DOI: 10.1534/genetics.104.026666

The Molecular Population Genetics of HIV-1 Group O

Philippe Lemey,*,[1] Oliver G. Pybus,[†] Andrew Rambaut,[†] Alexei J. Drummond,[†] David L. Robertson,[‡] Pierre Roques,[§,] Michael Worobey[††] and Anne-Mieke Vandamme***

**Rega Institute for Medical Research, KULeuven, B-3000 Leuven, Belgium, [†]Department of Zoology, University of Oxford, Oxford OX1 3PS, United Kingdom, [‡]School of Biological Sciences, University of Manchester, Manchester M13 9PT, United Kingdom, [§]Service de Neurovirologie, CEA, 92260 Fontenay-aux-Roses, France, **Service de Neurovirologie, CIRMF, BP 769 Franceville, Gabon and [††]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson Arizona 85721*

Manuscript received January 21, 2004
Accepted for publication April 5, 2004

ABSTRACT

HIV-1 group O originated through cross-species transmission of SIV from chimpanzees to humans and has established a relatively low prevalence in Central Africa. Here, we infer the population genetics and epidemic history of HIV-1 group O from viral gene sequence data and evaluate the effect of variable evolutionary rates and recombination on our estimates. First, model selection tools were used to specify suitable evolutionary and coalescent models for HIV group O. Second, divergence times and population genetic parameters were estimated in a Bayesian framework using Markov chain Monte Carlo sampling, under both strict and relaxed molecular clock methods. Our results date the origin of the group O radiation to around 1920 (1890–1940), a time frame similar to that estimated for HIV-1 group M. However, group O infections, which remain almost wholly restricted to Cameroon, show a slower rate of exponential growth during the twentieth century, explaining their lower current prevalence. To explore the effect of recombination, the Bayesian framework is extended to incorporate multiple unlinked loci. Although recombination can bias estimates of the time to the most recent common ancestor, this effect does not appear to be important for HIV-1 group O. In addition, we show that evolutionary rate estimates for different HIV genes accurately reflect differential selective constraints along the HIV genome.

THE human immunodeficiency virus type 1 (HIV-1) comprises three lineages, denoted M, N, and O, resulting from at least three separate introductions of simian immunodeficiency virus from chimpanzees (SIVcpz) into the human population (Gao *et al.* 1999; Corbet *et al.* 2000). The vast majority of HIV-1 infections worldwide belong to HIV-1 group M. In contrast, HIV-1 group O appears to be restricted to west-central Africa. A third (and very rare) lineage, group N, has been identified in Cameroon (Simon *et al.* 1998). A number of studies have used coalescent and molecular clock methods to understand the epidemic history of HIV-1. Molecular clock analyses have dated the most recent common ancestor of group M to around 1930 (Korber *et al.* 2000; Salemi *et al.* 2001; Sharp *et al.* 2001). Coalescent analysis of the HIV-1 group M population dynamics in Central Africa, where group M originated, indicated a history of exponential growth with an increasing exponential growth rate through time (Yusim *et al.* 2001). Similar analyses of HIV-2 in Guinea-Bissau revealed a transition from endemic infection to exponential growth during the independence war (Lemey *et al.* 2003). Our current knowledge of HIV-1 group O epidemiology is more limited and based primarily on serological studies (Nkengasong *et al.* 1993; Peeters *et al.* 1997; Zekeng *et al.* 1997; Ayouba *et al.* 2001). Group O was first identified in 1994 in Cameroon (Gurtler *et al.* 1994), which still appears to have the highest group O prevalence, albeit low compared to that of group M (Nkengasong *et al.* 1993; Peeters *et al.* 1997). Sporadic cases of group O infections have been documented in Europe and the United States (*e.g.*, De Leys *et al.* 1990; Charneau *et al.* 1994; Rayfield *et al.* 1996); however, the majority of these individuals had contacts with west-central Africa (Quinones-Mateu *et al.* 2000). Recent sequencing efforts have provided more detailed information about the phylogenetic relationships of group O strains (Roques *et al.* 2002; Yamaguchi *et al.* 2002). However, the origin and demographic history of this HIV-1 variant have not yet been elucidated.

Population genetic modeling provides a way to extract information about evolutionary and population genetic processes from sampled gene sequences. Major advances in this field have been made through use of the coalescent, a mathematical model that describes the

[1]*Corresponding author:* Rega Institute for Medical Research, Minderbroedersstraat 10, B-3000 Leuven, Belgium.
E-mail: philippe.lemey@uz.kuleuven.ac.be

statistical properties of the ancestral history of sampled sequences (Kingman 1982; Hudson 1990). This shared ancestry is usually formalized as a genealogy, or tree, which can be reconstructed using standard phylogenetic methods. The standard neutral coalescent model has been extended to uncover the evolutionary footprints of many processes, such as recombination (*e.g.*, Hudson 1990; Griffiths and Marjoram 1996), population subdivision (*e.g.*, Nath and Griffiths 1993), and variable population size (*e.g.*, Slatkin and Hudson 1991; Griffiths and Tavaré 1994). In the latter case, the coalescent model relates the shape of the genealogy to the demographic history of the sampled population.

If the gene sequences have been sampled from infectious organisms present in different individuals, then the coalescent model provides information about epidemic history or, more specifically, about the historical dynamics of the number of infected individuals (Holmes *et al.* 1995; Pybus *et al.* 2000). The statistical inference of pathogen epidemic history is aided by the use of "heterochronous" data—sequences that have been sampled at sufficiently different points in time that mutations have accumulated between those times (Drummond *et al.* 2003). Heterochronous sequences allow effective population size (N_e) and the rate of molecular evolution (μ) to be independently estimated from sequence data. In contrast, "isochronous" sequences—those that have been sampled at effectively the same point in time—contain only information about the composite parameter $\theta = 2N_e\mu$. Various statistical frameworks can be used to infer population genetic parameters from gene sequences (*e.g.*, Kuhner *et al.* 1998; Beerli and Felsenstein 2001; McVean *et al.* 2002), a subset of which can accommodate heterochronous data (*e.g.*, Rambaut 2000; Drummond and Rodrigo 2000; Pybus and Rambaut 2002; Seo *et al.* 2002). Most recently, Drummond *et al.* (2002) introduced a Bayesian approach to heterochronous data that uses Metropolis-Hastings Markov chain Monte Carlo (MCMC) sampling to integrate over different coalescent trees, thereby incorporating phylogenetic uncertainty. The feasibility of this approach has recently been demonstrated in a number of settings (see Drummond *et al.* 2003), including an analysis of hepatitis C virus in Egypt, the results of which correctly matched substantial *a priori* information about the epidemic history of the virus in that country (Pybus *et al.* 2003).

