# Impacts of Misspecifying the Evolutionary Model in Phylogenetic Tree Estimation

Tom Burr
Safeguards Systems Group, NIS-7
Los Alamos National Laboratory
Los Alamos, NM 87545

Gerry Myers
Biology Group, B-1
Los Alamos National Laboratory
Los Alamos, NM 87545

James M. Hyman
Mathematical Modeling, T-7
Los Alamos National Laboratory
Los Alamos, NM 87545

Alexei Skourikhine
Safeguards Systems Group, NIS-7
Los Alamos National Laboratory
Los Alamos, NM 87545

**Abstract** *We consider phylogenetic tree estimation with emphasis on estimating the number of groups (clades). We rarely know the full evolutionary model, so we want to understand the impact of model estimation errors. Sensitivity to misspecifying the model or model parameters depends on how distinct the clades are, so it is important to consider differing degrees of clade resolution. We do this by varying the macroscopic growth rate and microscopic mutation rate of the taxa. We simulate DNA sequence data using coalescent theory to simulate the sample genealogy, and one of several mutation models. For each case, we compute all pairwise distances between sequences using the true and several alternate models. The within-group variance (with distance data represented in principal coordinates) is used to choose the number of clades. We conclude that the estimated number of clades can be sensitive to model estimation errors, to an extent determined by clade resolution.*

*Keywords: evolutionary model, coalescence, clades, distance measures, HIV, AIDS.*

## 1. Introduction

This investigation was motivated by the unresolved question of why there are 11 approximately equi-distant subtypes (“clades” or groups) of HIV-1, type M [1, 2]. One plausible qualitative explanation is based on the rapid growth of the macroscopic AIDS epidemic and the large variance in the number of new HIV cases produced by existing cases. More quantitative evidence is available by computing the likelihood of the current DNA data from randomly selected sequences of each of the 11 subtypes under “forward models” that specify the growth rate and dynamics of the macroscopic epidemic and the microscopic mutation model. That is, if 11 equidistant subtypes can emerge for certain combinations of the macroscopic epidemic growth rate and microscopic mutation rate, then the 11 subtypes can be “explained” as a natural consequence of the dynamics of the disease.

Recently, a very different hypothesis has been suggested [3] involving the possibility of an inadvertent wide-scale spread of HIV-1 through oral polio vaccine (OPV) trials. The evidence for the OPV hypothesis includes classic epidemiology, including the facts that chimpanzee kidneys were used for some stocks of the virus in central Africa in 1957-1960, and the observed rate of incidence of initial HIV-1 cases clustered around the centers of the OPV trials. There is counter-evidence that the OPV trials could not have added to the AIDS epidemic [1].

