LA-UR- 09-00147

Approved for public release;
distribution is unlimited.

Title:	Identification of Full-length Transmitted/Founder Viruses and Their Progeny in Primary HIV-1 Infection.
Author(s):	B. Korber, Z# 108817, T-6/T-Division E. Giorgi, Z# 221633, T-6/T-Division P. Hraber, Z# 194500, T-6/T-Division T. Bhattacharya, Z# 111197, T-2/T-Division
Intended for:	Journal: Journal of Experimental Medicine

Form 836 (7/06)

DRAFT

Identification of Full-length Transmitted/Founder Viruses and Their Progeny in Primary HIV-1 Infection

Jesus F. Salazar-Gonzalez[*†], Maria G. Salazar[*†], Brandon F. Keele[*†], Gerald H. Learn[†], Elena E. Giorgi[‡§], Hui Li[†], Julie M. Decker[†], Shuyi Wang[†], Joshua Baalwa[†], Matthias H. Kraus[†], Nicholas F. Parrish[†], Katharina S. Shaw[†], M. Brad Guffey[†], Katharine J. Bar[†], Katie L. Davis[†], Christina Ochsenbauer-Jambor[†], John C. Kappes[†], Michael S. Saag[†], Myron S. Cohen[+], Cynthia A. Derdeyn[#], Susan Allen[#], Eric Hunter[#], Martin Markowitz[@], Peter Hraber[‡], Alan S. Perelson[‡], Tanmoy Bhattacharya[‡¶], Barton F. Haynes[¶¶], Bette T. Korber[‡¶], Beatrice H. Hahn[†], George M. Shaw[†^]

[*]Contributed equally
[†]University of Alabama at Birmingham, Birmingham, Alabama 35223
[‡]Los Alamos National Laboratory, Los Alamos NM 87545
[§]University of Massachusetts, Amherst, MA 01002
[+]The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599
[#]Emory University, Atlanta, GA 30329
[@]Aaron Diamond AIDS Research Center, Rockefeller University, NY, NY 10016
[¶]Santa Fe Institute, Santa Fe, NM 87501
[¶¶]Duke University Medical Center, Durham, NC 27710

[^]To whom correspondence should be addressed. E-mail: gshaw@uab.edu

ABSTRACT: Identification of transmitted/founder virus genomes and their progeny by is a novel strategy for probing the molecular basis of HIV-1 transmission and for evaluating the genetic imprint of viral and host factors that act to constrain or facilitate virus replication. Here, we show in a cohort of twelve acutely infected subjects (9 clade B; 3 clade C), that complete genomic sequences of transmitted/founder viruses could be inferred using single genome amplification of plasma viral RNA, direct amplicon sequencing, and a model of random virus evolution. This allowed for the precise identification, chemical synthesis, molecular cloning, and biological analysis of those viruses actually responsible for productive clinical infection and for a comprehensive mapping of sequential viral genomes and proteomes for mutations that are necessary or incidental to the establishment of HIV-1 persistence. Transmitted/founder viruses were CD4 and CCR5 tropic, replicated preferentially in activated primary T-lymphocytes but not monocyte-derived macrophages, and were effectively shielded from most heterologous or broadly neutralizing antibodies. By 3 months of infection, the evolving viral quasispecies in three subjects showed mutational fixation at only 2-5 discreet genomic loci. By 6-12 months, mutational fixation was evident at 18-27 genomic loci. Some, but not all, of these mutations were attributable to virus escape from cytotoxic T-lymphocytes or neutralizing antibodies, suggesting that other viral or host factors may influence early HIV-1 fitness.

Introduction

Transmission of HIV-1 generally results from virus exposure at mucosal surfaces followed by virus replication in submucosal and locoregional lymphoid tissues, and eventually, overt systemic infection (Shattock and Moore 2003; Pope and Haase 2003; Mehandru 2004; Brenchley 2004; Veazey and Lackner 2004; Haase 2005; Margolis and Shattock 2006). Because of the inaccessibility of early sites of replication, the molecular details of HIV-1 transmission and early virus evolution remain largely unknown. Analysis of individual HIV-1 genes from virus(es) that are responsible for productive clinical infection in humans can be instrumental in elucidating viral properties and biological events underlying the transmission process (Keele 2008; Salazar 2008; Abrahams 2009; Haaland 2009; Coffin 2009). However, identification and characterization of full-length genomes from such viruses could go much further in elucidating on a genome-wide basis those properties of transmitted/founder viruses, and their progeny, that are essential for virus transmission and the establishment of viral persistence. This is true for naïve individuals who become infected by HIV-1 and for subjects who are immunized with candidate HIV-1 vaccines but experience breakthrough infection (http://www.hvtn.org/science/1107.html).

Recently, we developed a mathematical model of HIV-1 sequence evolution in acute infection and an experimental strategy based on single genome amplification (SGA) of plasma vRNA followed by direct sequencing of uncloned cDNA that allowed us to infer the nucleotide sequences of full-length envelope (*env*) genes of transmitted/founder viruses in 98 of 102 consecutively studied patients (Keele 2008). The model assumes that a transmitted virus replicates exponentially with a generation

time of 2 days (Markowitz 2003), reproductive ratio (R_0) of 6 (Stafford 2000), and reverse transcriptase error rate of 2.16×10^{-5} (Mansky and Temin 1995); that it diversifies under no selection; and that it exhibits a constant mutation rate across positions and lineages and undergoes no back mutations. These assumptions were based on estimated parameters of virus replication and a timing of virus sampling prior to the development of detectable immune responses (Borrow 1997; Fiebig 2003; Wei 2003; Richman 2003; Jones 2004; Keele 2008). The model predicts a Poisson distribution of mutations and a star-like phylogeny that coalesces to an inferred consensus sequence at or near the time of virus transmission. We obtained direct experimental evidence in support of this model in Indian rhesus macaques mucosally infected by $SIV_{mac}251$ (B.F.K. CROI 2009) and in humans infected by a known sexual partner (J. Anderson CROI 2009): In each case, an *env* sequence in the SIV inoculum stock or in the semen of a chronically infected sexual partner was found to be *identical* to that of the transmitted/founder virus identified in the recipient. We also tested the model by Monte Carlo simulation (Lee 2008) and against an empirical data set of 3,449 SGA-derived complete *env* sequences from 102 acutely infected patients (Keele 2008). The results supported the model and its assumptions. Importantly, the model and the empirical findings allowed us to infer that in most cases of sexual transmission of HIV-1, a *single* virus (or infected cell) is responsible for establishing productive clinical infection, a conclusion now supported by studies in seven additional patient cohorts infected by HIV-1 subtypes A, B, C or A/D (Salazaar 2008; Hunter 2009; Williamson 2009; Hui CROI 2009; Ling CROI 2009; Baalwa CROI 2009; Coffin 2009).

In the present study, we asked if the experimental strategy for identifying transmitted/founder *env* sequences can be applied successfully to full-length HIV-1 vRNA genomes, which are nearly four times longer than *env* genes (9 kb versus 2.6 kb); whether early diversification of full-length HIV-1 genomes conforms to a model of random evolution with a Poisson distribution of mutations and star-like phylogeny; whether inferences regarding the identity and number of transmitted viruses and the estimated time to a most recent common ancestor (MRCA) based on full-length genome analyses correspond to findings based on *env*-only gene analyses; whether inferred transmitted/founder full-length sequences contain intact canonical gene open reading frames and encode replication competent viruses; and whether identification of transmitted/founder full-length genomes and their progeny can provide new insight into the biology of HIV-1 transmission and the kinetics and pathways of virus diversification and adaptation leading to viral persistence.

Results

Study subjects. Plasma specimens from twelve adult subjects (10M/2F) with acute HIV-1 infection were analyzed in this study (Table 1). Nine subjects were infected by HIV-1 subtype B and three by subtype C. At the initial sampling time point, 10 subjects were plasma vRNA+/Ab- (Fiebig stage II; see Fiebig 2003) and two subjects were vRNA+/ELISA+/WB indeterminant (Fiebig stage IV). Three subjects were sampled periodically through as much as 60 weeks of followup. Peak plasma viral loads ranged from 394,649 to 26,700,000 vRNA copies per milliliter. Four subjects admitted to

heterosexual exposure as their only HIV-1 risk factor and eight were men who had sex with men (MSM).

Single genome amplification and sequencing. Between 5 and 18 complete genomic cDNA amplicons (median of 9) were derived by amplification of individual plasma vRNA molecules from each subject (108 amplicons in total; see Table 2). Each of the 108 amplicons was sequenced in its entirety. Amplicons ranged in length from 7698bp to 9068bp. Sequence chromatograms of 63 cDNA amplicons were unambiguous at every position. Sequence chromatograms of 45 amplicons had mixed bases at 1 to 5 positions per sequence (geometric mean ± SD = 1.6 ± 1.1). Mixed bases were confirmed by repetition of sequencing reactions and by sequencing both strands of DNA. Because the proportion of PCR positive wells at endpoint cDNA dilution was <20%, and because mixed bases generally represented only a subset of polymorphisms in any one sequence, we could infer in most cases that mixed bases on chromatograms resulted from *Taq* polymerase error in the initial cycle(s) of PCR amplification and not from PCR amplification from more than one original vRNA/cDNA template. In rare instances where one or more mixed bases represented the only polymorphisms in a sequence, we could not distinguish between *Taq* error and mixed templates in the original amplification reaction. In sum, we could make an unambiguous assignment of nucleotides at each position in the nucleotide sequences of 103 HIV-1 genomes and at all but 9 positions in five others.

HIV-1 diversity. In a maximum likelihood phylogenetic tree, viral sequences from the nine US subjects clustered significantly with prototype B clade viruses, whereas sequences from the three Zambian subjects clustered with prototype C clade viruses (Figure 1). Maximum interstrain diversity among all 108 full-length genomes was >25%, reflecting differences typically observed between different clade B and C viruses. Within individual subjects, maximum virus diversity was far less, ranging from 0.04% in subject SUMA 0874 to 2.46% in subject ZM247F (Table 2). There was no interspersion of sequences between or among subjects. Maximum within-patient viral diversity was distinctly lower in 11 subjects (<0.14% in each) compared with the twelfth subject, ZM247F (2.46%). We postulated that the observed differences in maximum viral diversity observed within individuals might reflect the numbers of viruses responsible for establishing productive infection in these subjects, as shown previously for *env* diversity (Keele 2008). We formally tested this hypothesis by comparing observed viral genome diversities in each subject with estimates, based on model predictions, of the maximum diversity one could expect within 100 days following transmission of a single virus (0.60%; 0.54 – 0.68% C.I.; Keele 2008). Eleven of the 12 subjects had sequences that fell well below the 0.60% threshold, whereas one subject (ZM247F) had sequences that fell far above it (Table 2). We also used the model to estimate in each subject the minimum number of days that would be required to explain the observed within-patient HIV-1 genome diversification from a single most recent common ancestor (MRCA) sequence, again as we had done previously for *env* diversification (Keele 2008). In this analysis, we did not adjust for mutations that are selected against and go unobserved because they result in unfit viruses; as a consequence, the timing estimates based on a

comparison of the observed data to the model tend to be biased toward a low estimate. Eleven subjects with lower viral diversity had minimum estimates for days since a MRCA virus that fell well within model predictions for infection by a single virus (11 to 33 days; 95% C.I. = 7 to 38 days) and within a time frame consistent with each subject's Fiebig stage (Fiebig 2003; Table 2). Conversely, sequence diversity in subject ZM247F corresponded to a minimum estimate for a MRCA of 493 days, far beyond the range of plausibility for recent infection based on this subject's Fiebig stage II, which has an average duration from virus transmission of 22 days (95% C.I. = 16-39 days; Fiebig 2003; Keele 2008). Interestingly, sequences from ZM247F fell into two distinct low diversity phylogenetic lineages that differed from each other by an average of 2.40% (Figure 1). Sequence diversity within lineage 1 ranged from 0.01-0.09%, and within lineage 2 from 0.02-0.08%. Within each lineage, sequences exhibited a star-like phylogeny and a Poisson distribution of mutations. Model estimates for time from a MRCA for lineage 1 was 21 days (95% C.I. = 14-28 days), and for lineage 2 was 21 days (95% C.I. = 14-28 days). Based on this analysis, we concluded that ZM247F most likely had been infected by two viruses at the same time and from the same sexual partner (Figure 1 and Table 2).