The analyses of genetic diversity described above have typically made strong evolutionary assumptions, particularly regarding recombination and the molecular clock, and the quantitative effect of these assumptions on parameter estimates is largely unknown. Most analyses assume a constant-rate molecular clock. Unfortunately, this hypothesis is frequently rejected for HIV sequence data (Korber *et al.* 1998; Salemi *et al.* 2001). More generally, only 7 of 50 data sets from different RNA virus species complied with a strict molecular clock in a recent comprehensive study (Jenkins *et al.* 2002), although simulations suggest that estimated evolutionary rates can be reliable even when the strict clock is rejected, provided that rate heterogeneity among lineages is small (Jenkins *et al.* 2002). To accommodate for evolutionary rate variation among lineages, Thorne *et al.* (1998) proposed a parametric model for relaxing the clock that assumes autocorrelated rates across speciation/coalescence events. A variant of this method was applied to HIV-1 group M by Korber *et al.* (2000), resulting in estimates similar to those obtained under a strict molecular clock. This model has also been extended to data sets consisting of multiple gene sequences for each taxon of interest (Thorne and Kishino 2002).

Recombination is a frequent event in the evolution of HIV (Robertson *et al.* 1995), giving rise to a multitude of mosaic genomes, some of which are significant in the pandemic and termed "circulating recombinant forms" (Robertson *et al.* 2000). Coestimation of recombination rates, varying population sizes, substitution rates, and complex substitution models within a coalescent framework is expected to be technically challenging and no algorithms are currently available for this task. Previous coalescent analyses of HIV epidemic history have therefore assumed no recombination within the genome fragment under investigation. Frequent recombination will result in different phylogenies along the HIV genome and, at its most extreme, will lead to a total loss of linkage between genes. Worobey (2001) showed that if a single tree is estimated from recombining sequences then estimates of rate heterogeneity among sites are biased upward and the terminal tree branches are lengthened, resulting in a possible overestimation of the TMRCA and a possible demographic bias toward exponential growth. More detailed simulations by Schierup and Forsberg (2003) have confirmed this effect. The impact of these effects on demographic estimates has yet to be quantified. Furthermore, recombination—even at small levels—leads to a rejection of the molecular clock (Schierup and Hein 2000b).

The objective of this study is to investigate the population genetics and epidemic history of HIV-1 group O and examine the robustness of our estimates to variable evolutionary rates among lineages and recombination. In the first part, we select model components and test null hypotheses to specify a suitable coalescent framework. In the second part, we use MCMC methods to estimate the time to the most recent common ancestor (TMRCA), substitution rates, and population parameters. The effect of variable evolutionary rates on divergence time estimates is evaluated by comparing a strict molecular clock method against a relaxed molecular clock method. In some of the analyses it was necessary to use empirical priors for the TMRCA to estimate other parameters of interest. The final part explores the effect of recombination by implementing a model of unlinked

TABLE 1

Characteristics of the compiled data sets

	Genomic position[a]	Strains	Sites	Informative sites	Sampling time spread	% sampled in 1998–1999
env	6273–8859	46	2382	1046	1987–1999	63
envC2gp41	6999–8109	66	1245	589	1987–1999	44
envC2V3	7020–7328	92	291	222	1987–2000	38
envgp41	7959–8270	131	285	136	1987–2000	29
gag	1238–1918	47	681	208	1987–1999	64
gagp24	1304–1669	90	363	139	1987–2000	38
int	4285–5148	43	864	183	1987–1999	70

[a] The genomic position is numbered according to ANT70 (accession no. L20587).

loci in the Bayesian framework that allows different genes to have different genealogies. This provides an upper bound for the effects of recombination, to compare with the lower bound provided by assuming all genes are linked and share the same genealogy. Overall, our results show that HIV-1 group O evolution has a similar timescale to that of group M, but with a slower increase in population size: we estimate the number of group O infections has doubled approximately every 9 years.

BAYESIAN INFERENCE OF HIV-1 GROUP O POPULATION GENETIC PARAMETERS

Null hypothesis testing and model selection: The majority of sequences investigated here originate from the comprehensive studies of Yamaguchi *et al.* (2002) and Roques *et al.* (2002). In addition, we included all database sequences for which sampling dates were available. To maximize both the number of strains and sequence length used, different data sets were compiled in three major gene regions (Table 1). This is necessary since, for a lot of strains, only small stretches within *gag* and *env* have been sequenced. Previously identified recombinants within a single gene (Roques *et al.* 2002) were omitted and only one strain was included from groups of linked infections. Since almost all strains have been sampled in Cameroon, the data represent mainly Cameroonian group O diversity. Although several group O strains were isolated from patients in France, the majority of these individuals were born in Cameroon (Roques *et al.* 2002).

Sequences were aligned in CLUSTAL W (Thompson *et al.* 1994) and manually edited according to their codon-reading frame in Se-Al (http://evolve.zoo.ox.ac.uk). Appropriate nucleotide substitution models were determined with Modeltest v3.06 on the basis of hierarchical likelihood-ratio testing (Posada and Crandall 1998). Maximum-likelihood phylogenetic trees were reconstructed in PAUP* using a heuristic branch-swapping algorithm (Swofford 1998). Branch lengths under the molecular clock were estimated in PAML (Yang 1997), using the "single-rate dated-tip" constraint (Rambaut 2000). The molecular clock was tested using the likelihood-ratio test and was rejected for all data sets. Exploratory linear regression analyses revealed only weak correlations between genetic divergence and sampling time (data not shown). This is not surprising since the regression method assumes the phylogenetic tree is known without error—this is clearly not the case.

The demographic signal in the *gag*, *env*, *envC2gp41*, and *int* data sets was investigated using generalized skyline plots—nonparametric estimates of effective population size against time (Pybus *et al.* 2000; Strimmer and Pybus 2001). Figure 1 shows the generalized skyline plots; superimposed on the plots are parametric estimates under an exponential growth model (1), obtained using GENIEv3.5. For all genome regions, the exponential model provides a good fit to the data, as evaluated by likelihood-ratio testing.