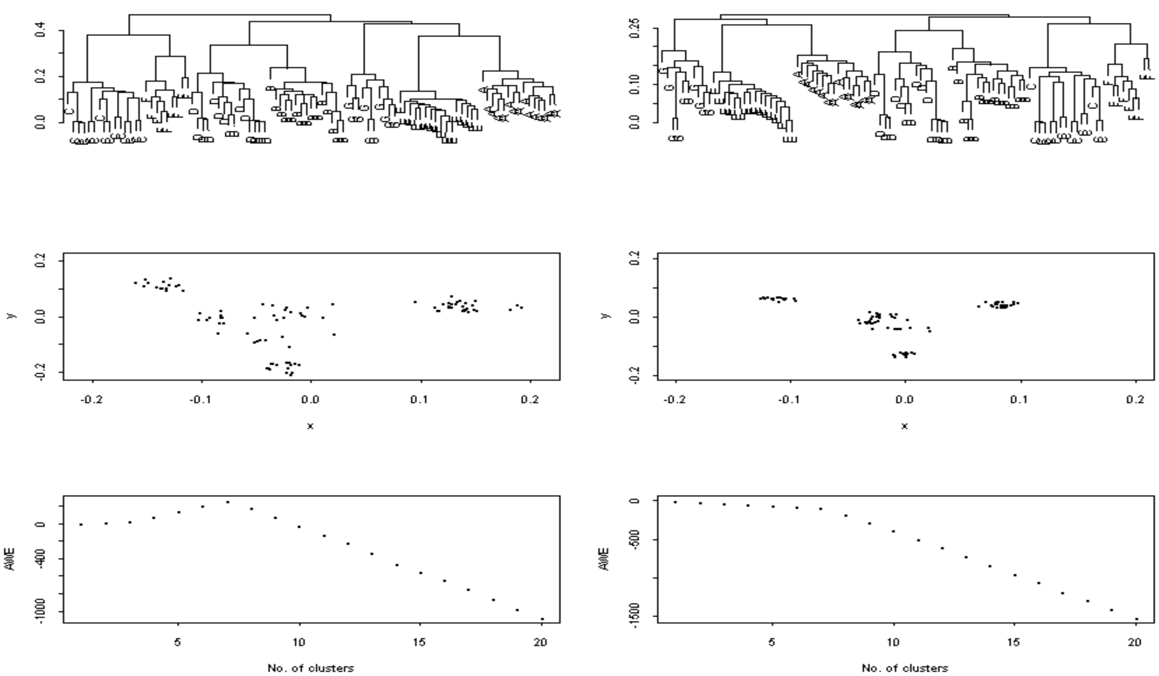
We do not take a position here on the OPV hypothesis. Instead, because the issue of how many groups are present in DNA data is generic, we consider the sensitivity of group assignments to inevitable misspecifications of the microscopic evolutionary model. We use simulated data from Treevolve ([4], *http://evolve.zoo.ox.ac.uk*), and then apply several distance measures to define clades under varying degrees of clade resolution. This approach allows us to study the impact of model estimation errors on the estimated number of clades that naturally arise. Section two gives an example using HIV data. We then describe the coalescent process we use to compute the probabilities of each possible sample genealogy. After describing the mutation models, we introduce distance measures derived from these models, and introduce one way to choose the number of clades. Finally, we present our simulation results and conclude that the estimated number of clades can be sensitive to model estimation errors, to an extent determined by clade resolution.

## 2. Background

This study addresses the impact of misspecifying the evolutionary model on the quantitative assessment of how many clades are present in a sample of DNA sequences. In practice, we cannot know the evolutionary model and all of its parameters exactly, so it is important to understand the impact of model misspecification and parameter estimation on the estimated number of clades.

In the hierarchical cluster plot in Figure 1a, the best estimate of the correct distance measure for the evolutionary model is used to compute all pairwise distances among 100 randomly selected HIV-1, env sequences (7 subtypes A-G are all used here; all are available at hiv-web.lanl.gov, and the accession numbers are available upon request). In Figure 1b, the distance measure for the evolutionary model is approximated by an alternate model (the Jukes-Cantor model). Sections 4, 7, and the Appendix have more detail about models and their approximations.

Notice that the dis tances in Figure 1a are more dispersed, leading to greater within group and between group variation. This dispersion difference is easier to observe in the middle plot (a principle coordinate plot) which displays the sequences in a way that the pairwise distances can be computed approximately using only the *x* and *y* coordinates. Notice (top plots) that even with these differences, the clade assignments would probably be the same for (a) and (b) if we chose 7 clades in each case. The bottom plot shows a maximum in the approximate weight of evidence (AWE, section 6) occurring for 7 groups in case (a) and for 1 group in case (b).




**(a)** **(b)**

Figure 1: 100 HIV env sequences using the best estimate of the true model (a) and an alternate (Jukes-Cantor) (b) model. The top plot is hierarchical clustering; the middle plot is a principal coordinate plot; and the bottom plot is the approximate weight of evidence for candidate numbers of clusters.

## 3. Coalescence

Coalescent theory [5, 6] arose from the need to infer aspects of the past by sampling from present-day populations. We start with a sample and trace a possible evolutionary path back in time to identify events that occurred en route to the most recent common ancestor (MRCA) of the sample. For example, suppose a population is size N for many successive generations. In the current generation, randomly select $n = 2$ cases and ask: what is the probability that both cases came from the same "parent" in the previous generation? A "randomly toss balls into boxes" argument is applicable, and for this simple example we readily see that both cases share the same parent with probability $1/N$. More generally, the probability that the coalescent to the MRCA occurred $t$ generations ago is given by the geometric distribution probability $1/N\ (1-1/N)^t$, and for large $N$ this is well approximated by an exponential distribution. The theory easily generalizes to compute the probability distribution for the times to each of the $n - 1$ coalescent events until all $n$ cases share a MRCA [5].