Model testing and analysis of HIV-1 evolution. To explore how full-length HIV-1 genomic sequences sampled near peak viremia conform to model predictions, we obtained for each subject the frequency distribution of all intersequence Hamming distances (HDs) (defined as the number of base positions at which two genomes differ) and determined whether it deviated from a Poisson model by using a chi-square goodness

of fit test. Four sequences from 3 subjects exhibited G-to-A hypermutation. Once these were eliminated from the analysis, sequences from 11 of 12 subjects conformed to the Poisson model (Table 2). Sequences from subject ZM247F did not conform to a Poisson model of variation but did so once the two low diversity lineages evident in the phylogenetic tree (Figure 1) were analyzed individually. We next investigated whether or not sequences evolved under a star-phylogeny model (i.e., all evolving sequences are equally likely and all coalesce at the founder) in the expected time frame based on Fiebig stage. Sequences from 6 of 12 subjects, including each lineage in subject ZM247F, exhibited a star phylogeny (Table 2). Among the samples that deviated from a star phylogeny were sequences from two subjects (700010040, 700010077) that exhibited early changes in predicted cytotoxic T-cell (CTL) epitopes corresponding to the HLA types of these subjects; CTL recognition and escape at these positions were subsequently demonstrated experimentally (Goonetilleke 2008). When CTL escape mutations and rare shared polymorphisms arising from early stochastic nucleotide substitutions (Keele 2008; Lee 2008), were excluded from the analysis, sequences from the remaining six subjects conformed to a star phylogeny.

We next analyzed the diversity among *env* genes within full-length genomes compared with a much larger number of *env* genes amplified as *env*-only sequences from the same subjects (Table 3). Since the latter sequences were amplified using different primers compared with complete genomes, the first question posed was whether both sets of sequences coalesced phylogenetically to the identical *env* (or e*nvs*) within each subject, as would be expected if sequences emanate from the same transmitted/founder virus(es). This was the case for each of the 12 subjects. The second question posed was,

given that there were fewer *env* sequences from full-length genomes available for analysis (average of 11 per subject; median 8) compared with *env*-only sequences (average of 36 per subject; median 34), were the estimated days since a MRCA comparable between the two data sets? Again the answer was affirmative with confidence intervals for the MRCAs overlapping in every subject except one (Table 3).

For each subject, we examined *gag*, *pol*, *env* and *nef* genes individually for levels of genetic diversity. Within individual subjects, there was considerable heterogeneity among gene regions in levels of genetic variation, a finding partly attributable to shorter gene lengths and smaller numbers of sequences compared to *env*-only sequences or complete genomes (Table 3). Three individuals lacked any variability in *gag* or *nef*. The maximum pairwise Hamming distance was greatest in *pol* (6 in subject 700010077) and *env* (6 in subjects 04013226-2 and 700010077). Correspondingly, the estimates of days since the MRCA varied among genes such that genes with greater variability, e.g., *env* and *nef*, generally agreed best with the estimates based on greater numbers of *env*-only sequences. Contrary to a previous report (Simmonds 1990a), there was no evidence of greater variability in *gag* compared with *env*. Table S1 shows the total number of synonymous (S_d) and nonsynonymous (S_n) differences of all sequences for each subject for each gene. There was an excess of nonsynonymous substitutions. However, when taking into account the fact that most random replacements in a coding sequence are nonsynonymous, the rate of synonymous divergence (p_s) exceeded that of the nonsynonymous divergence (p_n) for each gene. The p_n/p_s ratio, where values less than 1 indicate the intensity of negative selection for particular genes, reflected an expected pattern where *gag* and *pol* are under more intense pressure to preserve the amino acid

sequence of the transmitted/founder virus compared with *env*, which in turn was more strongly selected to preserve the transmitted/founder amino acid sequence than was *nef*. We also examined whether differences in positive, diversifying selection could be seen among subjects in various Fiebig stages. For these calculations we examined the difference between p_n and p_s (to avoid losing observations due to division by zero); p_n-p_s > 0 indicates an excess of nonsynonymous changes and would be evidence of positive selection. Subjects in later stages of primary infection (e.g., 700010040, 700010077, 700010058) tended to show a preponderance of positive selection, particularly in *env* and *nef* [stats needed -- Fisher Exact test?]. Again, there was no evidence of disproportionate positive selection of *gag* compared with *env*.

Identification of transmitted or early founder viral genomes. Figure 2 illustrates the phylogeny of full-length viral sequences from subject WITO4160 together with *Highlighter* plots depicting the positions and identities of nucleotide polymorphisms, insertions and deletions across the genomes. Together, the phylogenetic tree and *Highlighter* plots illustrate the salient features of SGA-direct amplicon sequencing that allow for an unambiguous identification of transmitted or early founder genomes. Among the 18 WITO4160 sequences depicted, no two were identical. Seven sequences contained a total of 15 double peaks due primarily, if not exclusively, to *Taq* polymerase errors in the initial one or two cycles of PCR amplification. This experimental result (15 *Taq* errors in 18 genomes x 3 cDNA strands x 9000 bp = 3×10^{-5}) is consistent with the error rate reported for *Taq* polymerase of $2.7 - 8.5 \times 10^{-5}$ (Bracho 1998). Excluding mixed bases (which are denoted by an IUPAC designation), single nucleotide insertions

(C1, C10, B3, C9), large deletions (G7 and C4), and G-to-A hypermutation (H3 and C3), each of the sequences differed from the others by 0 to 11 nucleotides. Nucleotide substitutions followed a Poisson distribution and star-like phylogeny (Table 2). The consensus of the sequences (WITO_CON) was the same whether or not sequences containing double peaks were included in the analysis; this is an expected result since *Taq* polymerase errors, like HIV-1 RT errors, are essentially randomly distributed across the genomes. Of the 18 genomes, 9 had intact open reading frames for all essential viral genes, while 9 others contained stop codons or insertions or deletions ranging in length from 1 to 1,329 nt. Figure 3 illustrates in a second subject, ZM247F, many of the same features of early virus diversification but from two transmitted/founder viruses rather than one. Each transmitted/founder virus in ZM247F was represented by a distinct low diversity lineage evident in both the phylogenetic tree and the *Highlighter* plot. Among all 13 sequences, no two were identical. Eight of 13 sequences contained mixed bases at one or two positions. Nine sequences contained all essential open reading frames intact. Three of 13 sequences contained deletions of between 1 and 1,083 nucleotides; one additional sequence contained a nucleotide substitution resulting in a translational stop codon. Neither APOBEC-related G-to-A hypermutation nor viral recombination between lineages was observed in these very early ZM247F sequences. Sequences comprising each viral lineage (with or without mixed bases included) coalesced to unique, unambiguous transmitted/founder genomes that differed by 2.36%.

Sequences from each of the two viral lineages in ZM247F and from the single lineages in WITO4160, SUMA0874, TRJO4551, 04013396-0, and ZM249M, exhibited a Poisson distribution of mutations and a star-like phylogeny (Figures 2, 3 and S1; Table

2), thus allowing for a definitive identification of the transmitted/founder virus. Sequences from six other subjects, however, exhibited shared polymorphisms, which can confound the identification of transmitted/founder sequences (Keele 2008; Lee 2008). Examples of shared polymorphisms are illustrated in Figure 4 for subjects 700010040 and 700010077, each of whom had just entered Fiebig Stage V [vRNA+/ELISA+/WB+(P31-)] at the time of sampling of these sequences. Three shared polymorphisms were evident in sequences of subject 700010040(Figure 4a) and five were evident in subject 700010077 (Figure 4b). Shared polymorphisms can result if two very closely related viruses are acquired during the transmission event, most commonly from a donor who himself/herself is acutely infected (Keele 2008; Hui 2008), or they can arise as a consequence of reverse transcriptase (RT) errors in early virus replication cycles of a single transmitted virus and persist alongside the transmitted viral lineage (Keele 2008; Salazar 2008; Lee 2008). A third possibility is that one or more of the many mutations that occur with sequential replication cycles provides a selective advantage to the virus that results in rapid preferential accumulation of its progeny. We have observed examples of all three of these scenarios in *env* gene analyses (Keele 2008; Salazar 2008; Hui 2008; Anderson 2008). In subjects 700010040 (Figure 4a), 700010077 (Figure 4b), ZM246F, 700010058, 04013226-2, and WEAU0575 (Figure S2), shared polymorphisms initially precluded a definitive identification of transmitted/founder virus sequences, but this uncertainty could be resolved by analysis of viral sequences from earlier or later timepoints. This is because a clear directionality in shifting proportions of shared polymorphisms could be established. In subject 700010040, for example, where three shared polymorphisms at position 6705 in *env* and positions 9360 and 9371 in *nef*

(Figure 4a) were evident in the Fiebig stage V sample dated 7/27/06, analysis of sequences obtained from a still earlier plasma specimen dated 7/11/06, when the subject was at Fiebig stage II, revealed no polymorphisms in *nef*; 4, 12, 24, and 60 weeks after enrollment, 64 of 64 (100%) sequences carried mutations at the two polymorphic sites or within a 27 nucleotide sequence spanning them (Figures 5 and S3). Thus, it was obvious from this analysis that the transmitted/founder sequence at these polymorphic *nef* positions was represented by sequence CH40_fl.CON (Figure 4a). We interrogated the shared polymorphism in *env* in a similar manner. Here, however, 5 of 43 (11%) of sequences from the 7/11/06 sample, 12 of 41 (29%) of sequences from the 7/26/06 sample, and none of 48 sequences 4, 12, 24 or 60 weeks post-enrollment contained the shared polymorphism found in the enrollment sample (Figures 5, S3 and data not shown). Thus, this shared nucleotide polymorphism represented an early stochastic mutation that occurred shortly after virus transmission as predicted in the model (Keele 2008; Lee 2008), was represented as a minor population in the earliest samples, exhibited no fitness advantage and did not accumulate; instead, it was lost as a minor variant in the expanding quasispecies. Thus, in subject 700010040, we could conclude that the transmitted/founder virus genome was represented by the CH40_fl.CON sequence.

Five shared polymorphisms in subject 700010077 could be similarly deconvoluted and the transmitted/founder virus identified (Figure 4b). The initial sample that we analyzed from this subject was also obtained at early Fiebig stage V, similar to the enrollment sample of subject 700010040. At this timepoint at position 7288 in *env*, 45/59 (76%) of 700010077 sequences shared a common nucleotide while 14 others shared a different nucleotide. Fourteen days earlier when the patient was at Fiebig stage

II, 23 of 23 sequences were identical in this region (Figures 5a and S4). Three to 24 weeks after enrollment, 29 of 29 (100%) sequences contained nucleotide mutations at this position or within a 27 nucleotide span that encompassed it. Thus, we could conclude that the transmitted/founder sequence at this position was represented by CH77_fl.A2 and not by what was the consensus sequence at the enrollment timepoint (Figure 4b). A second set of shared nucleotide polymorphisms was evident in *tat* sequences (position 6021) where two full-length genomes (C7 and C3) shared a common polymorphism. We could resolve which sequence represented the transmitted/founder virus by determining viral sequences 14 days before and 3 to 24 weeks after the enrollment full-genome analysis conducted on the 9/9/06 specimen. These results revealed that 18 of 18 (100%) sequences on 08/25/06 were identical to each other at position 6021 and in a 27 nucleotide region spanning it, but they were *different* from 54 of 56 (96%) sequences from the 9/9/06 specimen and from 29 of 29 (100%) sequences obtained an additional three to 24 weeks later (Figures 5 and S4). The two other shared polymorphisms in *pol* at positions 4021and 4104 represented uncommon sequences in the 09/09/06 sample that were not evident in the 8/25/06 sequences and were also absent in sequences from plasma samples obtained 3, 12 and 24 weeks later (Figure S5 and data not shown). Thus, they represented stochastic mutations that occurred sometime shortly after transmission and persisted at a detectable frequency only transiently. A fifth synonymous polymorphism in *pol* at position 2625 could be similarly resolved. From these analyses, we could infer that the CH77_fl.A2 sequence corresponded to the transmitted/founder genome in subject 700010077. For the three other subjects whose sequences exhibited rare shared polymorphisms and where sequential samples were available for analysis, all showed

evidence that their consensus sequences indicated in Figure S2 corresponded to transmitted/founder viruses. Thus, in 11 of 12 study subjects, we could identify unambiguously the transmitted/founder viral genome(s), and in the twelfth subject (04013226-2) we could identify the likely transmitted/founder viral genome with uncertainty at only one nucleotide position out of 9,067 (Figure S2).

