Elevated HIV viral load is associated with higher recombination rate *in vivo*

Elena V. Romero[1] and Alison F. Feder[1]

[1]Department of Genome Sciences, University of Washington, Seattle, WA, USA

Abstract

HIV's exceptionally high recombination rate drives its intra-host diversification, enabling immune escape and multi-drug resistance within people living with HIV. While we know that HIV's recombination rate varies by genomic position, we have little understanding of how recombination varies throughout infection or between individuals as a function of the rate of cellular co-infection. We hypothesize that denser intra-host populations may have higher rates of co-infection and therefore recombination. To test this hypothesis, we develop a new method (Recombination Analysis via Time Series Linkage Decay, or RATS-LD) to quantify recombination using autocorrelation of linkage between mutations across time points. We validate RATS-LD on simulated data under short read sequencing conditions and then apply it to longitudinal, high-throughput intra-host viral sequencing data, stratifying populations by viral load (a proxy for density). Among sampled viral populations with the lowest viral loads ($<$ 26,800 copies/mL), we estimate a recombination rate of 1.5×10^{-5} events/bp/generation (95% CI: $7 \times 10^{-6} - 2.9 \times 10^{-5}$), similar to existing estimates. However, among samples with the highest viral loads ($>$ 82,000 copies/mL), our median estimate is approximately 6 times higher. In addition to co-varying across individuals, we also find that recombination rate and viral load are associated within single individuals across different time points. Our findings suggest that rather than acting as a constant, uniform force, recombination can vary dynamically and drastically across intra-host viral populations and within them over time. More broadly, we hypothesize that this phenomenon may affect other facultatively asexual populations where spatial co-localization varies.

Introduction

Recombination is a key evolutionary driver, permitting organisms to purge deleterious mutations and combine beneficial ones. These functions are critical in human immunodeficiency virus (HIV) [32] where intra-host recombination promotes viral diversification, immune escape [36], multi-drug resistance [20, 28, 31] and the maintenance of genomic integrity in spite of an exceptionally high viral mutation rate [21, 35]. While average HIV recombination rates have been previously estimated on the order of 10^{-5}/bp/generation or higher [46, 29, 6], new investigations have revealed that this average fails to capture the full variation in recombination rate in multiple settings [43, 48, 50, 51]. Because of recombination's critical role in HIV's evolutionary success, it is important to investigate the factors that underlie its variation.

Two processes govern HIV's recombination rate: cellular co-infection and reverse transcriptase template switching (Figure 1A). When multiple viruses co-infect the same cell, the virions emerging

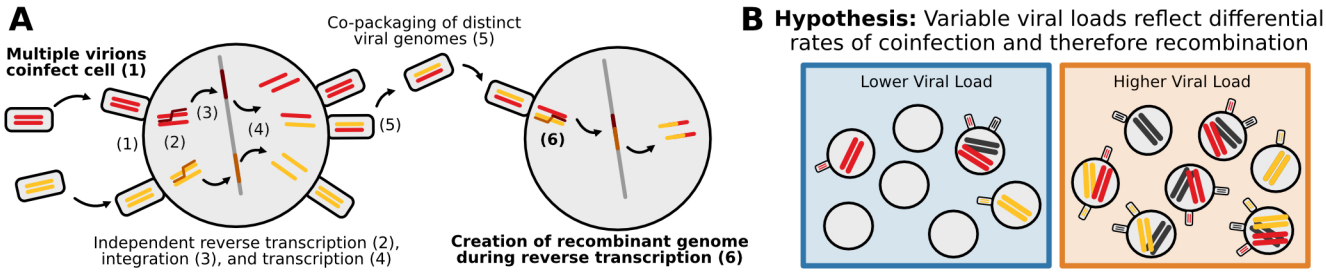

Figure 1: **Drivers of HIV recombination rate. (A)** Cartoon schematic of HIV replication cycle leading to recombination between distinct viral genomes. Steps that determine the rate of recombination are shown in bold. **(B)** HIV populations with higher viral loads (orange) may experience higher rates of viral coinfection and therefore higher rates of recombination than HIV populations with lower viral loads (blue).

from that cell can contain two distinct but co-packaged viral genomes (as opposed to virions produced by singly infected cells which contain two identical genomes) [19, 8]. When those virions infect new cells, template switching between the two co-packaged genomes can result in a recombinant genome inheriting alleles from both parent strands [33, 18, 17]. Increased rates of co-infection should result in more virions with distinct co-packaged genomes, in which recombination can create new viral genotypes, while increased rates of template switching should result in more recombination events per genome pair.

Variation in the template switching rate has already been recognized as an important mediator of viral recombination both *in vitro* and *in vivo*. Template switching and, by extension, recombination can be suppressed by decreased sequence homology between the two template strands [4, 3, 5, 30]. This effect is dose dependent [3, 30] and extremely low levels of sequence homology ($< 63\%$) near fully block template switching and recombination, leading to a breakdown of genomic integrity [35]. Recent sequence context aware work from human-derived HIV-1 sequences has identified particular genomic positions as hot and cold spots for template switching and recombination, which impact the ways in which viral populations diversify [27, 43, 48, 50].

Despite its potential importance, we know little about how variation in co-infection rates impacts viral recombination *in vivo*. While studies of HIV in mouse models and cell cultures have demonstrated that increased coinfection is associated with an increase in recombinant viruses [23], it is unknown if this effect persists among naturally-occurring HIV infections in humans. We expect coinfection to be mediated by the abundances and locations of both viruses and the CD4+ T cells they infect [19, 8]. Therefore, we expect viral load - a proxy for overall viral abundance - to be related to the rate at which co-infection occurs, even if non-linearly. Since viral load and CD4+ T cell count vary across individuals and within them over time, we hypothesize that viral intra-host demography modulates co-infection rates and may therefore play an important but previously uncharacterized role in determining the viral recombination rate in human hosts (Figure 1B).

Existing statistical approaches to estimate recombination rate have made this effect challenging to detect *in vivo*. Current approaches to estimate the HIV recombination rate fall into two broad categories: identifying specific breakpoints in sequencing data [47, 26, 50, 25], or employing overall patterns of linkage disequilibrium to infer intra-host recombination rates (breakpoint-free) [6, 29, 46]. Both rely on the analysis of single genome amplification and sequencing data [39], which substantially limits sampling depth and makes quantitative comparisons across different viral load groups challenging. Note, Song et al found no differences in recombination breakpoint counts at different viral loads, likely because the number of breakpoints visible at any specific point in time is small [50]. While bulk-sequencing data enables deeper sampling than single genome sequencing, the relatively short reads

commonly used in bulk-sequencing protocols limit linkage measurement except between nearby sites. However, HIV's high recombination rate results in significant breakdown of genetic linkage even within the relatively short length of a typical bulk sequencing paired-end read. Therefore, adapting breakpoint-free methods to work with bulk sequencing data has the potential to more sensitively quantify the effects of recombination, even when stratifying data sets by viral load.

Here, we test for viral load-associated differences in recombination rate through developing a new approach called Recombination Analysis via Time Series Linkage Decay (RATS-LD) which quantifies recombination rates from longitudinal bulk sequencing data. We validate that this method can successfully recover recombination rates in simulated short read paired-end data. To test for positive associations between HIV viral loads and recombination rate, we apply RATS-LD to longitudinal intra-host HIV bulk sequencing data from individuals with varying viral loads sequenced with careful controls for PCR recombination [53, 52]. We find that while HIV populations with viral loads in the lowest third of data set ($< 26,800$ copies/mL) have recombination rates in line with previous estimates (1.5×10^{-5}/bp/generation), populations with viral loads in the upper third ($> 82,000$ copies/mL) have a median recombination rate that is nearly six-fold higher. Furthermore, within single individuals we observe patterns of viral load and effective recombination rate increasing concurrently. These findings demonstrate that recombination rate is not a static parameter, and intra-host conditions can mediate the strength of its effect both across and within individuals over time. Since recombination mechanisms generally require that two genomes physically find each other in space, such contact-network mediated effects are also likely to occur in populations beyond HIV.

