection on *env* is provided by Fig 7 and Fig 8.

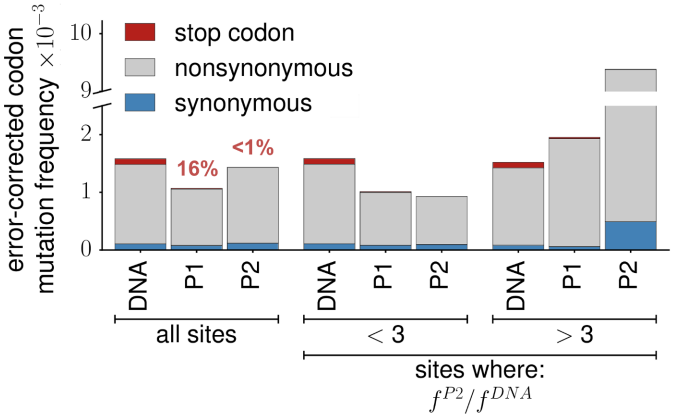

Figure 6: **Complete selection against stop codons requires two rounds of viral passage.** We deep sequenced the replicate 3 library after both one (P1) and two (P2) rounds of viral passaging. This figure is similar to Fig 5, but shows data for both P1 and P2. Purging of stop-codon mutations shows selection was only complete after two rounds of passaging. Whereas two rounds of passaging purged stop-codon mutations to $<$1% their frequency in the initial library (DNA), one round of passaging only purged stop-codon mutations to 16% their starting frequency (see the data for "all sites", where the red numbers above the bars for P1 and P2 indicate the percentage of stop codons after each passage relative to the starting library).

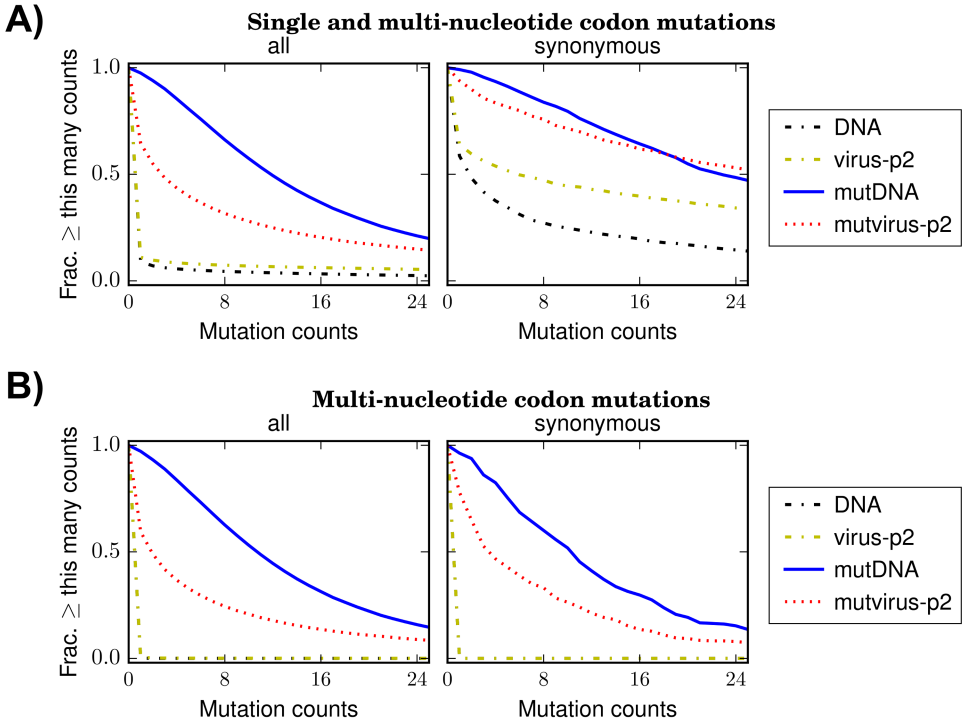

Figure 7: **Sampling of codon mutations in all replicates combined. (A)** Each plot shows the number of all (single and multi-nucleotide) codon mutations observed at least the indicated number of times in the sequencing of all replicates combined. We observed almost all mutations in the starting plasmid libraries (mutDNA), showing rich initial mutational diversity. Many mutations were depleted in the mutant virus libraries after two rounds of passaging (mutvirus-p2), consistent with purifying selection purging deleterious variants or bottlenecking diminishing library diversity. Examination of mutation counts in the wildtype plasmid (DNA) and wildtype virus (virus-p2) controls revealed a considerable fraction of mutations that were present at appreciable numbers due to errors from deep sequencing and PCR or *de novo* mutations from viral replication. This observation underscores the importance of using these wildtype controls to correct for background errors and *de novo* mutations. **(B)** If we examine only multi-nucleotide codon mutations, then there are negligible background errors in the wildtype controls. Similar data for each replicate individually are in Fig 8.

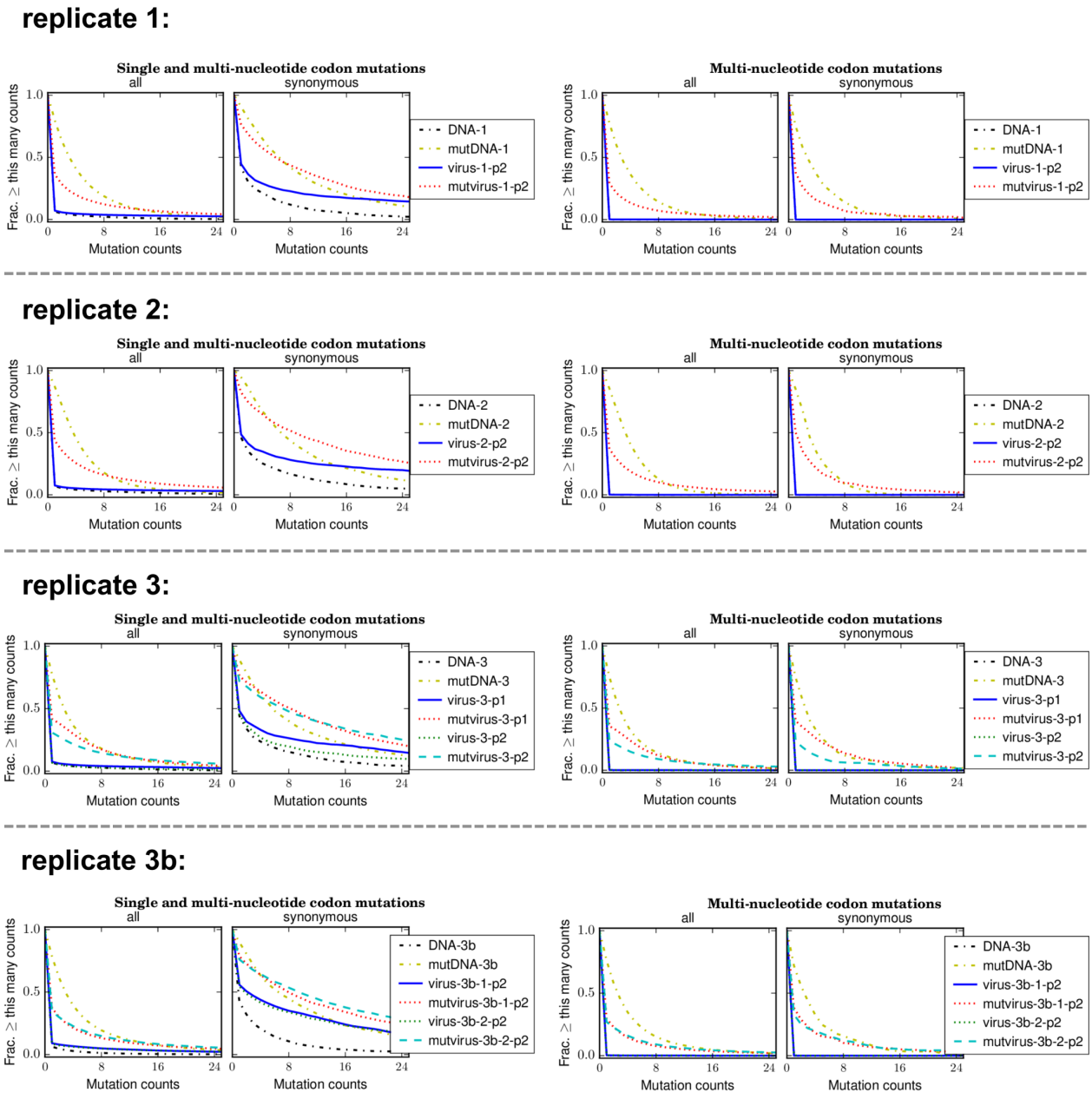

Figure 8: **Sampling of codon mutations in individual replicates.** These plots are the same as in Fig 7 but show each replicate individually.

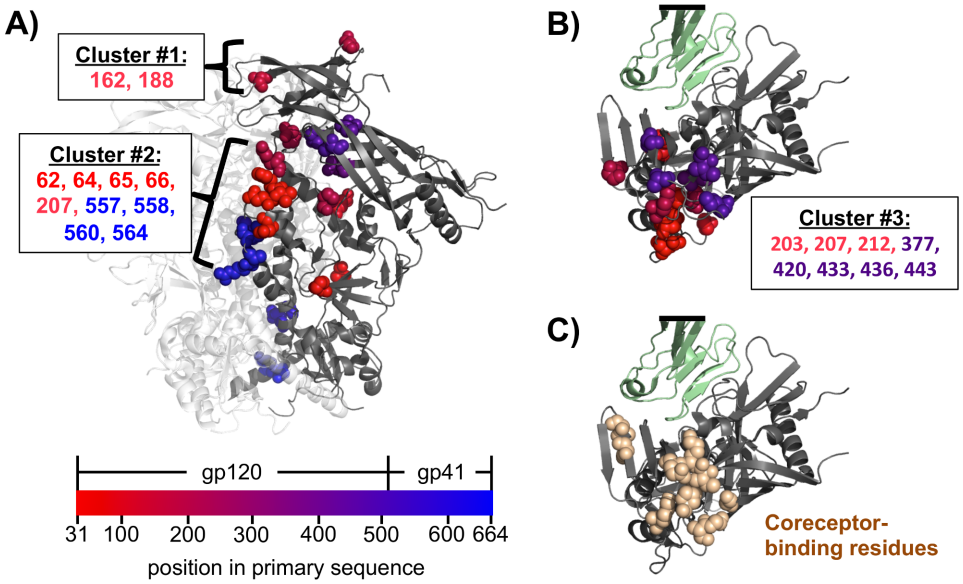

Figure 9: **Sites of recurrent cell-culture mutations mapped on Env's structure.** The 25 sites from Fig 5B where the mutation frequency increased >3-fold in at least two replicates after cell-culture passage. **(A)** Trimeric Env (PDB 5FYK [160]) with one monomer in grey and the others in white, oriented so the membrane-proximal region is at the bottom. Sites of cell-culture mutations are shown as spheres, colored red-to-blue according to primary sequence. Most of these sites fall in one of three clusters. Mutations in the first cluster disrupt potential glycosylation sites at Env's apex. The second cluster includes or is adjacent to sites where mutations are known to affect Env's conformational dynamics [145, 39]. **(B)** The third cluster is near the co-receptor binding surface. This panel shows an apex-down view of monomeric gp120 (grey) in complex with CD4 (green) (PDB 3JWD [123]). Sites of recurrent cell-culture mutations are shown as spheres colored according to primary sequence as in panel A. The black bar indicates cropping of CD4. **(C)** The same view as panel B, but the spheres now show sites known to affect binding to CXCR4 [10] or CCR5 [140]. Note the extensive overlap between the spheres in this panel and panel B.

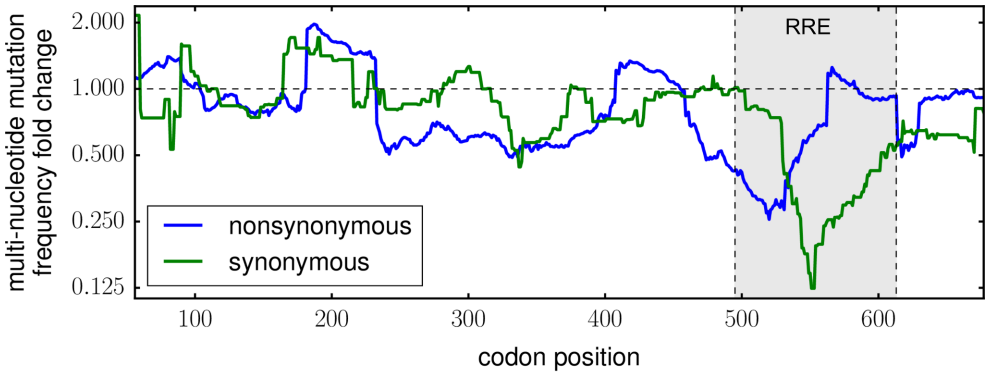

Figure 10: **Selection depleted multi-nucleotide codon mutations in the Rev-response element (RRE).** This plot shows a 51-codon sliding-window average of the fold change in per-codon multi-nucleotide mutation frequency after two rounds of viral passage, with data plotted for the center point in each window. The strongest depletion of both synonymous and nonsynonymous mutations occurred in the RRE, which is an RNA secondary structure important for viral replication.