Genetic and biological analysis of transmitted/founder viruses. If early viral sequences (eclipse phase through Fiebig stage II; ref. Fiebig 2003) coalesce to *actual* transmitted/founder virus(es) that lead to productive clinical infection, then we would expect three predictions to be borne out experimentally: (i) transmitted/founder *env* sequences inferred from full-length viral genome analysis must be identical to transmitted/founder sequences inferred from *env*-only (subgenomic) SGA analysis; (ii) all essential viral gene open reading frames in transmitted/founder full-length genomes must be intact; and (iii) inferred transmitted/founder full-length genome sequences must encode replication competent viruses. We tested and affirmed all three predictions: First, consensus transmitted/founder *env* genes derived by subgenomic amplifications of *env* and by full genome amplifications were identical in each of the 12 subjects (Table 2). Second, transmitted/founder complete genomes from each of the 12 subjects contained intact *gag*, *pol* (*rt*, *pro*, *int*) *env*, *tat*, *rev*, *vif*, *vpu*, *vpr* and *nef* open reading frames (www.HIV.lanl.gov/salazarfile). Third, three complete HIV-1 clade C proviral genomes corresponding to transmitted/founder viruses for subjects ZM246F and ZM247F were constructed either by chemical synthesis or by PCR amplification of viral nucleic acid followed by cloning into plasmid expression vectors (Figure 6A). All three genomes

(pZM246F-10, pZM247Fv1, and pZM247Fv2) yielded replication competent virus after transfection and expression in human 293T cells. Each of the three virus strains replicated in activated primary human $CD4^+$ T-lymphocytes with kinetics and yields comparable to six control viruses (YU2, SG3, NL4.3, BaL, ADA and JRCSF) (Figure 6B). Surprisingly, each of the three transmitted/founder viruses failed to replicate efficiently when passaged onto human monocyte-derived macrophages obtained from the same normal donor while three prototypic primary macrophage-tropic control viruses (YU2, BaL, and ADA) used as positive controls replicated efficiently in both cell types (Figure 6B). This experiment was repeated four times using four different normal blood donors as the source of lymphocytes and monocyte-derived macrophages and using virus initially generated either in 293T cells or in activated human lymphocytes, each time with similar results. In addition, seven full-length clade B transmitted/founder viruses were generated as part of a different study (C.O-J. and J.C.K., unpublished) and tested for replication in human cells. Again, without exception, transmitted/founder viruses replicated efficiently in activated $CD4^+$ T-lymphocytes but not in monocyte-derived macrophages.

To examine further the biological and antigenic properties of transmitted/founder viruses, we tested HIV-$1_{pZM46F-10}$, HIV-$1_{pZM247Fv1}$, and HIV-$1_{pZM247Fv2}$ for sensitivity to the receptor inhibitor soluble CD4 (sCD4); coreceptor inhibitors TAK-779 and AMD3100; fusion inhibitors T20 and T1249; and mAbs specific for the Env coreceptor binding surface (17b, 21c), CD4 binding site (b12), V3 loop (447-52D and F425-B4e8), membrane proximal external region (2F5 and 4E10), and cell surface CD4. We also tested the transmitted/founder viruses for sensitivity to heterologous subtype B and C

plasmas to assess their overall neutralization profile. The results are summarized in Table 4 and show that the three transmitted/founder viruses exhibit properties typical of primary virus strains including CD4 and CCR5 dependence, effective concealment of the coreceptor binding surface and V3 loop structures, and generalized resistance to neutralization by heterologous plasma antibodies.

Molecular pathways and kinetics of virus diversification and adaptation. HIV-1 virions in plasma have an exceedingly short lifespan (<6 hrs) as do productively infected lymphocytes (<1.2 days) (Wei 1995; Ho 1995; Ramratnam 1999; Markowitz 2003). As a result, the genetic composition of plasma virus can change quickly and provides a sensitive indicator of selection pressures acting on virus and virus-producing cells (Wei 1995; Ho 1995; Borrow 1997; Wei 2003; Jones 2004). These properties, combined with the identification of full-length transmitted/founder virus genomes, provided a unique opportunity to evaluate the kinetics and precise molecular pathways of HIV-1 sequence diversification and evolution, since mutations could be mapped to a specific viral genome at or near the moment of transmission. In none of the 12 subjects did we find evidence of positive selection in viral sequences obtained prior to first antibody detection (prior to Fiebig stage III). Instead, we found evidence of selection against amino acid conferring substitutions during this early period with pn-ps values less than predicted for neutral evolution Table S1). Sequence diversification during this early period of infection was notable for a high proportion of defective genomes that contained insertions or deletions (INDELs), inframe termination codons, or G-to-A hypermutation: 27 sequences had frameshifting INDELs; 3 sequences had large in-frame deletions 318 –

1,329 nt in length; 5 sequences encoded products that were truncated due to inframe stop codons arising from nucleotide point mutations; and 4 sequences contained evidence of APOBEC-mediated G-to-A hypermutation. Altogether, 34 of 108 sequences contained mutations that rendered virus progeny overtly defective. Even this high proportion of defective genomes is an underestimate of the proportion of viruses or proviruses that contain defective genomes, since we excluded from our analyses amplicons shorter than 7 kb in length and we did not examine the integrated proviral DNA compartment directly. In a separate study, we have done the latter and have found that approximately two-thirds of HIV-1 infected cells in acutely infected subjects harbor overtly defective genomes (Guffey 2009).[move or delete?]

We next examined sequences from a subset of subjects (700010040, 700010077, and 700010058) beginning prior to peak viremia (Fiebig stage II) through viral load setpoint 4 to 14 months later (Fiebig stage VI). We did this first by identifying the transmitted/founder viral genome in each subject and then (using the same SGA-direct sequencing approach to avoid *Taq*-induced nucleotide substitutions and *in vitro* generated recombination events; see Salazar-Gonzalez 2008) by determining genomic sequences of plasma virus amplified as overlapping half-genomes over a period of as much as 60 weeks follow-up (Figure 5). In contrast to sequences obtained during Fiebig stage II, which generally contained few if any shared polymorphisms (Figs. 2, 3, S1 and S2), sequences from subjects 700010040 and 700010077 at early Fiebig stage V (days 16 and 14 following screen, respectively) and from subject 700010058 at Fiebig stage III (day 9 following screen) exhibited in some instances multiple shared polymorphisms and non-star-like phylogenies (Figures 4 and 5 and S7-9). This included nucleotide positions

2625, 4021, 4104, 6002, 6021 and 7288 in subject 700010077 (day 14); nucleotide positions 6705, 9306 and 9371 in subject 700010040 (day 16); and nucleotide position 4408 in subject 700010058 (day 9). In these three subjects, it was necessary to extend our sequence analyses to earlier screening samples (Figs. 5 and S3-S6) and to later samples days 32 to 412 after screening (Figs. 5 and S3-S6) in order to confirm the identity of the transmitted/founder viruses and to determine which of the early shared polymorphisms were subsequently positively selected and which were not. By 14 – 45 days after the first screening timepoint, which was at Fiebig stage II just prior to antibody seroconversion and less than three months after acquisition of infection, clear patterns of nonrandom substitutions leading to mutational fixation became evident in the sequences from each subject Fig. 5. The most rapid and complete replacement of transmitted/founder virus sequences by mutant virus was in subject 700010077 where within a period of two weeks [between screen (day 0) and day 14; see CH77 in Fig. 5 and S7) virtually the entire replicating virus population in the body that contributed virus to the plasma compartment was replaced by what were subsequently proved to be viruses containing CTL escape mutations at two epitopes (ref. to NiLu 2008 and Fig. S7); equally remarkable was the finding that this mutant virus population was in turn completely replaced by still another population that contained two additional CTL escape mutations 18 days later (compare day 14 and day 32 sequences in CH77; Fig. 5 and S7). Using a sliding "window" of 27 nucleotides, we identified in the three subjects at days 32 to 45 between 2 and 5 loci where nonsynonymous substitutions became essentially completely fixed (Figures S7-9). If we extended the analysis of synonymous and nonsynonymous

nucleotide substitutions out to 159 to 412 days after the screening sample, we found evidence of fixed mutations at 18 to 27 loci across the genomes.

Based on the patterns of nucleotide substitutions that we observed across the viral genomes, we hypothesized that the genetic imprint of most if not all viral and host selection pressures might be evident or decipherable from the genomic sequences. We thus sought to develop a systematic, statistically-based approach to assessing virus diversification and evolution that would allow us to identify nonrandomly distributed changes across multiple genomes, including linked substitutions that were spatially separated. **[Bette et al. section goes here]**

Discussion

Understanding the precise molecular events underlying HIV-1 transmission and the subsequent steps leading to productive clinical infection and viral persistence could prove invaluable to the development of effective HIV-1 vaccines and microbicides. But elucidating these events in a biologically and physiologically relevant context has been problematic since neither the particular virus(es) responsible for productive infection nor their initial target cells are known with certainty (Shattock and Moore 2003; Haase 2005; Margolis and Shattock 2006; Pope and Haase 200X; Hladick 2007) . Here, we take a step toward addressing this challenge by showing that the complete genome of the transmitted/founder virus can be identified unambiguously, that it can be synthesized,

cloned, expressed and analyzed phenotypically, and that its evolution can be mapped precisely.

The current study builds on recent reports by our group and others that describe a model of HIV-1 evolution in acute infection and an empirical analysis of *env* gene diversity -- all based on SGA and direct amplicon sequencing of plasma vRNA/cDNA -- in over XXX acutely infected subjects from five different cohorts representing clade A, B, C, and D infections (Keele 2008; Lee 2008; Salazar-Gonzalez 2008; Haaland 2008; Williamson 2008; Anderson 2009; Li 2009; Baalwa 2009). The common finding from all of these studies is that the transmitted/founder *env* genes of viruses responsible for productive clinical infection can be inferred from viral sequences obtained early in infection, and that in most cases of sexual transmission of HIV-1, one or very few viruses are responsible. This is true for heterosexuals and for men who have sex with men (MSM), although in the latter instance transmission of more than one virus appears to be more common, accounting for ~40% of transmissions versus 10-20% for heterosexual transmissions (Ritola 2004; Li 2009).

The present study was designed to extend this work on transmitted/founder *env* sequences to an analysis of complete HIV-1 genomes and to address five questions specifically: First, can methods for SGA-direct sequencing of plasma vRNA/cDNA *env* sequences be applied successfully to full-length HIV-1 genomes? The answer is yes, but because of the four-fold increased length of the amplified genome, shared nucleotide polymorphisms and mixed bases are detected approximately four times more frequently. Such polymorphisms are predicted by the model (Keele 2008; Lee 2008) and by the known rates of *Taq* and HIV-1 RT polymerase error (Mansky and Temin 1995; Bracho

1998), and they can be reconciled by the analysis of sequential sequences. It is preferable to perform full-length SGA-direct sequence analysis on samples from as early in infection as possible (preferably before antibody detection), since the consequences of CTL selection are generally not evident in viral sequences at this time. Not unexpectedly, we observed that specimen storage conditions, sample integrity, and plasma vRNA load were more critical for successful SGA of full-length genomes compared with *env*-only or still shorter subgenomic fragments. We could circumvent this problem in our study by limiting the analysis of certain early Fiebig stage II samples to subgenomic fragments that spanned shared polymorphisms that we identified in subsequent samples samples (Figure 5). These limitations notwithstanding, we have not encountered in this or in other studies, suitably processed and stored samples of plasma from an acutely infected individual with clade A, B, C, or D infection where the primers and amplification conditions described herein did not yield full-length genomic sequences.