Although we excluded previously identified recombinants (Roques *et al.* 2002), we also investigated the evidence for further recombination in our data. To test for recombination events between the major gene regions, we compared the maximum-likelihood (ML) tree topology for each gene region against the ML trees for the other gene regions and the concatenated data set. This analysis used 42 strains sequenced in the *gag*, *int*, and *env* regions and employed the Kishino-Hasegawa (KH) test (Kishino and Hasegawa 1989), the Shimodaira-Hasegawa (SH) test (Shimodaira and Hasegawa 1999), and the approximately unbiased (AU) test (Shimodaira 2002). For each genome region, the inferred ML tree appeared to be significantly better than the topologies reconstructed for the remaining genome regions and this result is highly consistent among different tests (Table 2). However, the ML tree for the concatenated data is not rejected for the *env* and *gag* regions. Although conflicting phylogenetic signals among gene regions could arise from other evolutionary forces, the results probably indicate significant levels of recombination. For the *gag*, *int*, *env*, and *envC2gp41* data sets, we tested for recombination using the informative sites test (IST; Worobey 2001), which tests whether the ratio of two-state parsimony-informative sites to all polymorphic

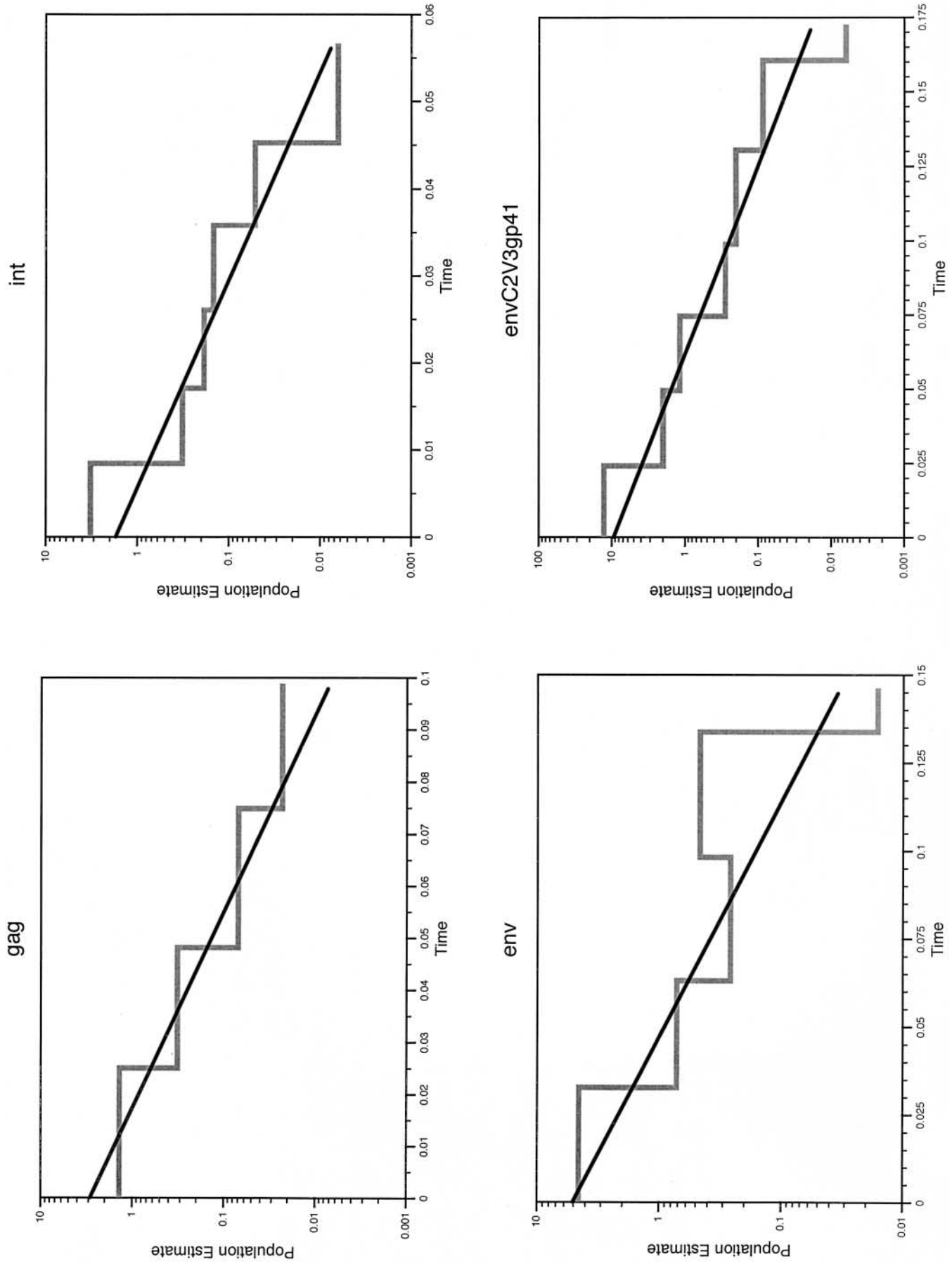
int
Population Estimate
Time
envC2V3gp41
Population Estimate
Time
gag
Population Estimate
Time
env
Population Estimate
Time

TABLE 2

***P*-values for the tree incongruence tests**

	Data								
	gag			*int*			*env*		
Tree	KH	SH	AU	KH	SH	AU	KH	SH	AU
gag	(Best ML tree)			0.002	0.002	0.001	0.001	0.004	<0.001
int	<0.001	<0.001	<0.001	(Best ML tree)			0	0	<0.001
env	0.004	0.008	<0.001	0.011	0.011	0.009	(Best ML tree)		
gagintenv	0.110	0.352	0.110	0.019	0.019	0.019	0.148	0.458	0.160

KH, Kishino-Hasegawa test; SH, Shimodaira-Hasegawa test; AU, approximately unbiased test.

sites is greater than expected from clonally generated data. All gene regions deviated significantly from clonality (Table 3), suggesting that recombination considerably shaped the sequence data in each gene.

