HIV Evolutionary Dynamics Within and Among Hosts

Philippe Lemey1,2, Andrew Rambaut1 and Oliver G. Pybus1
1Department of Zoology, University of Oxford, Oxford, UK; 2Rega Institute, Katholieke Universiteit Leuven, Leuven, Belgium

Abstract

The HIV evolutionary processes continuously unfold, leaving a measurable footprint in viral gene sequences. A variety of statistical models and inference techniques have been developed to reconstruct the HIV evolutionary history and to investigate the population genetic processes that shape viral diversity. Remarkably different population genetic forces are at work within and among hosts. Population-level HIV phylogenies are mainly shaped by selectively neutral epidemiologic processes, implying that genealogy-based population genetic inference can be useful to study the HIV epidemic history. Such evolutionary analyses have shed light on the origins of HIV, and on the epidemic spread of viral variants in different geographic locations and in different populations. The HIV genealogies reconstructed from within-host sequences indicate the action of selection pressure. In addition, recombination has a significant impact on HIV genetic diversity. Accurately quantifying both the adaptation rate and the population recombination rate of HIV will contribute to a better understanding of immune escape and drug resistance. Characterizing the impact of HIV transmission on viral genetic diversity will be a key factor in reconciling the different population genetic processes within and among hosts. (AIDS Reviews 2006;8:125-40)

Key words


Introduction

Simian immunodeficiency viruses (SIV) have frequently moved among primate species, and one such event resulted in the devastating HIV-1 pandemic in humans. Because the virus was not recognized until the early 1980s1, our documented record of the AIDS epidemic is largely limited to the past two decades of its transmission.

However, HIV-1 strains circulating today carry in their genome sequences a significant amount of information about the evolutionary and epidemiologic history of the virus. The high mutation rates2,3 and short generation times4-6 of HIV are the constant fuel for its rapid evolutionary change. The long-term fate of these abundant genetic changes depends on the interplay of effective population size and natural selection, resulting in an extremely high rate of HIV genomic evolution. Population-level processes such as selection, migration, population dynamics, and recombination shape HIV genetic diversity both among and within hosts. The genetic footprint of these processes may be complex, obscured or scrambled, which means that realistic evolutionary models are necessary to recover the useful information contained in HIV gene sequences sampled within and among hosts.

Understanding the processes that determine viral genetic diversity will undoubtedly assist in the struggle against viral infections and will contribute to our knowledge of past epidemiologic events. Here, we discuss evolutionary models and inference methods for HIV population genetic processes. In particular, we highlight some recent computation techniques that have particular utility for comparing HIV-1 epidemics among populations and for investigating the dynamics of intra-host HIV diversity7-10.

Correspondence to:
Philippe Lemey
Department of Zoology
University of Oxford
Oxford, England, UK
Email: philippe.lemey@zoo.ox.ac.uk
Since HIV populations evolve at a rate that is several orders of magnitude faster than that of their human hosts, HIV sequences sampled longitudinally will usually accumulate a significant amount of evolutionary change. Longitudinally sampled or “heterochronous” sequences can be obtained in one of two ways: either from a single patient over the course of infection, or from different patients over the duration of an epidemic (Fig. 1). Interestingly, phylogenies reconstructed from such sequences have distinctive features that reveal the differences in the dynamics of HIV evolution at the inter-host and intra-host levels.

Within each host, the viral population is targeted by both cellular and humoral immune responses, resulting in relatively strong diversifying selection that is most noticeable in the variable regions of the envelope (env) gene. It has been demonstrated that the rate of amino acid substitution in env correlates with the rate of phenotypic escape from neutralizing antibodies. This implies that neutralizing antibody responses cause the relative fitness of different strains within an infection to vary, thus constituting a major force that drives rapid lineage turnover. As a result, intra-host phylogenies of heterochronous env sequences exhibit an asymmetrical or “ladder-like” shape, with limited diversity at any one time (Fig. 1).

In contrast, HIV evolution at the inter-host level shows little evidence that HIV transmission is driven by a similar selective process. Inter-host phylogenies of HIV are not ladder-like and show the persistence of multiple lineages through time. The shape of inter-host phylogenies is primarily determined by (selectively) neutral demographic processes.

In summary, HIV lineages within a host vary in their ability to survive and infect new cells, whereas different HIV lineages among hosts show little genetic variation in their ability to infect new individuals. Some lineages...
may have more opportunity for onward infection than others, but such variation is not heritable and therefore does not generate natural selection. One possible exception may be subtype C, which has been hypothesized to be more sexually transmissible than other strains, but this has yet to be confirmed.

The two patterns outlined above suggest that different evolutionary and population genetic processes should be inferred from within and among host sequence data. Within hosts, we can focus on selection and adaptation and the variables involved in these processes, such as effective population size, virus generation time, and the distribution of selection coefficients. These processes are important from a clinical perspective since they contribute to the variability in disease progression and the development of drug resistance.