A second question we posed was if early diversification of full-length HIV-1 genomes conforms to a model of random evolution with a Poisson distribution of mutations and star-like phylogeny? This question was answered affirmatively once APOBEC-mediated hypermutations, early stochastic changes, and early CTL selected mutations were excluded and distinct low diversity sequence lineages in a subject (ZM247F) infected by two viruses were analyzed individually (Fig. 3). This result, together with clinical history, laboratory staging (Table1; see also Fiebig 2003 and Keele 2008) and model estimates of goodness of data fit and time to a MRCA (Table 2), provided strong evidence that viral genomic sequences sampled at or near peak viremia

coalesce to single MRCA sequence(s) at or near the moment of virus transmission (Lee 2008). There are, however, two important caveats: First, to identify transmitted/founder sequences and distinguish them from the earliest CTL selected escape variants, sequences from Fiebig stages I-III must generally be analyzed. Second, shared polymorphisms in two or more sequences must always be interrogated further in order to determine those that represent the transmitted/founder sequence versus the mutant sequence. The latter can be done efficiently by subgenomic analyses of earlier and/or later samples (Figure 5).

Thirdly, we asked if inferences regarding the identity and number of transmitted viruses and the estimated time to an MRCA based on full-length genome analyses corresponded to findings based on *env*-only gene analyses? Theortically this should be the case, although the sensitivity and precision of such determinations depend on the number of target sequences analyzed (Keele 2008; Lee 2008). Table 3 shows that in the present study the answer to all three queries was yes. We found that the numbers of transmitted/founder viruses were exactly the same whether based on analyses of complete genomes or *env*-only sequences, that the inferred nucleotide sequences of transmitted/founder *envs* were identical whether based on analyses of complete genomes or *env*-only sequences, and that there were no substantial differences in the estimated MRCA of sequences based on either analysis. We have since corroborated these findings in additional studies where we identified the transmitted/founder virus from PBMC proviral DNA and plasma vRNA at two different time points and using primer sets that amplified complete viral genomes, 5′ and 3′ half-genomes, or *env*-only genomes (Li 2008; Guffey 2008). Altogether, the findings indicate that sampling biases from primer selection or virus compartmentalization did not affect the identification of

transmitted/founder viruses in the setting of acute HIV-1 infection. Of note, we examined *gag*, *pol* and *nef* genes individually compared with *env* (Table 3) for model fitness, estimated MRCA, and evidence of selection, given that a previous report had suggested that nonsynonymous Gag mutations exceeded Env mutations in acute and early infection as a consequence of early CTL pressure (Simmonds 1990). We observed that variation in all four genes conformed to model predictions, that estimated MRCAs generally did not differ substantially among them, and that there was no indication of a higher mutation rate (synonymous or nonsynonymous) in *gag* than in any other gene (Tables 3 and S1). We attribute differences between our findings and those in the Simmonds study to differences in experimental strategy and methods and to our analysis of sequence changes against an unambiguous consensus transmitted/founder sequence.

Fourthly, do inferred transmitted/founder full-length sequences have canonical open reading frames and do they encode replication competent viruses as would be expected of transmitted/founder viruses responsible for spawning productive clinical infection? Indeed, each of the 13 transmitted/founder genomes that we identified had intact open reading frames for *gag*, *pol* (*rt*, *pro*, *int*) *env*, *tat*, *rev*, *vif*, *vpu*, *vpr* and *nef*. Of interest, we noted that two of the clade C viruses (both from subject ZM247F) were atypical in genome organization in having *vpu* and the first exon of *rev* expressed from the same reading frame without an intervening termination codon (Fig. 6). This gene arrangement presumably has no substantial impact on the biology and natural history, since viruses from about 20% of clade C infected patients have this genome organization. Moreover, peak plasma viral loads in subject ZM247F exceeded 10 million vRNA molecules per milliliter comprised of viruses having this genome arrangement. *In vitro*,

however, this uncommon genomic organization adversely affects Env expression from subgenomic *rev-vpu-env* clones, and in Env pseudotype expression assays, the infectivity of Env-pseudotyped viruses is impaired (Parrish 2009). This problem can be circumvented by site-directed mutagenesis of *rev-vpu-env* constructs either by placing a stop codon in frame between *rev* and *vpu* or by inserting a single nucleotide between the two genes so as to place them in different translational reading frames (Parrish 2009).[delete or modify]

To determine if transmitted/founder genomes encoded replication competent viruses, we selected the three full-length clade C sequences for analysis. Obtaining full-length recombinant clones of proviral genomes whose sequences match exactly those of transmitted/founder viruses can be challenging because of errors introduced by Superscript III MuLV reverse transcriptase or by *Taq* DNA polymerase in individual HIV-1 genomic sequences. Further complicating the cloning of replication competent virus is an inherent pathogenicity of some *env* genes when grown in bacteria (refs). We circumvented these problems by pursuing two different strategies. First, we chemically synthesized and subcloned the complete proviral genome in three overlapping subgenomic fragments, none of which included an intact *env* open reading frame. Then, in a final step, the three fragments were cleaved with restriction enzymes, ligated, and cloned into a low copy number plasmid vector (pBR246-10; Fig. 6). This was subsequently grown at reduced temperatures and for shorter durations than normal to prevent spontaneous deletions in *env* (refs). This approach avoided the RT and DNA polymerase steps in the cloning process altogether and minimized the effects of *env* gene expression in bacteria. As a second approach, we circumvented the Superscript III RT

step and used the high fidelity Phusion DNA polymerase to amplify in two overlapping fragments (neither containing an intact *env* gene) proviral sequences from pre-seroconversion high molecular weight PBMC DNA. This was followed by a ligation step and cloning of the provirus into a low copy plasmid vector (pXL247Fv1 and pXL247Fv2). The nucleotide sequence of each of the three proviruses was confirmed to be identical to the respective transmitted/founder viral sequences. After transfection into 293T cells, progeny viruses replicated efficiently in activated primary human $CD4^+$ T cells (Figure 6B upper panel). Surprisingly, however, none of the three transmitted/founder viruses replicated efficiently in monocyte-derived macrophages from the same donors (Figure 6B lower panel). This result was reproducible in lymphocytes and monocyte-derived macrophages from four different normal donors, and in a different study, was confirmed using 7 clade B transmitted/founder viruses (C.O-J. and J.C.K.). [eliminate redundancy in Results] We note that other investigators have observed that only a subset of primary R5 tropic viruses replicate efficiently in human monocyte-derived macrophages (Collman; Clapham). These findings suggest that prototypic macrophage-tropic HIV-1 strains such as YU2, ADA, and BaL, which have frequently been used to model HIV-1 transmission, may not be representative of a substantial proportion of transmitted/founder viruses with respect to cell tropism. Our findings further suggest that during the initial stages of infection between transmission and peak viremia, replication of the transmitted/founder virus and its progeny in macrophages is unlikely to contribute substantially to overall virus production compared with lymphocyte tropic viruses. This conclusion is consistent with other studies that have found that activated memory $CD4^+$ T cells, not macrophages, are principal targets of HIV-1 or SIV

infection in human and primate tissues studied *in vivo* or *ex vivo* as tissue explants (Haase 2005; Hladik 2009). [refs from haase and veazy]

Other biological properties of the three transmitted/founder viruses that we studied included their sensitivity or resistance to a large number of Env-specific ligands (Table 4). Viruses HIV-$1_{pZM246F-10}$, HIV-$1_{pZM247Fv1}$ and HIV-$1_{pZM247Fv2}$ were similar to other primary HIV-1 strains in CD4 and CCR5 dependence and in their sensitivity to the gp41 peptide fusion inhibitors T20 and T1249 and to the MPER mAb 4E10. Only HIV-$1_{pZM246F-10}$ was sensitive to B12 and none of the three were sensitive to 2G12 or 2F5, likely due to epitope variation of C clade viruses. The coreceptor binding surface of diverse HIV-1 (and HIV-2) strains is antigenically conserved (Decker 2005) as is the V3 loop (Davis 2009), but HIV-$1_{pZM246F-10}$, HIV-$1_{pZM247Fv1}$ and HIV-$1_{pZM247Fv2}$ were completely resistant to CD4i (17b, 21C) and V3 (447-52D, F425-B4e8) mAbs. This finding suggests that the CD4i bridging sheet and V3 epitopes on functional clade C Env trimers are effectively shielded. Finally, we found that 11 clade B plasmas and 22 clade C plasmas (all heterologous) exhibited low overall Nab titers against the three clade C viruses (<1:20 and <1:40, respectively). This again suggested that the functional Env trimer on these transmitted/founder clade C viruses is effectively masked from antibody recognition. Altogether, the findings indicate that transmitted/founder viruses from subjects ZM246F and ZM247F are preferentially replicative in primary human lymphocytes compared with monocyte-derived macrophages and that their sensitivity to Env-specific ligands resembles that of primary HIV-1 clade B and C viruses.

The last question we asked was if the identification of transmitted/founder full-length genomes and their progeny could provide new insight into the kinetics and

molecular pathways of virus diversification and adaptation leading to viral persistence? For this analysis, we utilized mathematical modeling and statistical analysis to . . . . **[Bette and colleagues to add a section here]** Between transmission and peak viremia (Fiebig stage II), we found that virus diversified in an essentially random fashion leaving no or little evidence of host-related selective pressures imprinted on its genome. However, between peak viremia and antibody seroconversion 9 – 16 days later (Fiebig stages III-V), evidence of striking selection on the virus quasispecies became evident (Figures 5 and S7-9). In subject 700010077, for example, nearly all of the circulating virus was replaced by a population that was mutant at two positions, one at ~6021 in Tat and another at position ~7288 in Env. At this very early time point, 20% (3/15) of plasma viral seequences also contained a region of multiple nonsynonymous replacements at position 8858-8869 in Nef where 18 days later (day 32 in Figure 5 and S7) 100% of sequences differed from the transmitted/founder sequence, as did a fourth mutation at position ~9AAA. It is notable that all four mutations exist colinearly on the same viral RNA molecules. This pattern of early rapid virus evolution to escape CTL pressure is also evident in subjects 700010040 and 7000100758 (Figures 5, S8 and S9). In studies reported elsewhere, we could attribute each of these very early mutational fixations to CTL recognition and escape (Goonetilleke 2008). The rapidity of CTL recognition and virus escape described here is substantially faster and earlier than previously reported (McLean 2006; Walker papers) and was possible to observe only because we compared early viral sequence diversity against the actual transmitted/founder virus. Had we not done this, these earliest escape mutations would have gone undetected (Goonetilleke 2008). Using a cost-to-fitness model described by

others (McLean 2006), we could calculate that these earliest CTL responses exacted a fitness cost on the transmitted/founder viruses in this study of XX to YY%, far higher than reported for CTL pressure in other studies (McLean 2006). A fitness cost of XX to YY% corresponds to a reduction in virus burst size of XX to YY%, and this explains the rapid fixation of escape mutations in the evolving viral quasipecies. **[Move to Nilu ms?]**

By 32 – 45 days post peak viremia, most patients have fully seroconverted to antibody positivity (Fiebig V) on a diagnostic western immunoblot and are approaching the semi-steady state virual load setpoint (Figure 5). We asked in our subjects how many loci of amino acid mutations within a span of 10 codons or less were associated with the establishment of virus persistence at this early time point approximately 2 to 2½ months after infection. The answer was notably few: 4 substitutions for 700010077, 5 for 700010040, and 2 for 700010058 (Figures 5 and S7-S9). Remarkably, all 11 of these mutated loci could be accounted for by HLA-restricted CTL pressure and virus escape (Figures S7-S9; Goonetilleke 2008). Thus, not only could we identify *all* amino acid changes in a transmitted/founder virus that were required for the establishment of productive infection in a new host, but each mutation could be accounted for based on escape from CTL pressure. There was no suggestion of other adaptive changes such as reversion from CTL pressure, adaptation to replication in different cell types, or evidence of selection by other immune effector modalities (e.g., innate or antibody responses). If any of these other immune activities are at play in the very early infection period, then their effects could not be recognized at a genetic level by virus escape in these three subjects. At later time points (days 85-412 after peak viremia), we observed evidence of mutational fixation at many more loci (15 in subject 700010077; 23 in 700010040; and

18 in 7000100058; see Figures S7-S9). Some of these mutations were in Env. In subject 700010040, we could precisely associate at day 111 loss of virus susceptibility to Nabs with a single amino acid mutation in V1 (position 6661) that conferred complete Nab escape (Bar 2008). Thus, in this subject, we could identify by simple inspection of sequential sequences followed by phenotypic analyses *all* mutations that were necessary and sufficient for the evolving viral quasispecies to avoid a potent host immune response. [expand to discuss V3 contribution to the epitope] At later time points, however, we noted an increasing number of mutations including those that were synonymous, others that were in noncoding regions of the LTR, and still others that were nonsynonymous changes in open reading frames that could not be readily explained by CTL or Nab escape. CTL reversion and other immune or nonimmune mediated selection pressures are likely responsible, and additional analyses will be required to elucidate the biological basis and importance of these imprinted changes in the evolving viral quasispecies.