Inference of the origin and demographic history of HIV-1 group O: A Bayesian coalescent framework was used for the joint estimation of population parameters, substitution parameters, dates of divergence, and tree topology (Drummond *et al.* 2002). MCMC is used to obtain parameter estimates by averaging over many genealogies and weighting the contribution of each genealogy by its likelihood given the sequence data. Here, we use an improved implementation of this algorithm in the program BEAST (Drummond and Rambaut 2003). In an analysis of this type the likelihood function is composed of both a phylogenetic model and a coalescent model. Analyses were performed using the Hasegawa-Kishino-Yano or general time-reversible substitution models with gamma-distributed rate variation and a proportion of invariable sites, as specified by the Modeltest results. For the complete *env* data set, a codon-position-specific model of rate heterogeneity was used. For the concatenated data set, different genes were given different substitution models. An exponential growth model was used to describe the HIV-1 group O epidemic history, as suggested by the skyline plots. This model is defined as

$$N(t) = N_0 e^{-rt}. \quad (1)$$

The current and ancestral effective numbers of infections are represented by N_0 and $N(t)$, respectively, and r is the exponential growth rate. Three independent MCMC chains were run for 10^7 generations with sampling every 100th generation. The burn-in was set at 10% of the posterior sample. We tested for convergence of the MCMC chains to stationarity as in Drummond *et al.* (2002).

The data exploration step revealed that some data sets were substantially less informative than others about substitution rates. Only for the complete *env* data set and a concatenated data set (consisting of 42 strains sampled in the *gag*, *int*, and *env* regions) were the results consistent among independent MCMC runs in BEAST. For these data sets, divergence dates were also estimated under a relaxed molecular clock using the program MULTIDIVTIME (Thorne *et al.* 1998; Thorne and Kishino 2002). MULTIDIVTIME takes into account both uncertainty in branch length estimation and lineage-specific rate variation, within a Bayesian framework (Thorne and Kishino 2002). For multilocus data, a test for correlated changes in evolutionary rates among genes is provided (Thorne and Kishino 2002). MULTIDIVTIME uses a Metropolis-Hastings MCMC algorithm to sample from the posterior distribution of the model parameters. The mean of the normally distributed prior for the substitution rate was set at 0.003 nucleotide substitutions/site/year for the multilocus data set and to 0.002 nucleotide substitutions/site/year for *env*, both with a standard deviation of 0.001. The mean of the normal prior for the TMRCA was set at 1930 with a standard deviation of 50 years. Different priors on the time to most recent common ancestor (1950 ± 50 years and 1880 ± 50 years) had little influence on the posterior probability (data not shown). The mean of the prior for the Brownian motion constant v was set at 0.02 with standard deviation 0.02. Two independent MCMC chains were run for 10^7 generations with sampling every 100th generation. The burn-in was set after sampling 10^5 generations.

Marginal posterior distributions for the TMRCA of HIV-1 group O are shown in Figure 2. The coalescent method with strict clock (BEAST) and the relaxed clock method (MULTIDIVTIME) produced overlapping marginal posterior densities, with posterior modes close to 1920. This comparison suggests that the effect of variable evolutionary rates on the TMRCA estimate was

Figure 1.—Generalized skyline plots for the *gag*, *int*, *env*, and *envC2gp41* data sets. Nonparametric estimates are shown as shaded lines, whereas the superimposed solid lines indicate parametric estimates under an exponential model.

TABLE 3

Statistics for the informative sites test

	ISI	*P*
gag	0.670588	<0.0001
int	0.375731	<0.0147
env	0.445652	<0.0001
envC2gp41	0.542447	<0.0001

ISI, informative sites index, which takes the value of zero for pure clonality and one for a complete loss of linkage between sites. All tests were performed on the third codon positions of the nonoverlapping gene regions.

limited for HIV-1 group O. Estimates for the concatenated data and the single *env* locus were also very similar. Interestingly, analysis of the concatenated data set revealed no significant correlation among genes of changes in evolutionary rate over time. The rank correlations between *env* and *gag*, *int* and *gag*, and *int* and *env* were 0.27, 0.19, and −0.24, respectively. Coalescent estimates of the population growth rate using BEAST resulted in 0.068 (0.041–0.095) year^{-1} and 0.075 (0.048–0.10) year^{-1} for *env* and the concatenated data, respectively. Collectively, these analyses suggest that group O infections have doubled approximately every 9 years since about 1920.

A multilocus model to evaluate the effect of recombination: Significant topological differences and loss of correlated evolutionary changes among genome regions could be signs of recombination. At the limit, with significant amounts of recombination, the gene loci can be regarded as independent realizations of the coalescent process. In this situation it may be more reasonable to consider a model in which each locus has a different genealogy, but all loci share the same demographic history. If the three genes, *env*, *gag*, and *int*, are represented by a vector of sequence alignments, $\mathbf{D} = \{D_1, D_2, D_3\}$, each associated with a unique genealogy, $\mathbf{G} = \{G_1, G_2, G_3\}$, then the posterior distribution of the substitution rate, μ, and demographic history $\psi = \{N_e, \pi\}$ is defined by

$$P(\mu, \psi|\mathbf{D}) = \int_{\mathbf{G},\Phi} \Pr\{\mathbf{D}|\mathbf{G}, \Phi, \mu\} f(\mathbf{G}|\psi)\, h(\psi, \Phi, \mu), \quad (2)$$

where $\Pr\{\mathbf{D}|\mathbf{G}, \Phi, \mu\} f(\mathbf{G}|\psi) = \prod_{i=1}^{3} \Pr\{D_i|G_i, \Phi, \mu\} f(G_i|\psi)$. This unlinked multilocus model was implemented in BEAST. Although this model allows different loci to have different TMRCAs, it resulted in posterior densities for the TMRCAs that were very similar among genes and with respect to the values obtained assuming a common genealogical history (Figure 2). The estimated growth rate of 0.070 [confidence interval (C.I.) 0.044–0.097] year^{-1} is also similar to the linked loci estimate (see Table 4).

We used the TMRCA results from the multiple unlinked loci analysis as prior distributions for the analysis of the remaining data. In particular, this empirical *a priori* distribution for the TMRCA was used in the MCMC runs for *gag*, *gagp24*, *int*, *envC2gp41*, *envC2V3*, and *envgp41*. The resulting estimates for the substitution rates and demographic parameters are listed in Table 4. Interestingly, growth rate estimates are highly consistent among the different gene regions, while the evolutionary rates vary and appear to reflect differential selective constraints on the HIV-1 genome. To test this more formally, d_N/d_S estimates were obtained under a codon substitution model (Yang *et al.* 2000) and plotted against the evolutionary rate estimates (Figure 3). As can be observed, there was a strong relationship between substitution rates and d_N/d_S estimates ($R^2 = 0.95$, $P < 0.001$). It should be noted that due to overlaps in the particular genome regions used, the data sets cannot be considered as completely independent and thus caution should be exercised when evaluating this relationship.

