AIDS Reviews 2006;8:125-40

HIV Evolutionary Dynamics Within and Among Hosts

Philippe Lemey[1,2], Andrew Rambaut[1] and Oliver G. Pybus[1]
[1]Department of Zoology, University of Oxford, Oxford, UK; [2]Rega 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)**
Corresponding author Philippe Lemey, philippe.lemey@zoo.ox.ac.uk

Key words

HIV. Evolution. Population dynamics. Coalescent. Selection. Recombination.

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 1980s[1], 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 rates[2,3] and short generation times[4-6] 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 diversity[7-10].

Correspondence to:
Philippe Lemey
Department of Zoology
University of Oxford
Oxford, England, UK
Email: philippe.lemey@zoo.ox.ac.uk

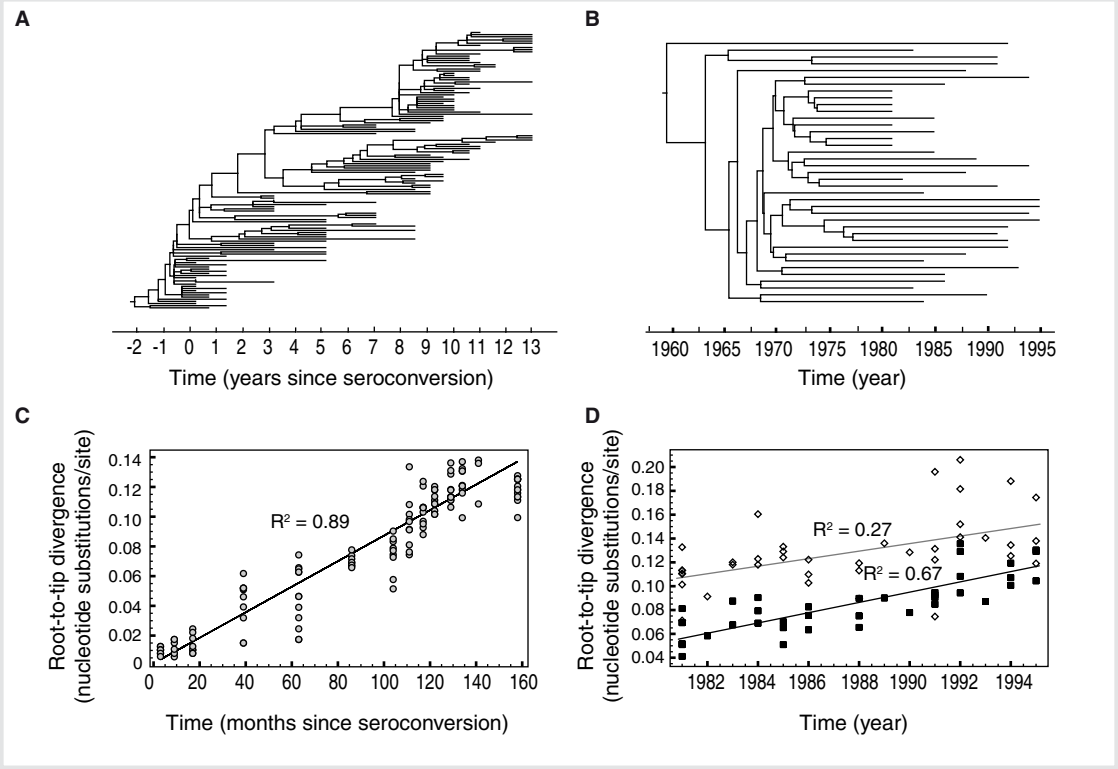

Figure 1. A: *HIV-1 within-host phylogeny with branch lengths in time units: partial* env *gene longitudinally sampled from a single patient over 155 months (subtype B, 129 sequences, 516 bp; patient 9)*[34,53]. **B:** *HIV-1 population phylogeny with branch lengths in time units: full length* env *gene sampled during the U.S. HIV epidemic from 1981-1995 (subtype B, 39 sequences, 2396 bp)*[35]. **C:** *Root-to-tip divergence as a function of sampling time for the within-host phylogeny. The R^2 value is indicated above the regression line.* **D:** *Root-to-tip divergence as a function of sampling time for the population phylogeny: divergence estimates for the full length and C2V5* env *gene are shown with black squares and open diamonds respectively.*

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"[11] 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[12].

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[13]. 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)[12].

In contrast, HIV evolution at the inter-host level shows little evidence that HIV transmission is driven by a similar selective process[14,15]. Inter-host phylogenies of HIV sampled through time are not ladder-like and show the persistence of multiple lineages through time (Fig. 1)[12]. The shape of inter-host phylogenies is primarily determined by (selectively) neutral demographic processes[12,15].

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[16,17], 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[18,19].

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[14], such analyses can help to elucidate the origin and epidemic spread of different HIV variants[20].

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[21,22] 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[23]. 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[24], 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[24]. This and other problems have now been overcome by the development of genealogy-based probabilistic methods[25,26].

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[27]. Fortunately, more sophisticated methods that can accommodate evolutionary rate variation among lineages have recently been developed[28,29], including, most promisingly, Bayesian "relaxed" molecular clocks[30-32].

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[18] (Lemey P, unpublished work)[33].

Using a molecular clock when evolutionary rates vary considerably may lead to underestimation of the rate and overestimation of divergence times[29], so caution should be taken when assessing differences in evolutionary rate. This is illustrated in figure 1 by regression plots of root-to-tip divergence for *env* sequences sampled (i) throughout the infection history of a single patient and (ii) across different patients during the U.S. epidemic[34,35]. Although the timescale of sampling

Table 1.

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

Evolutionary rate estimates were obtained using Bayesian MCMC implemented in BEASTv.1.3[50]. Relaxed clock estimates were estimated using an uncorrelated relaxed clock model[31]. 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.

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[36]. 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 under both 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[24].

