JOURNAL OF VIROLOGY, Feb. 2006, p. 1637–1644 Vol. 80, No. 4
0022-538X/06/$08.00+0 doi:10.1128/JVI.80.4.1637–1644.2006

Human Immunodeficiency Virus Type 1 *env* Evolves toward Ancestral States upon Transmission to a New Host

Joshua T. Herbeck,[1] David C. Nickle,[1] Gerald H. Learn,[1] Geoffrey S. Gottlieb,[2] Marcel E. Curlin,[2] Laura Heath,[1] and James I. Mullins[1,2]*

Departments of Microbiology[1] and Medicine,[2] University of Washington School of Medicine, Seattle, Washington 98195-8070

Received 20 August 2005/Accepted 21 November 2005

Selecting human immunodeficiency virus (HIV) sequences for inclusion within vaccines has been a difficult problem, as circulating HIV strains evolve relentlessly and become increasingly divergent over time. We report an assessment of this divergence from three perspectives: (i) across different hosts as a function of time of infection, (ii) between donors and recipients in known transmission pairs, and (iii) within individual hosts over time in relation to the initially replicating virus and to the deduced ancestral sequence of the intrahost viral population. Surprisingly, we consistently found less divergence between viruses from different individuals sampled in primary infection than in individuals sampled at more advanced stages of illness. Furthermore, longitudinal analysis of intrahost divergence revealed a 2- to 3-year period of evolution toward a common ancestral sequence at the start of infection, indicating that HIV recovers certain ancestral features when infecting a new host. These results have important implications for the study of HIV population genetics and rational vaccine design, including favoring the inclusion of viral gene sequences taken early in infection.

Human immunodeficiency virus type 1 (HIV-1) is a difficult target for vaccine development, due in part to its enormous capacity to mutate and escape immune recognition (27, 30), as well as the large and continually expanding (14) genetic diversity seen among circulating strains (29). In rational vaccine design, importance is given to viral strain selection as a means of optimizing vaccine reactivity to the largest number of potential infecting strains. We along with others have proposed strategies of finding optimal immunogens through computational and phylogenetic analyses of interhost sequence data sets (10, 20, 29, 31). These approaches attempt to minimize genetic divergence between possible immunogens and circulating viral strains.

It is also of critical public health importance to understand the size, growth rate, and distribution patterns of the HIV pandemic. In this context, interhost sequence data sets are analyzed with population genetic and genealogic methods to infer the past and present population dynamics of HIV (11, 33, 46), to measure HIV diversity in emerging local epidemics (1, 46), and to date the introduction of HIV-1 into the human population (18, 34). The main group of HIV-1, group M, the cause of most cases of AIDS worldwide, can be classified into nine subtypes and at least 16 circulating recombinant forms (22). Most epidemiological studies have focused on specific countries and the particular subtypes occurring in those countries, revealing large differences among countries and geographic regions (1, 34).

Biological processes that can influence patterns of genetic diversity and evolution in a population include mutation, recombination, migration, natural selection, and genetic drift. For HIV, as a pathogen with interhost transmission, an additional process involves the geography and dynamics of human interaction. All of these processes may affect the patterns of variation observed in HIV interhost sequence data sets. However, HIV accumulates genetic diversity only within individual hosts. The diversity of the intrahost viral population is generally low after transmission and increases during the course of infection at a rate of approximately 1% per year (in the C2 to V5 region of the envelope glycoprotein gene [36]), with sequence pairs reaching at least 15% difference in long-term survivors.

A main assumption in evolutionary analyses of HIV interhost sequence data sets is that genetic divergence accumulated within hosts is maintained through transmission events. If we accept this assumption, it follows that HIV interhost divergence from the estimated most recent common ancestor (MRCA) increases through time, with no punctuation at interhost transmission, and regardless of the time each viral lineage spends in an infected individual. In other words, all viral lineages in all infected individuals would be expected to be equally divergent from the MRCA at any given moment in time, regardless of the duration of infection within a given individual or how many transmission events have occurred in a given viral lineage. One then predicts that any two distinct, contemporaneously sampled interhost data sets from comparable cohorts (including treatment regimens, if applicable) within a local epidemic will contain equivalent levels of genetic diversity and divergence.

There is sufficient biological evidence to question the above assumption. For example, some cytotoxic T-lymphocyte (CTL) escape epitopes are known to revert after transmission to a new host (see, for example, reference 23), CXCR4 receptor-utilizing (X4) viruses are notably absent during early infection (see, for example, references 36 and 42), and intrahost viral populations undergo homogenization in acute infection (21), irrespective of the route of transmission (37). While the degree of genetic heterogeneity observed in primary infection differs

* Corresponding author. Mailing address: Department of Microbiology, SC 42, University of Washington, Seattle, WA 98195. Phone: (206) 732-6163. Fax: (206) 732-6167. E-mail: jmullins@u.washington.edu.

TABLE 1. Estimates of mean pairwise divergence (corrected and uncorrected π) and θ for HIV-1 subtype B interhost data sets

Cohort (time point)	Parameter				
	π	π (HKY)	π (ML)	θ_π	θ_w
MACS (1993)[b]					
Mean	0.137	0.155	0.21	80.23	96.51
Variance	1.80E-05	2.90E-05	1.70E-04	7.52	11.73
PIC (1993)[c]					
Mean	0.112	0.123	0.157	59.74	81.87
Variance	2.00E-06	3.00E-06	1.20E-05	0.56	0.76
Lyon (1993) mean[d]	0.119	0.132	0.162	71.25	87.55

[a] For MACS and PIC, there are 10 replicate data sets of nine sequences each (randomly chosen clonal sequences from nine subjects). No sequences from different subjects were found to display a close phylogenetic relationship consistent with epidemiologic linkage (data not shown). HKY, HKY85; ML, maximum likelihood; θ_w, diversity based on number of segregating sites.

[b] Late infection time points (except for one subject for which 1991 was the latest time point available).

[c] Primary infection time points.

[d] Lyon 1993, 1994, and 1995 primary infection time points.

between studies (7, 25, 44, 47), it has long been accepted that HIV populations undergo one or more substantial bottlenecks during early infection (45). For example, a recent study that documented viral genetic variability in eight heterosexual transmission pairs clearly demonstrated that an extreme bottleneck can accompany transmission from donor to recipient (8). Despite these observations, the stage of illness has not been explicitly considered as an important variable in analyses of interhost sequence data sets.

