Title

The genetic interaction between HIV and the antibody repertoire

Authors

Nicolas Strauli[1], Emily Kathleen Fryer[2,3], Olivia Pham[2], Mohamed Abdel-Mohsen[4], Shelley N. Facente[5,6.7], Christopher Pilcher[8], Pleuni Pennings[2], Satish Pillai[6,9], Ryan D. Hernandez[10,11,*]

Affiliations

1. Biomedical Sciences Graduate Program, University of California, San Francisco, CA, USA
2. Department of Biology, San Francisco State University, San Francisco, CA, USA
3. Carnegie Institution for Science, Plant Biology Department, Stanford, California, USA
4. The Wistar Institute, Philadelphia, PA, USA
5. University of California, San Francisco (UCSF), San Francisco, CA, USA
6. Vitalant Research Institute, San Francisco, California, USA
7. Facente Consulting, Richmond, CA, USA
8. Division of HIV, Infectious Diseases and Global Medicine, University of California, San Francisco (UCSF), San Francisco, CA, USA
9. Department of Laboratory Medicine, University of California, San Francisco (UCSF), San Francisco, CA, USA
10. Bioengineering & Therapeutic Sciences, University of California, San Francisco (UCSF), San Francisco, CA, USA
11. Department of Human Genetics, McGill University, Montreal, QC, Canada

* Correspondence: ryan.hernandez@me.com

Abstract

The interaction between human immunodeficiency virus (HIV) and the antibody repertoire (AbR) during chronic infection can provide important information for HIV vaccine research, yet has not been well-characterized on a systems level. We deeply sequenced the HIV population and the AbR of ten HIV-infected, antiretroviral (ART)-naïve individuals, each with 10-20 longitudinal samples spanning 4-14 years. Our unbiased sequencing approach identified partitions of AbRs showing evidence of interaction with autologous HIV populations. We show that these HIV-associated partitions are enriched for the V gene segments of known HIV broadly neutralizing antibodies (bnAbs), indicating that the HIV-responding component of the AbR can be identified via time-series genetic data. Despite this evidence for larger-scale AbR/HIV interactions at the sub-population level, we found little to no evidence for antagonistic coevolution (i.e. an arms

race). This suggests that antagonistic coevolution is either rare, or hard to detect, which has important vaccine design implications.

Introduction

Since the beginning of the modern pandemic in 1981 (Faria et al., 2014; Gottlieb et al., 1981), human immunodeficiency virus (HIV) has been the source of incredible scientific scrutiny. While there has been great progress in the development of antiretroviral therapy (ART), which can now manage the disease indefinitely (albeit only for those who have access to them), a cure remains elusive (Stein, Storcksdieck genannt Bonsmann, & Streeck, 2016), and little progress has been made toward a preventative vaccine (Trovato, D'Apice, Prisco, & De Berardinis, 2018). Prevention efforts have recently made significant headway by implementing preexposure prophylaxis (PrEP) to high risk individuals. However, in its current form, this strategy has its downsides, such as a reliance on daily self-administration, significant financial burden, and health side effects (Weinstein, Yang, & Cohen, 2017). Thus, cure and vaccine strategies remain the elusive goal for HIV research. Together, the distinct gains in treatment, yet relative lack of gains in HIV prevention, has resulted in a stalemate of sorts, where instead of HIV being triumphantly eradicated by modern science, it has settled into a persistent, yet treatable reality of human life.

The fervent hope for progress is particularly palpable in HIV vaccine research. This fervor is mainly fueled by the fact that effective HIV immunity is entirely possible and well documented, as a broadly neutralizing humoral immune response against HIV occurs naturally in 10-20% of those chronically infected (Binley et al., 2008; Hraber et al., 2014; Sather et al., 2009). It seems that while humans universally cannot clear an HIV infection, some can develop immunity to future infections. Moreover, potent humoral immunity has been associated with spontaneous immunological suppression of HIV following disease acquisition in the absence of ART (Freund et al., 2017). If a vaccine were to be designed that could somehow recapitulate the broad and potent anti-HIV humoral immune responses observed in naturally resistant individuals, then effective protective immunity may be achievable. While a good deal of progress has been made in this avenue of research, it has not yet resulted in an effective vaccine. Among the promising discoveries are broadly neutralizing antibodies (bnAbs), which are monoclonal antibodies (Abs) that can single-handedly neutralize up to 90% of heterologous HIV strains (McCoy & Burton, 2017; X. Wu et al., 2011). To study the development of these bnAbs, researchers have utilized a post-hoc deep-sequencing approach by which a set of primers are developed that will preferentially amplify a subset of the antibody repertoire (AbR)—the population of antibodies in an organism—that is known to contain a bnAb lineage (Bhiman et al., 2015; Liao et al., 2013; X. Wu et al., 2011). The advantage of this approach is that one can cut through the incredible noise and complexity of the AbR to focus on a particular lineage of importance. However, such approaches miss the diversity of Ab lineages interacting with HIV that may have important effects on

immunity outcomes. To our knowledge, only 2 studies (Hoehn et al., 2015; Setliff et al., 2018) have deeply sequenced the AbR in an unbiased fashion in the context of HIV infection, but they did not collect paired HIV sequence data to directly study genetic interactions.

The common narrative of bnAb development is that they are in a coevolutionary arms race (Dawkins & Krebs, 1979) with the autologous HIV population (Rantalainen et al., 2018; Zhou, Ton, Morriss, Nguyen, & Fera, 2018). This relationship specifically refers to 'antagonistic coevolution', which for simplicity, we will here to refer to as 'coevolution'. There is good reason to suspect that this coevolutionary hypothesis is true: bnAbs tend to be quite derived relative to their inferred naïve ancestors (Bhiman et al., 2015), they tend to take a long time to develop (years), and there tends to be a time dependence on neutralization capabilities (i.e. HIV-neutralizing Abs are more likely to neutralize autologous virus from the past, and less likely to neutralize contemporaneous or future autologous virus) (Gray et al., 2011; Liao et al., 2013). However, there is also evidence contrary to the arms race hypothesis: bnAbs can arise relatively quickly, and with few mutations (Nicole A. Doria-Rose et al., 2014; Goo, Chohan, Nduati, & Overbaugh, 2014; MacLeod et al., 2016; Simonich et al., 2016), and superinfections—multiple HIV infections in the same individual—don't necessarily drive further evolution in existing HIV-neutralizing Ab lineages, but rather promote the development of *de novo* HIV-targeting Ab lineages (Sheward et al., 2018; Williams et al., 2018). Whereas in a paradigm of coevolution, one might expect that novel epitopes introduced by a superinfection would spur the development of existing HIV-targeting Ab lineages to evolve innovations to regain affinity. Similarly, in the context of malaria infection, repeated immunizations with a complex malaria antigen promote the activation of *de novo* naïve Ab lineages, rather than the evolution of already existing malaria-targeting Abs (Murugan et al., 2018).

A better understanding of the interaction between HIV and the AbR over time could shed light on how HIV immunity is achieved. However, longitudinal sequence studies of HIV infections tend to either focus on HIV (Feder, Kryazhimskiy, & Plotkin, 2014; Fischer et al., 2010; Henn et al., 2012; Shankarappa et al., 1999; Zanini et al., 2015) or the AbR (Hoehn et al., 2015; Nourmohammad, Otwinowski, Luksza, Mora, & Walczak, 2018; Xueling Wu et al., 2015), but not both. To our knowledge, only three studies have deeply sequenced the AbR along with the autologous HIV population, but each of these studies consisted of a single individual with relatively few time-points (Bhiman et al., 2015; Landais et al., 2017; Liao et al., 2013). In this study, we ameliorate this dearth of data, and describe in unprecedented detail the genetic interaction between these putatively coevolving populations in 10 well-characterized HIV-infected individuals.

Results

Sequencing HIV *env* and IGH

We collected a total of 119 cryopreserved peripheral blood samples from the OPTIONS cohort at the University of California, San Francisco (UCSF). The samples all originated from male participants aged 25-48 years old at the estimated time of infection in San Francisco, and each participant had 10-20 longitudinal samples (Table 1, and Figure S1). The samples were all collected prior to administration of ART, with the exception of the last time-point of participants 1, 2, and 5. We chose to deeply sequence the C2-V3 region of the *env* gene because of its rich history of interactions with Abs, as evidenced by the HIV epitope map from Los Alamos National Labs (LANL) (Hatada et al., 2010; Ringe et al., 2012; Yusim et al., 2016). Of the 116 ART-naïve samples, we were unable to successfully amplify C2-V3 from 12. In 11 of these cases, low viral load was the presumed cause for lack of amplification, but the 9th time-point of participant 6 was unsuccessful despite high viral load (Figure 1, S2). We were also unable to amplify C2-V3 from the 3 ART-exposed samples. While viral load measurements were not available for these samples, the presumed cause for our inability to amplify C2-V3 was low viral load, given their ART status. The initial *env* sequencing depth ranged from 3,771-101,831 reads per sample, and after cleaning the data with several quality control (QC) steps, this ranged from 2,276-56,914 reads per sample (Figure S3).

We also deeply sequenced the variable region of the immunoglobulin heavy chain locus (IGH), the product of which we refer to as the antibody repertoire (AbR). The AbR was successfully sequenced in all 119 samples, with the exception of the fourth time-point of participant 3. Sequence data was generated for this sample, but it exhibited an abrupt clonal expansion of a magnitude that was a clear outlier for participant 3, and not seen in any other sample, so it was discarded (Figure S4, and 1). Initial AbR sequencing depth ranged from 669,331-2,669,662 reads per sample, and after QC this ranged from 160,291-552,479 reads per sample (Figure S5).

Characterizing the HIV population and AbR over time

In order to quantify the broad attributes of the AbR and HIV populations over time, we calculated a variety of summary statistics that characterized the genetic diversity, divergence, selection, and abundance for each of the populations. Selection pressure (positive or negative) in HIV was estimated as dN/dS (Nei & Gojobori, 1986), and in the AbR was estimated using the Σ values from BASELINe (Yaari, Uduman, & Kleinstein, 2012). As others have reported (Maldarelli et al., 2013; Zanini et al., 2015), we found that HIV diversity, divergence, and selection (dN/dS) all tend to increase with time since estimated date of infection, with large perturbations over smaller time-scales (Figures 2, S6, and S7). Of note, the high viral load of the first time-point of participant 7 suggests that the acute viremia phase of early HIV infection was captured. Additionally, participant 10 had very low viral load for the first 8 time-points, which

explains why amplification of HIV *env* was unsuccessful for these samples. Participant 6 exhibited strong evidence for a super-infection occurring between the 1st and 2nd time-points (Figure S8). Super infections are not uncommon with HIV (Redd, Quinn, & Tobian, 2013), and will cause a sudden injection of 'artificial' genetic divergence relative to the initial infecting virus. Thus, we accounted for this superinfection when calculating divergence summary statistics for participant 6 by comparing each sequence to both the initial infection, as well as the subsequent superinfection (see Methods).