Importantly, HIV substitution rate variability within hosts will be affected by transient polymorphisms[8]. 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 such phylogenies – models that provide a mathematic description of the statistical properties of phylogenies under different population scenarios. The theoretic foundation of this technique was originally developed by Kingman, and is generally known as "coalescent theory"[37,38]. 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

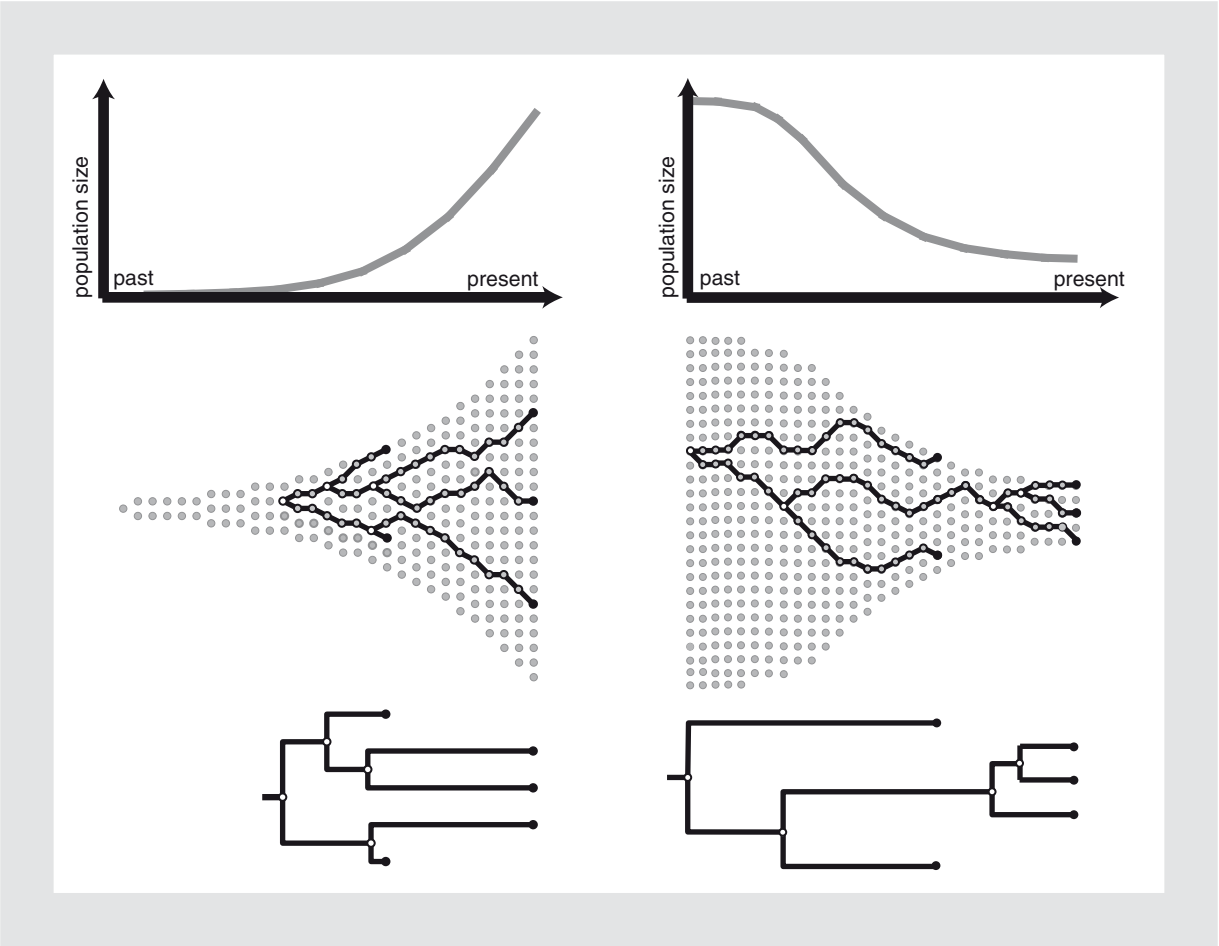

Figure 2. *Representation of the coalescent process for variable population sizes. In one, the population size has been exponentially increasing (left), and in the other it has recently decreased (right, a population 'bottleneck'). Moving back in time, from the present to the past, sampled lineages join together, or coalesce. The rate at which they do this is inversely proportional to the population size, as there are fewer possible ancestors for each lineage when the population size is small. The genealogies for five samples drawn at different times from this coalescent process are shown underneath. If substitutions accumulate during this genealogic process, then we can estimate the genealogy using phylogenetic approaches applied to gene sequences.*

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; 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[39,40], population subdivision[41], and changing population size[42,43], and can also be extended to heterochronous sequences[44].

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[45]. This situation is analogous to the problem in evolutionary biology of estimating species trees from gene trees[46]. 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[47-49], 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[128].

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[26]. The examples of HIV demography discussed below were analyzed using BEAST[50], a Bayesian MCMC program for genealogy-based population genetic inference that includes the recently developed Bayesian skyline plot model as well as relaxed clock models[7,31].

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[51] around 10^7-10^8, 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[52,53]. 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[8,54]. 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[8,54].

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[54] was compatible with values of N_e ranging between $2*10^{-4}$ to $5*10^{-5}$. 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[19].

More recently, Edwards, et al. (2006) used genealogy-based statistics to measure deviations from neu-

