Advances in understanding the evolution of HIV

Keith A. Crandall[a], Daniel Vasco[a], David Posada[a] and Hiromi Imamichi[b]

AIDS 1999, **13** (suppl A)S00–S00

Keywords: HIV, drug resistance, evolution, phylogeny, likelihood, model, candidate gene

Introduction

Phylogenetics is becoming increasingly more important in the analysis of HIV sequence data. Such analyses have been used to classify the diversity of HIV sequences into major groups and subtypes within these groups [1]. Phylogenetics has been used to track epidemiological change over time [2] and to detect epidemiological linkage among patients [3]. Indeed, phylogenetic evidence was recently used in a legal setting to convict a felon on the basis of epidemiological linkage supported by the phylogenetic relationships among sequences [4]. In recent years, there have been significant advances in phylogenetics, both in phylogeny reconstruction methodology and the estimation of population genetic parameters (coalescent theory) [5], as well as the use of phylogenetics in HIV studies. Here, we highlight four main areas of research that, over the past year, have used phylogenetic analyses to gain insights into the evolutionary biology and population dynamics of HIV: the discovery of a new group, the evolution of drug resistance, the dynamics of HIV within hosts and, finally, the impact of the host genotype on the evolution of HIV. Within each of these areas, we will discuss the advent of new phylogenetic tools used to test hypotheses concerning the evolution of HIV more rigorously.

New group of HIV

In 1995, a highly divergent strain of HIV-1 from Cameroon was isolated by Simon *et al*. [6]. This strain, YBF30, did not fall neatly into the two well-established groups of HIV-1 diversity: the major group (M; subtypes A–J) and the outliers group (O) (Fig. 1a). Instead, YBF30 clustered with a chimpanzee sequence (for the *env* sequences), with a 70% bootstrap support in their tree [6]. We have re-analyzed these sequences to determine the robustness of this conclusion.

First, we must clearly state the hypotheses being tested. In this case, we are interested in whether the YBF30 strain is a member of the M group or the O group or neither. Therefore, we must test these hypotheses within a phylogenetic hypothesis-testing framework. Second, since our hypotheses depend on the phylogenetic structure within the sequence data, we should take care to obtain as reasonable a representation of the phylogeny as we can.

Phylogeny reconstruction

When estimating phylogenetic relationships among sequences, one assumes a model of evolution. In many papers, researchers have not justified the model of evolution, and the model used [typically the Kimura 2-parameter model (K2P)] may not adequately reflect the underlying complexity of the data [7]. As models are developed with increasing complexity [8–11], it is to our advantage to take into account as much complexity as is needed in our estimation procedure. Determining how much complexity should be taken into account is a matter of statistical hypothesis testing [12–14]. In estimating the phylogenetic relationships among *env* sequences for placement of the YBF30 strain, Simon *et al*. [6] used 'the Kimura model', which we assume was the K2P model [15] (Kimura also developed a three-parameter model [16]). This

From the [a]Department of Zoology, Brigham Young University, Provo, Utah, USA and [b]SAIC Frederick, Frederick Cancer Research & Development Center, National Cancer Institute, Frederick, Maryland, USA.

Sponsorship: This work was supported by a James A. Shannon Award, an Alfred P. Sloan Award, and NIH grant #R01 HD34350-01A1.

Correspondence to Dr Keith A. Crandall, 574 Widtsoe Building, Department of Zoology, Brigham Young University, Provo, UT 84602-5255, USA. Tel: +1 801 378 3495; fax: +1 801 378 7423; e-mail: kac@email.byu.edu

© Lippincott Williams & Wilkins

Table 1. Likelihood ratio tests of models of molecular evolution [14,17].

Null hypothesis	Models compared[a]	-ln L_0	-ln L_1	-2ln l	d.f.	P[b]
Equal base frequencies	H_0: JC69 and H_1: F81	41473.5	41275.1	396.9	3	< 0.000001
Equal ti/tv[c] rates	H_0: F81 and H_1: HKY85	41275.1	40725.7	1098	1	< 0.000001
Equal ti and equal tv rates	H_0: HKY85 and H_1: GTR	40725.7	40634.8	181.7	3	< 0.000001
Equal rates among sites	H_0: GTR and H_1: GTR + Γ	40634.8	39197.2	2875	1	< 0.000001
Proportion of invariable sites	H_0: GTR + Γ and H_1:GTR + Γ + invar	39197.2	39173.0	48.5	1	< 0.000001 of

[a]JC69, Jukes–Cantor; F81, Felsenstein 81; HKY85, Hasegawa–Kishino–Yano 85; GTR, general time reversible. All models are detailed in [14,17]. [b]Due to the performance of multiple tests, the significance level of rejection of the null hypothesis should be adjusted via the Bonferroni correction to α = 0.01. [c]ti, transitions; tv, transversions. **[AU: Define L_1 and L_0 in header]**

model takes into account differences in transition/transversion bias. Does this model adequately explain the data?

We tested this hypothesis by obtaining sequences from the Los Alamos Database (http://hiv-web.lanl.gov/) and optimizing different models of evolution [14]. We collected 24 *env* sequences representing the major subtypes in the M group, the O group, SIV sequences, and HIV-2 sequences. Phylogenetic relationships among sequences were then estimated with the optimized model [17] and compared with the K2P model using the neighbor-joining method [18] as implemented in PAUP* [19]. Confidence in the resulting relationships was assessed using 1000 replications of the bootstrap procedure [20].