Results

Measuring intra-host recombination rates using short read data

We developed an approach to leverage short read sequencing data to quantify recombination rates in longitudinally-sampled intra-host viral populations. Our approach exploits the decay in genetic linkage following the emergence of new mutations in the population. New mutations will initially occur on a single genetic background and will be measurably linked to any other single nucleotide polymorphisms (SNPs) on that background. Over time, however, recombination will disassociate linked groups of mutations, leading to a relaxation back towards linkage equilibrium. This disassociation will be faster if the SNPs are separated by greater physical distances (d, measured in nucleotides), as the increased distance provides more opportunities for recombination events between the SNPs. Specifically, for a pair of segregating SNPs, the linkage decay, given a recombination rate ρ, after a period of time Δt is given as follows [16]:

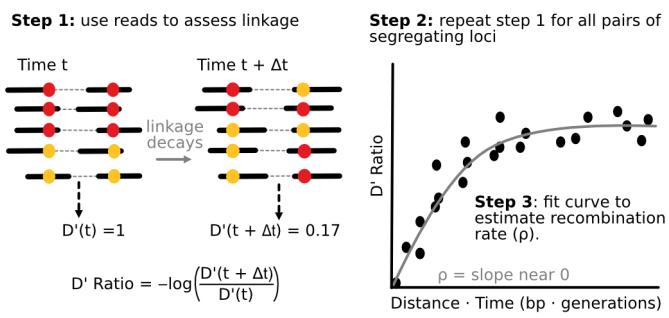

Figure 2: **Linkage decay patterns (quantified by the D' ratio) enable recombination rate estimation.** D' values quantify genetic linkage between pairs of SNPs at a given time point. Negative log ratios of D' values across two sampled time points are positively associated with the product of the time between sample points and distance between SNPs. The slope of this relationship can be used to estimate the recombination rate, $\hat{\rho}$. We call this method RATS-LD.

$$D'(t + \Delta t) = D'(t)e^{-\rho d \Delta t} \tag{1}$$

where D' is a measure of statistical association between alleles at two sites. If D' = 1, the allelic identity at site 1 maximally describes the identity at site 2, whereas if D' = 0, the allelic identity at site 1 reveals nothing about the identity at site 2. Thus, by examining the genetic association between pairs of SNPs longitudinally, we can quantify the rates of this statistical decay at different distances to estimate ρ:

$$D'_{\text{ratio}} = -\log\left(\frac{D'(t+\Delta t)}{D'(t)}\right) = \rho d\Delta t. \tag{2}$$

Our method works by computing the ratio on the left hand side of equation 2 (which we call the D' ratio) over many pairs of SNPs between multiple time points. Then, it uses these D' ratios to derive an empirical relationship between the linkage decay and the time and distance separating each ratio's corresponding pair of SNPs (Figure 2). The initial slope of this empirical relationship is the estimated recombination rate $\hat{\rho}$. This approach, which we call Recombination Analysis via Time Series Linkage Decay (or RATS-LD), is a generalization of existing methods for recombination rate estimation [29] which permits application in short read sequencing data.

Validating RATS-LD on simulated data

We first simulated evolving intra-host viral populations with known recombination rates and tested that RATS-LD could recover these known rates. To create input for RATS-LD, we sampled these simulations longitudinally and mimicked the short-read, paired-end sampling common in viral deep sequencing experiments. We centered our simulated recombination rates around existing HIV intra-host estimated recombination rates ($\rho = 2 \times 10^{-6}$ - 10^{-3} [6, 29, 46], see Materials & Methods for full simulation and sampling details). We observed that the relationship between the D' ratio and the time-scaled physical distance ($d\Delta t$) is not linear, but asymptotically approaches a given D' ratio at large $d\Delta t$ values. This relationship emerges because, at long time-scaled distances, the genetic association between two SNPs is dominated not by decay from the initial measured linkage, but by sampling noise and other ongoing evolutionary processes. The rate at which this asymptote is approached depends strongly on the recombination rate (Figure 3A).

We therefore modified our approach for estimating ρ by fitting a function of the form

$$\hat{D}'_{ratio}(d\Delta t) = c_0 + c_1(1 - e^{-c_2 d\Delta t}) \tag{3}$$

which can successfully capture this behavior. The derivative of $\hat{D}'_{ratio}$ with respect to $d\Delta t$ at 0 recaptures the true recombination rate in simulated data from both neutrally-evolving populations (Figure 3B) and populations under strong selection (Figure S1, Materials & Methods). We noted that estimation at high or low values of ρ was biased downward or upward, respectively.

This minor bias emerged not because the D' ratio and the time-scaled distance scale unexpectedly, but due to challenges fitting D' ratio curves with paired-end data. To demonstrate that the slope of the D' ratio at 0 does indeed recapture the true recombination rate, we fit linear relationships between $d\Delta t$ and the D' ratio for subsets of the data near $d\Delta t = 0$ (i.e., pre-curve saturation). We found that these linear fits accurately recaptured the true recombination rate when only pre-saturation data was included in the fit (Figures S2 & S3). However, the D' ratio curves saturate at different $d\Delta t$ for different values of ρ, rendering these linear fits unsuitable when the true recombination rate is unknown (Figures 3A & S3). In contrast, the curve-fitting approach can be applied without a known $d\Delta t$ saturation point. For low recombination rates, the linkage decay curves do not saturate in the time-scaled distances available from the relatively short simulated reads (all pairs of SNPs less than 700 bp apart) (Figure 3A). As a result, RATS-LD underestimates the asymptote of Equation 3, which