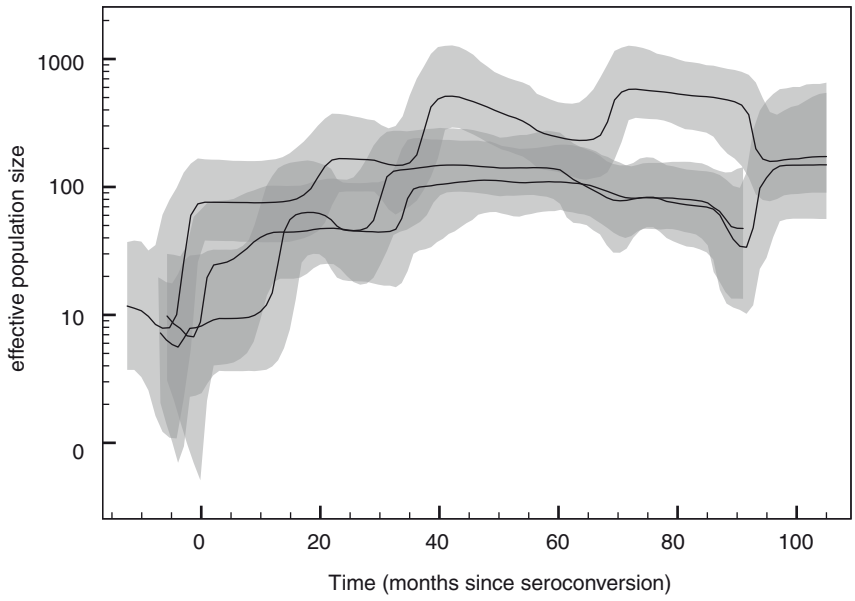

Figure 3. *Within-host Bayesian skyline plots: partial* env *gene longitudinally sampled from three different patients over 105, 91, and 105 months respectively (patient1: 129 sequences, 516 bp; patient 3: 109 sequences, 516 bp; patient 7: 198 sequences, 516 bp)[34,53]. The population estimate obtained by BEAST[50], 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.*

trality, explicitly accounting for the potential bias of demography. These statistics revealed a clear signal of selection that could not be generated by recombination[8]. 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[55,56]. In particular, 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[57,58]. Recently, a population genetic approximation to the coalescent with recombination, rather than a phylogenetic approach, has been proposed to estimate diversifying selection in the presence of recombination from isochronous sequences[10], highlighting the importance of adopting a population genetic perspective to tackle complex evolutionary problems. An extension of the nonparametric McDonald-Kreitman test[59] 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[18]. Such estimates, in addition to the strongly asymmetrical shape of within-host genealogies and the values of genealogy-based statistics[8], 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[60]. 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*[61,62].

Both experimental and epidemiologic studies have provided compelling evidence that genetic recombination is a significant factor in shaping HIV diversity[63,64]. Abundant coinfection of spleen cells[65], 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

methods applied to sequence data sampled across hosts as well as within hosts[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[68]. Shriner, et al. (2004) studied the HIV recombination rate within a single individual and applied stringent experimental procedures to avoid artifactual recombination whilst amplifying the 3' half of the genome[67]. 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[2]) and the coalescent estimate of θ ($= N_e * \mu$), they obtained a mean estimate of 1.38×10^{-4} recombination events per adjacent sites per generation[67]. The study also revealed considerable variation between two different coalescent estimators of the recombination rate[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[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[70].

Progress is being made towards the goal of co-estimating multiple evolutionary and population genetic forces[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[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)[74], but the statistical support for this has been recently questioned[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[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[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[77-79]. Within-host HIV migration events have typically been displayed graphically using phylogenetic trees, and sometimes quantitatively assessed using parsimony techniques[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[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[9], even when these parameters and the number of subpopulations change over time[82]. Moreover, co-estimating migration rates, population growth and recombination rates has been made available for isochronous sequences[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

