MIKKEL H. SCHIERUP[a], ROALD FORSBERG[a]

RECOMBINATION AND PHYLOGENETIC ANALYSIS OF HIV-1

INTRODUCTION

The acquired immunodeficiency syndrome (AIDS) caused by the human immunodeficiency virus (HIV) virus is one of the most devastating infectious diseases to emerge in modern times. To investigate the circumstances that led to this human catastrophe is an important scientific pursuit as it may provide valuable knowledge for the battle against HIV and enable us to take qualified precautions against the emergence of other novel diseases. The main cause of the present day pandemic is viruses from the M group of the HIV type 1. The viruses from this group are believed to originate in the chimpanzee (*Pan troglodytes troglodytes*) (Gao *et al.* 1999), and in order to understand the M group's emergence as a human pathogen much effort has therefore focused on resolving the circumstances surrounding the species shift from monkey to man (e.g. Korber *et al.* 2000).

For this purpose, great emphasis has been placed on the reconstruction and timing of events in the evolutionary history of the M group from historical and present-day samples of the viral genetic sequence diversity. Ideally, such studies provide an objective reconstruction of epidemiological events, thereby circumventing the reliance on non-genetic data that may be missing or biased. The most ambitious of these studies have used phylogenetic analysis and have employed statistically advanced maximum-likelihood methods to date the origin of diversity (i.e. the time to the most recent common ancestor (TMRCA) to around 1931+/-10 years, with most of the recognizable subtypes within the group arising just a few years later (Korber *et al.* 2000). This result does not reveal anything about which species the most recent common ancestor (MRCA) resided

[a] Bioinformatics Research Center (BiRC) – University of Aarhus – Ny Munkegade – Building 540 – 8000 Aarhus C. (Denmark).

in. However, based on the observation that inter-species transmission of viruses is rare, several authors have argued that present day viruses from the human population most likely originate from a single species-transfer event. Using this parsimony argument in conjunction with the estimated date of the MRCA from phylogenetic studies, these authors have then arrived at the conclusion that the transmission of the virus from the presumed chimpanzee ancestor happened some time before 1940 (Korber *et al.* 2000; Weiss 2001).

An underlying assumption of phylogenetic analyses, like the above, is that a single phylogenetic tree can describe the evolutionary history of the sequences. Exchange of genetic material among distinct viruses through polymerase template switching violates this assumption. An important point to make, is that if recombination occurs, different parts of the HIV sequence have different histories with different phylogenetic trees. It is therefore not meaningful to relate the epidemiological history of the virus to a single tree through phylogenetic analysis, to discuss a single most recent common ancestor (MRCA), or to attempt to date events based on a single phylogenetic tree. Instead, we would expect a whole set of most recent common ancestors in different individuals at different times.

The ability of HIV viruses to recombine through template switching has been repeatedly demonstrated in vitro (Mikkelsen-Pedersen 2000) and indirect evidence for recombination in the wild type has been inferred through sequence analysis (Robertson *et al.*, 1995), leading for instance to the definition of circulating recombinant strains (Robertson *et al.* 2000). Previous phylogenetic studies have tried to circumvent effects of recombination by filtering out detectable recombinant sequences, but it is unclear to what extent this approach alleviates the problem, since recombination within subtypes and recombination events happening early in the evolution are very difficult, if not impossible, to detect, while they may still have important consequences for phylogenetic analysis (Wiuf *et al.* 2001).

So we are faced with a dilemma: the phylogenetic approach is very powerful but should not be applied to recombining data. However, it is intrinsically a very hard problem to confirm or reject the occurrence of recombination in the evolutionary history of a sequence sample and therefore to confirm the appropriateness of using the phylogenetic methodology. This means that when interpreting results from phylogeny based analyses of HIV data, we must take into consideration that the results may potentially be biased by the failure of the model to incorporate recombination.

It was previously demonstrated by Schierup and Hein (2000 a,b), that if recombination is ignored and phylogenetic methods are applied to re-

combining data the analysis may result in biased parameter estimates and false conclusions concerning biological hypotheses. In particular it was shown that: 1) the molecular clock may be falsely rejected, 2) the rate heterogeneity of evolution over the sequence may be inferred to be much larger than it actually is, and 3) the TMRCA inferred from a single tree may be different from the "average" of the set of MRCAs over the sequence, and 4) the amount of recent divergence is underestimated.