The trajectories of the AbR summary statistics did not show any obvious stereotyped pattern across participants (Figures 3, S9, and S10). However, there were a couple data points that suggested interactions with the HIV population: the second time-point of participant 7 (0.41 years post infection) showed a large increase in selection (in both the framework regions, FWR, and complementarity determining regions, CDR), which could be a response to the initial viremia in the prior time-point; and the ninth time-point of participant 10 (3.37 years post infection) also showed a large increase in selection with a concomitant drop in diversity, which could have been a response to the large increase in viral load at that time-point.

To increase the accessibility of these data, we created a web application, available at https://ab-hiv-coevolution.github.io/HIV_AB_CoEvo/, for others to investigate this new dataset interactively. It allows users to explore the data by individual patient or all patients across various summary statistics including genetic diversity, divergence, selection, and abundance for HIV and/or the AbR.

Testing for whole-population level interactions

In order to gauge the effect HIV has on the AbRs of participants, we first pooled all the data across participants, and used a regression framework to test if any of the AbR summary statistics were significantly correlated with that of the HIV population (while controlling for participant-specific effects, see Methods). We performed this test in a pairwise fashion on all AbR summary statistics against all HIV summary statistics. Similar to Hoehn's work (Hoehn et al., 2015), we found no significant correlation between AbR diversity and viral load, yet we did find a modest, statistically significant correlation between AbR and HIV diversity ($p= 0.02$, Figure 4 A). While this association was marginally significant—indeed, it ceases to be significant after controlling for multiple tests—we found any association at the whole-population level surprising, and thus, worth reporting.

It is possible that some participants' AbRs interact with their autologous HIV population more than others, thus we also tested for interactions between summary statistics on a participant-by-participant approach. Because the number of data points per participant is relatively low, we opted to use a permutation based test in order to accurately estimate type 1 error (Ernst, 2004) (see Methods). In participant 2, we found that selection (dN/dS) and viral load in the

HIV population were associated with divergence and selection (FWR) in the AbR, respectively (p=0.009 and p=0.0425, Figure 4 B, and C). We also found that in participant 7, viral load and diversity in the HIV population were associated with selection (FWR) and divergence in the AbR, respectively (p=0.040 and p=0.040, Figure 4 D, and E). All of these associations were positive correlations, with the exception of HIV selection (dN/dS) and AbR divergence in participant 2, which was anticorrelated.

Together, these data raise the possibility that the effect that HIV has on the AbR (and vice versa) is large enough to be observable at the intra- participant, whole population level, although more samples would be needed to make definitive statements about this.

Identifying partitions of the AbR that interact with HIV

The AbR is an exceedingly complex population consisting of a myriad of Ab lineages capable of simultaneously binding and neutralizing a countless number of antigenic targets. To identify specific parts of the AbR that may be interacting with HIV, we first partitioned the AbR across time based on the germline identity of each sequence's V and J gene segments (Figure S11), and then tested each AbR partition for evidence of interactions with the autologous HIV population using similar summary statistics as the overall population (see Methods). Using an analogous permutation-based test as was used when comparing the overall populations, we found significant associations between AbR-partition trajectories and HIV trajectories in participants 3, 7, and 8 (Figures S12, S13, and 5). Of these associations, AbR-partition frequency tended to be most often associated with viral load. For example, participant 7 had a distinct viral load trajectory—presumably due to acute viremia—and the frequency trajectory of the IGHV4-31:IGHJ5 AbR partition was positively associated with the unique shape of this trajectory (while the diversity of this AbR partition was negatively associated with viral load). This suggests a clonal expansion occurred in this AbR-partition in response to HIV, which caused an increase in the partition's frequency with a concomitant drop in diversity. Similarly, in participant 8, the frequency trajectory of two AbR partitions with the same V gene segment (IGHV6-1:IGHJ5 and IGHV6-1:IGHJ4) were positively associated with viral load, suggesting that the IGHV6-1 gene segment in this participant may have had a predisposition to targeting HIV. In participant 3, the frequency trajectory of the IGHV3-30:IGHJ3 partition was negatively associated with both non-synonymous divergence and selection (dN/dS) in the HIV population, suggesting that escape mutations in HIV caused a drop in frequency of the interacting AbR partition.

Validating the HIV-associated AbR partitions

In order to establish that our permutation-based test is in fact identifying AbR partitions that had a biological response to HIV infection and was not the result of random noise in the data, we sought to compare our results to previous findings

in the literature. We first used Fisher's method to compile all the results of our permutation-based test into a single score for each V gene segment (see Methods), and then used the database of HIV bnAbs from bNAber (Eroshkin et al., 2014) to compare the incidence of known HIV-binding V gene segments in the literature, to how well these V gene segments score in our test (Table S1). We found that V gene segments that had been shown to bind HIV in the literature dataset were associated with significantly higher scores in our permutation-based test (p=1.52e-3, Mann-Whitney U test, Figure S14 A, and C). There are two reasonable explanations for this: i) these V gene segments have a predisposition to bind HIV (i.e. a 'public' or convergent response (Setliff et al., 2018)), or ii) These V gene segments have a predisposition to bind anything (due to high endogenous expression, having broad affinity for viruses, or some other unknown reason). In order to differentiate between these two possibilities we performed a similar test except instead of comparing our results to Abs known to target HIV, we compared our results to a literature dataset that we previously compiled of Abs that have been shown to bind to influenza (Strauli & Hernandez, 2016) (Table S1). Similar to the HIV literature dataset, we found that V genes that were well represented in the influenza literature dataset also tended to score highly in our permutation-based test (p=7.28e-4, Mann-Whitney U test, Figure S14 B, and D). Therefore, while we are likely identifying a biological response to HIV, the response may not be specific to HIV.

Testing for coevolution

Coevolution between HIV and a handful of well-known bnAbs has been extensively reported (Liao et al., 2013; Rantalainen et al., 2018) and reviewed (Bonsignori, Liao, et al., 2017). Coevolution provides an intellectually compelling explanation for the development of bnAbs against HIV, however, examples tend to be anecdotal and qualitative (likely due to small sample sizes). While we cannot be sure that bnAbs exist in our data, we sought to test if coevolution is a predominant driver of HIV-targeting Ab development generally. We tested for genetic signals of coevolution in our data by first dividing the AbR data of each participant into time-course lineages of Abs (Figure 6, S15, and S16). We then used MAFFT (Katoh, Misawa, Kuma, & Miyata, 2002) to create a multiple sequence alignment (MSA) of each Ab lineage, and compare each of these Ab lineage MSAs with a representative MSA of the HIV population overtime using a mutual information (MI) statistic (Brandman, Brandman, & Pande, 2012; Gloor, Martin, Wahl, & Dunn, 2005; Marino Buslje, Teppa, Di Doménico, Delfino, & Nielsen, 2010; Simonetti, Teppa, Chernomoretz, Nielsen, & Marino Buslje, 2013). Importantly, we reduced the complexity of the amino acid code to a code of 'change' or 'no-change' prior to calculating MI (see Methods).

If coevolution is a common characteristic of HIV-targeting Ab lineages, then we might expect that the HIV-associated AbR partitions that we previously identified will have abnormally high MI values. Thus, we compared the mean MI values of the lineages within all of the HIV-associated AbR partitions (IGHV3-30:IGHJ3 of

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

Lastly, we test for a global coevolutionary signal, agnostic to whether or not a lineage belongs to an HIV-associated partition. To do this we gathered the mean MI values across all of the observed lineages within a participant, and then compare this distribution to that of the mean MI values from the simulated null lineages (Figure 8). If Ab/HIV coevolution were taking place on a large scale, we would expect to see a shift towards higher MI values in the observed distribution relative to the simulated null. However, we saw no evidence of this, and instead saw that, if anything, the simulated null lineages tend to have higher MI values. This suggests that if coevolution is taking place at all, it is either a weak force or rare in these participants.

Discussion

In this study we have created a large dataset of Ab sequences and HIV sequences from 119 longitudinal samples. This is, to our knowledge, the largest dataset of its kind. The HIV literature currently encompasses an abundance of AbR sequence datasets from HIV+ individuals. However, these datasets primarily originate from the amplification of a particular Ab lineage that was known to contain an HIV-bnAb, prior to deep sequencing. These ‘biased’ AbR datasets are useful to home in on the development of a particularly interesting Ab lineage. However, we argue that it is equally important to understand the AbR response to HIV infection on a global/systems scale for the following reasons: i) It is just as important to understand why an HIV-targeting Ab lineage failed to develop broad neutralization ability, as it is to understand why a lineage succeeded in developing it. ii) There is increasing evidence that non-neutralizing Ab lineages play important roles in viral pathogenesis through Fc-mediated functions (e.g. antibody-dependent cellular cytotoxicity (Horwitz et al., 2017; Mayr, Decoville, et al., 2017; Mayr, Su, & Moog, 2017). iii) It is quite possible that a significant proportion of humans that develop broad immunity to future HIV infections, do so in a polyclonal manner (Williams et al., 2018), where broad neutralization depth against HIV is achieved via the cooperative action of many Ab lineages, each simultaneously targeting different epitopes, or different versions of the same

epitope. iv) The population dynamics of HIV-bnAbs (in addition to HIV-binding Abs in general) are poorly understood. For example, do these Abs persist at high or low frequency in the greater population? How does this frequency change over time? What type of selection drives their development (ex: positive, negative, balancing, etc.)? These types of questions are difficult, if not impossible to answer without understanding the larger population context for which these Ab lineages exist. In this study, we have taken the preliminary steps towards addressing these types of questions. Specifically, we have developed a statistical approach to identify the partitions of the AbR that are likely responding to HIV. Once this has been established, questions like those enumerated above can be answered. Further, we hope that the sequence datasets we have created here will provide a useful resource for others with similar lines of inquiry.

While (Hoehn et al., 2015) found no correlation between AbR diversity and viral load in their data, they did not have the means to address other characteristics of the HIV population, as they did not include HIV sequence data in their analyses. However, they did find that AbR diversity was lower in HIV+ individuals than healthy controls. This suggested that HIV may have a broad effect on the AbR, yet the details of this effect remained unclear. We have presented a small positive correlation between overall AbR and HIV diversity across all our samples. A possible scenario that would explain this observation is one where AbR diversity is decreased due to clonal expansions in Ab lineages that target HIV, which in turn causes a decrease in HIV diversity, due to positive selection for escape mutations. Once the HIV population has escaped, its diversity will return, and diversity in the AbR will also return because the previous clonal expansion will have vanished due to its target having escaped. However, we stress that this correlation had nominal significance and should be treated cautiously. The AbR is an especially complicated population that is capable of simultaneously responding to countless antigens, thus even a fleeting correlation with HIV at the whole-population level may be worthy of further follow up studies.