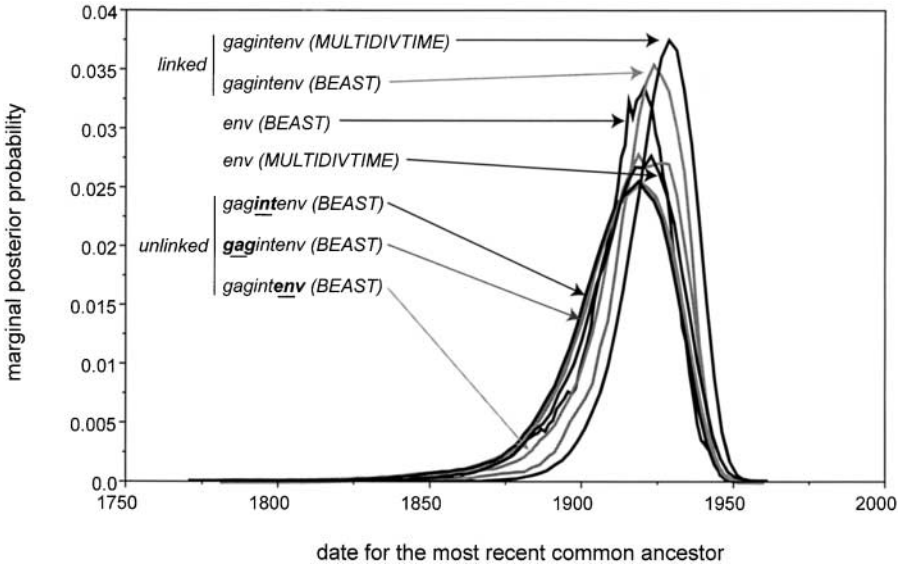

Figure 2.—Marginal posterior distributions for the date of the most recent common ancestor of HIV-1 group O. The Bayesian coalescent analysis and the Bayesian relaxed clock analyses were performed with the programs BEAST and MULTIDIVTIME, respectively. In the unlinked case, TMRCA estimates are shown for the single genes that are indicated in boldface type and underlined in the concatenated data set.

TABLE 4

Estimates of divergence dates, evolutionary rates, and population parameters

	MRCA (yr)	Evolutionary rate (nucleotide substitutions/site/yr)	Effective no. of infections[a]	Exponential growth rate (yr^{-1})
env	1914 (1880–1941)	0.0019 (0.0013–0.0026)	3,992 (1,548–7,130)	0.068 (0.041–0.095)
envC2gp41	1916 (1897–1933)	0.0022 (0.0017–0.0027)	6,135 (2,793–10,140)	0.080 (0.058–0.103)
envC2V3	1912 (1893–1931)	0.0034 (0.0026–0.0044)	9,556 (4,333–15,820)	0.082 (0.061–0.110)
envgp41	1925 (1907–1942)	0.0022 (0.0016–0.0029)	2,769 (1,555–4,222)	0.085 (0.064–1.11)
gag	1924 (1901–1945)	0.0013 (0.0008–0.0019)	3,023 (910–743)	0.082 (0.053–0.12)
gagp24	1914 (1892–1936)	0.0013 (0.0009–0.0018)	11,170 (4,077–20,060)	0.090 (0.063–0.12)
Int	1915 (1891–1937)	0.0009 (0.0006–0.0012)	3,712 (1,177–7,028)	0.072 (0.047–0.10)
gagintenv				
linked	1920 (1896–1942)	*gag*: 0.0011 (0.0009–0.0015) *int*: 0.0007 (0.0005–0.0010) *env*: 0.0022 (0.0015–0.0028)	4,008 (1,416–7,275)	0.075 (0.048–0.10)
unlinked	*gag*: 1912 (1877–1943) *int*: 1911 (1877–1942) *env*: 1917 (1885–1943)	*gag*: 0.0011 (0.0007–0.0015) *int*: 0.0007 (0.0004–0.0010) *env:* 0.0021 (0.0014–0.0029)	3,542 (1,820–5,720)	0.070 (0.044—0.097)

The values listed are the mean estimates with the 95% highest posterior density interval. Estimates for the *env* data set and the concatenated data set were obtained using uniform priors. Informative gamma priors were encoded on the TMRCA for the other data sets: shape is 28.1 and scale is 2.93 for *envC2gp41*, *envC2V3*, and *envgp41*; shape is 25.2 and scale is 3.48 for *gag*, *gagp24*, and *int*.

[a] Product of effective population size and generation time.

DISCUSSION

In this study we investigated the population genetics of HIV-1 group O using a Bayesian coalescent framework. Model selection tools helped us to set up a parameter-rich model that formalized epidemic history in terms of evolutionary and demographic parameters. Recombination analyses suggested that it was invalid to assume a single genealogy across the *gag*, *int*, and *env* genes. Therefore we implemented a model that assumed independent genealogical histories for each gene. This model is at best an approximation; in reality, HIV-1 recombination is likely to generate intermediate levels of linkage among genes. However, the model does allow us to analyze between-gene recombinants and to explore the potential bias of recombination on coalescent parameter estimation. In contrast to a previous simulation study (Schierup and Forsberg 2003), our results suggest that the estimated TMRCA assuming a single genealogy for all loci is not significantly different from that obtained when assuming different genealogies for each locus. Estimates of effective population size and growth rate were also similar for linked and unlinked analyses. A genealogical perspective is useful in interpreting such results (McVean 2002): recombination events in a very rapidly growing population will mostly occur on the terminal branches of the "star-like" sample genealogy. Although recombination among terminal branches will increase variance in the number of mutations on the terminal branches, resulting in the rejection of the molecular clock and an increase in the variance of TMRCA estimates (as shown in Table 3), importantly, it will not systematically bias estimates of the TMRCA in either direction. Thus the simulation results of Worobey (2001) and Schierup and Hein (2000a) are somewhat dependent on the use of "structured" tree topologies that have long internal branches. During phylogenetic reconstruction, recombination events among these internal branches are misinterpreted as homoplasies on terminal branches, thus biasing tree length and TMRCA upward. It could be said that the above argument is circular, since recombination makes structured trees appear more star like, but this is not the case because there is plenty of nongenetic evidence that HIV-1 has grown exponentially and therefore star-like genealogies are to be expected. Thus the quantitative effect of recombination on evolutionary analyses of HIV-1 may be less severe than initially thought (Schierup and Hein 2000a). The similarity of TMRCA estimates among loci also arises from the low variance in coalescence times when populations are rapidly growing.