Using likelihood ratio tests, we rejected the null hypotheses of these sequences having: (i) equal base frequencies (A = 0.3447, C = 0.1810, G = 0.2346, and T = 0.2397); (ii) equal transition and transversion rates; (iii) equal transition rates and equal transversion rates; and (iv) rate homogeneity (Table 1). Therefore, our optimized model of evolution is the general-time reversible model (GTR) with a gamma correction for rate heterogeneity. Clearly, the standard K2P model does not account for the complexity of these data. Using the GTR model of evolution, we estimated a neighbor-joining tree for the *env* sequences (Fig. 1a). Does this different model make a difference in the resulting tree topology? In this case, the model of evolution does make a difference in the tree topology, but not in the placement of the YBF30 strain (Fig. 1a, b). The difference in topology here is in the deep nodes and in the positioning of the F-subtype clade. In the GTR tree, the HIV-2 and SIV do not cluster together as they do in the K2P tree. The mixing of SIV and HIV-2 has been proposed in the past [21,22]. Muse [23] provides a thorough review of models of evolution in HIV sequence analyses.

We would also like to examine two other issues related to phylogeny estimation; both dealing with the treatment of gaps. The first is that the Los Alamos HIV Database provides a set of pre-aligned sequences. If these sequences are re-aligned (to add or delete certain sequences) and the gaps are not taken out prior to re-alignment, the results can have a significant effect on the resulting tree topology. We point this out because the default on the popular sequence alignment program Clustal X [24] does not delete ('reset') these gaps. Researchers should thus be cautious when performing this sort of analysis. In this case, it does change the positioning of the YBF30 sequence when comparing the reset gap trees [using the K2P model and the GTR model, giving the association with the CPZGAB sequence (Fig. 1a, b)] to the non-reset gap tree [using the K2P model, giving an association with the M group (Fig. 1c)]. The second issue deals with 'gapstripping', a common practice in HIV research. Gapstripping deletes all data associated with an insertion or deletion event and denies the biological reality of 'indels'. While these molecular evolutionary events are difficult to account for in models of evolution, some progress is being made in this area [25,26]. Once again, it is important to ask whether or not this common analytical procedure affects inferred tree topology. It does (Fig. 1d). If the positional homology of characters is in question, then it is appropriate to remove those positions from consideration in a phylogenetic analysis. However, simply to remove positions that contain gaps without a preliminary assessment of homology is throwing out potential phylogenetic information. Hillis [27] provides a detailed discussion on homology of molecular characters.

Hypothesis testing

Obtaining a phylogeny should not be the end of a statistical analysis in tests of phylogenetic hypotheses. With the YBF30 strain, the main hypothesis of interest is whether this strain falls within the O group, M group, or neither group. The best estimate of phylogenetic relationships (Fig. 1a) suggests that the strain does not cluster with either the M or O groups, but with CPZGAB. This result is supported by high bootstrap values (73% in our search and 77% by Simon

