Clonal competition in B-cell repertoires during chronic HIV-1 infection

Armita Nourmohammad[1,2,*], Jakub Otwinowski[3], Marta Łuksza[4],
Thierry Mora[5,6†], Aleksandra M. Walczak[6,7†]

[1] *Max Planck Institute for Dynamics and Self-organization, Am Faßberg 17, 37077 Göttingen, Germany*
[2] *Department of Physics, University of Washington, 3910 15th Avenue Northeast, Seattle, WA, USA*
[3] *Department of Biology, University of Pennsylvania, Lynch Laboratory, Philadelphia, PA, USA*
[4] *The Simons Center for Systems Biology, Institute for Advanced Study, Princeton, NJ, USA*
[5] *Laboratoire de Physique Statistique, CNRS, UPMC (Sorbonne University), Paris-Diderot University*
[6] *École Normale Supérieure (PSL), 24, rue Lhomond, 75005 Paris, France*
[7] *Laboratoire de Physique Théorique, CNRS, UPMC (Sorbonne University)*

During chronic infection, HIV-1 engages in a rapid coevolutionary arms race with the host's adaptive immune system. While it is clear that HIV exerts strong selection on the adaptive immune system, the modes of immune response are still unknown. Traditional population genetics methods fail to distinguish a chronic immune response from natural repertoire evolution in healthy individuals. Here, we infer the evolutionary modes of B-cell repertoire response and identify a complex dynamics where, instead of one winning clone, there is a constant production of new better mutants that compete with each other. A substantial fraction of mutations in pathogen-engaging CDRs of B-cell receptors are beneficial, in contrast to the many deleterious changes in structurally relevant framework regions. The picture is of a dynamic repertoire, where better clones may be outcompeted by new mutants before they fix, challenging current vaccine design and therapy ideas.

* correspondence should be addressed to: armita@ds.mpg.de
† equal contribution

The HIV-1 virus evolves and proliferates quickly within the human body [1–3], often recombining its genetic material among different viral genomes and rapidly mutating. These factors make it very hard for the host immune system to control an infection, leading to long-term chronic infection. While it is clear that the virus exerts strong selective pressure on the host immune system, the adaptive immune response during chronic infections remains unknown.

The immune system has a diverse set of B and T-cells with specialized surface receptors that recognize foreign antigens, such as virus epitopes, and protect the organism. We focus on the chronic phase of HIV infection, where the immune response is dominated by antibody-mediated mechanisms, following the strong response of the cytotoxic T-lymphocytes (i.e., CD8+ killers T-cells), around 50 days after infection [4]. During the chronic phase, the symptoms are minor and the viral load is relatively stable but its genetic composition undergoes rapid turnover. After an infection, B-cells undergo a rapid somatic hypermutation in lymph node germinal centers, with a rate that is approximately $4-5$ orders of magnitude larger than an average germline mutation rate in humans [5]. Mutated B-cells compete for survival and proliferation signals from helper T-cells, based on the B-cell receptor's binding to antigens. This process of *affinity maturation* is Darwinian evolution within the host and can increase binding affinities of B-cell receptors up to 10-100 fold [6]. It generates memory and plasma B-cells with distinct receptors, forming lineages that trace the evolutionary selection pressures inflicted by the virus [7] (see schematic in Fig. 1A). A B-cell repertoire consists of many such lineages forming a forest of co-existing genealogies.

Immune repertoire high-throughput sequencing has been instrumental in quantifying the diversity of B-cell repertoires [8, 9]. Statistical methods have been developed to characterize the processes involved in the generation of diversity in repertoires and to infer the underlying heterogenous hypermutation preferences in B-cell receptors (BCRs) [9–11]. Deviation of the observed mutations in BCRs from the expected hypermutation patterns are used to infer selection effects of mutations from repertoire snapshots in order to identify functional changes that contribute to the response against pathogens [10, 12].

Recently, longitudinal data, with repertoires sampled over multiple time points from the same individuals, has brought insight into the dynamics of affinity maturation in response to antigens [13–16]. The dynamics of affinity maturation and selection in response to HIV have also been characterized for chosen monoclonal broadly neutralising antibody lineages [3, 17]. Yet, the effect of a chronic infection on the dynamics of the whole BCR repertoire remains unknown.

Here, we compare the structure and dynamics of BCR repertoires sampled over 2.5 years in HIV patients (data from ref. [15] collected through the SPARTAC study [18]) with the repertoire structure in healthy individuals (data from ref. [19]). We reconstruct genealogical trees for B-cell receptor lineages inferred from BCR repertoires in each individual (SI). B-cell lineages of HIV patients, a few examples of which are shown in Fig. 1B, can persist over months to years of infection, which is much longer than the lifetime of a germinal center (weeks), indicating the recruitment of memory cells for further affinity maturation in response to the

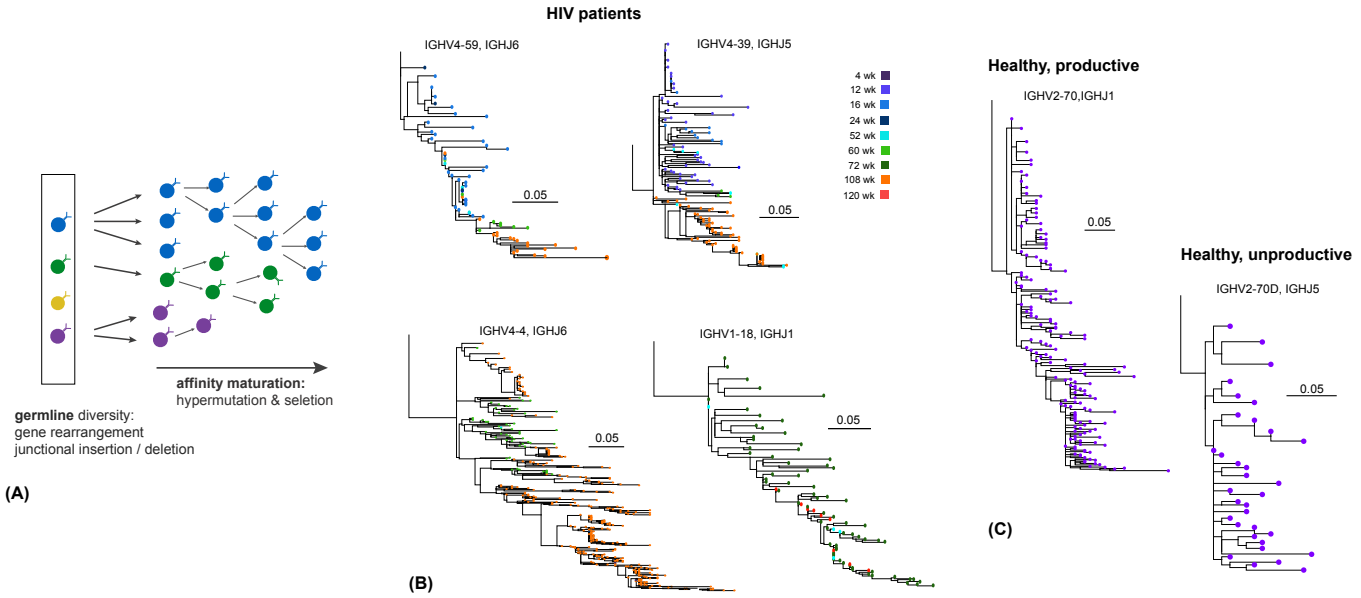

FIG. 1: **Affinity maturation forms B-cell lineages. (A)** Schematic of B-cell affinity maturation and lineage formation. The naive immune repertoire consists of a diverse set of B-cell receptors, generated by gene rearrangement (VDJ recombination) and junctional sequence insertion and deletion (distinct colored cells in the box). Affinity maturation with somatic hypermutations and selection for strong binding of BCRs to antigens forms lineages of BCRs stemmed from a germline progenitor, shown by three growing lineages in this figure. **(B)** Examples of B-cell lineages reconstructed from the heavy chain sequences of BCR repertoires in HIV patients (see SI). The distance between the nodes along the horizontal axis indicates their sequence hamming distance. The nodes are colored according to the time they were sampled from a patient over the period of ~ 2.5 yrs. **(C)** Examples of a productive (left) and unproductive (right) B-cell lineage reconstructed from the heavy chain repertoire of a healthy individual sampled at a single time point (SI).

evolving virus.