Although the unlinked analysis provides some assurance that recombination is not strongly biasing the estimates of TMRCA, it would be desirable to employ a model that explicitly accounts for recombination when estimating divergence times for heterochronous sequences. The development of a MCMC framework that could evaluate models of this type would require substantial effort and falls outside the scope of this article.

For some single-gene data sets, the MCMC was not always consistent among runs when uniform priors were used. Specifically, states of low substitution rate and older TMRCAs were not well distinguished from states

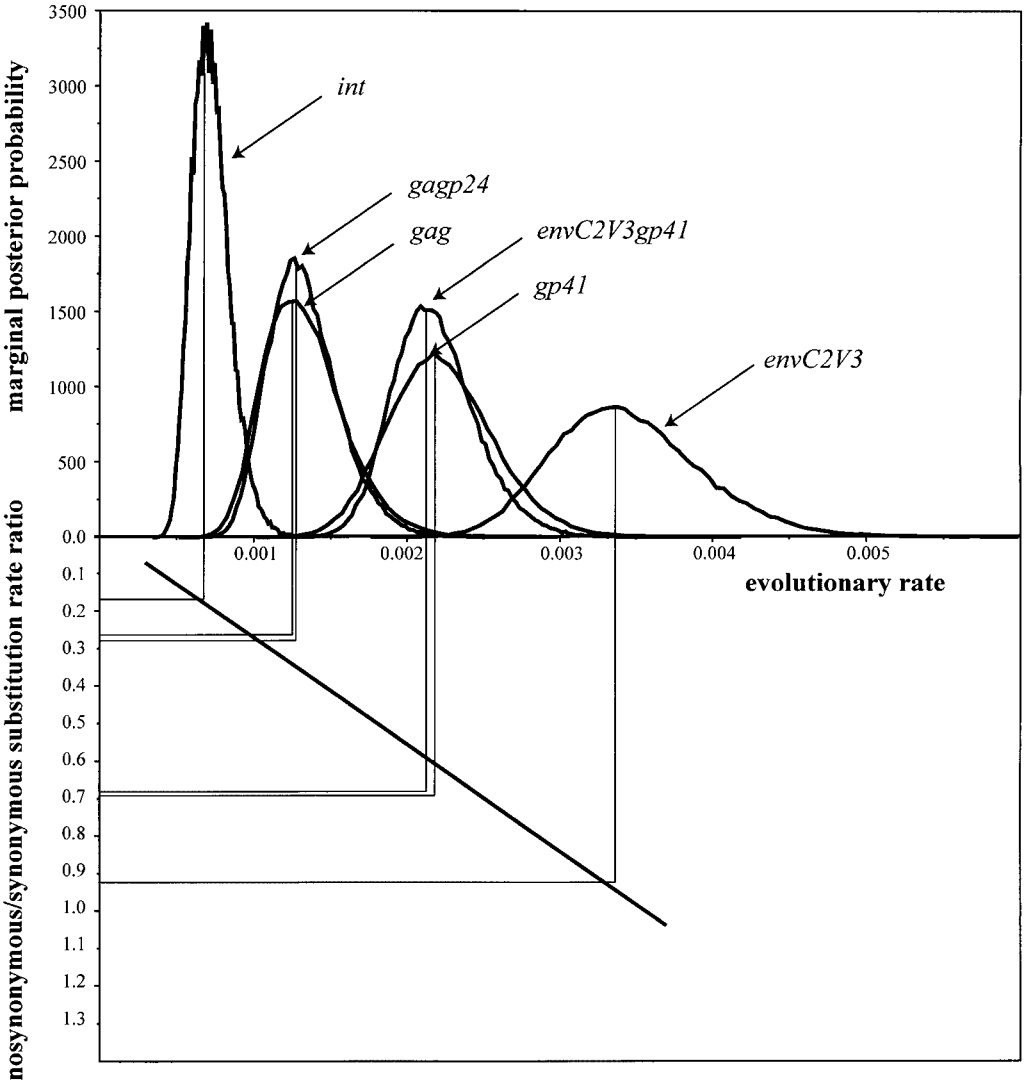

FIGURE 3.—Marginal posterior distributions for the substitution rates and maximum-likelihood estimates for the nonsynonymous/synonymous substitution rate ratios (d_N/d_S) in different genome regions of HIV-1 group O. Marginal posterior distributions for the substitution rates obtained in BEAST using empirical TMRCA priors are shown in the top part. In the bottom part, these substitution rates are plotted against the d_N/d_S estimates obtained using a codon substitution model.

with higher substitution rate and more recent TMRCAs. This has been previously identified as a property of data with weak statistical signal on substitution rates (for a discussion of this problem see DRUMMOND *et al.* 2002). In these cases, there is a problem of identifiability, such that the population size and substitution rate cannot be independently estimated and only estimation of their product is straightforward. To resolve this, we reduced the uncertainty on the substitution rate by using an empirical prior for the TMRCA, obtained from the multilocus analysis. It should be emphasized that this does not represent subjective prior knowledge but "data-residing" prior knowledge. In this situation, formalizing prior knowledge is equivalent to adding extra data.

The MRCA for HIV-1 group O was estimated to have existed around 1920 (1890–1940), in the same range as the TMRCA of group M (KORBER *et al.* 2000; SALEMI *et al.* 2001; SHARP *et al.* 2001). Evolutionary rate estimates for group O (*env*, 0.0019, C.I. 0.0013–0.0026) are also similar to previous estimates for group M (*env*, 0.0024, C.I. 0.0018–0.0028; KORBER *et al.* 2000). These estimates appear to agree with our knowledge of group M and O diversity. Several investigators have reported a similar diversity for both groups (CHARNEAU *et al.* 1994; LOUSSERT-AJAKA *et al.* 1995; KORBER *et al.* 1996; HACKETT *et al.* 1997). A more detailed analysis of larger data sets showed a somewhat higher diversity for group O, which led to the suggestion that 1930 is an upper limit for the MRCA of group O (ROQUES *et al.* 2002). Our point estimates might indeed suggest that group O is slightly older, but the wide overlapping confidence intervals are inconclusive. The effective number of HIV-1 group O infections has been increasing exponentially with a growth rate (r) of 0.08 (0.05–0.12). This is slower than the growth rate estimated for HIV-1 group M in Central Africa ($r = 0.17$; YUSIM *et al.* 2001). Not surprisingly, group O prevalence is much lower than that of group M at present in Cameroon (MAUCLERE *et al.* 1997).