It is evident that especially the two latter points may potentially have great influence on the conclusions of phylogeny based dating analyses.

On the basis of these previous results, we here attempt to evaluate specifically the bias that recombination can induce into the dating of events in the evolutionary history of the M group of HIV-1 viruses. We apply a newly developed method to estimate rates of recombination in two HIV-1 group M data sets previously used in phylogenetic analysis. We then in turn use the estimated parameter of recombination to estimate the consequence of recombination on previous timing estimates through a simulation approach designed to mimic the dynamics of the HIV virus.

Materials and methods

Analysis of sequences

Two sequence data sets were analysed: 1) the "Korber" data set of Envelope sequences used by Korber *et al.* (2000), and 2) the "Vidal" data set used by Vidal *et al.* 2000. The Korber data set was downloaded already aligned from the website http://www.santafe.edu/~btk/science-paper/bette.html, whereas the Vidal data set was downloaded from Genbank and aligned using ClustalX (Thompson *et al.* 1997) (followed by adjustments by eye). In the Korber data set, sequences were already assigned to subtypes. In the Vidal data set, we assigned sequences to subtypes after realigning to a set of reference subtypes and determining which reference subtype a given sequence was most closely related to, as measured by the Hamming distance.

For test of recombination and estimation of recombination rate, the LDhat program (McVean *et al.*, 2000) was used. This program estimates a composite maximum likelihood curve for the data set for various values of the recombination rate ϱ, and uses as the maximum-likelihood estimate the recombination rate yielding the highest composite likelihood score. The algorithm is a finite sites extension of Hudson

(2001). The LDhat program also allows a permutation-based test of significance (against the hypothesis of no recombination ($\varrho = 0$)), and calculates the R2 correlation test for recombination used by Awadalla *et al* (1999).

The size of the two data sets prohibits a simultaneous analysis of all sequences. Instead, we divided the sequences of each data set into three different types of sub sets: 1) each subtype separately, 2) one of each subtype, and 3) a random sample of 30 sequences from the total set. We used a threshold of $P = 0.1$ for inclusion of sites, and set the maximum population recombination rate at $\varrho = 500$.

Simulations

Here we used a two-step strategy: First, we created a number of artificial data sets by doing coalescent simulations under different key-parameters, hereunder varying the amount of recombination between sequences. Second, we subjected the simulated data sets to maximum-likelihood phylogenetic analysis, which implicitly (and wrongly) assumes that the sequences in the data sets are related by a single tree.

The parameters varied were: the mutation rate, the rate of exponential growth and the recombination rate. Simulations were done using the "Coalescent with Exponential growth and Recombination" program, which was previously used by Schierup and Hein (2000) and which can be accessed through http://www.birc.dk/~mheide.

Briefly, the program implements Hudson's (1983) algorithm for the continuous time coalescent with recombination, allowing for the population size to be exponentially increasing forward in time (see Slatkin-Hudson 1991).

In the simplest case without exponential growth of the population, a set of sampled sequences is followed back in time, waiting for either coalescent events (sequences find a common ancestor) or recombination events (which split up a given sequence on two ancestors). The expected time to either a coalescent or recombination event is exponentially distributed (Hudson 1983) with the sum of the coalescent and recombination intensities. The intensity of coalescence when i ancestral sequences are present is $i(i-1)/2$, whereas the recombination intensity is $i\varrho$, where ϱ is the scaled recombination parameter.

If the population is growing exponentially (i.e. decreasing exponentially when viewed back in time), then the waiting time until the next coalescent event depends on the time parameter t, as outlined in Hudson and Slatkin (1991). The exponential growth parameter is here termed b.

This stochastic process results in a number of phylogenetic trees (with branch lengths) relating different parts of the sequence. From this a nucleotide data set of aligned sequences was simulated by adding mutations to the genealogical tree at L equally spaced positions in the sequence. Note that different positions are likely to have different genealogical trees when recombination occurs. We use the simplest possible substitution model, which is the Jukes-Cantor model, which assumes that the number of mutations at a given branch is Poisson distributed with mean mt, where m is the mutation rate and t is the branch length. Thus, we assumed that all sites have the same mutation rate.

