DOI: 10.1534/genetics.103.023382

Pervasive Genomic Recombination of HIV-1 *in Vivo*

Daniel Shriner,*,[1] Allen G. Rodrigo,*,[2] David C. Nickle* and James I. Mullins*,†

**Department of Microbiology and †Departments of Medicine and Laboratory Medicine, University of Washington School of Medicine, Seattle, Washington 98195-8070*

Manuscript received October 20, 2003
Accepted for publication April 29, 2004

ABSTRACT

Recombinants of preexisting human immunodeficiency virus type 1 (HIV-1) strains are now circulating globally. To increase our understanding of the importance of these recombinants, we assessed recombination within an individual infected from a single source by studying the linkage patterns of the auxiliary genes of HIV-1 subtype B. Maximum-likelihood phylogenetic techniques revealed evidence for recombination from topological incongruence among adjacent genes. Coalescent methods were then used to estimate the *in vivo* recombination rate. The estimated mean rate of 1.38×10^{-4} recombination events/adjacent sites/generation is ~5.5-fold greater than the reported point mutation rate of 2.5×10^{-5}/site/generation. Recombination was found to be frequent enough to mask evidence for purifying selection by Tajima's *D* test. Thus, recombination is a major evolutionary force affecting genetic variation within an HIV-1-infected individual, of the same order of magnitude as point mutational change.

THE human immunodeficiency virus type 1 (HIV-1) is a member of the lentivirus genus of the retrovirus family. It contains genes that encode the core structural and enzymatic functions common to all retroviruses as well as six auxiliary genes within its ~10-kb single-stranded RNA genome (Figure 1A). Among the auxiliary genes, *tat* and *rev* encode proteins that have essential roles in regulating viral gene expression and *vif*, *vpu*, *vpr*, and *nef* encode proteins that enhance virulence (Cullen 1998). Compared to the three core functional genes (*gag*, *pol*, and *env*), little is known about how the auxiliary genes evolve, and even less is known about how these genes coevolve.

Lentiviruses are well known for accumulating vast levels of genetic diversity, and recombination is thought to be an important process that affects this diversity. Recombination requires coinfection or superinfection of the same target cell within a host. Subsequently, the divergent viral RNA genomes are copackaged into the same progeny virion, with genomic recombinants produced by template switching during reverse transcription in cells subsequently infected with the heterovirion (Hu and Temin 1990a,b; Stuhlmann and Berg 1992). The isolation of recombinants between HIV-1 group M subtypes (*e.g.*, Sabino *et al.* 1994; Leitner *et al.* 1995; Robertson *et al.* 1995a; Carr *et al.* 1996; Gao *et al.* 1996; Salminen *et al.* 1997), between HIV-2 strains (*e.g.*, Gao *et al.* 1992, 1994; Robertson *et al.* 1995b), between different strains of the same HIV-1 subtype (*e.g.*, Diaz *et al.* 1995; Zhu *et al.* 1995; Liu *et al.* 2002), and between viruses in one, singly infected individual (*e.g.*, Delassus *et al.* 1991; Howell *et al.* 1991; Groenink *et al.* 1992; Morris *et al.* 1999) provides evidence for widespread coinfection or superinfection at the cellular level. Despite these numerous case studies, there is only one *in vivo* estimate of the HIV-1 recombination rate, which was derived from a cross-sectional analysis of multiple individuals infected with viruses from the same subtype (McVean *et al.* 2002). Currently, there is no analogous estimate of the HIV-1 recombination rate from analyses of multiple sequences derived from the viral population within a single individual.

One study of sequences derived from the first coding exon of *tat* (*tat*1), *nef*/LTR, and the second coding exon of *rev* (*rev*2)/gp41 over a 4-year period of infection led to the conclusion that these three loci evolved independently (Meyerhans *et al.* 1989; Delassus *et al.* 1991; Martins *et al.* 1991). Recombination could explain these results; however, it should be noted that the different gene regions were amplified from independent PCR replicates; hence no linkage should have been expected (Meyerhans *et al.* 1989; Delassus *et al.* 1991; Martins *et al.* 1991). Zhang *et al.* (1997) coamplified *vif*, *vpr*, and *vpu* but proceeded to analyze the three genes separately and did not examine coevolution. Similarly, Michael *et al.* (1995) coamplified *vif*, *vpr*, *vpu*, *tat*1, and *rev*1 from a single individual but performed no phylogenetic analyses on the separate genes.

To address coevolution, we coamplified all six auxil-

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. AY496645–AY496684.

[1]*Corresponding author:* Department of Microbiology, Rosen Bldg., University of Washington School of Medicine, Box 358070, Seattle, WA 98195-8070. E-mail: dshriner@u.washington.edu

[2]*Present address:* School of Biological Sciences, University of Auckland, Private Bag 92019, Auckland, New Zealand.

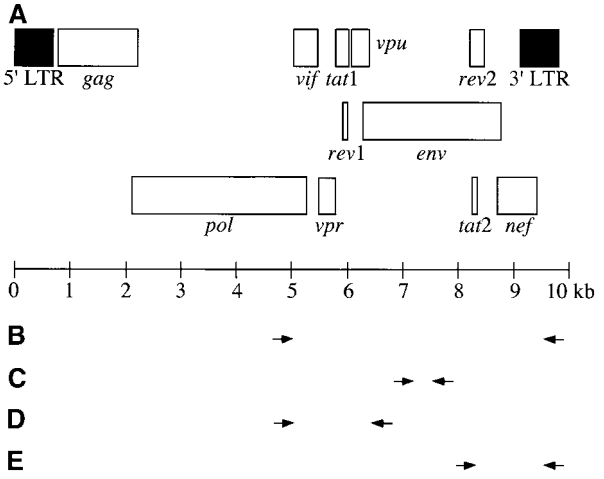

FIGURE 1.—(A) Schematic of the HIV-1 genome. The noncoding long terminal repeats (LTR) have the structure U3-R-U5, in which the U regions are unique sequences and the R region is a direct repeat sequence. (B) The 4.6-kb fragment amplified by first-round primers DS3 and DS8. (C) The second-round 600-bp fragment used for template quantification amplified by primers DR7 and DR8. (D) The central 1.3-kb fragment. (E) The 3′ 1.3-kb fragment.

iary genes after dilution of template molecules to obtain a single amplifiable template in each PCR replicate. Despite this step to justify the null assumption of linkage, high levels of recombination were confirmed by two independent computer simulation approaches. In the first approach, parametric bootstrapping in a maximum-likelihood framework was utilized to assess departures from the expected topological congruence across different genes in the presence of linkage. In the second approach, coalescent methods were used to estimate the recombination rate. Our work reveals that, during HIV-1 chronic infection, the mean recombination rate is ~5.5-fold greater than the mean mutation rate (MANSKY 1996). Hence, recombination has significant consequences for understanding the evolution of HIV-1.

MATERIALS AND METHODS

Subject data: The HIV-1-infected individual analyzed in this study was described previously (Pt. 6 in SHANKARAPPA *et al.* 1999). Plasma virus RNA derived from a clinical visit 2.96 years after seroconversion was used. As shown in Figure 2, genetic diversity at the *env* locus had peaked by 2.00 years after seroconversion in this individual (after which the slope was not different from zero, $P = 0.1833$), thus suggesting that population diversity was near equilibrium at the time of sampling.