A key first step towards illuminating the global interaction between the AbR and an HIV infection is to be able to identify the subset of the AbR that is responding to HIV. Similar to our previous work in the context of influenza vaccination (Strauli & Hernandez, 2016), we leveraged the time-series nature of our dataset to identify partitions of the AbR that seem to be associated with HIV. We purposefully made no prior assumptions about what types of interactions we might find. For example, the common narrative of HIV-targeting Ab lineages is that they are under intense positive selection. Thus, one might have the expectation that Ab selection will be positively correlated with HIV selection. However, it is also possible that an HIV-targeting Ab lineage is under intense negative selection, where there is a preference for amino acids *not* to change so as to not ablate their binding ability, or perhaps to not change a strict structural conformation that is required to access an epitope. In this case, one might expect Ab selection to be negatively correlated with HIV selection. We therefore compared all AbR summary statistics to all HIV summary statistics. This all-by-all

comparison gave us the privilege of an unbiased approach, yet greatly increased the number of tests that were performed, and hence the severity of our multiple-tests correction. As such, we were only able to identify a handful of AbR partitions that were significantly associated with HIV. This suggests that long-term interactions between Ab lineages and HIV are rare, and that Ab/HIV interactions may be of a more transitory nature, where an antibody binds to HIV, then HIV escapes, and then another unrelated Ab binds to the escape mutant, and so on. Another possibility is that our test was simply underpowered and had many false negatives. One way to ameliorate this would be to first filter AbR partitions based on some statistic (e.g. divergence) and then test for associations using a different, orthogonal statistic (e.g. diversity). Further, there remains a great deal of powerful analyses that could be done with the HIV sequence data. In principle, one could divide the HIV population into lineages and test each HIV lineage against each AbR partition. This would increase the number of tests, but could also illuminate interactions that would be otherwise hidden.

Lastly, we tested for a coevolutionary signal in our data. Coevolution in sequence data is notoriously hard to establish (Avila-Herrera & Pollard, 2015), and to our knowledge, reports of HIV/Ab coevolution to date have been universally qualitative, with little or no statistical analyses (Bhiman et al., 2015; Bonsignori, Kreider, et al., 2017; Bonsignori et al., 2016; Nicole A. Doria-Rose et al., 2014; Gao et al., 2014; Landais et al., 2017; Liao et al., 2013; MacLeod et al., 2016; Rantalainen et al., 2018). When it comes to claims of coevolution, there are two sources of uncertainty that we have attempted to account for in this study. i) When both the Ab lineage and the putative HIV epitope are under positive selection, it is very easy for mutations to be correlated by chance rather than by coevolution. ii) There is a huge amount of uncertainty when assigning Abs to lineages, especially when trying to link a given Ab sequence to other Ab sequences that existed months-years prior. AbRs are incredibly dynamic populations with high turnover, and high mutation rates. In a population such as this, where *de novo* lineages are continuously being added, it is important for one to account for the possibility that two similar Ab sequences—even if strikingly similar—may not be of the same lineage. By creating a simulated null dataset from shuffled Ab lineages, we were able to create a null distribution of MI values that took both of these confounders into account. After doing this we found no global signal for coevolution, yet we did find one isolated Ab lineage in participant 8 that showed compelling evidence for it. This suggests that while coevolution between Ab lineages and HIV is possible, it is likely exceedingly rare and/or hard to detect. Given that other sequence datasets of the AbR in the context of HIV infection have about the same or fewer time-points than the participants in our dataset, we suspect that claims of coevolution in these data would be equally hard to make.

Coevolution has been responsible for some of the most remarkable phenotypes known (e.g. the cheetah's speed, a flower’s beauty, the strangeness of genitalia (Brennan & Prum, 2015)), yet it remains unclear as to how much of a role it plays

in the development of HIV-targeting Abs. This has important implications for vaccine strategies. If coevolution is the predominant force in the development of HIV-bnAbs, then a vaccine regimen that mimics the natural evolution of the HIV epitope *in vivo* would be desired, as this would recapitulate the coevolutionary process leading to increased neutralization breadth. However, it is also possible that HIV-bnAbs occur as rare, random events. Examples of this include unlikely V(D)J recombination events whereby a naïve Ab lineage is created that happens to have a predisposition to target a conserved HIV epitope, or a (typically diverged) Ab lineage 'stumbling' upon broad neutralization breadth by chance. In this case, one might desire a vaccine regimen that has a very diverse array of HIV epitopes so as to maximize the chances that these rare events occur. This approach is supported by the fact that neutralization breadth is positively correlated with viral load and HIV diversity (N. A. Doria-Rose et al., 2010; Landais et al., 2016; Piantadosi et al., 2009). An interesting future study would be to use mathematical modeling (Nourmohammad, Otwinowski, & Plotkin, 2016), or simulation frameworks (Murugan et al., 2018) that include the introduction of novel naïve Ab lineages into the population, to gain a better understanding of which evolutionary parameters (e.g. population size, mutation rate, selection strength, population diversity, etc.) in the HIV and AbR populations promote coevolution, and which do not. We hope that this study has provided a sound example of how to go about formally testing for coevolution in order to differentiate between these two possibilities.

Methods

Participant selection and sample processing

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

Viral load

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

Estimated time of infection

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

C2V3 amplification and ultra-deep sequencing

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

IGH amplification and ultra-deep sequencing

Total RNA was extracted from PBMCs using the Qiagen RNeasy Mini Kit. To reverse transcribe, amplify IGH encoding RNA, and generate sequencing-ready libraries, we used iRepertoire's long read iR-Profile Kit and followed the procedure as described in the accompanying protocol (Wang et al., 2010). Paired-end sequencing (2×300 bp) was performed using the Illumina MiSeq instrument and the MiSeq Reagent Kit v3.

HIV sequence data QC

Sequences from different samples were de-multiplexed by barcode using the internal software on the Illumina machine. We used the software package pRESTO (Vander Heiden et al., 2014) to assemble read pairs, remove sequences shorter than 300bp, remove sequences with a mean quality score less than 30, mask the primer sequences, and remove sequences that only occur once in a given sample. We then use an in house implementation of BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990) to check that each sequence has at least 70% identity to at least one HIV subtype *env* reference sequence, which were downloaded from LANL (Foley et al., 2017). In order to check for possible contaminations from HIV sequences outside of our study, we again used BLAST to map each of our sequences to every *env* sequence within the LANL database. In this case, any of our sequences that had 99% identity or more to any sequence within the *env* database would be deemed a contaminant. We found that all the samples from participant 8 in our study had a significant amount of identity with sequences derived from patient ID: 9036 in the LANL database. We also found that all the samples from our participant 9 had significant identity with sequences in the LANL database derived from patient ID: 9018. There are two reasonable explanations for this: i) these samples had a large degree of contamination, or ii) that our participants 8 and 9 are the same as patients 9036 and 9018 in the LANL database, respectively. We conclude that the latter is the more likely explanation because of the following rationale. This large degree of 'contamination' only occurred in participants 8 and 9, and it occurred in all their samples, however, the samples from these participants were processed in different batches. These participants' diversity, and divergence trajectories showed a relatively steady increase over time, which would not be consistent with contamination (see results). Lastly, patients 9036 and 9018 from the LANL database both correspond to the study (Sturdevant et al., 2015) which also recruited participants from San Francisco, CA. None of the other samples in our study had detectable contamination using this method.

To check for cross contamination of sequences across samples in our study, we used a clustering approach. We first reduced the size of the dataset by grouping the sequences within each sample that have an edit distance less than or equal to 4 (see "Clustering sequences with samples" section below). We then choose the sequence within each group (or cluster) that had the highest count to be the 'representative sequence' for that cluster, after which we pooled all the representative sequences across all samples and clustered those pooled representative sequences using the same clustering algorithm. To identify clusters of sequences that were likely cross contaminants we used the following criteria: the cluster had to i) have a representative sequence that clustered closer with sequences from a different participant than with sequences from the same sample, and ii) have a frequency less than 0.001 within its sample. All sequences within clusters that satisfied these criteria were removed. This effectively identified low frequency sequences that were closer in genetic distance to sequences from another participant. Lastly, we used a phylogenetic approach (see "Making phylogenetic trees" section below) to i) make phylogenetic trees of

the representative sequences in each sample, and ii) remove any sequence that is more closely related to a representative sequence that was identified as a cross contaminant (as described above) then to the other representative sequences in the sample (Figure S17).

Clustering sequences within samples

We use the Needleman-Wunsch algorithm (Needleman & Wunsch, 1970) as implemented in the 'needle' program from the European Molecular Biology Open Software Suite (Rice, Longden, & Bleasby, 2000) to globally align each pair of sequences, and calculate the edit distances. Through an in-house algorithm, we then grouped sequences into a cluster that had an edit distance less than some provided threshold to any other sequence in the cluster. For HIV sequences, this threshold was 4; for AbR sequences, this threshold was 6.

Phylogenetic analysis

To make a phylogenetic tree of a group of sequences, we first made multiple sequence alignments using MAFFT (Katoh et al., 2002), and then constructed phylogenetic trees using FastTree (Price, Dehal, & Arkin, 2010). Visualization and analyses of newick formatted files was performed using the ETE toolkit (Huerta-Cepas, Serra, & Bork, 2016).

Calculating HIV divergence

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

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

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

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

To estimate non-synonymous and synonymous divergence we then use

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

(Nei & Gojobori, 1986). The notation above was heavily borrowed from Richard Orton's blog post (Orton, 2014). Because participant 6 had two reference sequences, divergence for a given query sequence was calculated relative to both reference sequences, and the lower value was used.

Calculating HIV selection

Selection in a given HIV sequence was estimated by taking the ratio of the non-synonymous over synonymous divergence values (described above), dN/dS.

Translating HIV sequences

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

Calculating diversity

To estimate diversity in both the HIV population and the AbR we calculated the statistic, π (Strauli & Hernandez, 2016). In words, π is the expected genetic distance between two randomly selected sequences from a given sample. Mathematically π can be expressed as

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

where N gives the total number of unique sequences in the sample, c_i gives the count of sequence i, M gives the total counts of sequences in that sample (i.e. $M = \sum_{i=1}^{N} c_i$), and $G(x, y)$ gives the genetic distance between sequences x and y. We used VSEARCH with the "--allpairs_global" option to globally align all pairs of sequences in a sample (Rognes, Flouri, Nichols, Quince, & Mahé, 2016). Genetic distance between a pair of sequences was then calculated as the percent of mismatches in the alignment.