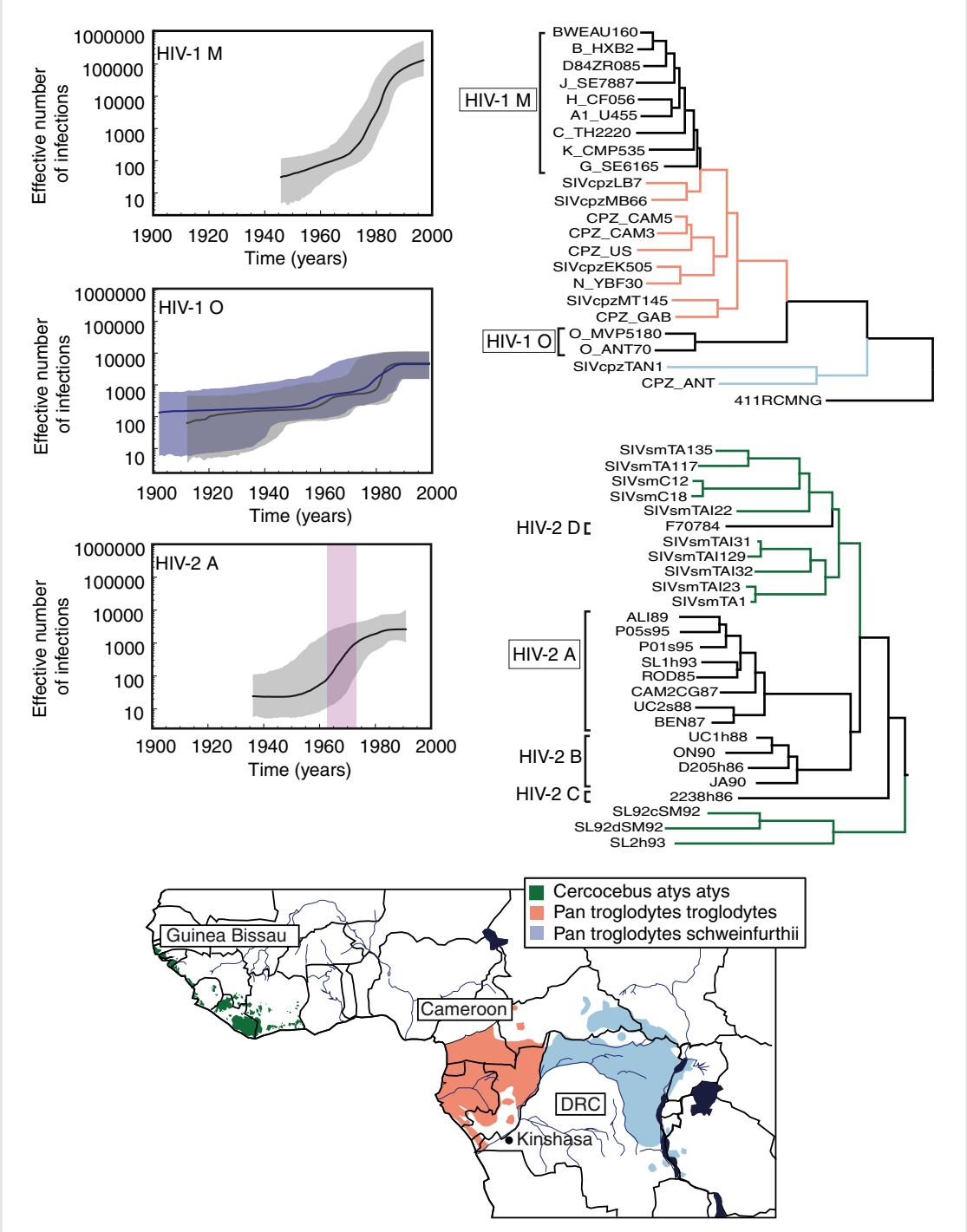

Figure 4. *Geographic range of two chimpanzee subspecies (*Pan troglodytes Schweinfurthii *and* Pan troglodytes troglodytes*) and sooty mangabey species (*Cercocebus atys atys*), phylogenetic tree of the SIVcpz/HIV-1 lineage and the SIVsmm/HIV-2 lineage and Bayesian skyline estimates for HIV-1 group 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[13]. Allowing for local extinction (e.g. as a consequence of a high turnover of productively infected T-cells[13]) 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[85]. Differences in cellular turnover first became evident when decay curves of plasma viremia following antiretroviral treatment were investigated[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[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[87]. The extremely long half-life of this compartment guarantees lifelong viral persistence and destroys the hope for eradication using the current antiretrovirals[88]. If viral lineages have gone through infection rounds in those cellular reservoirs with slow turnover, mean replication rates will considerably decrease[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 *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[85]. It should be noted, however, that the estimate of the mutation rate used in this study might have been too low ($\mu = 5.33 \times 10^{-4}$ 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[89]. Serial sample coalescent theory has also been employed to infer HIV-1 generation time *in vivo*[44], this time with an updated estimate of the mutation rate[2], providing estimates more agreement with virologic data (1.2 days/generation *vs.* 1.8 days/generation[5]).

HIV population dynamics among hosts

The ability to infer the dates of origin of epidemics and to investigate historic patterns of transmission from viral gene sequences now plays a key role in molecular epidemiology. Many studies have contributed to what can be considered a reasonably clear picture of how different HIV variants have spread in the past.

HIV-1 comprises three different lineages, groups M, N, and O, each of which is the result of a separate cross-species transmission of SIV from chimpanzees[90-92]. Group M has successfully founded epidemics worldwide and these founder events have led to the generation of several subtypes[15]. Evolutionary analyses using molecular clock techniques suggest that a common ancestor of HIV-1 group M existed around 1930[93,94]. The immediate precursor of HIV-1 group M infects chimpanzees of the subspecies *Pan troglodytes troglodytes* in West Central Africa[90,91] (Fig. 3). Not surprisingly, the highest degree of group M diversity has been found close to this area.

An epidemiologic survey by Vidal, et al. (2000) revealed a very high diversity of HIV strains present in the Democratic Republic of Congo (DRC)[98]; these strains showed much less distinction between intra- and intersubtype diversity in comparison to group M strains sampled globally[95]. Coalescent analysis of these sequences (albeit based on a single-tree estimate) suggested a logistic growth of HIV-1 population size over time[96,97].

In figure 4, we show a more up-to-date estimate for the same dataset, obtained using the Bayesian skyline plot method[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[96].

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[96]. 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 *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 (CI: 0.0016-0.0033)[96], which was based on analysis of the V3-V5 region of the Korber, et al. (2000) dataset[93]. 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 *env* gene, sug-

gesting 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[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[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, *Pan troglodytes troglodytes* (*Ptt*, Fig. 4), and that *Pan troglodytes Schweinfurthii* only carry an SIV distantly related to HIV-1[101], highlighted the importance of screening the *Ptt* subspecies.

Very recently, an analysis of about 600 fecal samples revealed several *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[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[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[104].

A Bayesian skyline plot estimated from concatenated *gag*, *int* and *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[105]. Although replication assays have found group O to be significantly less "fit" than group M[106], these *in vitro* differences are not reflected in *in vivo* differences in viral load[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 (*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[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[110,111]. In this and other populations, HIV-2 prevalence consistently peaked in older age groups[112,113]; 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[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[114]. We estimated a Bayesian skyline plot from HIV-2 *env* sequences sampled in northwestern Guinea-