Among hosts, we can use the information contained in population-level phylogenies to investigate the movement of HIV lineages among locations and risk groups, or to estimate change in viral effective population size over time. Since the latter depends on the changes in the number and density of infected hosts through time, such analyses can help to elucidate the origin and epidemic spread of different HIV variants.

**Phylogenetic and statistical models for HIV evolution**

To extract useful information from gene sequences we need (a) accurate models of evolutionary and population genetic processes, and (b) statistical methods to infer evolutionary parameters and their confidence limits. Tremendous advances have been made on both fronts in recent years. Evolutionary models for HIV can be classified as having either a phylogenetic or a population genetic basis. In reality, both types of model are closely related and are becoming increasingly integrated. The former models represent the shared ancestry of gene sequences using a bifurcating tree, and typically include complex descriptions of the pattern of nucleotide substitution along lineages. Population genetic methods are most interested in the distribution and frequency of polymorphic nucleotide sites within a set of sequences.

The advantages and disadvantages of each approach are clear: recombination is easier to accommodate in population genetic studies (see later), whereas accurate models of sequence evolution through time are more easily implemented in a phylogenetic framework.

Methods of phylogenetic estimation and models of molecular evolution have both been extensively reviewed elsewhere and will not be covered here. Typically, phylogenies are estimated directly from sequence data and make no assumptions about how population-level processes influence the shape of the phylogeny. To obtain phylogenies with branch lengths that are measured on a calendar timescale (months or years) the model of molecular evolution needs to assume some form of relationship between genetic divergence and time. The simplest is the so-called “strict” molecular-clock model, which assumes a precise linear relationship between the two, such that the rate of molecular evolution remains constant through time. There are several methods to infer constant rates of molecular evolution (substitution rates) from heterochronous sequence data, the simplest being the linear regression approach depicted in figure 1. However, the shared ancestry among the sampled sequences means that the standard regression assumption of independent data points is violated, and appropriate confidence intervals are very difficult to obtain. This and other problems have now been overcome by the development of genealogy-based probabilistic methods.

Even these approaches, however, still assume a strict molecular clock, and although constant rates of HIV evolution might be realistic over small timescales (e.g., within hosts, Fig. 1), statistical testing has demonstrated that this assumption is generally unrealistic for RNA viruses at epidemic scales. Fortunately, more sophisticated methods that can accommodate evolutionary rate variation among lineages have recently been developed, including, most promisingly, Bayesian “relaxed” molecular clocks.

Divergence appears to accumulate at a fairly constant rate within hosts, but at a more variable rate among hosts (Fig. 1). This can be explained by a combination of different replication rates among patients (and thus different HIV generation times; Lemey P, unpublished work) and different levels of immune response, which can influence non-synonymous substitution rates (Lemey P, unpublished work)
is comparable for both datasets, the within-host regression plot for the env C2V5 gene region suggests less divergence-rate variability compared to the complete env sequences sampled among patients.

When the among-patient regression is based on the same C2V5 region as was used for the within-host regression, the rate variability is even more pronounced. Although C2V5 is one of the most divergent gene regions in env, this is not reflected in a steeper regression slope (faster evolutionary rate). Hence the considerable rate-variability among hosts might result in an underestimation of the evolutionary rate at this level, if such rate variability is not explicitly modeled using a relaxed-clock approach (simulations indicated that there was no evidence for substitution saturation; data not shown).

It has been said that the effect of convergent or reversion mutations in C2V5 is to reduce estimated divergence among patients. However, the results of such approaches are likely to be sensitive to the position of the phylogenetic root used for each patient and uncertainty in root position must be considered. Reversion substitutions are likely more common in C2V5 than other HIV genome regions, hence among-host studies should aim to consider multiple genome regions. In addition, the problem of reversionary changes can be avoided by considering synonymous sites only.

To illustrate the effect of rate variability more precisely, evolutionary rate estimates of the strict and relaxed clock assumptions are listed in Table 1. Using a strict molecular clock, the estimate of the C2V5 rate is similar to that for the complete env rate. When a relaxed clock is used, however, the C2V5 rate is somewhat higher than the complete env rate, as expected for this gene region. The coefficient of variation, which quantifies the variability of the substitution rate among branches under a relaxed clock, indicates that the rate varies more among patients in the C2V5 gene region. This is also evident from the greater deviations from the fitted regression line (Fig. 1). The within-host coefficient of variation suggests that substitution rates vary considerably. This might initially seem at odds with the regression analysis (Fig. 1). However, the ladder-like structure of the within-host tree has a greater degree of non-independence of root-to-tip distances due to shared evolutionary history and will result in a lower apparent variance in rates.

Importantly, HIV substitution rate variability within hosts will be affected by transient polymorphisms. These mutations, which are likely to be deleterious, usually segregate on external branches and give rise to a faster rate for these branches (Lemey P, unpublished work). The regression plot, however, will not be affected considerably by different rates for internal and external branches, and will mainly be determined by mutations fixed between time points. Transient deleterious mutations might also explain why the within-host rate is faster than the among-host rate for the C2V5 gene region (Table 1). However, further statistical evaluation is required to test this hypothesis.