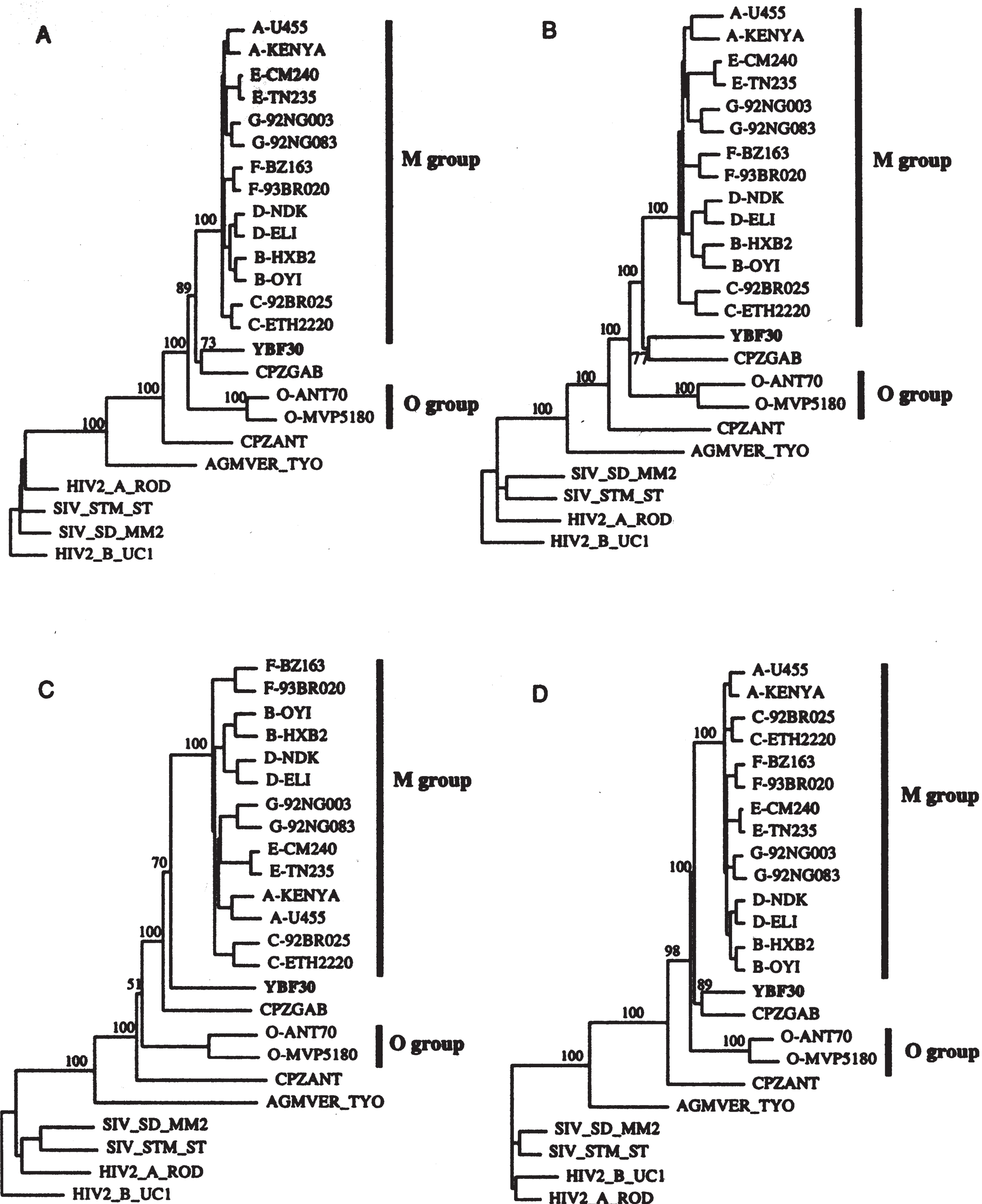


Fig. 1. (a) Neighbor-joining phylogeny based on the GTR model of evolution with invariable sites and rate heterogeneity incorporated. The estimated rate parameters used were $R(a) = 1.4366$, $R(b) = 2.8995$, $R(c) = 0.8917$, $R(d) = 1.1006$, $R(e) = 3.2988$, a proportion of invariable sites estimated to be 0.0962, and a gamma shape parameter of $\alpha = 1.2850$. (b) Neighbor-joining phylogeny based on the Kimura two-parameter model (K2P). (c) Neighbor-joining phylogeny using the K2P model as (b), but without removing gaps before the alignment procedure. (d) Neighbor-joining phylogeny using the GTR model as (b), but with gapstripping enforced. Bootstrap values are shown only for those clades associated with the placement of the YBF30 strain and are based on 1000 bootstrap replications.

Table 2. Statistical tests of the hypotheses that the YBF30 strain is a member of the M group or O group using the Wilcoxon signed-rank test for parsimony scores and the Kashino–Hasegawa (KH) test with likelihood scores.

Trees	Length	Wilcoxon signed-rank test *P* value	-ln likelihood	KH test *P* value
Estimated tree with no constraint	9141		39231.71100	
Estimated tree with constraint M group, YBF30	9130	0.5143	39217.88458	0.3612
Estimated tree with constraint O group, YBF30	9212	<0.0001***	39267.51935	0.0008***

[AU: Indicate significance level of *]**

et al. [6]). However, this is not an adequate test of the hypotheses in question. Can we reject the null hypotheses of YBF30 clustering with the M group or the O group? To test these hypotheses, we first need two constraint trees forcing YBF30 to be a member of the M group and the O group, respectively. Then we compare these trees with our best estimate of phylogenetic relationships using a statistical test. A number of tests have been proposed based on parsimony [28], likelihood [29] and parametric bootstrapping [30] to statistically test whether two alternative trees are significantly different. We tested these two hypotheses using the Kishino–Hasegawa test [29] and the Wilcoxon signed-rank test [31]. With both of these tests, we failed to reject the null hypothesis of YBF30 as a member of the M group, but strongly rejected the hypothesis of association with the O group (Table 2). Thus, while our best estimate supports the idea that YBF30 is distinct from groups M and O, we can only reject the O group membership hypothesis. The conclusion drawn by Simon *et al*. [6] that YBF30 is a new group is not supported by our statistical analysis of the *env* sequences. Clearly, YBF30 is evolutionarily very distinct, whether or not it is a member of the M group. These analyses are meant only as demonstrations of techniques and procedures in phylogenetic analyses. To make reasonable conclusions about the status of YBF30, entire genome analyses would be more appropriate, with heuristic searches and hypothesis testing. Clearly, results may also depend on the various sequences selected as representatives of the different taxa involved and the extent to which recombination has been a confounding factor in the history of these sequences [32]. Thus, the application of appropriate phylogenetic techniques is central to our understanding of the global diversity of HIV.

Drug resistance

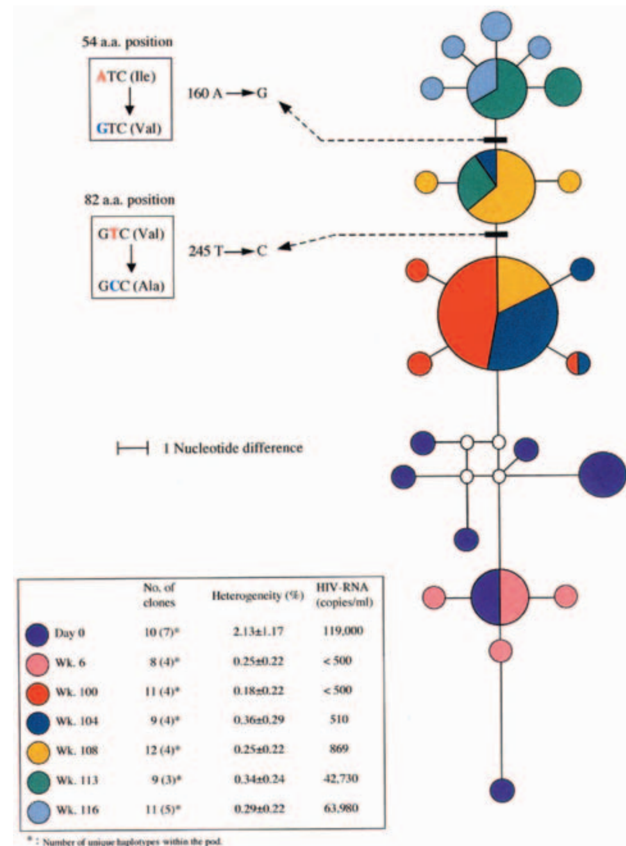
The advent of highly active antiretroviral therapy (HAART) in the treatment of HIV-infected individuals has been successful in prolonging lives and increasing quality of life. However, the question remains whether HAART represents a cure [33]. There is clear evidence for reservoirs of HIV-1 in patients on HAART over extended periods of time, even when the therapy has been extended over years [34–36]. What is not clear is whether the virus is evolving as a result of these latent reservoirs. Finzi *et al*. [34] state explicitly that this latent reservoir is not evolving. Wong *et al*. [37] also argue that replication in their patients with 'undetectable' levels of plasma virus 'was suppressed below the extent necessary for discernible evolutionary changes'. Yet, in their phylogeny, there is clear separation between the early and late timepoints, indicating that evolutionary change had occurred. If the virus is evolving, there remains a potential to evolve drug resistance.