This approach continues to expand into new application areas and has several attractive features. First, it is sample-based rather than population based. Second, it leads to highly efficient algorithms (example: Treevolve*)* for simulating samples of DNA sequences from populations that have been changing under a wide range of population genetics models (for example, allowing for migration among partially subdivided populations, and different macroscopic growth rates of populations). Third, it is suitable for DNA or other molecular data (provided the macroscopic branching process that describes the population growth is independent of the mutational process [4]).

Treevolve has several features that allow for a wide range in the number of clades and clade resolution. For example, it allows the evolutionary time period to be divided in time windows with different growth (or decline) rates in each window, and it allows the user to specify the variance in the number of "offspring" (new AIDS cases for example) that each "parent" (existing AIDS cases for example) produces. Qualitatively, clades will be strongly resolved if they are well separated from the tree root (long branches in the top plot of Fig 1 for example). Equivalently, from the coalescent point of view, if clade 1 has a recent MRCA that is well separated genetically from the MRCA for clade 2, and both MRCA's are recent enough to have small within-clade variation, then clades 1 and 2 are easily distinguished.

## 4. Mutation Models

We consider aligned sequences of DNA data with no gaps, such as the example with 3 taxa in Table 1.

Table 1: Example simulated (using Treevolve) aligned DNA sequence data**.**

| Taxa Code | Mutually Aligned DNA |
|---|---|
| 1 | CCCGATCAAATT… |
| 2 | CACGCTCAAATT… |
| 3 | CGCGAACAAATT… |

An evolutionary model specifies the probability per unit time of a given nucleotide mutating to another nucleotide. A fully parameterized 4-by-4 transition probability matrix would have 12 freely varying entries, but most current analyses and software restrict the number of free parameters to 1 to 5. Most of these models allow a distance measure (section 5) to be defined that properly accounts for the assumed model. One of the most common currently used is a 5-parameter model (HKY5) that specifies positive $\mu$, $\kappa$ and the four nonnegative nucleotide frequencies $\pi_A$, $\pi_C$, $\pi_G$, and $\pi_T$ (that sum to 1). The parameter $\mu$ determines the average rate of change (usually assumed to be constant over time and to be the same for each taxa). Typically, the number of changes per unit time is assumed to be Poisson distributed with mean $\mu$. The parameter $\kappa$ allows for purine-to-purine or pyrimidine-to-pyrimidine mutations (called transitions) to be different than purine-to-pyrimidine or pyrimidine-to-purine (called transversions). Although the $\pi$s can be estimated using the observed nucleotide frequencies, the best way to estimate $\mu$ and $\kappa$ requires access to an outgroup taxa.

Recently, it has been demonstrated that allowing $\mu$ to vary across site can be important

[7], and especially so in our context of estimating the number of clades [8] where long branches tend to attract (group together). Typically, $\mu$ is assumed to have a gamma distribution with the parameter $\gamma$ determining its variation across sites (large $\gamma$ means less variation).

It is currently believed [9] that substitution model parameter estimation is somewhat robust to misspecifying the tree topology. Still, at best we have a challenging statistical estimation problem, so in practice we will at least introduce errors in approximating the evolutionary model or model parameters. Also, currently almost no software is available for allowing site-to-site dependence in the mutation probabilities (limited exceptions ([10] and others) use hidden Markov models to identify DNA sections of similar rates). This is another reason that models are likely to be misspecified to some degree in practice. However, our simulated data does not have site-to-site dependence, so we can exactly specify the correct model and then deliberately misspecify a model by varying parameter values to study the impact of model misspecification on clade assignments.

## 5. Distance Measures

All distance measures attempt to compute a distance that is expected (on average) to increase approximately linearly or in a known way with time to the MRCA. The simplest model is Jukes-Cantor (JC1) which assumes all mutational possibilities are equally likely. For the JC1 model, the distance between taxa $x$ and $y$ is $d_{xy}$ = -3/4 log(1 - 4/3 $D$), where $D$ is the percent of sites that differ between $x$ and $y$. When D reaches its "saturation limit" of _ the distance is infinite (by chance and reversible mutation, eventually any two sequences should agree at approximately 25% of their sites).