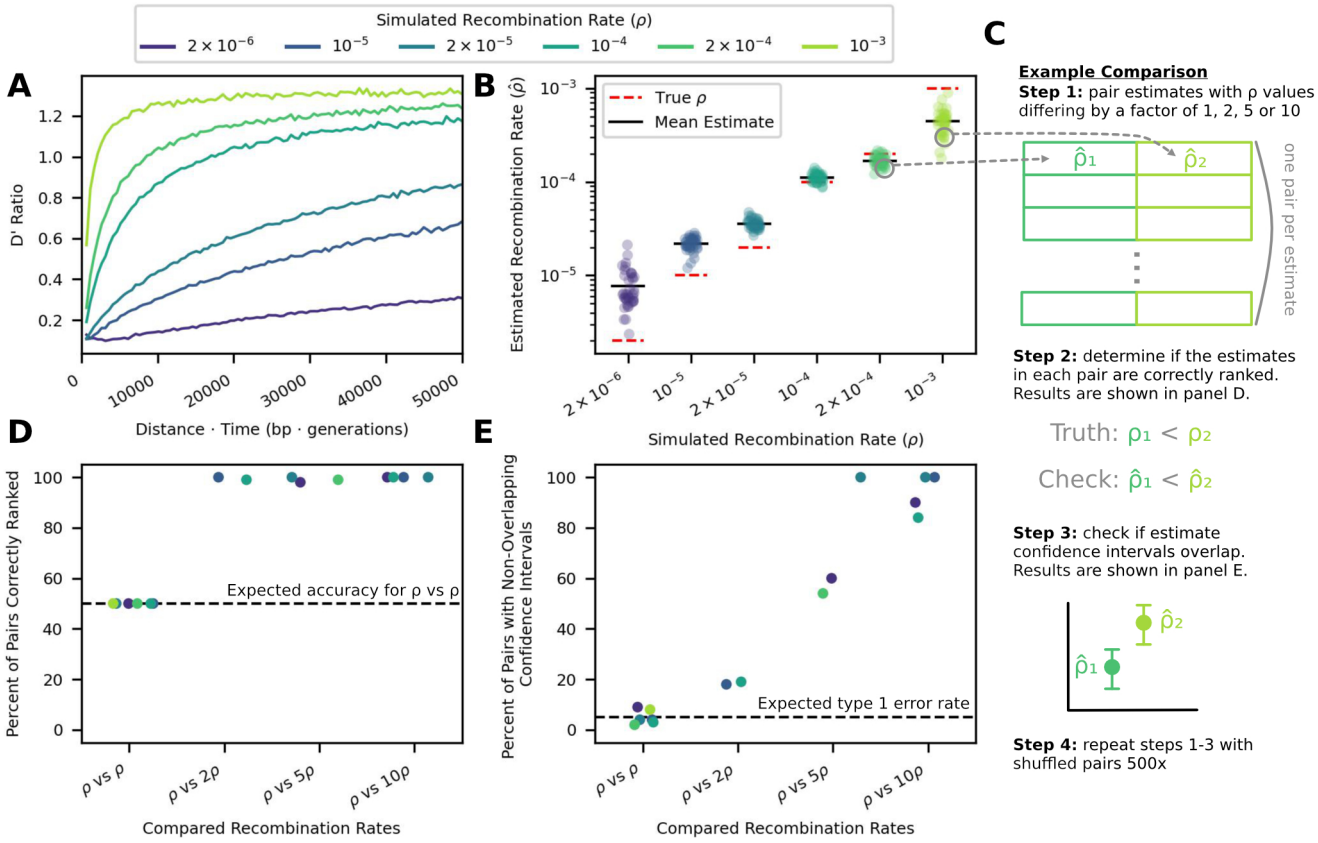

Figure 3: **RATS-LD recaptures true recombination rates in simulated data and accurately distinguishes between different rates.** **(A)** D' ratio curves calculated from simulated data exhibit saturation patterns with respect to $d\Delta t$ that depend on the underlying recombination rate, ρ (moving average displayed, window size = 500 bp · generations). **(B)** The slopes of these curves (D' ratio as a function of $d\Delta t$) at $d\Delta t = 0$ serve as RATS-LD's estimates and recapture the known ρ value in simulated data. Each point represents the median of a bootstrapped distribution of estimates containing 1000 bootstraps. **(C)** We performed a discrimination analysis to confirm that RATS-LD can distinguish different recombination rates. We simulated data with set recombination rates $(2 \times 10^{-6}, 10^{-5}, 2 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4}, 10^{-3})$ and estimated $\hat{\rho}$ using RATS-LD. We then randomly paired simulations with recombination rates that differed by a factor of 1, 2, 5, or 10 to assess how often RATS-LD correctly distinguished underlying recombination rates via simple rank order **(D)** and non-overlapping bootstrapped 95% confidence intervals **(E)**.

results in a slight overestimation of ρ (Figure 3B). Among very high recombination rates, little data is available before curve saturation, resulting in challenges fitting Equation 3's behavior near $d\Delta t = 0$ and underestimation of ρ. For full details on how these fits were performed, see Materials & Methods.

Despite the challenges of quantifying the recombination rate at more extreme ρ values, we found that RATS-LD could readily discriminate between differences in recombination rates in simulated data. We assessed this in two ways: first, we compared RATS-LD estimates ($\hat{\rho}$ values) across pairs of simulations and checked if the simulation with the higher true ρ also had the higher estimated $\hat{\rho}$ (Figures 3C & 3D). We found that pairs of simulations with a 2-fold or greater difference in ρ were almost always ordered correctly using RATS-LD estimates, regardless of their recombination rates (Figures 3C & 3E). Second, for each pair of simulations, we checked whether their bootstrapped $\hat{\rho}$ confidence intervals overlapped. We found that we had a low probability of type I errors (i.e., assessing differences in recombination rates when no such differences existed), and that RATS-LD could nearly always discriminate simulations with true ρ values a factor of 10 apart (Figure 3E). RATS-LD could also often discriminate ρ values a factor of 2 or 5 apart (Figure 3E).

Applying RATS-LD to *in vivo* HIV data

Thus established, we sought to use RATS-LD to quantify how recombination rate varied within and across intra-host HIV populations with different viral loads. We analyzed viral sequences from a cohort of 10 people living with HIV, who were sampled longitudinally over 6-12 years and were not on antiretroviral therapy (ART) [53], using a workflow illustrated in Figure S4. We chose this data set because the authors recorded contemporaneous viral load measurements and sequenced these data via a workflow that minimized PCR recombination (see Discussion) [52]. Therefore, we expect the detected signals of recombination in these data to be driven by *in vivo* processes, rather than amplification artifacts. RATS-LD estimated the average recombination rate across this data set to be 3.6×10^{-5} events/bp/generation (95% CI $2.8 \times 10^{-5} - 4.7 \times 10^{-5}$) which is slightly higher than previous estimates [6, 29, 46], but consistent under RATS-LD's slight upward bias in this range.

Within this data set, viral loads varied within and between individuals (Figure S5), as has been observed in other cohorts [37, 45]. For each pair of time points within an individual, we computed the mean viral load. Note, time points did not need to be consecutive to be compared, but we excluded any time point pairs where an intervening viral load measurement did not fall between the measurements at the two compared time points (see Materials & Methods). After labeling each D' ratio with the mean viral load between the two time points (t and $t + \Delta t$), we split the data based on the viral load tertiles which divided the data into three groups with approximately equal numbers of D' ratios. In Figure 4A, we show the viral load time courses for each participant with shading corresponding to the viral load tertiles.

We used RATS-LD to compute $\hat{\rho}$ separately for each of the viral load groups (Figure 4B-C). We found that $\hat{\rho}$ among time point pairs below the first tertile ($< 26,800$ copies/mL) was 1.5×10^{-5} recombination events/bp/generation (bootstrapped confidence intervals $7 \times 10^{-6} - 2.9 \times 10^{-5}$), a rate nearly identical to existing recombination rate estimates for HIV [29]. However, among pairs of time points with the highest average viral loads (>82,000 copies/mL), the estimated median recombination rate was nearly 6 times higher: 8.9×10^{-5} recombination events/bp/generation (bootstrapped confidence intervals $6.5 \times 10^{-5} - 1.2 \times 10^{-4}$). The middle tertile fell between the other two, although substantially closer to the lower tertile (bootstrapped confidence intervals $2 \times 10^{-5} - 4 \times 10^{-5}$). We also demonstrated that these differences were not solely driven by the data of any individual participant by conducting a leave one out analysis (Supplemental Figure S6).

To further quantify the relationship between viral load and estimated recombination rate, we also conducted an analysis in which we grouped D' ratios depending on their average viral load meeting a

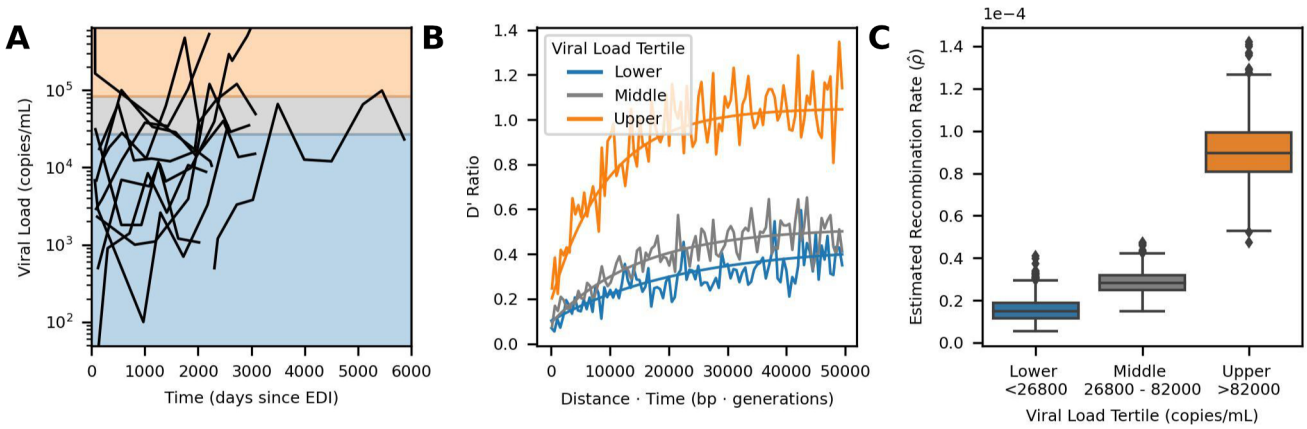

Figure 4: **Elevated recombination rates are associated with high viral load *in vivo*.** **(A)** Viral load time courses for each participant are plotted. The background shading indicates the viral load tertiles: $< 26,800$ (blue), $26,800 - 82,000$ (gray), and $> 82,000$ (orange) which divide the D' ratios into three approximately equal sized groups ($\approx$ 20k D' ratios per group). While viral loads at single time points are shown here, pairs of time points were used to form the D' ratio groups and were sorted based on their average viral load (see Figure S5). **(B)** Curve fits used by RATS-LD for estimation. Curve fits representing the median estimates of the bootstrap distribution (1000 bootstraps) for each group are overlaid on the binned average of the D' ratios from the *in vivo* data (moving average used for display purposes only, window size = 500 bp · generations). **(C)** Recombination rate estimates, $\hat{\rho}$, are plotted as a bootstrapped distribution (1000 bootstraps). Each box represents the quartiles of the $\hat{\rho}$ distribution and the median is indicated by the central line. Estimates differ between groups as viral load increases.