The first issue in this debate concerns detection of the virus. How close is our minimum detection level of 50 copies/ml to identifying an eradication point? Assuming that initial infection in a patient is of the order of 10^8 cells in productive cycle [38], our most sensitive assays would detect a viral load of only 10^5, due to the differentiation of virus trapped; for example, on follicular dendritic cells versus productive mononuclear cells [39]. Thus, one infected cell would correspond to 10^{-3} copies/ml, leaving the term 'undetectable' misleading at 50 copies/ml and a long way from eradication [40].

We recently examined the evolutionary change in the face of multi-drug therapy in eight patients, five of whom escaped drug therapy; the remaining three continued to show 'undetectable' levels of variation [41]. It was clear from this study with multiple clones sampled at the beginning of treatment and at least 54 weeks after the initiation of treatment that evolution had occurred. Indeed, the five patients who escaped drug therapy did so through parallel changes. Yet the standard method for detecting positive Darwinian selection by Nei and Gojobori [42] did not indicate an abundance of non-synonymous substitutions, suggesting that this is a weak test for detecting selection. Nielsen and Yang [43] came to a similar conclusion, and suggested the use of likelihood methods instead (see, for example, [8,44]).

The notion of continued evolution when viral loads are 'undetectable' has also been supported by other studies [36]. We can explore the dynamics of the ev-

olution of drug resistance by examining sequence evolution in a single host individual over time. We sequenced the protease gene from a patient, as described in Crandall *et al.* [41]. This patient is number two in both Crandall *et al.* [41] and Zhang *et al.* [45]. The sequence data used in this analysis have been deposited in GenBank under accession numbers AF101340–AF101360 and AF112486–AF112534. We then estimated genealogical relationships among sequences using the method described in Templeton *et al.* [46]. This method has been shown to have greater resolving power than traditional phylogenetic methods when sequence divergence is low [47–49]. The resulting genealogy indicates the evolutionary pathway of drug resistance (Fig. 2). Prior to combination antiretroviral therapy, there is a great deal of heterogeneity, with no dominant sequence type in the protease gene. An amino-acid replacement then occurs that confers some drug resistance, and there is virtually a wholesale replacement of the population, i.e. a selective sweep of this mutation into high frequency in the population. Three such events have occurred in the history of these sequences, leading to the three large circles representing multiple sequences. However, the first such event is not due to an amino-acid replacement in the protease gene, but to an amino-acid replacement in the Gag protein (Table 3). Thus, the evolutionary history of these sequences can demonstrate the effects of evolutionary events not even associated with the specific gene region in question. These Gag mutations have been observed in other patients who have also escaped drug therapy [45,50]. Viral fitness associated with escape of drug therapy has been shown to be lower than fitness before treatment [51]. Understanding the evolutionary mechanisms that allow HIV to evolve resistance to drug therapy will be a crucial area of study in the next few years.



	No. of clones	Heterogeneity (%)	HIV-RNA (copies/ml)
Day 0	10 (7)*	2.13±1.17	119,000
Wk. 6	8 (4)*	0.25±0.22	< 500
Wk. 100	11 (4)*	0.18±0.22	< 500
Wk. 104	9 (4)*	0.36±0.29	510
Wk. 108	12 (4)*	0.25±0.22	869
Wk. 113	9 (3)*	0.34±0.24	42,730
Wk. 116	11 (5)*	0.29±0.22	63,980

* : Number of unique haplotypes within the pod.

Fig. 2. Genealogical relationships among protease sequences from a single patient over a 116-week period, during which viral load decreased to undetectable levels and rebounded after the accumulation of drug-resistant mutations. The connections are supported with a probability of > 0.95 using the method of Templeton *et al.* [46]. Open circles represent missing intermediates; circles are drawn proportional to the number of sequences corresponding to that clone.

Dynamics within host

We have now touched on the dynamics of HIV associated with the evolution of drug resistance. However, the population dynamics of HIV and the relative influences of genetic drift and natural selection have been hotly debated over the past year. These debates were initiated by the publication of articles by Leigh Brown and Richman [52,53], which suggest that the effective population size of HIV is relatively small (10^3); they thus conclude that there is significant opportunity for genetic drift to play a role in the evolution of HIV. This view conflicts with that of Coffin [54], which argues that population sizes of HIV are large (10^{8-10}), and therefore the dynamics are steady state and are dominated by natural selection. These two arguments differ for a number of reasons, in part because of the distinction between census size and effective population size. Yet they also differ in their conclusions about the relative importance of natural selection versus genetic drift.

Many laboratories are now engaged in collecting data relevant to this issue. Furthermore, there has been a great effort in the development of theoretical frameworks for estimating effective population sizes and genetic diversity based on coalescent theory [55–59]. Genetic diversity is defined in population genetics as $\theta = 2N_{ei}\mu$, where N_{ei} is the inbreeding effective population size and μ is the mutation rate per nucleotide [60]. We have used these techniques to explore the dynamics of HIV evolution over time within individuals. We analyzed data from Holmes *et al.* [61] because the selection pressures of these sequences have been extensively explored [43,62]. This will allow us

Table 3. Amino-acid replacements associated with drug resistance accumulated through evolutionary change in the patient shown in Fig. 2.