AbR sequence data QC

As with the HIV data, we used pRESTO to assemble read pairs, remove reads less than 300bp, remove reads with a mean Q score less than 20, and remove reads that only occur once. We then use IgBLAST (Ye, Ma, Madden, & Ostell, 2013) to align each sequence to a database of germline immunoglobin genes downloaded from the IMGT website (imgt.org) (Giudicelli et al., 2006). After this, we used Change-O to annotate each sequence with its most likely V, D, and J

germline gene-segments, identify the FWRs and CDRs, and to construct the likely naïve antibody sequence from the germline gene-segment alignments.

Calculating Ab divergence

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

Estimating Ab Selection

We ran the BASELINe program on each individual Ab sequence and used the resulting Σ value to estimate selection pressure. For a detailed description of this tool see (Yaari et al., 2012). Very briefly, BASELINe compares the observed number of mutations in a sequence (relative to its inferred naïve ancestor) to a null distribution of the expected number of mutations under no selection. The program takes local nucleotide motifs into account when calculating mutation probabilities and returns a Σ value that indicates the distance between the observed number of mutations and the null distribution. A negative Σ indicates fewer mutations than expected (negative selection), and a positive Σ indicates more mutations than expected (positive selection). It does this separately for different regions of the Ab sequence (i.e. the FWR and CDR).

Creating AbR lineages

To cluster the AbR of a given participant, we first divided it into partitions by grouping together all sequences that use the same germline V and J gene-segments (as annotated by Change-O). We then clustered the sequences within a given partition/time-point (see “Clustering sequences within samples” section above), with an edit distance threshold of 6 (Figure S18 A). Once clusters were delineated, and similar to the clustering of HIV sequences, we assigned the most numerous sequence of each cluster as the ‘representative sequence’. We then linked clusters, within a given partition, across adjacent time-points using the following algorithm: We first found the representative sequence in the previous time-point that had the smallest edit distance to a given representative sequence in a contemporary time-point. If this edit distance was smaller than 30 (Figure S18 B), then the two representative sequences (and the clusters they represented) were linked as being part of the same lineage. This process was carried out independently in each participant, over each representative sequence, and for each adjacent time-point pair (see Figure S19 A). Finally, once all lineages were assigned in this manner, any lineage that did not rise above 0.0001 frequency, in any of the time-points was disregarded.

Simulating null AbR lineages

To simulate random AbR lineages, we began by using our clustered Ab sequences within the partitioned AbR data (see "Creating AbR lineages" section above). We then carried out an identical procedure as was done to create the observed lineages, with the exception that instead of finding the parent cluster in the previous time-point that had the minimum edit distance, a parent cluster was randomly chosen from the previous time-point (Figure S19 B). Additionally, in order to replicate the aspect of the observed data where each time-point brought a certain number of new lineages to the population, we estimated the probability of a new lineage as

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

where $l_{i,j}$ gives the number of new lineages, and $C_{i,j}$ gives the total number of clusters in participant, i, and time-point, j. For the simulated lineages, we then randomly assigned each cluster as being a new lineage (i.e. not having any connections with the previous time-point) with probability $r_{i,j}$. In order to have a null dataset that was sufficiently large, we duplicated the observed, clustered data N times and then simulated lineages using this N-fold larger dataset. When comparing all lineages to simulated data (i.e. Figure 8), N=10; when comparing lineages from particular AbR partitions to simulated data (i.e. Figure 7), N=100.

Linear modeling of population level interactions

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

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

where $Y_{i,j}$ gives the value of a given HIV summary statistic for the jth time-point of participant i, $X_{i,j}$ gives the value of a given AbR summary statistic for the same participant/time-point, β_i is a random intercept term to correct for participant specific effects, and $\epsilon_{i,j}$ is a random error term that is assumed to be normally distributed with a mean of 0. This model was implemented in the R programming language using the 'lmer' function of the 'lme4' package (Bates, Mächler, Bolker, & Walker, 2015). We then used a likelihood ratio test to determine if a model with $\beta \neq 0$ provided a significantly better fit than a model with $\beta=0$ ($p \leq 0.05$). If it did, then the given pair of HIV and AbR summary statistics was deemed to be interacting.

Trajectory permutation test

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

deviation. If v is a row vector representing a given trajectory, then v is standardized by

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

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

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

where r' is a standardized HIV trajectory vector. This gives the observed SSE values for each AbR/HIV trajectory pair. If either the HIV or AbR trajectories have missing values, then these time-points are disregarded in the SSE calculation, and a trajectory must have at least 75% of its values defined to be included in the test. We then permute the columns of the AbR trajectory matrix many times and calculate the SSE values for each permuted AbR trajectory after each permutation. This gives our permuted null distribution of SSE values. If an observed AbR/HIV trajectory pair have an SSE value that is significantly outside of this null distribution, then they are deemed to be significantly associated with one another, where significance is appropriately adjusted as based on the number of tests. When conducting this test on whole AbR population trajectories vs. whole HIV population trajectories (Figure 4 B-E), we performed 100,000 permutations for each participant. When conducting this test on AbR partition trajectories vs. whole HIV population trajectories (Figures 5, S12, and S13) we performed 1,000 permutations for each participant.

Comparison against published datasets

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

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

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

Where $\chi^2_{2|p|}$ is the chi-squared distribution with $2|p|$ degrees of freedom. Strictly speaking, $p_{overall}$ will tend to be inflated because not all the p values in p are independent (ex. AbR partition trajectories that are associated with HIV selection, will also tend to be associated HIV non-synonymous divergence). However,

because this inflation should be the same across V gene segments, and because we were not interested in actual significance but rather needed some reasonable method for combining p values into an overall score, Fisher's method should be sufficient. Finally, to avoid $p_{overall}$ being interpreted as a significance level, we took its log transform to arrive at a V gene score

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

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

Calculating mutual information (MI)

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

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

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

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

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

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

Correcting for multiple tests

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

Network plots

Networks of lineages were visualized using cytoscape (Shannon et al., 2003).

Software

All computer code written for this study is available on GitHub, at https://github.com/nbstrauli/abr_hiv_evo. This is accompanied by a flow chart which shows each data-processing step with the corresponding name of the script that carries it out (Figure S20).

Data availability

All sequence data associated with this study is available at the sequence read archive (SRA) under the BioProject ID: PRJNA543982.

Website

To create an interactive web application, we built a website using the Boostrap framework for the frontend interface along with the JavaScript graphing library plotly.js. The website is available at https://ab-hiv-coevolution.github.io/HIV_AB_CoEvo/

Acknowledgements

First and foremost, we would like to thank the anonymous participants who donated their data, time, and person to make this study (and those like it) possible. We will never know who you are, but we are eternally grateful for your contributions. We thank Dr. Manon Eckhardt for her edits and figure contributions. We would like to thank Alf Seccombe for driving a poor graduate student across town in order to get MiSeq kits into a proper freezer. Lastly, We thank Dr. Kenneth Hoehn for valuable manuscript feedback. This research was supported by grants from the National Institutes of Health (R01HG007644 to RDH), and University of California, San Francisco-Gladstone Institute of Virology & Immunology Center for AIDS Research (P30-AI027763)

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Avila-Herrera, A., & Pollard, K. S. (2015). Coevolutionary analyses require phylogenetically deep alignments and better null models to accurately detect inter-protein contacts within and between species. *BMC Bioinformatics*, *16*(1), 268. https://doi.org/10.1186/s12859-015-0677-y

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, *67*(1), 1–48.

https://doi.org/10.18637/jss.v067.i01
Bhiman, J. N., Anthony, C., Doria-Rose, N. A., Karimanzira, O., Schramm, C. A., Khoza, T., … Moore, P. L. (2015). Viral variants that initiate and drive maturation of V1V2-directed HIV-1 broadly neutralizing antibodies. *Nature Medicine*, *21*(11), 1332–1336. https://doi.org/10.1038/nm.3963
Binley, J. M., Lybarger, E. A., Crooks, E. T., Seaman, M. S., Gray, E., Davis, K. L., … Mascola, J. R. (2008). Profiling the Specificity of Neutralizing Antibodies in a Large Panel of Plasmas from Patients Chronically Infected with Human Immunodeficiency Virus Type 1 Subtypes B and C. *Journal of Virology*, *82*(23), 11651–11668. https://doi.org/10.1128/JVI.01762-08
Bonsignori, M., Kreider, E. F., Fera, D., Meyerhoff, R. R., Bradley, T., Wiehe, K., … Haynes, B. F. (2017). Staged induction of HIV-1 glycan–dependent broadly neutralizing antibodies. *Science Translational Medicine*, *9*(381), eaai7514. https://doi.org/10.1126/scitranslmed.aai7514
Bonsignori, M., Liao, H.-X., Gao, F., Williams, W. B., Alam, S. M., Montefiori, D. C., & Haynes, B. F. (2017). Antibody-virus co-evolution in HIV infection: paths for HIV vaccine development. *Immunological Reviews*, *275*(1), 145–160. https://doi.org/10.1111/imr.12509
Bonsignori, M., Zhou, T., Sheng, Z., Chen, L., Gao, F., Joyce, M. G., … Haynes, B. F. (2016). Maturation Pathway from Germline to Broad HIV-1 Neutralizer of a CD4-Mimic Antibody. *Cell*, *165*(2), 449–463. https://doi.org/10.1016/j.cell.2016.02.022
Brandman, R., Brandman, Y., & Pande, V. S. (2012). Sequence Coevolution between RNA and Protein Characterized by Mutual Information between Residue Triplets. *PLoS ONE*, *7*(1), e30022. https://doi.org/10.1371/journal.pone.0030022
Brennan, P. L. R., & Prum, R. O. (2015). Mechanisms and Evidence of Genital Coevolution: The Roles of Natural Selection, Mate Choice, and Sexual Conflict. *Cold Spring Harbor Perspectives in Biology*, *7*(7), a017749. https://doi.org/10.1101/cshperspect.a017749
Dawkins, R., & Krebs, J. R. (1979). Arms races between and within species. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, *205*(1161), 489–511. https://doi.org/10.1098/RSPB.1979.0081
Doria-Rose, N. A., Klein, R. M., Daniels, M. G., O'Dell, S., Nason, M., Lapedes, A., … Connors, M. (2010). Breadth of Human Immunodeficiency Virus-Specific Neutralizing Activity in Sera: Clustering Analysis and Association with Clinical Variables. *Journal of Virology*, *84*(3), 1631–1636. https://doi.org/10.1128/JVI.01482-09
Doria-Rose, Nicole A., Schramm, C. A., Gorman, J., Moore, P. L., Bhiman, J. N., DeKosky, B. J., … Mascola, J. R. (2014). Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature*, *509*(7498), 55–62. https://doi.org/10.1038/nature13036
Ernst, M. D. (2004). Permutation Methods: A Basis for Exact Inference. *Statistical Science*, *19*(4), 676–685.
Eroshkin, A. M., LeBlanc, A., Weekes, D., Post, K., Li, Z., Rajput, A., … Godzik, A. (2014). bNAber: database of broadly neutralizing HIV antibodies. *Nucleic*