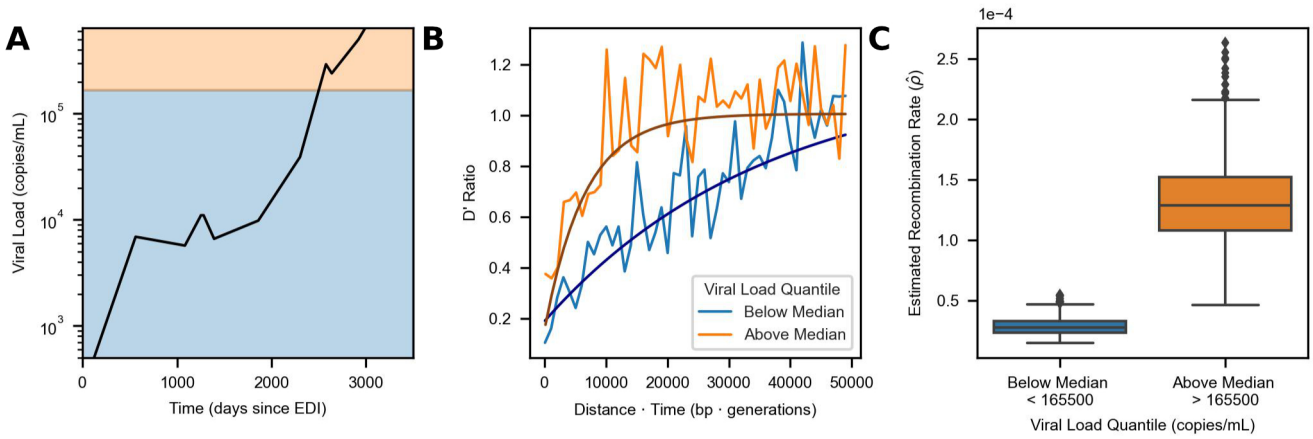

Figure 5: **High viral load is also associated with elevated recombination rate in Participant 1.** **(A)** Viral load trajectory over the number of days since estimated date of infection (EDI) for Participant 1. The background shading indicates the two groups (above median in orange and below median in blue) that pairs of time points were divided into based on their average viral load. Each group contains approximately $6,000$ D' ratios. **(B)** Curve fits representing the median estimates from each group's corresponding bootstrap distribution (1000 bootsraps) are overlaid on the binned average of the D' ratios in the *in vivo* data for Participant 1 (moving average used for display purposes only, window size = 1000 bp · generations). **(C)** Recombination rate estimates for times with average viral load measurements above or below the median using only data from Participant 1. Each box indicates the interquartile range of the bootstrapped distribution of $\hat{\rho}$ for the given viral load group (1000 Bootstraps). The median estimate is indicated via the central line. Identical analyses for Participants 3, 4, and 7 are shown in Figure S9.

given threshold (Figures S7 & S8). This permitted us to further determine if viral load has a dose-dependent effect on recombination rate. The estimated recombination rate became more extreme at higher viral load thresholds, and time points with a mean viral load above 200,000 copies/mL showed an estimated $\hat{\rho}$ nearly an order of magnitude higher than prevailing estimates (1.1×10^{-4}, confidence interval $7.5 \times 10^{-5} - 1.6 \times 10^{-4}$). Note, however, small sample sizes led to substantial variation in curve fitting and correspondingly large bootstrapped confidence intervals at more extreme thresholds. Taken together, these results suggest that HIV populations at high viral loads may be recombining significantly more often than previous estimates.

The analyses above suggest that higher viral loads are correlated with faster recombination across individuals, but viral loads can also vary significantly within individuals over time. We therefore set out to examine if RATS-LD could detect differences in recombination rates within single individuals. While most individuals did not have enough genetic variation to probe differences in recombination rate over time, Participant 1 had extensive genetic variation ($\approx$ 12,000 D' ratios) as well as a broad range of viral load measurements ($< 10^3$ to $> 10^6$ copies/mL). We divided the D' ratios for Participant 1 into two equally sized groups based on their viral loads and found a significant difference in recombination rates between time points with viral loads above and below the median of 165,500 copies/mL ($\hat{\rho}_{<165k} = 2.8 \times 10^{-5}$, confidence interval = 1.7×10^{-5} - 4.5×10^{-5}. $\hat{\rho}_{>165k} = 1.3 \times 10^{-4}$, confidence interval = 7.4×10^{-5} - 2×10^{-4}). We repeated this analysis with Participants 3, 4, and 7, who were the other best candidates due to the number of D' ratios available for estimation ($\approx 23,000$ ratios for Participant 7 and $\approx 8,000$ ratios for Participants 3 and 4). Consistent with the results from Participant 1, time points with viral loads below the median had lower estimated recombination rates than those with viral loads above the median (Figure S9). Although the bootstrapped confidence intervals between the high

and low viral load groups overlapped, these results are not unexpected as Participants 3, 4, and 7 did not have as broad of a range of viral load measurements as Participant 1 and RATS-LD is not well powered to detect differences in recombination rate among sample sizes this small. Nevertheless, the analyses within individuals suggest that as a population's viral load varies over time, evolutionarily important parameters such as the recombination rate can vary concurrently.

Discussion

HIV evolution within a host is a complex process driven by shifting selection pressures and variation in mutational availability as viral population sizes expand and contract over orders of magnitude. Here, we show that the recombination rate of HIV *in vivo* can also change dynamically over the course of infection in concert with changes in the viral load and effective population density. We introduce a method, RATS-LD, which estimates significantly elevated recombination rates at time points with the highest viral loads. Within an individual with both higher and lower viral load time points, we detect nearly an order of magnitude difference in recombination rates within the same viral population across time. Note, we observe this effect despite later stage infections having lower sequence homology, which might have been expected to suppress recombination [4, 5]. Although HIV recombination has long been appreciated to be a common event, these results suggest that in certain intra-host contexts, rates of recombination can be even more extreme than previously recognized [6, 29, 46]. While this effect has previously been demonstrated in cell culture where viral density can be experimentally manipulated [23], this represents the first evidence that intra-host viral density varies extensively enough to cause substantial changes to the recombination rate *in vivo* in human hosts.

This previously unappreciated recombination rate variation likely impacts both intra-host evolutionary dynamics and also interpretations of intra-host viral data. While most of the sequencing data we investigated were not from acute infection, the initial HIV expansion brings the viral load its most extreme values [44]. If elevated recombination rates broadly accompany high viral loads, recombination may drive initial population diversification to a previously unrecognized extent. The importance of this effect may be counteracted by relatively low population diversity in the early stages of infection [53]. This interaction merits future investigation. Recent work has also suggested that elevated recombination rates could mask signatures of spatial compartmentalization in intra-host HIV populations [42]. In contrast, low viral loads in individuals on incompletely suppressive ART may lead to correspondingly low recombination rates, mitigating recombination of distinct drug resistance mutations on different genetic backgrounds as a path to multi-drug resistance (speculated in [11]). Lower rates of recombination in lower viral load populations may also make phylogenetic approaches that rely on treelike evolutionary divergence more tractable to apply without requiring ancestral recombination graphs [7]. While we did not have access to samples with extremely low viral loads near the limit of detection, future investigations could determine if previously reported average recombination rates ($\approx 1.4 \times 10^{-5}$) can explain low viral load patterns of linkage decay.