Reconstructed lineage trees show a skewed and asymmetric structure, consistent with rapid evolution under positive selection (see Fig. S1A) [20]. To quantify these asymmetries, we estimated two indices of tree imbalance and terminal branch length anomaly. In both HIV patients and healthy individuals, we observe a significant branching imbalance at the root of the BCR lineage trees, indicated by the U-shaped distribution of the sub-lineage weight ratios (see SI), in contrast to the flat prediction of neutral evolution, calculated from Kingman's coalescent (Fig. 2A). Moreover, we observe elongated terminal branches in BCR trees compared to their internal branches, with the strongest effect seen in trees from HIV patients, again in violation of neutrality (Fig. 2B, Fig. S1). These asymmetric features of BCR trees are clear signs of intra-lineage positive selection. However, they only reflect the history of lineage replication and give limited insight into the mechanisms and dynamics of selection. For instance, tree asymmetry is also observed in unproductive BCR lineages, which lack any immunological function but are carried along with the productive version of the recombined gene expressed on the other chromosome (Fig. 2A,B).

To characterize the selection effect of mutations in more detail, we evaluate the spectrum of mutation frequencies in a lineage, known as the site frequency spectrum (SFS). We evaluate the SFS separately for synonymous and nonsynonymous mutations in different regions of BCRs (Fig. 2C, Fig. S2). We see a signifiant upturn of SFS polarized on non-synonymous mutations in pathogen-engaging CDR3 regions, consistent with rapid adaptive evolution [20], and in contrast to monotonically decaying SFS in neutrality (SI). This signal of positive selection is strongest in HIV patients with an order of magnitude increase in the high end of the spectrum, suggesting that the BCR population rapidly adapts in HIV patients.

To understand the dynamics and fate of these adaptive mutations, we use the longitudinal nature of the data to analyse the temporal structure of the lineages. We estimate the likelihood that a new mutation appearing in a certain region of the BCR reaches frequency x at some later time within the lineage (Fig. 3A), and evaluate a measure of selection $g(x)$ as the ratio of this likelihood between non-synonymous and synonymous mutations [21] (SI). At frequency $x = 1$ (i.e., substitution), this ratio is equivalent to the McDonald-Kreitman test for selection [22]. Generalizing it to $x < 1$ makes it a more flexible measure applicable to the majority of mutations that only reach intermediate frequencies. A major reason why many beneficial mutations never fix in a lineage is clonal interference, whereby BCR mutants within and across lineages compete with each other [7]. To quantify the prevalence of clonal interference, we also

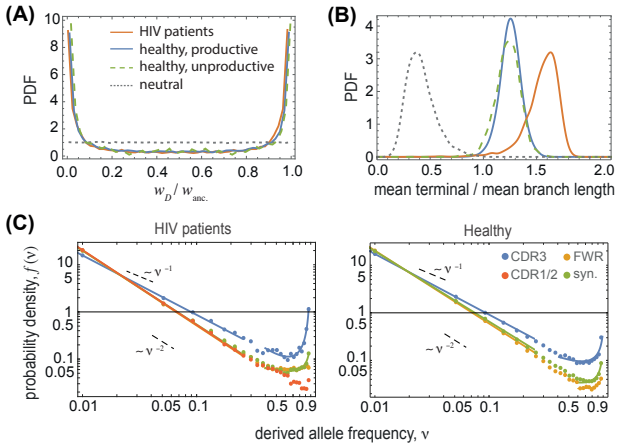

FIG. 2: **Statistics of BCR lineage genealogies indicate positive selection.** **(A)** The U-shaped distribution of sub-lineage weight ratios at the root of lineage trees (SI) $w_D/w_{\text{anc.}}$ and **(B)** the distribution of elongated mean terminal branch lengths (in units of divergence time) relative to the mean length of all branches in BCR lineages indicate positive selection in HIV patients and in healthy individuals (colors), in contrast to the neutral expectation (dotted lines); see Fig. S1 for comparison of tree statistics under different evolutionary scenarios. **(C)** The Site Frequency Spectrum (SFS) $f(\nu)$ is shown for mutations in different regions of BCRs (distinct colors) in HIV patients (left) and in healthy individuals (right); see Fig. S2 for SFS of unproductive BCR lineages. The upturn of SFS for non-synonymous mutations in CDR3 region is indicative of rapid evolution under positive selection.

evaluate the nonsynonymous-to-synonymous ratio $h(x)$ of the likelihood for a mutation to reach frequency x and later to go extinct (SI). In short, $g(x)$ identifies "surges" and $h(x)$ "bumps" in frequency trajectories of clones. These likelihood ratios have intuitive interpretations: $g(x) > 1$ indicates evolution under positive selection, with a fraction of at least $\alpha_{\text{benef.}} = 1 - 1/g$ strongly beneficial amino acid changes in a given region [23]. On the other hand, the likelihood ratio $g(x)$ smaller than 1 is indicative of negative selection, with a fraction of at least $\alpha_{\text{del.}} = 1 - g$ strongly deleterious changes (see SI for a derivation of these bounds). Likewise, $\kappa_{\text{benef.}} = 1 - 1/h$ or $\kappa_{\text{del.}} = 1 - h$ define a lower bound on the fraction of either beneficial or deleterious mutations that go extinct.

Fig. 3B shows the selection likelihood ratio $g(x)$ in an HIV patient (patient 4) for lineages belonging to a typical V-gene class IGHV2-70D (SI); see Fig. S3 for statistics in all individuals. In this gene family, we detect positive selection ($g > 1$) in the CDR3 region, with around a two fold larger fraction of non-synonymous compared to synonymous changes that reach frequencies $x > 0.6$, indicating at least $\alpha_{\text{benef.}} = 40\%$ of CDR3 mutations to be strongly beneficial. On the other hand, the likelihood ratio in FWR signals strong negative selection ($g < 1$), where non-synonymous changes reaching frequencies $x > 0.6$ are two times fewer than the synonymous changes, indicating at least $\alpha_{\text{del.}} = 35\%$ of these mutations to be strongly deleterious. Similarly, the interference likelihood ratio $h(x)$ for a V-gene class IGHV5-10-1 in an HIV patient with interrupted treatment (patient 5) indicates that about $\kappa_{\text{benef.}} = 47\%$ of CDR3 mutations in this gene family that go extinct due to clonal competition are strongly beneficial (Fig. 3B). In short, we observe a large fraction of adaptive mutations, and also a substantial amount of clonal interference which prevents some of the mutations from dominating within lineages.

To see how these observations generalize at the repertoire level, we quantify the region-specific fraction of beneficial and deleterious mutations within BCR lineages of distinct VJ-gene classes and also the fraction of selected mutations that are impeded by clonal interference (Fig. 3C and Table I). We infer a larger fraction of VJ-gene classes with positively selected amino acid changes in their CDR regions $\overline{\alpha}_{\text{benef.}} = 12\% - 30\%$ and negatively selected amino acid changes in FWRs $\overline{\alpha}_{\text{del.}} = 16\% - 20\%$. Moreover, the positively selected beneficial mutations in CDR3 and the pooled CDR1/CDR2 regions are strongly impacted by clonal inference, in contrast to mutations in FWR (Fig. 3C, Table I, Fig. S3). These observations confirm the pervasiveness of clonal interference in the regions of the BCR with the most important functional role.

In patients with interrupted ART, we infer a twice larger fraction of beneficial mutations to rise with strong clonal interference in pathogen-engaging CDR3 regions following the interruption of treatment, compared to the ART-naive patients with a stable chronic infection— such a shift is not present for mutations in CDR1, CDR2 and FWR (Fig. 3 and Table I). This pattern is consistent with the rate of HIV-1 evolution in patients with different states of therapy. Genome-wide analysis of HIV-1 has revealed that evolution of the virus within ART-naive patients slows down during chronic infections with limited clonal interference in viral populations [24]. The antibody response traces the evolution of the virus [1, 7] and forms a quasi-equilibrium balance. On the other hand, rapid expansion and evolution of HIV following the interruption of ART drives a strong immune response and affinity maturation in HIV-responsive B-cell lineages. Evolution of HIV-1 population during viral expansion introduces a time-dependent target for the adaptive immune system and opens room for many beneficial changes in the HIV-engaging CDR3 regions, as indicated in Fig. 3.