In this study, we examined the possible impact of HIV intrahost evolution on estimates of HIV interhost genetic diversity and divergence. First, we used data sets from three distinct subject cohorts representing different times since infection but sampled within the same calendar year (Multicenter AIDS Cohort Study [MACS] and University of Washington Primary Infection Cohort [PIC]) or roughly the same calendar period (Lyon cohort). Second, we used known transmission pairs to compare viral divergence from the interhost MRCA in the donor and the recipient. Third, we examined intrahost viral evolution in a longitudinal cohort.

MATERIALS AND METHODS

Interhost sequence data sets. We used serially sampled longitudinal gene sequence data from the *env* C2 to V5 region of HIV-1 in nine infected MACS subjects with typical times to progression to AIDS (15, 36). Subjects were sampled at roughly semiannual intervals ranging from about 2 months to 11.5 years postseroconversion (during the period 1984 to 1996), and an average of 12.3 clonal sequences were obtained from peripheral blood mononuclear cells (PBMC) and/or plasma at each time point (Table 1). From these longitudinal samples, we created interhost data sets containing one clonal sequence per subject, all representing the calendar year 1993 (with the exception of one subject in whom the last sample was from 1991). These sequences are available in the GenBank database under accession numbers AF13769 to AF138163, AF138166 to AF138263, and AF138305 to AF138703. Because multiple sequences were available for each sample, we created 10 replicate data sets by randomly sampling sequences from the given time point, without replacement when possible. In these MACS subjects, samples from 1993 represent relatively late time points in infection (ranging from 4.5 to 8.5 years after seroconversion).

We also created a comparable data set of HIV-1 *env* sequences collected from nine subjects in the University of Washington PIC (21, 35). These sequences were all sampled within 6 months of seroconversion, all in 1993; these early time point sequences are thus contemporaneous with the MACS late time point sequences sampled in 1993. The PIC sequences are available in the GenBank database under accession numbers AF418311 to AF418547 and AF418694 to AF418920. It should be noted that, whereas the two data sets (MACS and PIC) were collected at different geographic sites with the United States, there is no evidence that sites within the United States differ in terms of HIV genetic diversity (3). More importantly, no epidemiologic linkage was found in either cohort using the Slatkin-Maddison test (39), as implemented in MacClade version 4.06 (26). We also calculated bootstrap support (500 replicates) for the MACS 1993, PIC 1993, and Centers for Disease Control and Prevention combined data set (4) and found no sequence pairs in MACS or PIC with greater than 50% support, thus ensuring the independence of these data sets (data available at http://ubik.mullins.microbiol.washington.edu/HIV/Herbeck2006/). We aligned all MACS and PIC sequences to each other and then parsed replicate subsets from this alignment for analysis.

Although the PIC sequences are not epidemiologically linked and are, therefore, a valid comparison to MACS 1993 sequences for HIV-1 subtype B diversity and divergence estimates, we included another interhost set of primary infection sequences for corroboration. Ataman-Onal et al. (5) analyzed a cohort of 11 subjects in Lyon, France, with symptomatic primary infection of HIV-1 subtype B. We used previously published *env* sequences from eight of these subjects sampled in 1993 ($n = 5$), 1994 ($n = 2$), and 1995 ($n = 1$) and aligned these sequences to our PIC and MACS *env* data set. We found no epidemiologic linkage in this cohort using the tests described above. Using the Lyon primary infection sequences, we created a single interhost data set for comparison to our 10 PIC 1993 interhost data sets.

Diversity and divergence measurement. Population dynamics can be described using summary statistics such as the mean pairwise diversity, π, and the neutral parameter θ (defined as $2N_e\mu$ in a haploid population). In θ the effective population size (N_e) and the substitution rate (μ) are conflated, yet estimating either parameter (or estimating θ values where μ is not thought to vary) can provide useful epidemiological information. For each replicate interhost data set (all nucleotides, synonymous and nonsynonymous) we used SITES (43) to estimate π (uncorrected) and θ, calculated using both π (θ_π) and the number of segregating sites (θ_W). We then used Modeltest with the Akaike information criterion to identify appropriate substitution models for each data set. These substitution models were used to estimate maximum likelihood-corrected mean pairwise divergences in PAUP* version 4.0b10 (41); we also estimated HKY85 corrected distances. Because of the relatively small data sets (nine sequences in each replicate) and the potential for diversity estimates to be skewed by outlying subjects, we performed a jackknife procedure on all data sets by removing single sequences from each replicate and reestimating diversity values. We found no significant differences in π or θ between complete and jackknifed data sets.

We used PAUP* for maximum likelihood genealogy reconstruction for the MACS and PIC 1993 data sets, using the specific substitution models previously chosen in Modeltest. We produced two trees for each replicate, with and without a molecular clock enforced. From the genealogy without a clock enforced, we recorded the lengths of all external branches.

Genealogies and the coalescent approach. Coalescent methods focus on the rate at which lineages in a population coalesce backwards in time, making it possible to infer past population dynamics from a genealogy (17). This rate of lineage coalescence is related to evolutionary and demographic processes such as changes in effective population size, selection, migration, and recombination. These processes are reflected in the distribution of coalescent events in a genealogy and the distribution of substitutions along lineages (38). The coalescent is widely used to describe HIV population dynamics (11, 33, 46). Classic skyline plots are graphical depictions of coalescent analyses (33) that depict the relationship between time, measured in substitutions per site per year, and the coalescent estimate of $N_e\mu$ at time t. Estimates of N_e in coalescent analyses of HIV interhost data sets are intended to represent the effective number of infected individuals (11). For the MACS and PIC 1993 genealogies with the molecular clock enforced, we produced classic skyline plots with the Genie software package (32).

Transmission pairs. The study of donor-recipient transmission pairs can provide useful insights into the effects of transmission and primary infection on the evolution of HIV. To examine this, we utilized the HIV-1 V1 to V5 envelope region gene sequences from seven subtype C heterosexual pairs from Zambia (8). Donors had been infected for times ranging from 0.3 to 4.0 years before sampling. Recipients were sampled within 3 to 4 months of their last seronegative visit. This cohort has been described more fully elsewhere (8).