Time on therapy (weeks)	HIV-RNA (copies/ml)	Total CD4 (cells/μl)	Amino-acid residue associated with resistance in the protease gene (relative to consensus B sequence)											p7/p1 at P2	p1/p6 at P1′
			L10	K20	L24	V32	M46	I54	L63	A71	V82	I84	L90	A	L
0	119000	136	I	–	–	–	–	–	C/S	T	–	–	–	–	–
6	< 500	538	I	–	–	–	–	–	S	T	–	–	–	–	–
100	< 500	729	I	–	–	–	–	–	S	T	–	–	–	V	–
104	510	1032	I	–	–	–	–	–	S	T	–/A	–	–	V	–
108	869	1036	I	–	–	–	–	–	S	T	A/–	–	–	V	–
113	42730	599	I	–	–	–	–	–/V	S	T	A	–	–	V	–
116	63980	609	I	–	–	–	–	V	S	T	A	–	–	V	–

to relate the selection pressures to the population dynamics of HIV.

Using four different phylogenetic estimators of genetic diversity and two non-phylogenetic estimators, we examined changes in diversity over time for these *env* sequences (Fig. 3). Clearly, there are dynamic changes in genetic diversity, an indication that the steady-state model is not adequate to describe the ongoing population dynamics. Furthermore, assuming a constant mutation rate of 3×10^{-5} substitutions per site per generation **[AU: Confirm 'units']** [63], the inbreeding effective population size ranges from 2850 to 780. This estimate is on a par with that of Leigh Brown's estimate. Thus, the inbreeding effective population size is, as expected, much smaller than the census size. This small inbreeding effective population size indicates that genetic drift can play an important role in the evolutionary history of HIV. But is this small effective population size due to a population bottleneck of HIV-1?

The ratio of non-synonymous to synonymous substitutions has played a central role in the debate between the relative importance of selection and drift [64,65]. Researchers have argued that a ratio greater than one indicates the influence of positive selection [66]. However, others have argued that a ratio greater than one is not as important as a shift from the background ratio [67]. Either way, one must also be cautious about the method used to estimate these ratios, because pairwise methods have been shown to be biased under certain circumstances [41,68]. When we examined the dynamics of the ratio of the non-synonymous to synonymous substitutions over time, we noticed that genetic diversity decreases following an increase in this ratio (Fig. 3). These data support the interpretation of Fig. 2, i.e. these populations constantly go through selective sweeps of advantageous mutations into high frequency in the population. When this occurs, genetic diversity decreases (Fig. 3). Thus, this decrease

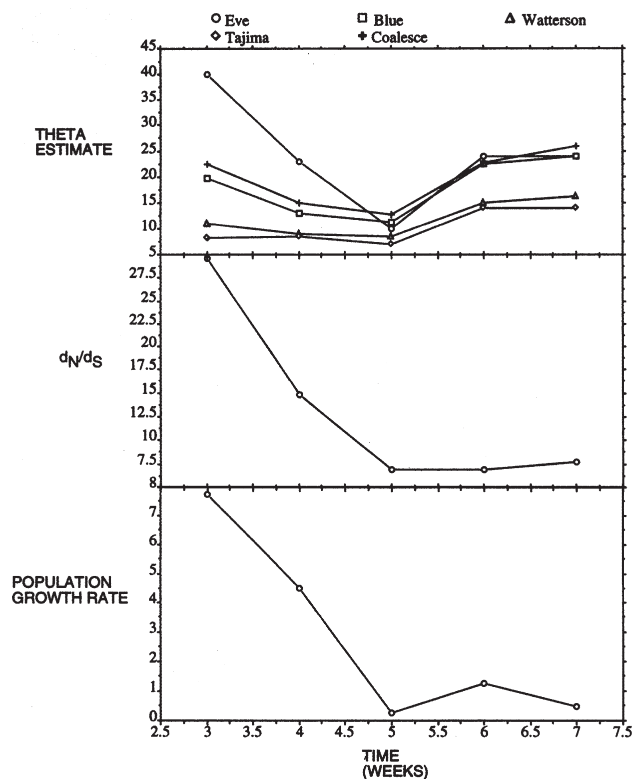


Fig. 3. Estimate of genetic diversity ($\hat{\theta}$) for the *env* sequences from Holmes *et al.* [61] using a number of different estimation procedures, including Eve (D.A. Vasco, Y.-X.Fu, manuscript submitted), BLUE [55], Coalesce [58], Tajima [85] and Watterson [86]. The ratios of non-synonymous to synonymous substitutions (d_N/d_S) were estimated using maximum likelihood [43]. Population growth rates based on the logistic growth model run backwards in time [69,70].

in genetic diversity is not necessarily due to bottlenecks at transmission, but to continual selective sweeps throughout the history of HIV infection in a patient. At the same time, we can estimate the population growth using a coalescent approach, such that $N_e(t)$

= $N_0 e^{-rt}$, where $N_e(t)$ is the effective population size at time t, N_0 is the initial effective population size, r is the growth rate (or decline rate when $r < 0$), and t is time since the initial generation [69,70]. Using this approach, we see that the population growth decreases during the most intense period of selection (Fig. 3) and then fluctuates slightly over the final three time-points. These data suggest that selection pressure is ongoing but varies in intensity over time, and this variation in intensity results in changes in genetic diversity and population growth. This idea of continued selective pressure has also been seen in other HIV-infected patients [71]. These results suggest an important role for the interaction of genetic drift and natural selection in shaping the diversity of HIV sequences. Much work remains to be performed in this area of research, especially in terms of exploring dynamics over different time scales and calibrating estimates of diversity accordingly.