In conclusion, the methods we have used here present a framework that goes some way toward a more realistic

description of viral evolution. In particular, several confounding factors such as substitution rate variation among lineages and recombination are considered. This might be essential when genetic data are used to investigate relevant features from epidemics of infectious diseases.

This work was supported by the Flemish Fonds voor Wetenschappelijk Onderzoek (FWO G.0288.01); P.L. was supported by the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT); O.G.P. was supported by the Wellcome Trust. A.R. was supported by the Royal Society.

LITERATURE CITED

Ayouba, A., P. Mauclere, P. M. Martin, P. Cunin, J. Mfoupouendoun *et al.*, 2001 HIV-1 group O infection in Cameroon, 1986 to 1998. Emerg. Infect. Dis. **7:** 466–467.

Beerli, P., and J. Felsenstein, 2001 Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc. Natl. Acad. Sci. USA **98** (8): 4563–4568.

Charneau, P., A. M. Borman, C. Quillent, D. Guetard, S. Chamaret *et al.*, 1994 Isolation and envelope sequence of a highly divergent HIV-1 isolate: definition of a new HIV-1 group. Virology **205:** 247–253.

Corbet, S., M. C. Muller-Trutwin, P. Versmisse, S. Delarue, A. Ayouba *et al.*, 2000 env sequences of simian immunodeficiency viruses from chimpanzees in Cameroon are strongly related to those of human immunodeficiency virus group N from the same geographic area. J. Virol. **74:** 529–534.

De Leys, R., B. Vanderborght, M. Vanden Haesevelde, L. Heyndrickx, A. Van Geel *et al.*, 1990 Isolation and partial characterization of an unusual human immunodeficiency retrovirus from two persons of west-central African origin. J. Virol. **64:** 1207–1216.

Drummond, A., and A. G. Rodrigo, 2000 Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA (sUPGMA). Mol. Biol. Evol. **17:** 1807–1815.

Drummond, A. J., and A. Rambaut, 2003 BEAST v1.0 (http://evolve.zoo.ox.ac.uk/beast/).

Drummond, A. J., G. K. Nicholls, A. G. Rodrigo and W. Solomon, 2002 Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics **161:** 1307–1320.

Drummond, A. J., O. G. Pybus, A. Rambaut, R. Forsberg and A. G. Rodrigo, 2003 Measurably evolving populations. Trends Ecol. Evol. **18:** 481–488.

Gao, F., E. Bailes, D. L. Robertson, C. Yalu, C. M. Rodenburg *et al.*, 1999 Origin of HIV-1 in Pan troglodytes troglodytes. Nature **397:** 436–441.

Griffiths, R. C., and P. Marjoram, 1996 Ancestral inference from samples of DNA sequences with recombination. J. Comp. Biol. **3:** 479–502.

Griffiths, R. C., and S. Tavaré, 1994 Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. Lond. B Biol. Sci. **344:** 403–410.

Gurtler, L. G., P. H. Hauser, J. Eberle, A. Von Brunn, S. Knapp *et al.*, 1994 A new subtype of human immunodeficiency virus type 1 (MVP-5180) from Cameroon. J. Virol. **68:** 1581–1585.

Hackett, J., Jr., L. Zekeng, C. A. Brennan, J. K. Lund, A. S. Vallari *et al.*, 1997 Genetic analysis of HIV type 1 group O p24gag sequences from Cameroon and Equatorial Guinea. AIDS Res. Hum. Retroviruses **13:** 1155–1158.

Holmes, E. C., S. Nee, A. Rambaut, G. P. Garnett and P. H. Harvey, 1995 Revealing the history of infectious disease epidemics through phylogenetic trees. Philos. Trans. R. Soc. Lond. B Biol. Sci. **349** (1327): 33–40.

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, Vol. 7, edited by D. Futuyma and J. Antonovics. Oxford University Press, Oxford.

Jenkins, G. M., A. Rambaut, O. G. Pybus and E. C. Holmes, 2002 Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. J. Mol. Evol. **54:** 156–165.

Kingman, J. F. C., 1982 On the genealogy of large populations. J. Appl. Probab. **19A:** 27–43.

Kishino, H., and M. Hasegawa, 1989 Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J. Mol. Evol. **29:** 170–179.

Korber, B., I. Loussert-Ajaka, J. Blouin and S. Saragosti, 1996 A comparison of HIV-1 group M and group O functional and immunogenic domains in the gag p24 protein and the C2V3 region of the envelope protein, pp. IV-61–77 in *HIV Moelcular Immunology Database 1996*, edited by B. Korber, C. Brander, B. Haynes, R. Koup, J. P. Moore *et al.* Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.

Korber, B., J. Theiler and S. Wolinsky, 1998 Limitations of a molecular clock applied to considerations of the origin of HIV-1. Science **280** (5371): 1868–1871.

Korber, B., M. Muldoon, J. Theiler, F. Gao, R. Gupta *et al.*, 2000 Timing the ancestor of the HIV-1 pandemic strains. Science **288:** 1789–1796.

Kuhner, M. K., J. Yamato and J. Felsenstein, 1998 Maximum likelihood estimation of population growth rates based on the coalescent. Genetics **149:** 429–434.

Lemey, P., O. G. Pybus, B. Wang, N. K. Saksena, M. Salemi *et al.*, 2003 Tracing the origin and history of the HIV-2 epidemic. Proc. Natl. Acad. Sci. USA **100** (11): 6588–6592.

Loussert-Ajaka, I., M. L. Chaix, B. Korber, F. Letourneur, E. Gomas *et al.*, 1995 Variability of human immunodeficiency virus type 1 group O strains isolated from Cameroonian patients living in France. J. Virol. **69:** 5640–5649.