We created separate donor and recipient sequence data sets for each transmission pair, each with 12 subtype C reference sequences, two subtype A outgroups, and five sequences chosen randomly from all sequences of each subject. We reconstructed genealogies using PAUP* by both maximum likelihood and neighbor-joining methods with the HKY85 distance correction, employing Model-

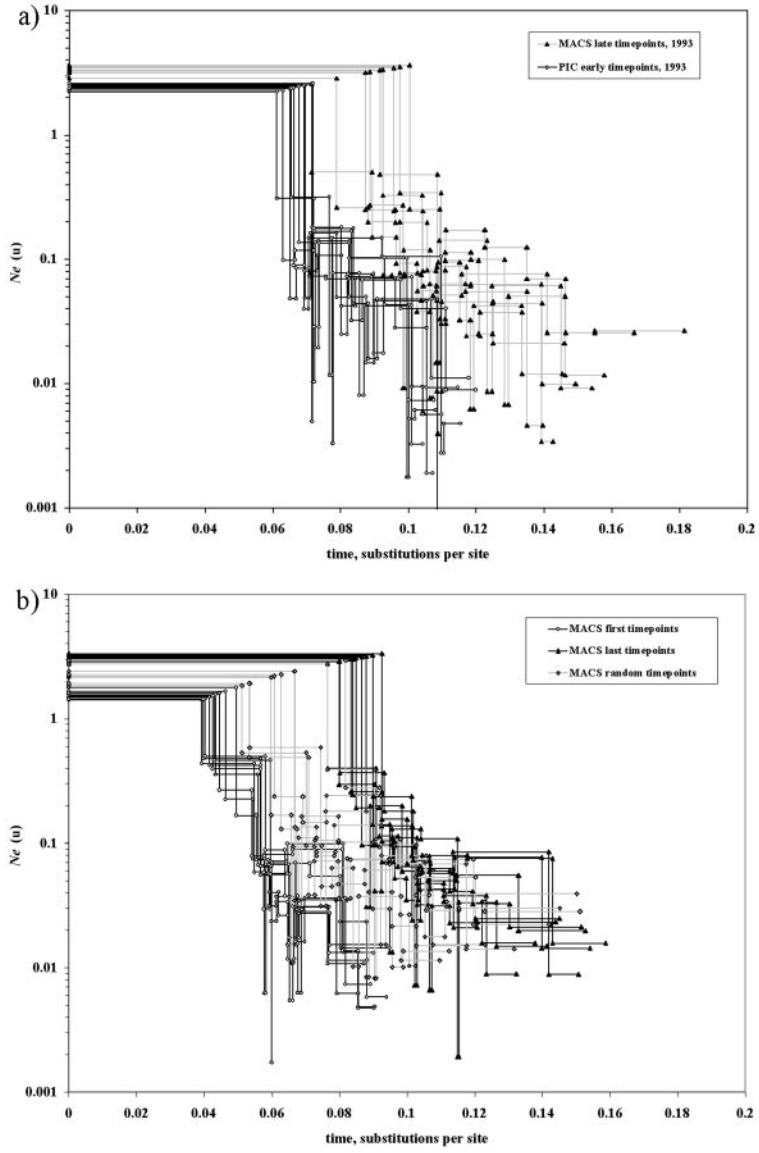

FIG. 1. Classic skyline plots inferred from comparisons of sequences between different subjects (interhost comparisons): MACS 1993 and PIC 1993 replicate genealogies (a) MACS first, last, and random time point replicate genealogies (b). MACS 1993 sequences are derived from epidemiologically unrelated subjects in later stages of disease progression; PIC 1993 sequences are derived from other, epidemiologically unrelated subjects early in infection, within 6 months of seroconversion.

test and the Akaike information criterion to choose appropriate likelihood substitution models for each individual data set. We estimated divergence in donor and recipient viral lineages by summing branch lengths from lineage tips to the node connecting the subtype C lineages to the subtype A outgroups, using the program TreeEdit, version 1.0a10.

Intrahost evolution. We estimated intrahost divergence of HIV-1 *env* C2 to V5 sequences over the course of infection for each of the nine MACS patients in two separate ways. First, we calculated mean pairwise distances between viral sequences at each time point and the consensus sequence from the initially infecting population (first virus-positive time point). Second, we calculated mean distance between viral sequences at each time point and the deduced MRCA sequence (10, 31) for a given patient. This ancestral sequence was derived using PAUP* (40) from the basal node on a tree of patient sequences from all time points, using an outgroup composed of a single first time point sequence (the sequence closest to the Los Alamos National Laboratory HIV Database subtype B consensus [19]) from each of the other patients. Sequence data and patient numbers have been previously described (36).

RESULTS

Comparison of interhost data sets representing early or late time points of infection. As discussed above, comparative interhost analyses assume that the accumulation of diversity in HIV is constant within and across hosts. Thus, HIV sequence data sets representing different time points of infection are expected to have equivalent measures of diversity (the accumulation of diversity should not be affected either by transmission or primary infection). All of these data sets are derived from patients with no evident epidemiological or phylogenetic linkage (data not shown). However, the PIC 1993 primary infection time point data set contains significantly less diversity than the MACS late time point data set

from 1993, with uncorrected π values of 0.112 and 0.137, respectively (Table 1) ($P < 0.001$, two-tailed Mann-Whitney U test). The Lyon primary infection time point data set, including 1993, 1994, and 1995 sequences, also contains less diversity than the MACS late time point data set, with an uncorrected π of 0.119 (Table 1). Thus, the duration of infection has a clear effect on observed *env* interhost diversity. In all data set replicates, the uncorrected π estimates are less than estimates corrected with various substitution models and, thus, can be considered conservative estimates that ignore multiple substitutions and rate heterogeneity across sites; maximum likelihood-corrected π estimates show an even greater discrepancy between primary infection and late infection time points: PIC, $\pi = 0.157$; MACS 1993, $\pi = 0.210$ (Table 1) ($P < 0.001$, Mann-Whitney U test); and Lyon, $\pi = 0.162$. Estimates of the neutral parameter θ are also less in primary infection than in later time point data sets: PIC, $\theta_W = 81.87$; MACS, $\theta_W = 96.51$ (Table 1) ($P < 0.001$, two-tailed Mann-Whitney U test); and Lyon, $\theta_W = 87.55$.

