Supplementary Materials for

The early spread and epidemic ignition of HIV-1 in human populations

Nuno R. Faria, Andrew Rambaut, Marc A. Suchard, Guy Baele, Trevor Bedford, Melissa J. Ward, Andrew J. Tatem, João D. Sousa, Nimalan Arinaminpathy, Jacques Pépin, David Posada, Martine Peeters, Oliver G. Pybus,* Philippe Lemey*

*Corresponding author. E-mail: philippe.lemey@rega.kuleuven.be (P.L.); oliver.pybus@zoo.ox.ac.uk (O.G.P.)

Published 3 October 2014, *Science* **346**, 56 (2014)
DOI: 10.1126/science.1256739

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:

Nucleotide accession numbers, BEAST XML input files, and BEAST phylogeny distributions have been deposited in the Dryad Repository (http://dx.doi.org/10.5061/dryad.nn952).

Materials and Methods

S1. Compilation of genetic, temporal and geographic data

We began by conducting a preliminary phylogenetic analysis to narrow the geographic focus of our study. We screened the HIV Sequence database (*65*) and retained all HIV-1 group M sequences from Cameroon (CM), Central African Republic (CF), Republic of Congo (RC) and Gabon (GA), plus a subset of sequences sampled from the Democratic Republic of Congo (DRC). These countries each include the Congo River basin and the range of the chimpanzee sub-species *Pan troglodytes troglodytes*. We focused on *env* C2V3 sequences because this genomic region is represented in the sequence database by the highest number of sequences (*65*). Each sequence was isolated from a sample taken from a single infected individual in a known year and geographic location. The data set for the preliminary analysis (fig. S1) comprised 814 sequences, sampled from the DRC (n=348), the RC (n=118), CM (n=304), GA (n=36), CF (n=7). Multiple sequence alignments for all data sets were constructed using MAFFT (*66*) and subsequently edited by hand in Se-Al (http://tree.bio.ed.ac.uk/). A time-scaled evolutionary history with ancestral location estimates from the preliminary data set clearly places the common ancestor of HIV-1 group M in the DRC (figs. S2 and S3).

Next we compiled a large set of envelope C2V3 sequence data comprising the full genetic diversity of HIV-1 group M in the DRC. Sampling dates and locations were confirmed by reference to the original literature. Data was available from six locations in the DRC, in Kinshasa, KN (the capital city), the north (Bwamanda, BW, and Kisangani, KS), and the south (Mbuji-Mayi, MM, Likasi, LK, and Lubumbashi, LB; main text Fig. 2). Because a remarkably high diversity of HIV-1 group M viruses has been found in the neighboring RC, we also included available C2V3 sequence data from Brazzaville, BR, and Pointe Noire, PN (table S1), resulting in a total of 792 nucleotide sequences (sites 7050-7541 relative to HXB2) collected between 1985-2004 (data set A). We focused on samples collected in the largest urban centers of the DRC and RC (*40, 41, 67*). It is possible that individuals living immediately outside these settlements may have attended the hospitals or clinics there.

Further phylogeographic hypotheses were tested by adding additional sequences to data set A, giving rise to data set B (fig. S1, table S1). Specifically, to establish the founder population of HIV-1 subtype C, the predominant lineage worldwide (*53*), we added 67 subtype C sequences from southeast Africa (Zambia, Botswana, Tanzania, Kenya, Uganda, Burundi, Ethiopia and South Africa) sampled between 1986-2005 (*65*). To identify the founder population of HIV-1 subtype B, we added 67 sequences from the Americas (Haiti, Trinidad and Tobago and the USA; sampled between 1978-1997) (*52*). Lastly, to validate our spatio-temporal estimates we added the portion of isolate ZR59 that overlapped the sequences in data set A (positions 7081-7235, HXB2). Isolate ZR59 was obtained in 1959 from blood collected in Kinshasa (*17*). Since the sequences for ZR59 only partly overlap the C2V3 alignment, we dealt with phylogenetic uncertainty by constraining ZR59 to be a sister lineage of subtype D (*16, 25*).

We subsequently down-sampled sequences from the most densely sampled location (KN, for which 422 sequences were available) in data sets A and B, resulting in data sets C and D, respectively (fig. S1, table S1). Down-sampling was undertaken to avoid

potential bias in spatial inference estimates that may arise from over-sampling a particular location (*68*), and ensured that the numbers of sequences from the three most densely sampled locations (KN, MM and BR) were equal (n=96). To obtain 96 sequences sampled from KN we first retained all KN sequences (n=49) from the earliest sampling time (1985) and then randomly selected other 47 sequences from KN. The 1985 sequences were retained in order to maximise temporal evolutionary signal (see main text for discussion of the effects of the removal of these sequences). The limited epidemiological data published to date suggests that KN, BR and MM have similar mean HIV prevalences in the general adult population, ~4-5% (*69, 70*).

S2. Time-scaled phylogenetic tree reconstruction using BEAST/BEAGLE

All five data sets (fig. S1, table S1) were analyzed using a general time-reversible (GTR) nucleotide substitution model (*71*) specifying a gamma distribution as a prior on each relative substitution rate, with a shape parameter = 0.05 and a scale parameter = 0.10 for all rates except for r_{AG}, for which a scale parameter of 0.20 was specified. We used a discrete gamma distribution to model among-site rate heterogeneity (*72*) with an exponential prior (mean = 0.5) on the shape parameter. We used a relaxed uncorrelated lognormal (UCLN) molecular clock model in order to infer the timescale of HIV evolution while accommodating among-lineage rate variation (*73*), with a gamma distribution prior on the mean clock rate (shape = 0.001, scale = 1000) and an exponential prior (mean = 1/3) on the standard deviation.

To confirm that our results are robust to the choice of the molecular clock model, we estimated the time of the most recent common ancestor (TMRCA) for data set C using (i) the UCLN model (*73*), (ii) the strict molecular clock, which assumes no among-lineage variation, (iii) the uncorrelated exponential relaxed clock (UCED) (*73*), and (iv) the random local clock (RLC) model (*74*), which allows a small number of discrete rate changes and hence permits different regions of the tree to evolve under different evolutionary rates. Crucially, all give highly consistent estimates (fig. S8a) demonstrating that the estimated epidemic timescale is robust to the molecular clock prior model chosen. No support was found for phylogenetically-correlated rate changes under the RLC (95% CIs of the estimated number of discrete rates = 0-4). The coefficient of variation of the UCLN, a measure of among-lineage rate variation, was estimated at 0.229 (95% CI: 0.198-0.259). We also tested whether estimates of the TMRCA, and of the age of the ZR59 strain used for statistical validation, were congruent across the data sets analysed. Estimates of these dates were highly consistent across data sets A, B, C and D (fig. S9).

To reconstruct the spatial dynamics of HIV-1 group M, we employed a Bayesian discrete phylogeographic approach (*75, 76*) using Markov chain Monte Carlo (MCMC) sampling, as implemented in the BEAST v1.8.0 software package (*77*). We also used the BEAGLE parallel computation library (*78, 79*) to enhance the speed of the likelihood calculations. For each data set, at least 3 MCMC chains of 250 million steps were computed. Parameters and trees were sampled every $50{,}000^{th}$ step. Samples were combined with LogCombiner (*77*) and between 10 to 30% of each MCMC chain was discarded as burn-in. MCMC mixing was diagnosed using visual trace inspection and calculation of effective sample sizes in Tracer (*77*). We report the posterior mean and

95% Bayesian credible intervals for evolutionary parameters. Using LogCombiner, we subsampled the posterior distribution of phylogenetic trees to generate an empirical distribution of 2,000 trees that is representative of the posterior sample. These were used for the phylogeographic analysis reported in Tables S2-S5. Sequence alignments, accession numbers, BEAST XML input files, and subsampled tree files are available in the DRYAD Repository (http://dx.doi.org/10.5061/dryad.nn952).

S3. Introduction to phylogeographic inference

Epidemiological processes such as geographic spread and population growth leave a measurable imprint on HIV-1 genomes sampled from infected individuals at different places and times (*24*, *80*). These processes can be recovered from genetic data using formal statistical inference methods that take into account both the sequences' shared ancestry and sources of statistical uncertainty (*76, 81*).

To illustrate the general principles of phylogeographic inference, fig. S2 shows virus sequences sampled from three different geographic locations (A, B and C). When individuals are infected in one location and then move to another, or infect someone whilst travelling, this is apparent as a "change" in the location ascribed to one branch of the tree. It is these *viral lineage movements*, which depend not only upon the location values at the tree tips but also on the particular shape of the phylogeny, that provide the information for phylogeographic inference. The pie charts in fig. S2 loosely illustrate statistical support for the inferred location of each internal (unobserved) node, including the most recent common ancestor (root). When movement is ubiquitous (left panel), there will be no relationship between phylogenetic clustering and geographic location, leading to equal support that A, B or C is the location of the root. On the other hand, if viral lineage movement is occasional (right panel) then one root location may be much more probable than others (in this case, location B). A more detailed and formal explanation of this approach is provided in (*76, 82, 83*).

S4. Discrete phylogeographic analyses and spatial structure

To perform ancestral reconstruction of the unobserved sampling locations, a variety of phylogeographic analyses were performed, utilizing the empirical distributions of trees obtained from data sets A to D in Tables S2-S5 (see section S2 above). For the preliminary analysis, two locations were considered: "DRC" and "Other locations in the Congo River basin" (fig. S3). Phylogeographic analyses of data sets A and C included 6 DRC locations and 2 RC locations. Phylogeographic analyses of data sets B and D included 6 DRC locations, 2 RC locations, plus southeast African (SEA) and American discrete locations to represent subtypes B and C (fig. S1). The location exchange process throughout the entire phylogeny was modeled using both symmetric (reversible) and asymmetric (non-reversible) continuous-time Markov chains (CTMCs) (*75*, *76*) with an approximate CTMC conditional reference prior on the overall rate scalar (*84*) and a uniform prior distribution [0,1] on the root state frequencies in the asymmetric models. The asymmetric model uses separate parameters for forward and reverse rates of movement between each pair of locations whilst the symmetric model sets the forward and reverse rates to be equal.