The methods outlined above describe how phylogenies measured on a real timescale can be estimated from sampled sequences. A separate set of evolutionary models can be used to estimate population-level processes from these phylogenies – models that provide a mathematical description of the statistical properties of phylogenies under different population scenarios. The theoretical foundation of this technique was originally developed by Kingman, and is generally known as “coalescent theory.” Central to population genetic theory is the relationship between genetic diversity (or patterns of polymorphism) and effective population size, effective population sizes (N_e) are used in

Table 1.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Strict clock rate (subst/site/year)</th>
<th>Relaxed clock rate (subst/site/year)</th>
<th>Coefficient of variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter-host</td>
<td>4.09E-3 (3.54E-3 - 4.70E-3)</td>
<td>4.65E-3 (3.22E-3 - 6.01E-3)</td>
<td>0.32</td>
</tr>
<tr>
<td>Complete env</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inter-host</td>
<td>3.91E-3 (2.45E-3 - 5.28E-3)</td>
<td>5.18E-3 (2.88E-3 - 7.65E-3)</td>
<td>0.51</td>
</tr>
<tr>
<td>C2V5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intra-host</td>
<td>6.88E-3 (5.66E-3 - 8.12E-3)</td>
<td>8.18E-3 (6.26E-3 - 1.01E-2)</td>
<td>0.93</td>
</tr>
<tr>
<td>C2V5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Evolutionary rate estimates were obtained using Bayesian MCMC implemented in BEAST v.1.3. Relaxed clock estimates were estimated using an uncorrelated relaxed clock model. In this approach, rates are drawn independently and identically from an underlying distribution, in this case a lognormal distribution. The data sets are described in the legend of figure 1.
population genetic models to avoid unnecessary mathematic complexity and can be thought of as the "genetic" size of a population.

One way to understand effective population sizes is to note that a real-life, complex, biologic population with effective size $X$ loses genetic diversity by drift at the same rate as a simple "theoretically perfect" population with actual population size $X$. In the latter, effective and actual population sizes are the same, but in the former the effective size is generally smaller. A theoretically perfect population has no natural selection, non-overlapping generations, and no recombination within the genome region under investigation. If the evolutionary dynamics of an organism can be measured on a real timescale of months or years (as outlined above), then effective population size will also be a function of the organism's generation time.

The original coalescent model has been subsequently extended so that it is now able to reflect the population processes of recombination, population subdivision, and changing population size, and can also be extended to heterochronous sequences.

The manner by which population processes affect the shape of phylogenies is illustrated in figure 2, which depicts two theoretically perfect populations with different demographic histories. Because the shape of the sample phylogeny depends on the demographic history of the population, the former can be used to estimate the latter.

How exactly can the model illustrated in figure 2 be used to infer the epidemic history of a viral strain? The first step is to consider that each individual in the population represents one infection, so that population size equals the number of infected individuals. (If the phylogeny represents within-host evolution then each...
individual represents one infected cell). Coalescence events occur when two infections in a sampled lineage are both infected by the same donor. In essence, the method assumes that the phylogeny estimated from virus sequences accurately reflects the underlying transmission tree. This situation is analogous to the problem in evolutionary biology of estimating species trees from gene trees. This assumption is unlikely to be too restrictive because (i) virus transmission, or among-virus competition after transmission, typically generates a strong population bottleneck, making it very unlikely that multiple viral lineages will be repeatedly transmitted, and (ii) viral gene trees often span several decades so the time when two viral lineages coalesce is close (relative to the timescale of the phylogeny) to the actual transmission times. This may not be true when small transmission chains occurring over a short timescale are analyzed.

Probabilistic inference under many population genetic models has been achieved using standard maximum likelihood (ML) estimation, where the aim is to find the population parameters that give maximum probability to the observed genealogy. The genealogy is, however, never directly observed, but usually it is itself the result of phylogenetic inference based on the sequence data. Therefore, if population genetic inference is conditioned on a single ML phylogeny, any stochastic or systematic errors in the phylogenetic estimation procedure are ignored.

Computationally intensive methods have recently been proposed to tackle this problem by averaging over a set of plausible genealogies using Monte Carlo integration. For serially sampled populations, full probabilistic genealogy-based modeling and Bayesian inference using Markov Chain Monte Carlo (MCMC) sampling have proven most useful. The examples of HIV demography discussed below were analyzed using BEAST, a Bayesian MCMC program for genealogy-based population genetic inference that includes the recently developed Bayesian skyline plot model for genealogy-based population genetic inference that includes the recently developed Bayesian skyline plot model as well as relaxed clock models.