Although the Lyon data set contains eight sequences while the PIC and MACS 1993 data sets contain nine, PIC and MACS data sets jackknifed down to eight sequences have no significant differences in diversity compared to full data sets. Diversity (in π and θ) is slightly higher in Lyon than in PIC (Table 1), likely due to later sampling of three of eight subjects in Lyon (two in 1994 and one in 1995), adding diversity as the epidemic continually expands (14).

To examine possible effects of intrahost divergence on interhost genealogic patterns, we reconstructed evolutionary relationships for each replicate data set. We used these genealogies to infer past and present HIV population dynamics using coalescent methods (11, 17). Classic skyline plots for all of the data sets we examined reflect in their shape an exponential growth rate of the epidemic, consistent with previous descriptions of HIV-1 subtype B. However, estimates of the effective number of infected individuals (N_e) and of the timing of the subtype B epidemic differ between MACS 1993 and PIC 1993 data sets (Fig. 1). Mean $N_e\mu$ estimates are greater for MACS than for PIC, at 3.23 (variance, 0.10) and 2.42 (variance, 0.02), respectively ($P < 0.001$, two-tailed Mann-Whitney U test). Mean MRCA estimates are earlier for MACS than for PIC, at 0.15 (variance, 2.0×10^{-4}) and 0.11 (variance, 2.5×10^{-5}) substitutions per site, respectively ($P < 0.001$).

Given the disparities in coalescent estimates of HIV population dynamics between MACS and PIC 1993 interhost data sets, and the coalescent relationship between genetic diversity, population size, and genealogy, it is reasonable to examine the shape of these particular genealogies. We compared the mean external branch lengths of MACS and PIC 1993 trees; branch lengths are significantly longer in MACS 1993 than PIC 1993 data sets ($P < 0.001$, two-tailed Mann-Whitney U test). These differences alter the probabilities of lineage coalescence and therefore alter the estimates of N_e and MRCA. The external branch length comparisons were based on phylogenies comprised only of the nine subjects in each replicate data set. However, this pattern of branch length differences holds in phylogenies that also include 100 CDC sequences (4).

Divergence in transmission pairs. We measured divergence (mean of summed branch lengths) from an interhost MRCA, in this case the node rooting subtype C to the subtype A outgroups, in donor and recipient sequences. Divergence is significantly less in recipient than in donor sequences ($P < 0.05$, two-tailed Mann-Whitney U test) in five of seven transmission pairs (Table 2). This corroborates the finding of decreased interhost divergence in early time point *env* relative to *env* sampled at time points after primary infection. Variation in the extent, or existence, of this pattern in these seven transmission pairs may be due to variation in the duration of infection in donor subjects prior to transmission, variation in infection and sampling in recipients, or to host factors such as HLA type affecting selection within the recipients.

Intrahost evolution. We next examined intrahost viral evolution in the longitudinal MACS cohort. As shown previously (36), the virus population diverges from the population found at the earliest time point at a consistent rate of about 1% per year throughout the first years of infection (Fig. 2, gray lines). However, in most cases, there is a delay in the development of this divergence over the first 2 to 3 years of infection when the virus population is compared to the deduced MRCA (Fig. 2, black lines). This suggests the existence of competing selective forces during the first 3 years of infection: one of recovery of ancestral genetic features shortly after transmission to a new host and one of diversifying selection as the viral population adapts to new local niches within the host. Even in the few cases in which there is no noted departure in the rate of divergence from the initial strain versus the MRCA, a clear evolutionary trend toward the MRCA early in infection was noted in phylogenetic reconstructions of patient viral sequences taken early in infection (Fig. 3). This supports the hypothesis of *env* evolution toward ancestral states after transmission or during primary infection.

DISCUSSION

The impact of intrahost evolutionary processes on interhost divergence estimates has implications for phylogenetic approaches to rational vaccine design. We found that HIV interhost genetic diversity and divergence are significantly less during early infection than at later times following seroconversion, with a potential difference of at least 2.5% (for uncorrected π; HKY and maximum likelihood-corrected π show 3.2% and 5.3% differences, respectively). Thus, if we consider HIV-1 B phylogenies to be nearly star-like, in which most circulating strains radiate from some central point within the phylogeny (4, 34), we expect a portion of external substitutions in each lineage to be substitutions accumulated after infection. Furthermore, if HIV transmission events occur predominantly during early infection, given the high viremia in acute infection (6) and the positive correlation between transmission rates and viremia (12), these external substitutions may be unimportant to transmission and thus irrelevant in immunogen design. Therefore, for the purposes of including *env* in rational immunogen design, it may be advantageous to incorporate only early time point sequences. It is currently unclear whether other HIV genes undergo similar evolutionary processes in transmission and primary infection.

Coalescent analyses of HIV interhost sequences assume that the accumulation of substitutions within hosts will not affect estimates of N_e or the timing of the MRCA, because the mean

TABLE 2. Transmission pair divergence analysis[a]

Method and patient pair no.	Divergence (mean branch length)	
	Donor	Recipient
NJ with ML-estimated distance matrix		
55	0.333	0.352
53	0.358	0.316*
13	0.352	0.335*
83	0.312	0.327
106	0.344	0.312*
109	0.372	0.365*
135	0.366	0.345*
NJ with HKY85 distance matrix		
55	0.194	0.197
53	0.221	0.194*
13	0.207	0.191*
83	0.197	0.199
106	0.198	0.182*
109	0.224	0.218*
135	0.205	0.200*

[a] For each patient pair, we reconstructed phylogenetic trees including 12 subtype C reference sequences, two subtype A outgroups, and five sequences (either all donor or all recipient, subtype C) chosen randomly from each subject. Divergence was measured as the mean branch lengths from the subtype C interhost MRCA (where C rooted to the A outgroups) to lineage tips. *, *P* values all <0.05, one-tailed Mann-Whitney U test for donor-recipient divergence distributions. NJ, neighbor-joining; ML, maximum likelihood.

sequence diversity and estimates of θ will remain the same. We have shown this not to be the case. Due to the coalescent relationship between genetic diversity, population size, and genealogy, such effects may cause erroneous descriptions of HIV epidemiologic dynamics and history. Our data sets are relatively small, and extrapolation to interhost coalescent analyses with larger numbers of sequences should be made with caution. However, the variance seen among the random time point data sets suggests that common HIV interhost data sets may result in underestimates of the variance associated with population genetic parameters.