Isolation, quantification, and amplification of HIV-1 virion RNA: A total of 200 μl of plasma was diluted with 800 μl of phosphate-buffered saline and ultracentrifuged at 125,000 × *g* at 4° for 1 hr. Plasma virus RNA was extracted using a standard guanidine method (AUSUBEL *et al.* 1990). RNA was then coprecipitated with yeast tRNA, redissolved in 0.3 M sodium acetate, extracted with phenol:chloroform:isoamyl alcohol, precipitated with isopropanol, washed with 75% ethanol, dissolved in diethyl pyrocarbonate-treated H_2O, and stored at −80°.

cDNA synthesis was performed using oligo(dT) and Superscript II RNase H-free reverse transcriptase (Life Technologies, Gaithersburg, MD) according to the recommendations of the manufacturers, with the addition of 5 mM spermidine, at 42° for 50 min. cDNA was serially diluted and subjected to nested PCR using the Expand high-fidelity PCR system as recommended by the manufacturers (Boehringer Mannheim, Indianapolis). First-round PCR amplified a 4.6-kb fragment encompassing the 3′ end of the viral genome (Figure 1B). These reactions included 1.5 mM $MgCl_2$ and primers DS3 (5′-GTT TCGGGTTTATTACAGGGACAGCAGAGA, 4895–4924, NL4-3) and DS8 (5′-GTTTGTCTAACCAGAGAGACCCAGTACAGG, 9550–9521, NL4-3). Cycling conditions were 94° for 15 sec, 55° for 45 sec, and 68° for 6 min for 10 cycles, followed by 20 cycles in which the extension step was incremented by 20 sec each cycle. The last cycle was followed by incubation at 72° for 30 min. For template quantification, these long, first-round products were subjected to second-round PCR using 1.25 mM $MgCl_2$ and primers DR7 and DR8 (Figure 1C; LIU *et al.* 1997). Cycle conditions were 3 cycles of 94° for 1 min, 55° for 1 min, and 72° for 1 min, followed by 32 cycles of 94° for 15 sec, 55° for 45 sec, and 72° for 1 min, with the final incubation at 72° extended to 5 min. cDNA copy number was calculated from the results of serial endpoint dilutions using the QUALITY application (RODRIGO *et al.* 1997).

To test primer sensitivity, nested PCR was performed using the molecular clone pNL4-3 as template. Template input ranged from 1 copy to 50 copies. In the first round, DS3 and DS4 (5′-GTTTCTTGTGGGTTGGGGTCTGTGGGTACA, 6471–6442, NL4-3) were used to amplify a 1.3-kb fragment encompassing the central region of the viral genome (Figure 1D); this reaction incorporated 1.5 mM $MgCl_2$ with the following cycle conditions: 94° for 15 sec, 57.7° for 45 sec, and 72° for 2 min and 15 sec for 25 cycles, followed by a final incubation at 72° for 10 min. In the second round, DS1 (5′-GTTTAAAGGTGAAGG GGCAGTAGTAATACA, 4955–4984, NL4-3) and DS2 (5′-GTT TCAGGTACCCCATAATAGACTGTGACC, 6354–6325, NL4-3) were used; this reaction incorporated 1.3 mM $MgCl_2$ with the following cycle conditions: 94° for 15 sec, 55.9° for 45 sec, and 72° for 2 min for 25 cycles, followed by a final incubation at 72° for 10 min.

For derivation of single template-derived amplicons for DNA sequencing, limiting dilution of cDNA was performed so that the average input into each individual PCR replicate was ~0.4 amplifiable templates. Thus, ~33% of replicates were expected to be positive (due to actually receiving template input), of which ~81% were expected to have received a single input cDNA copy (RODRIGO *et al.* 2000). Long first-round PCR conditions were as described above. Two separate second-round reactions were performed. In one, DS3 and DS4 were used to amplify a 1.3-kb fragment encompassing the central region of the viral genome as described above (Figure 1D). In the other reaction, DS7 (5′-GTTTAGAAAAGAATGAACAA GAATTATTGG, 8169–8198, NL4-3) and DS8 were used to amplify a 1.3-kb 3′ region (Figure 1E), with a final magnesium concentration of 1.7 mM and with the following cycle conditions: 94° for 15 sec, 56.8° for 45 sec, 72° for 2 min for 30 cycles, followed by a final incubation at 72° for 10 min. Second-round PCR products were directly sequenced on an Applied Biosystems (Foster City, CA) 377 automated sequencer.

Sequence analysis: Sequence editing and contig assembly were performed using SEQUENCHER, version 3 (Gene Codes, Ann Arbor, MI). As a further control against sequencing from multiple templates, any sequence that contained one or more sites at which the sequencing software detected signals of equal intensity from two different bases, thus indicating the presence of more than one template, was considered ambiguous and was excluded. Sequence alignments were generated using CLUSTAL W (THOMPSON *et al.* 1994) and were manually edited.

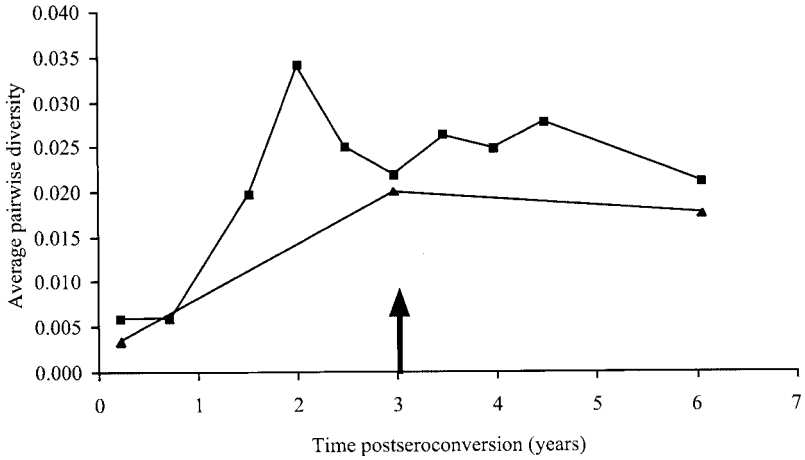

FIGURE 2.—Genetic diversity at the *env* locus. Squares represent average pairwise diversities from cell-associated viral DNA and triangles represent average pairwise diversities from cell-free viral RNA. The arrow represents the time of the sample used herein. Data are from SHANKARAPPA *et al.* (1999).

For analysis, sequences derived from the two second-round PCR products were concatenated to form "full sequences" of total length 2607 sites. The sequences were also segmented as follows: the *vif*, *vpr*, and *vpu* genes; the first exon of *tat*; the 3′ half of the *env* gene segment that encodes gp41; the *nef* gene; and the 5′ 88% of the U3 region (note that the first exon of *rev* corresponds to an alternate reading frame within the first exon of the *tat* gene, and that the second exons of both *tat* and *rev* are within the gp41 coding sequence). A coding subset of 2241 sites, from the *vif*, *vpu*, *tat*1, gp41, and *nef* segments, was further selected to exclude noncoding sites and to avoid double-counting sites in overlapping reading frames.

Phylogenetic analyses were performed using PAUP* (SWOFFORD 2002). This program reconstructs only branching phylogenies (the assumption is no recombination). Phylogenies were inferred by first estimating a neighbor-joining tree and then swapping branches under maximum likelihood using the subtree pruning and regrafting (SPR) algorithm (SWOFFORD 2002). For each sequence matrix, the SPR and the more rigorous tree bisection-reconnection (TBR) branch-swapping algorithms returned identical maximum-likelihood phylogenies (data not shown). Because the SPR algorithm is less time consuming than the TBR algorithm, we employed the former for all of the phylogenetic reconstruction in the computer simulation experiments reported. Models of sequence evolution were estimated under maximum likelihood using the general time-reversible model of substitution with unequal base frequencies and a gamma distribution of among-site rate variation (YANG 1994; SWOFFORD 2002).