Some models allow different base pair frequencies, the $\kappa$ parameter defined in Section 4, and the observed numbers of each type of mutation (A to C, A to G, A to T, etc.). Many of the more elaborate models also have distance measures [9], which we compute using DNADIST [10], PAML [11], or our own code using S-plus [12]. The Fig. 1 distances were computed using the general reversible model [7, 9] with rate heterogeneity in (a) and JC1 in (b).

## 6. Clade Assignments

There are several methods for deciding how many clades are present and which taxa belong to which clades. In a generic sense, this is an unsupervised learning problem, and cluster assignments often depend strongly on the distance measure used. One common way is to resample the sequences (bootstrap) $n$ times and count the fraction of times that the specified subsets remain clustered using any of several tree building methods [9].

In any method, there is at least an implicit assumption about the evolutionary model. Also, we anticipate that the assumed distance measure (or evolutionary model) will impact the clade assignments regardless of how those assignments are made.

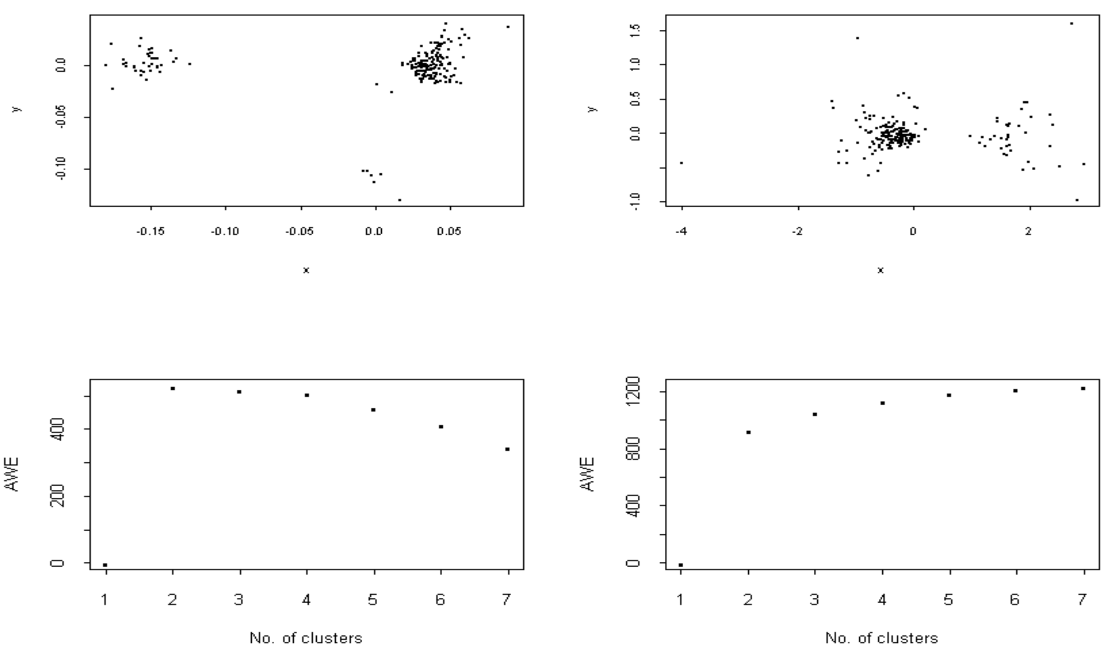
Here we report results from a novel and convenient way to choose clade assignments. This model-based clustering (mclust [13], as implemented in available software [12]), provides a semi-objective way to choose the number of clusters. Qualitatively, mclust is similar to the well-known "look for diminishing returns" approach that is often used in k-means clustering: add clusters until the reduction in within group sum of squares begins to diminish sharply.