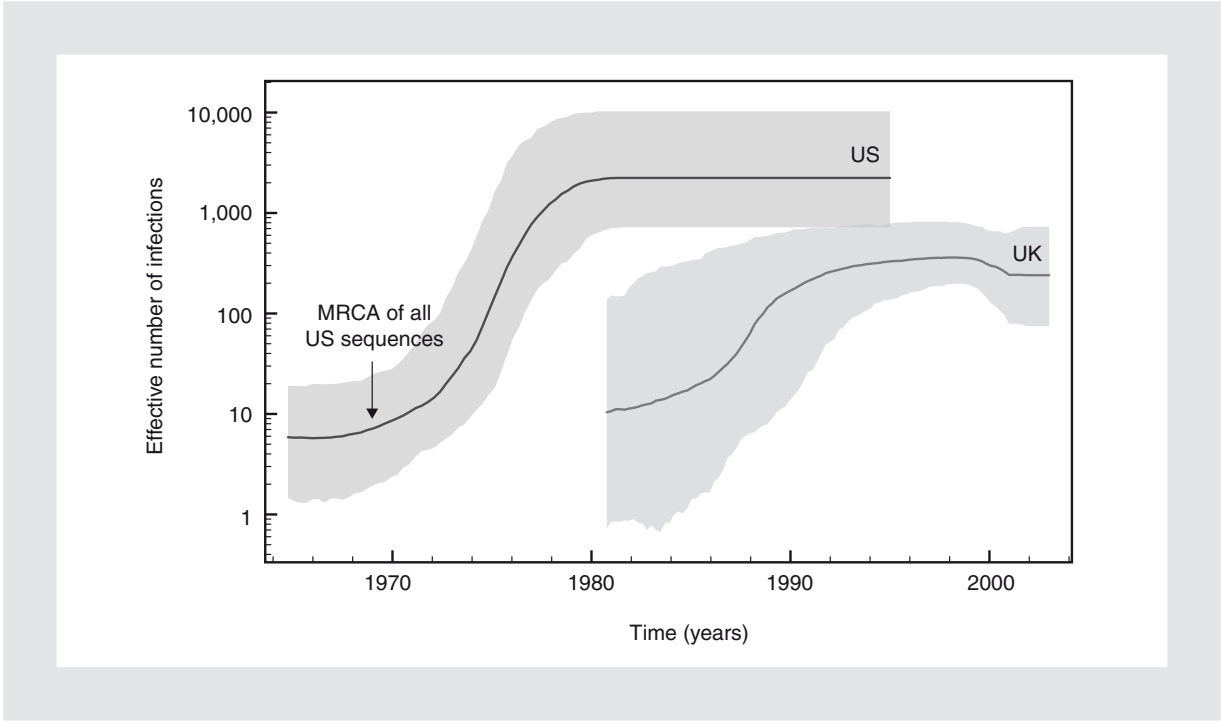

Figure 5. *Bayesian skyline plots for HIV-1 subtype B. The skyline plot representing the U.S. epidemic history was reconstructed using the serially sampled data set analyzed in Robbins K, et al.[35]. The date for the most recent common ancestor of sequences from US origin is indicated with an arrow. The UK skyline plot was inferred using the largest cluster ('Cluster 2') identified in Hue, et al. 2005[118]; the same prior distribution for the evolutionary rate was used in our analysis.*

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[129]. (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[115].

Although subtype distributions in different countries are continually changing[116], 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[117].

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)[118]. 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 (Fig. 4 and 5), which could be at least partly attributed

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[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[119,120], estimated rates of superinfection are generally very low[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 *ad hoc* way to deal with, or rather ignore, the problem of 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 data[70,105,124]. It is therefore not surprising that the assumption of a single phylogenetic history across the genome, or unlinked phylogenies for different genes, did not lead to large differences in the demographic estimates for HIV-1 group O[105].

(iv) Recombination that occurs among lineages within a single infection will not bias the topology of an among-host phylogeny. Such recombination may increase the variance in evolutionary rate among lineages, but this can now be adequately modeled using relaxed clock approaches.

In the context of our point (i) above, an interesting application of estimating intra-subtype population recombination rates was provided by Taylor and Korber (2004)[125]. They simulated sequence data using a structured coalescent model with recombination, reflecting transmission dynamics with varying levels of superinfection. By comparing population recombination rates inferred from these simulations with estimates from real sequence data, they concluded that superinfection rates might be as high as 15% of infections. This contrasts with much lower rates observed in epidemiologic surveys[121-123]. This difference could be the result of assumptions made in the simulations, such as constant numbers of infected individuals, homogeneous substitution rates among sites, neutral evolution, or epidemiologic heterogeneity[125]. Small networks of individuals, with large superinfection rates relative to the total population, can severely impact genealogic estimates[125]. Recent findings of a significantly higher frequency of dual infections in high-risk populations seem to confirm this[126]. Interestingly, simulation studies have specifically assessed the impact of epidemiologic mixing patterns on demographic inference[127]. Although the networks of behavior that spread HIV can affect the relationship between N_e and census population size, parametric models for estimating growth rate – the parameter of interest from an epidemiologic perspective – seem to be highly robust to violations of panmixis, even for small samples[127]. More research is needed, however, to evaluate the influence of social network structures together with geographic distance on HIV diversity in larger populations.

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 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[14,15]. In addition, selectively advantageous mutations might "miss the boat" for transmission if they occur late in infection[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%)[47]. Whether HIV transmission is selectively neutral is, however, still the subject of debate[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)[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