Acids Research, *42*(D1), D1133–D1139. https://doi.org/10.1093/nar/gkt1083
Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., … Lemey, P. (2014). HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. *Science (New York, N.Y.)*, *346*(6205), 56–61. https://doi.org/10.1126/science.1256739
Feder, A. F., Kryazhimskiy, S., & Plotkin, J. B. (2014). Identifying Signatures of Selection in Genetic Time Series. *Genetics*, *196*(2), 509–522. https://doi.org/10.1534/genetics.113.158220
Fischer, W., Ganusov, V. V., Giorgi, E. E., Hraber, P. T., Keele, B. F., Leitner, T., … Korber, B. T. (2010). Transmission of Single HIV-1 Genomes and Dynamics of Early Immune Escape Revealed by Ultra-Deep Sequencing. *PLoS ONE*, *5*(8), e12303. https://doi.org/10.1371/journal.pone.0012303
Foley, B., Leitner, T., Apetrei, C., Hahn, B., Mizrachi, I., Mullins, J., … Korber, B. (Eds.). (2017). *HIV Sequence Compendium 2017*. Los Alamos: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, NM. Retrieved from http://www.hiv.lanl.gov/
Freund, N. T., Wang, H., Scharf, L., Nogueira, L., Horwitz, J. A., Bar-On, Y., … Nussenzweig, M. C. (2017). Coexistence of potent HIV-1 broadly neutralizing antibodies and antibody-sensitive viruses in a viremic controller. *Science Translational Medicine*, *9*(373), eaal2144. https://doi.org/10.1126/scitranslmed.aal2144
Gao, F., Bonsignori, M., Liao, H.-X., Kumar, A., Xia, S.-M., Lu, X., … Haynes, B. F. (2014). Cooperation of B Cell Lineages in Induction of HIV-1-Broadly Neutralizing Antibodies. *Cell*, *158*(3), 481–491. https://doi.org/10.1016/j.cell.2014.06.022
Giudicelli, V., Duroux, P., Ginestoux, C., Folch, G., Jabado-Michaloud, J., Chaume, D., & Lefranc, M.-P. (2006). IMGT/LIGM-DB, the IMGT(R) comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Research*, *34*(90001), D781–D784. https://doi.org/10.1093/nar/gkj088
Gloor, G. B., Martin, L. C., Wahl, L. M., & Dunn, S. D. (2005). Mutual Information in Protein Multiple Sequence Alignments Reveals Two Classes of Coevolving Positions†. https://doi.org/10.1021/BI050293E
Goo, L., Chohan, V., Nduati, R., & Overbaugh, J. (2014). Early development of broadly neutralizing antibodies in HIV-1-infected infants. *Nature Medicine*, *20*(6), 655–658. https://doi.org/10.1038/nm.3565
Gottlieb, M., Schanker, H., Fan, P., Saxon, A., Weisman, J., & Pozalski, I. (1981). Pneumocystis pneumonia--Los Angeles. *MMWR. Morbidity and Mortality Weekly Report*, *30*(21), 250–252. Retrieved from https://www.popline.org/node/421156
Gray, E. S., Madiga, M. C., Hermanus, T., Moore, P. L., Wibmer, C. K., Tumba, N. L., … CAPRISA002 Study Team, and the C. S. (2011). The neutralization breadth of HIV-1 develops incrementally over four years and is associated with CD4+ T cell decline and high viral load during acute infection. *Journal of Virology*, *85*(10), 4828–4840. https://doi.org/10.1128/JVI.00198-11

Hatada, M., Yoshimura, K., Harada, S., Kawanami, Y., Shibata, J., & Matsushita, S. (2010). Human immunodeficiency virus type 1 evasion of a neutralizing anti-V3 antibody involves acquisition of a potential glycosylation site in V2. *Journal of General Virology*, *91*(5), 1335–1345. https://doi.org/10.1099/vir.0.017426-0

Henn, M. R., Boutwell, C. L., Charlebois, P., Lennon, N. J., Power, K. A., Macalalad, A. R., … Allen, T. M. (2012). Whole Genome Deep Sequencing of HIV-1 Reveals the Impact of Early Minor Variants Upon Immune Recognition During Acute Infection. *PLoS Pathogens*, *8*(3), e1002529. https://doi.org/10.1371/journal.ppat.1002529

Hoehn, K. B., Gall, A., Bashford-Rogers, R., Fidler, S. J., Kaye, S., Weber, J. N., … Pybus, O. G. (2015). Dynamics of immunoglobulin sequence diversity in HIV-1 infected individuals. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *370*(1676). https://doi.org/10.1098/rstb.2014.0241

Horwitz, J. A., Bar-On, Y., Lu, C.-L., Fera, D., Lockhart, A. A. K., Lorenzi, J. C. C., … Nussenzweig, M. C. (2017). Non-neutralizing Antibodies Alter the Course of HIV-1 Infection In Vivo. *Cell*, *170*(4), 637-648.e10. https://doi.org/10.1016/j.cell.2017.06.048

Hraber, P., Seaman, M. S., Bailer, R. T., Mascola, J. R., Montefiori, D. C., & Korber, B. T. (2014). Prevalence of broadly neutralizing antibody responses during chronic HIV-1 infection. *AIDS*, *28*(2), 163–169. https://doi.org/10.1097/QAD.0000000000000106

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, *33*(6), 1635–1638. https://doi.org/10.1093/molbev/msw046

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12136088

Landais, E., Huang, X., Havenar-Daughton, C., Murrell, B., Price, M. A., Wickramasinghe, L., … Poignard, P. (2016). Broadly Neutralizing Antibody Responses in a Large Longitudinal Sub-Saharan HIV Primary Infection Cohort. *PLOS Pathogens*, *12*(1), e1005369. https://doi.org/10.1371/journal.ppat.1005369

Landais, E., Murrell, B., Briney, B., Murrell, S., Rantalainen, K., Berndsen, Z. T., … Poignard, P. (2017). HIV Envelope Glycoform Heterogeneity and Localized Diversity Govern the Initiation and Maturation of a V2 Apex Broadly Neutralizing Antibody Lineage. *Immunity*, *47*(5), 990-1003.e9. https://doi.org/10.1016/j.immuni.2017.11.002

Liao, H.-X., Lynch, R., Zhou, T., Gao, F., Alam, S. M., Boyd, S. D., … Haynes, B. F. (2013). Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature*, *496*(7446), 469–476. https://doi.org/10.1038/nature12053

MacLeod, D. T., Choi, N. M., Briney, B., Garces, F., Ver, L. S., Landais, E., … Bekker, L.-G. (2016). Early Antibody Lineage Diversification and Independent Limb Maturation Lead to Broad HIV-1 Neutralization Targeting

the Env High-Mannose Patch. *Immunity*, *44*(5), 1215–1226. https://doi.org/10.1016/j.immuni.2016.04.016

Maldarelli, F., Kearney, M., Palmer, S., Stephens, R., Mican, J., Polis, M. A., … Coffin, J. M. (2013). HIV populations are large and accumulate high genetic diversity in a nonlinear fashion. *Journal of Virology*, *87*(18), 10313–10323. https://doi.org/10.1128/JVI.01225-12

Marino Buslje, C., Teppa, E., Di Doménico, T., Delfino, J. M., & Nielsen, M. (2010). Networks of High Mutual Information Define the Structural Proximity of Catalytic Sites: Implications for Catalytic Residue Identification. *PLoS Computational Biology*, *6*(11), e1000978. https://doi.org/10.1371/journal.pcbi.1000978

Mayr, L. M., Decoville, T., Schmidt, S., Laumond, G., Klingler, J., Ducloy, C., … Moog, C. (2017). Non-neutralizing Antibodies Targeting the V1V2 Domain of HIV Exhibit Strong Antibody-Dependent Cell-mediated Cytotoxic Activity. *Scientific Reports*, *7*(1), 12655. https://doi.org/10.1038/s41598-017-12883-6

Mayr, L. M., Su, B., & Moog, C. (2017). Non-Neutralizing Antibodies Directed against HIV and Their Functions. *Frontiers in Immunology*, *8*, 1590. https://doi.org/10.3389/fimmu.2017.01590

McCoy, L. E., & Burton, D. R. (2017). Identification and specificity of broadly neutralizing antibodies against HIV. *Immunological Reviews*, *275*(1), 11–20. https://doi.org/10.1111/imr.12484

Murugan, R., Buchauer, L., Triller, G., Kreschel, C., Costa, G., Pidelaserra Martí, G., … Wardemann, H. (2018). Clonal selection drives protective memory B cell responses in controlled human malaria infection. *Science Immunology*, *3*(20), eaap8029. https://doi.org/10.1126/sciimmunol.aap8029

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Nei, M., & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, *3*(5), 418–426. https://doi.org/10.1093/oxfordjournals.molbev.a040410

Nourmohammad, A., Otwinowski, J., Luksza, M., Mora, T., & Walczak, A. M. (2018). Clonal competition in B-cell repertoires during chronic HIV-1 infection. *BioRxiv*, 271130. https://doi.org/10.1101/271130

Nourmohammad, A., Otwinowski, J., & Plotkin, J. B. (2016). Host-Pathogen Coevolution and the Emergence of Broadly Neutralizing Antibodies in Chronic Infections. *PLOS Genetics*, *12*(7), e1006171. https://doi.org/10.1371/journal.pgen.1006171

Orton, R. (2014). Calculating dNdS for NGS datasets | Bioinformatics I/O. Retrieved March 4, 2019, from http://bioinformatics.cvr.ac.uk/blog/calculating-dnds-for-ngs-datasets/

Piantadosi, A., Panteleeff, D., Blish, C. A., Baeten, J. M., Jaoko, W., McClelland, R. S., & Overbaugh, J. (2009). Breadth of Neutralizing Antibody Response to Human Immunodeficiency Virus Type 1 Is Affected by Factors Early in