Our method of defining clades involves three steps: (1) assume an evolutionary model and use it to compute distances among all pair of taxa; (2) represent the pairwise distance matrix using principal coordinates (Fig 1), which provides a low-dimensional representation of the data that closely preserves the distances, and (3) choose the number of clades based on the mclust algorithm when applied to the principal coordinate [12] data (Fig.1). We have compared step (3) to the more common "diminishing returns with k-means approach" and have noticed a tendency for mclust to suggest fewer groups than k-means. Because our focus is on the impact of model misspecification, we report results for only one "clade assignments" method (mclust). However, we are currently investigating other methods to choose the number of clades and the stated evidence for that choice.

In Fig. 2 we present the results of our three steps and note that the 2 clades in Fig. 2a (using the true model) are more distinct than the 2-7 clades in Fig. 2b (using a wrong model, with $\kappa$ = 2 rather than $\kappa$ = 1.2 and rate heterogeneity parameter $\gamma$ = 2 (smaller rate variance in $\mu$ across sites) rather than $\gamma$ = 0.4 (larger rate variance)). Qualitatively, we note that the case in Fig. 2a has relatively strong evidence for 2 tightly clustered clades, while the case in Fig. 2b has weaker evidence for between 2 to 7 weakly clustered clades. We provide a semi-quantitative measure of the evidence $E$ for the chosen number of clades in the section 7 simulation results that gives E = 0.09 for $c$ = 2 clades in Fig. 2a and E = 0.05 for $c$ = 7 to 13 clades in Fig. 2b. We select the clade number(s) $c$ that maximize the estimated approximate weight of evidence (AWE) as defined by mclust. And we define $E$ by normalizing AWE to NAWE (NAWE is nonnegative and sums to one), and $E$ = maximum(NAWE), which is a measure of how peaked the AWE curve is.

# 7. Simulation Results

In this section we present results of assigning simulated DNA from 100 taxa to clades under several models for both the macroscopic branching process that describes the population growth and the microscopic mutational process. We illustrate that when clades are well resolved, the clade assignments are not sensitive to the choice of distance measure. However, when clades are not well resolved, the identification of specific clades is sensitive to the evolutionary models. The advantage of using simulated data for this type of study is that we know the exact mutation model (and all its parameters) used in the simulations. We refer to the exact model as the “baseline” model or “true” model; all other models are referred to as “alternate” models.




**(a)** **(b)**

Figure 2. 200 simulated sequences with distances estimated using the (a) true ($\kappa$= 2, $\gamma$ = 0.4) and (b) alternate ($\kappa$= 1.2, $\gamma$=2) models. The top plot is a principal coordinate plot and the bottom plot is the approximate weight of evidence (AWE) for candidate numbers of clusters.

Table 2 presents results ($c$ and $E$) for the baseline HKY5 model with $\pi_A = \pi_C = \pi_T = \pi_G = 0.25$, $\kappa = 2$, $\mu = 0.003$ per site per year plus one rate heterogeneity parameter $\gamma$ that describes how $\mu$ varies across nucleotide sites, and results for five alternative models for 3 main cases with 3 subcases (each with different model parameters) per case. Because we chose equal $\pi$s, HKY5 is equivalent to Kimura-2 [9].

Subcases 1.1-1.3 assume zero population growth followed by exponential population growth, and have $\gamma = 0.3$, 1, and 2, respectively. Subcases 2.1-2.3 assume exponential population growth, and have $\gamma = 0.3$, 1, and 2, respectively. Subcases 3.1-3.3 assume zero population growth, and have $\gamma = 0.3$, 1, and 2, respectively.

The alternate models A-D are chosen to be within the range of statistical uncertainty (for 400 nucleotide sites) due to model parameter estimation if the phylogenetic software can calculate distance measures under the correct model and therefore only needs to estimate model parameters. Case E is the JC1 model with $\gamma$ assumed to be infinite. The $(\gamma,\kappa)$ values for each model are specified in Table 2. Note that the impact of misspecifying $\gamma$ is larger for the $\gamma = 0.3$ cases (cases 1.1 (B), 3.1 (D), env (A-E)).