To infer a minimum set of location exchange rates that provides an adequate description of the process of viral dissemination, we use a Bayesian stochastic search variable selection (BSSVS) procedure (*76*). This approach assumes *a priori* that many location state transitions remain unobserved and their corresponding rates are therefore zero. This is formally represented by a truncated Poisson prior that places a 50% prior mass on minimal rate configurations (i.e. only *k*-1 among-location rates are needed to connect *k* locations). By comparing the prior with the posterior odds for a particular rate being non-zero, this procedure offers a Bayes factor (BF) test to identify the most adequate parsimonious description of the process of spatial spread. To test whether the down-sampling of sequences from Kinshasa impacted upon the estimation of the support for viral dispersal links, we compare data sets B and D. Specifically, for these data sets, we compare the *k*-1 among-location rates that are most strongly supported by BSSVS procedure, using both symmetric and asymmetric diffusion models (table S3). In addition, we estimate the expected number of location changes along the branches of a posterior tree distribution using a "robust counting" procedure (*85-87*), as implemented in BEAST v1.8.0 (*77*). This approach uses stochastic mapping techniques to infer, on a branch-by-branch basis, the history of viral movement events between each pair of locations. Fig. S5 shows the percentage of virus movement events from Kinshasa to other locations (conditional on high BF support for the link between the two locations). Fig. S6 shows the percentage of all virus movement events from each location. We used R to summarize posterior probability distribution densities for these transitions and used the ggplot2 package to plot the estimated number/proportion of changes through time to/from particular locations (*88, 89*).

To further explore whether our estimates of the location of the HIV-1 group M ancestor are affected by differences in the number of samples per location, we randomized tip-to-location assignments during the Bayesian inference of each data set. This procedure results in posterior probabilities for each location being the location of the group M common ancestor that are approximately equal (fig. S4). This confirms that the source location that we infer for the real data emerges from the association between phylogenetic clustering and sample location and not from the relative frequency of sampling locations.

S5. Demographic model selection

We evaluated the performance of different coalescent tree priors using a formal model selection procedure. Marginal likelihoods of parametric models (constant, expansion, exponential, logistic growth), and a recently developed flexible nonparametric model (*90*), were compared on data set C using an improved variant of path sampling (PS) (*91*, *92*) called stepping-stone (SS) (*93*) sampling. For the PS method, we selected β values from the path between the posterior and the prior according to evenly-spaced quantiles of a Beta distribution, with exponent = 0.3 and shape parameter = 1.0, as recommended by Xie et al. (*93*).

Statistical model selection indicated that the Bayesian skygrid model fitted the data better than simple parametric models such as constant size, exponential growth, and logistic growth (table S7). Based on the dynamics inferred by the nonparametric approach, we subsequently developed a two-phase exponential-logistic growth

parametric model that provided the best fit among all parametric models (table S7). This model estimates growth rate parameters for each growth period independently and provides an estimate of the time of transition between the exponential and logistic periods. Prior distributions for the exponential-logistic growth coalescent model were as follows: for the effective population size at present we used a log-normal distribution (mean=1.0, standard deviation=2.5), for both the exponential phase and logistic phase growth rate parameters we used a Laplace distribution (mean=0.0, scale parameter=0.05), for the exponential-to-logistic transition time we used a gamma distribution (shape=0.001, scale=1000), and for the logistic half-life time parameter we used a gamma distribution (shape=0.001, scale=1000).

To test whether our results are robust to the coalescent prior used, we estimated the TMRCA parameter under six different coalescent prior models. The coalescent prior chosen had no significant effect on the TMRCA estimates (fig. S8a, table S7). Further, we also estimated the three key parameters of the exponential-logistic growth model (transition time, exponential growth rate, logistic growth rate) under three different molecular clock models (UCLN, RCL and strict clock). Again, the choice of prior model had no significant effect on the estimated parameter values (figs. S8b,c). The BEAST XML definition of the exponential-logistic model is available in the Dryad Repository (http://dx.doi.org/10.5061/dryad.nn952).

S6. The impact of recombination on Bayesian genealogical inference

We employed custom coalescent simulation software (source code available upon request) to assess the impact of recombination on parameter estimates in BEAST. This approach generates non-contemporaneous sequence data using a two-step population genetic procedure that includes recombination and variable demography. First, a genealogy relating the sample is simulated under the coalescent framework with recombination and population size change through time. Second, sequences are evolved along this genealogy or ancestral recombination graph under a particular nucleotide substitution model.

We modeled our simulations after estimates obtained for a subset (n=150) of the HIV-1 group M sequences sampled between 1985-2004 from data set C. We performed a Bayesian inference analysis on this data set using a GTR substitution model with gamma-distributed among-site rate heterogeneity and an exponential growth demographic function. Based on this analysis, we generated sequence data following the empirical sampling time distribution and using empirical HIV-1 base frequencies, the estimated GTR and gamma distribution parameters, a substitution rate of 0.003 substitutions per site per year, an effective population size at present of 500,000 and an exponential growth rate of 0.147 yr^{-1}. Five hundred replicates were generated for four different recombination rates: 0.0, 0.000075, 0.00015 and 0.0003 recombinations per site per year. Table S8 lists the number of recombination events estimated as well as the difference between the expected ϕ-statistic under the assumption of no recombination and the observed ϕ-statistic for the data sets generated using the different recombination rates (*94*). We note that the difference between the expected and observed ϕ-statistic for the real DRC data set (-0.02), for which no significant evidence of recombination is detected (P=0.98), is in the 25^{th} percentile for the simulated data without recombination.

The results of the simulations outlined above are presented in table S9, which reports the coverage probability, mean squared error (MSE) and bias of estimates of three parameters: (i) the TMRCA, (ii) the evolutionary rate and (iii) the exponential growth rate. For the TMRCA, the coverage decreases from close to the nominal value (95%) without recombination, to about 82% for the highest recombination rate. There is also an increase in MSE for higher recombination rates, and this seems to be mostly attributable to an increase in variance of the estimates, as the bias is relatively limited, even for high recombination rates (mean of 5.27% of the TMRCA for r=0.0003). We note that there is a tendency to slightly underestimate TMRCAs in our framework in contrast to previous findings demonstrating that recombination may lead to an overestimation of divergence times using maximum likelihood methods (*95*).

Concomitant with the slight underestimation of TMRCAs, we also observe a small overestimation of evolutionary rates for high recombination rates (about 8.91% for r = 0.0003). So, the major consequence of recombination on estimation of both the TMRCA and evolutionary rate appears to be that one will be somewhat too confident in these estimates. A stronger effect is observed for the exponential growth rate in the coalescent model, which is overestimated by a mean of 52.29% or r=0.0003, which reduces the coverage (to 15% for r=0.0003). This can be explained by the fact that trees estimated from recombinant sequences tend to be more star-shaped than expected from the underlying demographic model, which gives the false impression of higher growth rates (*59*).

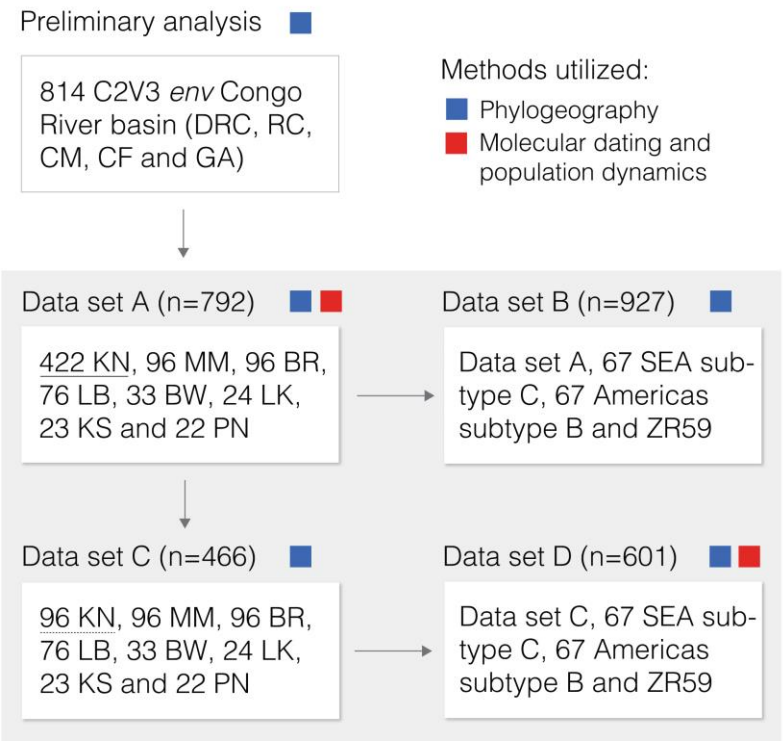

Fig. S1.

Flow diagram illustrating the data sets and analysis approach undertaken in this study. Sampling locations in the DRC are KN=Kinshasa, MM=Mbuji-Mayi, LB=Lubumbashi, BW=Bwamanda, LK=Likasi, KS=Kisangani. Sampling locations in the RC are BR=Brazzaville, PN=Pointe Noire. SEA=Southeast Africa subtype C. The numbers of sequences from KN (underlined) were downsampled in data sets C and D to achieve an identical number of sequences for the three most densely sampled locations (KN, MM, BR; table S1). Data sets subjected to molecular dating and population dynamic analyses are indicated with red squares, whilst those used for phylogeographic analyses are indicated with blue squares. All analyses are described in Materials and Methods. Sequence alignments, accession numbers and BEAST XML input files have been deposited in the Dryad Repository (http://dx.doi.org/10.5061/dryad.nn952).