Finally, we have begun to explore if the identification of transmitted/founder sequences and their progeny by SGA-direct sequencing methods can be an effective experimental strategy for studying other viral infections besides. We hypothesized that SGA analysis might elucidate key features of the SIV-macaque model of HIV infection relevant to vaccine development and assessment. Specifically, we posited that SIVsmE660 and SIVmac251 inoculation onto rectal or vaginal mucosa of rhesus macaques would recapitulate features of mucosal HIV-1 transmission in humans, that low-dose SIV challenge would result in the transmission of one or few viruses, that SIV diversification would adhere to the same model of acute infection as does mucosal HIV-1 infection, and finally, that transmitted/founder SIV *env* sequences and complete viral

genomes could be identified, cloned, characterized, and potentially used as genetically-defined virus inocula. All of these predictions have been borne out (Keele 2009), thus highlighting the potential value of SGA-direct sequencing approaches in the molecular analysis of a variety of human or animal pathogens including, but not limited to, HIV-1 and SIV.

METHODS

Study subjects: Plasma samples were obtained from 12 subjects with acute or very recent HIV-1 infection. All subjects gave informed consent, and plasma collections were performed with institutional review board and other regulatory approvals. Blood specimens were generally collected in acid citrate dextrose and plasma separated and stored at -70°C.

Laboratory staging: Plasma samples were tested for HIV-1 RNA, p24 antigen, and viral specific antibodies by a battery of commercial tests. These included quantitative Chiron bDNA 3.0 or Roche Amplicor vRNA assays; Coulter or Roche p24 Ag assays; Genetic Systems Anti-HIV-1/2 Plus O; and Genetic Systems HIV-1 Western Blot Kit. Based on these test results, subjects were staged according to the Fiebig laboratory classification system for acute and early HIV-1 infection [Fiebig 2003] (4). The duration of the eclipse phase (prior to the detection of plasma viral RNA) was estimated to be 10 days (range 7-21 days) [Keele 2008] (5-10).

Viral RNA extraction and cDNA synthesis: For each sample approximately 20,000 viral RNA copies were extracted using the QIAamp Viral RNA Mini Kit (Qiagen, Valencia, CA) or using the Qiagen BioRobot EZ1 Workstation with EZ1 Virus Mini Kit v2.0 (Qiagen, Valencia, CA). Samples with low vRNA loads (200 - 10,000 copies/ml) were concentrated by centrifugation at 23,600 x g for 1 h at 4°C prior to the same extraction procedure. RNA was recovered from the spin columns in a final elution volume of 60 microliters (μl). RNA was either frozen at -80°C or immediately used to synthesize cDNA. ~5,000 vRNA molecules or less were reverse transcribed. Reverse transcription of RNA to single-stranded cDNA was performed by using the SuperScript III protocol according to the manufacturer's instructions (Invitrogen Life Technologies, Carlsbad, CA) and modified as follows. RT reactions were carried out in 40 ul reactions; final concentration of reagents are shown in parentheses. First, 15 microliters of RNA were mixed with 11 microliters of a master mix A containing deoxynucleoside triphosphates (0.5 millimolar each), and primer 1.R3.B3R (5' – ACTACTTGAAGCACTCAAGGCAAGCTTTATTG -3' corresponding to nucleotides 9642 to 9611, HXB2 numbering) (0.25 micromolar) and incubated for 5 min at 65°C to denature secondary RNA structure. Following denaturation, the tube was incubated on ice for 1 min. Then, 26 microliters of the reaction mixture were combined with a master mix B containing reverse transcriptase buffer (1X), dithiothreitol (5 millimolar), RNase inhibitor RNaseOUT (2 units/microliter), and SuperScript III (5 units/ul). Tubes were then incubated at 50°C for 1.5 h. An additional 1 microliter of SuperScript III was then added and tubes incubated for another 1.5 h at 55°C. Following the completion of the reverse transcription step, the reaction mixture was inactivated by heat (70°C for 15 min)

followed by addition of 1 microliter of RNase H and incubated at 37°C for 20 min. The resulting cDNA was then aliquoted and frozen at -80°C until further analysis or used immediately for PCR.

Single genome amplification: An aliquot of cDNA was serially diluted in replicates of 8 PCR wells and submitted to nested PCR amplification with HIV-specific primers that yield a 9 Kb fragment. cDNA dilutions that yielded 40% or fewer PCR positive wells were retested in 96 well plates to identify a dilution where <20% of wells were positive for amplification. First-round PCR was carried out in 1x Expand Long Template buffer 1 (1.75 mM $MgCl_2$ final concentration), 0.35 millimolar of each deoxynucleoside triphosphate, 0.3 micromolar of forward primer 1.U5.B1F 5'-CCTTGAGTGCTTCAAGTAGTGTGTGCCCGTCTGT-3' (HXB2 positions 538-571), reverse primer 1.R3.B3R and 3.75 units/microliter of Expand Long Template Mix (Roche Applied Science, Mannheim, Germany) in a 50 microliter reaction mixture. The PCR mixtures were set up in MicroAmp optical 96-well reaction plates (Applied Biosystems, Foster City, CA) and sealed with ABI MicroAmp adhesive film. The following PCR conditions were used: 94°C for 2 min followed by 10 cycles of 94°C for 15s, 55°C for 30 s, and 68°C for 8 min, followed by 25 cycles of 94°C for 15 s, 55°C for 30 s, and 68°C for 8 min with cumulative increments of 20 sec at 68oC with each successive cycle, followed by a final extension of 68°C for 10 min. Second-round PCR was carried out by transferring 1 microliter of the first-round product into a final volume of 50 microliter of a reaction mixture containing 1x Expand Long Template buffer, 0.35 millimolar of each deoxynucleoside triphosphate, 0.3 micromolar of forward primer 2.U5.B4F 5'-

AGTAGTGTGTGCCCGTCTGTTGTGTGACTC-3' (nt552-581) and reverse primer 2.R3.B6R 5'-TGAAGCACTCAAGGCAAGCTTTATTGAGGC-3' (nt9636-9607). The PCR conditions were the same indicated for the first round PCR. The amplicons were sized on precast 1% agarose E-gel 48 (Invitrogen Life Technologies, Carlsbad, CA). All products derived from cDNA dilutions yielding < 20% PCR positive wells and appearing to be >7 kb by analytical gel electrophoresis were sequenced. All PCR procedures were carried out under PCR clean room conditions using procedural safeguards against sample contamination, including pre-aliquoting of all reagents, use of dedicated equipment, and physical separation of sample processing from pre- and post-PCR amplification steps.

DNA sequencing: Amplicons were directly sequenced by using BigDye Terminator chemistry and protocols recommended by the manufacturer (Applied Biosystems, Foster City, CA). Sequencing reaction products were analyzed with an ABI 3730xl genetic analyzer (Applied Biosystems; Foster City, CA). Both DNA strands were sequenced using partially overlapping fragments. Individual sequence fragments for each amplicon were assembled and edited using the Sequencher program 4.7 (Gene Codes; Ann Arbor, MI). All chromatograms were inspected for sites of ambiguous sequence (double peaks). These were recorded in the finished by IUPAC code.

Sequence alignments: All alignments were initially made with GeneCutter (www.hiv.lanl.gov) to compensate for frame shifting mutations. Sequences of complete genome were aligned using Clustalw (Thompson 1994) and alignments were visually

adjusted to optimize codon alignment using MacClade version 4.08 (Maddison and Maddison 2001). To generate the final consensus sequence for each patient, ties near regions of insertion and deletions were resolved by considering the proximal codons and context. The full alignment is available in a supplemental data file, and the sequences are also available through GenBank. All 108 genomic sequences were deposited in GenBank and transmitted/founder full-genome sequences can be accessed at www.hiv.lanl.gov/content/sequence/hiv/user_alignments/Salazar.

Sequence diversity analysis. We classified two very distinctive levels of within-patient diversity that we observed in the 12 study subjects as either "homogeneous" or "heterogeneous." This was done using three different strategies which all concurred. First we visually inspected the samples using neighbor-joining or maximum likelihood phylogenies and the *Highlighter* tool (www.hiv.lanl.gov/content/sequence/HIGHLIGHT/highlighter.html) and found that one sample clearly had much greater diversity than the other 11. We then looked at all pairwise Hamming Distances (HD, defined as the number of base positions at which the two genomes differ, excluding gaps) within each sample. The same heterogeneous sample compared with the homogeneous samples exhibited distinct peaks inconsistent with expansion from single infecting virus. Lastly, to formalize the criteria and test whether the heterogeneous sample reflected transmission of more than one variant, we used the mathematical model described below to predict the expected maximum HD that could be observed under a homogeneous infection assumption (i.e., infection by a single virus), given the Fiebig stage of the sample. If the maximum HD in the sample was

much greater than the expected, the observed diversity was considered to have originated at a time prior to transmission, *i.e.* in the donor indicating that multiple strains transmitted from the donor to the recipient established the infection; this was the case for the sample. For the homogeneous samples, we considered the possibility that these individuals had been infected by a single virus (or infected cell) or by two very closely related viruses. Either scenario could result in a low overall *env* diversity, but in the case of transmission of two very closely related viruses, the distribution of HDs would not fit model expectations. We found this to be the case in none of the subjects with homogeneous infection.

Star phylogeny. With no selection pressure, one can expect homogeneous viral populations to evolve from a founder strain following a star-like phylogeny, *i.e.* all evolving sequences coalesce at the founder. The validity of this proposition can be investigated by inspecting the sequence alignment. Since mutations are rare, one does not expect shared mutations in a star phylogeny. When this is indeed the case, the distribution of intersequence HD's is constrained to be a self-convolution of the distribution of the HD's from the sequences to the ancestral sequence. In particular, for every pair of sequences s_1 and s_2, let $HD[s_1,s_2]$ be the number of base positions at which the two differ and the probability distribution it follows be $P_I(HD)$. Next, we compare each sequence in the sample with the consensus sequence (which we assume to be the founder strain), and compute the corresponding HD distribution. Denoting s_0 the founder strain, for every sequence s_1 we compute $HD[s_0,s_1]$ and we denote $P_C(HD)$ the distribution it follows.