**Within-Host HIV population dynamics**

Determining the effective population size of HIV populations within patients is a key goal towards understanding within-patient evolutionary dynamics. It decides whether genetic drift or natural selection is the most important evolutionary process, and also determines whether HIV genetic variation should be modeled stochastically or deterministically. Evolutionary theory predicts that in small populations, mutations will be produced more rarely and their fixation will largely depend on chance stochastic events (genetic drift). In large populations, however, mutations occur more frequently, but their fate will ultimately be decided by the deterministic action of natural selection.

Distinguishing between these scenarios is essential to understanding processes such as drug resistance and immune escape. Although the number of HIV-infected cells within hosts is estimated around $10^2$-$10^8$, the fraction of cells that contribute to future generations of viruses is unknown (this fraction is a key determinant of effective population size).

Estimates of $N_e$ using coalescent approaches applied to HIV within-host gene sequences have typically been much lower than the number of infected cells ($10^3$), and this has been interpreted as support for stochastic models of HIV evolution. Figure 3 shows within-host Bayesian skyline plots for three, different, longitudinally sampled patients with estimates of $N_e$ ranging between $10^2$ and $10^4$. Proponents of stochastic within-host HIV evolution might also consider these estimates as being in favor of neutrality. However, all coalescent-based estimators of $N_e$ assume neutrality a priori, but when this assumption cannot be upheld, conclusions about the evolutionary processes based on the size of $N_e$ will be inappropriate. There are also several problems in trying to justify this assumption using standard neutrality tests, including weak statistical power to reject the neutral regime, and non-independency when applied to serial sampled data.

With these limitations in mind, Rouzine and Coffin (1999) adopted a different approach, inspired by classical population genetics, that evaluates the pattern of linkage disequilibrium in HIV sequence data. The basis of their test is the deviation of the frequency at which four possible genetic variants at two loci are observed from the product of the corresponding one-locus frequencies. The extent of linkage disequilibrium is critically dependent on $N_e$, and according to simulation experiments, the pattern observed in HIV sequence data was compatible with values of $N_e$ ranging between $2^{10}^4$ to $5^{10}^4$. These estimates are closer to values that can be expected under a deterministic regime. They note, however, that the population size might be significantly reduced under HAART, which could explain the variability in time to develop drug resistance.

More recently, Edwards, et al. (2006) used genealogy-based statistics to measure deviations from neu-
trality, explicitly accounting for the potential bias of demography. These statistics revealed a clear signal of selection that could not be generated by recombination. Under such deviations from neutrality, estimates of $N_e$, like the ones plotted in figure 3, should merely be regarded as a measure of diversity.

The strength of natural selection in HIV gene sequences has also been repeatedly demonstrated using the non-synonymous/synonymous substitution rate ratio ($d_N/d_S$), especially in the $env$ gene region. Site-specific estimates of $d_N/d_S$ using genealogy-based codon substitution models have proven useful in investigating the distribution of selective coefficients in protein coding sequences. These approaches are prone to overestimation of the number of positively selected sites when recombination is in effect. Recently, a population genetic approximation to the coalescent, rather than a phylogenetic approach, has been proposed to estimate diversifying selection in the presence of recombination from isochronous sequences, highlighting the importance of adopting a population genetic perspective to tackle complex evolutionary problems. An extension of the nonparametric McDonald-Kreitman test has also been applied to HIV to estimate within-host adaptation rates, assuming free recombination, and revealed a staggering adaptation rate of, on average, one adaptive fixation event in $env$ every ~2.5 months. Such estimates, in addition to the strongly asymmetrical shape of within-host genealogies and the values of genealogy-based statistics, argue for strong selection acting on the immunodominant $env$ gene. It has been argued that genes coding for targets of the immune response can be affected by frequency dependent selection, where haplotypes coding for a new epitope will have a selective advantage until an appropriate immune response has been developed. Model-based inference will be necessary to obtain quantitative insights into this process, which has already been reported to occur for HIV in vivo and in vitro.

Both experimental and epidemiologic studies have provided compelling evidence that genetic recombination is a significant factor in shaping HIV diversity. Abundant coinfection of spleen cells, the diploid HIV genome, and reverse transcriptase strand transfer events constitute the mechanism by which recombination within hosts occurs. Not surprisingly, high recombination rates have been estimated using coalescent

---

Figure 3. Within-host Bayesian skyline plots: partial $env$ gene longitudinally sampled from three different patients over 105, 91, and 105 months respectively (patient1: 129 sequences, 516 bp; patient 3: 109 sequences, 516 bp; patient 7: 198 sequences, 516 bp). The population estimate obtained by BEAST, $N_e$, generation time, was rescaled to $N_e$ using a HIV generation time of two days. The bold line represents the mean population size, while the grey area represents the 95% credibility interval.
methods applied to sequence data sampled across hosts as well as within hosts\textsuperscript{66,67}.