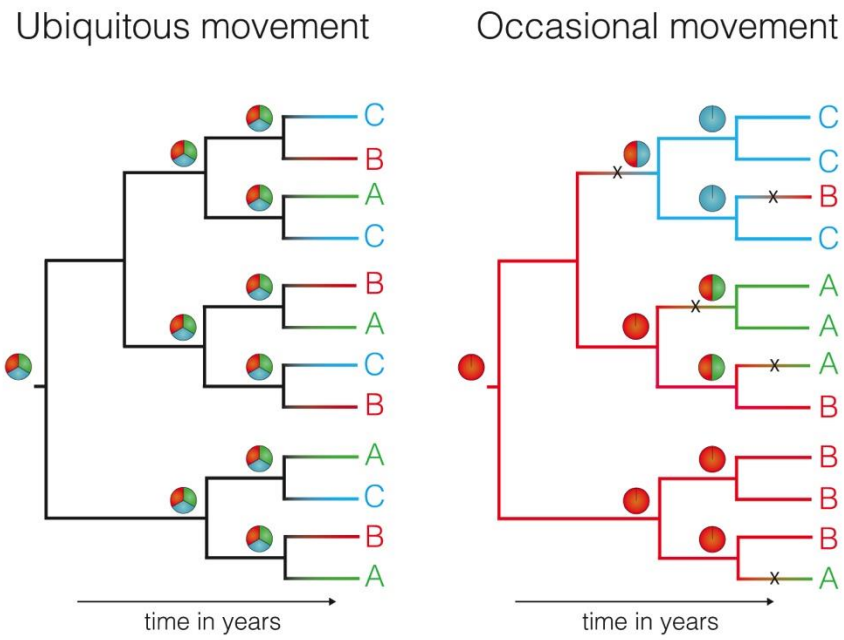

Fig. S2.

Illustration of the general principles of phylogeographic inference. Traits A, B and C at the tips of the phylogeny represent hypothetical geographic locations from which genetic sequence data was collected. The color-coded pie charts represent posterior probability support for the location estimates.

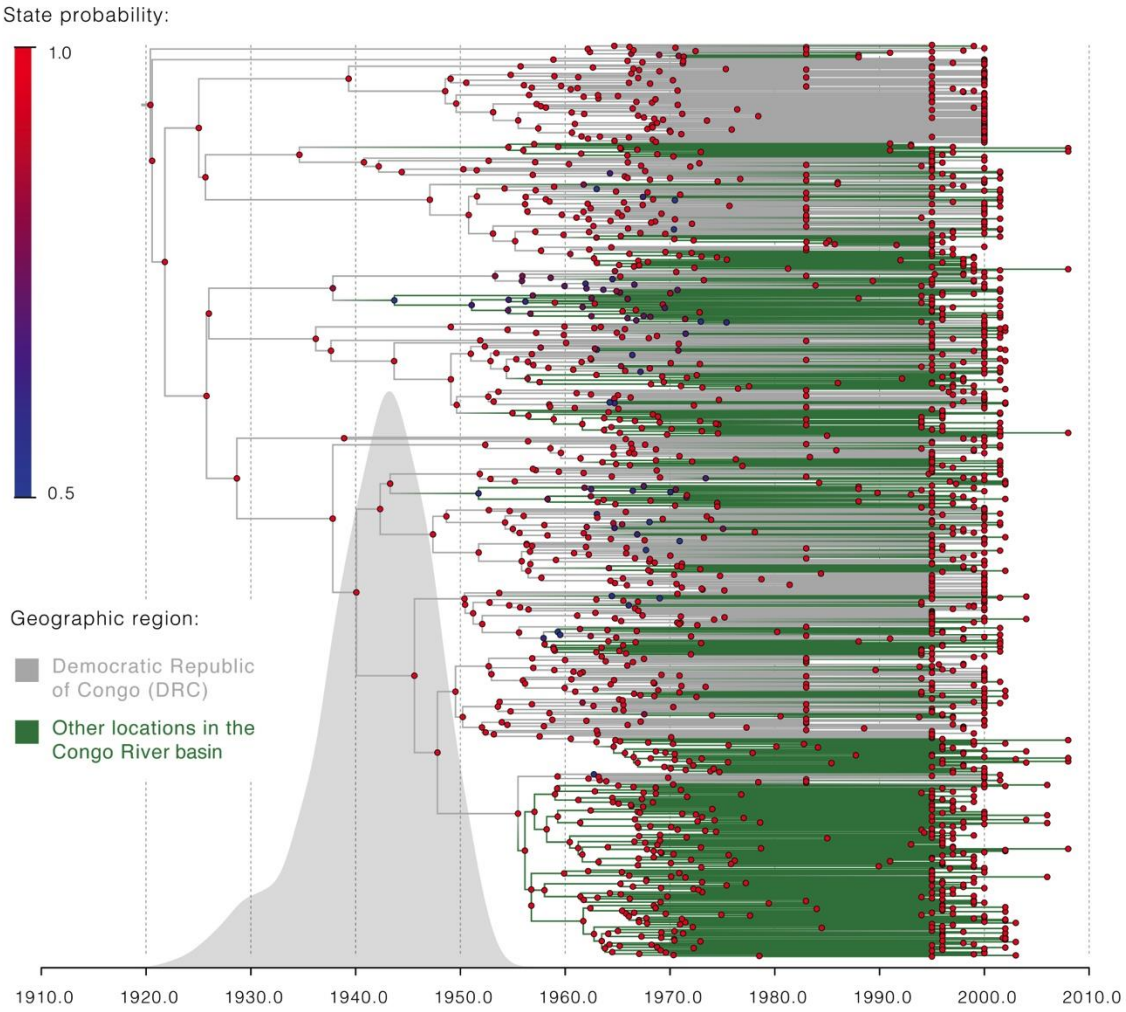

Fig. S3.

Time-calibrated maximum clade credibility tree of the preliminary analysis data set (see fig. S1). Branches in the phylogeny are colored according to their most probable geographic location. DRC locations are in gray; other countries located in the Congo River basin are shown in green (Cameroon, Central African Republic, Republic of Congo, Gabon; see Materials and Methods). At each node, circles are colored according to the posterior probability of the node location (see color gradient, top-left). Bayesian inference was performed as described in Materials and Methods. The posterior probability support for DRC as the ancestral root location was >0.99. The posterior probability density of the earliest migration event out of the DRC is superimposed in gray and is centered around 1944.

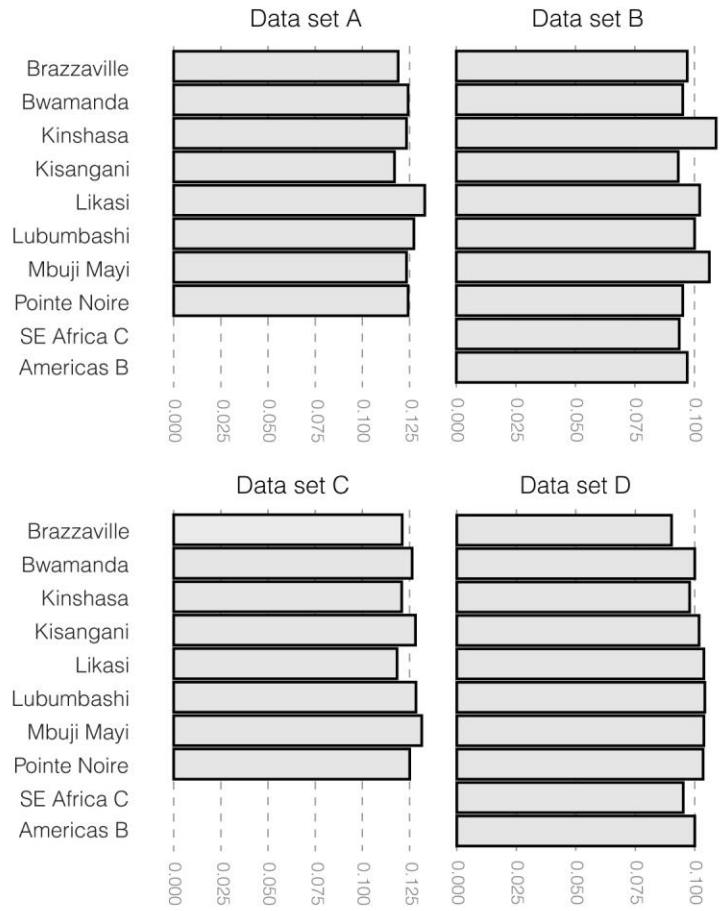

Fig. S4.

Results of the randomization procedure, used to assess the robustness of the estimated location of the HIV-1 group M common ancestor. During the analysis of each data set, the assignment of locations to tree tips was randomized. In each panel, sampling locations are shown on the left, and the posterior probability for each of location being the root location is shown on the horizontal axis.

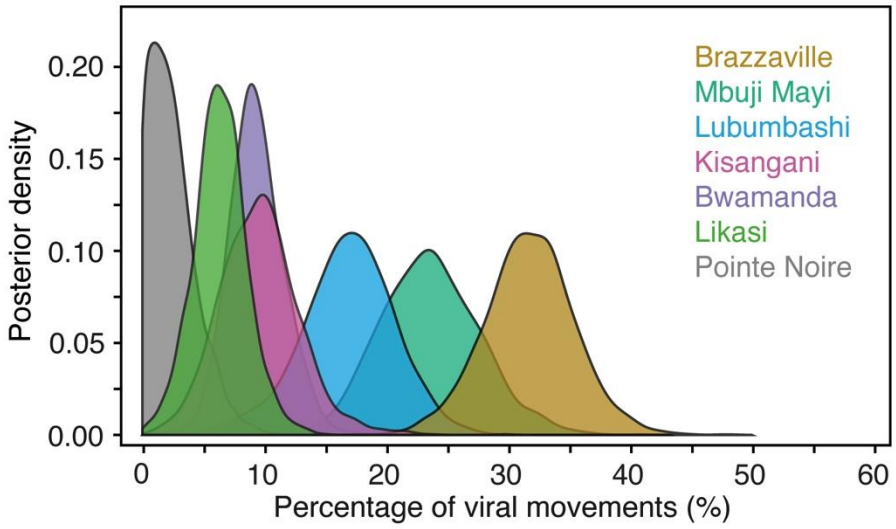

Fig. S5.