Somatic evolution during affinity maturation is complex: there is no one winner of the race for the best antibody. We show that rapid and strong affinity maturation upon sudden pathogenic challenges, and a quasi-stationary response during chronic infections are a feature of the B-cell response to infections. Somatic evolution of BCRs is similar to rapid evolution in asexual populations where many beneficial mutations rise to intermediate frequencies leading to complex clonal compe-

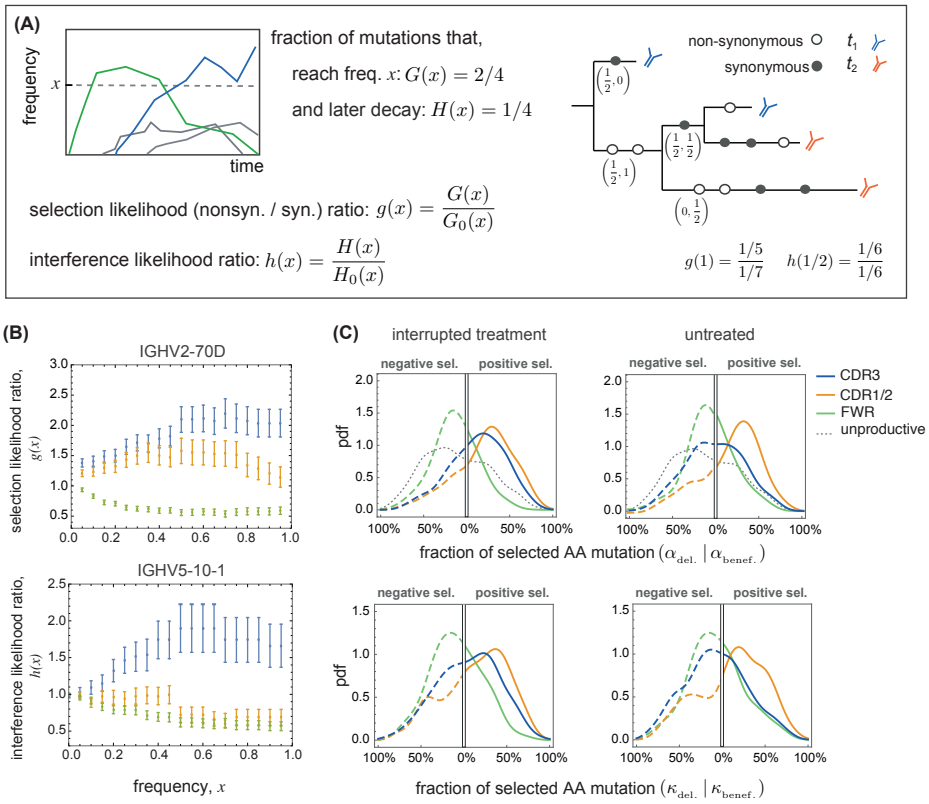

FIG. 3: **Inference of selection and clonal interference in BCR lineages. (A)** Schematic shows time-dependent frequencies of mutations that rise within a population (left). We denote the fraction of mutations that reach frequency x within a population (or lineage) by $G(x)$ (blue and green) and the subset that later goes extinct due to clonal interference by $H(x)$ (green). The likelihood ratios between non-synonymous and synonymous mutations ($g(x)$, $h(x)$) quantify the strength of selection in each case. A schematic B-cell genealogy (right) is shown for a lineage sampled at two time points (colors). Non-synonymous and synonymous mutations are shown by empty and filled circles and their frequencies (x_{t_1}, x_{t_2}), as observed in the sampled tree leaves, are indicated below each branch. **(B)** Selection likelihood ratio $g(x)$ in the V-gene class IGHV2-70D in patient 4 (top) and the interference likelihood ratio $h(x)$ for the V-gene class IGHV5-10-1 in patient 5 (bottom) are plotted against frequency x for mutations in different BCR regions (colors). The likelihood ratios indicate positive selection and strong clonal interference in the CDR3 region, negative selection on the FWR region and positive selection on mutations that rise to intermediate frequencies in the joint CDR1 / CDR2 regions. We do no observe interference in the FWR and the joint CDR1 / CDR2 region. **(C)** Each panel shows the probability density across distinct VJ-gene classes in HIV patients with interrupted treatment (left) and without treatment (right), for the fraction of beneficial and deleterious mutations (($\alpha_{\text{benef.}}$ / $\alpha_{\text{del.}}$) on right x-axis and left inverted x-axis) that reach frequency $x = 80\%$ (top), and similarly, for beneficial / deleterious mutation fractions ($\kappa_{\text{benef.}}$ / $\kappa_{\text{del.}}$) that reach frequency $x = 60\%$ within a lineage and later go extinct (bottom). The dotted grey line indicates the null distribution from unproductive lineages of healthy individuals (Fig. S4). The Color code for distinct BCR regions in all panels is consistent with the legend. See Figs. S3, S4.

tition and genetic hitchhiking. Such evolutionary dynamics is prominent in microbial populations [25], in viruses including HIV within a patient [24, 26] and global influenza [21, 27, 28]. In the immune system, clonal competition in BCR repertoires is also observed on short time scales ($\sim$ weeks) in response to the influenza vaccine [16].

Clonal interference among beneficial mutations not only makes selection slower and less efficient, but it also makes the outcome of the evolutionary process less predictable [25]. This is of significant consequence for designing targeted immune-based therapies. Currently, the central challenge in HIV vaccine research is to devise a means to stimulate a lineage producing highly potent broadly neutralizing antibodies (BnAbs). A combination of successive immunization and ART has been suggested as an approach to elicit a stable and effective BnAb response; see e.g. ref. [29]. An optimal treatment strategy should account for clonal interference among BCRs during a rapid immune response to antigen stimulation, which could hamper the emergence of a desired BnAb within the repertoire.

	HIV infected untreated				HIV infected interrupted treatment				Healthy / productive		Healthy / unproductive	
	$\overline{\alpha}_{\text{benef.}}$	$\overline{\alpha}_{\text{del.}}$	$\overline{\kappa}_{\text{benef.}}$	$\overline{\kappa}_{\text{del.}}$	$\overline{\alpha}_{\text{benef.}}$	$\overline{\alpha}_{\text{del.}}$	$\overline{\kappa}_{\text{benef.}}$	$\overline{\kappa}_{\text{del.}}$	$\overline{\alpha}_{\text{benef.}}$	$\overline{\alpha}_{\text{del.}}$	$\overline{\alpha}_{\text{benef.}}$	$\overline{\alpha}_{\text{del.}}$
CDR3	12%	16%	11%	18%	20%	9%	17%	12%	30%	3%	9%	21%
CDR1/2	23%	8%	22%	11%	26%	8%	23%	10%	NA	NA	NA	NA
FWR	8%	14%	9%	17%	6%	19%	8%	17%	7%	20%	11%	24%

TABLE I: **Fraction of beneficial and deleterious mutations in BCRs.** The average fraction of beneficial $\overline{\alpha}_{\text{benef.}}$ and deleterious $\overline{\alpha}_{\text{del.}}$ mutations that reach frequency $x = 80\%$ (based on selection likelihood ratio $g(0.8)$) in different regions of BCRs among VJ-gene classes are reported for HIV patients (with interrupted and without treatment) and for healthy individuals (productive and unproductive lineages). Similarly, the fraction of beneficial $\overline{\kappa}_{\text{benef.}}$ and deleterious $\overline{\kappa}_{\text{del.}}$ mutations that reach frequency $x = 60\%$ followed by extinction (based on interference likelihood ratio $h(0.6)$) are reported for HIV patients with interrupted and without treatment; we cannot estimate the interference likelihood ratio in healthy individuals due to the lack of time-resolved data. The corresponding distributions are presented in Fig. 3 and Fig. S4.