Testing for topological incongruence: To test for statistical differences among topologies, the Shimodaira-Hasegawa test (SHIMODAIRA and HASEGAWA 1999) was performed as implemented in PAUP* (SWOFFORD 2002). This test is a multiple comparison version of the Kishino-Hasegawa test (KISHINO and HASEGAWA 1989). The set of possible topologies to be compared consisted of the seven segment topologies and the concatenated full-sequence topology. Each sequence matrix with its corresponding model of evolution was then tested against this set to see if other topologies were equally likely explanations of the data as compared to the null topology estimated for the given sequence matrix. A significant result indicates that the alternative topology is statistically worse than the null topology. Significance was determined by comparison to null distributions generated by 1000 bootstrap replicates with the full optimization option. In this test, branch lengths are optimized separately for each test topology.

Analysis of multiple hits: Multiple hits were analyzed to distinguish statistically between parallel evolution and recombination as explanations for recurrent mutation. The term "multiple hits" encompasses both real mutational events at a site and artifacts induced when reconstructing a single, branching phylogeny that ignores recombination. We made use of the following mutually exclusive categorization of multiple hits: (1) a parallelism, which is an identical change from one of the four nucleotide states to a second state on different branches of the phylogeny, (2) a reversal, which is a change from the second state back to the first state, (3) a third state change, which is a change from either the first or the second state to a third state (this category also includes changes to the fourth nucleotide state), (4) the case in which reversals occur in parallel, and (5) the case in which third state changes occur in parallel. Thus, all multiple hits can be parsed into exactly one of the five categories.

HUDSON and KAPLAN (1985) described the estimation of the number of recombination events that can be parsimoniously inferred in the history of a sample of DNA sequences. They assumed a neutral, infinite-sites model of mutation, according to which at most one mutation occurred at a given site (hence no third state changes occurred), and neither back mutations (yielding true reversals) nor recurrent mutations (yielding true parallelisms) occurred. In such a model, for any pair of sites at most four haplotypes can exist, and recombination provides the only way to explain the existence of all four haplotypes. If recombination had occurred but was ignored during phylogenetic reconstruction, then an excess of multiple hits is expected, which can be accounted for as apparent among-site rate variation (SCHIERUP and HEIN 2000).

MAYNARD SMITH and SMITH (1998) modified this method by relaxing the infinite-sites assumption. Thus, under a neutral Poisson mutational process, more than one mutation is expected to occur at some number of sites, meaning that there is an expectation for some number of multiple hits even in the case of complete linkage. Maynard Smith and Smith defined the "effective number of sites" in such a way as to incorporate the probability that a site was truly hit twice, accounting for the probabilities of observing real parallelisms, reversals, and third state changes. Thus, a limited amount of real rate variation was allowed.

This method was extended and modified (WOROBEY 2001) to include phylogenetic reconstruction and rate variation estimation in a maximum-likelihood framework. In this study, we further extended and modified this method. We expected that recombination should induce an excess of reversals not expected under parallel evolution. To test this expectation, we examined the distribution of multiple hits as reconstructed on the maximum-likelihood branching phylogeny for the full-length sequences. Multiple hits were counted in the reconstructed phylogeny and parsed into the five categories described

above. To count multiple hits, we used PAUP* to map all site changes onto the branches of the maximum-likelihood phylogeny using the reconstructed ancestral sequences at each node. All events at a site beyond one initial mutational event were counted as multiple hits. Next, we generated 100 parametric bootstrap replicates conditioned upon this phylogeny and the corresponding estimated model of evolution (Rambaut and Grassly 1997). For each simulated data set, we estimated the phylogeny and determined the distribution of multiple hits. Because we performed these simulations under complete linkage, multiple hits had to reflect true multiple hits resulting from among-site rate variation.

We also expected that the distribution of multiple hits should be proportionally identical at nonsynonymous *vs.* synonymous sites under a neutral recombination process. To test this expectation, we tested for identical distributions of multiple hits at nonsynonymous and synonymous sites using a test similar to a χ^2 goodness-of-fit test. To avoid problems with low expected frequencies, the null distribution was created using a resampling method rather than a χ^2 distribution. Following the procedure for a standard χ^2 test, we conditioned upon both the total number of events and the expected probabilities of the different categories of multiple hits at the two types of sites. Events were shuffled and resampled 10,000 times. The probability for a given category of multiple hits was determined and the differences between the replicate probabilities at nonsynonymous and synonymous sites, p_N and p_S, respectively, were recorded. The null distribution represents the sums of the absolute values of the differences in probabilities, *i.e.*, $\Sigma|p_N - p_S|_i$, for the ith category of multiple hits. The *P*-value was the rank of the sum from the actual data compared to the null distribution.

In coalescent theory, population-wide mutation in a haploid organism is described by the relationship $\theta = 2N_e\mu L$, in which θ is the population-scaled mutation rate, N_e is the effective population size, μ is the neutral mutation rate per site per generation, and L is the number of sites. "Generation" refers to one passage through the entire viral life cycle. Watterson's point estimate of θ is $\theta_W = S/a_n$, in which S is the number of observed variable sites and $a_n = \Sigma_{i=1}^{n-1}(1/i)$ for n sequences (Watterson 1975). To determine an expected value for the number of real multiple hits, we performed a coalescent simulation based on the coding subset of sites. We generated 1000 realizations of the coalescent conditioned on Watterson's point estimate of θ, the estimated parameters of the general time-reversible model of substitution with unequal base frequencies and gamma-distributed among-site rate variation, a sample size of 20 sequences, and a length of 2241 sites (N. C. Grassly and A. E. Rambaut, http://evolve.zoo.ox.ac.uk, TREEVOLVE version 1.32). For each replicate, we reconstructed the phylogeny and counted the number of multiple hits. The mean of these replicates reflects the expected number of real multiple hits given the observed shape parameter of gamma-distributed among-site rate variation.

Estimation of recombination rates: To determine a recombination rate, we used three coalescent theory-based approaches. In coalescent theory, population-wide recombination is described by the relationship $\rho = 2N_eCL$, in which ρ is the population-scaled recombination rate, N_e is the effective population size, C is the recombination rate per adjacent sites per generation, and L is the number of sites. The program RECOMBINE (Kuhner *et al.* 2000) was used to estimate θ and $r = C/\mu$ (thus $\rho = r\theta$). The transition/transversion ratio and the among-site rate variation parameters were estimated using PAUP*. Whereas Watterson's estimate of θ assumes infinite sites and no recombination, RECOMBINE relaxes both of these assumptions. Using a mean mutation rate for point substitutions μ of 2.5×10^{-5}/site/generation (Mansky 1996), we can derive an estimate of C. Variance in this estimate of μ was not taken into account.