Estimated percentages of HIV-1 group M migration events *from* Kinshasa *to* each location in the DRC and RC, obtained using data set C. Posterior mean and 95% Bayesian credible intervals (BCIs) of the percentage of jumps from each location obtained using a robust counting approach (see Materials and Methods). Color-coded locations are shown on the right.

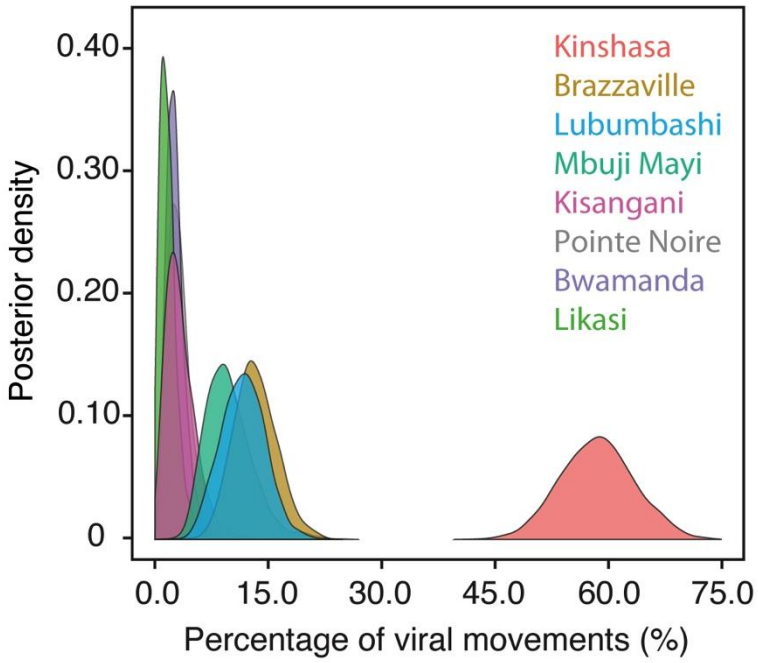

Fig. S6.

Estimated percentages of HIV-1 group M migration events *from* each location, obtained using data set C. Posterior mean and 95% Bayesian credible intervals (BCIs) for each percentage were obtained using a robust counting approach (see Materials and Methods). Color-coded locations are shown on the right.

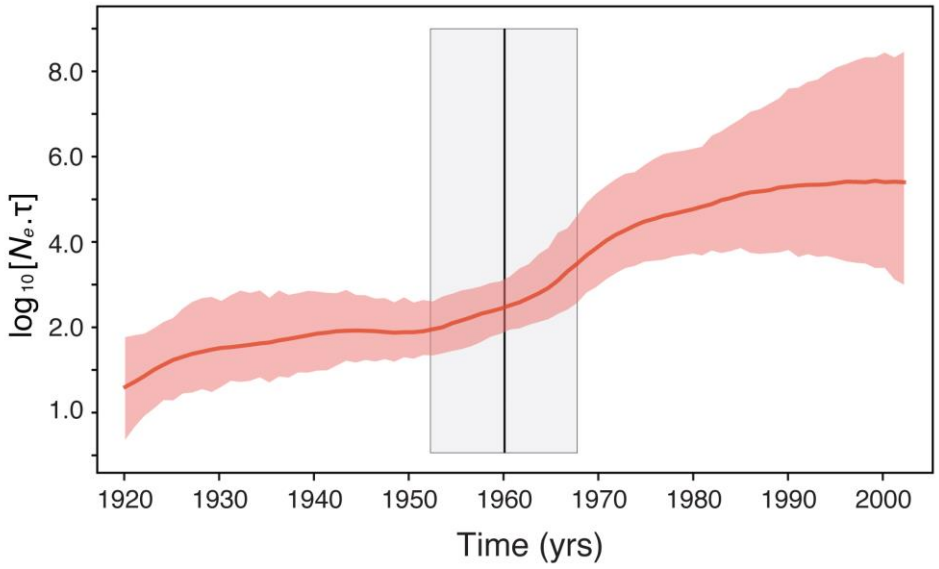

Fig. S7.

Bayesian skygrid plot estimated for the subset of sequences (*n*=84) in data set C that maintain ancestry in Kinshasa throughout their entire phylogenetic history. Mean and respective 95% BCIs of the compound parameter [effective population size x generation time] are plotted in red. The gray area represents the BCI of the posterior estimate of the transition time between the slow and fast growth rate phases of the exponential-logistic demographic model (main text Fig. 4). The vertical black line represents the mean estimate of this parameter.

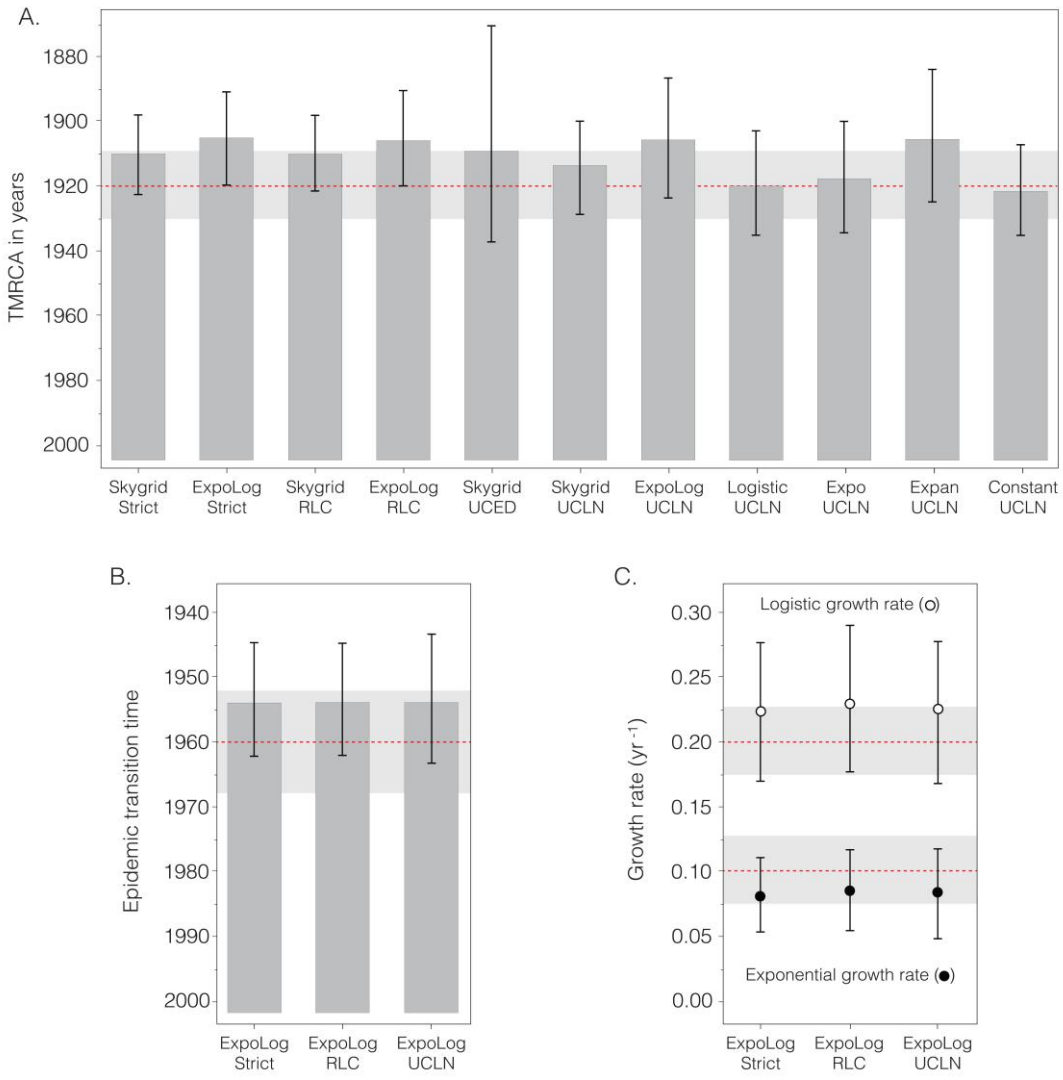

Fig. S8.

Robustness of estimates of key parameters to the choice of coalescent and molecular clock priors in our Bayesian inference framework (TMRCA, panel A; epidemic transition time, panel B; exponential and logistic growth rates, panel C). Vertical boxes show mean parameter estimates and vertical bars represent the corresponding 95% CIs. Red dashed lines and gray horizontal bars indicate the mean and 95% CIs of the parameter estimates that were reported in the main text. Model definitions are provided beneath each estimate; the top line of text specifies the coalescent prior used and the bottom line of text specifies the clock model used. All analyses were conducted using data set C. In panel C, the logistic phase growth rates were ~2.9 times faster than the exponential phase growth rates. RLC: random local clock; UCLN: uncorrelated relaxed lognormal clock; UCED: uncorrelated relaxed exponential clock; Expo: exponential growth coalescent prior, ExpoLog: exponential-logistic coalescent prior; Expan: expansion coalescent prior; Constant: constant size coalescent prior. The expoLog+UCED model combination did not converge.