[1] Richman DD, Wrin T, Little SJ, Petropoulos CJ (2003) Rapid evolution of the neutralizing antibody response to HIV type 1 infection. Proc Natl Acad Sci USA 100: 4144–4149.
[2] Moore PL, Ranchobe N, Lambson BE, Gray ES, Cave E, et al. (2009) Limited neutralizing antibody specificities drive neutralization escape in early HIV-1 subtype C infection. PLoS Pathog 5: e1000598.
[3] Liao HX, Lynch R, Zhou T, Gao F, Alam SM, et al. (2013) Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. Nature 496: 469–476.
[4] McMichael AJ, Borrow P, Tomaras GD, Goonetilleke N, Haynes BF (2010) The immune response during acute HIV-1 infection: clues for vaccine development. Nature Rev Immunol 10: 11–23.
[5] Campbell CD, Eichler EE (2013) Properties and rates of germline mutations in humans. Trends Genet 29: 575–584.
[6] Victora GD, Nussenzweig MC (2012) Germinal centers. Annu Rev Immunol 30: 429–457.
[7] Nourmohammad A, Otwinowski J, Plotkin JB (2016) Host-pathogen coevolution and the emergence of broadly neutralizing antibodies in chronic infections. PLoS Genet 12: e1006171.
[8] Weinstein JA, Jiang N, White RA, Fisher DS, Quake SR (2009) High-throughput sequencing of the zebrafish antibody repertoire. Science 324: 807–810.
[9] Elhanati Y, Sethna Z, Marcou Q, Callan CG, Mora T, et al. (2015) Inferring processes underlying B-cell repertoire diversity. Phil Trans R Soc B 370.
[10] Yaari G, Vander Heiden JA, Uduman M, Gadala-Maria D, Gupta N, et al. (2013) Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. Front Immunol 4: 358.
[11] McCoy CO, Bedford T, Minin VN, Bradley P, Robins H, et al. (2015) Quantifying evolutionary constraints on B-cell affinity maturation. Phil Trans R Soc B 370.
[12] Uduman M, Shlomchik MJ, Vigneault F, Church GM, Kleinstein SH (2014) Integrating B cell lineage information into statistical tests for detecting selection in Ig sequences. J Immunol 192: 867–874.
[13] Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR (2013) Genetic measurement of memory B-cell recall using antibody repertoire sequencing. Proc Natl Acad Sci USA 110: 13463–13468.
[14] Laserson U, Vigneault F, Gadala-Maria D, Yaari G, Uduman M, et al. (2014) High-resolution antibody dynamics of vaccine-induced immune responses. Proc Natl Acad Sci USA 111: 4928–4933.
[15] Hoehn KB, Gall A, Bashford-Rogers R, Fidler SJ, Kaye S, et al. (2015) Dynamics of immunoglobulin sequence diversity in HIV-1 infected individuals. Phil Trans R Soc B 370.
[16] Horns F, Vollmers C, Dekker CL, Quake SR (2017) Signatures of selection in the human antibody repertoire: selective sweeps, competing subclones, and neutral drift. bioRxiv doi.org/10.1101/145052.
[17] Vieira MC, Zinder D, Cobey S (2017) Selection and neutral mutations drive pervasive mutability losses in long-lived B cell lineages. bioRxiv : 163741.
[18] SPARTAC Trial Investigators, Fidler S, Porter K, Ewings F, Frater J, et al. (2013) Short-course antiretroviral therapy in primary HIV infection. N Engl J Med 368: 207–217.
[19] DeWitt WS, Lindau P, Snyder TM, Sherwood AM, Vignali M, et al. (2016) A Public Database of Memory and Naive B-Cell Receptor Sequences. PLoS ONE 11: e0160853.
[20] Neher RA, Hallatschek O (2013) Genealogies of rapidly adapting populations. Proc Natl Acad Sci USA 110: 437–442.
[21] Strelkowa N, Lässig M (2012) Clonal interference in the evolution of influenza. Genetics 192: 671–682.
[22] McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652–654.
[23] Smith NGC, Eyre-Walker A (2002) Adaptive protein evolution in Drosophila. Nature 415: 1022–1024.
[24] Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, et al. (2015) Population genomics of intrapatient HIV-1 evolution. eLife 4: e11282.
[25] Lässig M, Mustonen V, Walczak AM (2017) Predicting evolution. Nat Ecol Evol 1: 77.
[26] Pandit A, De Boer RJ (2014) Reliable reconstruction of HIV-1 whole genome haplotypes reveals clonal interference and genetic hitchhiking among immune escape variants. Retrovirology 11: 56.
[27] Luksza M, Lässig M (2014) A predictive fitness model for influenza. Nature 507: 57–61.
[28] Neher RA, Russell CA, Shraiman BI (2014) Predicting evolution from the shape of genealogical trees. eLife 3: e03568.
[29] Caskey M, Klein F, Nussenzweig MC (2016) Broadly neutralizing antibodies for HIV-1 prevention or immunotherapy. N Engl J Med 375: 2019–2021.

Supplementary Information

1 B-cell repertoire data

HIV patients. We analyze B-cell repertoire data from 6 HIV patients from ref. [1] with raw sequence reads accessible from the European Nucleotide Archive under study accession numbers, ERP009671 and ERP000572. We study the repertoire data in two untreated HIV patients with sample accession numbers, ERS664994 - 5001 (patient 1) and ERS139291- 9298 (patient 2) and in four patients with ART interruption at week 48, ERS664966 - 4974 (patient 3), ERS664975 - 4983 (patient 4), ERS664984 - 4992 (patient 5), ERS664976 - 5002 (patient 6). The data covers ~ 2.5 years of study with 6-8 sampled time points per patient; see Table S1 for details.

The B-cell repertoire sequences consist of 150bp non-overlapping paired-end reads (Illumina MiSeq), with one read covering much of the V gene and the other read covering the area around the CDR3 region and the J gene. For the initial processing of the raw reads we use pRESTO [2] (version 0.5.2) with the following steps: We filter sequences for quality (> 32) and length (> 100). The paired end reads that overlap are assumed to be anomalous, and are discarded from the analysis. We assemble the paired reads by aligning against the IMGT reference database of V genes [3], such that an appropriate size gap is inserted between the non-overlapping paired reads. Duplicate sequences are collapsed into unique sequences. The sequences contained a large number of singletons, that is sequences with no duplicates. With an R script, we calculate the minimum hamming distance of each singleton to any non-singleton, H_0. The distribution of H_0 is bimodal, and singletons with $H_0 < 5$ (the minimum between the modes) are discarded, since sequences with few changes are more likely to have appeared due to sequencing errors. Due to lack of barcoding for individual molecules, we only use the unique BCR sequences for analysis and do not incorporate the information on the multiplicity of each sequence.

Healthy individuals. We analyze memory B-cell repertoire data of 3 individuals published in ref. [4]: https://clients.adaptivebiotech.com/pub/robins-bcell-2016. The published data in healthy individuals is already preprocessed for quality control and corrected for sequencing error.

BCR annotation. In both datasets, we annotate the BCR repertoire sequences of each individual (pooled time points) by Partis [5]. Partis uses very large amounts of memory, so the initial (cache-parameters) stage is run on a subset of 200,000 random sequences, and the annotation stage is run on the full set of sequences. We process the output of Partis in R, which includes the estimated V gene/allele, J gene/allele, location of the CDR3 region, and an inferred naive sequence (germline before hyper-mutation). Sequences which have indels outside of the CDR3 are discarded. We partition the sequences into two groups: productive BCRs, which are in-frame and have no stop codons, and the unproductive BCRs. The sequences are further annotated by processing the inferred naive sequences with MiXCR [6, 7], which gives the CDR1, CDR2 and framework regions.

Lineage reconstruction. To identify BCR lineages, we first group sequences by the assigned V gene, J gene and CDR3 length, and then used single linkage clustering with a threshold of 90% hamming distance. A similar threshold has been previously suggested by ref. [8] to identify BCR lineages. Clusters of small size (< 20) are discarded from our analysis. For each cluster, there may be multiple inferred naive sequences, as this is an uncertain estimate, and the most common naive sequence is chosen to be the outgroup for genealogy reconstruction. See Table S1 for detailed statistics of BCR lineages in each individual.

Unproductive BCRs. Due to a larger sequencing depth in healthy individuals, we are able to reconstruct relatively large unproductive BCR lineages. Unproductive sequences are BCRs that were generated but due to a frameshift or insertion of stop codons were never expressed. These BCRs reside with productive (functional) BCRs in a nucleus and undergo hypermutation during B-cell replication, and therefore, provide a suitable null expectation for somatic

evolution during affinity maturation.

2 Inference of lineage phylogenies

Lineage genealogy reconstruction. For each lineage and its aligned sequences we reconstruct its underlying genealogical tree. We use FastTree [9] to construct the initial tree by maximum parsimony. We use this tree as seed for the maximum likelihood construction of the phylogeny with RAxML [10], using the GTRCAT substitution model. In the last step of tree topology reconstruction, we use the GTRGAMMA substitution model to optimize sequence divergence along the tree (i.e., branch lengths). We use a maximum likelihood approach to reconstruct nucleotide sequences of internal nodes on the tree [11]. We do not include the positions with gaps in the multi-sequence alignment in inference of tree topology and the nucleotide mutations along the tree.