1. CDC. Pneumocystis pneumonia–Los Angeles. Morb Mortal Wkly Rep 1981;30:250-2.
2. Mansky L. Forward mutation rate of HIV-1 in a T lymphoid cell line. AIDS Res Hum Retroviruses 1996;12:307-14.
3. Mansky L, Temin H. Lower *in vivo* mutation rate of HIV-1 than that predicted from the fidelity of purified reverse transcriptase. J Virol 1995;69:5087-94.
4. Ho D, Neumann A, Perelson A, Chen W, Leonard J, Markowitz M. Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. Nature 1995;373:123-6.
5. Perelson A, Neumann A, Markowitz M, Leonard J, Ho D. HIV-1 dynamics *in vivo*: virion clearance rate, infected cell life-span, and viral generation time. Science 1996;271:1582-6.
6. Wei X, Ghosh S, Taylor M, et al. Viral dynamics in HIV-1 infection. Nature 1995;373:117-22.
7. Drummond A, Rambaut A, Shapiro B, Pybus O. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 2005;22:1185-92.
8. Edwards C, Holmes E, Wilson D, et al. HIV-1 envelope gene evolution during chronic infection is deterministic and dominated by negative selection. Genetics (In press).
9. Ewing G, Nicholls G, Rodrigo A. Using temporally spaced sequences to simultaneously estimate migration rates, mutation rate and population sizes in measurably evolving populations. Genetics 2004;168:2407-20.
10. Wilson D, McVean G. Estimating diversifying selection and functional constraint in the presence of recombination. Genetics 2006; 172:1411-25.
11. Drummond A, Pybus O, Rambaut A, Forsberg R, Rodrigo A. Measurably evolving populations. Trends in Ecology and Evolution 2003;18:481-8.
12. Grenfell B, Pybus O, Gog J, et al. Unifying the epidemiological and evolutionary dynamics of pathogens. Science 2004;303:327-32.
13. Frost S, Dumaurier M, Wain-Hobson S, Brown A. Genetic drift and within-host metapopulation dynamics of HIV-1 infection. Proc Natl Acad Sci USA 2001;98:6975-80.
14. Holmes E. The phylogeography of human viruses. Mol Ecol 2004; 13:745-56.
15. Rambaut A, Posada D, Crandall K, Holmes E. The causes and consequences of HIV evolution. Nat Rev Genet 2004;5:52-61.
16. John-Stewart G, Nduati R, Rousseau C, et al. Subtype C is associated with increased vaginal shedding of HIV-1. J Infect Dis 2005; 192:492-6.
17. Iversen A, Learn G, Skinhoj P, Mullins J, McMichael A, Rambaut A. Preferential detection of HIV subtype C' over subtype A in cervical cells from a dually infected woman. Aids 2005;19:990-3.
18. Williamson S. Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression. Mol Biol Evol 2003;20:1318-25.
19. Frost S, Nijhuis M, Schuurman R, Boucher CA, Brown A. Evolution of lamivudine resistance in HIV1-infected individuals: the relative roles of drift and selection. J Virol 2000;74:6262-8.
20. Pybus O. Inferring evolutionary and epidemiologic processes from molecular phylogenies. University of Oxford, Oxford 2000.
21. Swofford D, Olsen G, Waddell P, Hillis D. Phylogenetic inference. In: Molecular Systematics. Ed. Hillis D, Moritz C, Mable B: Sinauer Associates 1996:407-514.
22. Salemi M, Vandamme A-M. The phylogenetic handbook: a practical approach to DNA and protein phylogeny. Cambridge University Press 2003.
23. Wilson D, Falush D, McVean G. Germs, genomes and genealogies. TRENDS in Ecology and Evolution 2005;20:39-45.
24. Drummond A, Pybus O, Rambaut A. Inference of viral evolutionary rates from molecular sequences. Advances In Parasitology 2003; 54:331-58.
25. Rambaut A. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics 2000;16:395-9.
26. Drummond A, Nicholls G, Rodrigo A, Solomon W. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 2002;161:1307-20.
27. Jenkins G, Rambaut A, Pybus O, Holmes E. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. J Mol Evol 2002;54:156-65.
28. Sanderson M. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol Biol Evol 1997;14:1218-31.
29. Yoder A, Yang Z. Estimation of primate speciation dates using local molecular clocks. Mol Biol Evol 2000;17:1081-90.
30. Aris-Brosou S, Yang Z. Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. Syst Biol 2002;51:703-14.
31. Drummond A, Ho S, Phillips M, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biol 2006;4.