Then, under a star-phylogeny evolution, $P_I(HD)$ is given by the self-convolution of $P_C(HD)$:

$$P_I(HD=n)=\sum_{k=0}^{n}P_C(HD=k)P_C(HD=n-k)$$

Occasional deviations from a star phylogeny are, however, expected. The sampling of 30 sequences, for example, from a later generation of an exponentially growing population with six-fold growth per generation has about 5% chance of including a pair of sequences which shares five initial generations, a 25% chance of those sharing the first four, and overwhelmingly likely to include sequences that share three ancestors. But since the rate of mutations in the region under study is about 1 per 5 generations (see next section), this leads to about 40% chance of finding sequences sharing a pair of mutations, and less than X% chance of sharing more than that. The probabilities are slightly enhanced by the early stochastic events that can lead to the virus producing less than six descendants in some generations, but it remains overwhelmingly likely that the sequences share few mutations.

Mathematical model. The model employed in the present study has been described (Keele 2008). It assumes a homogeneous infection in which the virus grows exponentially with no selection pressure, no recombination, no occurrence of back mutations and a constant mutation rate across positions and across lineages. Under this scenario, the HD frequency distribution is given by a Poisson distribution whose mean depends linearly on the number of generations since the founder strain.

Hypermutated samples. Enrichment for APOBEC3G/F mutations violates the assumption of constant mutation rate across positions, as the editing performed by these enzymes are base and context sensitive. Enrichment for mutations with APOBEC3G/F signatures was assessed using Hypermut 2.0 (www.hiv.lanl.gov/content/sequence/HYPERMUT/hypermut.html), which compares each sequence in the sample to the consensus sequences. Hypermut detects an enrichment for G→ A mutations that occur in the context of the APOBEC3G/F signature pattern, where the G is followed by either G or A, then by a base that is not C (in IUPAC code, the pattern GRD where the first G changes to A). A contingency table is constructed which is then used to obtain Fisher exact p-values that test whether the extent of hypermutation is more than would be expected by chance. Single sequences that yielded a p-value of 0.05 or lower were considered significantly hypermutated and therefore were not included in other analyses.

Synonymous and nonsynonymous mutation analysis. Coding regions for the four largest HIV-1 genes, *gag*, *pol*, *env*, and *nef*, were extracted from the complete genome sequences from the enrollment sample for analysis of the distribution of sequence differences from the transmitted/founder virus over synonymous and nonsynonymous sites. Regions where these sequences overlapped with different reading frames were removed from the gene sequences. The proportion of nonsynonymous differences per nonsynonymous sites (pn) and synonymous differences per synonymous sites (ps) were

calculated using the SNAP tool from the LANL HIV Sequence Database (www.hiv.lanl.gov/content/sequence/SNAP/SNAP.html) (Korber 2000).

Comprehensive summary of observed evolution patterns, timing estimates, and statistics based on a model of random evolution in acute infection under no selection. Table 2 contains demographic and sample information corresponding to each study subject along with summaries of model parameters and statistics, including the estimated time from the MRCA of the sequences found in each sample. The confidence intervals in the table do not take into account uncertainty in the viral mutation rate or the possibility of selection against a proportion of the non-synonymous mutations. Estimates based on the random evolution model were based on all sequences or edited sequence sets (see below). Substitutions with APOBEC3G/F signatures were excluded the Poisson/Star model; when this was done the observed sequence variation became compatible with this model and the estimated days to the MRCA was reduced. For patients with low diversity that significantly deviated from this model, we noted in the table whether the best explanation for the deviation was CTL selection or early stochastic events. For the subject with high diversity and two infecting strains, the timing estimates from a random evolution model are based on times to the MRCA restricted to sequences from each clade in the subject, as this provided a strategy to test if each such clade plausibly represented the outgrowth of a single transmitted virus, which they did.

Analysis of patterns of mutational fixation in sequential viral sequences. [Bette and colleagues methods go here]

Power studies. To better understand our likelihood of missing infrequent transmitted variants, we did a power study to explore the probability of sampling limitations. We show that with a sample of at least n=20 plasma vRNA sequences, we could be 95% confident that a given missed variant comprised less than 15% of the virus population. For samples for which n≥30, we could be 95% confident not to have missed any variant that comprised at least 10% of the total viral population. Because it is not practically feasible from a cost or time perspective to generate more than 10 full-length genomes, it is preferable to perform SGA-direct sequencing on ~30 subgenomic fragments to establish the number of transmitted/founder viruses. This in turn allows the full-length genomic analyses to be focused on a predetermined number of transmitted/founder. This was done in for the present study where we had previously identified the transmitted/founder viruses by *env*-only SGA analyses (Salazar 2008; Keele 2008; Li 2009).

Proviral DNA cloning. To obtain an infectious molecular clone of the transmitted/founder virus from subject ZM246F, the entire 9.7 kb proviral consensus sequence was chemically synthesized in three subgenomic fragments of 4 kb, 2.6 kb and 3.1 kb (Blue Heron Biotechnology). These fragments overlapped at unique NheI (position 3982) or DraIII (position 6586) sites, respectively, and were propagated in plasmid vectors. A 5' terminal MluI cloning site that was attached during chemical synthesis and a BamHI polylinker site at the 3' end of the genome were utilized for subsequent manipulation. To facilitate assembly of a full-length proviral clone, a unique

MluI site was engineered by oligonucleotide adapter ligation between EcoRI and HindIII of pBR322. The 4kb 5' fragment was then cloned in MluI-NheI of the modified pBR322 vector. Following restriction enzyme digestion of this recombinant plasmid with NheI and BamHI, the full-length proviral clone pBR246F-10 was generated by simultaneous ligation with the 2.6kb middle NheI-DraIII and the 3.1kb 3' DraIII-BamHI fragments. The integrity of the entire proviral consensus sequence was confirmed by physical mapping and nucleotide sequence analysis.

To obtain infectious molecular clones of the two transmitted/founder viruses from subject ZM247F, we employed a high fidelity DNA polymerase with proviral DNA as a template to obtain both 5' and 3' halves each with the entire long terminal repeat (LTR). High molecular weight genomic DNA was isolated from PBMCs obtained on the same day that plasma virus was used to identify the two unique transmitted lineages. The entire genome for both lineages were amplified and cloned in two overlapping halves encompassing a unique restriction site PacI. Single-round PCR amplification was carried out in the presence of 1x Phusion Hot Start HF Buffer, 0.2mM each deoxynucleoside triphosphate, 0.5 uM of each primer, 3% final concentration of DMSO and 0.02 units/ul Phusion Hot Start High Fidelity DNA polymerase (New England BioLabs) in a 40ul reaction. PCR primers for the variant 1 are (V15'F) 5'-TGGAAGGGTTAATTTACTCCAAGAAAAGG-3' and (V15'R) 5'-CACTGTCTTCTGCCCTTTCTCTAATTCTTT-3' and (V13'F) 5'-TAGGAAATTGGTAAGACAAAGAAAAATAGACTGG-3' and (V13'R) 5'-GCTAGAGATTTTCCACACTACCAAAATGG-3'. PCR primers for variant 2 are (V15'F) 5'-TGGAAGGGTTAGTTTACTCCAAGAAAAGG-3' and (V15'R) 5'-

CATGGTGTTGGTATTATTGCTGGTAGCAG-3' and (V13'F) 5'-GGCTCCATAGCTTAGGGCAACATATCTATA-3' and (V13'R) 5'-GCTAGAGATTTTCCACACTACCAAAATGG-3'. PCR was performed under the following conditions: 1 cycle of 99°C for 1 min; 35 cycles of a 99°C for 8 s, 65°C for 30 s, and 72°C extension for 4 min; followed by a final extension of 72°C for 10 min. Correctly sized fragments (~6kb for 5' and ~4kb for 3') were identified by gel electrophoresis. An adenine over-hang was added to each purified fragment using *Taq* polymerase with 1x buffer (Promega) and 0.2mM each deoxynucleoside triphosphate incubated at 94°C for 2 min followed by a single extension of 72°C for 10 min. Each half genome was then independently T/A cloned into the TOPO-XL vector (Invitrogen) and transformed into XL2-Blue MRF' competent bacteria (Stratagene). Bacteria were plated on LB agar plates supplemented with 50 ug/ml of kanamycin and cultured overnight at 30°C. Single colonies were selected and grown overnight in liquid LB broth at 30°C with 225 rpm shaking followed by plasmid isolation. Each molecular clone was sequence confirmed to be identical to the transmitted viral sequence for each lineage. The 5' clone for each lineage was ligated to the cognate 3'-XL vector by using *Pac*I and *Mlu*I (variant 1) or *Not*I (variant 2) restriction digestion and ligation (New England BioLabs) thereby generating the infectious clones pXL247Fv1 and pXL247Fv2.

Virus phenotypic analysis: The ability of cloned viral genomes to express replication competent virus was assessed as previously described using 293T cells for DNA

transfection and human JC53BL-13 cells (NIH AIDS Research and Reference Reagent Program catalogue #8129, TZM-bl), a HeLa-derived line which has been genetically-modified so as to constitutively express CD4, CCR5 and CXCR4, as a target cell to assess virus entry. JC53BL-13 cells contain integrated luciferase and β-galactosidase (β-gal) genes under tight regulatory control of an HIV-1 LTR and thus virus entry can be quantitatively assessed over a broad range (23). It has been used extensively in the analysis of anti-HIV-1 neutralizing antibodies (20-22). $7x10^3$ JC53BL-13 cells were plated in 96-well tissue culture plates (Falcon) and cultured overnight in DMEM supplemented with 10% fetal calf serum (FCS). For analysis of virus entry, 293T derived virions were quantified by p24Ag or RT activity and assessed directly for infectivity. For analysis of virus neutralization, 3,000 infectious units of virus were combined in a total volume of 60 µl with or without a 2X concentration of sCD4 in DMEM with 6% FCS and 80 µg/ml DEAE-dextran. After 1 hr at 37°C, an equal volume of test or control plasma (10% vol/vol in DMEM plus 6% FCS or five-fold dilutions thereof), monoclonal antibody, fusion inhibitor, or chemokine coreceptor inhibitor was added. Monoclonal antibodies as described (2) were kindly provided by the following individuals: Dennis Burton provided b12 and 2G12; Michael Zwick and Dennis Burton provided Z13e1; Herman Katinger provided 2F5 and 4E10; Susan Zolla-Pazner provided 447-52D; Lisa Cavacini provided F425-B4e8, James Robinson provided 17b; and David Montefiori provided HIVIG. The following reagents were obtained commercially: soluble CD4 (R&D systems, 514-CD); T1249 (Triangle Pharmaceuticals); and anti-CD4 monoclonal antibody (BD Pharmingen, 555344). The coreceptor inhibitors TAK779 and AMD3100 were obtained from the NIH AIDS Research and Reference Reagent Program (4983 and

8128). The addition of ligand or antibody This brought the final concentration of DEAE dextran to 40 μg/ml. When sCD4 was used to trigger a conformation change in gp120 prior to cell attachment (24), the concentration was chosen so that the final 1X concentration after the addition of test antibody corresponded to the IC_{50} of sCD4 specific for each virus. The virus + sCD4 + test antibody mixture was incubated for 1 hr at 37°C. Media was removed entirely from the adherent JC53BL-13 monolayer just before the addition of the virus + sCD4 + test antibody to it. Cells were incubated at 37°C for 2 days and then analyzed for luciferase expression. Controls included cells exposed to no virus and to virus pretreated with NHP or control monoclonal antibodies only. Relative infectivity was calculated by dividing the number of luciferase units at each dilution of test plasma or monoclonal antibodies by values in wells containing NHP but no test plasma or monoclonal antibodies. Neutralization was assessed by 50% inhibitory concentration (IC_{50}) determined by linear regression using a least-squares method. All samples were tested in duplicate and all experiments repeated at least three times to ensure reproducibility.