Like other *in vivo* methods, RATS-LD relies on naturally segregating sites as genetic markers to track the breakdown of linkage. While the rate of this breakdown is not affected by population diversity, our ability to accurately quantify its average does depend on the number of segregating sites we can identify. Therefore, having more or fewer mutations stands only to impact the confidence of RATS-LD rather than bias the estimated recombination rates themselves. Less diverse populations require more pooling to have the requisite number of pairs of data points. This limits our ability to powerfully estimate recombination rates in smaller, less diverse populations.

Importantly, we cannot detect recombination events between identical sequences. If genetically identical viral bursts are largely confined to a given spatial location, rates of coinfection and recombination

may be high but undetectable to this method. Similar effects could be driven by cell-to-cell transmission [22]. While important for understanding the viral life history and phenomena like complementation [14] and replication fidelity [21, 35], these undetectable recombination events are unlikely to contribute meaningfully to the population's evolution because they do not generate new viral haplotypes. As such, our estimated $\hat{\rho}$ values reflect the effective rates of recombination and mixing discounting population structure.

Although we analyzed data with careful controls for PCR recombination, we cannot completely rule it out as a complicating factor [40]. In order to minimize PCR recombination, Zanini et al., 2015 only used one round of PCR (rather than a more standard nested approach) and optimized polymerases and extension times, among other measures [53, 52]. They validated this approach by comparing linkage decay in a 50/50 mix of two laboratory strains. While no linkage decay was observed in this mixture, the reduced homology between the two strains may also have discouraged recombination [4, 5]. As further validation, they analyzed a natural control of an intra-host population founded by two distinct viruses. Despite the distinct founder viruses having high homology, nearly complete linkage was preserved after a single round of PCR amplification. Additionally, these control samples were from relatively high viral load time points ($\approx 10^5$ copies/mL) indicating that PCR recombination is extremely low even in the high viral load samples we analyzed. While we cannot definitively say that PCR recombination did not in any way affect our results, these controls provide reassurance that rates of *in vitro* PCR recombination were extremely low.

It is important to note several additional caveats with our analysis: 1) Correlation does not imply causation. Although we report an association between recombination rate and viral load, it is possible that rapid recombination potentiates a higher viral load, and not that higher viral loads cause more coinfection and subsequent recombination. It is also possible that a third underlying factor causes both higher viral load and more frequent recombination. While we cannot directly disentangle causation in this study, the association still suggests that we must consider how recombination rates vary within and between individuals. 2) We analyzed a relatively small cohort of untreated participants, many of whom showed some degree of virologic control [53]. The same features that may permit these individuals to control the virus may also affect recombination processes within these viral populations. However, we note that individuals with low viral loads showed recombination rates consistent with other estimates in the literature [29, 6], while time points with less control exhibited elevated recombination rates more consistent with *in vitro* studies [23, 43].

While RATS-LD is a broadly useful technique for quantifying rapid recombination from time series data, the sequencing strategy (especially the read length) limits the recombination rates that can be effectively inferred. Similar limitations have been noted in other work estimating linkage structure from pooled read data [12]. Low recombination rates may permit few or no recombination events to occur within the context of single reads. As a result, little breakdown of linkage will be observable, and longer read sequencing data might be necessary to capture these events. As these sequencing approaches proliferate [13], we expect the utility of linkage-based techniques for quantifying recombination and other evolutionary forces to expand, even in populations with lower recombination rates. For the purpose of HIV, RATS-LD is well suited to the current state of short read paired-end next generation sequencing data.

More broadly, the biological phenomenon that we report here - association between higher recombination rate and higher population density - could be much more widespread in facultatively asexual populations where physical proximity can mediate the rate of recombination. Potentially relevant cases include viral homologous recombination (as shown here) and reassortment, but also rates of plasmid conjugation in bacteria, or sexual reproduction in hermaphroditic organisms or plants that do not self or outcross exclusively. Explicit attention to population density in mediating recombination rates stands to impact our understanding of species diversification in viruses and more broadly.

Acknowledgements

The authors would like to thank Richard Neher for assistance accessing data from Participant 6, in addition to Kelley Harris, Benjamin Kerr, Daniel Reeves, Lillian Cohn, Elena Giorgi, Nandita Garud, the Cohn lab at Fred Hutchinson Cancer Center and the Feder lab at the University of Washington for useful discussions on this project and feedback on the manuscript. EVR was supported by the UW Statistical Genetics Training Grant.

Materials & Methods

Measuring Linkage Over Time

RATS-LD quantifies the linkage between pairs of segregating loci by using the D' statistic [24] (4).

$$D' = \frac{|p_{12} - p_1 p_2|}{D_{max}} \tag{4}$$

where p_1 and p_2 are the frequencies of the majority nucleotide at the corresponding locus, p_{12} is the frequency of the haplotype made up by the majority nucleotides at the two loci, and D_{max} is given by

$$D_{max} = \begin{cases} \min(p_1 p_2, (1-p_1)(1-p_2)) & \text{if } p_{12} < p_1 p_2 \\ \min(p_1(1-p_2), (1-p_1)p_2) & \text{if } p_{12} > p_1 p_2 \end{cases}.$$

Estimation of ρ can also be performed using D rather than D', but we found D' was less affected by noise (Figure S10). Note, because linkage information is necessary to compute D', pairs of SNPs must be close enough to be spanned by a sequencing read.

RATS-LD then computes the decay of linkage between those pairs of segregating loci at a later time point. For a given set of time points, D' is expected to decay according to the following equation [16]:

$$D'(t + \Delta t) = D'(t) e^{-\rho d \Delta t} \tag{1}$$

where d is the distance between the loci in nucleotides and Δt is the time between the two time points in generations (assuming that D_{max} is similar over time). For the *in vivo* analysis, we converted the number of days between samples to generations by assuming an approximate rate of 1 generation per 2 days [34, 38]. We can rewrite (1) as

$$D'_{\text{ratio}} = -\log\left(\frac{D'(t + \Delta t)}{D'(t)}\right) = \rho d \Delta t. \tag{2}$$

Both the left hand side of (2), which we will denote as the D' ratio, and $d\Delta t$ can be measured from the data directly, providing the opportunity to estimate ρ. However, the decay of linkage over time between two pairs of SNPs is a noisy measurement, so we fit this relationship over many pairs of SNPs and many time points, as described in the next section.

Recombination Rate Estimation

After filtering the data as described below, we calculated D' ratios (2) for each pair of segregating loci at each pair of time points. Then we estimated recombination rates by fitting the decay of these D' ratios over the distance between loci and time between samples ($d\Delta t$).

Equation (2) suggests a linear relationship between the D' ratio and $d\Delta t$ with slope ρ. However, two factors complicate this expected pattern: 1) sampling error makes it difficult to estimate small differences between D' values accurately. As a result, a decrease of D' from 0.01 to 0.001 may be driven entirely by sampling error, whereas a decrease from 0.5 to 0.05 likely reflects true decay in linkage, although they result in equal D' ratios. At large values of $d\Delta t$, D' values are expected to become increasingly small, so the effect of sampling error becomes pronounced. 2) Over long periods of time, other evolutionary forces - selection, recurrent mutation - are more likely to affect linkage relationships between two loci. Therefore, we fit the functional form

$$f(x) = c_0 + c_1(1 - e^{-c_2 x}), \tag{5}$$

where D' ratio $= f(x)$ and $x = d\Delta t$. The slope of the curve as $d\Delta t \to 0$ should still be ρ since, in this range, there is minimal time for evolutionary processes to act and tight linkage leads to large D' values which are not as affected by sampling noise as detailed above. This yields a final estimate, given by the derivative of (5) at $d\Delta t = 0$ which is

$$\hat{\rho} = f'(x = 0) = c_1 c_2. \tag{6}$$

Thus, RATS-LD quantifies the recombination rate by fitting the relationship between linkage decay and time-scaled distance ($d\Delta t$) then calculating the derivative of the fit near $d\Delta t = 0$.

Curve Fitting

For all analyses except those shown in Figures S2 & S3, we fit the curve form (5) to the D' ratio data via non-linear least squares [9]. Prior to performing the fits, we restricted the data to pairs of SNPs with $d\Delta t < 5 \times 10^4$ bp $\cdot$ generations to balance RATS-LD's performance at both high and low values of ρ (Figure S11A).