In conclusion, we have demonstrated that time since infection can be a significant factor affecting the diversity and divergence observed in HIV interhost sequence data sets. Currently, interhost sequence data sets are not typically assembled with regard to the stage of disease progression within each sampled subject. This result has consequences for population genetic studies of HIV molecular epidemiology that rely on the assumption of continual interhost divergence from the interhost MRCA and certainly may affect the choice of sequences for phylogenetic reconstruction used in rational vaccine design.

We have also shown that the source of the discontinuity of evolutionary divergence in *env* is likely the evolution toward ancestral states that takes place upon transmission to a new

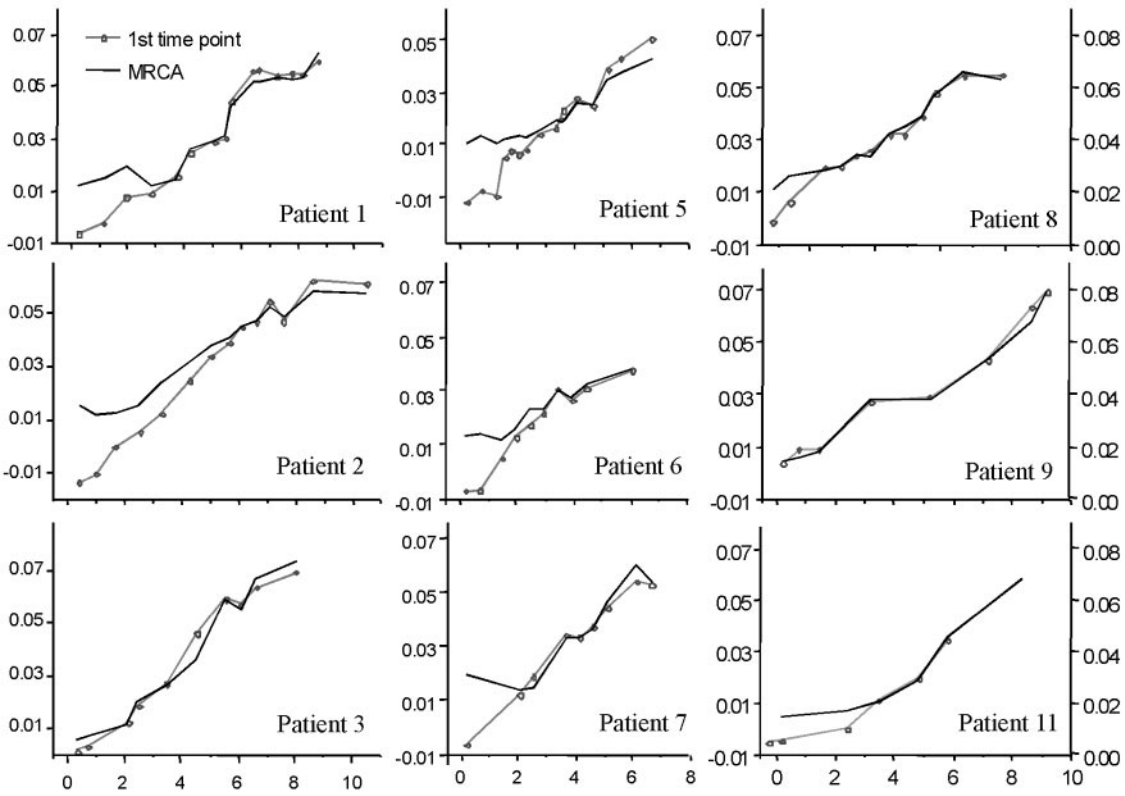

FIG. 2. Divergence of HIV-1 *env* C2 to V5 sequences over time within individual subjects (intrahost comparisons). Mean pairwise comparisons of viral sequence distances to the initially infecting population (first virus-positive time point) are shown by the points connected by a gray line, with values given on the common right-side *y* axis. The black lines and left-side *y* axes show the mean distance between viral sequences at each time point and the deduced MRCA sequence (10, 31) for each patient. This ancestral sequence was derived in the program PAUP* (40) from the basal node on a tree of patient sequences from all time points, defined by an outgroup composed of a single first time point sequence (the sequence closest to the Los Alamos National Laboratory HIV Database subtype B consensus [19]) from each of the other patients. Sequence data and patient numbers were described previously (36). The *y* axes on the left side were adjusted up or down to provide maximal alignment with plots using the right-side *y* axis.