The real data for env sequences (plotted in Fig. 1) are analyzed as the last case in Table 2. We observe that the estimated number of clades (ranging from 7 to 1) is extremely sensitive to the model misspecifications considered here for env, but only slightly sensitive for our simulated data. Also, because the env sequences are real data, we can not know the true model, so we refer to the best model as the baseline model. Therefore, for our purposes here, the baseline model case is the same as for subcases 1.1, 2.1, and 3.1. From our own estimates and from [7] we know that a better model for env sequences is the general reversible model [9] with $\gamma = 0.3$, but we have found that this model (used in Figure 1a) also estimates 7 clades with 0.09 weight of evidence. To make all our table entries consistent it was more convenient to always use HKY-5 as described as the baseline model.

## 8. Conclusions

To our knowledge, this is the first quantitative study that begins to "calibrate" the effect of the difference between distance measures (all based on evolutionary models) as applied to DNA sequences for the purpose of choosing the number of clades. We have begun to confirm that different distance measures give more varied results when clade resolution is vague. Therefore, we have also begun to quantify how well separated the clades must be to ensure that all distance measures give essentially the same

Table 2: Estimated number of clades, $c$, and a measure of clade resolution/evidence, $E$, using the baseline model and 5 alternate models A, B, C, D, and E for 10 cases (9 simulated, 1 real example).

| **Case** | **Baseline HKY5 Model** $\kappa$**=2,** $\pi$**s equal** | | **Alternate Models** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | **A** $\kappa = 1.3$ | | **B** $\kappa = 1.3$ | | **C** $\kappa = 2.6$ | | **D** $\kappa = 2.6$ | | **E** (JC1) $\kappa = 1$ | |
| | $\gamma$ | **c, E** | $\gamma$ | **c, E** | $\gamma$ | **c,E** | $\gamma$ | **c,E** | $\gamma$ | **c,E** | $\gamma$ | **c,E** |
| **1.1** | 0.3 | 3, 0.09 | 0.2 | 3, 0.09 | 0.4 | **2**, 0.09 | 0.2 | 3, 0.09 | 0.4 | 3, 0.09 | | 3, 0.09 |
| **1.2** | 1 | 4, 0.09 | 0.9 | 4, 0.09 | 1.1 | 4, 0.09 | 0.9 | 4, 0.09 | 1.1 | 4, 0.09 | | 4, 0.09 |
| **1.3** | 2 | 3, 0.09 | 1.4 | 3, 0.09 | 3.1 | 3, 0.09 | 1.4 | 3, 0.09 | 3.1 | 3, 0.08 | | 3, 0.09 |
| **2.1** | 0.3 | 1, 0.10 | 0.2 | 1, 0.10 | 0.4 | 1, 0.10 | 0.2 | 1, 0.10 | 0.4 | 1, 0.10 | | 1, 0.10 |
| **2.2** | 1 | 1, 0.10 | 0.9 | 1, 0.10 | 1.1 | 1, 0.10 | 0.9 | 1, 0.10 | 1.1 | 1, 0.10 | | 1, 0.10 |
| **2.3** | 2 | 1, 0.10 | 1.4 | 1, 0.10 | 3.1 | 1, 0.10 | 1.4 | 1, 0.10 | 3.1 | 1, 0.10 | | 1, 0.10 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **3.1** | 0.3 | 2, 0.08 | 0.2 | 2, 0.10 | 0.4 | 2, 0.07 | 0.2 | 2, 0.10 | 0.4 | **4**, 0.07 | | 2, 0.09 |
| **3.2** | 1 | 3, 0.08 | 0.9 | 3, 0.10 | 1.1 | 3, 0.07 | 0.9 | 3, 0.10 | 1.1 | 3, 0.07 | | 3, 0.09 |
| **3.3** | 2 | 3, 0.09 | 1.4 | 3, 0.09 | 3.1 | 3, 0.09 | 1.4 | 3, 0.09 | 3.1 | 3, 0.09 | | 3, 0.09 |
| **env** | 0.3 | 7, 0.08 | 0.2 | **1**, 0.09 | 0.4 | **1**, 0.09 | 0.2 | **1**, 0.09 | 0.4 | **1**, 0.09 | | **1**, 0.08 |

results. To do so, we used Treevolve to rapidly experiment with different clade resolutions, obtained from different combinations of macroscopic growth rate of the population and microscopic mutation rate. This also allowed us to better understand the "forward" processes that lead to well-resolved clades. The only case that gave the estimated number of clades for all distance measures was the case with one clade (which occurred with exponential growth). Concerning model estimation, many existing codes assume an infinite $\gamma$parameter (which means that all sites have the same mutation rate $\mu$). Because of the potential impact on the estimated number of clades, particularly for $\gamma$ < ,1this study illustrates the importance of estimating $\gamma$. Future work will include alternate ways to choose the number of clades (such as bootstrap percentages) because here we used only the "model-based" (mclust) evidence.