To confirm that the slope of the D' ratio data near $d\Delta t = 0$ does capture the true recombination rate, we performed linear fits to the D' ratio data which are shown in Figures S2 & S3. Since the linear fits are only accurate when considering D' ratios prior to saturation, we performed these fits to data only including D' ratios with $d\Delta t < 500, 1000, 2000, 10000,$ or 50000 bp $\cdot$ generations so that at least one $d\Delta t$ fit threshold would capture the pre-saturation behavior for each value of ρ in the simulated data. We fit the slope m and intercept b of the relationship

$$f(x) = mx + b, \tag{7}$$

where D' ratio $= f(x)$ and $x = d\Delta t$, via linear regression [9]. The slope of the fit was the estimated $\hat{\rho}$.

Bootstrapped Confidence Intervals

To create bootstrapped confidence intervals for RATS-LD's recombination rate estimates, we sampled the segregating loci with replacement and repeated estimation using only the D' ratios where both loci were sampled. Each bootstrap sample included the same number of segregating loci as was initially used for estimation. The example fits displayed for the bootstrapped confidence intervals use the set of coefficients that provided the estimate at the given quantile (2.5%, 50%, 97.5%) of the $\hat{\rho}$ distribution (rather than the quantiles of the coefficient distributions themselves). The primary $\hat{\rho}$ reported for each analysis is the median of the bootstrapped distribution.

Simulated Data

To validate RATS-LD's ability to quantify recombination rates, we simulated populations with known recombination rates using SLiM [15]. To approximate intra-host viral conditions, we simulated a genomic region of length 700bp, with a mutation rate of $\mu = 10^{-6}$ mutations/bp/generation [2, 1, 41] and effective population size of $N_e = 10^4$ [2, 10]. The simulated recombination rates ρ included 2×10^{-6}, 10^{-5}, 2×10^{-5}, 10^{-4}, 2×10^{-4}, and 10^{-3} recombinations/bp/generation to broadly encapsulate the range of previously estimated HIV recombination rates [6, 29, 46]. Each simulation was run for 5×10^4 generations before samples from five evenly spaced time points were collected at intervals of 150 generations. Each sample contained $M = 800$ genomes to approximately match the median number of templates in each of the *in vivo* sequencing reactions [53]. Each simulation produced approximately $20,000 - 25,000$ D' ratios.

To match the sequencing coverage in the *in vivo* data, we created a procedure to downsample the simulations to mimic the pairwise coverage of two loci d nucleotides apart. To do this, we created an empirical distribution of reads overlapping a locus at an arbitrary position 0 (Figure S12A). First, we simulated the expected positions of 10^5 reads with start positions uniformly distributed between genomic positions -700 and 0 (in bp). For each start position, we uniformly sampled a total end to end read length between 400 and 700 bp, which was the observed read length distribution in Zanini et al., 2015. In addition, we only considered 300 bp of coverage extending inwards from each end of the read, to match the paired end sequencing study design [53, 52]. As a result, longer reads did not have coverage in their centers. We discarded any simulated reads that did not capture genomic position 0. At this point, we formed the empirical distribution by computing the number of reads that overlapped position 0 and a position d nucleotides away in either direction. The resulting empirical distribution had high coverage for nearby loci and decreasing coverage as d increased (Figure S12C). This paired-end read sampling technique mimicked the *in vivo* sampling scheme and provided coverage similar to the *in vivo* data ([53]) (Figure S12B). Then, for each pair of SNPs in the SLiM data at a distance of d nucleotides apart, we downsampled their coverage to match the empirical number of overlapping reads for two SNPs at a distance of d nucleotides.

To assess RATS-LD's performance in the presence of selection, we performed additional simulations in which all mutations had a selection coefficient drawn from an empirically-derived distribution of HIV intra-host fitness effects [49]. These selection simulations were used to produce Figure S1 while all other validation simulations included only neutral mutations. Under the empirically determined distribution of fitness effects, fewer sites segregated due to both positive and negative selection.

RATS-LD requires many pairs of SNPs across multiple time points to fit Equation 3. Therefore, some data pooling was required for estimation. To ensure all data of the validation analyses used equal numbers of D' ratios, we grouped the simulation runs and randomly downsampled segregating loci in each group. Each group of neutrally simulated data was comprised of 5 simulation runs prior to downsampling while the selection simulation runs were put in groups of 20, 408, and 686 at $\rho = 10^{-3}$, $\rho = 10^{-4}$ and $\rho = 10^{-5}$ respectively. We then downsampled each group by including a D' ratio in the downsampled data set if both of its loci had been randomly drawn. We continued drawing new segregating loci until the total number of D' ratios reached the desired data set size. Each point in Figures 3B, S1, S2, and S10 represents the application of RATS-LD to data from one of these downsampled data sets with $25,000$ loci each. To assess the limits of RATS-LD to estimate recombination rates with fewer segregating sites, we downsampled the number of D' ratios even further and found that at least 5000 D' ratios should be used for estimation, but fit accuracy is best when there are $\geq 2\times10^4$ D' ratios (Figure S11B).

In vivo Data

We used longitudinal HIV deep-sequencing data from Zanini et al. 2015 for *in vivo* recombination rate estimation [52, 53]. The samples were collected from the blood plasma of 11 participants who were diagnosed with HIV-1 in Sweden from 1990 – 2003 and were not on antiretroviral treatment. Six to twelve time points are included per participant with both sequencing and viral load information at most time points. The sequences span the full HIV genome and were generated using 6 overlapping amplicons, but we excluded fragment 5 (which contains *env*) in our analysis due to its lower rates of template recovery. The sequencing reactions used paired-end reads amplified via a protocol designed to minimize PCR recombination (see Discussion). In particular, the protocol used only a single round of PCR (rather than a nested approach). We used arrays of co-occurrences of alleles across pairs of loci (available at https://hiv.biozentrum.unibas.ch), which had undergone Zanini et al.'s quality assurance process for read mapping and filtering. This process included rigorous checks for base quality, mapping errors, and sample contamination. We also used the estimated dates of infection (EDI) as provided by Zanini et al., and each time point was measured relative to the EDI for the corresponding participant.

To assess if higher viral load is associated with a difference in recombination rate, we sorted pairs of time points into groups based on their average viral load. We started by matching genotype information from each sampling time point with a viral load measurement near the same time. For the majority of the data, the date of sample collection exactly matched the date of the viral load measurement. However, some samples from participants 4 and 7 did not have a concurrent viral load measurement and were therefore matched to viral load measurements within 100 days of the sample's collection. Additionally, data from Participant 10 was ommitted from the analysis due to nearly all of the time points lacking viral load measurements within 100 days. For each pair of time points, we calculated the average viral load by taking the mean of the viral load measurements matched to those two samples. For pairs of time points with intervening viral load measurements, we required that those viral loads stayed within the bounds of the viral loads measured at the starting and ending time points. Pairs of time points with an intervening measurement that violated these boundaries were excluded from the analysis to ensure that the average viral loads used to group the data are as accurate as possible.

For the quantile analyses, the D' ratios were labeled with the average viral loads across their corresponding time point pairs and divided based on average load to make two (Figures 5 & S9) or three (Figures 4 & S6) groups with approximately equal numbers of D' ratios. Since a single pair of time points could have many corresponding D' ratios, the size of the groups sometimes varied by ≈ 1000 ratios since all D' ratios per time point pair were included in the same group. For the threshold analysis shown in Figure S7 and Figure S8,the pairs of time points were placed into groups based on whether their average viral load was above or below a given threshold, and then we applied RATS-LD separately to the high and low viral load groups.

To ensure the pattern observed in Figure 4C was not solely driven by data from a single participant, we performed a leave-one-out analysis in which we repeated estimation excluding all data from one participant at a time.