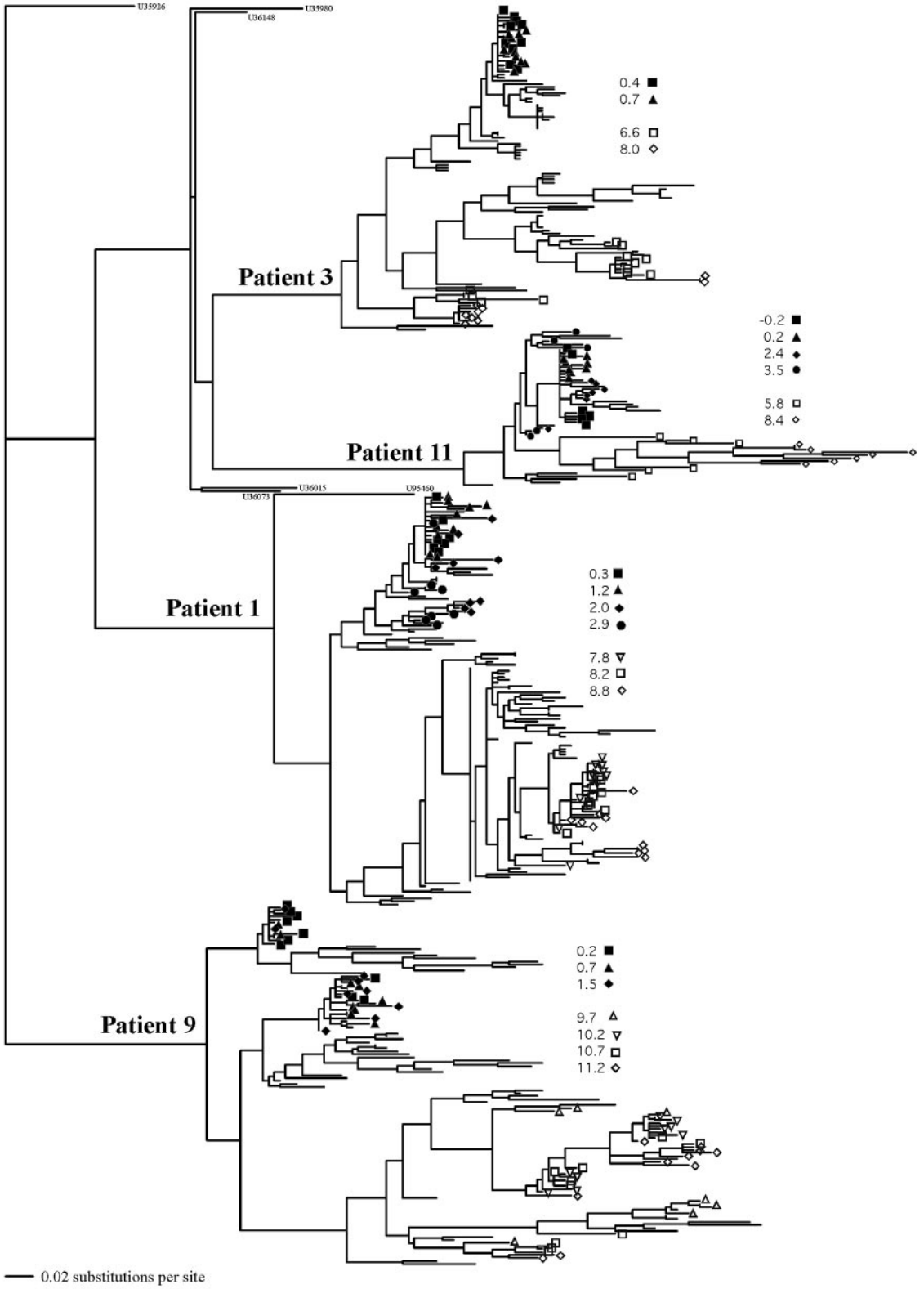

FIG. 3. Phylogenetic analysis of viral sequences in patients 1, 3, 9, and 11 from the study of Shankarappa et al. (36). Viral sequences taken from the first two to four biannual visits following infection are shown at tree tips marked with filled symbols; those taken from the final two to four follow-up visits are shown with open symbols. The legend at the right of each patient's clade indicates the time, in years, the specimen was taken following the estimated time of seroconversion. GenBank locus names are given to subtype B HIV-1 sequences used as outgroups.

host. Recent data reveal some of the selective mechanisms that are likely to account for these observations. One the one hand, a consistently paced forward evolution (36) most visibly results from the effect of mutations leading to immunologic escape from CTL responses (2, 23, 24). Opposing this trend is the reversion of CTL escape mutations upon transmission to a new host (2, 9, 23, 24). Escape mutants from restricted epitopes that arose within donor individuals having a given HLA type appear to revert to a susceptible epitope sequence in recipients with different HLA alleles that are unable to present these viral peptides to the immune system. Adaptation and concentration of escape mutants in HIV circulating in populations with common HLA alleles have also been postulated (16, 28), which are likely to mute the effect we have shown here. Furthermore, changes in glycosylation patterns associated with escape from host antibody responses also occur early in infection (8). It is also known that drug resistance mutations in HIV-1 can impart decreased relative replication capacity in cell cultures lacking antiretroviral drugs (13) and that resistant forms are selectively lost in vivo following removal of the drug (6a, 8a). Thus, recovery of ancestral states may reflect restoration of fitness lost as a result of immunological escapes in the previous host, as well as replication-advantageous mutations within an HIV immunologically naïve host.

The convergence of viral sequences and specific mechanisms of reversion suggest, circumstantially, that there are strong sequence constraints that are important to viral reproduction across patients. This provides impetus to the development of vaccine immunogens that favor inclusion of viral sequences from early in infection and that embody ancestral or consensus features of viruses circulating in a given population (for a review, see reference 29).

ACKNOWLEDGMENTS

The authors thank Morgane Rolland and two reviewers for helpful comments on the manuscript.

This work was supported by grants from the U.S. Public Health Service, including a training fellowship (NIH T32AI07140) to J.H.

REFERENCES

1. **Abebe, A., V. V. Lukashov, G. Pollakis, A. Kliphuis, A. L. Fontanet, J. Goudsmit, and T. F. de Wit.** 2001. Timing of the HIV-1 subtype C epidemic in Ethiopia based on early virus strains and subsequent virus diversification. AIDS **15:**1555–1561.
2. **Allen, T. M., M. Altfeld, X. G. Yu, K. M. O'Sullivan, M. Lichterfeld, S. Le Gall, M. John, B. R. Mothe, P. K. Lee, E. T. Kalife, D. E. Cohen, K. A. Freedberg, D. A. Strick, M. N. Johnston, A. Sette, E. S. Rosenberg, S. A. Mallal, P. J. Goulder, C. Brander, and B. D. Walker.** 2004. Selection, transmission, and reversion of an antigen-processing cytotoxic T-lymphocyte escape mutation in human immunodeficiency virus type 1 infection. J. Virol. **78:**7069–7078.
3. **Anderson, J. P., G. H. Learn, A. G. Rodrigo, X. He, Y. Wang, H. Weinstock, M. L. Kalish, K. E. Robbins, L. Hood, and J. I. Mullins.** 2003. Predicting demographic group structures based on DNA sequence data. Mol. Biol. Evol. **20:**1168–1180.
4. **Anderson, J. P., A. G. Rodrigo, G. H. Learn, Y. Wang, H. Weinstock, M. L. Kalish, K. E. Robbins, L. Hood, and J. I. Mullins.** 2001. Substitution model of sequence evolution for the human immunodeficiency virus type 1 subtype B gp120 gene over the C2–V5 region. J. Mol. Evol. **53:**55–62.
5. **Ataman-Onal, Y., C. Coiffier, A. Giraud, A. Babic-Erceg, F. Biron, and B. Verrier.** 1999. Comparison of complete *env* gene sequences from individuals with symptomatic primary HIV type 1 infection. AIDS Res. Hum. Retrovir. **15:**1035–1039.
6. **Daar, E. S., T. Moudgil, R. D. Meyer, and D. D. Ho.** 1991. Transient high levels of viremia in patients with primary HIV-1 infection. N. Engl. J. Med. **324:**961–964.