Infection but Does Not Influence Disease Progression. *Journal of Virology*, *83*(19), 10269–10274. https://doi.org/10.1128/JVI.01149-09
Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, *5*(3), e9490. https://doi.org/10.1371/journal.pone.0009490
Rantalainen, K., Berndsen, Z. T., Murrell, S., Cao, L., Omorodion, O., Torres, J. L., … Ward, A. B. (2018). Co-evolution of HIV Envelope and Apex-Targeting Neutralizing Antibody Lineage Provides Benchmarks for Vaccine Design. *Cell Reports*, *23*(11), 3249–3261. https://doi.org/10.1016/j.celrep.2018.05.046
Redd, A. D., Quinn, T. C., & Tobian, A. A. (2013). Frequency and implications of HIV superinfection. *The Lancet Infectious Diseases*, *13*(7), 622–628. https://doi.org/10.1016/S1473-3099(13)70066-5
Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, *16*(6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2
Ringe, R., Das, L., Choudhary, I., Sharma, D., Paranjape, R., Chauhan, V. S., & Bhattacharya, J. (2012). Unique C2V3 sequence in HIV-1 envelope obtained from broadly neutralizing plasma of a slow progressing patient conferred enhanced virus neutralization. *PloS One*, *7*(10), e46713. https://doi.org/10.1371/journal.pone.0046713
Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584
Sather, D. N., Armann, J., Ching, L. K., Mavrantoni, A., Sellhorn, G., Caldwell, Z., … Stamatatos, L. (2009). Factors Associated with the Development of Cross-Reactive Neutralizing Antibodies during Human Immunodeficiency Virus Type 1 Infection. *Journal of Virology*, *83*(2), 757–769. https://doi.org/10.1128/JVI.02036-08
Setliff, I., McDonnell, W. J., Raju, N., Bombardi, R. G., Murji, A. A., Scheepers, C., … Georgiev, I. S. (2018). Multi-Donor Longitudinal Antibody Repertoire Sequencing Reveals the Existence of Public Antibody Clonotypes in HIV-1 Infection. *Cell Host & Microbe*, *23*(6), 845-854.e6. https://doi.org/10.1016/j.chom.2018.05.001
Shankarappa, R., Margolick, J. B., Gange, S. J., Rodrigo, A. G., Upchurch, D., Farzadegan, H., … Mullins, J. I. (1999). Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *Journal of Virology*, *73*(12), 10489–10502. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10559367
Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303
Sheward, D. J., Marais, J., Bekker, V., Murrell, B., Eren, K., Bhiman, J. N., … Williamson, C. (2018). HIV Superinfection Drives De Novo Antibody Responses and Not Neutralization Breadth. *Cell Host & Microbe*, *24*(4), 593-

599.e3. https://doi.org/10.1016/J.CHOM.2018.09.001
Simonetti, F. L., Teppa, E., Chernomoretz, A., Nielsen, M., & Marino Buslje, C. (2013). MISTIC: Mutual information server to infer coevolution. *Nucleic Acids Research*, *41*(Web Server issue), W8-14. https://doi.org/10.1093/nar/gkt427
Simonich, C. A., Williams, K. L., Verkerke, H. P., Williams, J. A., Nduati, R., Lee, K. K., & Overbaugh, J. (2016). HIV-1 Neutralizing Antibodies with Limited Hypermutation from an Infant. *Cell*, *166*(1), 77–87. https://doi.org/10.1016/j.cell.2016.05.055
Stein, J., Storcksdieck genannt Bonsmann, M., & Streeck, H. (2016). Barriers to HIV Cure. *HLA*, *88*(4), 155–163. https://doi.org/10.1111/tan.12867
Strauli, N. B., & Hernandez, R. D. (2016). Statistical inference of a convergent antibody repertoire response to influenza vaccine. *Genome Medicine*, *8*(1), 60. https://doi.org/10.1186/s13073-016-0314-z
Sturdevant, C. B., Joseph, S. B., Schnell, G., Price, R. W., Swanstrom, R., & Spudich, S. (2015). Compartmentalized Replication of R5 T Cell-Tropic HIV-1 in the Central Nervous System Early in the Course of Infection. *PLOS Pathogens*, *11*(3), e1004720. https://doi.org/10.1371/journal.ppat.1004720
Trovato, M., D'Apice, L., Prisco, A., & De Berardinis, P. (2018). HIV Vaccination: A Roadmap among Advancements and Concerns. *International Journal of Molecular Sciences*, *19*(4), 1241. https://doi.org/10.3390/ijms19041241
Vander Heiden, J. A., Yaari, G., Uduman, M., Stern, J. N. H., O'connor, K. C., Hafler, D. A., … Kleinstein, S. H. (2014). PRESTO: A toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics*, *30*(13), 1930–1932. https://doi.org/10.1093/bioinformatics/btu138
Wang, C., Sanders, C. M., Yang, Q., Schroeder, H. W., Wang, E., Babrzadeh, F., … Han, J. (2010). High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *Proceedings of the National Academy of Sciences*, *107*(4), 1518–1523. https://doi.org/10.1073/pnas.0913939107
Weinstein, M., Yang, O. O., & Cohen, A. C. (2017). Were we prepared for PrEP? Five years of implementation. *AIDS*, *31*(16), 2303–2305. https://doi.org/10.1097/QAD.0000000000001626
Williams, K. L., Wang, B., Arenz, D., Williams, J. A., Dingens, A. S., Cortez, V., … Overbaugh, J. (2018). Superinfection Drives HIV Neutralizing Antibody Responses from Several B Cell Lineages that Contribute to a Polyclonal Repertoire. *Cell Reports*, *23*(3), 682–691. https://doi.org/10.1016/J.CELREP.2018.03.082
Wu, X., Zhou, T., Zhu, J., Zhang, B., Georgiev, I., Wang, C., … Mascola, J. R. (2011). Focused Evolution of HIV-1 Neutralizing Antibodies Revealed by Structures and Deep Sequencing. *Science*, *333*(6049), 1593–1602. https://doi.org/10.1126/science.1207532
Wu, Xueling, Zhang, Z., Schramm, C. A., Joyce, M. G., Do Kwon, Y., Zhou, T., … Young, A. (2015). Maturation and Diversity of the VRC01-Antibody Lineage over 15 Years of Chronic HIV-1 Infection. *Cell*, *161*(3), 470–485. https://doi.org/10.1016/j.cell.2015.03.004

Yaari, G., Uduman, M., & Kleinstein, S. H. (2012). Quantifying selection in high-throughput Immunoglobulin sequencing data sets. *Nucleic Acids Research*, *40*(17), e134–e134. https://doi.org/10.1093/nar/gks457

Ye, J., Ma, N., Madden, T. L., & Ostell, J. M. (2013). IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Research*, *41*(W1), W34–W40. https://doi.org/10.1093/nar/gkt382

Yusim, K., Korber, B. T. M., Brander, C., Barouch, D., Boer, R. de, Haynes, B. F., … Watkins, D. I. (Eds.). (2016). *HIV Molecular Immunology 2016*. Los Alamos: Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico. Retrieved from https://www.hiv.lanl.gov/content/immunology/contents2016.html

Zanini, F., Brodin, J., Thebo, L., Lanz, C., Bratt, G., Albert, J., & Neher, R. A. (2015). Population genomics of intrapatient HIV-1 evolution. *ELife*, *4*. https://doi.org/10.7554/eLife.11282

Zhou, J. O., Ton, T., Morriss, J. W., Nguyen, D., & Fera, D. (2018). Structural Insights from HIV-Antibody Coevolution and Related Immunization Studies. *AIDS Research and Human Retroviruses*, *34*(9), 760–768. https://doi.org/10.1089/AID.2018.0097

Tables

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

Table 1. Participant demographics.
MSM – men who have sex with men.
* at estimated time of infection

Figures

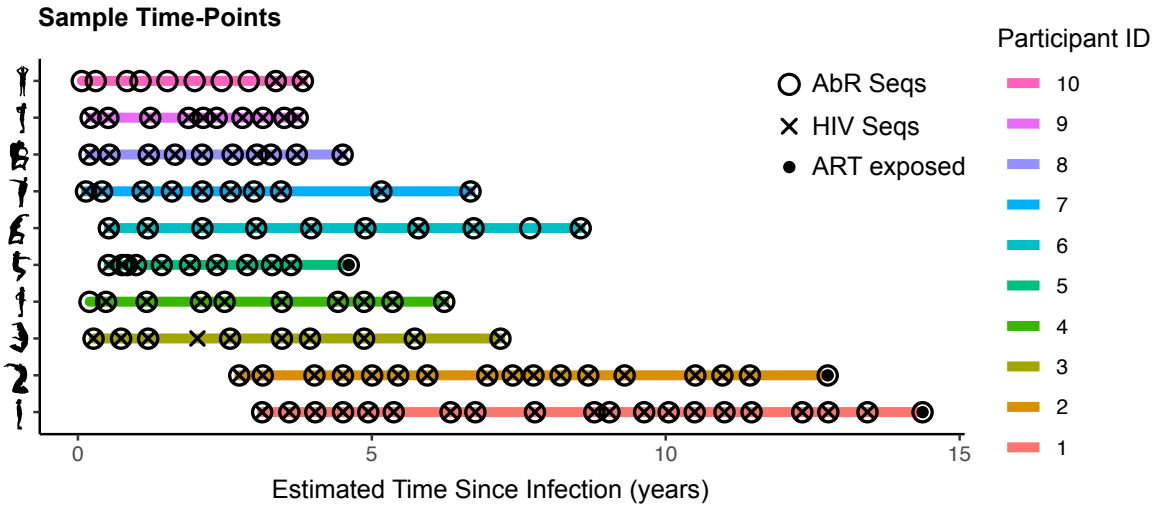

Figure 1. Sampled time-points.
Depicts the time since infection for each sample, in each of the participants. Open circles indicate that the AbR was successfully sequenced, crosses indicate that HIV was successfully sequenced, and solid circles indicate that the participant was on ART at this time.