In the second approach, we performed neutral coalescent simulations with a fixed S (S here was taken to be the total number of events expected from the coalescent simulation with among-site rate variation, so that it includes real multiple hits) and analyzed the effect of ρ on the number of multiple hits in the subsequently reconstructed branching phylogeny when recombination was ignored. Thus, we conditioned upon a sample size of 20 sequences, a locus of 2241 sites, S of 120, and ρ of 0, 1, 2, 4, 8, 16, 32, 64, 128, 256, and 512. Using the program "ms," 1000 independent data sets were generated for each ρ value (Hudson 2002). For each replicate data set, the maximum-likelihood branching phylogeny was reconstructed and the number of multiple hits was counted. To interpolate ρ for the sample from the observed number of multiple hits, we obtained the least-squares fit to the equation $y = c_1(1 - e^{-c_2x})$ using Mathematica, version 4.1.1.0 (Wolfram Research, Champaign, IL). In this equation, c_1 represents the maximum number of multiple hits, which is realized on a star phylogeny (Archie and Felsenstein 1993). We derived the following formulas for the expected value of c_1. Under the standard coalescent model, the expected number of total mutational events is $a_n\theta = \theta(1/1 + 1/2 + 1/3 + \cdots + 1/(n-1))$ (Watterson 1975). In this distribution, $1/1$ corresponds to singletons, $1/2$ corresponds to doubletons, and so forth (Wakeley and Takahashi 2003). When mapping such sequence data onto a star phylogeny, singletons require one event, doubletons require two events, and so forth. Thus, the total number of events on a rooted star phylogeny is given by $\theta(1/1 + 2/2 + 3/3 + \cdots + (n-1)/(n-1))$. However, if the star phylogeny is unrooted, there is no information regarding which state is ancestral and which is derived. In this case, it is more parsimonious to invoke multiple events to describe whichever state is at a frequency <50%. Thus, for an even number of n sequences, $c_1 = \theta((n/2) + \Sigma_{i=(n/2)+1}^{n-1}(n-i)/i - a_n)$ and for an odd number of n sequences, $c_1 = \theta((n-1)/2 + \Sigma_{i=(n+1)/2}^{(n-1)}(n-i)/i - a_n)$.

In the third approach, we used the program "pairwise" from the LDhat package (McVean *et al.* 2002). This program obtains an approximate-likelihood estimate of ρ given a point estimate of θ by combining the coalescent likelihoods of all pairwise comparisons of segregating sites, using a likelihood function based on linkage disequilibrium (Hudson 2001).

Detection of natural selection: We tested neutrality by comparing nonsynonymous and synonymous mutations. The numbers of potential nonsynonymous and synonymous sites for coding segments were separately calculated using CODEML from the Phylogenetic Analysis by Maximum Likelihood package, version 3.1 (Yang *et al.* 2000), which allowed us to account for gene-specific codon usage biases. The observed numbers of nonsynonymous and synonymous mutations were tested against expectations based upon the numbers of potential nonsynonymous and synonymous sites, using a χ^2 test with 1 d.f.

We also explored the issue of neutrality using Tajima's D test (Tajima 1989). This test is known to be conservative in the presence of recombination (Tajima 1989). We therefore created a null distribution of 10,000 replicates conditioned upon the values of n, L, S, and $\rho = 25.9$ (the smallest of the three point estimates of ρ we obtained), using DnaSP, version 3.53 (Rozas and Rozas 1999).

RESULTS

The null assumption of linkage: To begin to assess coevolution of the six HIV-1 auxiliary genes, we sought to control for the possibility of PCR-mediated recombina-

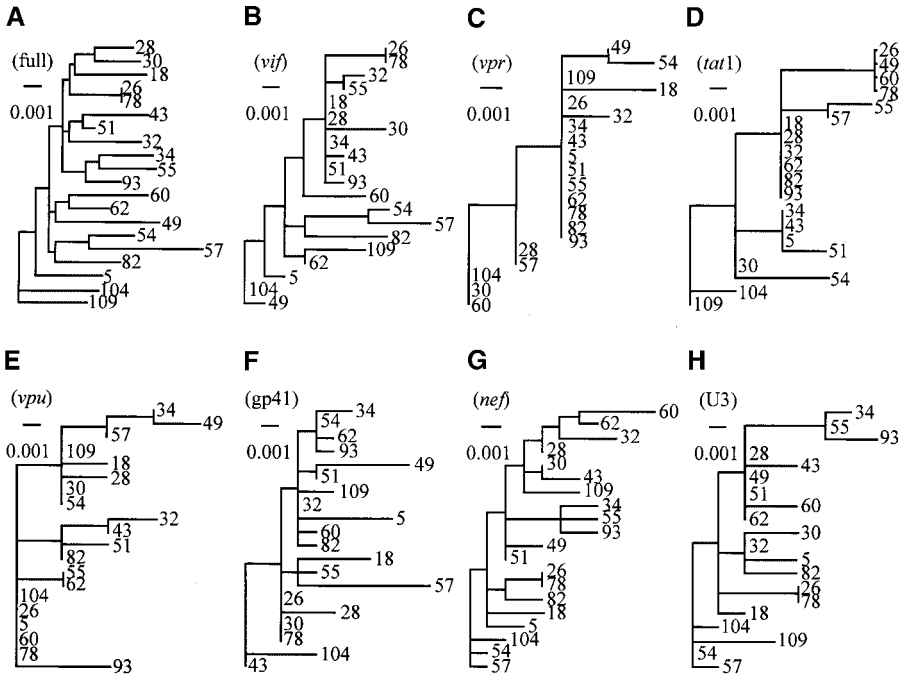

FIGURE 3.—Reconstructed maximum-likelihood phylogenies for each gene segment evaluated. The phylogenies represent (A) the full set of 2607 sites; (B) *vif*, corresponding to sites 1–579; (C) *vpr*, corresponding to sites 519–812; (D) *tat*1, corresponding to sites 793–1007; (E) *vpu*, corresponding to sites 1024–1317; (F) gp41, corresponding to sites 1318–1869; (G) *nef*, corresponding to sites 1871–2515; and (H) U3, corresponding to sites 2184–2607. Under each letter is the name of the segment followed by a bar representing 0.001 substitutions/site.

tion. We first assessed the sensitivity of the PCR primers. First-round primers DS3 and DS8 amplified a 4.6-kb fragment encompassing the 3′ half of the viral genome (Figure 1B), whereas primers DS3 and DS4 amplified a 1.3-kb fragment (Figure 1D). When nested second-round PCR was conducted using primers DS1 and DS2 on the 1.3-kb template, we detected single bands of the expected size from a nominally one-input template (data not shown). Nested amplification using DR7 and DR8 on the 4.6-kb fragment was marginally more sensitive than nested amplification using DS1 and DS2 (data not shown). This result indicated that our endpoint dilution protocol would err on the conservative side, meaning that we would have more negative replicates than expected. Nested amplification using DS1 and DS2 on the 4.6-kb fragment was positive down to 10 input copies (data not shown), indicating that short PCR in the first round was somewhat more efficient than long PCR. This raised the possibility that multiple templates were still present in the positive half-genome PCR replicates, even after limiting dilution. However, multiple coamplified templates in a real sample should give rise to ambiguous sequence when PCR products are directly sequenced. Of 125 PCR replicates using a nominal average input of 0.4 copies/replicate, 24 (19.2%) were positive, of which 20 (83%) yielded unambiguous sequence. The four positives that did not yield unambiguous sequence were discarded from further analysis and were attributed to coamplification of multiple templates during PCR. The concordance of these results with the expectations based on a Poisson distribution of input copies indicated that those replicates that ultimately yielded unambiguous sequences were likely to have been derived from single templates (RODRIGO *et al.* 2000).

Detection of recombination: Reconstructed phylogenies for the full-length sequences as well as for each segment are shown in Figure 3. We assessed whether the topologies of the segments varied statistically, either between themselves or compared to the full sequence topology, using the Shimodaira-Hasegawa multiple-comparison test (SHIMODAIRA and HASEGAWA 1999). As shown in Table 1, the only segments for which the topologies were not statistically different were the *nef*-U3 pair. However, this was expected given their considerable genetic overlap (the 3′ 50% of *nef* encompasses the 5′ 75% of U3). Three of seven of the full sequence *vs.* segment topology tests also failed. These results demonstrated significant incongruence of topologies along the viral genome.