## References


[1] Hahn, B. Shaw, G., De, K., and Sharp, P., "AIDS as a Zoonosis: Scientific and Public Health Implications," *Science* **287**, 607-614 (2000).

[2] Myers, G., Korber, B., Foley, B., Jeang, K., and Mellors, J., et al, editors, Human Retroviruses and AIDS, Theoretical Biology and Biophysics Group, Los Alamos, NM (1996).

[3] Hooper E., The River: A Journey to the Source of HIV and AIDS, Little, Brown, New York (1999).

[4] Grassley, N., Harvey, P., and Holmes, E., "Population Dynamic of HIV-1 Inferred from Gene Sequences" *Genetics* **151**, 427-438 (1999).

[5] Fu, Y., and Li, W., "Coalescing into the 21st Century: An Overview and Prospects of Coalescent Theory," *Theoretical Population Biology* **56**, 1-10 (1999).

[6] Kingman, J.C., "On the Genealogy of Large Populations," *J. Appl. Probab.* **A 19**, 27-43 (1982).

[7] Leitner, T. et al, "Tempo and Mode of Nucleotide Substitutions in gag and env Gene Fragments in HIV Type 1 Populations with a Known Transmission History," *Virology* **71(6)**: 4761-4770, (1997).

[8] Van de Peer, Y., Janssens, W., et al, "Phylogenetic Analysis of the env Gene of HIV-1: Isolates Taking into Account Individual Nucleotide Substitution Rates," *AIDS* **10**, 1485-1494 (1996).

[9] Swofford et. al., "Phylogeny Inference," Molecular Systematics, 2nd edition, eds. Hillis et. al. (1996).

[10] Felsenstein, J. PHYLIP: Phylogeny Inference Package, ver 3.5c, Univ. of Washington, Seattle (1993).

[11] Yang, Z., PAML: Phylogenetic Analysis by Maximum Likelihood, Version 2.0, 1999

[12] S-Plus 5.1, MathSoft, Seattle Wash. (1999).

[13] Banfield, J., and Raftery, A., "Model-based Gaussian and non-Gaussian Clustering," *Biometrics* **49 (3)**, 803-822 (1993).


## Appendix. Treevolve Model Details

We describe the three macroscopic models, each with three similar microscopic models used in Table 2. All baseline models were simulated using Treevolve 1.30 with inputs: seq. length = 400, sample size = 100, mutation rate $\mu$ = 0.003, microscopic substitution model = HKY5; transition/transversion ratio = 2; $\pi_A = \pi_C = \pi_G =$

$\pi_T$ = 0.25, discrete gamma rate heterogeneity shape $\gamma$ = 0.3 (subcase 1), 1 (subcase 2), or 2 (subcase 3) with 8 rate categories; haploid model, (generation time) / (variance in offspring no.) = 0.001. All 3 cases used population size at time 0 of $N_0$ =2.5 $x$ $10^6$, and $N = N_0\ e^{-rt}$ as time moves backward, with case 1 having an 8–year $r$ = 0.693 period ($N$ doubles yearly going forward in time toward the present time 0) followed by 20 year $r$ =0.05 period, followed by 10000 year $r$ = 0 period; case 2 used one period with $r$ = 0.693; and case 3 used one period with $r$ = 0 (no growth).

We expected and observed: a "star-phylogeny" having 1 clade for case 2 (2.1, 2.2, and 2.3 in Table 2); two or more clades for cases 1 and 3, with better clade resolution for case 1. We did observe slightly better clade resolution for case 1 than case 3, and the strongest clade resolution (evidence $E$) occurred for case 2.