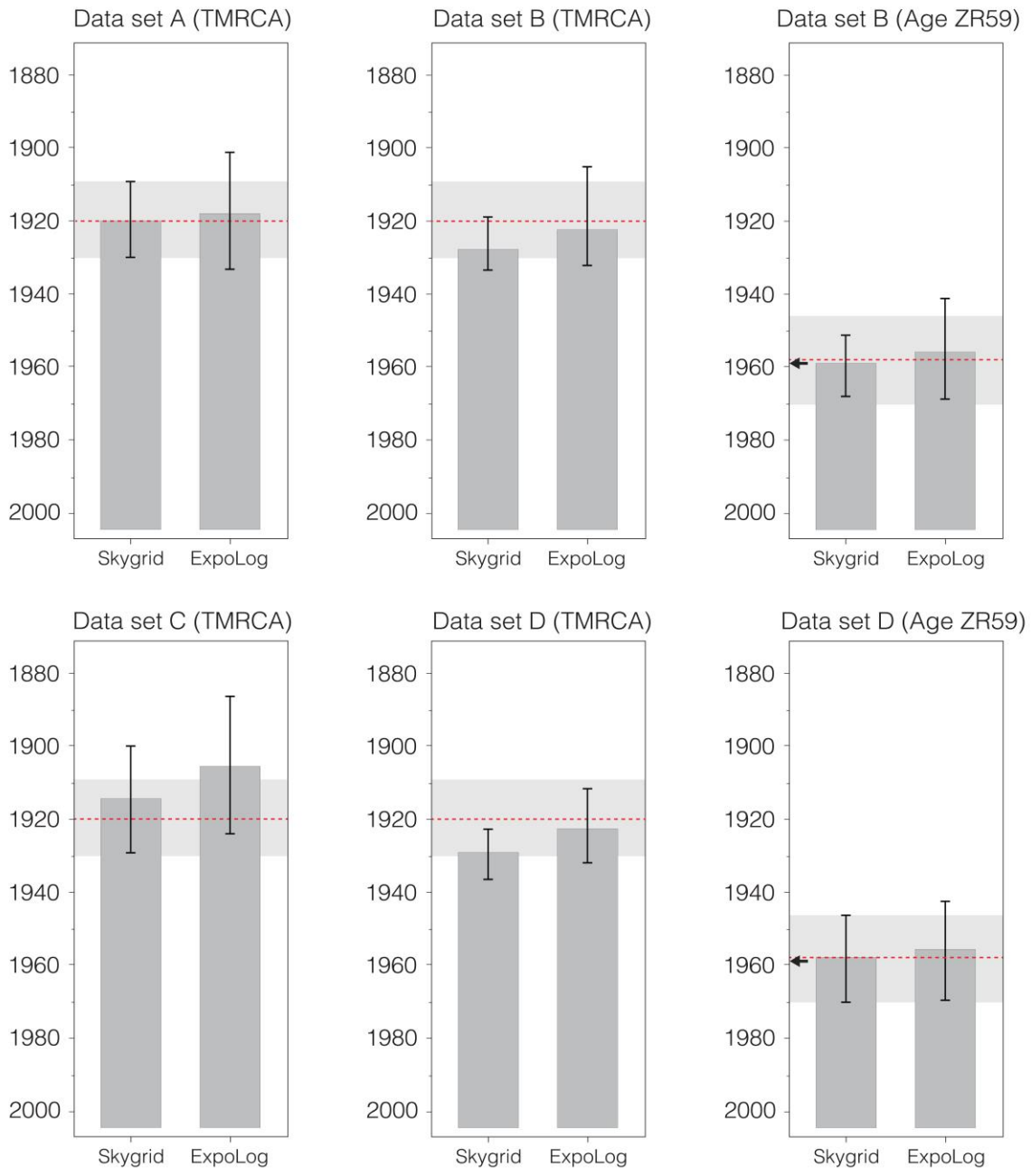

Fig. S9.

Robustness of molecular clock dating estimates for different data sets (table S1) and coalescent priors. All analyses were performed using an uncorrelated relaxed lognormal molecular clock model. Red dashed lines and gray horizontal bars indicate the mean and 95% CIs of the parameter estimates that were reported in the main text. Small arrows within the right hand side panels indicate the true age of the ZR59 isolate.

Table S1.

HIV-1 group M data sets used in this study. The numbers of sequences from Kinshasa after down-sampling are highlighted in bold. Data sets B and D include US/Haiti/Trinidad subtype B and southeast African (SEA) subtype C sequences. The ZR59 sequence was added to data sets B and D and its date and location of sampling was estimated as a control. RC: Republic of Congo. Brazzaville is separated by Kinshasa only by the Congo River. Pointe Noire is the second largest city of the RC and is connected since 1934 to Brazzaville via the Congo-Ocean railway (*22*).

	Data sets				
Locations	**A**	**B**	**C**	**D**	**References**
Kinshasa	422	422	**96**	**96**	(*12, 13, 32, 95*)
Kisangani	23	23	23	23	(*32*)
Mbuji-Mayi	96	96	96	96	(*13, 32*)
Bwamanda	33	33	33	33	(*13*)
Likasi	24	24	24	24	(*31*)
Lubumbashi	76	76	76	76	(*32*)
Brazzaville (RC)	96	96	96	96	(*18, 19*)
Pointe Noire (RC)	22	22	22	22	(*96*)
Americas subtype B	-	67	-	67	(*52,97*)
SEA subtype C	-	67	-	67	(*65*)
ZR59	-	1	-	1	(*17*)
Total	792	927	466	601	

Table S2.
Root state posterior probabilities estimated for data set A. Root location state probabilities were estimated for a representative sample of 2,000 genealogies drawn from the posterior distributions using the discrete phylogeographic models implemented in BEAST v.1.8.0 (*77*). Symmetric (*76*) and asymmetric (*75*) models of trait diffusion were used, with and without Bayesian stochastic search variable selection (BSSVS) (*76*). Modal state probability support values are highlighted in bold. SEA=Southeast Africa.

	Reversible	**Reversible BSSVS**	**Non-Reversible**	**Non-Reversible BSSVS**
Kinshasa	**1.0000**	**0.998**	**1.0000**	**0.998**
Kisangani	0.0000	0.0000	0.0000	0.0000
Mbuji-Mayi	0.0000	0.0000	0.0000	0.0000
Bwamanda	0.0000	0.0000	0.0000	0.0000
Likasi	0.0000	0.0000	0.0000	0.0000
Lubumbashi	0.0000	0.002	0.0000	0.002
Brazzaville	0.0000	0.0000	0.0000	0.0000
Pointe Noire	0.0000	0.0000	0.0000	0.0000
Americas subtype B	0.0000	0.0000	0.0000	0.0000
SEA subtype C	0.0000	0.0000	0.0000	0.0000

Table S3.

Root state posterior probabilities estimated for data set B. See table S2 for details.

	Reversible	Reversible BSSVS	Non-Reversible	Non-Reversible BSSVS
Kinshasa	**0.991**	**0.9906**	**0.9972**	**1.0000**
Kisangani	0.0000	0.0000	0.0000	0.0000
Mbuji-Mayi	0.0025	0.0007	0.0000	0.0000
Bwamanda	0.0000	0.0000	0.0000	0.0000
Likasi	0.0049	0.0074	0.0009	0.0000
Lubumbashi	0.0008	0.0000	0.0009	0.0000
Brazzaville	0.0008	0.0013	0.0000	0.0000
Pointe Noire	0.0000	0.0000	0.0009	0.0000
Americas subtype B	0.0000	0.0000	0.0000	0.0000
SEA subtype C	0.0000	0.0000	0.0000	0.0000

Table S4.

Root state posterior probabilities estimated for data set C. See table S2 for details.

	Reversible	Reversible BSSVS	Non-Reversible	Non-Reversible BSSVS
Kinshasa	**0.998**	**0.996**	**0.9959**	**0.9573**
Kisangani	0.0000	0.0007	0.0000	0.0000
Mbuji-Mayi	0.0000	0.0007	0.0000	0.0000
Bwamanda	0.0000	0.0000	0.0000	0.0000
Likasi	0.0013	0.0027	0.0000	0.0000
Lubumbashi	0.0007	0.0000	0.0000	0.0000
Brazzaville	0.0000	0.0000	0.0000	0.0000
Pointe Noire	0.0000	0.0000	0.0000	0.0000
Americas subtype B	0.0000	0.0000	0.0000	0.0000
SEA subtype C	0.0000	0.0000	0.0041	0.0427

Table S5.

Root state posterior probabilities estimated for data set D. See table S2 for details.

	Reversible	Reversible BSSVS	Non-Reversible	Non-Reversible BSSVS
Kinshasa	**0.9875**	**0.997**	**0.9953**	**0.9962**
Kisangani	0.0000	0.0000	0.0000	0.0000
Mbuji-Mayi	0.0062	0.0000	0.0000	0.0000
Bwamanda	0.0000	0.0000	0.0000	0.0000
Likasi	0.0031	0.0000	0.0047	0.0047
Lubumbashi	0.0000	0.0000	0.0000	0.0000
Brazzaville	0.0031	0.003	0.0000	0.0000
Pointe Noire	0.0000	0.0000	0.0000	0.0000
Americas subtype B	0.0000	0.0000	0.0000	0.0000
SEA subtype C	0.0000	0.0000	0.0000	0.0038

Table S6.

The highest supported pathways of HIV-1 group M spread. Bayes Factor support for non-zero viral migration pathways between pairs of geographic locations was obtained using the BSSVS procedure (*76*). Bayes factors were summarized using the SPREAD v.1.0.6 application (*98*). Links supported by Bayes Factor (BF) values below ten are highlighted in italics. Note that the first nine rates are concordant for both data sets and supported by decisive BF support (>100). The finding of a significant rate between SEA subtype C and Kinshasa is consistent with a report highlighting the introduction of this lineage in the capital of the DRC in recent years (*32*).

From	To	Bayes factor	
		Data set D	Data set B
Kinshasa	Kisangani	>1000	>1000
Kinshasa	Mbuji-Mayi	>1000	>1000
Kinshasa	Bwamanda	>1000	>1000
Kinshasa	Lubumbashi	>1000	>1000
Kinshasa	Likasi	>1000	>1000
Kinshasa	Brazzaville	>1000	>1000
Brazzaville	Pointe Noire	>1000	702.6
SEA subtype C	Mbuji-Mayi	>1000	>1000
SEA subtype C	Lubumbashi	>1000	>1000
Mbuji-Mayi	Lubumbashi	562.5	*3.88*
Brazzaville	Kinshasa	219.6	271.89
SEA subtype C	Kinshasa	60.27	>1000

Table S7.