Finally, since it is required by some of the subsequent phylogenetic analysis, we also simulated an outgroup sequence using the same substitution model.

Following the simulation of a sequence data set we used phylogenetic analysis to estimate various time parameters. Clearly, phylogenetic analysis is not appropriate when recombination occurs, but the focus here is to estimate any possible bias in doing just this (i.e. ignoring recombination). We chose to use maximum-likelihood methods since this is believed to be the least biased method, and it is this type of analysis Korber *et al.* (2000) used. Maximum-likelihood analysis allows one to estimate and incorporate rate heterogeneity (measured as α) in the analysis. This is important even though no rate heterogeneity was assumed in the simulation of the sequences, because ignoring recombination leads to a large apparent rate heterogeneity (Schierup-Hein 2000a; Worobey 2001). Thus, this study differs from Schierup and Hein (2000a), which did not consider artificially created apparent rate heterogeneity in their analysis.

We used the programs DNAdist and Fitch of Phylip 3.572 (Felsenstein 1995) to estimate a rooted (using the outgroup) neighbour joining topology of the tree. This topology was subsequently fed to the program PAML 3.0 (Yang 1999). PAML was used to estimate the rate heterogeneity a and the branch length of the tree assuming a molecular clock. From the resulting tree we then calculated various measures, such as the height of the tree (the time to the most recent common ancestor, TMRCA), the length of the terminal branches S, the pairwise difference P, and the total length of the tree T.

A similar analysis was done for data sets without an outgroup using the Phylip program Kitsch in place of Fitch.

For each parameter set investigated we repeated this process 100 times and we report mean and standard deviations over these runs.

Results

Analysis of HIV data sets

Table 1 shows results from applying the LDhat program to the Vidal data set. In all cases, the maximum-likelihood estimate of the recombination rate ϱ is extremely high (greater than the maximum rate of 500). If recombination occurs at a rate greater than 500, sites at opposite ends of the sequence are effectively uncoupled. However, only in 4 out of 7 cases was the rate found to be significantly different from zero as determined by the permutation test (McVean *et al.* in press). This implies that the likelihood curve as a function of ϱ is relatively flat and the estimated recombination rates will have very broad confidence intervals. Nevertheless, Table 1 does provide evidence that recombination occurs both within and among subtypes even within a short fragment of the Envelope gene (572 base pairs). The R2 test also provides evidence for recombination in the three subtypes where the result using LDhat was not significant. This implies that the likelihood test is not always the most powerful test.

While the existence of recombinant sequences was already suggested in the Vidal data set (Vidal *et al.* 2000), the second data set we have analysed has been actively "cleaned" for signs of recombination (Korber *et al.* 2000). Table 2 shows that this data set still appear to show signs of recombination both within and between subtypes. Recombination within subtypes is not surprising since the method used to remove recombinants only works for between-subtype recombinants, but the magnitude of the estimated recombination rate is surprising. A greater surprise is the evidence of recombination (apparently at a very high rate) between subtypes. This is important, because it indicates that recombination may have

Table 1

Partial envelope data set from Democratic Republic of Congo (Vidal *et al.* 2000)

Data analysed	# sequences	ϱ	P-value	R2	P-value
Subtype A	30	>500	0.2	-0.12	0.01*
Subtype C	20	>500	0.03*	-0.04	0.18
Subtype D	21	>500	0.06	-0.09	0.02*
Subtype F	16	>500	0.03*	0.02	0.79
Subtype H	16	>500	0.46	-0.04	0.05*
Random sample	30	>500	0.04*	-0.04	0.72
One of each subtype	08	>500	0.001*	-0.03	0.06

TABLE 2

Full length Envelope sequences analysed by Korber et al. 2000.