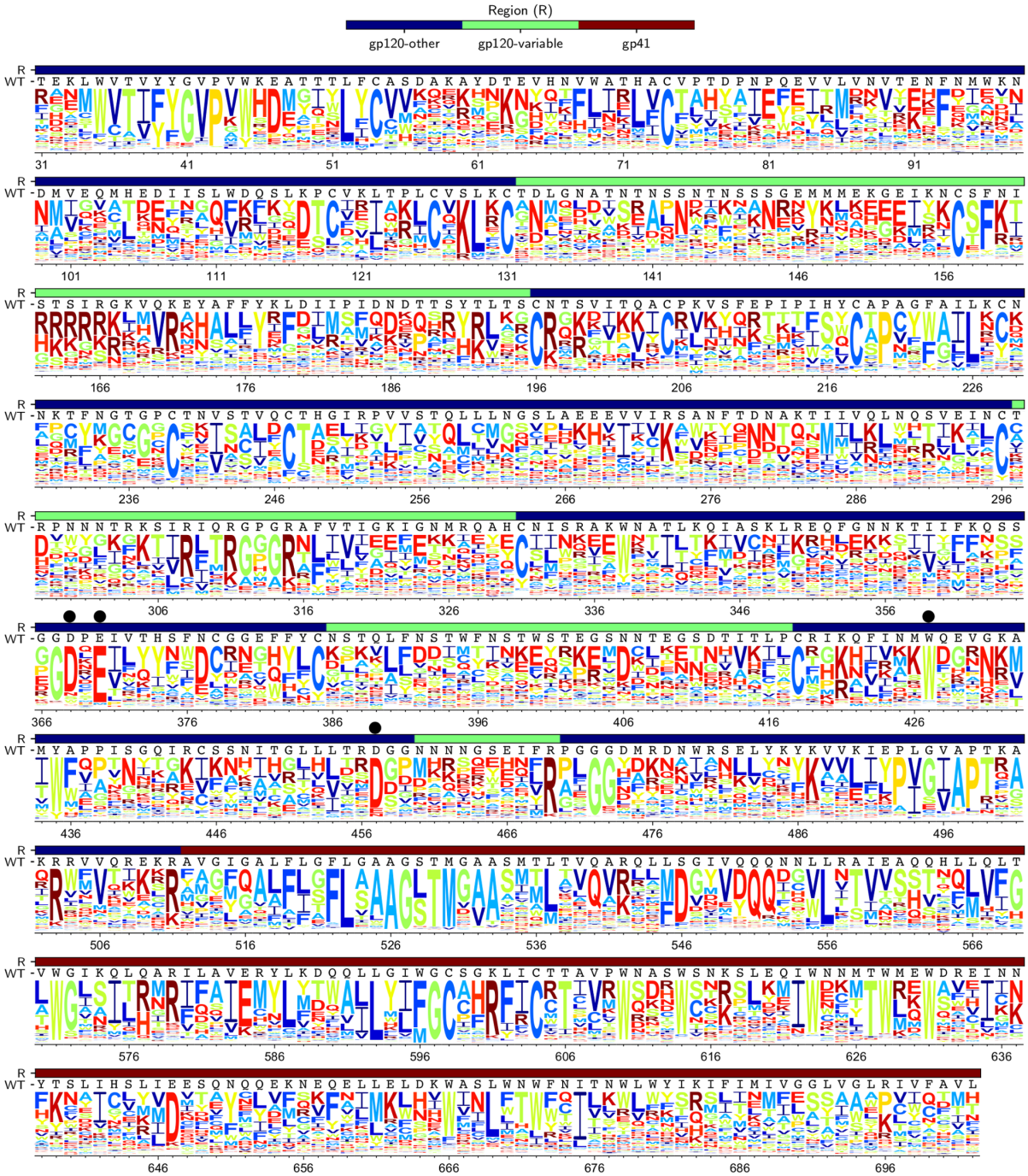

Figure 11: **Env's site-specific amino-acid preferences.** The amino-acid preferences averaged across replicates and re-scaled to account for differences in the stringency of selection between our experiments and natural selection. At each site, the preferences for each of the 20 amino acids sum to one, with the height of each letter proportional to the preference for that amino acid at that site. Letters are colored according to hydrophobicity. The overlay bar indicates the gp120 variable loops, other regions of gp120, and gp41. The LAI wildtype (WT) sequence is shown below the overlay bar. Black dots indicate sites where mutations are known to disrupt CD4 binding (Table 3). Sites are numbered using the HXB2 scheme [84]. Numerical values of the preferences before and after re-scaling are in S1 File and S2 File, respectively.

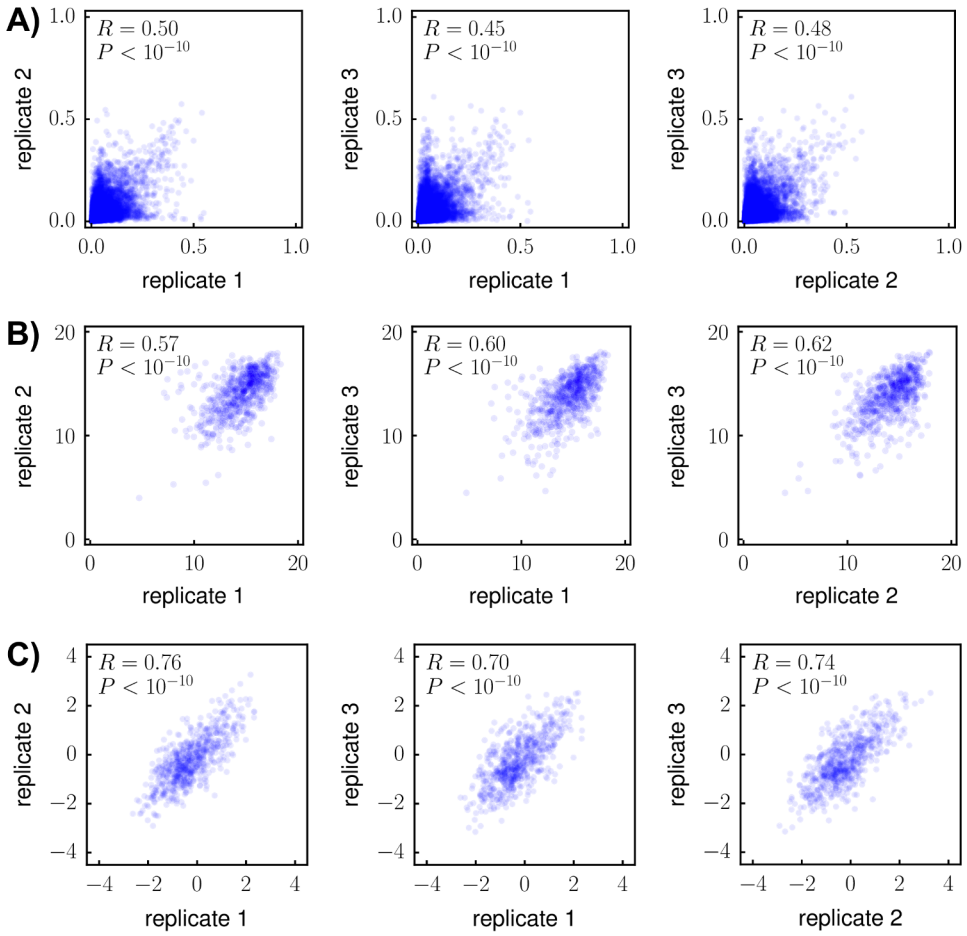

Figure 12: **The amino-acid preferences are modestly correlated among experimental replicates, but the sites tolerate similar numbers of amino acids and prefer similar amino acids across replicates. (A)** Correlations between the site-specific amino-acid preferences from each replicate. **(B)** Correlations between the effective number of amino acids tolerated per site. For each site r, the effective number of tolerated amino acids is e^{H_r}, where H_r is the Shannon entropy of that site's amino-acid preferences. This number ranges between 1 and 20, with 20 indicating all amino acids are preferred equally and 1 indicating only a single amino acid is preferred. **(C)** Correlations between the preference-weighted hydrophobicities. For each site r, the preference-weighted hydrophobicity is $\sum_a \pi_{r,a} \times X_a$ where $\pi_{r,a}$ is the preference of r for amino acid a, and X_a is the Kyte-Doolittle hydropathy [90] of a. The fact that both the effective number of tolerated amino acids and the hydrophobicities are more correlated than the amino-acid preferences means that when different amino acids are preferred at a site in different experimental replicates, the number and chemical properties of the preferred amino acids are similar. Each plot shows the Pearson correlation coefficient and associated P-value. Similar data for replicates 3b-1 and 3b-2 are in Fig 13. The plots in this and subsequent figures show all 20 amino-acid preferences for each site; although only 19 of these preferences are independent parameters, all 20 values are shown because otherwise the correlation will depend on which value is excluded.

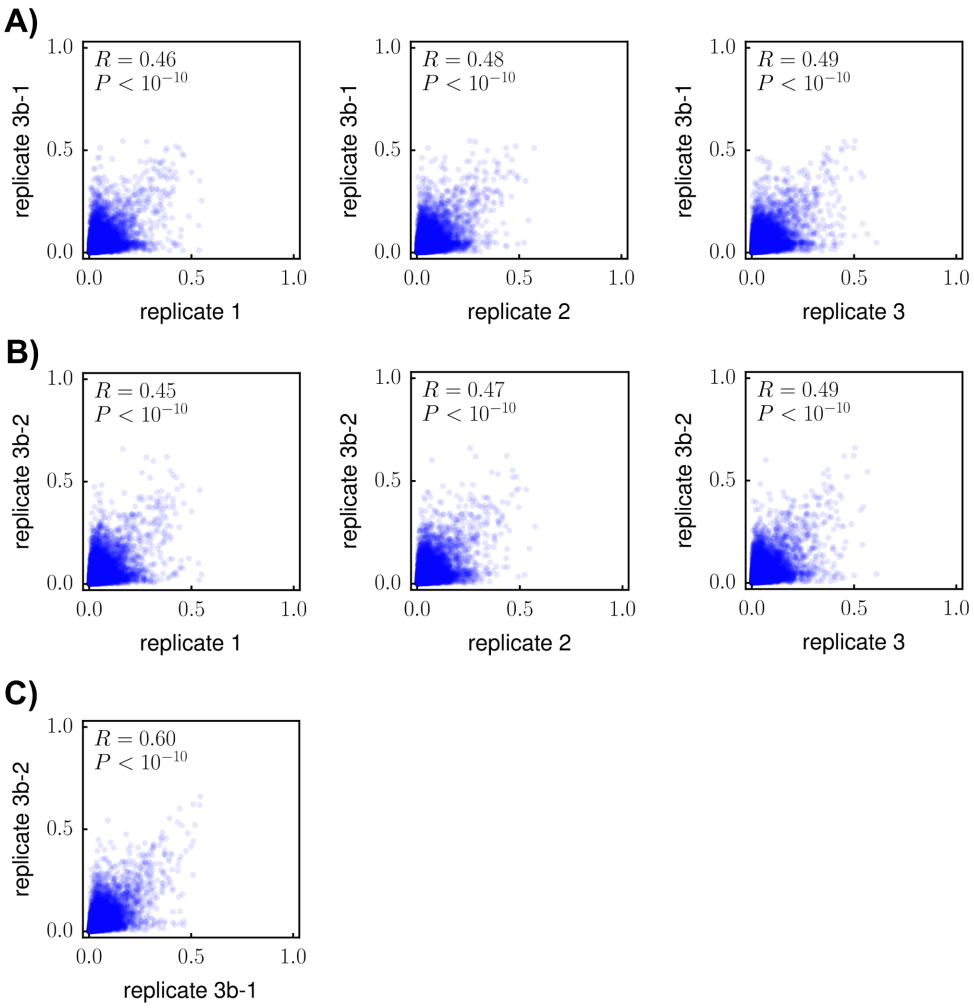

Figure 13: **Correlation of site-specific amino-acid preferences between replicates, including 3b-1 and 3b-2.** **(A)** The correlation between replicate 3b-1 and replicate 1, 2, or 3. **(B)** The correlation between replicate 3b-2 and replicate 1, 2, or 3. **(C)** The correlation between replicates 3b-1 and 3b-2.