6a. **Deeks, S. G.** 2001. Durable HIV treatment benefit despite low-level viremia: reassessing definitions of success or failure. JAMA **286:**224–226.

7. **Delwart, E. L., H. W. Sheppard, B. D. Walker, J. Goudsmit, and J. I. Mullins.** 1994. Human immunodeficiency virus type 1 evolution in vivo tracked by DNA heteroduplex mobility assays. J. Virol. **68:**6672–6683.
8. **Derdeyn, C. A., J. M. Decker, F. Bibollet-Ruche, J. L. Mokili, M. Muldoon, S. A. Denham, M. L. Heil, F. Kasolo, R. Musonda, B. H. Hahn, G. M. Shaw, B. T. Korber, S. Allen, and E. Hunter.** 2004. Envelope-constrained neutralization-sensitive HIV-1 after heterosexual transmission. Science **303:**2019–2022.

8a. **Frenkel, L. M., and J. I. Mullins.** 2001. Should patients with drug-resistant HIV-1 continue to receive antiretroviral therapy? N. Engl. J. Med. **344:**520–522.

9. **Friedrich, T. C., E. J. Dodds, L. J. Yant, L. Vojnov, R. Rudersdorf, C. Cullen, D. T. Evans, R. C. Desrosiers, B. R. Mothe, J. Sidney, A. Sette, K. Kunstman, S. Wolinsky, M. Piatak, J. Lifson, A. L. Hughes, N. Wilson, D. H. O'Connor, and D. I. Watkins.** 2004. Reversion of CTL escape-variant immunodeficiency viruses in vivo. Nat. Med. **10:**275–281.
10. **Gaschen, B., J. Taylor, K. Yusim, B. Foley, F. Gao, D. Lang, V. Novitsky, B. Haynes, B. H. Hahn, T. Bhattacharya, and B. Korber.** 2002. Diversity considerations in HIV-1 vaccine selection. Science **296:**2354–2360.
11. **Grassly, N. C., P. H. Harvey, and E. C. Holmes.** 1999. Population dynamics of HIV-1 inferred from gene sequences. Genetics **151:**427–438.
12. **Gray, R. H., M. J. Wawer, R. Brookmeyer, N. K. Sewankambo, D. Serwadda, F. Wabwire-Mangen, T. Lutalo, X. Li, T. van Cott, and T. C. Quinn.** 2001. Probability of HIV-1 transmission per coital act in monogamous, heterosexual, HIV-1-discordant couples in Rakai, Uganda. Lancet **357:**1149–1153.
13. **Harrigan, P. R., S. Bloor, and B. A. Larder.** 1998. Relative replicative fitness of zidovudine-resistant human immunodeficiency virus type 1 isolates in vitro. J. Virol. **72:**3773–3778.
14. **Hu, D. J., T. J. Dondero, M. A. Rayfield, J. R. George, G. Schochetman, H. W. Jaffe, C. C. Luo, M. L. Kalish, B. G. Weniger, C. P. Pau, C. A. Schable, and J. W. Curran.** 1996. The emerging genetic diversity of HIV. The importance of global surveillance for diagnostics, research, and prevention. JAMA **275:**210–216.
15. **Kaslow, R. A., D. G. Ostrow, R. Detels, J. P. Phair, B. F. Polk, and C. R. Rinaldo, Jr.** 1987. The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants. Am. J. Epidemiol. **126:**310–318.
16. **Kiepiela, P., A. J. Leslie, I. Honeyborne, D. Ramduth, C. Thobakgale, S. Chetty, P. Rathnavalu, C. Moore, K. J. Pfafferott, L. Hilton, P. Zimbwa, S. Moore, T. Allen, C. Brander, M. M. Addo, M. Altfeld, I. James, S. Mallal, M. Bunce, L. D. Barber, J. Szinger, C. Day, P. Klenerman, J. Mullins, B. Korber, H. M. Coovadia, B. D. Walker, and P. J. Goulder.** 2004. Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. Nature **432:**769–775.
17. **Kingman, J. F. C.** 1982. On the genealogy of large populations. J. Appl. Probability **19A:**27–43.
18. **Korber, B., M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B. H. Hahn, S. Wolinsky, and T. Bhattacharya.** 2000. Timing the ancestor of the HIV-1 pandemic strains. Science **288:**1789–1796.
19. **Kuiken, C., B. Foley, E. Freed, B. Hahn, B. Korber, P. Marx, F. McCutchan, J. Mellors, and S. Wolinksy (ed.).** 2002. HIV sequence compendium 2002. Publication LA-UR 03–3564. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, N. Mex.
20. **Learn, G., and J. I. Mullins.** 2000. Presented at the Seventh International Discussion Meeting on HIV Dynamics and Evolution, Seattle, Wash., 28 to 30 April 2000.
21. **Learn, G. H., D. Muthui, S. J. Brodie, T. Zhu, K. Diem, J. I. Mullins, and L. Corey.** 2002. Virus population homogenization following acute human immunodeficiency virus type 1 infection. J. Virol. **76:**11953–11959.
22. **Leitner, T., B. Foley, B. Hahn, P. Marx, F. McCutchan, J. Mellors, S. Wolinsky, and B. Korber (ed.).** 2004. HIV sequence compendium 2003. Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, N. Mex.
23. **Leslie, A. J., K. J. Pfafferott, P. Chetty, R. Draenert, M. M. Addo, M. Feeney, Y. Tang, E. C. Holmes, T. Allen, J. G. Prado, M. Altfeld, C. Brander, C. Dixon, D. Ramduth, P. Jeena, S. A. Thomas, A. S. John, T. A. Roach, B. Kupfer, G. Luzzi, A. Edwards, G. Taylor, H. Lyall, G. Tudor-Williams, V. Novelli, J. Martinez-Picado, P. Kiepiela, B. D. Walker, and P. J. Goulder.** 2004. HIV evolution: CTL escape mutation and reversion after transmission. Nat. Med. **10:**282–289.
24. **Liu, Y., J. McNevin, J. Cao, H. Zhao, I. Genowati, K. Wong, S. McLaughlin, D. C. Nickle, D. Shriner, M. J. McElrath, and J. I. Mullins.** Unpublished data.
25. **Long, E. M., H. L. Martin, Jr., J. K. Kreiss, S. M. Rainwater, L. Lavreys, D. J. Jackson, J. Rakwar, K. Mandaliya, and J. Overbaugh.** 2000. Gender differences in HIV-1 diversity at time of infection. Nat. Med. **6:**71–75.
26. **Maddison, W. P., and D. R. Maddison.** 2001. MacClade: analysis of phylogeny and character evolution, version 4. Sinauer Associates, Inc., Sunderland, Mass.
27. **McMichael, A., and S. Rowland-Jones.** 2001. Cellular immune responses to HIV. Nature **410:**980–987.
28. **Moore, C. B., M. John, I. R. James, F. T. Christiansen, C. S. Witt, and S. A.**