Data analysed	# sequences	ϱ	P-value	R2	P-value
Subtype A	15	362	0.23	0.01	0.72
Subtype B	59*	>500	0.07	-0.02	0.14
Subtype C	28	>500	0.9	0.01	0.72
Subtype D	13	>500	0.005*	-0.02	0.09
Subtype E	16	>500	0.02*	0.01	0.65
Random sample	30	137	0.20	-0.02	0.09
One of each subtype	8	344	0.04*	-0.02	0.08

been prevalent also before the diversification into subtypes, i.e. early on in the evolution of the present diversity. Note that none of the R2 tests for this data set are significant. This may reflect that the cleaning-for-recombinants process has detected sequences, which are compatible with different trees in different regions (and thus contribute most to the correlation detected by R2). It may also reflect that recombination is so prevalent that the R2 test loses power when applied to the longer sequences of the Korber data set.

Simulation

The parameters of the simulations were chosen to encompass the unknown parameters of HIV, i.e., the HIV parameters are expected to lie somewhere within the parameter space investigated. We varied the mutation rate, the recombination rate and the exponential growth rate. The low and high mutation rates correspond to an average nucleotide diversity of 4% and 20%, respectively. Four different recombination rates were chosen, all smaller than the maximum-likelihood estimates of Tables 1 and 2.

Table 3 shows various statistics for the case with and without exponential growth, with standard deviations for the most important statistics (standard errors of the mean are approximately 10% of the standard deviations shown). For the cases without exponential growth, expected results calculated from the neutral coalescent are also shown. It can be seen that when $\varrho = 0$, phylogenetic analysis recover these values to a good approximation. However, when increasing recombination (while keeping the other parameters constant), all the different measures are biased to various degrees.

TABLE 3

Simulation results

Mutation rate	ϱ	α	TMRCA	*S*	*D*	*T*	T5	T10
No growth								
0.02	Expected		0.038	0.040	0.020	0.143	0.006	0.002
	0	18	0.037 +/- 0.019	0.048 +/- 0.020	0.021 +/- 0.010	0.156 +/- 0.053	0.011	0.003
	10	0.2	0.032 +/- 0.012	0.091 +/- 0.033	0.022 +/- 0.008	0.199 +/- 0.058	0.018	0.006
	50	0.08	0.034 +/- 0.007	0.168 +/- 0.047	0.027 +/- 0.006	0.294 +/- 0.065	0.025	0.014
	200	0.07	0.038+/- 0.006	0.313 +/- 0.063	0.032 +/- 0.005	0.436 +/- 0.070	0.031	0.022
0.1	Expected		0.190	0.200	0.100	0.720	0.030	0.010
	0	32	0.192+/- 0.087	0.230 +/- 0.099	0.107 +/- 0.043	0.777 +/- 0.238	0.050	0.013
	10	1.7	0.155+/- 0.046	0.388 +/- 0.140	0.107 +/- 0.033	0.929 +/- 0.255	0.082	0.029
	50	0.6	0.144 +/- 0.023	0.757 +/- 0.213	0.114 +/- 0.017	1.281 +/- 0.225	0.105	0.063
	200	0.44	0.135 +/- 0.017	1.152 +/- 0.204	0.117 +/- 0.013	1.600 +/- 0.206	0.113	0.083
Exp. growth								
0.02	0	35	0.026+/- 0.003	0.183 +/- 0.034	0.020 +/- 0.002	0.264 +/- 0.029	0.019	0.013
	10	2.1	0.029 +/- 0.004	0.289 +/- 0.044	0.024 +/- 0.003	0.360 +/- 0.043	0.024	0.019
	50	0.6	0.032 +/- 0.004	0.391 +/- 0.051	0.026 +/- 0.003	0.445 +/- 0.052	0.027	0.023
	200	0.42	0.033 +/- 0.004	0.447 +/- 0.052	0.027 +/- 0.003	0.487 +/- 0.055	0.029	0.025
0.1	0	46.3	0.128 +/- 0.013	0.892 +/- 0.119	0.102 +/- 0.009	1.300 +/- 0.114	0.094	0.064
	10	10.1	0.130 +/- 0.012	1.374 +/- 0.201	0.114 +/- 0.009	1.731 +/- 0.174	0.115	0.093
	50	2.4	0.135 +/- 0.009	1.835 +/- 0.167	0.121 +/- 0.008	2.085 +/- 0.146	0.124	0.110
	200	1.6	0.137 +/- 0.007	2.081 +/- 0.129	0.125 +/- 0.006	2.260 +/- 0.112	0.127	0.117