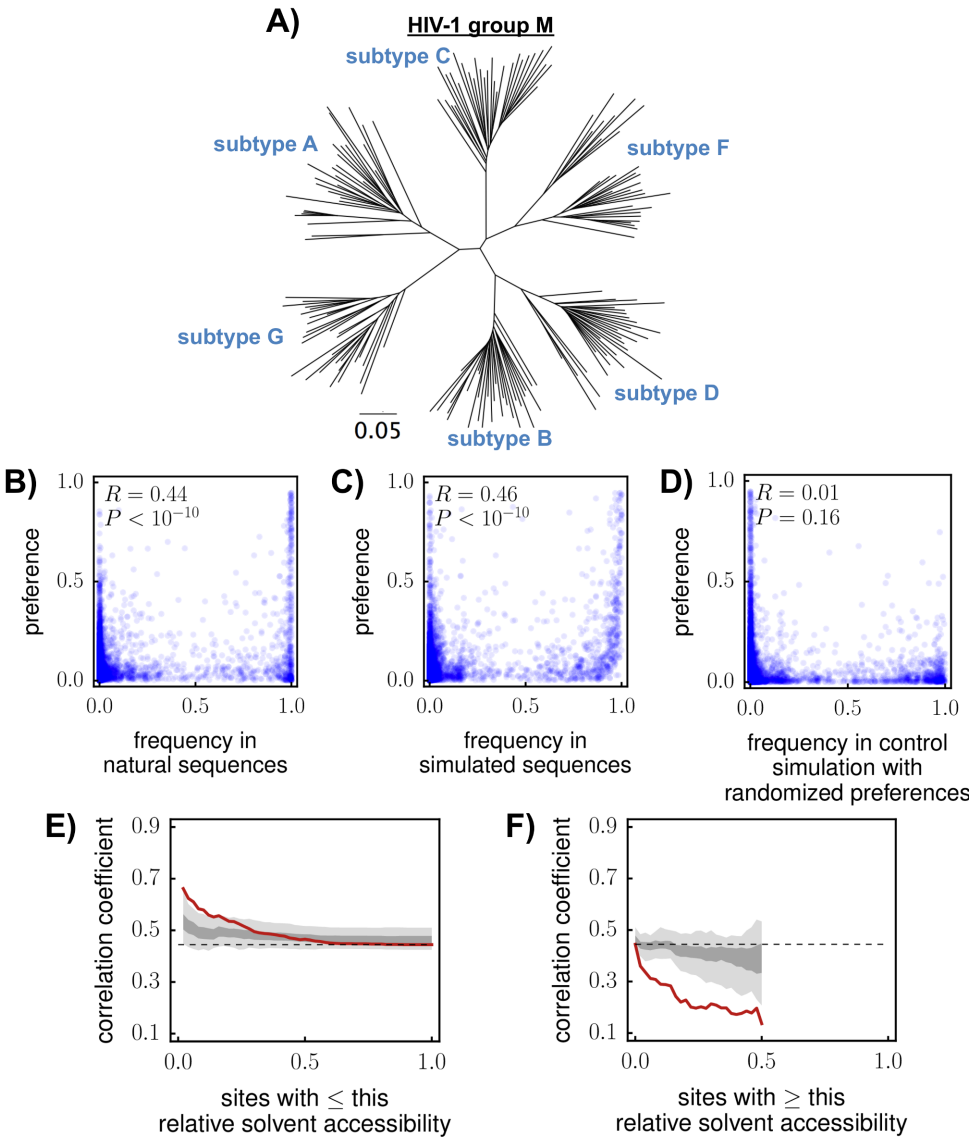

Figure 14: **Correlations between amino-acid preferences and frequencies in natural HIV sequences. (A)** Phylogenetic tree of the HIV-1 group-M sequences in the alignment. **(B)** Correlation between alignment frequencies and preferences. The preferences are the replicate averages rescaled by the stringency parameter in Table 4. **(C)** The correlation if evolution is simulated along the phylogenetic tree assuming that the preferences correctly describe the actual selection. **(D)** There is no correlation in a control simulation in which preferences are randomized among sites. **(E), (F)** Correlation between preferences and alignment frequencies as a function of relative solvent accessibility (RSA). Red lines show the actual correlation. Dark and light gray show the range of correlations in the middle 80% and 100% of 100 simulations. For both plots, data are shown until the subset of sites that meets the RSA cutoff becomes less than 10% of all sites in Env; this is why neither x-axis extends all the way from 0 to 1. Correlation coefficients are Pearson's R.

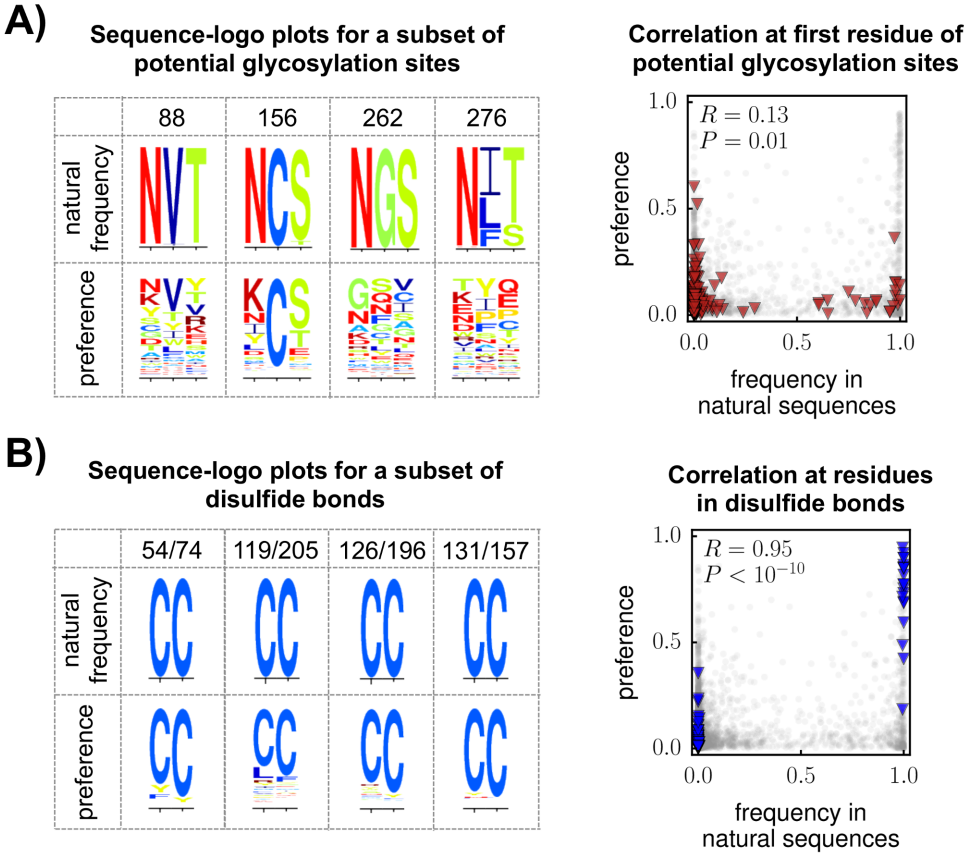

Figure 15: **The correlation between the experimentally measured preferences and amino-acid frequencies in natural sequences is low at glycosylation sites, but high at disulfide-bonded cysteines.** **(A)** The logo plots show the frequencies of amino acids in the group-M alignment or the amino-acid preferences from our experiments at a subset of potential N-linked glycosylation sites (see Fig 16 for all 30 sites). The glycosylation sites are conserved in nature, but tolerant of mutations in our experiment. The scatter plot shows that there is a poor correlation between the preferences and natural amino-acid frequencies at all 22 alignable glycosylation sites: red triangles represent the first position in each glycosylation site, whereas gray circles represent all other sites. **(B)** There is much better concordance between the preferences and natural amino-acid frequencies for Env's disulfide-bonded cysteines. The logo plots show each pair of cysteines for a subset of disulfides (see Fig 16 for all 10 disulfides). The scatter plot shows that there is a strong correlation between the preferences and natural amino-acid frequencies at all disulfide-bonded cysteines.

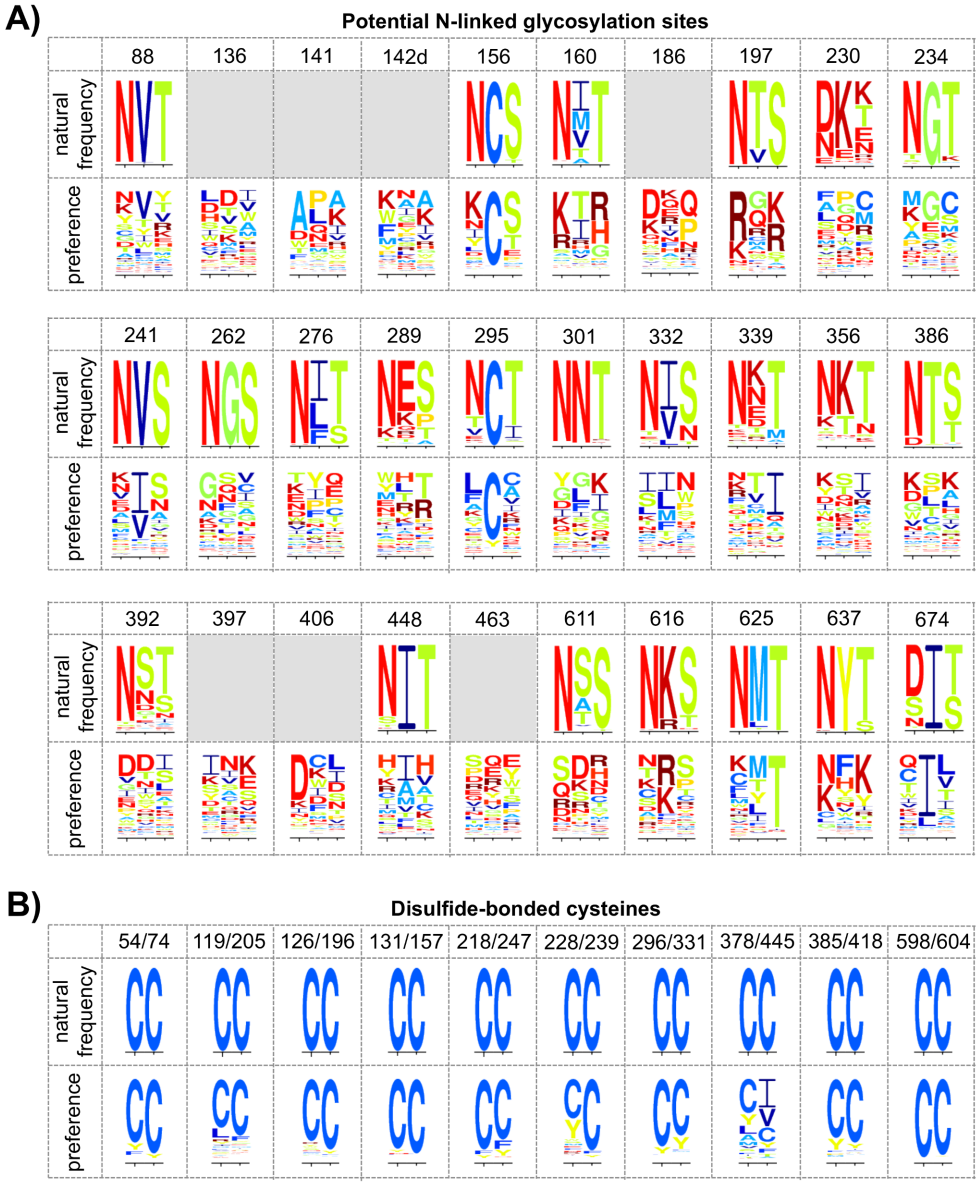

Figure 16: **Amino-acid frequencies and preferences for all potential N-linked glycosylation sites and disulfide bonds.** This figure is similar to Fig 15, but shows logo plots for all 30 glycosylation sites (defined using the N-GlycoSite tool [186] from the HIV sequence database, `http://www.hiv.lanl.gov/`) and all 10 disulfide bonds [94] in LAI. **(A)** Most glycosylation sites are highly conserved in natural sequences, but highly tolerant of mutations in our experiments. Logo plots showing amino-acid frequencies in nature are replaced by grey boxes for sites in the alignment of group-M sequences that were masked because the site had >5% deletions relative to HXB2 or because the region looked unalignable by eye (for details, see IPython notebook `CurateLANLMultipleSequenceAlignment.ipynb` within S3 File). **(B)** Disulfide-bonded cysteines are absolutely conserved in nature. Most of these positions have a strong preference for cysteine in our experiments. A previous study [168] found that only the C378-C445 disulfide bond tolerated alanine mutations at individual cysteines while supporting robust viral replication in cell culture. In accordance with this previous work, these cysteines are the most mutationally tolerant ones in our experiment.