Data Filtering

We started with data from all time points and all pairs of SNPs which were close enough to be spanned by sequencing reads. Then, we used several filtering steps in both the simulated and *in vivo* data sets to reduce noise. 1) We required that at least 10 reads span a pair of loci for that pair to be included in the analysis. 2) SNPs at each time point were included only if the minor allele had a frequency $> 1\%$ at that time point. We chose this cutoff since Zanini et al., 2015 report that SNPs' frequencies can be estimated down to 1% accuracy in the *in vivo* data set we analyzed [53, 52]. Any SNPs transiently under 1%

frequency were not included in the analysis at the corresponding time point, but were included at other time points where they reached the frequency threshold. 3) All four possible haplotype combinations of the major and minor alleles at a pair of loci had to be observed at a given time point to be included in the analysis at that time point. Filtering steps 2 and 3 prevented false linkage signals from very low frequency alleles appearing and disappearing between time points. 4) Because the investigation focused on the decay of linkage across pairs of time points, we required that $D' > 0.2$ at the first time point in each pair (or $D > 0.0075$ for the analysis using D) since decay in larger D' values is less likely to be a function of noise. RATS-LD's ability to discriminate between underlying recombination rates is robust to changes in this threshold. 5) Lastly, we excluded time point pairs that spanned less than 50 days, since the signal of linkage decay at such short timescales is also very susceptible to noise.

Discrimination Analysis

To validate that RATS-LD can discriminate between different recombination rates, we randomly paired simulations with recombination rates that were a factor of 1, 2, 5, or 10 apart. For each value of ρ we used 40 data sets of $\approx 25,000$ neutrally evolving loci which were simulated in SLiM as described above. We applied RATS-LD to each simulated data set to produce a $\hat{\rho}$ estimate given by the median of its corresponding bootstrapped distribution (1000 bootstraps). To compare two given ρ values, we randomly paired these estimates to create 40 matched pairs. For each matched pair with true recombination rates ρ_1 and ρ_2, we computed two comparison metrics. 1) Simple rank order: we determined whether the individual estimates $\hat{\rho}_1$ and $\hat{\rho}_2$ were correctly ordered such that $\hat{\rho}_1 > \hat{\rho}_2$ if $\rho_1 > \rho_2$. If $\rho_1 = \rho_2$, we checked agreement against an arbitrary ordering. 2) Bootstrapped significance: we determined whether the bootstrapped $\hat{\rho}$ distributions from the data sets overlapped. We reshuffled the pairings and recomputed the two metrics 500 times.

Data availability

All data used in this study (viral loads and paired participant read counts from [53]) can be downloaded from `https://hiv.biozentrum.unibas.ch/`.

Code availability

All code to reproduce these analyses can be found at `https://github.com/federlab/hiv-recombination`.

References

[1] M. E. Abram, A. L. Ferris, W. Shao, W. G. Alvord, and S. H. Hughes. Nature, Position, and Frequency of Mutations Made in a Single Cycle of HIV-1 Replication. *Journal of Virology*, 84(19):9864–9878, 2010.

[2] G. Achaz, S. Palmer, M. Kearney, F. Maldarelli, J. W. Mellors, J. M. Coffin, and J. Wakeley. A Robust Measure of HIV-1 Population Turnover Within Chronically Infected Individuals. *Molecular Biology and Evolution*, 21(10):1902–1912, 06 2004.

[3] W. An and A. Telesnitsky. Effects of varying sequence similarity on the frequency of repeat deletion during reverse transcription of a human immunodeficiency virus type 1 vector. *J Virol*, 76(15):7897–7902, Aug 2002.

[4] M. Balakrishnan, P. J. Fay, and R. A. Bambara. The Kissing Hairpin Sequence Promotes Recombination within the HIV-I 5 Leader Region*. *Journal of Biological Chemistry*, 276(39):36482–36492, 2001.

[5] M. Balakrishnan, B. P. Roques, P. J. Fay, and R. A. Bambara. Template dimerization promotes an acceptor invasion-induced transfer mechanism during human immunodeficiency virus type 1 minus-strand synthesis. *Journal of Virology*, 77(8):4710–4721, 2003.

[6] R. Batorsky, M. F. Kearney, S. E. Palmer, F. Maldarelli, I. M. Rouzine, and J. M. Coffin. Estimate of effective recombination rate and average selection coefficient for HIV in chronic infection. *Proceedings of the National Academy of Sciences*, 108(14):5661–5666, 2011.

[7] L. A. Castro, T. Leitner, and E. Romero-Severson. Recombination smooths the time-signal disrupted by latency in within-host HIV phylogenies. *bioRxiv[Preprint]*, 2022.

[8] J. Chen, O. Nikolaitchik, J. Singh, A. Wright, C. E. Bencsics, J. M. Coffin, N. Ni, S. Lockett, V. K. Pathak, and W.-S. Hu. High efficiency of HIV-1 genomic RNA packaging and heterozygote formation revealed by single virion analysis. *Proceedings of the National Academy of Sciences*, 106(32):13535–13540, 2009.

[9] V. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[10] A. F. Feder, C. Kline, P. Polacino, M. Cottrell, A. D. M. Kashuba, B. F. Keele, S.-L. Hu, D. A. Petrov, P. S. Pennings, and Z. Ambrose. A spatio-temporal assessment of simian/human immunodeficiency virus (SHIV) evolution reveals a highly dynamic process within the host. *PLoS Pathog*, 13(5):e1006358, May 2017.

[11] A. F. Feder, P. S. Pennings, and D. A. Petrov. The clarifying role of time series data in the population genetics of HIV. *PLOS Genetics*, 17(1):1–10, 01 2021.

[12] A. F. Feder, D. A. Petrov, and A. O. Bergland. LDx: estimation of linkage disequilibrium from high-throughput pooled resequencing data. *PloS one*, 7(11):e48588, 2012.

[13] C. M. Gallardo, S. Wang, D. J. Montiel-Garcia, S. J. Little, D. M. Smith, A. L. Routh, and B. E. Torbett. MrHAMER yields highly accurate single molecule viral sequences enabling analysis of intra-host evolution. *Nucleic acids research*, 49(12):e70–e70, 2021.

[14] H. C. Gelderblom, D. N. Vatakis, S. A. Burke, S. D. Lawrie, G. C. Bristol, and D. N. Levy. Viral complementation allows HIV-1 replication without integration. *Retrovirology*, 5(1):1–18, 2008.

[15] B. C. Haller and P. W. Messer. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Molecular Biology and Evolution*, 36(3):632–637, 01 2019.

[16] D. L. Hartl, A. G. Clark, and A. G. Clark. *Principles of population genetics*, volume 116. Sinauer associates Sunderland, 1997.

[17] W.-S. Hu and S. H. Hughes. HIV-1 reverse transcription. *Cold Spring Harbor perspectives in medicine*, 2(10), Oct. 2012.

[18] W. S. Hu and H. M. Temin. Genetic consequences of packaging two rna genomes in one retroviral particle: pseudodiploidy and high rate of genetic recombination. *Proc Natl Acad Sci U S A*, 87(4):1556–1560, Feb 1990.

[19] A. Jung, R. Maier, J.-P. Vartanian, G. Bocharov, V. Jung, U. Fischer, E. Meese, S. Wain-Hobson, and A. Meyerhans. Multiply infected spleen cells in HIV patients. *Nature*, 418(6894):144–144, July 2002.

[20] P. Kellam and B. A. Larder. Retroviral recombination can lead to linkage of reverse transcriptase mutations that confer increased zidovudine resistance. *J Virol*, 69(2):669–674, Feb 1995.

[21] King Steven R., Duggal Nisha K., Ndongmo Clement B., Pacut Crystal, and Telesnitsky Alice. Pseudodiploid Genome Organization Aids Full-Length Human Immunodeficiency Virus Type 1 DNA Synthesis. *Journal of Virology*, 82(5):2376–2384, Mar. 2008. Publisher: American Society for Microbiology.

[22] J. Kreger, J. Garcia, H. Zhang, N. L. Komarova, D. Wodarz, and D. N. Levy. Quantifying the dynamics of viral recombination during free virus and cell-to-cell transmission in HIV-1 infection. *Virus Evolution*, 7(1), 03 2021.

[23] D. N. Levy, G. M. Aldrovandi, O. Kutsch, and G. M. Shaw. Dynamics of HIV-1 recombination in its natural target cells. *Proceedings of the National Academy of Sciences*, 101(12):4204–4209, 2004.