The time to the most recent common ancestor is estimated to be shorter with increasing recombination when there is no exponential growth, whereas a smaller bias in the opposite direction is observed for the case of exponential growth (also illustrated in fig. 1). This suggests that if recombination occurs, then timing of the most recent common ancestor from phylogenetic analysis will give date that is too late (too close to the present) if HIV has a stable population size and a too early date if HIV has an exponentially growing population size.

The length of the terminal branches *S* is strongly biased upwards (irrespective of exponential growing or stable population size). This is because recombination makes sequences more equidistant from each other, rendering the inferred phylogeny more "star-shaped" (Fig. 1). The average pairwise difference *D*, and the total length of the tree *T* are also generally biased upwards with recombination rate. This reflects that recombination leads to incompatibilities in the data set and

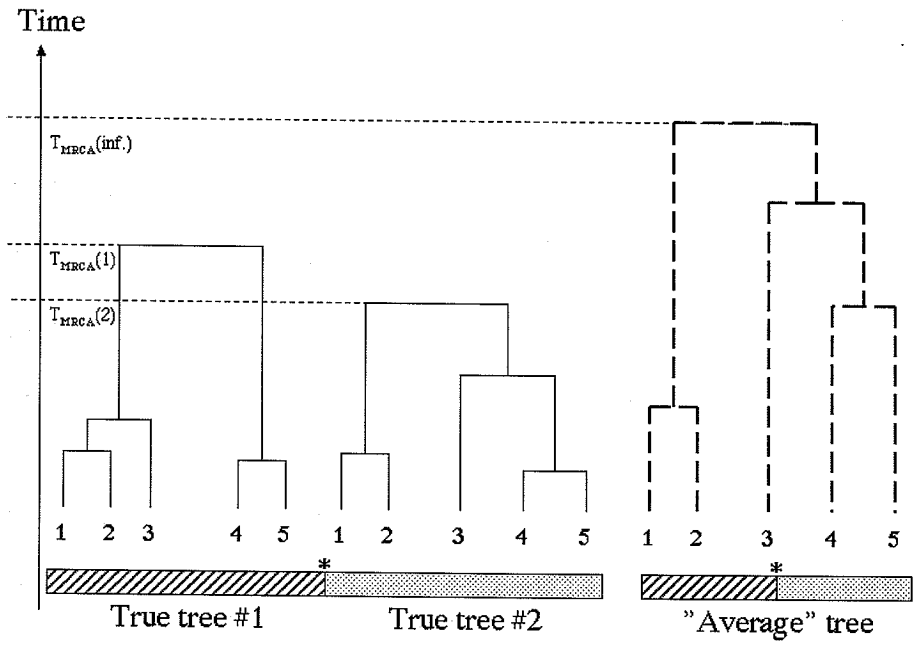

Fig. 1 – An illustration of the combined effect of recombination and exponential growth on phylogenetic analysis. Shown is the sequence of genetic material residing in five individuals. In the evolutionary history of the sequence, a recombination event has occurred at the breakage point (marked by an asterisk). The true relationship between the two regions of the sequence is therefore through two different trees. Time as measured in units of evolutionary time is indicated to the left, and the time to the most recent common ancestor (TMRCA) is indicated for the trees. When recombination is ignored and an "average" phylogenetic tree is inferred using the full sequence, a biased tree is obtained. As illustrated, the TMRCA on this "average" tree is estimated further back in time than the average of the true trees, and the timing of divergence events in the "average" tree are similarly biased upwards.

to accommodate these in a single tree, more mutation events need to be postulated.

T5 and T10 are the times (from the present) until there are 5 and 10 sequences left in the tree, respectively. Since ignoring recombination leads to long terminal branches (increased *S*), one would expect these times to increase with recombination rate. Table 3 shows that this effect is substantial. T5 was chosen as a measure to mimic diversification of, say, five subtypes. The results of Table 3 in this case imply that the timing of such diversification events from a phylogenetic analysis would give a much too early (too far back) date of diversification. The analogy in HIV-1 would be that the diversification events of the major subtypes have been dated much too early if recombination occurs.