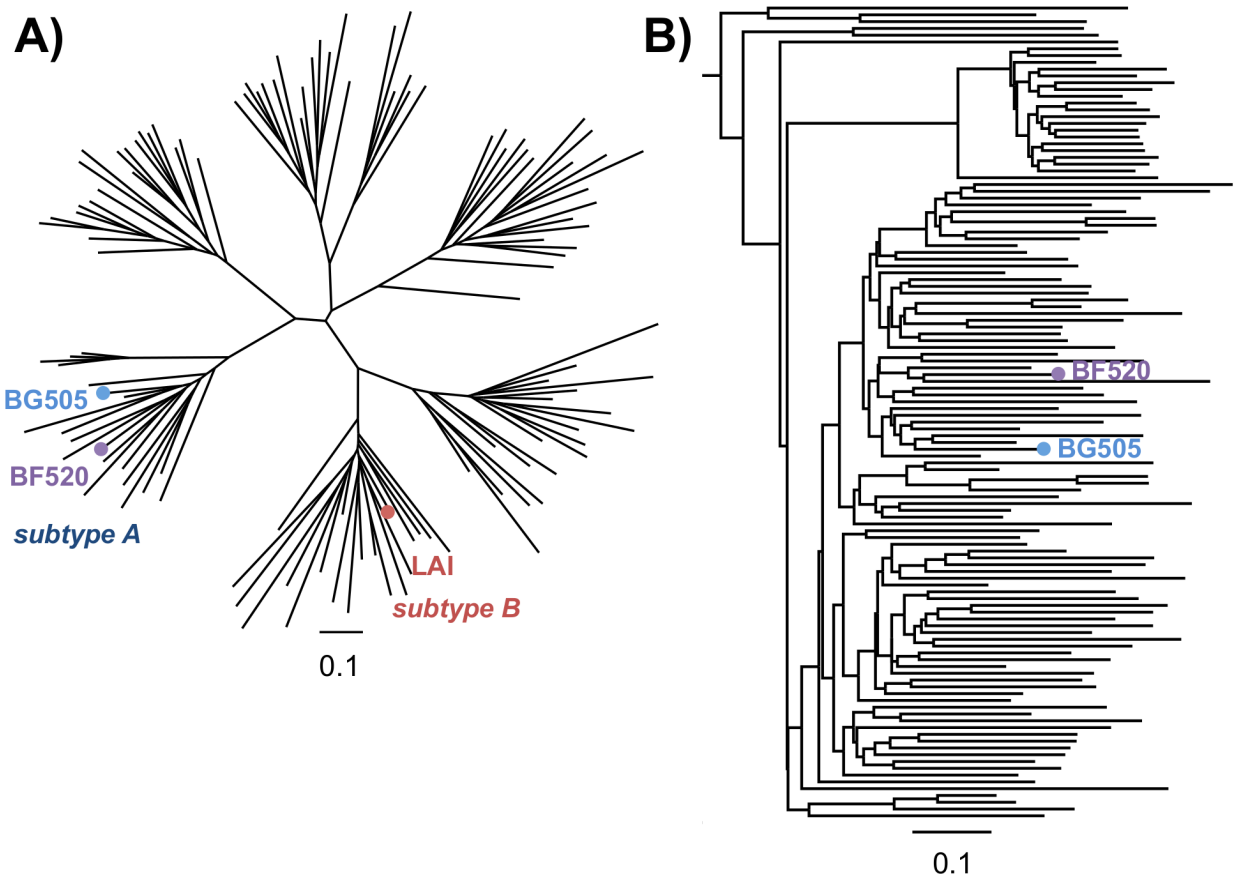

Figure 17: **Phylogenetic trees of HIV-1 *env* sequences showing the relationship between BG505, BF520, and LAI. (A)** A tree of group-M sequences with 20 sequences per subtype for subtypes A, B, C, D, F, and G. **(B)** A tree of subtype-A sequences with 120 sequences total, rooted using a subtype-B sequence as an outgroup. BG505 and BF520 cluster within subtype A, while LAI clusters within subtype B. The alignments we used to make these trees are the same alignments we used in the below `phydms` analysis, excluding the outgroup sequence used to root the subtype-A tree. The scale bars are in units of codon substitutions per site.

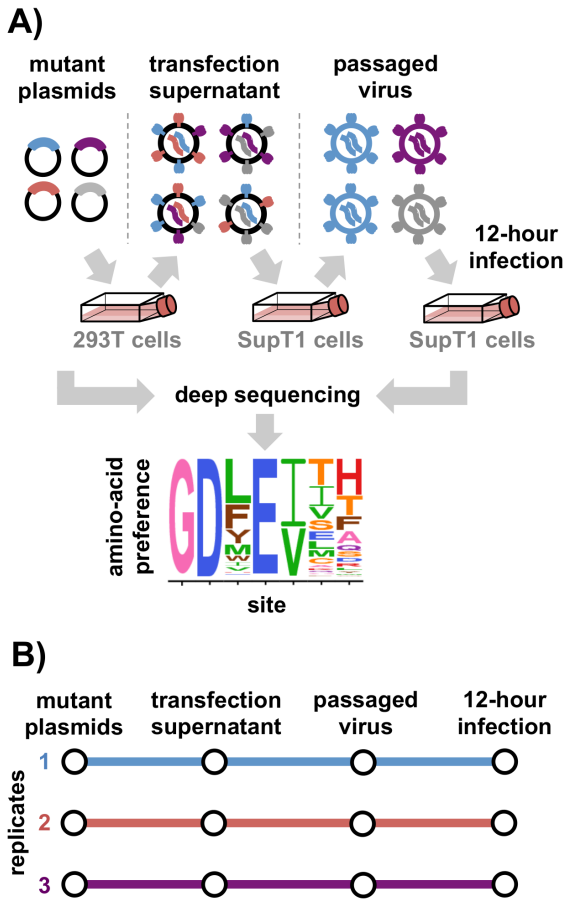

Figure 18: **Deep mutational scanning workflow. (A)** For each homolog, we made a library of proviral HIV plasmids with random codon-level mutations in *env*. We then transfected the plasmids into 293T cells to generate mutant viruses, which may lack a genotype-phenotype link since transfected cell are expected to each receive multiple plasmids. To establish this link and select for functional variants, we first passaged the libraries in SupT1 cells at a low MOI. Then, we imposed a second round of selection by infecting the passaged viruses into SupT1 cells at a high MOI and then harvesting reverse-transcribed unintegrated viral DNA at 12 hours post-infection. Finally, we deep sequenced the libraries before and after selection. We also deep sequenced wildtype controls to estimate error rates due to PCR, deep sequencing, and viral replication. Using these sequencing data, we then inferred each site's preference for each of the 20 amino acids. **(B)** For each homolog, we conducted this experiment in full biological triplicate, beginning at the stage of independently creating the plasmid mutant libraries.

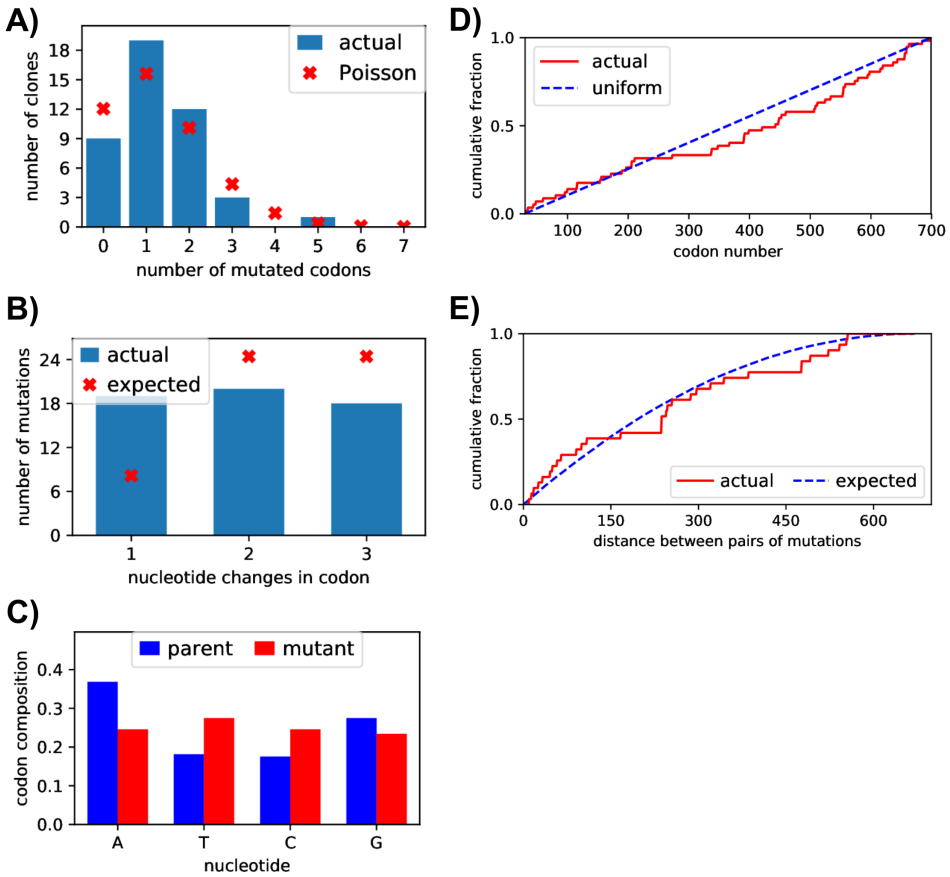

Figure 19: **Sanger sequencing of the BG505 mutant plasmids revealed that codon mutations were distributed roughly uniformly, with an average of 1.5 mutations per gene.** We Sanger sequenced 44 clones of BG505 Env. We sampled these clones roughly evenly from the three replicate mutant plasmid libraries before these libraries had undergone any functional selection. **(A)** There was an average of 1.5 mutant codons per clone, with the number of mutations per clone roughly following a Poisson distribution. **(B)** The mutant codons had a mix of single-, double-, and triple-nucleotide changes, with an elevated number of single-nucleotide changes than expected. **(C)** Nucleotide frequencies were fairly uniform in the mutant codons, as expected from random mutagenesis. **(D)** Mutations were distributed roughly evenly along the mutagenized region of *env* (30-699 in BG505 numbering). **(E)** For clones with multiple mutations, we computed the pairwise distance in primary sequence between each codon mutation and plotted the cumulative distribution of these distances (red line). We also simulated the expected distribution of pairwise distances if mutations occurred entirely independently (blue line). The observed distribution is close to the expected distribution.