We use the inferred naive sequence (germline) as the outgroup of the genealogy. The root of the tree may be some mutations away from the last common ancestor of the sampled sequences. This may be due to a number of initial rounds of hypermutation prior to secretion of the first selected B-cell, or alternatively, due to incorrect assignment of the germline allele during annotation; a fraction of V, D and J alleles circulating in the human population are missing from the existing reference datasets like IMGT [3]. In order to minimize the effect of such allele mis-assignments, we discard the mutations that separate the inferred germline sequence and the last common ancestor (root) of the tree from our analysis (i.e., mutations common to all sequences).

Inference of branching time along a phylogeny. To characterize the branch length statistics of lineages in units of divergence time (used in Fig. 2B and Fig. S1B), we use a maximum-likelihood approach and a probabilistic model to annotate internal nodes of a tree with times of occurrence, given the topology of the tree and the mutations on the branches. Internal nodes represent replication events, which may carry new mutations assigned to branches by the ancestral sequence reconstruction procedure. The model can also correct observation times of external nodes on the tree, given sufficient evidence. The model assumes a tree with n nodes, $[o_1, \ldots, o_m, i_{m+1}, i_n]$, where o_k are the sampled sequences in the leafs of the tree, and i_k are the internal nodes of the tree. The observed nodes are annotated with their sampling times, $\mathbf{T} = [T_1, \ldots, T_m]$. Branches of the tree are annotated with their mutational distances, $\mathbf{d} = [d_1, \ldots, d_n]$. We only consider synonymous mutations for computation of mutational distances. The model estimates mutation rate, μ and the times $\mathbf{t} = [t_1, \ldots, t_n]$ of the nodes, by maximizing the likelihood:

$$P(\mathbf{d}, \mathbf{T}|\mathbf{t}, \mu) = \prod_{k=1}^{m} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(t_k - T_k)^2}{2\epsilon}\right] \prod_{j=m+1}^{n} \frac{(\mu\tau_j)_j^d}{d_j!} \exp[\mu\tau_j], \tag{1}$$

where $\tau_j = t_j - t_{\mathcal{A}(j)}$, the time difference between a node in the tree and its parent, is constrained to be positive. The model assumes Gaussian measurement error with standard deviation σ of the sampling time for the observed nodes, and the Poisson model for mutations on tree branches, with rate μ. To limit the search space of the optimization algorithm we constrain the times $\mathbf{t}$ to be discrete; the units of time used in our data are weeks. Here we set $\sigma = 5$. The optimization is solved by an iterative procedure in which times of nodes are changed by one unit in the direction of score increase, until convergence.

3 Inference of selection from lineage tree statistics

We compare genealogies of B-cell lineages in HIV patients with healthy individuals to characterize the evolutionary selection during affinity maturation in response to chronic infection. Structure of genealogies has been linked to evolutionary modes in a population [12]. Rapid evolution under positive selection leads to skewed tree topologies

and elongated terminal branches [11, 13–15], compared to neutral evolution [12, 16] (Fig. S1A).

Asymmetric tree branching. We characterize the asymmetry of trees by the branching imbalance of the last common ancestor at the root of the tree - the last common ancestor may be a number of mutations away from the germline progenitor. We define the weight of each node in a tree by the number of leaves (terminal nodes) within its clade; see Fig. S1A [15]. The weight of the last common ancestor $w_{\text{anc..}}$, i.e., the total number of leaves in a tree, and its daughters, w_{D_1}, w_{D_2}, are indicated in the simulated trees of Fig. S1A. In rapid evolution under selection, the first branching event produces highly imbalanced sub-clades, and hence, extreme values for the weight of the first daughter nodes [15]. In contrast, neutral evolution predicts a uniform distribution of tree weights of ancestral sub-clades. Fig. 2A shows a U-shaped distribution of relative weights of daughters to the ancestor $w_D/w_{\text{anc.}}$ in lineages reconstructed from BCRs in HIV patients and in healthy individuals, indicating intra-lineage selection during affinity maturation.

Terminal branch statistics. In lineages under selection, the descendent sequences (leaves of a tree) are likely to coalesce to an ancestor with high fitness, resulting in long terminal branches in a tree (Fig. S1A) [14, 15] — branch length statistics are estimated in units of divergence time, rather than sequence hamming distance; see Section 2 of SI for inference of branching times along a phylogeny. In Fig. 2B we compare the ratio of mean terminal branch length of a lineage to the averaged length of all the branches in a lineage. The distribution of this branch length ratio in HIV patients show an excess of lineages with relatively long terminal branches, compared to the expected distribution for simulated neutral lineages of the same size (Kingman's coalescence); see Section 5 of SI. A similar trend with a weaker signal is seen in healthy individuals (Fig. 2B). To make sure that the strong signal in HIV patients is not only due to sampling of the repertoire over multiple time points, we repeat the same analysis on the subset of lineages which are sampled in only one time point. Fig. S1B shows a comparable over-representation of long terminal branches in this subset.

Site frequency spectrum (SFS). The SFS is the probability density $f(\nu)$ of observing a derived mutation (allele) with a given frequency ν in a lineage. A mutation that occurs along the phylogeny of a lineage forms a clade and is present in all the descendent nodes (leaves) of its clade (see Fig. S1A). Therefore, SFS carries information about the shape of the phylogeny, including both the topology and the branch lengths. In neutrality, mutations rarely reach high frequency, and hence, the SFS decays monotonically with allele frequency as, $f(\nu) \sim \nu^{-1}$ [16]. In phylogenies with skewed branching, many mutations reside on the larger sub-clade following a branching event, and hence, are present in the majority of the descendent leaves on the tree. The SFS of such lineages is often non-monotonic with an upturn in the high frequency part of the spectrum and a steeper drop ($\sim \nu^{-\beta}$ with $\beta > 1$) in the low frequency part of the spectrum [14]. To identify the targets of selection, we classify mutations based on the region they occur in. BCRs are made up of the three immunologically important complementarity-determining regions (CDRs) [17], CDR1, CDR2 and CDR3 and the remaining part of the V and J genes referred to as the framework region (FWR).

Fig. 2C shows SFS in lineages of HIV patients and in healthy individuals. We see a signifiant upturn of SFS polarized on non-synonymous mutations in pathogen-engaging CDR3 regions; this signal of selection is strongest in HIV patients with an order of magnitude increase in the high end of the spectrum (Fig. 2C). SFS polarized on mutations in other regions show steeper drop in the low frequency side of the spectrum compared to neutral expectation. Fig. S2 shows SFS of the unproductive BCR lineages in healthy individuals, with a comparable steep drop in low frequencies. A similar pattern of SFS has recently been reported for lineages of B-cell repertoires following influenza vaccination [18].

For analysis of lineage tree statistics, i.e., the weight imbalance, terminal branch statistics and SFS, we only rely on the relatively large lineages with size (> 50) leaves.

4 Selection and clonal interference likelihood ratios

Selection likelihood ratio. Hypermutations during affinity maturation create new clades within a lineage. The frequency x of these clades change over time, as shown by the schematic in Fig. 3A. A mutation under positive (or negative) selection should reach a higher (lower) frequency than a neutral mutation. Many population genetics tests, such as the McDonnell-Kreitman test for positive selection [19], rely on a comparison between statistics of substitutions (i.e., mutations that fix within a population) and the circulating polymorphisms within species. Unlike phylogenies based on the species divergence, B-cell lineages form genealogies with many mutations that rise to intermediate frequencies as polymorphisms but often do not fix within a lineage. Here, instead of relying on the substitution statistics, we use the history of polymorphisms to quantify selection in B-cell lineages. In particular, we estimate the frequency propagator $G(x)$ [20] as the likelihood that a new mutation (allele) appearing in a lineage reaches frequency x at some later time within a lineage (see schematic in Fig. 3A).

To estimate intra-lineage selection, we compare the likelihood of an amino acid changing non-synonymous mutation reaching a given frequency x at any point in its time trace, $G(x)$ to that likelihood for a synonymous mutation in the same lineage, $G_0(x)$, and detemine the selection likelihood ratio (Fig. 3A) [20],

$$g(x) = \frac{G(x)}{G_0(x)}. \tag{2}$$

Due to heterogeneity and context dependence of mutation rates in different regions of BCRs, we evaluate the likelihood ratio separately for each region, namely the CDR3, CDR1 & CDR2 (pooled together) and framework regions (FWR). In the Fig. S5 we show the robustness of the region-specific selection likelihood ratio with respect to such mutational biases.