Discussion

Are phylogenetic trees useful for virus data?

Use of the traditional phylogenetic method has seen a recent increase in studies of intra-specific population sequence data, particularly in the fields of virology and microbiology. One reason for the popularity of phylogenetic methods in this context is that the incorporation of phylogeny allows models of evolution to include the correlation induced by common descent, and thereby allows the use of the full information contained in the data (Huelsenbeck-Rannala 1997; Whelan *et al.* 2001). However, the phylogenetic method can only be applied to population genetic data under the assumption that the genetic region under study does not undergo any form of recombination or gene conversion. Violation of this assumption means that a single tree cannot describe the evolutionary history of the region, and thus, application of a phylogenetic method is inherently inappropriate. Additionally, the use of phylogenetic methodology on recombining data may easily lead one to ask questions which are not well-posed, e.g. "what is the time of the common ancestor?", when in fact there are several common ancestors at different times, or "which sequence type (subtype) does a sequence belong to?", when in fact various parts of the sequence belong to different sequence types.

If recombination has indeed occurred in the data, one should abstain from using phylogenetic methods and instead apply methods, which are insensitive to recombination or alternatively, use methods that model the recombination process explicitly. The latter option is preferable, since it potentially uses all information in the data. However, even though promising progress has been reported (Fearnhead-Donnelly 2001) such methods are not yet sufficiently sophisticated and fast to be of practical use in the study of virus evolution.

Recombination in HIV?

The considerations in the last paragraph have little relevance if recombination very rarely or never occurs in viruses in general and HIV-1 specifically. However, different lines of evidence suggest that recombination is relatively frequent in HIV-1.

HIV-1 and other retroviruses have the capacity to undergo recombination during the process of reverse transcription. This is a consequence of the dimeric nature of the genome, which allows the proceeding polymerase an opportunity to switch templates during the copying process. Thus, recombinant proviruses are generated that contain a mix of the

genetic information in the two parental strains (Mikkelsen-Pedersen 2000). Experiments in vitro have shown that this happens readily at very high rates throughout the HIV-1 genome (Jetzt *et al.* 2000). Recombination can therefore be expected to happen continuously in intra-patient populations. The amount of recombination between divergent sequences depends on the population frequency of double infections. It has previously been assumed that this frequency was low. However, recombination between subtypes does occur in the HIV population (Robertson *et al.* 1995), and quite a number of inter-subtype recombinant types (some of them circulating) have been classified. It has been shown that the frequency of recombination events increases with the degree of homology between the parental RNA strains (Mikkelsen-Pedersen 2000). Thus, existence of the many clear inter-subtype recombinants would suggest an even more common occurrence of recombination between closely related strains such as those found within the subtypes of the M group.

Another line of evidence for recombination comes from tests of the molecular clock hypothesis in HIV data. HIV evolution of the envelope and gag genes has been shown to follow a molecular clock in a phylogenetic study of a sequence data set obtained from the individuals of a limited and known transmission chain, where recombination between the strains residing in different individuals could be ruled out (Leitner-Albert 1999). However, in the population-wide phylogenetic study by Korber et al. of the same genetic regions, the presence of a strict molecular clock was rejected. Schierup and Hein (2000b) showed that ignoring recombination when it occurs leads to false rejection of the molecular clock hypothesis in phylogeny-based tests. Hence, the failure to demonstrate a molecular clock in the population data set could easily be a consequence of recombination in the evolutionary history of the sample.

Finally, application of two statistical tests in this study suggests high rates of recombination in two different data sets. The estimates of the recombination rates have very wide confidence limits and the extremely high maximum-likelihood estimates should therefore be treated with great caution. Nevertheless, the evidence taken as a whole suggests significant amounts of recombination, even in a data set where any possible inter-subtype recombinants have been removed. Even though the results are not entirely conclusive, we believe that phylogenetic analysis of population wide data sets from the M group have a great risk of being misleading and therefore we conducted a simulation study to address this question.