We then investigated topological incongruence by analyzing the patterns of multiple hits. As shown in Figure 4, the observed data (depicted by horizontal lines) yielded significantly more parallelisms, reversals, and parallel reversals, and fewer third state changes, for a comparable number of total multiple hits observed in a simulation in which there was complete linkage with gamma-distributed rate variation ($P < 0.01$ for the comparison of the joint distributions). Both of these results are consistent with expectations of a neutral recombination process, but inconsistent with expectations of both complete linkage and parallel evolution (in which scenario reversals would be unexpected). Furthermore, the relative lack of third state changes is inconsistent with the presence of positive, diversifying selection.

Estimation of the recombination rate: We used RECOMBINE to estimate θ and r, the recombination rate relative to the point mutation rate, for the coding subset of sites. The mean estimate of θ was 74.2 (95% confi-

TABLE 1

Results of the Shimodaira-Hasegawa multiple-comparison test

	Full	*vif*	*vpr*	*tat*1	*vpu*	gp41	*nef*	U3
Full		0.431	*0.044*	*0.006*	*0.003*	0.051	0.070	0.132
vif	0.291		*0.046*	*0.007*	*0.005*	*0.029*	*0.006*	*0.008*
vpr	*0.001*	*0.011*		*0.003*	*0.003*	*0.033*	*0.003*	*0.003*
*tat*1	*0.009*	*0.010*	*0.046*		*0.004*	*0.046*	*<0.001*	*0.007*
vpu	*0.005*	*0.010*	*0.046*	*0.004*		*0.040*	*0.001*	*0.003*
gp41	*0.003*	*0.006*	*0.046*	*0.003*	*0.003*		*<0.001*	*0.003*
nef	0.267	*0.008*	*0.047*	*0.004*	*0.002*	*0.030*		0.493
U3	0.185	*0.010*	*0.046*	*0.002*	*0.006*	*0.038*	0.254	

The test compared the alternative topology for the segment listed in the column against the null topology and the sequence data from the segment listed in the row. The values given are the resultant *P*-values from 1000 replicates with full-likelihood optimization, with italic values indicating significance at the 5% level.

dence interval 48.0–108.0) and the mean estimate of r was 0.349 (95% confidence interval 0.224–0.506), yielding the mean estimate $\rho = 25.9$ (95% confidence interval 10.8–54.6; Table 2). Fixing the mutation rate for point substitutions at 2.5×10^{-5}/site/ generation (Mansky 1996), we estimated the mean recombination rate C to be 8.73×10^{-6}/adjacent sites/generation, ranging from 5.60×10^{-6} to 1.27×10^{-5}.

We also estimated a recombination rate based on the curve of multiple hits as a function of ρ. The expected number of multiple hits can be estimated from the observed values of S and the shape parameter for gamma-distributed among-site rate variation, leading to an estimate of 9.4 with the mean estimate of the shape parameter equal to 0.0145, 9.2 with the upper bound estimate of the shape parameter (0.09), and 8.5 with the lower bound estimate of the shape parameter (1.5×10^{-9}). From this simulation, we estimated an effective value of S that included among-site rate variation as 120, yielding an effective value of θ of 33.8. Although this inference assumed no recombination, the expected number of mutational events is unaffected by recombination (compare Ewens 1974 to Watterson 1975). We subtracted 9 from the 75 multiple hits inferred from the maximum-likelihood phylogeny reconstructed from the coding subset of sites. We then estimated from the least-squares fit to the curve in Figure 5 that 66 recombination-induced artifacts corresponded to $\rho = 77.7$ (95% confidence interval 33.3–140.5), yielding the mean estimate $C = 5.75 \times 10^{-5}$/adjacent sites/generation (95% confidence interval 2.46×10^{-5}–1.04×10^{-4}; Table 2).

Third, using LDhat, ρ was estimated to be 187 (Figure 6). We then estimated C to be 1.38×10^{-4}/adjacent sites/generation (Table 2). There is no known analytical expression for the variance of C using this method. Because the three estimates of C varied by over an order of magnitude, we performed the following simulation. Sequences were simulated under the neutral coalescent model with recombination (http://www.brics.dk/compbio/meta_hudson/sim/combined.html). Coalescent parameter estimates were then obtained from both RECOMBINE and LDhat (Table 3). From these simulations, as

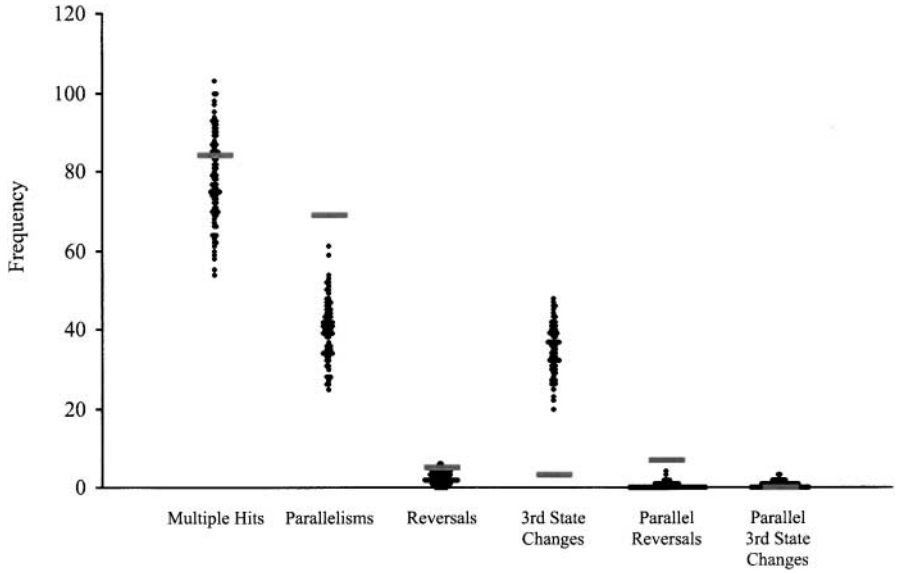

Figure 4.—Distribution of categorized multiple hits. A total of 100 parametric bootstraps were created using the reconstructed full-sequence phylogeny and its corresponding model of evolution under the null assumption of complete linkage. The diamonds represent the values for each of the replicate simulated data sets and the shaded horizontal line represents the values observed from the original data.

TABLE 2

Mean estimates of the coalescent recombination parameters for the coding subset of sites

Coalescent parameter	RECOMBINE	Multiple hits	LDhat
θ	74.2	33.8	33.8
r	0.349	2.30	5.53
ρ	25.9	77.7	187
C	8.73×10^{-6}	5.75×10^{-5}	1.38×10^{-4}

θ is the estimated population-scaled mutation rate per locus. *r* is the estimated relative recombination-to-mutation rate. ρ is the estimated population-scaled recombination rate per locus. *C* is the estimated recombination rate per adjacent sites per generation. The neutral mutation rate μ was assumed to be 2.5×10^{-5}/site/generation.

ρ increased, RECOMBINE more severely overestimated θ and underestimated ρ/θ, whereas LDhat slightly overestimated θ and underestimated ρ. We thus conclude that the LDhat estimates of ρ are the most reliable of the three, indicating that there were from one to two recombination events per genome per generation *in vivo*.

Detection of natural selection: The coding subset of sites contains 58 nonsynonymous and 54 synonymous mutations, across 1643 potential nonsynonymous sites and 598 potential synonymous sites. Synonymous mutations were in significant excess over nonsynonymous mutations ($P = 2.6 \times 10^{-7}$), thus suggesting the presence of negative, purifying selection. If we assume that nine real multiple hits all were nonsynonymous, synonymous mutations were still in excess ($P = 8.1 \times 10^{-6}$). Furthermore, when nonsynonymous sites and synonymous sites were considered independently, the distributions of multiple hits were not statistically different ($P = 0.290$; Figure 7). This result suggested that recombination was acting similarly at both nonsynonymous and synonymous sites, even in the presence of purifying selection.