Interference likelihood ratio (time-ordered selection). Clonal competition among beneficial mutations on different genetic backgrounds is a characteristic of evolution in asexual populations. In the absence of clonal interference, beneficial mutations can readily fix in a population after they rise to intermediate frequencies, beyond which stochastic effects would not impact their fate [21]. Clonal interference reduces the efficacy of selection, resulting in a quasi-neutral regime of evolution [22].

To examine the amount of clonal competition among BCRs of a lineage, we consider time ordered selection propagators (interference propagators) indicating the likelihood that a mutation reaches frequency x and later goes extinct, $H(x) = G(x) \times G(0|x)$; here $G(0|x)$ is the conditional probability that a mutation trajectory decays to frequency 0 given that it starts from frequency x; see schematic in Fig. 3A. We estimate the interference likelihood ratio by comparing the probability of a non-synonymous mutation to reach a frequency x and later go extinct $H(x)$ to the same scenario for synonymous mutations, $H_0(x)$ (Fig. 3A),

$$h(x) = \frac{H(x)}{H_0(x)} \equiv \frac{G(x) \times G(0|x)}{G_0(x) \times G_0(0|x)}. \tag{3}$$

Fraction of selected mutations based on the likelihood ratios. Following the well established tradition of population genetics, we assume that synonymous mutations that do not change the amino acid provide a neutral gauge for evolution. In the case of frequency $x = 1$ the propagator ratio $g(x)$ becomes equal to the ratio of the fixation probability $(d/n)\Big/(d_0/n_0)$ where d and d_0 are respectively the number of fixed non-synonymous and synonymous polymorphisms and n and n_0 are total number of polymorphisms is each class. In other words, $g(x = 1)$ is equivalent to the McDonald-Kreitman test for selection based on the observed polymorphisms [19].

Selection likelihood ratio $g(x) = G(x)/G_0(x)$ larger than 1 implies an over-representation of non-synonymous compared to synonymous changes that reach frequency x and is indicative of beneficial amino acid changes in a

given region. Assuming a total of N non-synonymous mutations, we expect $NG_0(x)$ of these mutations to reach frequency x by neutral evolution, and at least a fraction $\alpha_{\text{benef.}}(x) = (N(x) - NG_0(x))/N(x) = (g(x) - 1)/g(x)$ of these mutations to be beneficial [23]. On the other hand, a selection likelihood ratio smaller than 1 indicates negatively selected amino acid changes in a given region. The deviation from the expected number of non-synonymous mutations in neutrality, $NG_0(x) - N(x)$, is an estimate for the number of mutations that were suppressed due deleterious fitness effects, indicating that at least a fraction $\alpha_{\text{del.}}(x) = 1 - g(x)$ of non-synonymous mutations to be under negative selection [23]. Similarly, we can compute the fraction of beneficial and deleterious mutations that are impacted by clonal interference, $\kappa_{\text{benef.}}(x) = (h(x) - 1)/h(x)$ for $h(x) > 1$, and $\kappa_{\text{del.}}(x) = 1 - h(x)$ for $h(x) < 1$.

Robustness of selection inference. It should be noted that the heterogenous and context dependent somatic hypermutation rates during affinity maturation [24–28] introduce BCR-specific biases that could influence inference of selection. In order to verify the robustness of our method, we have simulated the process of affinity maturation based on two distinct BCR-specific hyper-mutation models [24,26,28] along the inferred BCR lineage phylogenies; see Section 5 of SI for details. Fig. S5 shows that the region-specific likelihood ratios $g(x)$, $h(x)$ are insensitive to the heterogenous hyper-mutation statistics and such biases do not produce spurious evidence for selection and clonal interference. In addition, the likelihood ratio is insensitive to the initial frequency of an allele within a lineage [20] and provides a robust measure for inference of selection in evolving genealogies.

Inference of likelihood ratio statistics from data. The descendants of a given mutation α on a lineage tree define the clade $\mathcal{C}^\alpha$. We evaluate the frequency of mutation α at time t, $x^\alpha(t)$ as the fraction of the observed sequences (leaves of the tree) from time point t that reside within the clade $\mathcal{C}^\alpha$. The fraction of non-synonymous and synonymous mutations that reach frequency x during their history define the selection propagators $G(x)$ and $G_0(x)$, respectively. To infer statistically significant evidence for selection, we estimate propagators based on the mutations pooled from lineages of common gene classes, e.g. lineages with common V gene (Fig. 3B) or common V & J genes (Fig. 3C and Fig. S4). We evaluate the expected error of a propagator at frequency x, by assuming binomial sampling from the total of N non-synonymous and N_0 synonymous mutations (i.e., all the mutations observed in a given gene class). This results in the sampling errors $\sigma^2(x) = G(x)(1 - G(x))/N$ for non-synonymous and $\sigma_0^2(x) = G_0(x)(1 - G_0(x))/N_0$ for synonymous mutations, and a corresponding propagated error for the ratio, $g(x)$ in eq. 2. We use a similar approach to estimate the error for the interference likelihood ratio, $h(x)$ in eq. 3.

5 Simulations

Simulated trees. In Fig. 2B, we compare the branch length characteristics of the BCR genealogies with the neutral expectation from 2000 simulated trees with Kingman's coalescence, generated by the beta coalescent algorithm with parameter $\alpha = 1$ [29]. The sizes of the simulated trees in the neutral ensemble are drawn from the BCR lineage size distribution in HIV patients. The schematic trees in Fig. S1A are also generated similarly by the beta coalescent algorithm [29] with parameters $\alpha = 1$ for neutral evolution and $\alpha = 2$ for rapid adaptation.

Null model for context-dependent affinity maturation. We simulate mutations along BCR lineage trees according to two context-dependent models of hyper-mutation, (i) IGoR statistics [28] and (ii) S5F [24]. For a given branch on a lineage tree, we draw a number of mutations equal to the branch length from a multinomial distribution with position-specific weights determined by the hypermutation models. Due to the changes in the sequence, we update the position weights at each internal node of the tree to account for context-dependent hyper-mutation rates. This procedure reshuffles the identify of mutations along BCRs according to the neutral hyper-mutation models, while preserving the shape of tree. Fig. S5 shows that the propagator statistics do not recover evidence for region-specific selection in BCRs in the simulated lineages. Therefore, the original selection signal in Fig. 3 is not

reflecting any spurious effect due to heterogenous mutation rates.

Supplemental references

[1] Hoehn KB, Gall A, Bashford-Rogers R, Fidler SJ, Kaye S, et al. (2015) Dynamics of immunoglobulin sequence diversity in HIV-1 infected individuals. Phil Trans R Soc B 370.

[2] Vander Heiden JA, Yaari G, Uduman M, Stern JNH, O'Connor KC, et al. (2014) pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. Bioinformatics 30: 1930–1932.

[3] Lefranc MP, Lefranc G (2001) The Immunoglobulin FactsBook. Academic Press.

[4] DeWitt WS, Lindau P, Snyder TM, Sherwood AM, Vignali M, et al. (2016) A Public Database of Memory and Naive B-Cell Receptor Sequences. PLoS ONE 11: e0160853.

[5] Ralph DK, Matsen FA (2016) Consistency of VDJ Rearrangement and Substitution Parameters Enables Accurate B Cell Receptor Sequence Annotation. PLoS Comput Biol 12: e1004409.

[6] Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, et al. (2015) MiXCR: software for comprehensive adaptive immunity profiling. Nat Methods 12: 380–381.

[7] Bolotin DA, Poslavsky S, Davydov AN, Frenkel FE, Fanchi L, et al. (2017) Antigen receptor repertoire profiling from RNA-seq data. Nat Biotechnol 35: 908–911.

[8] Gupta NT, Adams KD, Briggs AW, Timberlake SC, Vigneault F, et al. (2017) Hierarchical Clustering Can Identify B Cell Clones with High Confidence in Ig Repertoire Sequencing Data. J Immunol 198: 2489–2499.

[9] Price MN, Dehal PS, Arkin AP (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS ONE 5: e9490.

[10] Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 1312–1313.

[11] Neher RA, Russell CA, Shraiman BI (2014) Predicting evolution from the shape of genealogical trees. eLife 3: e03568.

[12] Wakeley J (2007) Coalescent Theory: an introduction. Roberts Publishers.

[13] Desai MM, Walczak AM, Fisher DS (2013) Genetic diversity and the structure of genealogies in rapidly adapting populations. Genetics 193: 565–585.

[14] Neher RA, Hallatschek O (2013) Genealogies of rapidly adapting populations. Proc Natl Acad Sci USA 110: 437–442.