Population genetic methods estimate the population recombination rate, i.e., the rate at which recombinant genome regions become fixed in the population. This rate is a function of both the per-generation “molecular” recombination rate and effective population size\textsuperscript{68}. Shriner, et al. (2004) studied the HIV recombination rate within a single individual and applied stringent experimental procedures to avoid artificial recombination whilst amplifying the 3' half of the genome\textsuperscript{73}. By rescaling an estimate of the population recombination rate, using the experimentally determined mutation rate ($\mu = 2.5 \times 10^{-5}$ per site per generation)\textsuperscript{67} and the coalescent estimate of $\theta = N_e \mu$, they obtained a mean estimate of $1.38 \times 10^{-4}$ recombination events per adjacent sites per generation\textsuperscript{75}. The study also revealed considerable variation between two different coalescent estimators of the recombination rate\textsuperscript{66,69}, indicating that caution should be taken when interpreting such estimates.

Several biologic factors might give rise to biases in the estimate of population recombination rates. Attempts have been made to implement more complex nucleotide substitution models in recombination rate estimators\textsuperscript{70}, but the specifics of within-host demographic history and deviations from neutrality outlined above need to be accounted for. Simulations have shown that these processes do affect estimates of recombination rate\textsuperscript{73}.

Progress is being made towards the goal of co-estimating multiple evolutionary and population genetic forces\textsuperscript{10,71,72}, but there is a need for these approaches to be validated and extended to serially sampled data.

It appears that HIV is well adapted to recombine within hosts and it has been proposed that the recombination process is a form of sexual reproduction. In asexual populations, the accumulation rate of beneficial mutations in a gene will be restricted by their linkage to other segregating mutations\textsuperscript{73}. It has been suggested that the biologic role of recombination is to counteract the adverse effects of linkage and accelerate adaptation rates. However, recombination can both create and break up favorable combinations of viruses, so the net effect of recombination depends on the interaction of the fitness effects of different mutations within the genome (referred to as ‘epistasis’).

Recombination would be beneficial in situations of negative epistasis, where the combination of two detrimental mutations results in a greater loss of fitness than expected from the single mutations (synergy), and where beneficial mutations may act antagonistically. However, an extensive analysis of HIV protease and partial RT sequences with associated fitness values indicated that there is a predominant signal of positive epistasis (the opposite scenario)\textsuperscript{74}, but the statistical support for this has been recently questioned\textsuperscript{75}.

Rouzine and Coffin (2005) presented a model for HIV dynamics under selection and weak recombination, and derived an accumulation rate of beneficial mutations that increases with higher recombination rates and increased population size. They also provided important predictions for HIV evolution under antiviral therapy by showing that drug-resistance evolution can be prevented if the HIV population is suppressed below a critical value, and that drug concentrations required to prevent rebound of resistant virus can be significantly decreased if the number of target sites in the HIV genome is large\textsuperscript{76}.

HIV forms distinct subpopulations in different anatomical sites as well as in different cell types, often referred to as compartmentalization. Even within a specific organ, genetic analysis has revealed appreciable population substructure\textsuperscript{13,77,78}.

The pattern of HIV migration and colonization in different body tissues and cell types has important evolutionary and clinical repercussions. For example, HIV dynamics between blood and brain tissue, and within different brain compartments, have been implicated in the pathogenesis of HIV-associated dementia and the independent development of drug resistance\textsuperscript{77,79}. Within-host HIV migration events have typically been displayed graphically using phylogenetic trees, and sometimes quantitatively assessed using parsimony techniques\textsuperscript{78}. Although the impact of migration on expected phylogenetic tree shape has been adequately modeled and several software tools to estimate migration rates are available\textsuperscript{80,81}, such coalescent approaches have yet to be seriously applied to HIV gene sequences. This might change now that migration rates, substitution rates, and population sizes can be simultaneously estimated from heterochronous sequences\textsuperscript{9}, even when these parameters and the number of subpopulations change over time\textsuperscript{82}. Moreover, co-estimating migration rates, population growth and recombination rates has been made available for isochronous sequences\textsuperscript{80,81,83,84}.

It should be noted that all current implementations assume an “island model”, which does not allow for extinction of subpopulations. Although this might be a reasonable assumption in many applications, within-host HIV dynamics have been shown to fit a meta-population
Figure 4. Geographic range of two chimpanzee subspecies (Pan troglodytes Schweinfurthii and Pan troglodytes troglodytes) and sooty mangabey species (Cercocebus atys atys), phylogenetic tree of the SIVcpz/HIV-1 lineage and the SIVsmm/HIV-2 lineage and Bayesian skyline estimates for HIV-1 group M, HIV-1 group O and HIV-2 group A. The SIVcpz/HIV-1 phylogenetic tree was reconstructed from pol amino acid sequences, including recently obtained Ptt samples from Cameroon 102; The SIVsmm/HIV-2 phylogenetic tree was reconstructed from gag nucleotide sequences. The color of the branches in the trees matches the color of the geographic ranges of the chimpanzee subspecies. The timeframe of the independence war in Guinea-Bissau (1963-1974) is superimposed as a magenta rectangle onto the HIV-2 group A population dynamics. For HIV-1 group O, the skyline plot for the concatenated gag, int. and env gene sequences and the env sequences separately are shown in black and blue respectively.
model in particular cases\textsuperscript{13}. Allowing for local extinction (e.g. as a consequence of a high turnover of productively infected T-cells\textsuperscript{10}) might therefore be an interesting extension of structured coalescent models for HIV.