[24] R. C. Lewontin. On measures of gametic disequilibrium. *Genetics*, 120(3):849–852, 11 1988.

[25] D. P. Martin, A. Varsani, P. Roumagnac, G. Botha, S. Maslamoney, T. Schwab, Z. Kelz, V. Kumar, and B. Murrell. RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evolution*, 7(1), 04 2020. veaa087.

[26] J. Maydt and T. Lengauer. Recco: recombination analysis using cost optimization. *Bioinformatics*, 22(9):1064–1071, 02 2006.

[27] A. Moumen, L. Polomack, B. Roques, H. Buc, and M. Negroni. The HIV-1 repeated sequence R as a robust hot-spot for copy-choice recombination. *Nucleic Acids Research*, 29(18):3814–3821, Sept. 2001.

[28] L. Moutouh, J. Corbeil, and D. D. Richman. Recombination leads to the rapid emergence of HIV-1 dually resistant mutants under selective drug pressure. *Proc Natl Acad Sci U S A*, 93(12):6106–6111, Jun 1996.

[29] R. A. Neher and T. Leitner. Recombination rate and selection strength in HIV intra-patient evolution. *PLoS computational biology*, 6(1):e1000660, 2010.

[30] O. A. Nikolaitchik, A. Galli, M. D. Moore, V. K. Pathak, and W.-S. Hu. Multiple barriers to recombination between divergent HIV-1 variants revealed by a dual-marker recombination assay. *J Mol Biol*, 407(4):521–531, Apr 2011.

[31] T. Nora, C. Charpentier, O. Tenaillon, C. Hoede, F. Clavel, and A. J. Hance. Contribution of recombination to the evolution of human immunodeficiency viruses expressing resistance to antiretroviral treatment. *J Virol*, 81(14):7620–7628, Jul 2007.

[32] A. Onafuwa-Nuga and A. Telesnitsky. The remarkable frequency of human immunodeficiency virus type 1 genetic recombination. *Microbiology and Molecular Biology Reviews*, 73(3):451–480, 2009.

[33] A. T. Panganiban and D. Fiore. Ordered Interstrand and Intrastrand DNA Transfer During Reverse Transcription. *Science*, 241(4869):1064–1069, Aug. 1988.

[34] A. S. Perelson, A. U. Neumann, M. Markowitz, J. M. Leonard, and D. D. Ho. Hiv-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science*, 271(5255):1582–1586, Mar 1996.

[35] J. M. Rawson, O. A. Nikolaitchik, B. F. Keele, V. K. Pathak, and W.-S. Hu. Recombination is required for efficient HIV-1 replication and the maintenance of viral genome integrity. *Nucleic acids research*, 46(20):10535–10545, 2018.

[36] A. J. Ritchie, F. Cai, N. M. Smith, S. Chen, H. Song, S. Brackenridge, S. S. Abdool Karim, B. T. Korber, A. J. McMichael, F. Gao, and N. Goonetilleke. Recombination-mediated escape from primary CD8+ T cells in acute HIV-1 infection. *Retrovirology*, 11(1):69, Sept. 2014.

[37] M. L. Robb, L. A. Eller, H. Kibuuka, K. Rono, L. Maganga, S. Nitayaphan, E. Kroon, F. K. Sawe, S. Sinei, S. Sriplienchan, et al. Prospective study of acute HIV-1 infection in adults in east africa and thailand. *New England Journal of Medicine*, 374(22):2120–2130, 2016.

[38] A. G. Rodrigo, E. G. Shpaer, E. L. Delwart, A. K. N. Iversen, M. V. Gallo, J. Brojatsch, M. S. Hirsch, B. D. Walker, and J. I. Mullins. Coalescent estimates of HIV-1 generation time in vivo. *Proceedings of the National Academy of Sciences*, 96(5):2187–2191, Mar. 1999.

[39] J. F. Salazar-Gonzalez, E. Bailes, K. T. Pham, M. G. Salazar, M. B. Guffey, B. F. Keele, C. A. Derdeyn, P. Farmer, E. Hunter, S. Allen, O. Manigart, J. Mulenga, J. A. Anderson, R. Swanstrom, B. F. Haynes, G. S. Athreya, B. T. M. Korber, P. M. Sharp, G. M. Shaw, and B. H. Hahn. Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. *J Virol*, 82(8):3952–3970, Apr 2008.

[40] J. F. Salazar-Gonzalez, E. Bailes, K. T. Pham, M. G. Salazar, M. B. Guffey, B. F. Keele, C. A. Derdeyn, P. Farmer, E. Hunter, S. Allen, O. Manigart, J. Mulenga, J. A. Anderson, R. Swanstrom, B. F. Haynes, G. S. Athreya, B. T. M. Korber, P. M. Sharp, G. M. Shaw, and B. H. Hahn. Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. *J Virol*, 82(8):3952–3970, Apr 2008.

[41] R. Sanjuán, M. R. Nebot, N. Chirico, L. M. Mansky, and R. Belshaw. Viral mutation rates. *Journal of Virology*, 84(19):9733–9748, 2010.

[42] S. Sarkar, E. Romero-Severson, and T. Leitner. Migration coupled with recombination explains disparate HIV-1 anatomical compartmentalization signals. *bioRxiv*, 2023.

[43] T. E. Schlub, R. P. Smyth, A. J. Grimm, J. Mak, and M. P. Davenport. Accurately measuring recombination between closely related HIV-1 genomes. *PLoS computational biology*, 6(4):e1000766, 2010.

[44] R. M. Selik and L. Linley. Viral loads within 6 weeks after diagnosis of HIV infection in early and later stages: Observational study using national surveillance data. *JMIR Public Health Surveill*, 4(4):e10770, Nov 2018.

[45] R. Shankarappa, J. B. Margolick, S. J. Gange, A. G. Rodrigo, D. Upchurch, H. Farzadegan, P. Gupta, C. R. Rinaldo, G. H. Learn, X. He, et al. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *Journal of virology*, 73(12):10489–10502, 1999.

[46] D. Shriner, A. G. Rodrigo, D. C. Nickle, and J. I. Mullins. Pervasive genomic recombination of HIV-1 in vivo. *Genetics*, 167(4):1573–1583, 2004.

[47] A. C. Siepel, A. L. Halpern, C. Macken, and B. T. Korber. A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. *AIDS Research and Human Retroviruses*, 11(11):1413–1416, 1995. PMID: 8573400.

[48] R. P. Smyth, T. E. Schlub, A. J. Grimm, C. Waugh, P. Ellenberg, A. Chopra, S. Mallal, D. Cromer, J. Mak, and M. P. Davenport. Identifying recombination hot spots in the HIV-1 genome. *Journal of virology*, 88(5):2891–2902, 2014.

[49] M. S. Sohail, R. H. Y. Louie, M. R. McKay, and J. P. Barton. MPL resolves genetic linkage in fitness inference from complex evolutionary histories. *Nature Biotechnology*, 39(4):472–479, Apr. 2021.

[50] H. Song, E. E. Giorgi, V. V. Ganusov, F. Cai, G. Athreya, H. Yoon, O. Carja, B. Hora, P. Hraber, E. Romero-Severson, et al. Tracking HIV-1 recombination to resolve its contribution to HIV-1 evolution in natural infection. *Nature communications*, 9(1):1–15, 2018.

[51] K. Y. Taylor, I. Agu, I. José, S. Mäntynen, A. J. Campbell, C. Mattson, T.-W. Chou, B. Zhou, D. Gresham, E. Ghedin, and S. L. Díaz Muñoz. Influenza a virus reassortment is strain dependent. *PLoS Pathog*, 19(3):e1011155, Mar 2023.

[52] F. Zanini, J. Brodin, J. Albert, and R. A. Neher. Error rates, PCR recombination, and sampling depth in HIV-1 whole genome deep sequencing. *Virus Research*, 239:106–114, 2017. Deep sequencing in virology.

[53] F. Zanini, J. Brodin, L. Thebo, C. Lanz, G. Bratt, J. Albert, and R. A. Neher. Population genomics of intrapatient hiv-1 evolution. *Elife*, 4:e11282, 2015.