Log marginal likelihood estimates for different coalescent tree priors applied to HIV-1 group M. The best fitting parametric and nonparametric coalescent tree priors are highlighted with superscripts 1 and 2, respectively, for each marginal likelihood estimator. Duplicate analyses were run using equal numbers of MCMC iterations, resulting in identical results and low variance of the estimators (not shown). These estimates did not change with longer run times. HME: harmonic mean estimator, sHME: stabilized/smoothed harmonic mean estimator, PS: path sampling, SS: stepping-stone.

Method	**Constant**	**Expansion**	**Exponential**	**Logistic**	**Expo-Log[1]**	**Skygrid[2]**
HME	-54675.31	-54695.07	-54686.90	-54665.65	-54657.80	-52532.57
sHME	-54629.34	-54640.59	-54626.52	-54627.01	-54610.74	-52489.26
PS	-57507.68	-56913.42	-56898.19	-56890.70	-56662.06	-54653.82
SS	-57525.65	-56935.26	-56918.06	-56908.73	-56682.33	-54675.26

Table S8.

Number of recombination events and differences in expected and observed ϕ-statistics estimated for simulations under different recombination rates.

Recombination rate	No. recombination events (95% CI)	$\phi_e - \phi_o$ (stdev)
0	0	-4.60×10^{-4} (3.16×10^{-2})
0.000075	135 (114, 160)	1.69×10^{-2} (1.11×10^{-3})
0.000150	276 (240, 314)	1.89×10^{-4} (3.36×10^{-2})
0.000300	547 (504, 567)	1.89×10^{-4} (3.50×10^{-2})

Table S9.

Influence of recombination rate on estimates of times to the most recent common ancestor (TMRCA), evolutionary rates and growth rates. Coverage=coverage probability and MSE=Mean squared error.

Parameter	Property	Recombination rate			
		0.0	**0.000075**	**0.000150**	**0.000300**
TMRCA	Coverage	0.945	0.883	0.845	0.824
	MSE	43.979	54.178	64.741	71.423
	Bias	-0.737	2.261	3.112	4.296
Evolutionary rate	Coverage	0.916	0.845	0.833	0.838
	MSE	7.624E-08	1.362E-07	1.667E-07	1.952E-07
	Bias	-1.702E-05	-2.147E-04	-2.413E-04	-2.672E-04
Growth rate	Coverage	0.941	0.840	0.531	0.148
	MSE	2.432E-04	7.276E-04	2.081E-03	6.840E-03
	Bias	-1.964E-03	-1.825E-02	-3.891E-02	-7.686E-02

References and Notes

1. UNAIDS, “Global Reports - UNAIDS report on the global AIDS epidemic 2013” (UNAIDS, Geneva, 2013).

2. M. S. Gottlieb, M. D. Schanker, P. T. Fan, M. D. Saxon, J. D. Weisman; Centers for Disease Control (CDC), Pneumocystis pneumonia—Los Angeles. *MMWR Morb. Mortal. Wkly. Rep.* **30**, 250–252 (1981). Medline

3. F. Barré-Sinoussi, J. C. Chermann, F. Rey, M. T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vézinet-Brun, C. Rouzioux, W. Rozenbaum, L. Montagnier, Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* **220**, 868–871 (1983). Medline doi:10.1126/science.6189183

4. R. C. Gallo, P. S. Sarin, E. P. Gelmann, M. Robert-Guroff, E. Richardson, V. S. Kalyanaraman, D. Mann, G. D. Sidhu, R. E. Stahl, S. Zolla-Pazner, J. Leibowitch, M. Popovic, Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). *Science* **220**, 865–867 (1983). Medline doi:10.1126/science.6601823

5. P. Piot, H. Taelman, K. Bila Minlangu, N. Mbendi, K. Ndangi, K. Kalambayi, C. Bridts, T. C. Quinn, F. M. Feinsod, O. Wobin, P. Mazebo, W. Stevens, S. Mitchell, J. B. Mccormick, Acquired immunodeficiency syndrome in a heterosexual population in Zaire. *Lancet* **2**, 65–69 (1984). Medline doi:10.1016/S0140-6736(84)90241-1

6. P. Van de Perre, D. Rouvroy, P. Lepage, J. Bogaerts, P. Kestelyn, J. Kayihigi, A. C. Hekker, J. P. Butzler, N. Clumeck, Acquired immunodeficiency syndrome in Rwanda. *Lancet* **2**, 62–65 (1984). Medline doi:10.1016/S0140-6736(84)90240-X

7. B. F. Keele, F. Van Heuverswyn, Y. Li, E. Bailes, J. Takehisa, M. L. Santiago, F. Bibollet-Ruche, Y. Chen, L. V. Wain, F. Liegeois, S. Loul, E. M. Ngole, Y. Bienvenue, E. Delaporte, J. F. Brookfield, P. M. Sharp, G. M. Shaw, M. Peeters, B. H. Hahn, Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science* **313**, 523–526 (2006). Medline doi:10.1126/science.1126531

8. F. Van Heuverswyn, Y. Li, E. Bailes, C. Neel, B. Lafay, B. F. Keele, K. S. Shaw, J. Takehisa, M. H. Kraus, S. Loul, C. Butel, F. Liegeois, B. Yangda, P. M. Sharp, E. Mpoudi-Ngole, E. Delaporte, B. H. Hahn, M. Peeters, Genetic diversity and phylogeographic clustering of SIVcpz*Ptt* in wild chimpanzees in Cameroon. *Virology* **368**, 155–171 (2007). Medline doi:10.1016/j.virol.2007.06.018

9. A. Ayouba, S. Souquières, B. Njinku, P. M. V. Martin, M. C. Müller-Trutwin, P. Roques, F. Barré-Sinoussi, P. Mauclère, F. Simon, E. Nerrienet, HIV-1 group N among HIV-1-seropositive individuals in Cameroon. *AIDS* **14**, 2623–2625 (2000). Medline doi:10.1097/00002030-200011100-00033

10. M. Peeters, A. Gueye, S. Mboup, F. Bibollet-Ruche, E. Ekaza, C. Mulanga, R. Ouedrago, R. Gandji, P. Mpele, G. Dibanga, B. Koumare, M. Saidou, E. Esu-Williams, J.-P. Lombart, W. Badombena, N. Luo, M. Vanden Haesevelde, E. Delaporte, Geographical distribution of HIV-1 group O viruses in Africa. *AIDS* **11**, 493–498 (1997). Medline doi:10.1097/00002030-199704000-00013

11. A. Vallari, V. Holzmayer, B. Harris, J. Yamaguchi, C. Ngansop, F. Makamche, D. Mbanya, L. Kaptué, N. Ndembi, L. Gürtler, S. Devare, C. A. Brennan, Confirmation of putative HIV-1 group P in Cameroon. *J. Virol.* **85**, 1403–1407 (2011). Medline doi:10.1128/JVI.02005-10

12. M. L. Kalish, K. E. Robbins, D. Pieniazek, A. Schaefer, N. Nzilambi, T. C. Quinn, M. E. St Louis, A. S. Youngpairoj, J. Phillips, H. W. Jaffe, T. M. Folks, Recombinant viruses and early global HIV-1 epidemic. *Emerg. Infect. Dis.* **10**, 1227–1234 (2004). Medline

13. N. Vidal, M. Peeters, C. Mulanga-Kabeya, N. Nzilambi, D. Robertson, W. Ilunga, H. Sema, K. Tshimanga, B. Bongo, E. Delaporte, Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. *J. Virol.* **74**, 10498–10507 (2000). Medline doi:10.1128/JVI.74.22.10498-10507.2000

14. A. Rambaut, D. L. Robertson, O. G. Pybus, M. Peeters, E. C. Holmes, Human immunodeficiency virus: Phylogeny and the origin of HIV-1. *Nature* **410**, 1047–1048 (2001). Medline doi:10.1038/35074179

15. A. Rambaut, D. Posada, K. A. Crandall, E. C. Holmes, The causes and consequences of HIV evolution. *Nat. Rev. Genet.* **5**, 52–61 (2004). Medline doi:10.1038/nrg1246

16. M. Worobey, M. Gemmel, D. E. Teuwen, T. Haselkorn, K. Kunstman, M. Bunce, J. J. Muyembe, J. M. Kabongo, R. M. Kalengayi, E. Van Marck, M. T. Gilbert, S. M. Wolinsky, Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* **455**, 661–664 (2008). Medline doi:10.1038/nature07390

17. T. Zhu, B. T. Korber, A. J. Nahmias, E. Hooper, P. M. Sharp, D. D. Ho, An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. *Nature* **391**, 594–597 (1998). Medline doi:10.1038/35400

18. B. Bikandou, M. Y. Ndoundou-Nkodia, F. R. Niama, M. Ekwalanga, O. Obengui, R. Taty-Taty, H. J. Parra, S. Saragosti, Genetic subtyping of gag and env regions of HIV type 1 isolates in Republic of Congo. *AIDS Res. Hum. Retroviruses* **20**, 1005–1009 (2004). Medline doi:10.1089/aid.2004.20.1005

19. F. R. Niama, C. Toure-Kane, N. Vidal, P. Obengui, B. Bikandou, M. Y. Ndoundou Nkodia, C. Montavon, H. Diop-Ndiaye, J. V. Mombouli, E. Mokondzimobe, A. G. Diallo, E. Delaporte, H. J. Parra, M. Peeters, S. Mboup, HIV-1 subtypes and recombinants in the Republic of Congo. *Infect. Genet. Evol.* **6**, 337–343 (2006). Medline doi:10.1016/j.meegid.2005.12.001