Besides population structuring, infection of different cell-types that have different turnover rates can also considerably influence viral generation times\textsuperscript{85}. Differences in cellular turnover first became evident when decay curves of plasma viremia following antiretroviral treatment were investigated\textsuperscript{5,86}. After an initial rapid decay for 1-2 weeks, largely representing the decline in productively infected CD4+ T-lymphocytes, plasma virus decreases at a lower rate and ultimately drops below the viral-load detection limit\textsuperscript{5,86}. The second decay phase might be due to macrophages, which are less sensitive to viral cytopathogenic effects and delayed HIV release from dendritic cells. Even under the viral-load detection limit, a stable HIV reservoir remains in the form of resting memory T-cells, which can still release replication-competent virus upon reactivation\textsuperscript{87}. The extremely long half-life of this compartment guarantees lifelong viral persistence and destroys the hope for eradication using the current antiretrovirals\textsuperscript{88}. If viral lineages have gone through infection rounds in those cellular reservoirs with slow turnover, mean replication rates will considerably decrease\textsuperscript{85}.

It has been shown that replication rate (and its reciprocal, generation time) can be inferred using an estimate of the synonymous substitution rate and the \textit{in vitro} estimate of the mutation rate per replication cycle. These estimates agree with mathematic modeling of virologic data when subpopulations of latently infected cells are taken into account\textsuperscript{85}. It should be noted, however, that the estimate of the mutation rate used in this study might have been too low (μ = 5.33 x 10^{-4} per site per replication cycle vs. 2). Kelly, et al. (2003) provided a more formalized framework to bridge the gap between dynamic models and population genetic models of HIV infection, confirming the prediction of a reduced evolutionary rate when infection involves multiple cell-types. Several-sample coalescent theory has also been employed to infer HIV-1 generation time \textit{in vivo}\textsuperscript{89} and from data extracted from single-rate coalescent theory, which is the result of a separate coalescent analyses\textsuperscript{5,86}. This method not only incorporates phylogenetic uncertainty but also takes into account the uncertainty in the estimate of evolutionary rate, whereas in the original analysis both the phylogeny and the rate were fixed to a point estimate\textsuperscript{86}.

In figure 4, we show a more up-to-date estimate for the same dataset, obtained using the Bayesian skyline plot method\textsuperscript{7}. This method not only incorporates phylogenetic uncertainty but also takes into account the uncertainty in the estimate of evolutionary rate, whereas in the original analysis both the phylogeny and the rate were fixed to a point estimate\textsuperscript{86}.

Figure 4 confirms an early period of slow HIV-1 spread, followed by a subsequent period of more rapid epidemic transmission in more recent decades. The timescale of this estimate, which suggests an origin of HIV-1 group M around 1940, is slightly more recent than in previous coalescent analyses\textsuperscript{86}. The slightly different timescale results from a different treatment of the evolutionary rate parameter in the two analyses. Here, we used Bayesian inference to estimate evolutionary rate from the full length \textit{env} gene subtype B data set of Robbins, et al. 2003, whilst allowing for a separate rate for the V3-V5 gene region.

In the previous analysis, the evolutionary rate was fixed to 0.0023 nucleotide substitutions per site per year (QL: 0.0016-0.0033)\textsuperscript{96}, which was based on analysis of the V3-V5 region of the \textit{Korber}, et al. (2000) dataset\textsuperscript{95}. Although this region is one of the most variable in the HIV genome, it gave a roughly similar rate to that obtained for the complete \textit{env} gene, sug-
ggesting that this estimate may be subject to the problem of underestimation arising from the assumption of rate constancy, as discussed above and illustrated in figure 1.

Given that heterochronous data is now available for this DRC population\textsuperscript{98-100}, it would be more appropriate to estimate the rate from the data rather than fixing it to one value or specifying a strong prior distribution.

The geographic location, timing, and phylogenetic structure of the HIV-1 epidemic all provide evidence against the hypothesis that HIV has emerged due to SIVcpz-contaminated oral polio vaccines in the DRC in the late 1950s\textsuperscript{93,95,96}. In the past, low SIV infection rates in chimpanzees have prevented a very strong case being made for the natural transfer hypothesis of SIV to humans. The finding that all three HIV-1 lineages were spawned by an SIV progenitor carried by the central chimpanzee subspecies, \textit{Pan troglodytes troglodytes} (Ptt), and that \textit{Pan troglodytes Schweinfurthii} only carry an SIV distantly related to HIV-1\textsuperscript{101}, highlighted the importance of screening the \textit{Ptt} subspecies.