Given the presence of a high rate of recombination, we next used Tajima's *D* test (Tajima 1989). The test statistic of −1.64 obtained indicated the presence of an excess of low-frequency mutants. This excess was not significant if recombination was ignored ($P = 0.0650$), but was significant ($P = 0.0010$) under a null hypothesis that incorporated recombination at the smallest mean rate we obtained (see materials and methods). We conclude that a significant departure from neutrality was masked by the presence of recombination.

DISCUSSION

In this study, we demonstrated that recombination occurred frequently between closely related HIV-1 sequences within an individual infected from a single source. Our results suggest that the *in vivo* recombination rate is high (ignoring variance in the mutation rate), approximately one to two recombination events per genome per generation on average, one of the highest rates among pathogens yet reported (Awadalla 2003). This *in vivo* estimate is comparable to the estimate of two to four crossovers per genome per generation obtained using a cell culture system (Jetzt *et al.* 2000; Zhuang *et al.* 2002; Onafuwa *et al.* 2003). A high recombination rate for viruses from the same strain within an infected individual requires a relatively high number of multiply infected cells (Gratton *et al.* 2000; Jung *et al.* 2002), which may occur as a function of highly localized foci of infection as described by metapopulation dynamics (Frost *et al.* 2001).

Three elements of our experimental procedures were designed to prevent artifactual, recombination-like events that might occur during PCR or cloning. The use of molecular endpoints as input for PCR was intended to minimize recombination during PCR (Meyerhans *et al.* 1990; Yang *et al.* 1996). The positive control experiments performed to validate the long, nested, limiting-dilution PCR approach demonstrated that, to the best of our abil-

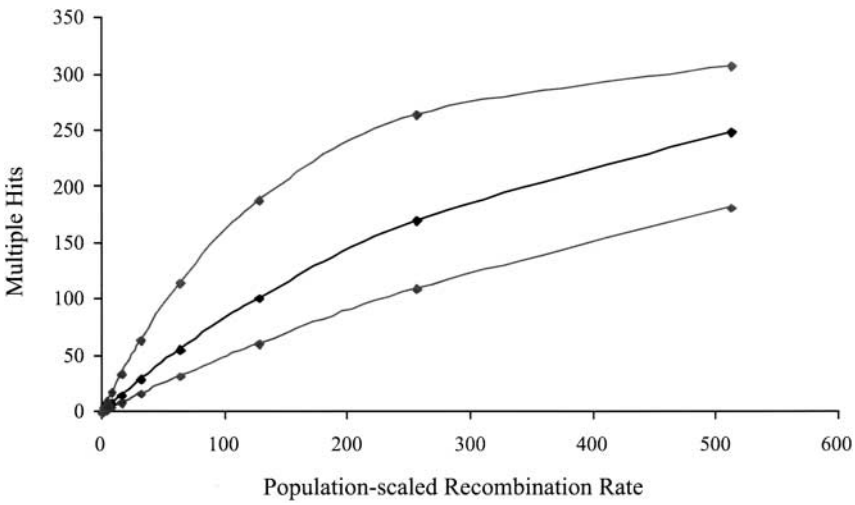

Figure 5.—Effect of recombination rate on the number of multiple hits. Shown are the means (solid curve) and 95% ranges (shaded curves) for 1000 independent replicates.

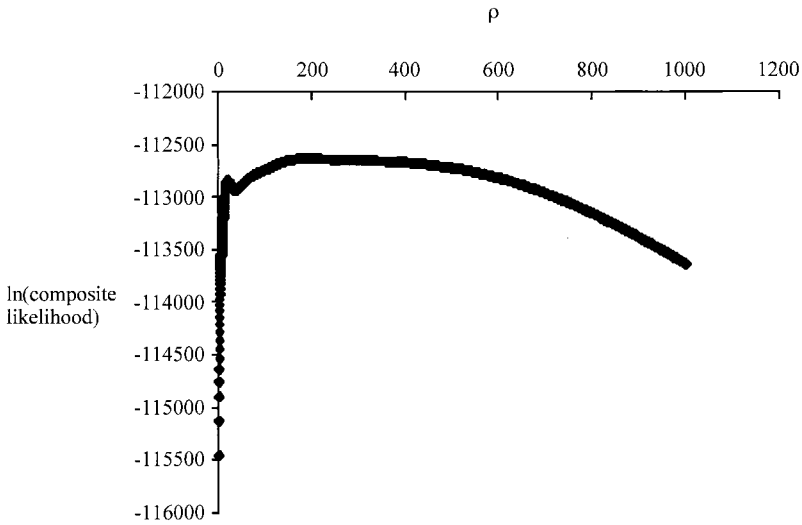

FIGURE 6.—Likelihood surface for the approximate-likelihood estimator of ρ. ρ was evaluated over the range 0–1000, separated into intervals of 0.1. θ was taken to be 33.8. On the *y*-axis, values closer to 0 indicate more likely values of ρ.

ity, all 20 sequences analyzed were generated from single amplifiable templates, thus excluding PCR-mediated recombination from having anything other than a minor contribution to the estimated recombination rate. Direct sequencing of PCR products was used to prevent detection of misincorporations that may have occurred during PCR.

One experimental issue we did not directly address was that of template switching during cDNA synthesis. Previous studies have demonstrated that efficient minus-strand transfer (*i.e.*, transfer of nascent DNA across RNA donor and acceptor templates) depends on RNase H activity (LUO and TAYLOR 1990; TANESE *et al.* 1991; DESTEFANO *et al.* 1992; PELISKA and BENKOVIC 1992; TELESNITSKY *et al.* 1992). Because most recombination events occur during minus-strand synthesis (ZHANG *et al.* 2000), the implication is that RNase H activity is likely to be important, which implies that recombination during cDNA synthesis with an RNase H-deficient reverse transcriptase (as we used) is unlikely. Further, since cDNA synthesis involves only a single incubation step, template switching would not occur from repeated denaturation and renaturation as occurs during PCR. Finally, a mean ρ value of 187 corresponds to an expected $187 \times \Sigma_{i=1}^{19}(1/i) = 663$ recombination events in the history of 20 sequences back to their most recent common ancestor (HUDSON and KAPLAN 1985). It therefore seems highly unlikely that the envisioned experimental artifacts influenced our estimates of the numbers of mutation and recombination events.

In this work, we used coalescent theory to describe the history of sequences derived from one population. All three recombination rate estimators assumed panmixis, a constant population size, and neutrality. As far as panmixis is concerned, we note that the individual studied represented a case of single-source infection. Thus, any population subdivision had to exist between anatomical or temporal compartments that still shared a most recent common ancestor for the entire infection within this individual. Furthermore, population subdivision generally yields a positive Tajima's *D* statistic (FU 1996), whereas we observed a negative Tajima's *D* statistic. A negative Tajima's *D* statistic is consistent with population growth (TAJIMA 1989). However, the assumption of a constant population size was likely not violated given that our sample was derived from a time in chronic infection when population diversity appeared stable.

Regarding the assumption of neutrality, the excess of synonymous mutations suggests the presence of purifying selection. This excess is consistent with negative Tajima's *D* if the low-frequency mutants represent delete-

TABLE 3

Simulation study of RECOMBINE and LDhat recombination rate estimator performance

	RECOMBINE		LDhat	
ρ	θ	ρ/θ	θ	ρ
0	50.8 (13.6)	0.004 (0.010)	53.5 (17.0)	0.00 (0.00)
1	54.0 (16.2)	0.020 (0.019)	53.1 (22.5)	0.70 (1.06)
10	66.9 (11.1)	0.136 (0.021)	53.7 (8.6)	9.31 (5.32)
100	79.7 (10.6)	0.472 (0.109)	51.1 (6.0)	97.10 (43.09)

For each value of ρ, 10 sets of 20 sequences of length 1000 were generated and evaluated. θ was fixed at 50. Shown are the means (standard deviation).