20. I. Pandrea, D. L. Robertson, R. Onanga, F. Gao, M. Makuwa, P. Ngari, I. Bedjabaga, P. Roques, F. Simon, C. Apetrei, Analysis of partial *pol* and *env* sequences indicates a high prevalence of HIV type 1 recombinant strains circulating in Gabon. *AIDS Res. Hum. Retroviruses* **18**, 1103–1116 (2002). Medline doi:10.1089/088922202320567842

21. J. K. Carr, N. D. Wolfe, J. N. Torimiro, U. Tamoufe, E. Mpoudi-Ngole, L. Eyzaguirre, D. L. Birx, F. E. McCutchan, D. S. Burke, HIV-1 recombinants with multiple parental strains in low-prevalence, remote regions of Cameroon: Evolutionary relics? *Retrovirology* **7**, 39 (2010). Medline doi:10.1186/1742-4690-7-39

22. J. Pepin, *The Origins of AIDS* (Cambridge Univ. Press, Cambridge, 2011).

23. D. Sauter, M. Schindler, A. Specht, W. N. Landford, J. Münch, K. A. Kim, J. Votteler, U. Schubert, F. Bibollet-Ruche, B. F. Keele, J. Takehisa, Y. Ogando, C. Ochsenbauer, J. C. Kappes, A. Ayouba, M. Peeters, G. H. Learn, G. Shaw, P. M. Sharp, P. Bieniasz, B. H. Hahn, T. Hatziioannou, F. Kirchhoff, Tetherin-driven adaptation of Vpu and Nef function and the evolution of pandemic and nonpandemic HIV-1 strains. *Cell Host Microbe* **6**, 409–421 (2009). Medline doi:10.1016/j.chom.2009.10.004

24. O. G. Pybus, A. Rambaut, Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.* **10**, 540–550 (2009). Medline doi:10.1038/nrg2583

25. B. Korber, M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B. H. Hahn, S. Wolinsky, T. Bhattacharya, Timing the ancestor of the HIV-1 pandemic strains. *Science* **288**, 1789–1796 (2000). Medline doi:10.1126/science.288.5472.1789

26. M. Salemi, K. Strimmer, W. W. Hall, M. Duffy, E. Delaporte, S. Mboup, M. Peeters, A. M. Vandamme, 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 (2001). Medline

27. K. Yusim, M. Peeters, O. G. Pybus, T. Bhattacharya, E. Delaporte, C. Mulanga, M. Muldoon, J. Theiler, B. Korber, 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. London Ser. B* **356**, 855–866 (2001). Medline doi:10.1098/rstb.2001.0859

28. G. Baele, P. Lemey, T. Bedford, A. Rambaut, M. A. Suchard, A. V. Alekseyenko, Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol. Biol. Evol.* **29**, 2157–2167 (2012). Medline doi:10.1093/molbev/mss084

29. D. L. Robertson, P. M. Sharp, F. E. McCutchan, B. H. Hahn, Recombination in HIV-1. *Nature* **374**, 124–126 (1995). Medline doi:10.1038/374124b0

30. Materials and methods are available as supplementary materials on *Science* Online.

31. K. Kita, N. Ndembi, M. Ekwalanga, E. Ido, R. Kazadi, B. Bikandou, J. Takehisa, T. Takemura, S. Kageyama, J. Tanaka, H. J. Parra, M. Hayami, H. Ichimura, Genetic diversity of HIV type 1 in Likasi, southeast of the Democratic Republic of Congo. *AIDS Res. Hum. Retroviruses* **20**, 1352–1357 (2004). Medline doi:10.1089/aid.2004.20.1352

32. N. Vidal, C. Mulanga, S. E. Bazepeo, J. K. Mwamba, J. W. Tshimpaka, M. Kashi, N. Mama, C. Laurent, F. Lepira, E. Delaporte, M. Peeters, 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.* **40**, 456–462 (2005). Medline doi:10.1097/01.qai.0000159670.18326.94

33. A. Chitnis, D. Rawls, J. Moore, Origin of HIV type 1 in colonial French Equatorial Africa? *AIDS Res. Hum. Retroviruses* **16**, 5–8 (2000). Medline doi:10.1089/088922200309548

34. J. Pépin, The expansion of HIV-1 in colonial Leopoldville, 1950s: Driven by STDs or STD control? *Sex. Transm. Infect.* **88**, 307–312 (2012). Medline doi:10.1136/sextrans-2011-050277

35. D. Vangroenweghe, The earliest cases of human immunodeficiency virus type 1 group M in Congo-Kinshasa, Rwanda and Burundi and the origin of acquired immune deficiency syndrome. *Philos. Trans. R. Soc. London Ser. B* **356**, 923–925 (2001). Medline doi:10.1098/rstb.2001.0876

36. J. D. de Sousa, C. Alvarez, A. M. Vandamme, V. Müller, Enhanced heterosexual transmission hypothesis for the origin of pandemic HIV-1. *Viruses* **4**, 1950–1983 (2012). Medline doi:10.3390/v4101950

37. P. M. Sharp, B. H. Hahn, AIDS: Prehistory of HIV-1. *Nature* **455**, 605–606 (2008). Medline doi:10.1038/455605a

38. A. Huybrechts, *Transports et Structures de Development au Congo: Etude du Progres Economique de 1900-1970* (Mouton, Paris, 1970).

39. R. R. Gray, A. J. Tatem, S. Lamers, W. Hou, O. Laeyendecker, D. Serwadda, N. Sewankambo, R. H. Gray, M. Wawer, T. C. Quinn, M. M. Goodenow, M. Salemi, Spatial phylodynamics of HIV-1 epidemic emergence in east Africa. *AIDS* **23**, F9–F17 (2009). Medline doi:10.1097/QAD.0b013e32832faf61

40. J. Flouriot, *Introduction a la Geographique Physique et Humaine du Zaire* (Lyon, France, 1994, mimeographed).

41. W. A. Hance, *Population, Migration, and Urbanization in Africa* (Columbia Univ. Press, New York, 1970).

42. T. C. Quinn, Population migration and the spread of types 1 and 2 human immunodeficiency viruses. *Proc. Natl. Acad. Sci. U.S.A.* **91**, 2407–2414 (1994). Medline doi:10.1073/pnas.91.7.2407

43. M. Ngimbi, *Kinshasa, 1881–1981: 100 Ans Après Stanley: Problèmes et Avenir d'une Ville* (Ed Centre de Recherches Pédagogiques, Kinshasa, 1982).

44. P. Lemey, A. Rambaut, O. G. Pybus, HIV evolutionary dynamics within and among hosts. *AIDS Rev.* **8**, 125–140 (2006). Medline

45. C. Mulanga-Kabeya, N. Nzilambi, B. Edidi, M. Minlangu, T. Tshimpaka, L. Kambembo, L. Atibu, N. Mama, W. Ilunga, H. Sema, K. Tshimanga, B. Bongo, M. Peeters, E. Delaporte, Evidence of stable HIV seroprevalences in selected populations in the Democratic Republic of the Congo. *AIDS* **12**, 905–910 (1998). Medline doi:10.1097/00002030-199808000-00013

46. N. Nzilambi, K. M. De Cock, D. N. Forthal, H. Francis, R. W. Ryder, I. Malebe, J. Getchell, M. Laga, P. Piot, J. B. McCormick, The prevalence of infection with human immunodeficiency virus over a 10-year period in rural Zaire. *N. Engl. J. Med.* **318**, 276–279 (1988). Medline doi:10.1056/NEJM198802043180503

47. S. M. Duke-Sylvester, R. Biek, L. A. Real, Molecular evolutionary signatures reveal the role of host ecological dynamics in viral disease emergence and spread. *Philos. Trans. R. Soc. London Ser. B* **368**, 20120194 (2013). Medline doi:10.1098/rstb.2012.0194

48. G. Magiorkinis, V. Sypsa, E. Magiorkinis, D. Paraskevis, A. Katsoulidou, R. Belshaw, C. Fraser, O. G. Pybus, A. Hatzakis, Integrating phylodynamics and epidemiology to

estimate transmission diversity in viral epidemics. *PLOS Comput. Biol.* **9**, e1002876 (2013). Medline doi:10.1371/journal.pcbi.1002876

49. J. C. Iles, G. L. Abby Harrison, S. Lyons, C. F. Djoko, U. Tamoufe, M. Lebreton, B. S. Schneider, J. N. Fair, F. M. Tshala, P. K. Kayembe, J. J. Muyembe, S. Edidi-Basepeo, N. D. Wolfe, P. Klenerman, P. Simmonds, O. G. Pybus, Hepatitis C virus infections in the Democratic Republic of Congo exhibit a cohort effect. *Infect. Genet. Evol.* **19**, 386–394 (2013). Medline doi:10.1016/j.meegid.2013.01.021

50. P. Beheyt, Contribution a l'etude des hepatites en Afrique. L'hepatite epidemique et l'hepatite par inoculation. *Ann. Soc. Belg. Med. Trop.* **33**, 297–340 (1953).

51. J. D. de Sousa, V. Müller, P. Lemey, A. M. Vandamme, High GUD incidence in the early 20 century created a particularly permissive time window for the origin and initial spread of epidemic HIV strains. *PLOS ONE* **5**, e9936 (2010). Medline doi:10.1371/journal.pone.0009936

52. M. T. Gilbert, A. Rambaut, G. Wlasiuk, T. J. Spira, A. E. Pitchenik, M. Worobey, The emergence of HIV/AIDS in the Americas and beyond. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 18566–18570 (2007). Medline doi:10.1073/pnas.0705329104

53. J. Hemelaar, E. Gouws, P. D. Ghys, S. Osmanov; WHO-UNAIDS Network for HIV Isolation and Characterisation, Global trends in molecular epidemiology of HIV-1 during 2000-2007. *AIDS* **25**, 679–689 (2011). Medline doi:10.1097/QAD.0b013e328342ff93

54. C. Kuyu, *Les Haitiens au Congo* (L'Harmattan, Paris, 2006).

55. Institut National de la Statistique, *Étude Socio-Demographique de Kinshasa, 1967: Rapport General* (Institut National de la Statistique, Kinshasa, 1969).