Very recently, an analysis of about 600 fecal samples revealed several \textit{Pan troglodytes troglodytes} communities in southern Cameroon with widespread SIVcpz-Ptt infection, indicating that the outbreak began in rural Cameroon and then traveled to Kinshasa, DRC\textsuperscript{102}. This agrees with our genetic estimate of HIV-1 group M epidemic history (Fig. 4). We hypothesize low and unrecognized HIV transmission in remote African areas during the early phase of the history of the epidemic, followed by more rapid epidemic spread within a changing and increasingly connected African population, possibly assisted by some level of iatrogenic human-to-human transmission\textsuperscript{103}.

In contrast to HIV-1 group M, group O infections have mostly remained restricted to Cameroon, with some movement to neighboring countries in West Central Africa. Current group O seroprevalence is relatively modest; however, a higher genetic diversity for group O viruses compared to group M viruses prompted the suggestion group O has been circulating in Central Africa for longer\textsuperscript{104}. A Bayesian skyline plot estimated from concatenated \textit{gag}, \textit{int} and \textit{env} gene sequences does indeed suggest an earlier common ancestor for group O than for group M (Fig. 4). The group O estimate of epidemic history has been plotted on the same scale as the group M estimate; the plots indicate that group O in Cameroon has not undergone the same explosive spread as group M in the DRC.

Previous analyses of this heterochronous data used a multi-locus model that allowed each gene region to have a different phylogenetic history and suggested that such estimates are not heavily biased by recombination among gene regions\textsuperscript{105}. Although replication assays have found group O to be significantly less “fit” than group M\textsuperscript{106}, these \textit{in vitro} differences are not reflected in \textit{in vivo} differences in viral load\textsuperscript{107} and therefore should not be extrapolated to the epidemiologic level, despite the temptation to ascribe the different epidemic outcomes of groups O and M to potential differences in transmissibility.

The relative contributions of viral genetic differences and epidemiologic circumstances to the variation in epidemic history among strains must therefore remain an unanswered question. The skyline plot results indicate that the number of effective infections were roughly similar around 1960, but that group O did not benefit as greatly from the extrinsic factors that led to the increased transmission of group M infection after 1960. HIV-1 group N might have found itself in even less favorable conditions for epidemic spread.

The second type of HIV, HIV-2, clusters with SIV from sooty mangabeys (\textit{Cercocebus atys atys}) to form a separate lineage in primate lentivirus phylogeny. Different HIV-2 lineages appear to be the result of separate cross-species transmissions, but only two HIV-2 strains (A and B) give rise to an appreciable number of infections in humans. Viral load within asymptomatic patients and transmission probability between individuals are both significantly lower for HIV-2 than for HIV-1 group M\textsuperscript{108,109}. It is therefore unsurprising that HIV-2 has not spread much further than the West African countries that coincide with the historic range of the sooty mangabey (Fig. 4).

The densest focus of HIV-2 prevalence is in Guinea-Bissau. A community study carried out in northwestern Guinea-Bissau revealed a seroprevalence as high as 10% among adults\textsuperscript{110,111}. In this and other populations, HIV-2 prevalence consistently peaked in older age groups.\textsuperscript{112} This observation plus further investigation of HIV-2 risk factors led to the hypothesis that HIV-2 was mainly disseminated by a generation that was sexually active during the independence war in Guinea-Bissau which took place during the 1960s and early 1970s\textsuperscript{112}. Coalescent analysis enabled this hypothesis to be tested using genetic data and revealed that HIV-2 subtype A switched from endemic transmission to epidemic growth sometime around 1955-1970\textsuperscript{112}. We estimated a Bayesian skyline plot from HIV-2 env sequences sampled in northwestern Guinea-
Bissau; this confirms that the period of epidemic growth coincides with the time frame of the Guinea-Bissau independence war (1963-1974, Fig. 4). In this analysis, a prior distribution for the rate of evolution was provided by an analysis of the same gene region of serially sampled HIV-2 sequences, published by Shi, et al. (2005). Both sexual and blood-borne HIV-2 transmission might have drastically increased during the independence war, and large-scale inoculation campaigns have been recorded at the local hospital where the sequence data was obtained.

Coalescent analyses similar to those presented above have been used several times to study specific subtypes and epidemic strains of HIV-1 group M, both in Africa and in other continents. The methods have also been used to compare the epidemic potential of different subtypes co-circulating in the same population. Although subtype distributions in different countries are continually changing, the developed world has mainly been burdened by HIV-1 subtype B infections. Until the early 1980s, this strain spread unnoticed among high-risk groups, notably homosexual men, in the USA. Phylogenetic-based reconstruction of the epidemic history of subtype B in the USA, including strains sampled relatively early in the epidemic, revealed an explosive spread in the 1970s that slowed down towards the present (a logistic growth trend; Fig. 5). The estimated time for the population to double in size at the onset of the epidemic is amongst the shortest reported for HIV population dynamics (0.84 years⁻¹, 0.74-0.96)³⁵, consistent with its propagation through standing networks of injecting drug-users and homosexual men.³⁷