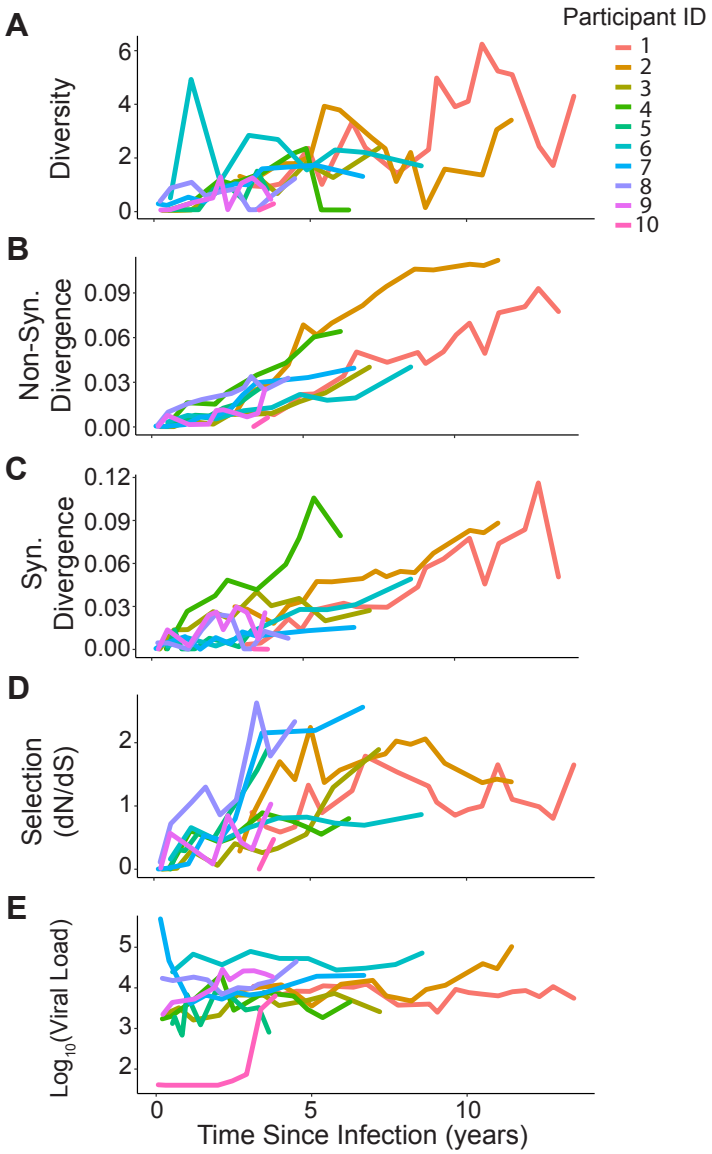

Figure 2. HIV summary statistic trajectories for each participant.
Each line shows the trajectory of a given summary statistic, in a given participant over time. Each participant has a unique color. (A) Diversity. (B) Nonsynonymous divergence. (C) Synonymous divergence. (D) Selection. (E) Viral load.

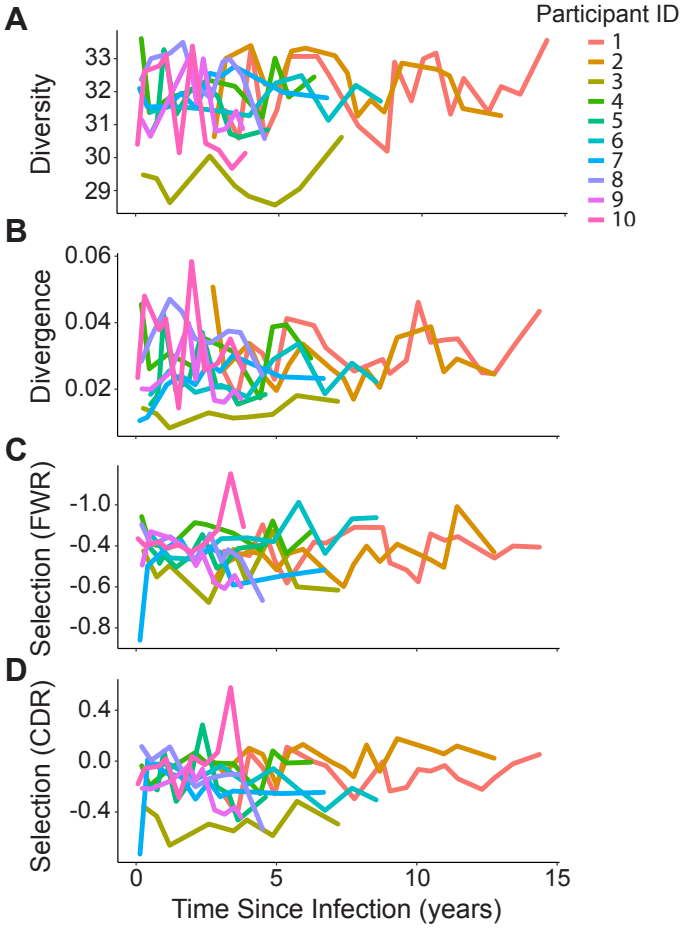

Figure 3. AbR summary statistic trajectories for each participant.
Each line shows the trajectory of a given summary statistic, in a given participant over time. Each participant has a unique color. (A) Diversity. (B) Divergence. (C) Selection in the framework region (FWR). (D) Selection in the complementarity determining region (CDR).

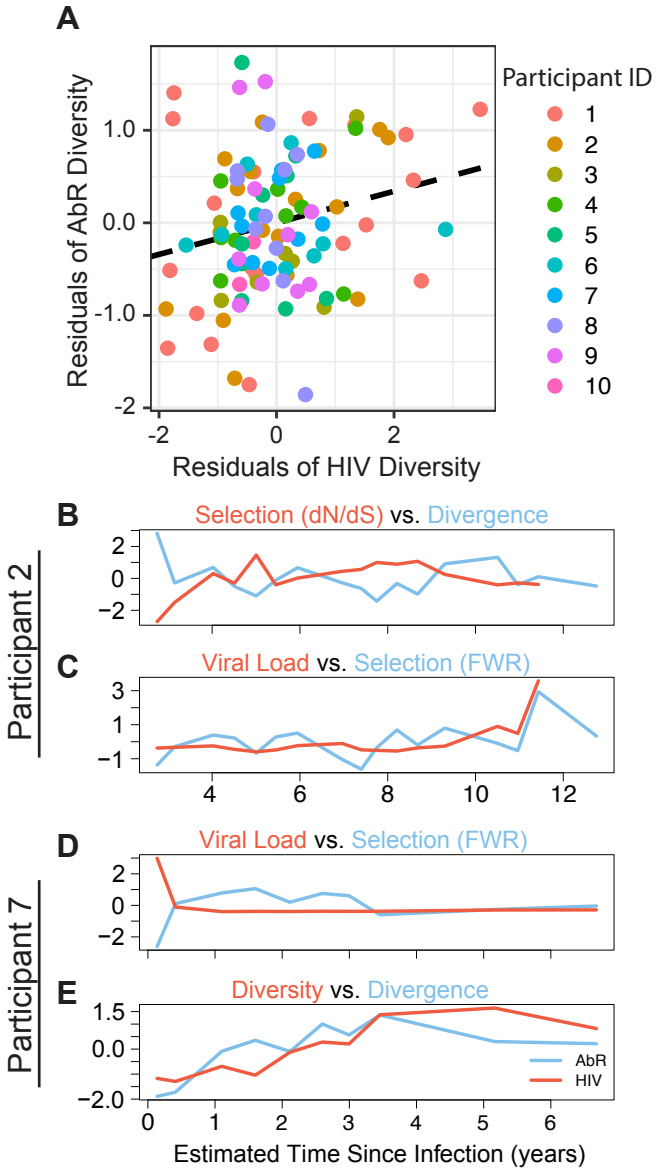

Figure 4. Whole population level associations between AbR and HIV summary statistics.
(A) Scatter plot showing positive correlation between HIV diversity (x-axis) and AbR diversity (y-axis). Each point represents a sample with diversity values from both the AbR and HIV sequence data. The axes show diversity values after participant specific effects have been regressed out. Dashed line shows positive relationship between AbR and HIV diversity, as given by our linear regression (see Methods). (B-E) Shows associations between summary statistic trajectories in AbR (blue) and HIV (red) at the individual participant level. (B) HIV selection with AbR Divergence, in participant 2. (C) Viral load with AbR FWR-selection, in participant 2. (D) Viral load with AbR FWR-selection, in participant 7. (E) HIV diversity with AbR divergence, in participant 7.

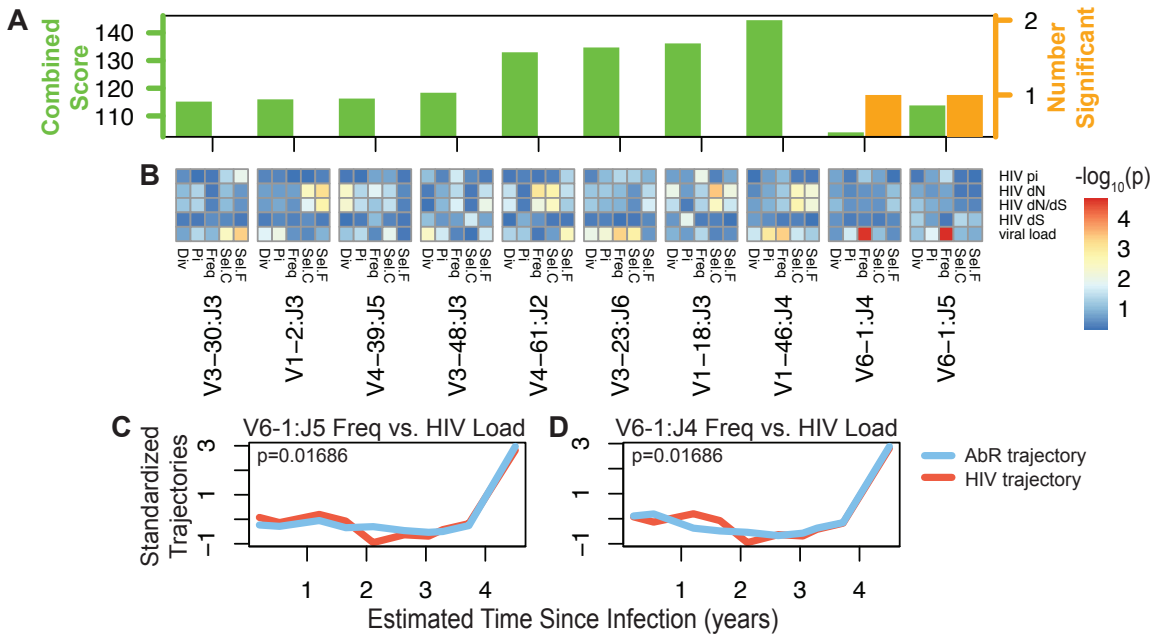

Figure 5. Results of permutation-based test to identify HIV-associated AbR partitions in participant 8.