Consequences of recombination for phylogenetic timing in HIV-1

The simulations were conducted to mimic as many aspects of HIV-1 evolution as was found possible, including population growth. Yet, a number of simplifications are not necessarily justified. For instance, population structure may have played a large role in establishing the major group M subtypes, and different functional constraints and selection over the sequence is likely in HIV-1 but was not modelled. However, there is no a priori reason to expect that inclusion of further factors such as these in the model would decrease or cancel biases in timings caused by recombination.

The most difficult question is the potential bias in estimating the time to the most recent common ancestor. However, since with recombination there is no single MRCA the real question is how the timing estimate obtained from the single tree compares to the average over the sequence of the different TMRCAs (see Fig. 1). The direction of the bias in TMRCA was found to depend on the chosen parameters, but the bias can be quite substantial. If exponential growth is allowed, the estimated TMRCA is further back in time than the average TMRCA over the tree. This result is the opposite of the general conclusion of Schierup and Hein (2000), which considered only stable population size. A model including population growth would seem to be the most realistic for the M group of HIV-1 based on knowledge of the epidemic. However, the true direction and size of the bias in HIV-1 can only be evaluated when the recombination rate, population growth rate and other important factors, such as population sub-division, are known with greater accuracy. Until then, the only solid conclusion we can make is that previous studies using phylogenetic dating of TMRCA are likely to have been too confident in their estimates (and their standard errors in particular). A more general conclusion can be made regarding the implication of the simulations for dating of the subtype diversification in HIV-1 group M. Since "the average tree" inferred through phylogenetic analysis always has very long terminal branches (compared to each of the true trees over the sequence), most diversification events will appear to have happened further back in time than they actually have in the individual and "true" trees. This implies that most of the present-day HIV-1 subtypes may be much younger than the phylogenetic dating of Korber *et al.* (2000) suggests, and hence, it becomes unnecessary to invoke a "sunburst" diversification of all the subtypes early after the origin of HIV-1 group M diversity (Burr *et al.* 2001). The simulation results (T5 in particular) are very consistent in predict-

ing the direction of this bias, but accurate quantification would also require more knowledge about the values of the various evolutionary parameters in HIV-1.

TMRCAs and species transmission

One reason for discussing the TMRCA is that it may yield insight about when HIV-1, group M, was transmitted from the presumed chimpanzee host to humans. If only a single viral particle was transmitted, then the MRCA of all present day M group sequences must have existed in humans. If this scenario is correct it is therefore valid to attempt a timing of the MRCA and such an estimate could provide a lower bound on the time until the species transmission occurred. However, if recombination has occurred in the viral population originating from the MRCA, it is not valid to use a phylogenetic method to obtain the time estimate, and our results suggest that doing so would give a certain overconfidence in the previous estimate of 1931+/-10 years.

If however, more than one viral particle has been transmitted, the situation is much more complex. In this case, some parts of the sequence may have a MRCA in humans, whereas other parts may have a MRCA in Chimpanzees further back in time. The TMRCA of an "average tree" in this situation would be virtually uninformative about the time of the species transmission event. Another way of illustrating this point is the following example: Imagine two sequences (say 10% different) were transmitted from Chimpanzee to humans in year x (or two different years close in time). If these sequences meet and recombine in the new human host, then a subsequent phylogenetic analysis of the recombinant population would create a phylogeny with many branches appearing to date back before year x. Based on this phylogeny it would then wrongly be inferred that many more species transmission events than the actual two occurred in year x.

ACKNOWLEDGEMENTS – We thank Jotun Hein and Lars Aagaard for discussions, Gil McVean for comments on the manuscript and Anders M. Mikkelsen for programming assistance. The study was supported by grant no. 9901522 from the Danish Agricultural and Veterinary Research Council (to R.F.) and by grant no. 00001262 from the Danish Natural Sciences Research Council (to M.H.S.).

References

Awadalla P., Eyre-Walker A., Smith J.M., 1999. *Linkage disequilibrium and recombination in hominid mitochondrial DNA.* Science, 286: 2524-2525.

Burr T., Hyman J.M., Myers G., 2001. *The origin of acquired immune deficiency syndrome: Darwinian or Lamarckian?.* Philos. Trans. R. Soc. Lond., B Biol Sci., 356: 877-887.