Interestingly, similar growth rates were estimated for several UK homosexual transmission clusters, emphasizing again the importance of high-risk group dynamics in the onset of an epidemic. Although the virus established an epidemic in the UK by several independent and more recent introductions, the inferred demography in the homosexual “sub-epidemics” is qualitatively (logistic growth) and quantitatively (growth rate and ratio of effective number of infections over prevalence) similar to the U.S. epidemic history (Fig. 5).³⁸ The leveling off towards an equilibrium state is consistent with behavioral interventions and HIV prevention strategies in both populations. However, it should be noted that all Bayesian skyline plots show some signal of steady-state dynamics towards the present (Figs. 4 and 5), which could be at least partly attributed

Figure 5. Bayesian skyline plots for HIV-1 subtype B. The skyline plot representing the U.S. epidemic history was reconstructed using the serially sampled data set analyzed in Robbins K, et al. The date for the most recent common ancestor of sequences from US origin is indicated with an arrow. The UK skyline plot was inferred using the largest cluster (“Cluster 2”) identified in Hue, et al. 2005; the same prior distribution for the evolutionary rate was used in our analysis.
to within-host HIV evolution. If all sequences are sampled from different hosts, then transmission events (and thus coalescent events) are expected to occur at least some years into the past. In addition, such sequences probably carry recent (slightly) deleterious mutations, which have not yet been eliminated at the population level by purifying selection, leading to an overestimation of the time to the most recent coalescence event in the tree.

Further research is needed to quantify and model these factors. For the time being, we recommend that the uncertainty of any estimate is always taken into account when coalescent analyses are being interpreted, and we suggest that very recent epidemic history should be interpreted with caution, especially when estimated phylogenies contain long external branches.

What is the influence of recombination on HIV demographic inferences? Although recombination is undoubtedly pervasive within hosts\textsuperscript{65,67}, this does not necessarily invalidate estimating a transmission tree using HIV gene trees for three main reasons:

(i) The ability to reconstruct genealogies of sequences sampled across hosts will mainly be hampered by recombination events between distinct variants harbored by different patients, therefore requiring coinfection or superinfection. Although cases of superinfection have been reported and several mosaic HIV genomes are the “circulating” proof of their occurrence\textsuperscript{119,120}, estimated rates of superinfection are generally very low\textsuperscript{121-123}. Rates of superinfection will also vary considerably among risk groups, being greater in high-risk groups such as injecting drug users, and commercial sex workers, who make up only a small fraction of overall prevalence.

(ii) The more divergent parental sequences of recombinants are, the more impact they are expected to have on tree reconstructions. Those recombinants are, however, the easiest to identify using recombination detection programs and can be omitted from the analysis (although this is an \textit{ad hoc} way to deal with recombination).

(iii) HIV demographics are generally characterized by exponential growth, at least in some stage of the epidemic history, generating star-like trees with long external and short internal branches (Fig. 1 B). In these growing populations, fewer recombination events can scramble the topologic information contained in the sequences.

Conclusions and perspectives

Different population genetic processes are shaping viral diversity within and between hosts. Model-based inference of HIV gene sequences now enables quantitative insights into the effects of these processes to be obtained. Among hosts, the change in viral effective population size, which has not been estimated before, is considerable.
population size over time can be modeled, revealing historic changes in transmission dynamics. In contrast to within-host HIV dynamics, there is little influence of immune-driven natural selection at this level. Extensive variation in partner exchange and transmission- associated bottlenecks can both contribute to genetic drift at the population level\textsuperscript{13,14,15}. In addition, selectively advantageous mutations might "miss the boat" for transmission if they occur late in infection\textsuperscript{14,15}. Modeling and analysis of transmission chain data might help to explain how intra-host evolution is transformed into HIV evolution at the population level. For example, a recent coalescent analysis of a homosexual transmission pair revealed that transmission was associated with a severe loss of diversity (> 99\%)\textsuperscript{17}. Whether HIV transmission is selectively neutral is, however, still the subject of debate\textsuperscript{33,48}.

The population genetic inferences discussed in this review indicate that model complexity required depends on the level at which the virus population is sampled. Different processes need to be modeled within and among hosts. Although the complexity of HIV intra-host, inter-host and transmission dynamics is not unique among human pathogens (cfr. Hepatitis C)\textsuperscript{12}, HIV is by far the most extensively studied pathogen and is represented by the greatest amount of genetic data. Therefore, HIV presents an opportunity to truly understand viral genetic diversity and the population genetic processes that shape it.

Acknowledgements

We would like to thank Beatrice Hahn for providing the figure of the natural range of sooty mangabey and two chimpanzee subspecies. We thank Stephane Hué for providing the sequence data of the UK transmission cluster. Philippe Lemey was supported by a long-term EMBO fellowship. Andrew Rambaut and Oliver G. Pybus were supported by the Royal Society.

References