32. Thorne J, Kishino H, Painter I. Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol 1998;15:1647-57.
33. Frost S, Liu Y, Pond S, et al. Characterization of HIV-1 envelope variation and neutralizing antibody responses during transmission of HIV-1 subtype B. J Virol 2005;79:6523-7.
34. Shankarappa R, Margolick J, Gange S, et al. Consistent viral evolutionary changes associated with the progression of HIV-1 infection. J Virol 1999;73:10489-502.
35. Robbins K, Lemey P, Pybus O, et al. U.S. HIV-1 epidemic: date of origin, population history, and characterization of early strains. J Virol 2003;77:6359-66.
36. Herbeck J, Nickle D, Learn G, et al. HIV-1 env evolves toward ancestral states upon transmission to a new host. J Virol 2006;80:1637-44.
37. Kingman J. The coalescent. Stochastic Processes and their Applications 1982;13:235-48.
38. Kingman J. On the genealogy of large populations. J Appl Probab 1982;19A:27-43.
39. Hudson R. Gene genealogies and the coalescent process. In: Oxford Surveys in Evolutionary Biology. Ed. Futuyama D, Antonovics J. Oxford University Press 1990.
40. Griffiths R, Marjoram P. Ancestral inference from samples of DNA sequences with recombination. J Comput Biol 1996,3:479-502.
41. Nath H, Griffiths R. The coalescent in two colonies with symmetric migration. J Math Biol 1993;31:841-51.
42. Slatkin M, Hudson R. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 1991;129:555-62.
43. Griffiths R, Tavare S. Sampling theory for neutral alleles in a varying environment. Philos Trans. R Soc Lond B Biol Sci 1994; 344:403-10.
44. Rodrigo A, Felsenstein J. Coalescent approaches to HIV population genetics. In: The Evolution of HIV. Ed. Crandall K. Baltimore: John Hopkins University Press 1999.
45. Leitner T, Fitch W. The Phylogenetics of Known Transmission Histories. In: The Evolution of HIV. Ed. Crandall K. Baltimore: Johns Hopkins University Press 1999:315-45.
46. Slowinski J, Page R. How should species phylogenies be inferred from sequence data? Syst Biol 1999;48:814-25.
47. Edwards C, Holmes E, Wilson D, et al. Population genetic estimation of the loss of genetic diversity during horizontal transmission of HIV-1. BMC Evol Biol 2006;6:28.
48. Derdeyn C, Decker J, Bibollet-Ruche F, et al. Envelope-constrained neutralization-sensitive HIV-1 after heterosexual transmission. Science 2004;303:2019-22.
49. Wolinsky S, Wike C, Korber B, et al. Selective transmission of HIV-1 variants from mothers to infants. Science 1992;255:1134-7.
50. Drummond A, Rambaut A. BEAST v1.3, Available from http://evolve.zoo.ox.ac.uk/beast/. 2003.
51. Haase A, Henry K, Zupancic M, et al. Quantitative image analysis of HIV-1 infection in lymphoid tissue. Science 1996;274:985-9.
52. Brown A. Analysis of HIV-1 env gene sequences reveals evidence for a low effective number in the viral population. Proc Natl Acad Sci USA 1997;94:1862-5.
53. Shriner D, Shankarappa R, Jensen M, et al. Influence of random genetic drift on HIV-1 env evolution during chronic infection. Genetics 2004;166:1155-64.
54. Rouzine I, Coffin J. Linkage disequilibrium test implies a large effective population number for HIV *in vivo*. Proc Natl Acad Sci USA 1999;96:10758-63.
55. Bonhoeffer S, Holmes E, Nowak M. Causes of HIV diversity. Nature 1995;376:125.
56. Yamaguchi Y, Gojobori T. Evolutionary mechanisms and population dynamics of the third variable envelope region of HIV within single hosts. Proc Natl Acad Sci USA 1997;94:1264-9.
57. Anisimova M, Nielsen R, Yang Z. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 2003;164:1229-36.
58. Shriner D, Nickle D, Jensen M, Mullins J. Potential impact of recombination on sitewise approaches for detecting positive natural selection. Genet Res 2003;81:115-21.
59. McDonald J, Kreitman M. Adaptive protein evolution at the Adh locus in drosophila. Nature 1991;351:652-4.
60. Nielsen R. Changes in ds/dn in the HIV-1 env gene. Mol Biol Evol 1999;16:711-4.
61. Yuste E, Moya A, Lopez-Galindez C. Frequency-dependent selection in HIV-1. J Gen Virol 2002;83:103-6.
62. Holmes E, Zhang L, Simmonds P, Ludlam C, Brown A. Convergent and divergent sequence evolution in the surface envelope glycoprotein of HIV-1 within a single infected patient. Proc Natl Acad Sci USA 1992;89:4835-9.
63. Levy D, Aldrovandi G, Kutsch O, Shaw G. Dynamics of HIV-1 recombination in its natural target cells. Proc Natl Acad Sci USA 2004;101:4204-9.
64. Robertson D, Sharp P, McCutchan F, Hahn B. Recombination in HIV-1. Nature 1995;374:124-6.
65. Jung A, Maier R, Vartanian J, et al. Multiply infected spleen cells in HIV patients. Nature 2002;418:144.
66. McVean G, Awadalla P, Fearnhead P. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 2002;160:1231-41.
67. Shriner D, Rodrigo A, Nickle D, Mullins J. Pervasive genomic recombination of HIV-1 *in vivo*. Genetics 2004;167:1573-83.
68. Stumpf M, McVean G. Estimating recombination rates from population-genetic data. Nat Rev Genet 2003;4:959-68.
69. Kuhner M, Yamato J, Felsenstein J. Maximum likelihood estimation of recombination rates from population data. Genetics 2000;156: 1393-401.
70. Carvajal-Rodriguez A, Crandall K, Posada D. Recombination estimation under complex evolutionary models with the coalescent composite-likelihood method. Mol Biol Evol 2006;23:817-27.
71. Kuhner M. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics 2006;22:768-70.
72. Przeworski M, Charlesworth B, Wall J. Genealogies and weak purifying selection. Mol Biol Evol 1999;16:246-52.
73. Fisher R. The Genetic Theory of Natural Selection. Oxford: Clarendon Press 1930.
74. Bonhoeffer S, Chappey C, Parkin N, Whitcomb J, Petropoulos C. Evidence for positive epistasis in HIV-1. Science 2004;306:1547-50.
75. Wang K, Mittler J, Samudrala R. Comment on "Evidence for positive epistasis in HIV-1". Science 2006;312:848 [author reply 848].
76. Rouzine I, Coffin J. Evolution of HIV under selection and weak recombination. Genetics 2005;170:7-18.
77. Shapshak P, Segal D, Crandall K, et al. Independent evolution of HIV-1 in different brain regions. AIDS Res Hum Retroviruses 1999; 15:811-20.
78. Salemi M, Lamers S, Yu S, de Oliveira T, Fitch W, McGrath M. Phylodynamic analysis of HIV-1 in distinct brain compartments provides a model for the neuropathogenesis of AIDS. J Virol 2005; 79:11343-52.
79. Smit T, Wang B, Ng T, Osborne R, Brew B, Saksena N. Varied tropism of HIV-1 isolates derived from different regions of adult brain cortex discriminate between patients with and without AIDS dementia complex (ADC): evidence for neurotropic HIV variants. Virology 2001;279:509-26.
80. Beerli P, Felsenstein J. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics 1999;152:763-73.
81. Beerli P, Felsenstein J. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc Natl Acad Sci USA 2001; 98:4563-8.
82. Ewing G, Rodrigo A. Coalescent-based estimation of population parameters when the number of demes changes over time. Mol Biol Evol 2006;23:988-96.
83. Bahlo M, Griffiths RC. Inference from gene trees in a subdivided population. Theor Popul Biol 2000;57:79-95.
84. Nielsen R, Wakeley J. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics 2001;158:885-96.
85. Kelly J. Replication rate and evolution in HIV. J Theor Biol 1996; 180:359-64.