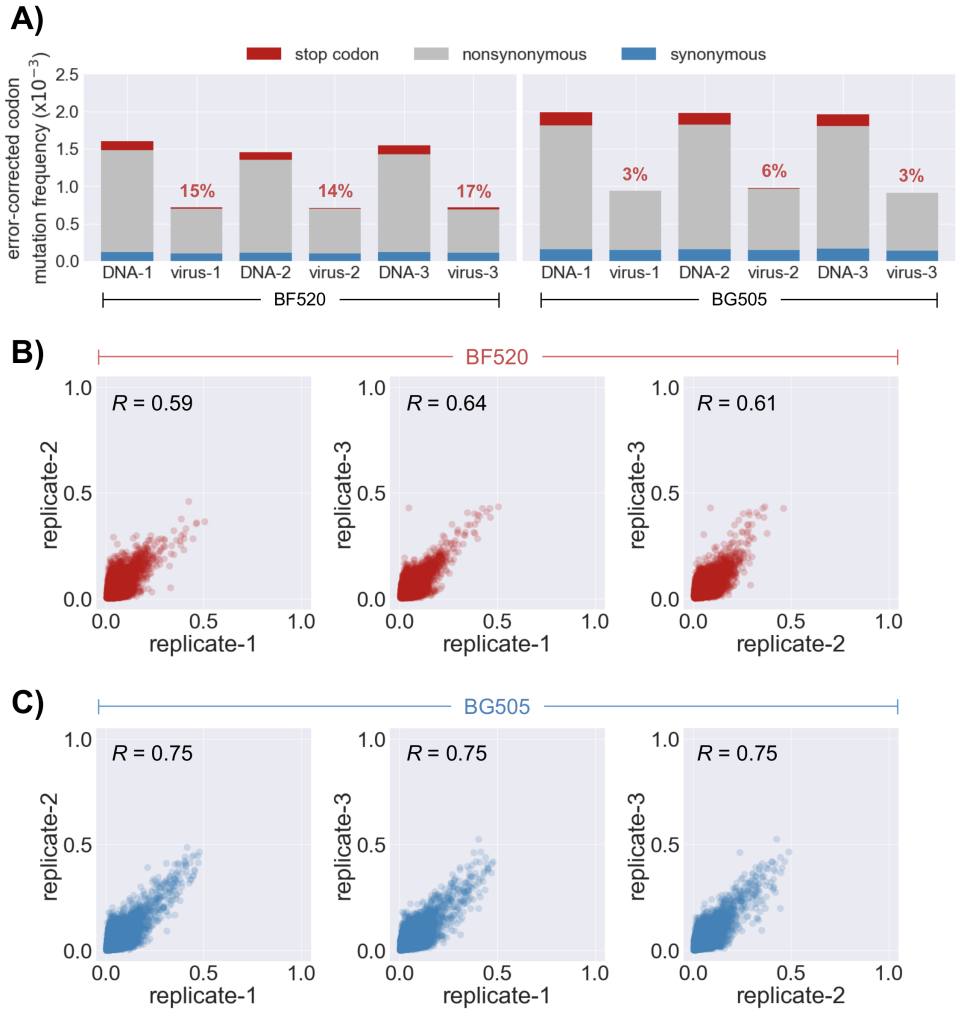

Figure 20: **The deep mutational scanning experiments imposed strong purifying selection and led to reproducible estimates of each homolog's amino-acid preferences. (A)** For each replicate of each homolog, we deep sequenced the starting plasmid libraries (DNA) and the selected virus libraries (virus). Bars show the per-codon mutation frequency averaged across all sites after subtracting mutation frequencies determined using wildtype controls (Fig 21). Mutations leading to stop codons were purged during the selection step to 3-17% their starting frequencies (see red numbers), indicating strong purifying selection. Nonsynonymous mutations decreased to 40-50% their starting frequencies, consistent with a large fraction of amino-acid mutations being deleterious to Env. Synonymous mutations only decreased to 87-95% their starting frequencies, as would be expected if synonymous mutations tend to have more neutral effects than nonsynonymous changes. Panels **(B)** and **(C)** show the correlation in our estimates of each homolog's site-specific amino-acid preferences between experimental replicates, with **(B)** showing the correlation between BF520 replicates and **(C)** showing the correlation between BG505 replicates. Each correlation plot reports the associated Pearson correlation coefficient, which ranged from 0.59-0.75, indicating that our estimates are largely repeatable between replicates.

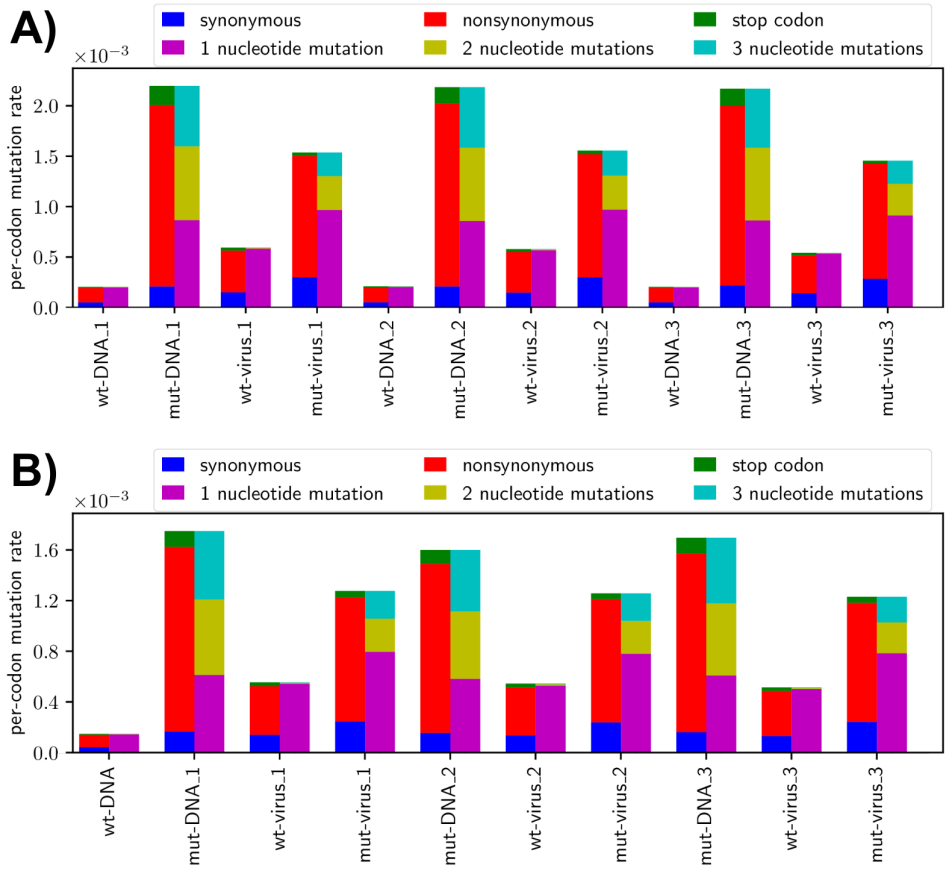

Figure 21: **Per-codon mutation frequencies of the mutant libraries and wildtype controls before and after selection.** This figure is similar to Fig 20, but it reports *non*-error-corrected per-codon mutation frequencies for each of the mutant libraries and wildtype controls before and after selection for **(A)** BG505 and **(B)** BF520. Specifically, for each each replicate, wt-DNA and mut-DNA refer to the pre-selection wildtype and mutant plasmids, respectively; and wt-virus and mut-virus refer to the post-selection wildtype and mutant viruses, respectively. Note, there is only a single replicate of the wt-DNA sample for BF520. The bars show frequencies of different codon-level mutations. The left bar for each sample groups mutations as being either synonymous, non-synonymous, or leading to the introduction of a stop codon. The right bar for each sample groups mutations as introducing a single-nucleotide codon mutation (e.g., aaa $\rightarrow$ aaT), double-nucleotide codon mutation (e.g., aaa $\rightarrow$ aTT), or triple-nucleotide mutation (e.g., aaa $\rightarrow$ TTT). Each wildtype plasmid shows a low rate of single-nucleotide codon mutations, consistent with errors from PCR and deep sequencing. Importantly, the wildtype plasmids have a much lower mutation rate than the mutant plasmids. The mutation frequency of single-nucleotide mutations substantially increases for each wildtype virus, which is expected due to errors from viral replication. Comparing the wildtype viruses to the mutant viruses suggests that roughly a third to a half of mutations in the mutant viruses arise due to to errors from PCR, deep sequencing, and viral replication, underlying the importance of estimating these error rates using the controls. When inferring Env's amino-acid preferences, we use the wildtype controls to statistically correct for such errors, as described in the Methods.

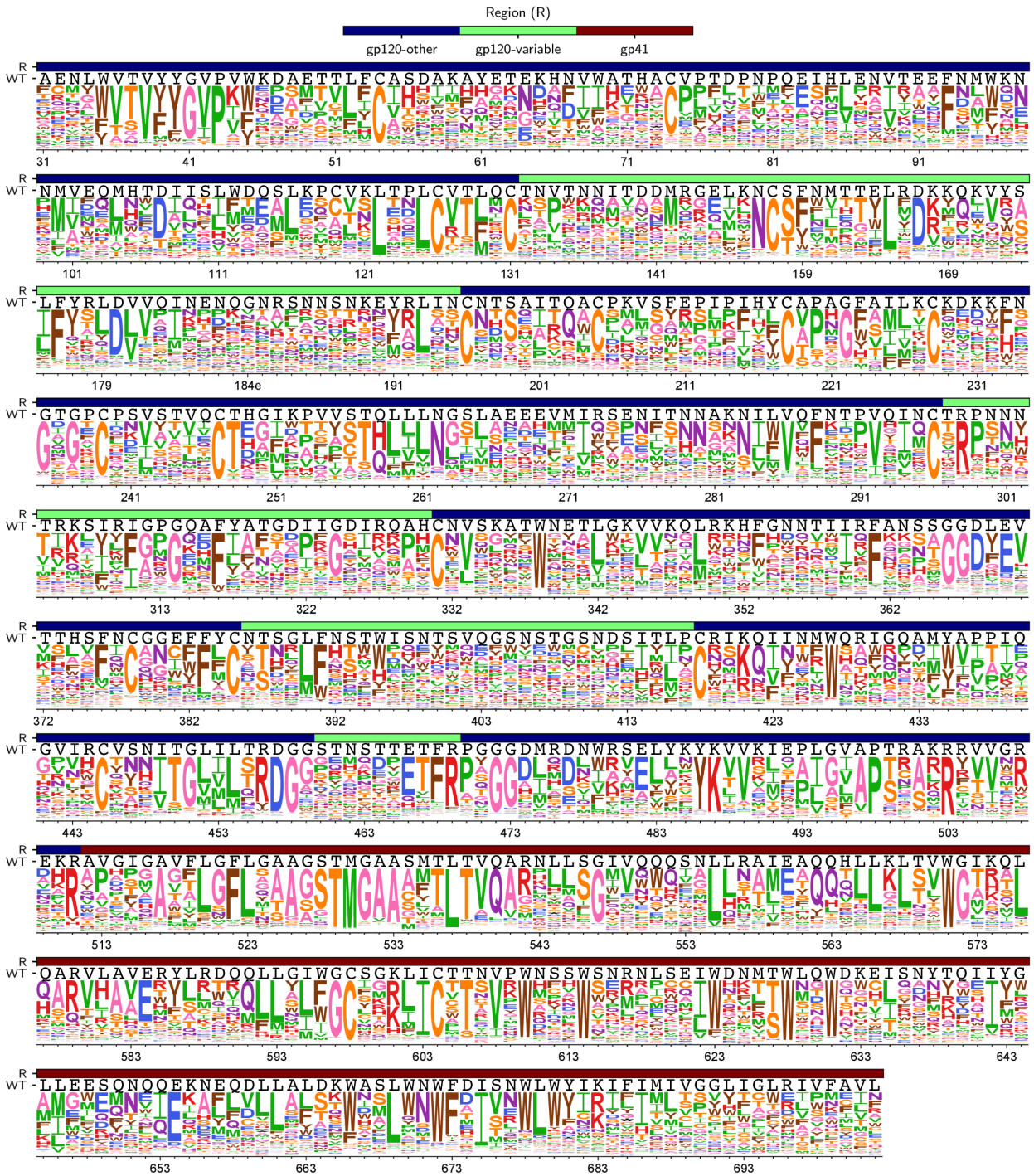

Figure 22: **The rescaled averaged site-specific amino-acid preferences for BG505.** This logo plot shows the site-specific amino-acid preferences for BG505 after averaging between replicates and then rescaling them using the stringency parameter from Table 7 inferred for BG505 from the group-M alignment. Each site has a stack of 20 letters corresponding to the 20 amino acids. Letter heights, which sum to one at each site, are proportional to the site's preference for each amino acid. The top bar (R) indicates gp120 variable loops, other regions in gp120, or gp41. The bottom bar (WT) shows the wildtype amino-acid sequence for BG505. Sites are numbered according to the HXB2 numbering scheme [84]).