Host–genotype interactions

We conclude with one of the most exciting areas of research in the evolutionary biology of HIV: the interaction of the virus with the host in a host–genotype-specific way. A number of candidate loci for host–virus interactions have been identified; for example, chemokine receptors (CCR5, CCR2, and CXCR4) and the major histocompatibility locus (MHC). The detection of the 32-bp deletion in the CCR5 associated with long-term non-progressors was the first indication of the significant role the host genotype can play in determining rates of disease progression [72,73]. Because this deletion failed to explain 80% of the long-term progressors (i.e. only 20% had this deletion), the CCR2 region was explored for the effects of its variation on disease progression [74]. An amino-acid replacement at position 64 (*CCR*2-64I) is associated with a 2–4-year decrease in the progression toward AIDS [74]. Recently, an amino-acid replacement in the stromal-derived factor (SDF-1) was shown to be associated with delaying onset of AIDS [75]. Even variation in the promoter region of CCR5 has been implicated in determining the rate of disease progression [76]. New data also suggest that CTLs exert significant selective pressure on HIV at many stages of infection [77]. There is evidence of strong selection acting on the CTL variants themselves [78]. No doubt this is just the tip of the iceberg in terms of host genotypes affecting disease progression. Data from HIV-2-infected patients suggest that a number of coreceptors can be used by a genetically diverse group of viral strains [79], whereas other data from HIV-1-infected patients suggest that the virus may utilize a single coreceptor (in this case CXCR4), even when other coreceptors are available [80].

Fortunately, an abundance of statistical tools from quantitative genetics are available to use in establishing correlations between genotype (host) and phenotype (e.g. disease progression) [81,82]. These tools have been used effectively in other human disease-related settings to determine the extent of genetic contributions to disease (see, for example, [83]). Preliminary work has been performed to explore the variation in chemokine coreceptor usage and how this variation is associated with different HIV-1 subtypes [84]. As genetic variations in these candidate loci are characterized for their association with disease progression and other disease-related phenotypes, an improved picture of the role of the host genotype as a determinant of disease progression in HIV infection will emerge.

Acknowledgements

We thank Greg Burton, Tom Folks, Eddie Holmes, Marsha Kalish, Matt Rain and Ray Viscidi for helpful discussions and comments on the manuscript.

References

1. McCutchan FE. **Global diversity in human immunodeficiency viruses.** In *The Evolution of HIV*. Edited by Crandall KA. Baltimore, MD: Johns Hopkins University Press; 1999:41–101.
2. Holmes EC, Zhang LQ, Robertson P, *et al.* **The molecular epidemiology of human immunodeficiency virus type 1 in Edinburough.** *J Infect Dis* 1995, **171**:45–53.
3. Ou C-Y, Ciesielski CA, Myers G, *et al.* **Molecular epidemiology of HIV transmission in a dental practice.** *Science* 1992, **256**:1165–1171.
4. Vogel G. **HIV strain analysis debuts in murder trial.** *Science* 1998, **282**:851–853.
5. Rodrigo AG, Felsenstein J. **Coalescent approaches to HIV population genetics.** In *Evolution of HIV*. Edited by Crandall KA. Baltimore, MD: Johns Hopkins University Press; 1999:233–272.
6. Simon F, Mauclere P, Roques P, *et al.* **Identification of a new human immunodeficiency virus type 1 distinct from group M and group O.** *Nat Med* 1998, **4**:1032–1037.
7. Goldman N. **Simple diagnostic statistical tests of models for DNA substitution.** *J Mol Evol* 1993, **37**:650–661.
8. Goldman N, Yang Z. **A codon-based model of nucleotide substitution for protein-coding DNA sequences.** *Mol Biol Evol* 1994, **11**:725–736.
9. Muse SV. **Evolutionary analyses of DNA sequences subject to constraints on secondary structure.** *Genetics* 1995, **139**:1429–1439.
10. Thorne JL, Goldman N, Jones DT. **Combining protein evolution and secondary structure.** *Mol Biol Evol* 1996, **13**:666–673.
11. Goldman N, Thorne JL, Jones DT. **Assessing the impact of secondary structure and solvent accessibility on protein evolution.** *Genetics* 1998, **149**:445–458.
12. Goldman N. **Statistical tests of models of DNA substitution.** *J Mol Evol* 1993, **36**:182–198.
13. Rzhetsky A, Nei M. **Tests of applicability of several substitution models for DNA sequence data.** *Mol Biol Evol* 1995, **12**:131–151.
14. Huelsenbeck JP, Crandall KA. **Phylogeny estimation and hypothesis testing using maximum likelihood.** *Annu Rev Ecol Syst*