[15] Dayarian A, Shraiman BI (2014) How to infer relative fitness from a sample of genomic sequences. Genetics 197: 913–923.

[16] Kingman J, 1982 On the genealogy of large populations, in "Essays in Statistical Science"(J. Gani and EJ Hannan, Eds.) J 19: 27–43.

[17] Janeway CA, Travers P, Walport M, Shlomchik M (2005) Immunobiology: the immune system in health and disease (Garland Science, New York).

[18] Horns F, Vollmers C, Dekker CL, Quake SR (2017) Signatures of selection in the human antibody repertoire: selective sweeps, competing subclones, and neutral drift. bioRxiv doi.org/10.1101/145052.

[19] McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652–654.

[20] Strelkowa N, Lässig M (2012) Clonal interference in the evolution of influenza. Genetics 192: 671–682.

[21] Desai MM, Fisher DS (2007) Beneficial mutation–selection balance and the effect of linkage on positive selection. Genetics 17: 385–394.

[22] Schiffels S, Szöllősi GJ, Mustonen V, Lässig M (2011) Emergent neutrality in adaptive asexual evolution. Genetics 189: 1361–1375.

[23] Smith NGC, Eyre-Walker A (2002) Adaptive protein evolution in Drosophila. Nature 415: 1022–1024.

[24] Yaari G, Vander Heiden JA, Uduman M, Gadala-Maria D, Gupta N, et al. (2013) Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. Front Immunol 4: 358.

[25] Yaari G, Benichou JIC, Vander Heiden JA, Kleinstein SH, Louzoun Y (2015) The mutation patterns in B-cell immunoglobulin receptors reflect the influence of selection acting at multiple time-scales. Philos Trans R Soc Lond, B, Biol Sci 370: 20140242.

[26] Elhanati Y, Sethna Z, Marcou Q, Callan CG, Mora T, et al. (2015) Inferring processes underlying B-cell repertoire diversity. Phil Trans R Soc B 370.

[27] Hoehn KB, Lunter G, Pybus OG (2017) A phylogenetic codon substitution model for antibody lineages. Genetics 206: 417–427.

[28] Marcou Q, Mora T, Walczak AM (2017) IGoR: a tool for high-throughput immune repertoire analysis .

[29] Neher RA, Kessinger TA, Shraiman BI (2013) Coalescence and genetic diversity in sexual populations under selection. Proc Natl Acad Sci USA 110: 15836–15841.

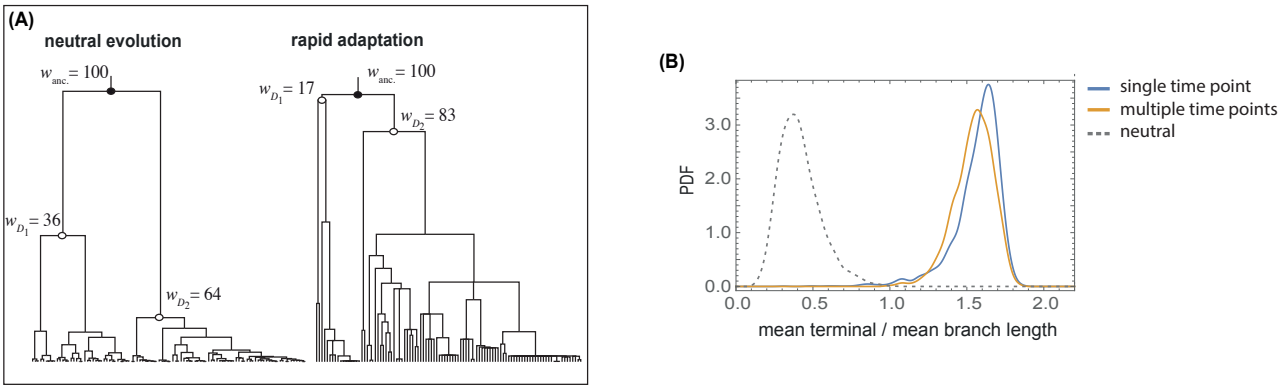

Figure S1: **Impact of selection on lineage tree statistics. (A)** Simulated phylogenies for neutral evolution (left) and rapid adaptation with positive selection (right), generated by the coalescence package [29] and plotted with FastTree [9]. The weights of the ancestral node $w_{\text{anc.}}$ and its daughters w_{D_1}, w_{D_2} (i.e., their clone size) are indicated in each phylogeny. **(B)** Branch length statistics for lineages sampled in only one time point from HIV patients. The distribution of mean terminal branch length (in units of divergence time) relative to the mean length of all branches is comparable between BCR lineages of HIV patients that are present in only a single time point (blue) and lineages sampled over multiple time points (orange). The corresponding distribution for the simulated neutral trees (similar to Fig. 2B) is shown by a dotted line. The elongated terminal branches in BCR lineages is indicative of positive selection.

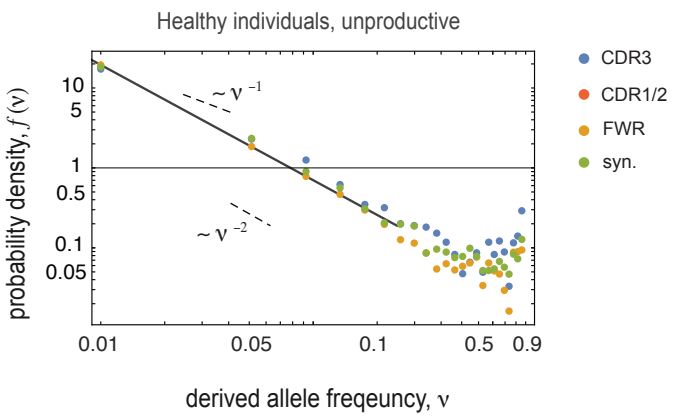

Figure S2: **Site frequency spectrum of the unproductive BCR lineages.** SFS $f(\nu)$ is shown for mutations in different regions of BCRs (distinct colors) in unproductive lineages of healthy individuals; see Fig. 2C for SFS of productive lineages.