86. Cavert W, Notermans D, Staskus K, et al. Kinetics of response in lymphoid tissues to antiretroviral therapy of HIV-1 infection. Science 1997;276:960-4.
87. Finzi D, Hermankova M, Pierson T, et al. Identification of a reservoir for HIV-1 in patients on HAART. Science 1997;278:1295-300.
88. Strain M, Gunthard H, Havlir D, et al. Heterogeneous clearance rates of long-lived lymphocytes infected with HIV: intrinsic stability predicts lifelong persistence. Proc Natl Acad Sci USA 2003;100:4819-24.
89. Kelly J, Williamson S, Orive M, Smith M, Holt R. Linking dynamical and population genetic models of persistent viral infection. Am Nat 2003;162:14-28.
90. Gao F, Bailes E, Robertson D, et al. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. Nature 1999;436-41.
91. Corbet S, Muller-Trutwin M, Versmisse P, et al. env sequences of SIV from chimpanzees in Cameroon are strongly related to those of HIV group N from the same geographic area. J Virol 2000;74:529-34.
92. Hahn B, Shaw G, De Cock K, Sharp P. AIDS as a zoonosis: scientific and public health implications. Science 2000;287:607-14.
93. Korber B, Muldoon M, Theiler J, et al. Timing the ancestor of the HIV-1 pandemic strains. Science 2000;288:1789-96.
94. Salemi M, Strimmer K, Hall W, et al. Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes using a new method to uncover clock-like molecular evolution. FASEB J 2001;15:276-8.
95. Rambaut A, Robertson D, Pybus O, Peeters M, Holmes E. HIV. Phylogeny and the origin of HIV-1. Nature 2001;410:1047-8.
96. Yusim K, Peeters M, Pybus O, et al. Using HIV-1 sequences to infer historical features of the AIDS epidemic and HIV evolution. Philos Trans R Soc Lond B Biol Sci 2001;356:855-66.
97. Strimmer K, Pybus O. Exploring the demographic history of DNA sequences using the generalized skyline plot. Mol Biol Evol 2001; 18:2298-305.
98. Vidal N, Peeters M, Mulanga-Kabeya C, et al. Unprecedented degree of HIV-1 group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. J Virol 2000;74:10498-507.
99. Kalish M, Robbins K, Pieniazek D, et al. Recombinant viruses and early global HIV-1 epidemic. Emerg Infect Dis 2004;10:1227-34.
100. Vidal N, Mulanga C, Bazepeo S, et al. Distribution of HIV-1 variants in the Democratic Republic of Congo suggests increase of subtype C in Kinshasa between 1997 and 2002. J Acquir Immune Defic Syndr 2005;40:456-62.
101. Worobey M, Santiago M, Keele B, et al. Origin of AIDS: contaminated polio vaccine theory refuted. Nature 2004;428:820.
102. Keele B, van Heuverswyn F, Li Y, et al. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. Science 2006;313:523-6.
103. Chitnis A, Rawls D, Moore J. Origin of HIV type 1 in colonial French Equatorial Africa? AIDS Res Hum Retroviruses 2000;16:5-8.
104. Roques P, Robertson D, Souquiere S, et al. Phylogenetic analysis of 49 newly derived HIV-1 group O strains: high viral diversity but no group M-like subtype structure. Virology 2002;302:259-273.
105. Lemey P, Pybus O, Rambaut A, et al. The molecular population genetics of HIV-1 group O. Genetics 2004;167:1059-68.
106. Arien K, Abraha A, Quinones-Mateu M, Kestens L, Vanham G, Arts E. The replicative fitness of primary HIV-1 group M, HIV-1 group O, and HIV-2 isolates. J Virol 2005;79:8979-90.
107. Gueudin M, Plantier J, Damond F, Roques P, Mauclere P, Simon F. Plasma viral RNA assay in HIV-1 group O infection by real-time PCR. J Virol Methods 2003;113:43-9.
108. Shanmugam V, Switzer W, Nkengasong J, et al. Lower HIV-2 plasma viral loads may explain differences between the natural histories of HIV-1 and HIV-2 infections. J Acquir Immune Defic Syndr 2000; 24:257-63.
109. O'Donovan D, Ariyoshi K, Milligan P, et al. Maternal plasma viral RNA levels determine marked differences in mother-to-child transmission rates of HIV-1 and HIV-2 in The Gambia. MRC/Gambia Government/University College London Medical School working group on mother-child transmission of HIV. Aids 2000;14:441-8.
110. Wilkins A, Ricard D, Todd J, Whittle H, Dias F, Paulo Da Silva A. The epidemiology of HIV infection in a rural area of Guinea-Bissau. AIDS 1993;7:1119-22.
111. Ricard D, Wilkins A, N'Gum P, et al. The effects of HIV-2 infection in a rural area of Guinea-Bissau. AIDS 1994;8:977-82.
112. Poulsen A, Aaby P, Jensen H, Dias F. Risk factors for HIV-2 seropositivity among older people in Guinea-Bissau. A search for the early history of HIV-2 infection. Scand J Infect Dis 2000; 32:169-75.
113. Poulsen A, Kvinesdal B, Aaby P, et al. Prevalence of and mortality from HIV-2 in Guinea-Bissau, West Africa. Lancet 1989;1:827-31.
114. Lemey P, Pybus O, Wang B, Saksena N, Salemi M, Vandamme A-M. Tracing the origin and history of the HIV-2 epidemic. Proc Natl Acad Sci USA 2003;100:6588-92.
115. Salemi M, de Oliveira T, Soares M, et al. Different epidemic potentials of the HIV-1B and C subtypes. J Mol Evol 2005;60:598-605.
116. Snoeck J, Van Laethem K, Hermans P, et al. Rising prevalence of HIV-1 non-B subtypes in Belgium: 1983-2001. J Acquir Immune Defic Syndr 2004;35:279-85.
117. Robertson J, Bucknall A, Welsby P, et al. Epidemic of AIDS related virus (HTLV-III/LAV) infection among intravenous drug abusers. Br Med J (Clin Res Ed) 1986;292:527-9.
118. Hue S, Pillay D, Clewley J, Pybus O. Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. Proc Natl Acad Sci USA 2005;102:4425-9.
119. Altfeld M, Allen T, Yu X, et al. HIV-1 superinfection despite broad CD8+ T-cell responses containing replication of the primary virus. Nature 2002;420:434-9.
120. Najera R, Delgado E, Perez-Alvarez L, Thomson M. Genetic recombination and its role in the development of the HIV-1 pandemic. AIDS 2002;16(suppl 4):S3-16.
121. Gonzales M, Delwart E, Rhee S, et al. Lack of detectable HIV-1 superinfection during 1072 person-years of observation. J Infect Dis 2003;188:397-405.
122. Tsui R, Herring B, Barbour J, et al. HIV-1 superinfection was not detected following 215 years of injection drug user exposure. J Virol 2004;78:94-103.
123. Chakraborty B, Valer L, De Mendoza C, Soriano V, Quinones-Mateu M. Failure to detect HIV-1 superinfection in 28 HIV-seroconcordant individuals with high risk of reexposure to the virus. AIDS Res Hum Retroviruses 2004;20:1026-31.
124. McVean G. A genealogical interpretation of linkage disequilibrium. Genetics 2002;162:987-91.
125. Taylor J, Korber B. HIV-1 intra-subtype superinfection rates: estimates using a structured coalescent with recombination. Infect Genet Evol 2005;5:85-95.
126. Herbinger K, Gerhardt M, Piyasirisilp S, et al. Frequency of HIV-1 dual infection and HIV diversity: analysis of low- and high-risk populations in Mbeya region, Tanzania. AIDS Res Hum Retroviruses 2006;22:599-606.
127. Goodreau S. Assessing the effects of human mixing patterns on HIV-1 interhost phylogenetics through social network simulation. Genetics 2006;172:2033-45.
128. Leitner T, Albert J. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. Proc Natl Acad Sci USA 1999,96:10752-7.
129. Shi Y, Brandin E, Vincic E, et al. Evolution of human immunodeficiency virus type 2 coreceptor usage, autologous neutralization, envelope sequence and glycosylation. J Gen Virol 2005,86:3385-96.