Fearnhead P., Donnelly P., 2001. *Estimating recombination rates from population genetic data.* Genetics, 59: 1299-1318.

Felsenstein J., 1995 *PHYLIP (Phylogeny inference package) version 3.572.* Distributed over the World Wide Web, Seattle.

Gao F., Bailes E., Robertson D.L., Chen Y., Rodenburg C.M., Michael S.F., Cummins L.B., Arthur L.O., Peeters M., Shaw G.M., Sharp P.M., Hahn B.H., 1999. *Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes.* Nature, 397: 436-441.

Hudson R.R., 1983. *Properties of a neutral allele model with intragenic recombination.* Theor. Popul. Biol., 23: 183-201.

Hudson R.R., 2001. *Two locus sampling distributions and their application.* Genetics, 159.

Huelsenbeck J.P., Rannala B., 1997. *Phylogenetic methods come of age: testing hypotheses in an evolutionary context.* Science, 276: 218-219.

Jetzt A.E., Yu H., Klarmann G.J., Ron Y., Preston B.D., Dougherty J.P., 2000. *High rate of recombination throughout the human immunodeficiency virus type 1 genome.* Journal of Virology, 74: 1234-1240.

Korber B., Muldoon M., Theiler J., Gao F., Gupta R., Lapedes A., Hahn B.H., Wolinsky S., Bhattacharya T., 2000. *Timing the ancestor of the HIV-1 pandemic strains.* Science, 288: 1789-1796.

Leitner T., Albert J., 1999. *The molecular clock of HIV-1 unveiled through analysis of a known transmission history.* Proc. Natl. Acad. Sci. USA, 96: 10752-10757.

McVean G.A.T., Awadalla P., Fearnhead P., 2001. *A coalescent-based method for detecting and estimating recombination from gene sequences.* Genetics, in press.

Mikkelsen J.G., Pedersen F.S., 2000 *Genetic reassortment and patch repair by recombination in retroviruses.* Journal of Biomedical Science, 7: 77-99.

Robertson D.L., Sharp P.M., McCutchan F.E., Hahn B.H., 1995. *Recombination in HIV-1.* Nature, 374: 124-126.

Robertson D.L., Anderson J.P., Bradac J.A., Carr J.K., Foley B., Funkhouser R.K., Gao F., Hahn B.H., Kalish M.L., Kuiken C., Learn G.H., Leitner T., McCutchan F., Osmanov S., Peeters M., Pieniazek D., Salminen M., Sharp P.M., Wolinsky S., Korber B., 2000. *HIV-1 nomenclature proposal.* Science, 288: 55-56.

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

Schierup M.H., Hein J., 2000b. *Recombination and the molecular clock.* Molecular Biology and Evolution, 17: 1578-1579.

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

Thompson J.D., Gibson T.J., Plewniak F., Jeanmougin F., Higgins D.G., 1997. *The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.* Nucleic. Acids. Res., 25: 4876-4882.

VIDAL N., PEETERS M., MULANGA-KABEYA C., NZILAMBI N., ROBERTSON D., ILUNGA W., SEMA H., TSHIMANGA K., BONGO B., DELAPORTE E., 2000. *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.

WEISS R.A., 2001. *Natural and iatrogenic factors in human immunodeficiency virus transmission.* Philos. Trans. R. Soc. Lond., B Biol Sci., 356: 947-953.

WHELAN S., LIO P., GOLDMAN N., 2001. *Molecular phylogenetics: state-of-the-art methods for looking into the past.* Trends Genet., 17: 262-272.

WIUF C., CHRISTENSEN T., HEIN J., 2001. *A simulation study of the reliability of recombination detection methods.* Mol. Biol. Evol., 18: 1929-1939.

WOROBEY M., 2001. *A novel approach to detecting and measuring recombination: New insights into evolution in viruses, bacteria, and mitochondria.* Molecular Biology and Evolution, 18: 1425-1434.

YANG Z., 1999. *Phylogenetic Analysis by Maximum Likelihood (PAML), Version 2.0g.* University College, London.