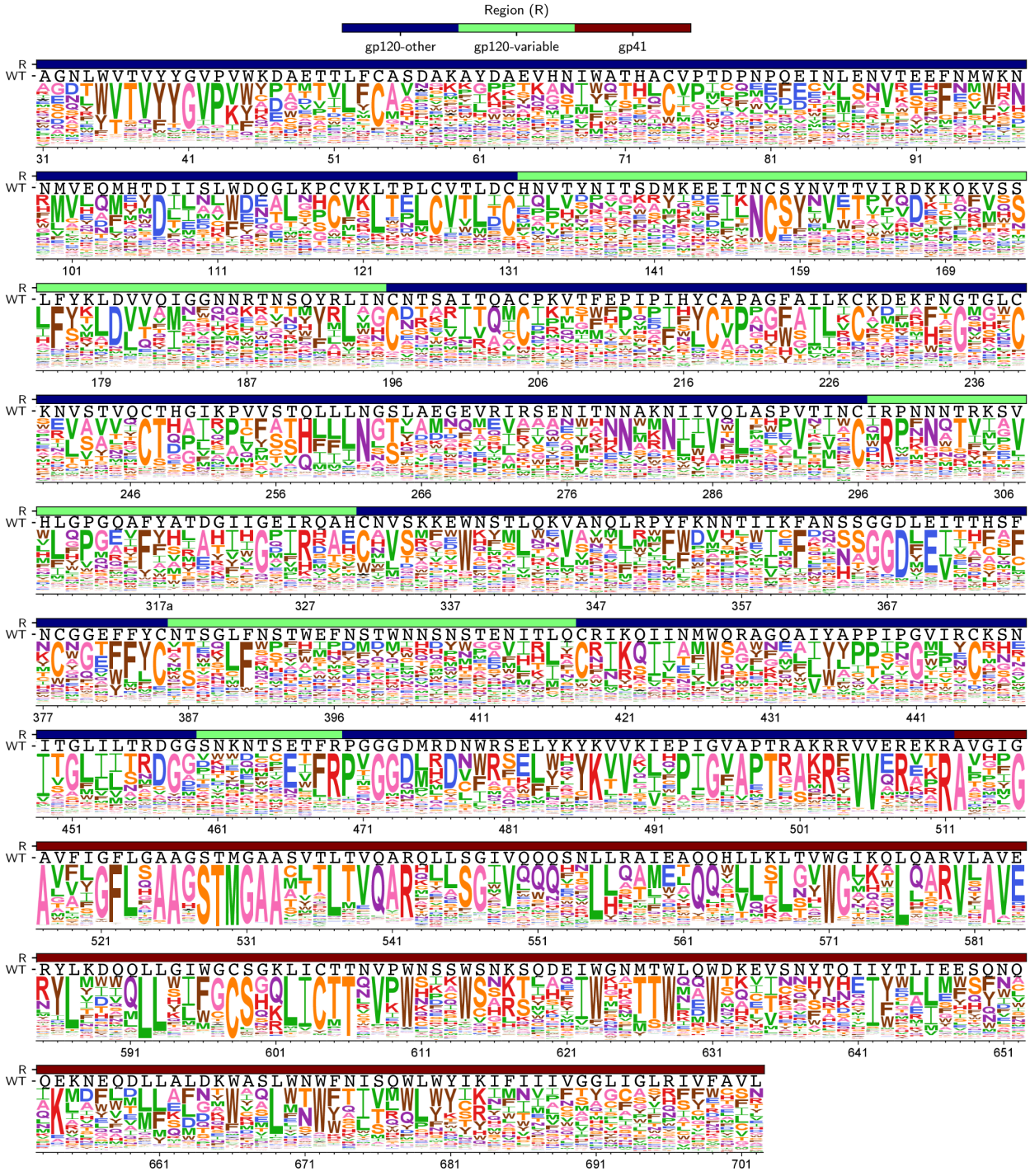

Figure 23: **The re-scaled averaged site-specific amino-acid preferences for BF520.** This figure is the same as Fig 22, but for BF520 instead of BG505.

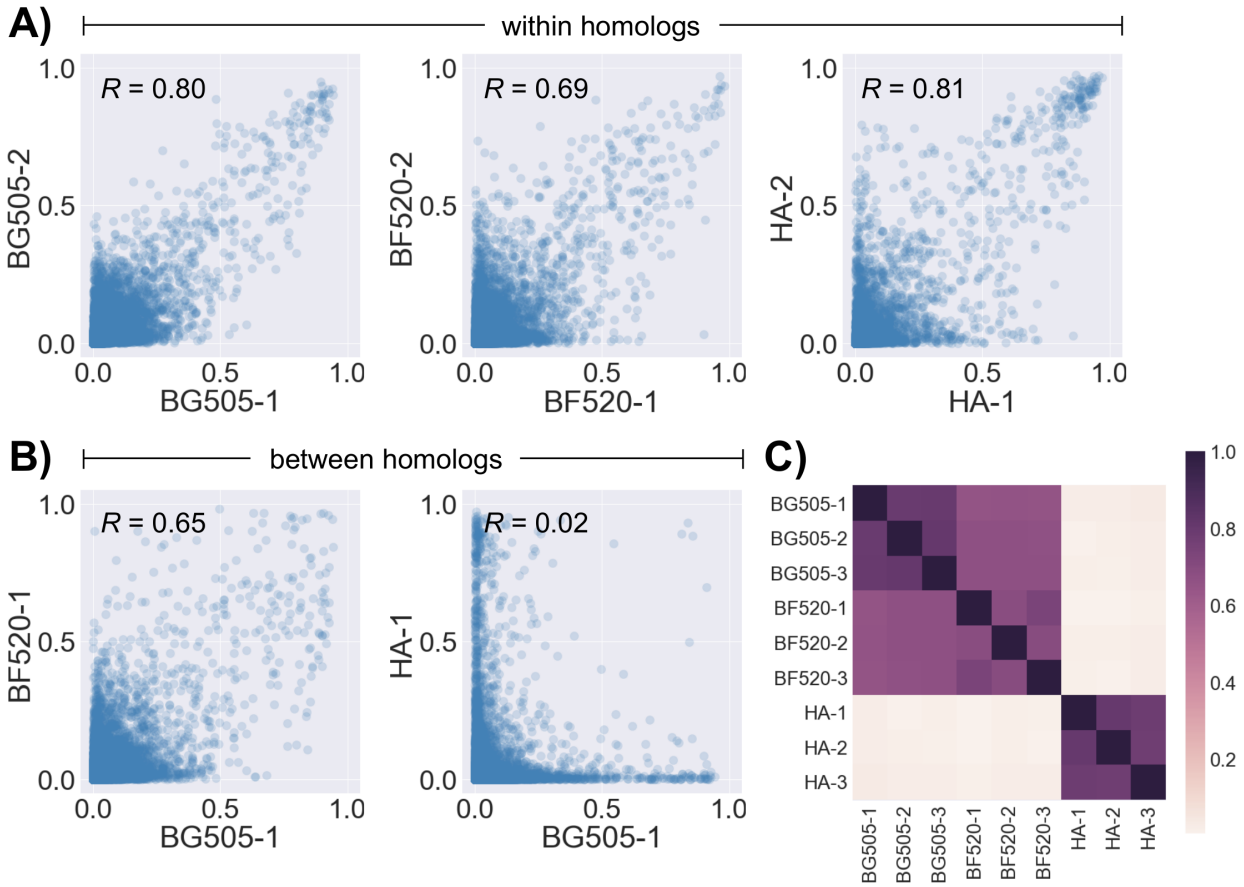

Figure 24: **Env's preferences are well correlated between BG505 and BF520.** This figure shows the correlation of the preferences between replicates, both within and between Env homologs. In this figure, the preferences for each replicate have been rescaled using the stringency parameter for the corresponding homolog from the group-M analysis in Table 7. We analyzed 616 sites that are shared between homologs and were in readily alignable regions of the group-M multiple-sequence alignment. As a control, we also compared our estimates of Env's preferences with the preferences of a non-homologous protein – influenza HA – across the 480 sites where these proteins overlap in sequential numbering. **(A)** Plots showing the correlation of preferences between two replicates from the same protein. Here, differences reflect experimental noise. **(B)** Plots showing the correlation of preferences between Env homologs, or between BG505 and HA. Here, differences reflect a combination of experimental noise and biological differences between proteins. The Env homologs are well correlated, whereas BG505 and HA are not correlated. **(C)** The heat map shows the Pearson correlation coefficient for all pairwise comparisons of all experimental replicates of each homolog (coefficients are also shown on the correlation plots). This heat map indicates that the trends in **(A)** and **(B)** hold for all replicates.

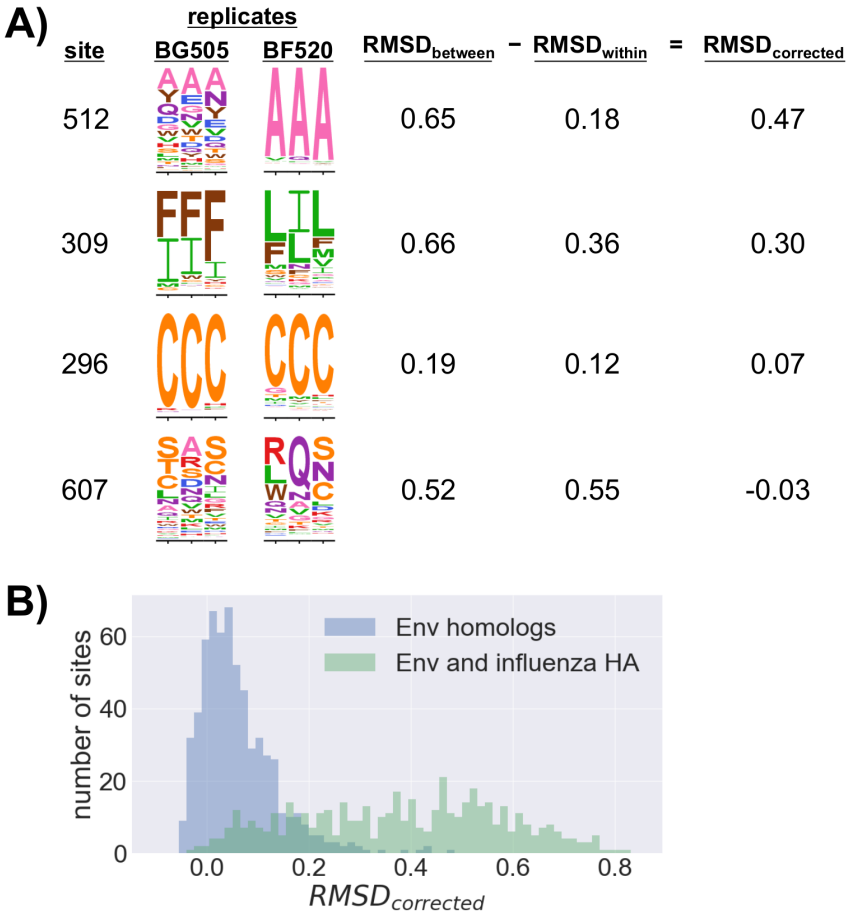

Figure 25: **Most shifts in amino-acid preference between Env homologs are small-to-intermediate in effect size after correcting for experimental noise.** We quantified the shift in each site's preferences using a distance metric that corrects for experimental noise, as quantified by experimental replicates. **(A)** This panel shows how the distance metric is calculated for a subset of sites. **(B)** The blue histogram shows the distribution of site-specific $RMSD_{corrected}$ values between Env homologs for all 616 sites being compared. Most sites have a distance greater than zero. The overlayed green histogram shows the distribution of $RMSD_{corrected}$ values between BG505 Env and influenza HA for all 480 sites that overlap in sequential numbering. Most site-specific distances in the Env-HA comparison are much larger than the distances in the BG505-BF520 comparison.