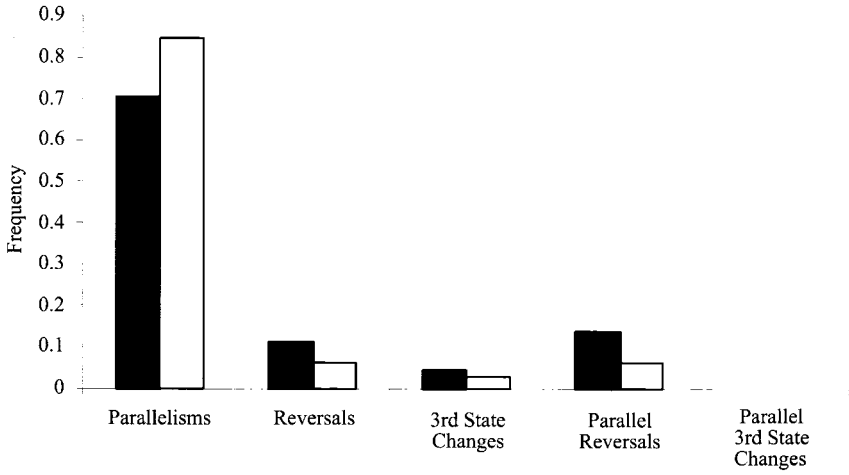

FIGURE 7.—Proportional distribution of categorized multiple hits. Solid bars represent nonsynonymous sites and open bars represent synonymous sites.

rious mutants. The lower amount of nucleotide diversity in the sequences we examined, 1.3% as compared to 2.1% in *env*, may represent a lower neutral mutation rate and/or stronger functional constraints in the auxiliary genes. Because Tajima's *D* test yielded strong significance when recombination was accounted for, the test was powerful enough to detect departures from neutral expectations in these data. However, because the test yielded no significance with unacknowledged recombination, recombination was revealed to be prevalent enough to mask selection in the history of these sequences.

Currently, McVEAN *et al.* (2002) provide the only other *in vivo* estimate of the recombination rate for HIV-1; they estimated $\rho > 100$. Their sequences reflect sampling within a viral subtype across many individuals, whereas our sequences reflect sampling within the population of a single individual. It is highly likely that these different population levels have different effective population sizes, so the estimates of the population-scaled recombination rate ρ are not expected to be comparable. Also, McVEAN *et al.* (2002) did not decompose their estimate of ρ to estimate *C*, the recombination rate per adjacent sites per generation; hence, our mean estimate of 1.38×10^{-4} recombination events/adjacent sites/generation is the first *in vivo* estimate reported for HIV-1.

We thank John E. Mittler and Mark A. Jensen for several helpful discussions and S. Otto and the anonymous reviewers for numerous comments that helped us greatly improve the manuscript. D.S. was a Howard Hughes Medical Institute Predoctoral Fellow. This work was supported by grants from the United States Public Health Service, including support from the University of Washington Center for AIDS Research.

LITERATURE CITED

ARCHIE, J. W., and J. FELSENSTEIN, 1993 The number of evolutionary steps on random and minimum length trees for random evolutionary data. Theor. Popul. Biol. **43:** 52–79.

AUSUBEL, F. M., R. BRENT, R. E. KINGSTON, D. D. MOORE, J. G. SEIDMAN *et al.*, 1990 *Current Protocols in Molecular Biology*. John Wiley & Sons, New York.

AWADALLA, P., 2003 The evolutionary genomics of pathogen recombination. Nat. Rev. Genet. **4:** 50–60.

CARR, J. K., M. O. SALMINEN, C. KOCH, D. GOTTE, A. W. ARTENSTEIN *et al.*, 1996 Full length sequence and mosaic structure of a HIV-1 isolate from Thailand. J. Virol. **70:** 5935–5943.

CULLEN, B. R., 1998 HIV-1 auxiliary proteins: making connections in a dying cell. Cell **93:** 685–692.

DELASSUS, S., R. CHEYNIER and S. WAIN-HOBSON, 1991 Evolution of human immunodeficiency virus type 1 *nef* and long terminal repeat sequences over 4 years *in vivo* and *in vitro*. J. Virol. **65:** 225–231.

DESTEFANO, J. J., L. M. MALLABER, L. RODRIGUEZ-RODRIGUEZ, P. J. FAY and R. A. BAMBARA, 1992 Requirements for strand transfer between internal regions of heteropolymer templates by human immunodeficiency virus reverse transcriptase. J. Virol. **66:** 6370–6378.

DIAZ, R. S., E. C. SABINO, A. MAYER, J. W. MOSLEY and M. P. BUSCH, 1995 Dual human immunodeficiency virus type 1 infection and recombination in a dually exposed transfusion recipient. J. Virol. **69:** 3272–3281.

EWENS, W. J., 1974 A note on the sampling theory for infinite alleles and infinite sites models. Theor. Popul. Biol. **6:** 143–148.

FROST, S. D., M. J. DUMAURIER, S. WAIN-HOBSON and A. J. BROWN, 2001 Genetic drift and within-host metapopulation dynamics of HIV-1 infection. Proc. Natl. Acad. Sci. USA **98:** 6975–6980.

FU, Y.-X., 1996 New statistical tests of neutrality for DNA samples from a population. Genetics **143:** 557–570.

GAO, F., L. YUE, A. T. WHITE, P. G. PAPPAS, J. BARCHUE *et al.*, 1992 Human infection by genetically diverse SIVsm-related HIV-2 in West Africa. Nature **358:** 495–499.

GAO, F., L. YUE, D. L. ROBERTSON, S. C. HILL, H. HUI *et al.*, 1994 Genetic diversity of human immunodeficiency virus type 2: evidence for distinct sequence subtypes with differences in virus biology. J. Virol. **68:** 7433–7447.

GAO, F., D. L. ROBERTSON, S. G. MORRISON, H. HUI, S. CRAIG *et al.*, 1996 The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin. J. Virol. **70:** 7013–7029.

GRATTON, S., R. CHEYNIER, M. J. DUMAURIER, E. OKSENHENDLER and S. WAIN-HOBSON, 2000 Highly restricted spread of HIV-1 and multiply infected cells within splenic germinal centers. Proc. Natl. Acad. Sci. USA **97:** 14566–14571.

GROENINK, M., A. ANDEWEG, R. FOUCHIER, S. BROERSEN, R. VAN DER JAGT *et al.*, 1992 Phenotype-associated *env* gene variation among eight related human immunodeficiency virus type 1 clones: evidence for *in vivo* recombination and determinants of cytotropism outside the V3 domain. J. Virol. **66:** 6175–6180.

HOWELL, R. M., J. E. FITZGIBBON, M. NOE, Z. REN, D. GOCKE *et al.*, 1991 *In vivo* sequence variation of the human immunodeficiency virus type 1 *env* gene: evidence for recombination among

variants found in a single individual. AIDS Res. Hum. Retroviruses **7:** 869–876.

Hu, W.-S., and H. M. Temin, 1990a Genetic consequences of packaging two RNA genomes in one retroviral particle: pseudodiploidy and high rate of genetic recombination. Proc. Natl. Acad. Sci. USA **87:** 1556–1560.

Hu, W.-S., and H. M. Temin, 1990b Retroviral recombination and reverse transcription. Science **250:** 1227–1233.