Virus replication was assessed in activated $CD4^+$ T cells and in monocyte-derived macrophages each obtained from the same normal human donors. Briefly, peripheral blood was collected into ACD-A anticoagulant by venipuncture and peripheral blood mononuclear cells (PBMCs) purified by standard density gradient methods. For $CD4^+$ T cell isolation, pbmcs were incubated with human anti-CD4 coated magnetic beads (Militenyi Biotek, Auburn, CA). $CD4^+$ T cells were positively selected using an AUTOMACS cell separator (Militenyi Biotek, Auburn, CA) and then incubated in XXcm polystyrene tissue culture plates for 2 hours at 37C in HBSS with 10 mM Ca and Mg to

remove adherent monocytes. Nonadherent cells collected, washed in XXX medium, and incubated in YYY medium with 3 ug/ml *Staphyloccocal enterotoxin B* (Sigma-Adridge, St. Louis, MO) for 48 hours at 37C to activate the lymphocytes. 5 x 10^5 cells were incubated with 50,000 IU of virus overnight at 37C in XXX medium. Cells were washed three times and plated in 24 well polystyrene tissue culture plates in a volume of 2 ml RPMI with 15% FBS and 30 units IL-2 / ml. 60ul media was removed for day 1 p24/RT baseline analysis. Every three days, 60ul media was removed and frozen for p24/RT analysis and half of the media was removed from each well and replaced with fresh media. For monocyte-derived macrophage isolation, 3 x 10^6 PBMCs per well were plated in 24 well plates in HBSS plus 10 mM Ca and Mg plus 10% human AB serum (Sigma-Aldrich, St. Louis, MO) and incubated at 37C for 2 hours. Nonadherent cells were removed and DMEM with 10% GCT media (Irvine Scientific, Santa Ana, CA) plus 10% human AB serum was added to wells with 5 units/ml rhMCSF (R&D Systems, Minneapolis, MN). After 3 days incubation, wells were washed vigorously with PBS three times and media containing DMEM with 10% GCT, 10% FBS and 5 units rhMSCF was added to wells. After 3 more days incubation, media was removed and macrophages were incubated with 100,000 IU of virus in XXX ul per well. After a 2 hour incubation, 500ul media was added per well. After overnight incubation, each well was washed three times, 1.5ml media was added and 60ul media was removed for day 1 p24/RT baseline analysis. Every three days, 60ul supernatant frozen for p24/RT analysis while the remainder of the media was removed from each well and replaced with 1.5 ml fresh media. Cultures were continued for 16 days.

ACKNOWLEDGMENTS. This work was supported by the Center for HIV/AIDS Vaccine Immunology and by grants from the NIH (AI067854, AI061734, AI27767) and the Bill and Melinda Gates Foundation (#37874). We thank D. Burton, M. Zwick, J. Mascola, S. Zolla-Pazner, H. Katinger, L. Cavacini and J. Robinson for monoclonal antibodies; D. McPherson, Y. Chen and B. Cochran for technical assistance; clinical cores of the CHAVI and UAB CFAR; and J. White for manuscript preparation.

REFERENCES

FIGURE LEGENDS

Figure 1. Maximum likelihood phylogenetic tree of full-length HIV-1 genome sequences from 12 subjects. Individual sequences are shown as filled ovals. HIV-1 clade B and C reference sequences from the LANL database are shown in gray. Numerals at nodes indicate phylogenetic support for sequences included at that node: maximum likelihood bootstrap support $\geq 70\%$ are shown in italics. Bayesian posterior probability ≥ 0.9 are shown in bold. Scale bar indicates 5% genetic distance (5 nucleotide differences per 100 sites).

Figure 2. Phylogenetic and *Highlighter* analyses of HIV-1 sequences from subject WITO4160. WITO_CON is the inferred transmitted/founder sequence. In the phylogram (left), asterisks indicate sequences that had IUPAC ambiguous state assignments due to mixed bases, which were interpreted to correspond to the consensus sequence (see text). Scale bar indicates 0.0001 (0.01%) genetic distance. In the *Highlighter* plots (center and right), nucleotide differences from the transmitted/founder consensus sequence are indicated by tic marks. The horizontal axis indicates base

position in the alignment. Brackets represent insertions of single bases. Gray boxes indicate deleted sequence. The consensus (CON) is the same whether or not mixed bases are included. The center *Highlighter* plot shows sequences without ambiguous (mixed) base assignments. The *Highlighter* plot on the right includes sequences with one or more IUPAC ambiguous bases (shown as dark blue tic marks). Sequences H3 and C3 contained APOBEC related G-to-A hypermutation.

Figure 3. Phylogenetic and *Highlighter* analyses of HIV-1 sequences from subject ZM247F. See legend to Figure 2. Subject ZM247F had sequences comprising two distinct lineages that were 2.4% different from each other shown by the phylogram (left) and *Highlighter* plot (right). ZM247F_CON1 is the consensus of lineage 1, the predominant lineage sampled at enrollment. The second lineage (sequences B8, E10, E11, and G11) differs from ZM427F_CON1 by 213 to 217 differences, indicated by the numerous tic marks shown for these sequences.

Figure 4. Phylogenetic and *Highlighter* analyses of HIV-1 sequences from subjects 700010040 (CH40) and 700010077 (CH77). See legend to Figure 2. In *Highlighter* plots (B and D), the boxed tic marks indicate differences shared among multiple sequences that are reflected by branches between internal nodes in the phylograms (A and C). In panel C and D, sequences are rooted on sequence A2, not a consensus of sequences. This is because additional analyses (see text and Figure 5) demonstrated that sequence A2 was identical to the transmitted/founder virus and sequences D5, A1, B3, C4, C9, C7 and C3 evolved progeny of sequence A2.

Figure 5. *Highlighter* plots of sequences from longitudinal samples from subjects 700010077 (CH77), 700010040 (CH40) and 700010058 (CH58). The dark line (labeled T/F) indicates the transmitted/founder sequence for each subject. Sequences from different sample dates are indicated by the grouped horizontal lines, with tic marks indicating differences from the T/F sequence. The screening sample (S) was in each case obtained from a preseroconversion Fiebig stage II sample near peak viremia. The sampling times in days post-screening and corresponding viral loads are indicated on the left. Light blue tic marks indicate differences in noncoding LTR regions of the genomes.

Gray tic marks and boxes indicate deletions. Green tic marks indicate differences that are synonymous in all reading frames. Red tic marks indicate differences that are nonsynonymous in any coding gene. Short horizontal lines labeled A, B and C indicate short (~ 1kb) sequences from a screening sample used to verify the presence or absence of a shared polymorphism (see supplemental figures S3-S6). The horizontal axis indicates base position in the alignment relative to HXB2 beginning with the U5 region of the 5' LTR and extending to the end of R in the 3' LTR. The heavy lines at the bottom of the figure show the location of protein coding regions.

Figure 6. Cloning strategy and functional analysis of 3 subtype C full-length molecular clones from subjects ZM246F and ZM247F. (A) The full-length transmitted/founder proviral sequence for ZM246F was chemically synthesized in three overlapping fragments. The two transmitted/founder proviruses in subjected ZM247F were PCR amplified from PBMC DNA in two overlapping halves. Cloned fragments for all proviruses were cloned into the indicated plasmid vectors. (B) Plasmid DNAs were transfected into 293T cells and progeny virus examined for the ability to replicate in activated CD4+ lymphocytes and monocyte-derived macrophages from the same donor. Culture supernatants were analyzed at day 1 through 16 for p24 production.

Legends for Supplemental Figures

Figure S1. Phylogenetic and *Highlighter* analyses of full-length HIV-1 sequences from subjects SUMA0847 (A), TRJO4551 (B), 04013396-0 (C) and ZM249M (D). The sequences from these subjects showed no shared mutations. Sequences with mixed bases leading to ambiguous base calls, but which could be inferred to represent consensus sequence, are shown by asterisks in the phylograms and dark blue tic marks in the *Highlighter* plots. Gray tic marks or boxes indicate deletions. Brackets indicate insertions of single bases. Scale bars indicate 0.0001 (0.01%) genetic distance (1 nucleotide difference) in the phylograms. The horizontal axis for the *Highlighter* plot indicates base position in the alignment.

Figure S2. Phylogenetic and *Highlighter* analyses of HIV-1 sequences from subjects ZM246F (A), 700010058 (CH58) (B), 04013226-2 (C) and WEAU0575 (D). See legend to Figure S1. Sequences from the subjects in this figure showed shared mutations, indicated by boxed tic marks. These are reflected by branches between internal nodes in the phylograms.

Figure S3. Alignments of polymorphic regions in *env* (left) and *nef* (right) in early samples from 700010040 (CH40). The single letter code for the inferred amino acid sequence is shown above the inferred transmitted/founder sequence (CON). CH40S_Lx and Mx sequences are from the screening sample. CH40E_flxx sequences are from full-length sequences obtained the from the enrollment sample. CH40d45xx, CH40d111xx, CH40d181xx, and CH40d412xx are from 3´ half genome sequences from samples taken 45, 111, 181, and 412 days post-screening, respectively.

Figure S4. Alignments of polymorphic regions in *tat* (left) and *env* (right) in early samples from 700010077 (CH77). The single letter code for the inferred amino acid sequence is shown above the inferred transmitted/founder sequence (CON). CH77S_xx sequences are from the screening sample. CH77E_flxx sequences are from full-length sequences obtained the from the enrollment sample. CH77d32xx, CH77d102xx, and CH77d159xx are from 3´ half-genome sequences from samples taken 32, 102, and 159 days post-screening, respectively.

Figure S5. Alignments of polymorphic regions in *pol* in early samples from 700010077 (CH77). The single letter code for the inferred amino acid sequence is shown above the inferred transmitted/founder sequence (CON). CH77S_xx sequences are from the screening sample. CH77E_flxx sequences are from full-length sequences obtained the from the enrollment sample. CH77d32xx, CH77d102xx, and CH77d159xx are from 5´ half-genome sequences from samples taken 32, 102, and 159 days post-screening, respectively.

Figure S6. Alignments of polymorphic regions in *pol* in early samples from 700010058 (CH58). The single letter code for the inferred amino acid sequence is shown above the inferred transmitted/founder sequence (CON). CH58S_xx sequences are from the screening sample. CH58E_flxx sequences are from full-length sequences obtained the from the enrollment sample. CH58d45xx, CH58d85xx, CH58d154xx, and CH77d350xx are from 5´ half-genome sequences from samples taken 45, 85, 154, and 350 days post-screening, respectively.

Figure S7. *Highlighter* plots for sequences from longitudinal samples from subject 700010077 (CH77). The dark bar (labeled T/F) indicates the transmitted/founder sequence. The grouped horizontal lines indicate sequences from different sample dates with tic marks indicating differences from the T/F sequence. The sampling times in days post-screening and corresponding viral loads are indicated on the left. Light blue tic marks indicate differences in noncoding regions of the genomes. Gray tic marks and boxes indicate single base deletions or larger regions of deleted sequence, respectively. Green tic marks indicate differences that are synonymous (silent with respect to amino acid sequence) in all reading frames. Red tic marks indicate differences that are nonsynonymous in any inferred coding gene. Boxes above the T/F sequence line indicate sequences used to infer peptides used for ELISpot assays in Goonetilleke et al. (ref). Vertical lines descending from these show inferred escape. Short horizontal lines labeled A, B and C indicate short (~ 1kb) sequences from an earlier, screening sample (S) used to verify the presence or absence of each shared polymorphism (see supplemental figures S4-S5). Vertical arrows indicate inferred selection in the absence of demonstrated CTL activity. The frequency plot indicates the relative frequency of noncoding (blue), synonymous (in all reading frames) (green) and nonsynonymous (in any inferred coding gene) (red) differences in a window of 27 nucleotides. The horizontal axis indicates base position in the alignment relative to HXB2. The heavy lines at the bottom of the figure show the location of protein coding regions.