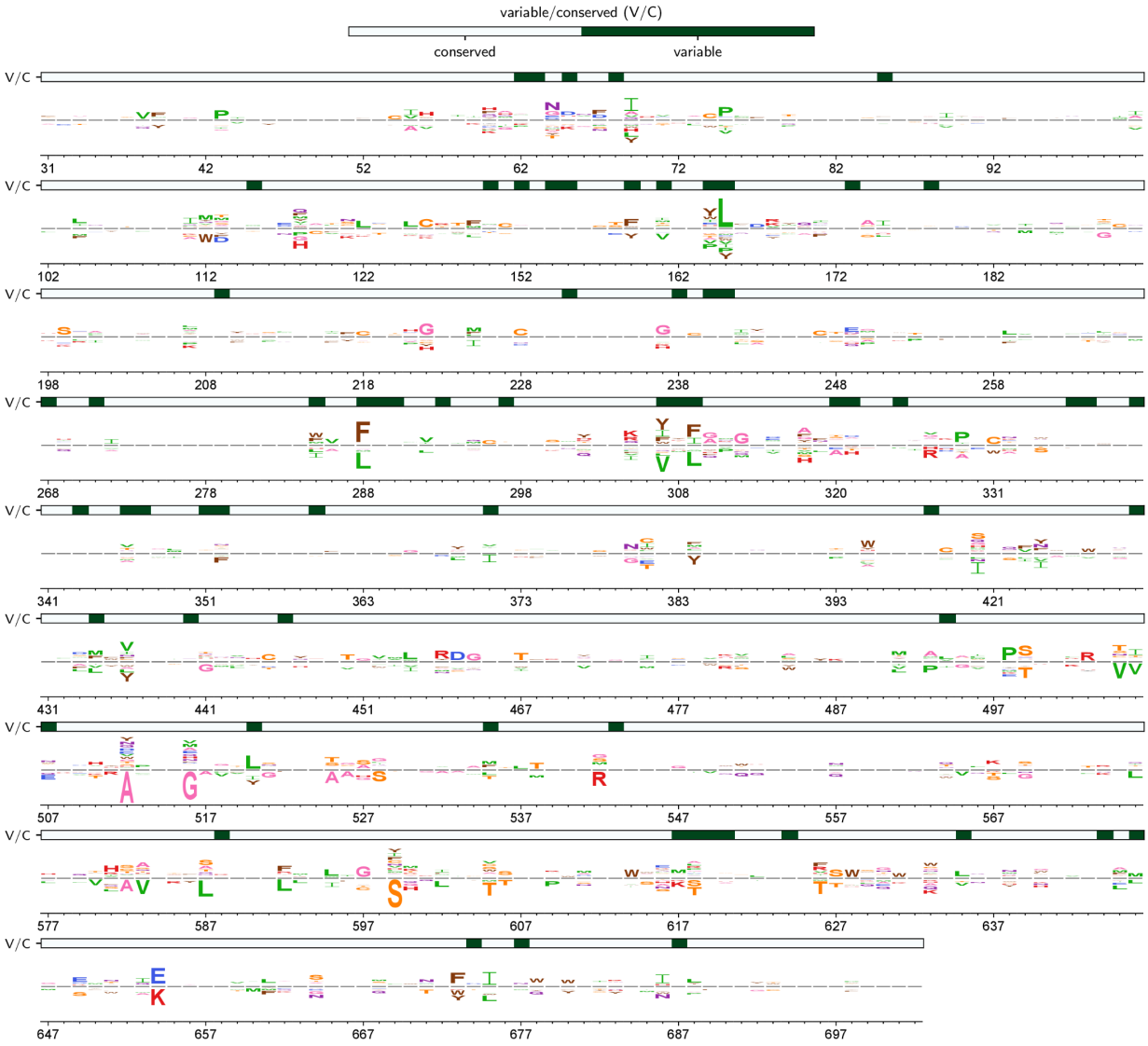

Figure 26: **The difference in Env's site-specific amino-acid preferences between homologs, scaled by site-specific $RMSD_{corrected}$ values.** This figure shows the estimated preference shift for each amino acid across all 616 sites in Env being compared. For each site, we plot the difference in the rescaled averaged preferences between homologs ($\Delta\pi_{r,a} = \pi_{r,a}^{BG505} - \pi_{r,a}^{BF520}$), where each site has 20 letters corresponding to the 20 amino acids, and where negative and positive values are shown below and above the central black line, respectively. Thus, amino acids above the central line are more preferred in BG505, whereas amino acids below the central line are more preferred in BF520. At each site, we adjusted the total height of the letters in both directions to equal that site's $RMSD_{corrected}$ from Fig 25B (i.e., $\sum_a |\Delta\pi_{r,a}| = RMSD_{corrected}$). In effect, the sites with the largest stack heights are the sites where we observe the largest differences, after accounting for noise. The bar indicates which sites are variable or conserved in amino-acid sequence between homologs. Note that adjacent sites are not always contiguous in primary sequence, since we masked sites that are either not shared between homologs, or are not readily alignable in the group-M multiple-sequence alignment.

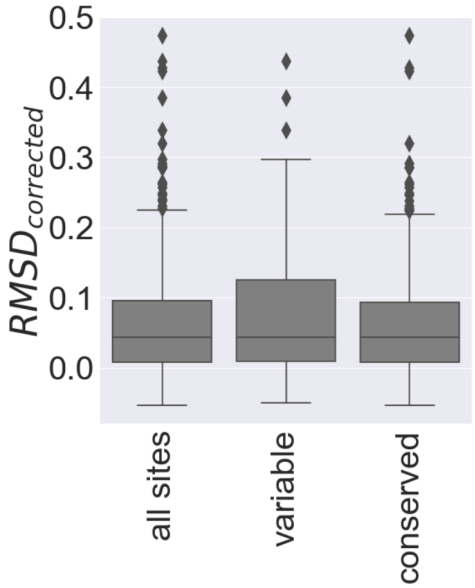

Figure 27: **The median shift in preferences between Env homologs is similar at variable vs. conserved sites.** The box plots compare show the shifts at all sites, just sites that differ in wildtype amino acids between BG505 and BF520 (variable), or sites that are conserved between the homologs. Shifts are quantified by site-specific $RMSD_{corrected}$ values from Fig 25. The median shift is roughly equivalent for each group of sites.

LIST OF TABLES

Table Number Page

Table 1: **Sites of mutations recurrently selected in cell culture.**

	error-corrected mutation frequency (P2:DNA)				hydropathy				
sites	1	2	3	WT amino acid	WT amino acid	preferences	difference	RSA	entropy of preferences
48	3.4	2.1	3.4	A	1.8	-0.4	2.2	0.2	2.9
62	1.6	4.2	8.2	D	-3.5	-1.6	-1.9	0.6	3.6
64	14.4	6.8	10.3	E	-3.5	-1.6	-1.9	0.6	2.8
65	1.3	3.3	3.1	V	4.2	-2.1	6.3	0.6	3.3
66	6.1	3.0	13.7	H	-3.2	-0.2	-3.0	0.6	3.5
81	4.1	4.5	4.1	P	-1.6	-1.9	0.3	0.6	2.9
105	2.2	3.1	7.0	H	-3.2	0.7	-3.9	0.0	3.0
162	11.2	4.9	1.2	S	-0.8	-2.6	1.8	0.2	2.7
188	5.8	2.9	4.8	T	-0.7	-2.1	1.4	0.5	2.9
203	-8.6	6.7	5.4	Q	-3.5	0.0	-3.5	0.0	2.5
207	15.6	19.2	21.1	K	-3.9	2.7	-6.6	0.5	2.7
212	5.2	-27.2	9.8	P	-1.6	-0.3	-1.3	0.2	3.4
377	10.5	1.1	3.9	N	-3.5	-2.7	-0.8	0.2	2.2
420	3.0	3.8	4.8	I	4.5	-1.7	6.2	0.0	2.8
433	7.2	7.2	8.4	A	1.8	2.2	-0.4	0.0	2.2
436	3.1	2.8	5.2	A	1.8	1.1	0.7	0.0	2.4
443	-5.4	4.2	3.2	I	4.5	-0.9	5.4	0.1	3.4
557	4.7	11.4	5.6	R	-4.5	0.6	-5.1	nd	3.5
558	2.9	6.0	3.8	A	1.8	-0.7	2.5	nd	2.2
560	5.9	5.2	6.0	E	-3.5	0.7	-4.2	nd	3.2
564	42.5	6.3	-4.0	H	-3.2	-0.8	-2.4	nd	4.1
588	9.4	11.4	11.2	K	-3.9	1.5	-5.4	0.2	3.0
591	6.2	4.2	5.0	Q	-3.5	2.2	-5.7	0.0	2.3
653	1.6	3.4	3.2	Q	-3.5	1.7	-5.2	0.5	3.1
655	10.8	8.2	4.7	K	-3.9	1.5	-5.4	0.1	3.3

The 25 sites (HXB2 numbering) from Fig 5B for which the error-corrected mutation frequency increased by $>$3-fold in at least two replicates upon two rounds of passaging in cell culture. We report the change in mutation frequency for each site as a ratio of mutation frequency post- vs. pre- selection (P2:DNA). Negative ratios arise when the mutation frequency in the wildtype control is greater than in the mutant plasmid or virus library. For many sites, there is a large difference between the hydropathy of the wildtype amino acid and the hydropathy averaged across the site's amino-acid preferences, suggesting pressure to change the chemical character of the amino-acid. We also report the relative solvent accessibility for each site as computed using PDB structure 4TVP [124]. Adaptation at each site could occur through a single highly beneficial amino-acid change or though numerous roughly equally beneficial changes. For many sites, we observe the latter scenario, as indicated by the entropy of the preferences, which ranges from 2.2-4.1 in this list of 25 sites, compared to 0.5-4.3 for all sites.

Table 2: **Sites that differ between LAI and HXB2 tend to prefer the HXB2 identity.**

site	HXB2 identity	LAI identity	HXB2 preference	LAI preference	difference (=HXB2-LAI)
135	K	G	0.047	0.046	0.001
137	D	A	0.202	0.043	0.159
146	R	E	0.083	0.054	0.028
148	I	M	0.118	0.038	0.080
192	K	T	0.316	0.021	0.295
275	V	A	0.203	0.043	0.160
290	T	Q	0.110	0.036	0.074
306	R	S	0.027	0.043	-0.016
340	N	A	0.024	0.055	-0.030
423	I	F	0.250	0.186	0.064
429	K	E	0.085	0.006	0.078
461	S	N	0.074	0.039	0.035
464	E	G	0.133	0.051	0.082
625	H	N	0.028	0.041	-0.013
626	T	M	0.155	0.281	-0.126
684	L	I	0.028	0.048	-0.021

This table shows all sites (in HXB2 numbering) that differ between HXB2 and LAI for which we have estimates of Env's preferences. At each site, we report the wildtype amino-acid identity for each strain and its corresponding preference from Fig 11. Most sites favor the HXB2 identity more than the LAI identity. Three sites (137, 192, and 275) strongly prefer the HXB2 identity, while only a single (626) site strongly prefers the LAI identity.

Table 3: **Our experimental estimates are mostly concordant with existing knowledge about the effects of mutations to functionally or structurally important parts of Env.**

Env function	Site (HXB2 numbering)	Mutation(s) known to disrupt function	Citation	Amino-acid preferences from our experiments
Disulfide bond	C at 54, 74, 119, 126, 131, 157, 196, 205, 218, 228, 239, 247, 296, 331, 385, 418, 598, 604	A	[168]	Preference for C at each of these sites is >30-fold higher than for A
CD4 binding	D368	P, R, N, K, E	[119]	Preference for D is >10-fold higher than for these other amino acids
CD4 binding	E370	Q, R	[119]	Preference for E is >100-fold higher than for these other amino acids
CD4 binding	W427	V, S	[119, 34]	Preference for W is >100-fold higher than for these other amino acids
CD4 binding	D457	A	[119]	Preference for D is >100-fold higher than for A
Co-receptor binding	R298	A	[10]	Preference for A is actually higher than for R
Co-receptor binding	R308	A	[10]	Preference for R is >100-fold higher than for A
Co-receptor binding	R315	A	[10]	Preference for R is >100-fold higher than for A
Co-receptor binding	F317	A	[10]	Preference for F is >100-fold higher than for A
Co-receptor binding	K421	A	[10]	Preference for K is >100-fold higher than for A
Co-receptor binding	Q422	A	[10]	Preference for A is actually higher than for Q
Protease cleavage site	R511	T	[56]	Preference for R is >100-fold higher than for T

The preferences listed in the last column are the average from all replicates, re-scaled by the stringency parameter in Table 4.

Table 4: **Correlation of amino-acid preferences with amino-acid frequencies in nature.**

	correlation		
replicate	preferences	rescaled preferences	stringency parameter (β)
1	0.32	0.33	1.7
2	0.31	0.32	1.6
3	0.29	0.29	1.4
3b-1	0.36	0.37	1.5
3b-2	0.35	0.36	1.5
average	0.40	0.44	2.1

Pearson correlation between experimentally measured amino-acid preferences and frequencies of amino acids in an alignment of HIV-1 group-M sequences. Correlations are shown for both raw preferences and preferences re-scaled by the stringency parameter that maximizes the correlation. The correlation is highest when the preferences are averaged across replicates and re-scaled by a stringency parameter > 1. Because insertions and deletions make some sites difficult to align, we masked columns that had $>$5% gap characters, or columns in variable loops that appeared poorly aligned by eye.

Table 5: **Broadly neutralizing antibody epitopes have significantly lower mutational tolerance than other sites in Env**

variable	coefficient	95% confidence interval
RSA	1.38	1.05 to 1.70
RRE	-0.80	-0.98 to -0.62
variable loop	0.09	-0.07 to 0.25
bNAbs	-0.15	-0.27 to -0.04

Multiple linear regression of mutational tolerance against relative solvent accessibility (RSA), whether a site is in the RRE, whether a site is in a variable loop, and the number of the anti-CD4 binding site broadly neutralizing antibody (bNAb) epitopes in which it is found. Positive coefficients indicate an association with increased mutational tolerance; negative coefficients indicate an association with reduced mutational tolerance. The units of the predictor variables are not standardized, and so coefficients are in units of entropy / (predictor variable unit).