Hudson, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159:** 1805–1817.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18:** 337–338.

Hudson, R. R., and N. L. Kaplan, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

Jetzt, A. E., H. Yu, G. J. Klarmann, Y. Ron, B. D. Preston *et al.*, 2000 High rate of recombination throughout the human immunodeficiency virus type 1 genome. J. Virol. **74:** 1234–1240.

Jung, A., R. Maier, J.-P. Vartanian, G. Bocharov, V. Jung *et al.*, 2002 Multiply infected spleen cells in HIV patients. Nature **418:** 144.

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

Kuhner, M. K., J. Yamato and J. Felsenstein, 2000 Maximum likelihood estimation of recombination rates from population data. Genetics **156:** 1393–1401.

Leitner, T., D. Escanilla, S. Marquina, J. Wahlberg, C. Brostrom *et al.*, 1995 Biological and molecular characterization of subtype D, G, and A/D recombinant HIV-1 transmissions in Sweden. Virology **209:** 136–146.

Liu, S.-L., T. Schacker, L. Musey, D. Shriner, M. J. McElrath *et al.*, 1997 Divergent patterns of progression to AIDS after infection from the same source: human immunodeficiency virus type 1 evolution and antiviral responses. J. Virol. **71:** 4284–4295.

Liu, S.-L., J. E. Mittler, D. C. Nickle, T. M. Mulvania, D. Shriner *et al.*, 2002 Selection for human immunodeficiency virus type 1 recombinants in a patient with rapid progression to AIDS. J. Virol. **76:** 10674–10684.

Luo, G., and J. Taylor, 1990 Template switching by reverse transcriptase during DNA synthesis. J. Virol. **64:** 4321–4328.

Mansky, L. M., 1996 Forward mutation rate of human immunodeficiency virus type 1 in a T lymphoid cell line. AIDS Res. Hum. Retroviruses **12:** 307–314.

Martins, L. P., N. Chenciner, B. Åsjö, A. Meyerhans and S. Wain-Hobson, 1991 Independent fluctuation of human immunodeficiency virus type 1 *rev* and gp41 quasispecies *in vivo*. J. Virol. **65:** 4502–4507.

Maynard Smith, J., and N. H. Smith, 1998 Detecting recombination from gene trees. Mol. Biol. Evol. **15:** 590–599.

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

Meyerhans, A., R. Cheynier, J. Albert, M. Seth, S. Kwok *et al.*, 1989 Temporal fluctuations in HIV quasispecies *in vivo* are not reflected by sequential HIV isolations. Cell **58:** 901–910.

Meyerhans, A., J. P. Vartanian and S. Wain-Hobson, 1990 DNA recombination during PCR. Nucleic Acids Res. **18:** 1687–1691.

Michael, N. L., G. Chang, L. A. d'Arcy, P. K. Ehrenberg, R. Mariani *et al.*, 1995 Defective accessory genes in a human immunodeficiency virus type 1-infected long-term survivor lacking recoverable virus. J. Virol. **69:** 4228–4236.

Morris, A., M. Marsden, K. Halcrow, E. S. Hughes, R. P. Brettle *et al.*, 1999 Mosaic structure of the human immunodeficiency virus type 1 genome infecting lymphoid cells and the brain: evidence for frequent *in vivo* recombination events in the evolution of regional populations. J. Virol. **73:** 8720–8731.

Onafuwa, A., W. An, N. D. Robson and A. Telesnitsky, 2003 Human immunodeficiency virus type 1 genetic recombination is more frequent than that of Moloney murine leukemia virus despite similar template switching rates. J. Virol. **77:** 4577–4587.

Peliska, J. A., and S. J. Benkovic, 1992 Mechanism of DNA strand transfer reactions catalyzed by HIV-1 reverse transcriptase. Science **258:** 1112–1118.

Rambaut, A., and N. C. Grassly, 1997 Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. **13:** 235–238.

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

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

Rodrigo, A. G., P. C. Goracke, K. Rowhanian and J. I. Mullins, 1997 Quantitation of target molecules from polymerase chain reaction-based limiting dilution assays. AIDS Res. Hum. Retroviruses **13:** 737–742.

Rodrigo, A. G., E. W. Hanley, P. C. Goracke and G. H. Learn, 2000 Sampling and processing HIV molecular sequences: a computational evolutionary biologist's perspective, pp. 1–17 in *Computational and Evolutionary Analyses of HIV Sequences*, edited by A. G. Rodrigo and G. H. Learn. Kluwer Academic Publishers, Boston.

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.

Sabino, E. C., E. G. Shpaer, M. G. Morgado, B. T. Korber, R. S. Diaz *et al.*, 1994 Identification of human immunodeficiency virus type 1 envelope genes recombinant between subtypes B and F in two epidemiologically linked individuals from Brazil. J. Virol. **68:** 6340–6346.

Salminen, M. O., J. K. Carr, D. L. Robertson, P. Hegerich, D. Gotte *et al.*, 1997 Evolution and probable transmission of intersubtype recombinant human immunodeficiency virus type 1 in a Zambian couple. J. Virol. **71:** 2647–2655.

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

Shankarappa, R., J. B. Margolick, S. J. Gange, A. G. Rodrigo, D. Upchurch *et al.*, 1999 Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J. Virol. **73:** 10489–10502.

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

Stuhlmann, H., and P. Berg, 1992 Homologous recombination of copackaged retrovirus RNAs during reverse transcription. J. Virol. **66:** 2378–2388.

Swofford, D. L., 2002 *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods)*, Version 4b10. Sinauer Associates, Sunderland, MA.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Tanese, N., A. Telesnitsky and S. P. Goff, 1991 Abortive reverse transcription by mutants of Moloney murine leukemia virus deficient in the reverse transcriptase-associated RNase H function. J. Virol. **65:** 4387–4397.

Telesnitsky, A., S. W. Blain and S. P. Goff, 1992 Defects in Moloney murine leukemia virus replication caused by a reverse transcriptase mutation modeled on the structure of *Escherichia coli* RNase H. J. Virol. **66:** 615–622.

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

Wakeley, J., and T. Takahashi, 2003 Gene genealogies when the sample size exceeds the effective size of the population. Mol. Biol. Evol. **20:** 208–213.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

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

Yang, Y. L., G. Wang, K. Dorman and A. H. Kaplan, 1996 Long polymerase chain reaction amplification of heterogeneous HIV type 1 templates produces recombination at a relatively high frequency. AIDS Res. Hum. Retroviruses **12:** 303–306.

Yang, Z., 1994 Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39:** 306–314.

Yang, Z., R. Nielsen, N. Goldman and A. M. Pedersen, 2000 Codon-

substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155:** 431–449.

Zhang, J., L. Y. Tang, T. Li, Y. Ma and C. M. Sapp, 2000 Most retroviral recombinations occur during minus-strand DNA synthesis. J. Virol. **74:** 2313–2322.

Zhang, L., Y. Huang, H. Yuan, S. Tuttleton and D. D. Ho, 1997 Genetic characterization of *vif*, *vpr*, and *vpu* sequences from long-term survivors of human immunodeficiency virus type 1 infection. Virology **228:** 340–349.

Zhu, T., N. Wang, A. Carr, S. Wolinsky and D. D. Ho, 1995 Evidence for coinfection by multiple strains of human immunodeficiency virus type 1 subtype B in an acute seroconvertor. J. Virol. **69:** 1324–1327.

Zhuang, J., A. E. Jetzt, G. Sun, H. Yu, G. Klarmann *et al.*, 2002 Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. J. Virol. **76:** 11273–11282.

Communicating editor: S. P. Otto