56. G. Bonacci, "Kuyu, Camille. – Les Haïtiens au Congo," in *Cahiers d'Etudes Africaines* (L'Harmattan, Paris, 2008), vol. 192, p. 895.

57. K. Jochelson, M. Mothibeli, J. P. Leger, Human immunodeficiency virus and migrant labor in South Africa. *Int. J. Health Serv.* **21**, 157–173 (1991). Medline doi:10.2190/11UE-L88J-46HN-HR0K

58. M. H. Schierup, R. Forsberg, in *Proceedings of the Conference: Origins of HIV and Emerging Persistent Viruses*, 28 to 29 September 2001 (Accademia Nazionale dei Lincei, Rome, 2003), vol. 187, pp. 231–245.

59. M. H. Schierup, J. Hein, Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**, 879–891 (2000). Medline

60. R. A. Neher, T. Leitner, Recombination rate and selection strength in HIV intra-patient evolution. *PLOS Comput. Biol.* **6**, e1000660 (2010). Medline doi:10.1371/journal.pcbi.1000660

61. M. J. Ward, S. J. Lycett, M. L. Kalish, A. Rambaut, A. J. Leigh Brown, Estimating the rate of intersubtype recombination in early HIV-1 group M strains. *J. Virol.* **87**, 1967–1973 (2013). Medline doi:10.1128/JVI.02478-12

62. P. Lemey, O. G. Pybus, A. Rambaut, A. J. Drummond, D. L. Robertson, P. Roques, M. Worobey, A. M. Vandamme, The molecular population genetics of HIV-1 group O. *Genetics* **167**, 1059–1068 (2004). Medline doi:10.1534/genetics.104.026666

63. J. O. Wertheim, M. Fourment, S. L. Kosakovsky Pond, Inconsistencies in estimating the age of HIV-1 subtypes due to heterotachy. *Mol. Biol. Evol.* **29**, 451–456 (2012). Medline doi:10.1093/molbev/msr266

64. L. de Saint-Moulin, *Villes et Organisation de l'Espace en République Démocratique du Congo* (L'Harmattan, Paris, 2010).

65. HIV Sequence Database, http://hiv.lanl.gov/ .

66. K. Katoh, K. Kuma, H. Toh, T. Miyata, MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005). Medline doi:10.1093/nar/gki198

67. L. Baeck, "Léopoldville, phénomène urbain Africain," in *Zaïre: Revue Congolaise* (Belgian African Review, Leopoldville, 1956), pp. 613–636.

68. J. C. Avise, *Phylogeography: The History and Formation of Species* (Harvard Univ. Press, Cambridge, MA, 2000).

69. UNAIDS/WHO, "Congo - UNAIDS/WHO epidemiological fact sheets on HIV and AIDS, 2008 Update" (UNAIDS/WHO Working Group on Global HIV/AIDS and STI, Geneva, 2008).

70. UNAIDS/WHO, "Democratic Republic of Congo - UNAIDS/WHO epidemiological fact sheets on HIV and AIDS, 2008 Update" (UNAIDS/WHO Working Group on Global HIV/AIDS and STI, Geneva, 2004).

71. S. Tavaré, "Some probabilistic and statistical problems in the analysis of DNA sequences," in *Some Mathematical Questions in Biology: DNA Sequence Analysis*, M. S. Waterman, Ed. (American Mathematical Society, Providence, RI, 1986), pp. 57–86.

72. Z. Yang, Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **39**, 105–111 (1994). Medline doi:10.1007/BF00178256

73. A. J. Drummond, S. Y. Ho, M. J. Phillips, A. Rambaut, Relaxed phylogenetics and dating with confidence. *PLOS Biol.* **4**, e88 (2006). Medline doi:10.1371/journal.pbio.0040088

74. A. J. Drummond, M. A. Suchard, Bayesian random local clocks, or one rate to rule them all. *BMC Biol.* **8**, 114 (2010). Medline doi:10.1186/1741-7007-8-114

75. C. J. Edwards, M. A. Suchard, P. Lemey, J. J. Welch, I. Barnes, T. L. Fulton, R. Barnett, T. C. O'Connell, P. Coxon, N. Monaghan, C. E. Valdiosera, E. D. Lorenzen, E. Willerslev, G. F. Baryshnikov, A. Rambaut, M. G. Thomas, D. G. Bradley, B. Shapiro, Ancient hybridization and an Irish origin for the modern polar bear matriline. *Curr. Biol.* **21**, 1251–1258 (2011). Medline doi:10.1016/j.cub.2011.05.058

76. P. Lemey, A. Rambaut, A. J. Drummond, M. A. Suchard, Bayesian phylogeography finds its roots. *PLOS Comput. Biol.* **5**, e1000520 (2009). Medline doi:10.1371/journal.pcbi.1000520

77. A. J. Drummond, M. A. Suchard, D. Xie, A. Rambaut, Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012). Medline doi:10.1093/molbev/mss075

78. D. L. Ayres, A. Darling, D. J. Zwickl, P. Beerli, M. T. Holder, P. O. Lewis, J. P. Huelsenbeck, F. Ronquist, D. L. Swofford, M. P. Cummings, A. Rambaut, M. A. Suchard, BEAGLE: An application programming interface and high-performance computing library for statistical phylogenetics. *Syst. Biol.* **61**, 170–173 (2012). Medline doi:10.1093/sysbio/syr100

79. M. A. Suchard, A. Rambaut, Many-core algorithms for statistical phylogenetics. *Bioinformatics* **25**, 1370–1376 (2009). Medline doi:10.1093/bioinformatics/btp244

80. B. T. Grenfell, O. G. Pybus, J. R. Gog, J. L. Wood, J. M. Daly, J. A. Mumford, E. C. Holmes, Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327–332 (2004). Medline doi:10.1126/science.1090727

81. P. Lemey, A. Rambaut, J. J. Welch, M. A. Suchard, Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* **27**, 1877–1885 (2010). Medline doi:10.1093/molbev/msq067

82. N. R. Faria, M. A. Suchard, A. Rambaut, P. Lemey, Toward a quantitative understanding of viral phylogeography. *Curr. Opin. Virol.* **1**, 423–429 (2011). Medline doi:10.1016/j.coviro.2011.10.003

83. E. C. Holmes, Evolutionary history and phylogeography of human viruses. *Annu. Rev. Microbiol.* **62**, 307–328 (2008). Medline doi:10.1146/annurev.micro.62.081307.162912

84. M. A. R. Ferreira, M. A. Suchard, Bayesian analysis of elapsed times in continuous-time Markov chains. *Can. J. Stat.* **36**, 355–368 (2008). doi:10.1002/cjs.5550360302

85. V. N. Minin, M. A. Suchard, Counting labeled transitions in continuous-time Markov models of evolution. *J. Math. Biol.* **56**, 391–412 (2008). Medline doi:10.1007/s00285-007-0120-8

86. V. N. Minin, M. A. Suchard, Fast, accurate and simulation-free stochastic mapping. *Philos. Trans. R. Soc. London Ser. B* **363**, 3985–3995 (2008). Medline doi:10.1098/rstb.2008.0176

87. J. D. O'Brien, V. N. Minin, M. A. Suchard, Learning to count: Robust estimates for labeled distances between molecular sequences. *Mol. Biol. Evol.* **26**, 801–814 (2009). Medline doi:10.1093/molbev/msp003

88. R Core Team, R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, 2013); www.r-project.org/.

89. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer, New York, 2009).

90. M. S. Gill, P. Lemey, N. R. Faria, A. Rambaut, B. Shapiro, M. A. Suchard, Improving Bayesian population dynamics inference: A coalescent-based model for multiple loci. *Mol. Biol. Evol.* **30**, 713–724 (2013). Medline doi:10.1093/molbev/mss265

91. A. Gelman, X. L. Meng, Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Stat. Sci.* **13**, 163–185 (1998).

92. Y. Ogata, A Monte Carlo method for high dimensional integration. *Numer. Math.* **55**, 137–157 (1989).

93. W. Xie, P. O. Lewis, Y. Fan, L. Kuo, M. H. Chen, Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.* **60**, 150–160 (2011). Medline doi:10.1093/sysbio/syq085

94. T. C. Bruen, H. Philippe, D. Bryant, A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172**, 2665–2681 (2006). Medline doi:10.1534/genetics.105.048975

95. C. Yang, M. Li, J. L. Mokili, J. Winter, N. M. Lubaki, K. M. Mwandagalirwa, M. J. Kasali, A. J. Losoma, T. C. Quinn, R. C. Bollinger, R. B. Lal, Genetic diversification and recombination of HIV type 1 group M in Kinshasa, Democratic Republic of Congo. *AIDS Res. Hum. Retroviruses* **21**, 661–666 (2005). Medline doi:10.1089/aid.2005.21.661

96. B. Bikandou, J. Takehisa, I. Mboudjeka, E. Ido, T. Kuwata, Y. Miyazaki, H. Moriyama, Y. Harada, Y. Taniguchi, H. Ichimura, M. Ikeda, P. J. Ndolo, M. Y. Nzoukoudi, R. M'Vouenze, M. M'Pandi, H. J. Parra, P. M'Pelé, M. Hayami, Genetic subtypes of HIV type 1 in Republic of Congo. *AIDS Res. Hum. Retroviruses* **16**, 613–619 (2000). Medline doi:10.1089/088922200308837

97. B. Foley, H. Pan, S. Buchbinder, E. L. Delwart, Apparent founder effect during the early years of the San Francisco HIV type 1 epidemic (1978–1979). *AIDS Res. Hum. Retroviruses* **16**, 1463–1469 (2000). Medline doi:10.1089/088922200750005985

98. F. Bielejec, A. Rambaut, M. A. Suchard, P. Lemey, SPREAD: Spatial phylogenetic reconstruction of evolutionary dynamics. *Bioinformatics* **27**, 2910–2912 (2011). Medline doi:10.1093/bioinformatics/btr481