1997, **28**:437–466.
15. Kimura M. **A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences.** *J Mol Evol* 1980, **16**:111–120.
16. Kimura M. **Estimation of evolutionary distances between homologous nucleotide sequences.** *Proc Natl Acad Sci USA* 1981, **78**:454–458.
17. Posada D, Crandall KA. **Modeltest: testing the model of DNA substitution.** *Bioinformatics* 1998, **14**:817–818.
18. Saitou N, Nei M. **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4**:406–425.
19. Swofford DL. *PAUP* Phylogenetic analysis using parsimony and other methods*. 4.0.0beta edition. Sunderland, MA: Sinauer Associates, 1998.
20. Felsenstein J. **Confidence limits on phylogenies: an approach using the bootstrap.** *Evolution* 1985, **39**:783–791.
21. Gao F, Yue L, White AT*, et al.* **Human infection by genetically diversise SIVsm-related HIV-2 in West Africa.** *Nature* 1992, **358**:495–499.
22. Sharp PM, Robertson DL, Gao F, Hahn BH. **Origins and diversity of human immunodeficiency viruses.** *AIDS* 1994, **8**:S27–S42.
23. Muse S. **Modeling the molecular evolution of HIV sequences.** In *The Evolution of HIV*. Edited by Crandall KA. Baltimore, MD: Johns Hopkins University Press; 1999:122–152.
24. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. **The clustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucl Acids Res* 1997, **24**:4876–4882.
25. Thorne JL, Kishino H, Felsenstein J. **Inching toward reality: an improved likelihood model of sequence evolution.** *J Mol Evol* 1992, **34**:3–16.
26. Thorne JL, Kishino H, Felsenstein J. **An evolutionary model for the maximum likelihood alignment of sequence evolution.** *J Mol Evol* 1991, **33**:114–124.
27. Hillis DM. **Homology in molecular biology.** In *Homology: The Hierarchical Basis of Comparative Biology*. Edited by Hall BK. New York: Academic Press; 1994:339–368.
28. Templeton AR. **Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes.** *Evolution* 1983, **37**:221–244.
29. Kishino H, Hasegawa M. **Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea.** *J Mol Evol* 1989, **29**:170–179.
30. Huelsenbeck JP, Hillis DM, Jones R. **Parametric bootstrapping in molecular phylogenetics: applications and performance.** In *Molecular Zoology: Advances, Strategies, and Protocols*. Edited by Ferraris JD, Palumbi SR. New York, NY: Wiley–Liss; 1996:19–42.
31. Templeton AR. **Convergent evolution and nonparametric inferences from restriction data and DNA sequences.** In *Statistical Analysis of DNA Sequence Data*. Edited by Weir BS. New York: Marcel Dekker; 1983:151–179.
32. Crandall KA, Templeton AR. **Statistical methods for detecting recombination.** In *The Evolution of HIV*. Edited by Crandall KA. Baltimore, MD: Johns Hopkins University Press; 1999:153–176.
33. Wain-Hobson S. **Down or out in blood and lymph?** *Nature* 1997, **387**:123–124.
34. Finzi D, Hermankova M, Pierson T*, et al.* **Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy.** *Science* 1997, **278**:1295–1300.
35. Chun T-W, Carruth L, Finzi D*, et al.* **Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection.** *Nature* 1997, **387**:183–188.
36. Chun T-W, Stuyver L, Mizell SB*, et al.* **Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy.** *Proc Natl Acad Sci USA* 1997, **94**:13193–13197.
37. Wong JK, Hezareh M, Gunthard HF*, et al.* **Recovery of replication-competent HIV despite prolonged suppression of plasma viremia.** *Science* 1997, **278**:1291–1295.
38. Ho DD, Neumann AU, Perelson AS, Chen W, Leonar JM, Markowitz M. **Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection.** *Nature* 1995, **373**:123–126.
39. Cavert W, Notermans DW, Staskus K*, et al.* **Kinetics of response in lymphoid tissues to antiretroviral therapy of HIV-1 infection.** *Science* 1997, **276**:960–964.
40. Coffin JM. **Molecular biology of HIV.** In *Evolution of HIV*. Edited by Crandall KA. Baltimore, MD: Johns Hopkins University Press; 1999:3–39.
41. Crandall KA, Kelsey CR, Imamichi H, Salzman NP. **Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection.** *Mol Biol Evol* 1999:(in press). **[AU: Any further publication details?]**
42. Nei M, Gojobori T. **Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions.** *Mol Biol Evol* 1986, **3**:418–426.
43. Nielsen R, Yang Z. **Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene.** *Genetics* 1998, **148**:929–936.
44. Muse SV, Gaut BS. **A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome.** *Mol Biol Evol* 1994, **11**:715–724.
45. Zhang Y-M, Imamichi H, Imamichi T*, et al.* **Drug resistance during indinavir therapy is caused by mutations in the protease gene and its gag substrate cleavage site.** *J Virol* 1997, **71**:6662–6670.
46. Templeton AR, Crandall KA, Sing CF. **A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation.** *Genetics* 1992,**132**:619–633.
47. Crandall KA. **Intraspecific cladogram estimation: accuracy at higher levels of divergence.** *Syst Biol* 1994, **43**:222–235.
48. Crandall KA. **Intraspecific phylogenetics: support for dental transmission of human immunodeficiency virus.** *J Virol* 1995, **69**:2351–2356.
49. Crandall KA. **Multiple interspecies transmissions of human and simian T-cell leukemia/lymphoma virus type I sequences.** *Mol Biol Evol* 1996, **13**:115–131.
50. Mammano F, Petit C, Clavel F. **Resistance-associated loss of viral fitness in human immunodeficiency virus type 1: phenotypic analysis of protease and *gag* coevolution in protease inhibitor-treated patients.** *J Virol* 1998, **72**:7632–7637.
51. Zennou V, Mammano F, Paulous S, Mathez D, Clavel F. **Loss of viral fitness associated with multiple Gag and Gag-Pol processing defects in human immunodeficiency virus type 1 variants selected for resistance to protease inhibitors in vivo.** *J Virol* 1998, **72**:3300–3306.
52. Leigh Brown AJ. **Analysis of HIV-1 *env* gene sequences reveals evidence for a low effective number in the viral population.** *Proc Natl Acad Sci USA* 1997, **94**:1862–1865.
53. Leigh Brown AJ, Richman DD. **HIV-1: gambling on the evolution of drug resistance?** *Nature Medicine* 1997, **3**:268–271.
54. Coffin J. **HIV population dynamics in vivo: implications for genetic variation, pathogenesis and therapy.** *Science* 1995, **267**:483–489.
55. Fu Y-X. **A phylogenetic estimator of effective population size or mutation rate.** *Genetics* 1994, **136**:685–692.
56. Fu Y-X. **Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences.** *Genetics* 1994, **138**:1375–1386.
57. Felsenstein J. **Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates.** *Genet Res Camb* 1992, **59**:139–147.
58. Kuhner MK, Yamato J, Felsenstein J. **Estimating effective population size and mutation rate from sequence data using Metropolis–Hastings sampling.** *Genetics* 1995, **140**:1421–1430.
59. Kuhner MK, Yamato J, Felsentein J. **Maximum likelihood estimation of population growth rates based on the coalescent.** *Genetics* 1998, **149**:429–434.
60. Li W-H. *Molecular Evolution.* Sunderland, MA: Sinauer Associates; 1997.
61. Holmes EC, Zhang LQ, Simmonds P, Ludlam CA, Leigh Brown AJ. **Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient.** *Proc Natl Acad Sci USA* 1992, **89**:4835–4839.
62. Bonhoeffer S, Holmes EC, Nowak MA. **Causes of HIV diversity.** *Nature* 1995, **376**:125.
63. Mansky LM, Temin HM. **Lower in vivo mutation rate of human**

immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J Virol* 1995, **69**:5087–5094.
64. Holmes EC, Zanotto PMdA. **Genetic drift of human immunodeficiency virus type 1?** *J Virol* 1998, **72**:886–887.
65. Plikat U, Nieselt-Struwe K, Meyerhans A. **Genetic drift can dominate short-term human immunodeficiency virus type 1 *nef* quasispecies evolution in vivo.** *J Virol* 1997, **71**:4233–4240.
66. Messier W, Stewart C-B. **Episodic adaptive evolution of primate lysozymes.** *Nature* 1997, **385**:151–154.
67. Sharp PM. **In search of molecular darwinism.** *Nature* 1997, **385**:111–112.
68. Yang Z, Nielsen R. **Synonymous and nonsynonymous rate variation in nuclear genes of mammals.** *J Mol Evol* 1998, **46**:409–418.
69. Griffiths RC, Tavaré S. **Sampling theory for neutral alleles in a varying environment.** *Philos Trans R Soc London, Ser B* 1994, **344**:403–410.
70. Slatkin M, Hudson RR. **Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations.** *Genetics* 1991, **129**:555–562.
71. Shankarappa R, Gupta P, Learn GH Jr, *et al.* **Evolution of human immunodeficiency virus type 1 envelope sequences in infected individuals with differing disease progression profiles.** *Virology* 1998, **241**:251–259.
72. Dean M, Carrington M, Winkler C, *et al.* **Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the *CKR5* structural gene.** *Science* 1996, **273**:1856–1862.
73. Samson M, Libert F, Doranz BJ, *et al.* **Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene.** *Nature* 1996, **382**:722–725.
74. Smith MW, Dean M, Carrington M, *et al.* **Contrasting genetic influence of CCR2 and CCR5 variants on HIV-1 infection and disease progression.** *Science* 1997, **277**:959–965.
75. Winkler C, Modi W, Smith MW, *et al.* **Genetic restriction of AIDS pathogenesis by an SDF-1 chemokine gene variant.** *Science* 1998, **279**:389–393.
76. Martin MP, Dean M, Smith MW, *et al.* **Genetic acceleration of AIDS progression by a promoter variant of *CCR5*.** *Science* 1998, **282**:1907–1911.
77. Ogg GS, Jin X, Bonhoeffer S, *et al.* **Quantitation of HIV-1-specific cytotoxic T lymphocytes and plasma load of viral RNA.** *Science* 1998, **279**:2103–2106.
78. da Silva J, Hughes AL. **Conservation of cytotoxic T lymphocyte (CTL) epitopes as a host strategy to constrain parasite adaptation: evidence from the *nef* gene of human immunodeficiency virus 1 (HIV-1).** *Mol Biol Evol* 1998, **15**:1259–1268.
79. Owen SW, Ellenberger D, Rayfield M, *et al.* **Genetically divergent strains of human immunodeficiency virus type 2 use multiple coreceptors for viral entry.** *J Virol* 1998, **72**:5425–5432.
80. Michael NL, Nelson JAE, Kewalramani VN, *et al.* **Exclusive and persistent use of the entry coreceptor CXCR4 by human immunodeficiency virus type 1 from a subject homozygous for *CCR5* D[GREEK]32.** *J Virol* 1998, **72**:6040–6047.
81. Crandall KA. **Identifying links between genotype and phenotype using marker loci and candidate genes.** In *The Impact of Plant Molecular Genetics*. Edited by Sobral BWS. Boston, MA: Birkhauser; 1996:137–157.
82. Templeton AR, Boerwinkle E, Sing CF. **A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*.** *Genetics* 1987, **117**:343–351.
83. Clark AG, Weiss KM, Nickerson DA, *et al.* **Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase.** *Am J Hum Genet* 1998, **63**:595–612.
84. Zhang L, He T, Huang Y, *et al.* **Chemokine coreceptor usage by diverse primary isolates of human immunodeficiency virus type 1.** *J Virol* 1998, **72**:9307–9312.
85. Tajima F. **Evolutionary relationship of DNA sequences in finite populations.** *Genetics* 1983, **105**:437–460.
86. Watterson GA. **On the number of segregating sites in genetical models without recombination.** *Theor Popn Biol* 1975, **7**:256–276.

[AU: The electronic figure files that you provided could not be opened by our software. Thus the figures have been scanned in. To gain optimal print quality, electronic figures are preferable. Could you resend your figures as either PC format eps files, tif files or CorelDraw files, or inserted into a word document]