Mallal. 2002. Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. Science **296:**1439–1443.
29. **Mullins, J. I., D. C. Nickle, L. Heath, A. G. Rodrigo, and G. H. Learn.** 2004. Immunogen sequence: the fourth tier of AIDS vaccine design. Expert Rev. Vaccines **3**(Suppl. 1)**:**S151–S159.
30. **Nabel, G., W. Makgoba, and J. Esparza.** 2002. HIV-1 diversity and vaccine development. Science **296:**2335.
31. **Nickle, D. C., M. A. Jensen, G. S. Gottlieb, D. Shriner, G. H. Learn, A. G. Rodrigo, and J. I. Mullins.** 2003. Consensus and ancestral state HIV vaccines. Science **299:**1515–1518.
32. **Pybus, O. G., and A. Rambaut.** 2002. GENIE: estimating demographic history from molecular phylogenies. Bioinformatics **18:**1404–1405.
33. **Pybus, O. G., A. Rambaut, and P. H. Harvey.** 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics **155:**1429–1437.
34. **Robbins, K. E., P. Lemey, O. G. Pybus, H. W. Jaffe, A. S. Youngpairoj, T. M. Brown, M. Salemi, A. M. Vandamme, and M. L. Kalish.** 2003. U.S. human immunodeficiency virus type 1 epidemic: date of origin, population history, and characterization of early strains. J. Virol. **77:**6359–6366.
35. **Schacker, T., A. C. Collier, J. Hughes, T. Shea, and L. Corey.** 1996. Clinical and epidemiologic features of primary HIV infection. Ann. Intern. Med. **125:**257–264.
36. **Shankarappa, R., J. B. Margolick, S. J. Gange, A. G. Rodrigo, D. Upchurch, H. Farzadegan, P. Gupta, C. R. Rinaldo, G. H. Learn, X. He, X. L. Huang, and J. I. Mullins.** 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J. Virol. **73:**10489–10502.
37. **Shpaer, E. G., E. L. Delwart, C. L. Kuiken, J. Goudsmit, M. H. Bachmann, and J. I. Mullins.** 1994. Conserved V3 loop sequences and transmission of human immunodeficiency virus type 1. AIDS Res. Hum. Retrovir. **10:**1679–1684.
38. **Slatkin, M., and R. R. Hudson.** 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:**555–562.
39. **Slatkin, M., and W. P. Maddison.** 1989. A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics **123:**603–613.
40. **Swofford, D. L.** 1999. PAUP* 4.0: phylogenetic analysis using parsimony (*and other methods), version 4.0b2a. Sinauer Associates, Inc., Sunderland, Mass.
41. **Swofford, D. L.** 2002. PAUP* 4.0: phylogenetic analysis using parsimony (*and other methods), version 4.0b10. Sinauer Associates, Inc., Sunderland, Mass.
42. **van 't Wout, A. B., N. A. Kootstra, G. A. Mulder Kampinga, N. Albrecht van Lent, H. J. Scherpbier, J. Veenstra, K. Boer, R. A. Coutinho, F. Miedema, and H. Schuitemaker.** 1994. Macrophage-tropic variants initiate human immunodeficiency virus type 1 infection after sexual, parenteral, and vertical transmission. J. Clin. Investig. **94:**2060–2067.
43. **Wakeley, J., and J. Hey.** 1997. Estimating ancestral population parameters. Genetics **145:**847–855.
44. **Wolfs, T. F., G. Zwart, M. Bakker, and J. Goudsmit.** 1992. HIV-1 genomic RNA diversification following sexual and parenteral virus transmission. Virology **189:**103–110.
45. **Wolinsky, S. M., C. M. Wike, B. T. M. Korber, C. Hutto, W. P. Parks, L. L. Rosenblum, K. J. Kunstman, M. R. Furtado, and J. L. Muñoz.** 1992. Selective transmission of HIV-1 variants from mother to infants. Science **255:** 1134–1137.
46. **Yusim, K., M. Peeters, O. G. Pybus, T. Bhattacharya, E. Delaporte, C. Mulanga, M. Muldoon, J. Theiler, and B. Korber.** 2001. Using human immunodeficiency virus type 1 sequences to infer historical features of the acquired immune deficiency syndrome epidemic and human immunodeficiency virus evolution. Philos. Trans. R. Soc. Lond. B **356:**855–866.
47. **Zhang, L. Q., P. MacKenzie, A. Cleland, E. C. Holmes, A. J. Leigh Brown, and P. Simmonds.** 1993. Selection for specific sequences in the external envelope protein of HIV-1 upon primary infection. J. Virol. **67:**3345–3356.