Mauclere, P., I. Loussert-Ajaka, F. Damond, P. Fagot, S. Souquieres *et al.*, 1997 Serological and virological characterization of HIV-1 group O infection in Cameroon. AIDS **11:** 445–453.

McVean, G., P. Awadalla and P. Fearnhead, 2002 A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics **160:** 1231–1241.

McVean, G. A., 2002 A genealogical interpretation of linkage disequilibrium. Genetics **162:** 987–991.

Nath, H., and R. C. Griffiths, 1993 The coalescent in two colonies with symmetric migration. J. Math. Biol. **31:** 841–851.

Nkengasong, J. N., M. Peeters, M. Vanden Haesevelde, S. S. Musi, B. Willems *et al.*, 1993 Antigenic evidence of the presence of the aberrant HIV-1ant70 virus in Cameroon and Gabon. AIDS **7:** 1536–1538.

Peeters, M., A. Gueye, S. Mboup, F. Bibollet-Ruche, E. Ekaza *et al.*, 1997 Geographical distribution of HIV-1 group O viruses in Africa. AIDS **11:** 493–498.

Posada, D., and K. A. Crandall, 1998 MODELTEST: testing the model of DNA substitution. Bioinformatics **14:** 817–818.

Pybus, O. G., and A. Rambaut, 2002 GENIE: estimating demographic history from molecular phylogenies. Bioinformatics **18:** 1404–1405.

Pybus, O. G., A. Rambaut and P. H. Harvey, 2000 An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics **155:** 1429–1437.

Pybus, O. G., A. J. Drummond, T. Nakano, B. H. Robertson and A. Rambaut, 2003 The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. Mol. Biol. Evol. **20:** 381–387.

Quinones-Mateu, M. E., S. C. Ball and E. J. Arts, 2000 Role of human immunodeficiency virus type 1 group O in the AIDS pandemic. AIDS Rev. **2:** 190–202.

Rambaut, A., 2000 Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics **16:** 395–399.

Rayfield, M. A., P. Sullivan, C. I. Bandea, L. Britvan, R. A. Otten *et al.*, 1996 HIV-1 group O virus identified for the first time in the United States. Emerg. Infect. Dis. **2:** 209–212.

Robertson, D. L., B. H. Hahn and P. M. Sharp, 1995 Recombination in AIDS viruses. J. Mol. Evol. **40:** 249–259.

Robertson, D. L., J. P. Anderson, J. A. Bradac, J. K. Carr, B. Foley *et al.*, 2000 HIV-1 nomenclature proposal. Science **288** (5463): 55–56.

Roques, P., D. L. Robertson, S. Souquiere, F. Damond, A. Ayouba *et al.*, 2002 Phylogenetic analysis of 49 newly derived HIV-1 group O strains: high viral diversity but no group M-like subtype structure. Virology **302:** 259–273.

Salemi, M., K. Strimmer, W. W. Hall, M. Duffy, E. Delaporte *et al.*, 2001 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. **15:** 276–278.

Schierup, M., and R. Forsberg, 2003 Proceedings of the Conference: Origins of HIV and Emerging Persistent Viruses, September 2001, Rome, Vol. 187, pp. 231–245.

Schierup, M. H., and J. Hein, 2000a Consequences of recombination on traditional phylogenetic analysis. Genetics **156:** 879–891.

Schierup, M. H., and J. Hein, 2000b Recombination and the molecular clock. Mol. Biol. Evol. **17:** 1578–1579.

Seo, T. K., J. L. Thorne, M. Hasegawa and H. Kishino, 2002 Estimation of effective population size of HIV-1 within a host: a pseudomaximum-likelihood approach. Genetics **160:** 1283–1293.

Sharp, P. M., E. Bailes, R. R. Chaudhuri, C. M. Rodenburg, M. O. Santiago *et al.*, 2001 The origins of acquired immune deficiency syndrome viruses: where and when? Philos. Trans. R. Soc. Lond. B Biol. Sci. **356:** 867–876.

Shimodaira, H., 2002 An approximately unbiased test of phylogenetic tree selection. Syst. Biol. **51:** 492–508.

Shimodaira, H., and M. Hasegawa, 1999 Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. **16:** 1114–1116.

Simon, F., P. Mauclere, P. Roques, I. Loussert-Ajaka, M. C. Muller-Trutwin *et al.*, 1998 Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. Nat. Med. **4:** 1032–1037.

Slatkin, M., and R. R. Hudson, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555–562.

Strimmer, K., and O. G. Pybus, 2001 Exploring the demographic history of DNA sequences using the generalized skyline plot. Mol. Biol. Evol. **18:** 2298–2305.

Swofford, D. L., 1998 *PAUP* 4.0—Phylogenetic Analysis Using Parsimony (*and Other Methods).* Sinauer Associates, Sunderland, MA.

Thompson, J. D., D. G. Higgins and T. J. Gibson, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:** 4673–4680.

Thorne, J. L., and H. Kishino, 2002 Divergence time and evolutionary rate estimation with multilocus data. Syst. Biol. **51:** 689–702.

Thorne, J. L., H. Kishino and I. S. Painter, 1998 Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. **15:** 1647–1657.

Worobey, M., 2001 A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria, and mitochondria. Mol. Biol. Evol. **18:** 1425–1434.

Yamaguchi, J., A. S. Vallari, P. Swanson, P. Bodelle, L. Kaptue *et al.*, 2002 Evaluation of HIV type 1 group O isolates: identification of five phylogenetic clusters. AIDS Res. Hum. Retroviruses **18:** 269–282.

Yang, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13:** 555–556.

Yang, Z., R. Nielsen, N. Goldman and A. M. Pedersen, 2000 Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155:** 431–449.

Yusim, K., M. Peeters, O. G. Pybus, T. Bhattacharya, E. Delaporte *et al.*, 2001 Using human immunodeficiency virus type 1 sequences to infer historical features of the acquired immune deficiency syndrome epidemic and human immunodeficiency virus evolution. Philos. Trans. R. Soc. Lond. B Biol. Sci. **356:** 855–866.

Zekeng, L., S. J. Obiang, H. Hampl, J. M. Ndemesogo, J. Ntutumu *et al.*, 1997 Update on HIV-1 group O infection in Equatorial Guinea, Central Africa. AIDS **11:** 1410–1412.

Communicating editor: S. Yokoyama