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

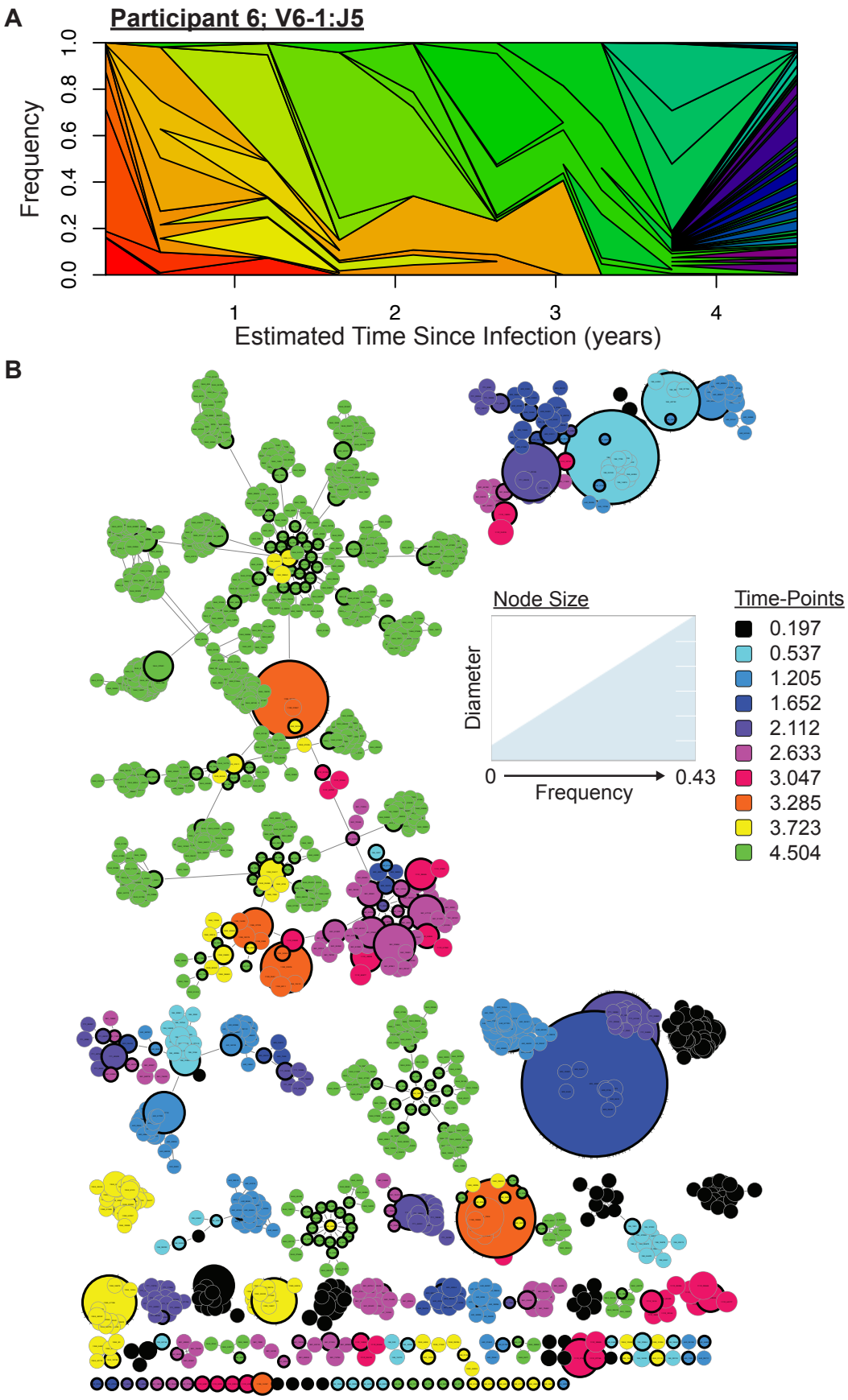
A
Participant 6; V6-1:J5
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
Estimated Time Since Infection (years)
B
Node Size
Diameter
0
0.43
Frequency
Time-Points
0.197
0.537
1.205
1.652
2.112
2.633
3.047
3.285
3.723
4.504

Figure 6. AbR lineage structure.
Illustrates the lineages in AbR partition IGHV6-1:IGHJ5, participant 8. (A) Muller plot depicting the relative frequency of each lineage within the partition. Each color represents a lineage. If a new lineage arises within the bounds of a preexisting lineage, then the new lineage is a daughter of the preexisting, parent lineage. Lineages that began earlier in the time-course have colors closer to the red side of the spectrum, while lineages that began later in the time-course have colors closer to violet. Only lineages that exceeded 0.0001 frequency are included in the plot. (B) Network plot. Each node represents a unique sequence. Nodes with thicker black borders are the 'representative sequence' of their sequence-cluster. Nodes of the same color (i.e. same time-point) are linked with an edge if they were in the same cluster. Nodes that have different colors are linked with an edge if they were assigned to the same lineage. This inter-time-point linking only occurs between 'representative sequences'. Taken together, each isolated grouping of nodes shows a family of related lineages.

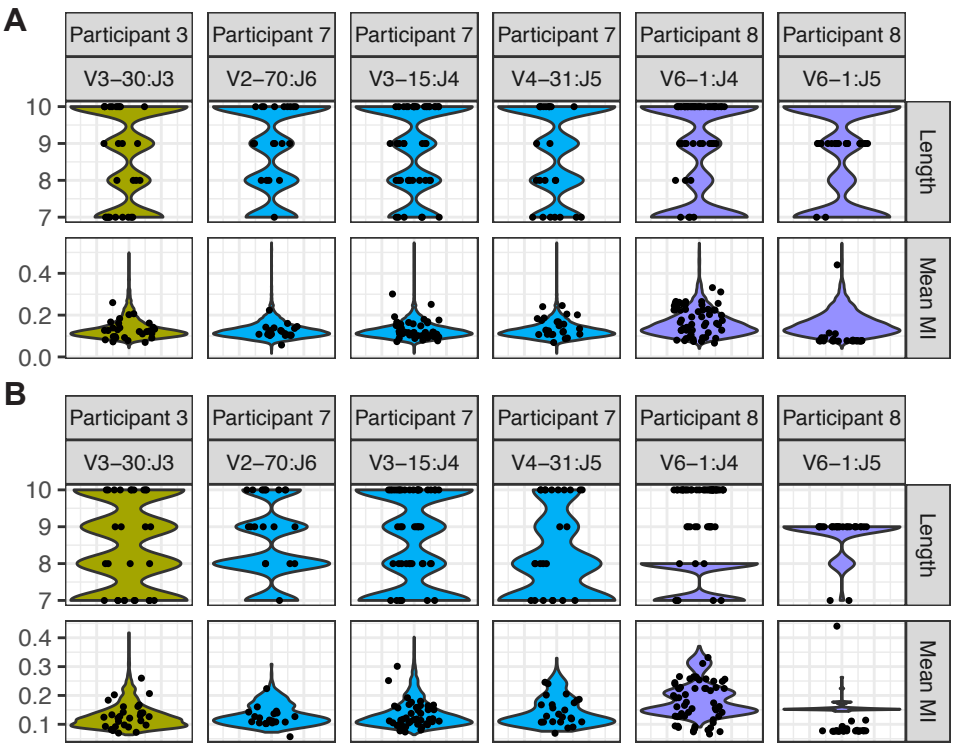

Figure 7. Coevolution test for individual lineages.
Each black dot corresponds to a value for a specific lineage within an HIV-associated AbR partition, and are jittered across the x-axis. The identifying information for the specific HIV-associated partition is given by the column labels for the plots (ex. The values for the lineages within the IGHV3-30:IGHJ3 AbR partition of participant 3 are found in the left-most column). The colored violin plots behind the points represent a given null-comparison for the points, where color corresponds to participant. The top row of plots shows the values and null distributions for the length of lineages (i.e. how many time-points they were present). The bottom row shows the values and null distributions for the mean MI of lineages. (A) The null distributions are made up of the values from the lineages in AbR partitions that were not HIV-associated. (B) The null distributions are made up of the values from the null simulation for each given AbR partition.

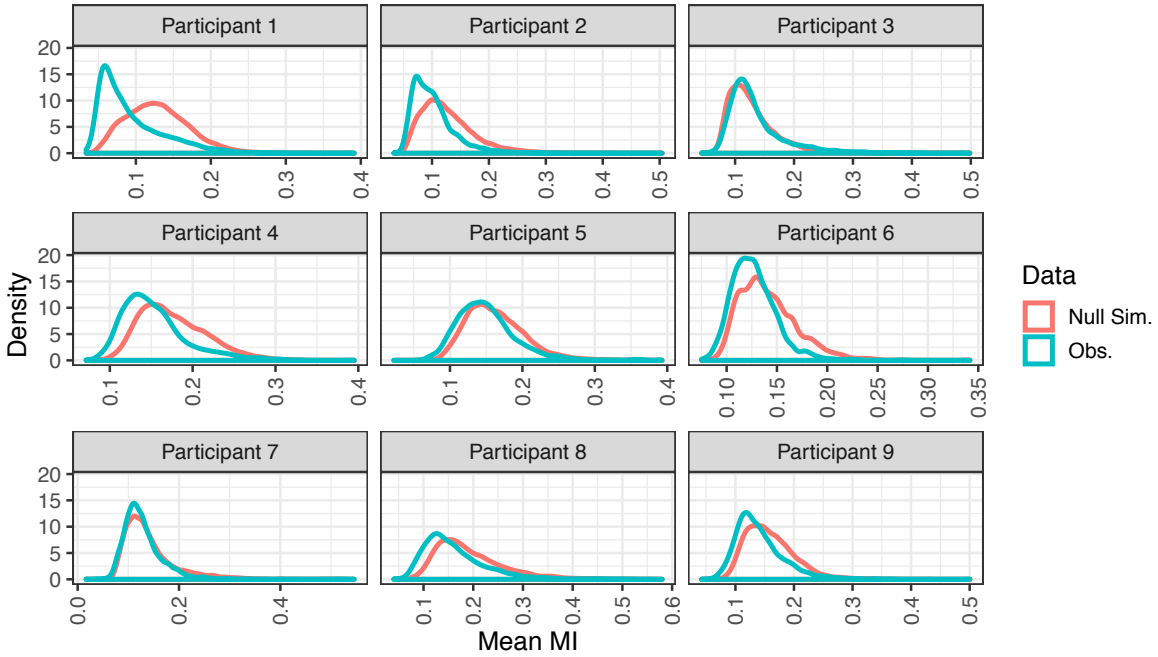

Figure 8. Global tests for coevolution.
Each line gives the distribution of mean MI values for a given participant. Turquoise colored lines correspond to observed lineages (i.e. lineages inferred from the data), and salmon colored lines correspond to lineages from the null simulations. Participant 10 is omitted because of limited HIV sequence data.