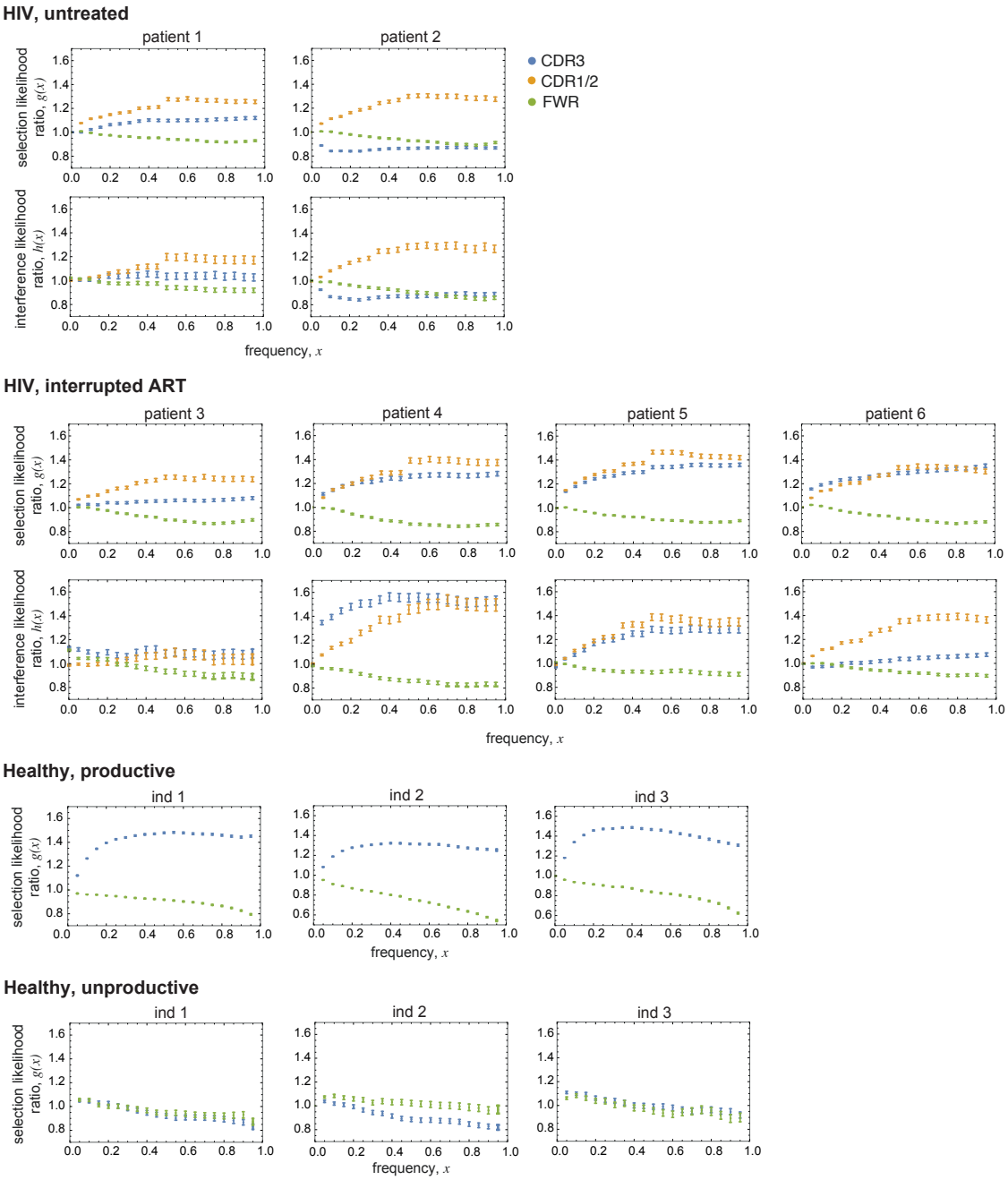

Figure S3: **Selection and interference likelihood ratios in all individulas.** Panels show selection and interference likelihood ratios $g(x)$, $h(x)$ in HIV patients (untreated and with interrupted ART) and the selection likelihood ratio $g(x)$ in productive and unproductive lineages of healthy individuals, estimated from all lineages in each individual. We consistently see strong evidence for negative selection in FWR regions of productive lineages and positive selection in both or either of CDR regions. We do not see such distinction in unproductive lineages. Note that the repertoire level averaged likelihood ratios are highly coarse grained statistics and miss the gene-specific evidence for selection and clonal interference, as shown in Fig. 3.

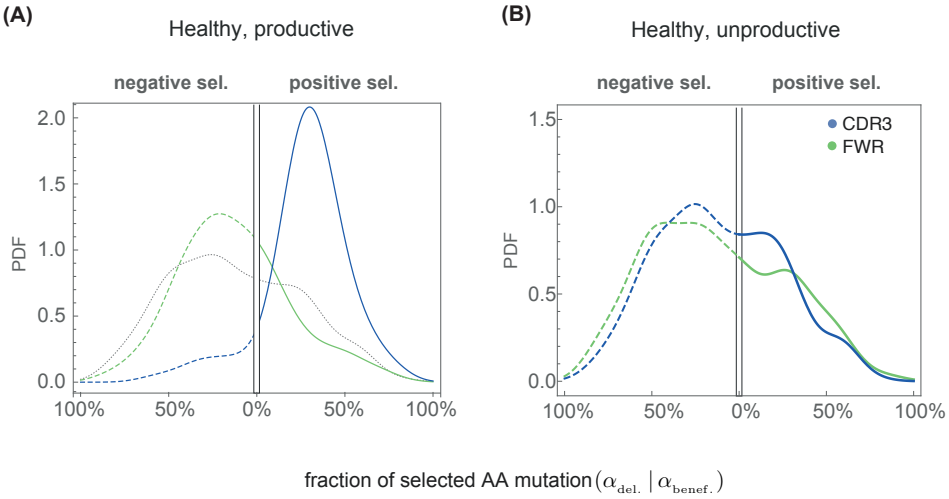

Figure S4: **Fraction of selected BCR mutations in healthy individuals.** The probability density across distinct VJ-gene classes for the (minimum) fraction of beneficial (right) $\alpha_{\text{benef.}}$ and deleterious (left; inverted x-axis) $\alpha_{\text{del.}}$ amino acid changes that reach frequency $x = 80\%$ within a lineage is shown for different regions of BCRs in **(A)** productive and **(B)** unproductive lineages of healthy individuals. Similar to HIV patients (Fig. 3), the CDR3 mutations in productive lineages of healthy individuals are under positive selection, whereas the FWR mutations are under negative selection. We do not infer any significant differences between selection patterns in CDR3 and in FWR region of unproductive lineages. The probability density for mutations pooled from both regions of unproductive lineages is shown as the null expectation (dotted gray line), similar to Fig. 3.

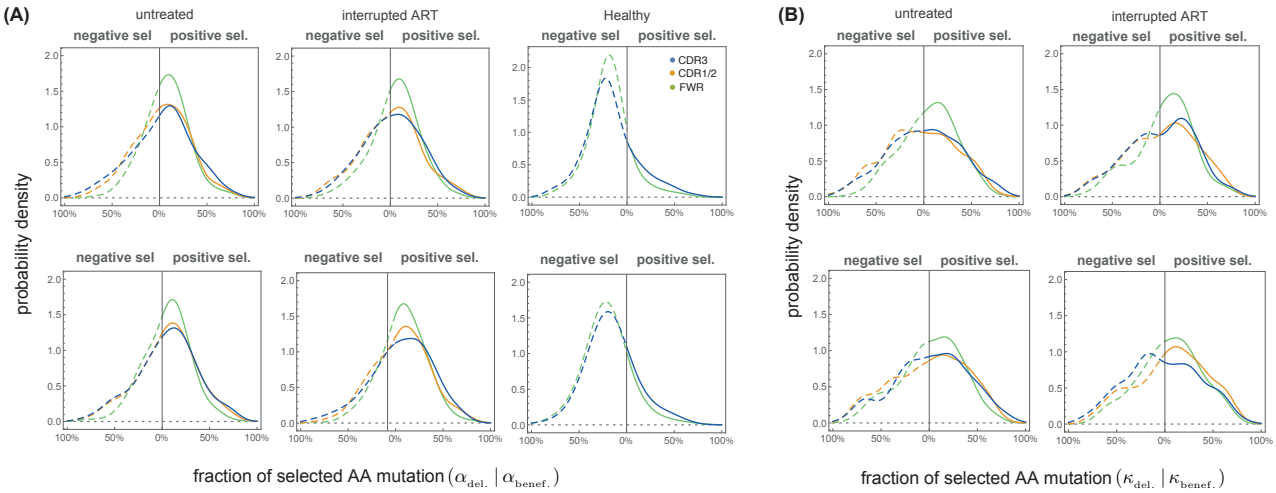

Figure S5: **Robustness of selection inference to BCR hypermutation biases.** The figure shows the statistics of selected mutations for simulated neutral hypermutation processes along the inferred B-cell lineages, as described in Section 5 of SI. We use two hypermutation models, IGoR statistics [28] (top row) and the S5F model [24] (bottom row). Similar to Fig. 3, each panel shows the probability density across distinct VJ-gene classes for **(A)** the minimum fraction of beneficial / deleterious mutations ($\alpha_{\text{benef.}} \,/\, \alpha_{\text{del.}}$) that reach frequency $x = 80\%$, and **(B)** for beneficial / deleterious mutation fractions ($\kappa_{\text{benef.}} \,/\, \kappa_{\text{del.}}$) that reach frequency $x = 60\%$ and later go extinct. These statistics are estimated for mutations in different regions of BCRs (colors) in healthy individuals, HIV patients with interrupted ART and in untreated HIV patients. The region-specific pattern of selection seen in Fig. 3 is simulated lineages with context-dependent hypermutation models.

		HIV infected (untreated)		HIV infected (interrupted ART at week 48)				Healthy		
		patient 1	patient 2	patient 3	patient 4	patient 5	patient 6	ind 1	ind 2	ind 3
	size									
# productive lineages	> 20	3,335	3,702	3,164	2,125	4,342	4,155	21,486	15,755	15,782
	> 50	785	773	613	485	1,246	1,205	5,242	3,978	3,390
# unproductive lineages	> 20	0	0	0	0	0	0	897	1041	962
	> 50	0	0	0	0	0	0	177	198	155
time samples (weeks)		0, 4, 16, 24, 52, 72, 120	0, 4, 12, 16, 24, 52 60, 108	4, 12, 16, 24, 52, 60, 108	4, 12, 16, 24, 52, 60, 108	4, 12, 16, 24, 52, 60, 108	0, 4, 16, 24, 52, 60, 108	0	0	0

Table S1: Statistics of reconstructed BCR lineages with size (> 20) and (> 50) and the sampled time points after the start of the study in HIV infected patients and in healthy individuals.