Table 6: **When considered individually, none of the variable loops have a statistically significant association with mutational tolerance.**

variable	coefficient	95% confidence interval
RSA	1.37	1.04 to 1.70
RRE	-0.80	-0.98 to -0.62
bNAbs	-0.16	-0.28 to -0.05
V1	0.09	-0.28 to 0.46
V2	-0.13	-0.39 to 0.13
V3	0.15	-0.13 to 0.42
V4	0.32	-0.03 to 0.66
V5	0.28	-0.21 to 0.77

A multiple linear regression as in Table 5, except the five variable loops (V1-V5) are analyzed independently from one another.

Table 7: **Phylogenetic models that incorporate Env's preferences indicate that selection was less stringent in the lab than in nature.**

Group M

model	ΔAIC	log likelihood	parameters: optimized values
BG505	0.00	-58185.53	7: β = 2.07, α_ω = 0.69, β_ω = 0.49, κ = 3.16
BF520	70.26	-58220.66	7: β = 2.44, α_ω = 0.71, β_ω = 0.47, κ = 3.11
LAI	2994.80	-59682.93	7: β = 1.00, α_ω = 0.55, β_ω = 0.46, κ = 3.09
BF520 site averaged	3913.08	-60142.07	7: β = 2.24, α_ω = 0.50, β_ω = 0.41, κ = 3.21
BG505 site averaged	3922.08	-60146.57	7: β = 1.53, α_ω = 0.51, β_ω = 0.42, κ = 3.21
LAI site averaged	4083.20	-60227.13	7: β = 0.83, α_ω = 0.51, β_ω = 0.41, κ = 3.2
YNGKP M5	4423.42	-60392.24	12: α_ω = 0.53, β_ω = 0.55, κ = 3.20

Subtype A

model	ΔAIC	log likelihood	parameters: optimized values
BF520	0.00	-48166.45	7: β = 2.80, α_ω = 0.59, β_ω = 0.33, κ = 3.53
BG505	234.70	-48283.8	7: β = 2.19, α_ω = 0.58, β_ω = 0.32, κ = 3.50
LAI	2860.64	-49596.77	7: β = 1.06, α_ω = 0.48, β_ω = 0.30, κ = 3.41
BG505 site averaged	3909.68	-50121.29	7: β = 0.93, α_ω = 0.47, β_ω = 0.30, κ = 3.61
BF520 site averaged	3913.70	-50123.3	7: β = 1.22, α_ω = 0.47, β_ω = 0.30, κ = 3.61
LAI site averaged	3961.18	-50147.04	7: β = 0.01, α_ω = 0.46, β_ω = 0.30, κ = 3.67
YNGKP M5	3980.00	-50151.45	12: α_ω = 0.42, β_ω = 0.42, κ = 3.62

We used `phydms` to incorporate each homolog's averaged preferences into ExpCMs. This table shows the results of using these models to optimize the branch lengths of the group-M and subtype-A trees shown in Fig 17. For both trees, the ExpCMs dramatically outperformed a standard codon-substitution model (YNGKP M5), based on differences in Akaike information criterion (AIC), with BG505's and BF520's preferences describing Env's evolution better than LAI's preferences. The stringency parameters (β) inferred for BG505 and BF520 were both $>$1, indicating that selection is more stringent in nature than our experiments for these homologs. In contrast, the stringency parameters inferred for LAI were $\sim$1, which may be due to experimental or biological differences, as discussed in the main text. As a control, for each homolog, we created an ExpCMs using preferences averaged across all sites in the protein. This averaging substantially decreased phylogenetic fit, indicating that the preferences capture site-specific selective pressures on Env in nature. The results for individual experimental replicates are shown in Tables 8 and 9.

Table 8: **Results of the `phydms` analysis with group-M sequences for individual experimental replicates.**

model	ΔAIC	log likelihood	parameters: optimized values
BG505-avg	0.00	-58185.53	7: β = 2.07, α_ω = 0.69, β_ω = 0.49, κ = 3.16
BF520-avg	70.26	-58220.66	7: β = 2.44, α_ω = 0.71, β_ω = 0.47, κ = 3.11
BF520-1	1149.48	-58760.27	7: β = 1.69, α_ω = 0.62, β_ω = 0.48, κ = 3.18
BG505-1	1169.24	-58770.15	7: β = 1.35, α_ω = 0.55, β_ω = 0.52, κ = 3.16
BG505-3	1188.94	-58780.0	7: β = 1.58, α_ω = 0.69, β_ω = 0.52, κ = 3.21
BG505-2	1215.68	-58793.37	7: β = 1.44, α_ω = 0.66, β_ω = 0.50, κ = 3.18
BF520-3	1550.10	-58960.58	7: β = 1.69, α_ω = 0.66, β_ω = 0.54, κ = 3.08
BF520-2	1613.64	-58992.35	7: β = 1.50, α_ω = 0.65, β_ω = 0.50, κ = 3.18
LAI-3b	2792.12	-59581.59	7: β = 0.87, α_ω = 0.56, β_ω = 0.44, κ = 3.17
LAI-avg	2994.80	-59682.93	7: β = 1.00, α_ω = 0.55, β_ω = 0.46, κ = 3.09
LAI-1	3402.28	-59886.67	7: β = 0.59, α_ω = 0.52, β_ω = 0.49, κ = 3.17
LAI-2	3561.38	-59966.22	7: β = 0.53, α_ω = 0.53, β_ω = 0.46, κ = 3.15
LAI-3	3594.92	-59982.99	7: β = 0.45, α_ω = 0.54, β_ω = 0.40, κ = 3.18
BF520-3 site averaged	3904.60	-60137.83	7: β = 2.42, α_ω = 0.50, β_ω = 0.42, κ = 3.21
BF520-avg site averaged	3913.08	-60142.07	7: β = 2.24, α_ω = 0.50, β_ω = 0.41, κ = 3.21
BG505-1 site averaged	3914.92	-60142.99	7: β = 1.51, α_ω = 0.51, β_ω = 0.42, κ = 3.21
BF520-1 site averaged	3920.58	-60145.82	7: β = 2.17, α_ω = 0.51, β_ω = 0.41, κ = 3.22
BF520-2 site averaged	3921.04	-60146.05	7: β = 2.09, α_ω = 0.50, β_ω = 0.41, κ = 3.21
BG505-3 site averaged	3921.64	-60146.35	7: β = 1.61, α_ω = 0.51, β_ω = 0.42, κ = 3.20
BG505-avg site averaged	3922.08	-60146.57	7: β = 1.53, α_ω = 0.51, β_ω = 0.42, κ = 3.21
BG505-2 site averaged	3933.90	-60152.48	7: β = 1.44, α_ω = 0.51, β_ω = 0.42, κ = 3.22
LAI-3b site averaged	4056.98	-60214.02	7: β = 1.05, α_ω = 0.51, β_ω = 0.40, κ = 3.26
LAI-avg site averaged	4083.20	-60227.13	7: β = 0.83, α_ω = 0.51, β_ω = 0.41, κ = 3.26
LAI-3 site averaged	4085.00	-60228.03	7: β = 0.73, α_ω = 0.51, β_ω = 0.41, κ = 3.27
LAI-1 site averaged	4093.64	-60232.35	7: β = 0.73, α_ω = 0.51, β_ω = 0.41, κ = 3.27
LAI-2 site averaged	4100.52	-60235.79	7: β = 0.61, α_ω = 0.51, β_ω = 0.41, κ = 3.27
YNGKP M5	4423.42	-60392.24	12: α_ω = 0.53, β_ω = 0.55, κ = 3.20

This table is similar to Table 7, but shows the results of ExpCMs for *all* experimental replicates of each homolog, instead of just the ExpCMs made after averaging a homolog's amino-acid preferences among its replicates. This table only shows the results of the analysis with group-M sequences; see 9 for the results of the analysis with subtype-A sequences. With the exception of LAI replicate 3b, the ExpCMs for individual experimental replicates always perform worse (based on ΔAIC) than the ExpCMs for the averaged replicates. Moreover, the inferred stringency parameters are always lower for individual replicates than for the averaged replicates. These patterns are both consistent with the idea that averaging across replicates increases the accuracy of our estimates by decreasing experimental noise. As in 7, the ExpCMs for individual replicates all outperform the standard model and decrease in performance when the preferences are averaged across all sites in the protein to create model that is not site-specific.

Table 9: **Results of the `phydms` analysis with subtype-A sequences for individual experimental replicates.**

model	ΔAIC	log likelihood	parameters: optimized values
BF520-avg	0.00	-48166.45	7: β = 2.80, α_ω = 0.59, β_ω = 0.33, κ = 3.53
BG505-avg	234.70	-48283.8	7: β = 2.19, α_ω = 0.58, β_ω = 0.32, κ = 3.50
BF520-1	1027.22	-48680.06	7: β = 1.90, α_ω = 0.57, β_ω = 0.30, κ = 3.61
BG505-3	1236.20	-48784.55	7: β = 1.63, α_ω = 0.53, β_ω = 0.33, κ = 3.59
BG505-1	1270.98	-48801.94	7: β = 1.48, α_ω = 0.55, β_ω = 0.33, κ = 3.53
BG505-2	1392.52	-48862.71	7: β = 1.53, α_ω = 0.54, β_ω = 0.39, κ = 3.41
BF520-3	1468.76	-48900.83	7: β = 1.88, α_ω = 0.61, β_ω = 0.34, κ = 3.47
BF520-2	1591.72	-48962.31	7: β = 1.68, α_ω = 0.54, β_ω = 0.34, κ = 3.65
LAI-3b	2813.66	-49573.28	7: β = 0.89, α_ω = 0.49, β_ω = 0.30, κ = 3.54
LAI-avg	2860.64	-49596.77	7: β = 1.06, α_ω = 0.48, β_ω = 0.30, κ = 3.41
LAI-1	3173.34	-49753.12	7: β = 0.68, α_ω = 0.46, β_ω = 0.33, κ = 3.49
LAI-3	3397.38	-49865.14	7: β = 0.45, α_ω = 0.49, β_ω = 0.30, κ = 3.49
LAI-2	3451.26	-49892.08	7: β = 0.51, α_ω = 0.47, β_ω = 0.34, κ = 3.48
BF520-3 site averaged	3900.54	-50116.72	7: β = 1.53, α_ω = 0.47, β_ω = 0.30, κ = 3.61
BG505-1 site averaged	3908.02	-50120.46	7: β = 0.92, α_ω = 0.47, β_ω = 0.30, κ = 3.61
BG505-2 site averaged	3908.50	-50120.7	7: β = 0.90, α_ω = 0.47, β_ω = 0.30, κ = 3.61
BG505-avg site averaged	3909.68	-50121.29	7: β = 0.93, α_ω = 0.47, β_ω = 0.30, κ = 3.61
BF520-avg site averaged	3913.70	-50123.3	7: β = 1.22, α_ω = 0.47, β_ω = 0.30, κ = 3.61
BG505-3 site averaged	3914.34	-50123.62	7: β = 0.94, α_ω = 0.47, β_ω = 0.30, κ = 3.60
BF520-2 site averaged	3917.84	-50125.37	7: β = 1.18, α_ω = 0.47, β_ω = 0.30, κ = 3.62
BF520-1 site averaged	3922.20	-50127.55	7: β = 1.20, α_ω = 0.47, β_ω = 0.30, κ = 3.63
LAI-3b site averaged	3959.56	-50146.23	7: β = 0.20, α_ω = 0.46, β_ω = 0.30, κ = 3.65
LAI-3 site averaged	3961.16	-50147.03	7: β = 0.03, α_ω = 0.46, β_ω = 0.30, κ = 3.67
LAI-avg site averaged	3961.18	-50147.04	7: β = 0.01, α_ω = 0.46, β_ω = 0.30, κ = 3.67
LAI-1 site averaged	3961.22	-50147.06	7: β = 0.01, α_ω = 0.46, β_ω = 0.30, κ = 3.67
LAI-2 site averaged	3961.30	-50147.1	7: β = 0.01, α_ω = 0.46, β_ω = 0.30, κ = 3.67
YNGKP M5	3980.00	-50151.45	12: α_ω = 0.42, β_ω = 0.42, κ = 3.62

This table is similar to Table 8, but shows the results of the analysis with subtype-A sequences. The trends seen for subtype A are similar to those for group M.