Figure S8. *Highlighter* plots for sequences from longitudinal samples from subject 700010040 (CH40). The dark bar (labeled T/F) indicates the transmitted/founder

sequence. The grouped horizontal lines indicate sequences from different sample dates with tic marks indicating differences from the T/F sequence. The sampling times in days post-screening and corresponding viral loads are indicated on the left. Light blue tic marks indicate differences in noncoding regions of the genomes. Gray tic marks and boxes indicate single base deletions or larger regions of deleted sequence, respectively. Green tic marks indicate differences that are synonymous (silent with respect to amino acid sequence) in all reading frames. Red tic marks indicate differences that are nonsynonymous in any inferred coding gene. Boxes above the T/F sequence line indicate sequences used to infer peptides used for ELISpot assays in Goonetilleke et al. (ref). Vertical lines descending from these show inferred escape. Short horizontal lines labeled A and B indicate short (~ 1kb) sequences from an earlier, screening sample (S) used to verify the presence or absence of each shared polymorphism (see supplemental figure S3). Vertical arrows indicate inferred selection in the absence of demonstrated CTL activity. The frequency plot indicates the relative frequency of noncoding (blue), synonymous (in all reading frames) (green) and nonsynonymous (in any inferred coding gene) (red) differences in a window of 27 nucleotides. The horizontal axis indicates base position in the alignment relative to HXB2. The heavy lines at the bottom of the figure show the location of protein coding regions.

Figure S9. *Highlighter* plots for sequences from longitudinal samples from subject 700010058 (CH58). The dark bar (labeled T/F) indicates the transmitted/founder sequence. The grouped horizontal lines indicate sequences from different sample dates with tic marks indicating differences from the T/F sequence. The sampling times in days post-screening and corresponding viral loads are indicated on the left. Light blue tic marks indicate differences in noncoding regions of the genomes. Gray tic marks and boxes indicate single base deletions or larger regions of deleted sequence, respectively. Green tic marks indicate differences that are synonymous (silent with respect to amino acid sequence) in all reading frames. Red tic marks indicate differences that are nonsynonymous in any inferred coding gene. Boxes above the T/F sequence line indicate sequences used to infer peptides used for ELISpot assays in Goonetilleke et al. (ref). Vertical lines descending from these show inferred escape. Short horizontal lines labeled

A indicate short (~ 1kb) sequences from an earlier, screening sample (S) used to verify the presence or absence of each shared polymorphism (see supplemental figure S6). Vertical arrows indicate inferred selection in the absence of demonstrated CTL activity. The frequency plot indicates the relative frequency of noncoding (blue), synonymous (in all reading frames) (green) and nonsynonymous (in any inferred coding gene) (red) differences in a window of 27 nucleotides. The horizontal axis indicates base position in the alignment relative to HXB2. The heavy lines at the bottom of the figure show the location of protein coding regions.

Additional References:

(Maddison and Maddison 2001)

Table 1. Subject Demographics, Risk Group, and Baseline Laboratory Data.

Subject	HIV-1 subtype	Geographic location	Gender	Risk group	Sampling date	VL	Clinical specimen	EIA	WB	Fiebig stage
WITO4160	B	Alabama	M	heterosexual	08.04.00	325,063	plasma	neg	neg	II
SUMA0874	B	Alabama	M	MSM	05.13.91	939,260	plasma	neg	neg	II
WEAU0575	B	Alabama	M	MSM	05.30.90	216,415	plasma	neg	neg	II
TRJO4551	B	Alabama	M	MSM	10.10.01	8,121,951	plasma	neg	neg	II
04013226-2	B	New York	M	MSM	11.20.02	26,700,000	plasma	neg	neg	II
04013396-0	B	New York	M	MSM	08.16.05	1,600,000	plasma	pos	Ind	IV
700010040.S	B	North Carolina	M	MSM	07.11.06	2,197,248	serum	neg	neg	II
700010040.E					07.27.06	298,026	plasma	pos	pos (p31-)	V
700010058.S	B	North Carolina	M	MSM	08.22.06	92,581	serum	neg	neg	II
700010058.E					08.31.06	394,649	plasma	neg	neg	III
700010077.S	B	North Carolina	M	MSM	08.25.06	3,565,728	serum	neg	neg	II
700010077.E					09.08.06	144,145	plasma	pos	pos (p31-)	V
ZM246F	C	Zambia	F	heterosexual	01.14.03	10,013,800	plasma	neg	neg	II
ZM249M	C	Zambia	M	heterosexual	08.05.03	>2,000,000	plasma	Pos	Ind	IV
ZM247F	C	Zambia	F	heterosexual	10.28.03	10,823,500	plasma	Neg	Neg	II

Table 2. Diversity Analysis of Full-Length HIV-1 Genomes Derived from Patients with Primary Infection.

Subject	Fiebig stage	Total number of HIV-1 genomes	Maximum nt length of viral genome	Minimum nt length of viral genome	Number of genomes analyzed[a]	Nucleotide sequence diversity, mean %	Nucleotide sequence diversity, range %	Maximum HD	Poisson Estimated days since MRCA[c]	Lambda[d]	Goodness of fit p value	HD fit to Poisson	Star phylogeny	Deviation from model of random evolution	number of transmitted viruses: env-only[e]	number of transmitted viruses: Full-length genome
WITO4160	II	18	9027	7698	14	0.04	0.00 - 0.09	8	17 (12, 22)	3.570	0.923	yes	yes		1	1
SUMA0874	II	6	9033	9020	6	0.03	0.01 - 0.04	4	11 (7, 15)	2.330	0.546	yes	yes		1	1
WEAU0575	II	9	9028	9022	9	0.03	0.00 - 0.08	7	14 (8, 21)	3.060	0.490	yes	no	early, stochastic	1	1
TRJO4551	II	7	9051	9046	7	0.06	0.03 - 0.07	7	20 (14, 26)	4.290	0.836	yes	yes		1	1
04013226-2	II	9	9068	9038	8	0.08	0.04 - 0.11	10	33 (28, 38)	7.040	0.269	yes	no	early, stochastic	1	1
04013396-0	IV	5	9049	9048	4	0.03	0.00 - 0.06	5	12 (7, 17)	2.500	0.704	yes	yes		1	1
700010040.S	II	26[b]	3593	2977	26	0.04	0.00 - 0.11	4	16 (12, 21)	1.372	0.251	yes	yes		1	1
700010040.E	V	9	8990	8989	9	0.05	0.01 - 0.10	9	21 (14, 28)	4.440	0.460	yes	no	CTL selection	1	1
700010058.E	III	7	9027	8717	7	0.06	0.02 - 0.10	9	24 (14, 33)	5.050	0.941	yes	no	early, stochastic	1	1
700010077.S	II	23[b]	2785	2785	23	0.05	0.00 - 0.14	4	20 (13,26)	1.300	0.449	yes	yes		1	1
700010077.E	V	8	9048	9037	8	0.06	0.02 - 0.13	12	27 (16, 37)	5.750	0.460	yes	no	CTL selection	1	1
ZM246F	II	10	8978	8977	10	0.04	0.00 - 0.09	8	17 (10, 24)	3.560	0.880	yes	no	early, stochastic	1	1
ZM249M	IV	7	9054	8735	7	0.03	0.01 - 0.06	5	15 (9, 20)	3.100	0.671	yes	yes		1	1
ZM247F	II	13	9040	7957	12	1.19	0.01 - 2.46	220	493 (314, 673)	106.197	0	no	no	2 strains	2	2
ZM247F lng 1[f]	II	9	9040	7957	8	0.05	0.01 - 0.09	8	21 (14, 28)	4.500	0.967	yes	yes		1	1
ZM247F lng 2[f]	II	4	9034	9033	4	0.05	0.02 - 0.08	7	21 (14, 28)	4.500	0.939	yes	yes		1	1

[a]Genomes with multiple G-to-A or large deletions were excluded from model analysis.
[b]SGA sequences representing 3'half subgenomic fragments from the screening sample.
[c]Predicted minimum number of days needed to achieve observed within-patient diversity.
[d]Mean of best fitting Poisson distribution. Lambda is the free parameter that defines a Poisson.
[e]Keele B (PNAS 2008); Salazar-Gonzalez J (JV 2008); Li H (2009).
[f]lng 1= lineage 1; lng 2= lineage 2.

Table 3. Nucleotide Gene Diversity in Four Protein-Coding Regions in Transmitted/Early Founder Viruses.

		gag					pol				env				nef			SGA env-only					
Subject	Fiebig stage	Number of sequences	Maximum HD	Poisson Estimated days since MRCA	Lambda	Goodness of fit p-value	Maximum HD	Poisson estimated days since MRCA	Lambda	Goodness of fit p-value	Maximum HD	Poisson estimated days since MRCA	Lambda	Goodness of fit p-value	Maximum HD	Poisson estimated days since MRCA	Lambda	Goodness of fit p-value	Total number of env	Maximum HD	Poisson estimated days since MRCA	Lambda	Goodness of fit value	Conform a model of random evolution
WITO4160	II	14	2	20 (10, 30)	0.714	0.348	3	20 (13, 27)	1.429	0.393	5	17 (5, 28)	1.000	0.486	2	19 (0, 38)	0.286	0.671	16	5	24 (12, 36)	1.375	0.940	yes
SUMA0874	II	6	1	10 (-4, 23)	0.333	0.630	2	11 (0, 22)	0.8	0.186	2	11 (1, 21)	0.667	0.608	0	nd	nd	nd	35	2	10 (6, 13)	0.568	0.243	yes, w/o APOBEC3G/F
WEAU0575	II	9	1	6 (-3, 16)	0.222	0.671	2	13 (5, 20)	0.889	0.321	4	21 (1, 41)	1.278	0.041	2	46 (12, 80)	0.667	0.550	44	6	23 (17, 29)	1.340	0.462	yes
TRJO4551	II	7	1	8 (-4, 20)	0.286	0.645	4	28 (15, 41)	2.000	0.694	2	23 (15, 32)	1.429	0.22	0	nd	nd	nd	18	6	36 (22, 49)	2.058	0.902	yes
04013226-2	II	8	3	35 (14, 56)	1.250	0.736	3	25 (16, 33)	1.750	0.277	6	54 (33, 74)	3.286	0.885	2	49 (16, 83)	0.750	0.513	33	9	25 (16, 35)	1.542	0	non-star; Poisson w/o APOBEC3G/F
04013396-0	IV	4	0	nd	nd	nd	2	14 (5, 23)	1.000	0.531	1	8 (-2, 19)	0.500	0.590	0	nd	nd	nd	39	5	16 (11, 21)	0.974	0.897	yes
700010040.S	II	26	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	4	17 (11, 22)	0.986	0.715	2	21 (7, 35)	0.308	0.634	46	14	23 (11, 36)	1.407	0.000	non-star; stochastic
700010040.E	IV	9	1	6 (-3, 16)	0.222	0.671	4	16 (3, 28)	1.111	0.679	4	25 (10, 40)	1.500	0.943	2	64 (34, 94)	0.944	0.280	29	5	23 (17, 30)	1.338	0.867	non-star; stochastic
700010058.E	III	7	3	32 (7, 57)	1.143	0.838	2	15 (7, 22)	1.048	0.282	6	33 (4, 62)	2.000	0.427	1	19 (-10, 48)	0.286	0.645	46	10	36 (28, 44)	2.072	0.004	non-star, not Poisson; CTL pressure
700010077.E	V	8	1	7 (-3, 17)	0.250	0.659	6	37 (20, 54)	2.643	0.965	2	20 (11, 28)	1.179	0.066	2	33 (4, 62)	0.500	0.636	51	6	35 (29, 41)	2.013	0.417	non-star; CTL pressure
ZM246F	II	10	0	nd	nd	nd	3	11 (3, 20)	0.800	0.863	5	29 (14, 45)	1.756	0.912	1	14 (-5, 33)	0.200	0.683	27	5	21 (13, 29)	1.236	0.726	non-star; stochastic
ZM249M	IV	7	0	nd	nd	nd	1	4 (-2, 10)	0.286	0.645	3	18 (4, 32)	1.095	0.789	1	19 (-9, 47)	0.286	0.645	24	7	30 (19, 40)	1.804	0.031	non-star; stochastic
ZM247F Ing 1	II	8	1	7 (-3, 17)	0.250	0.659	3	21 (12, 31)	1.500	0.575	3	16 (4, 20)	1.000	0.853	2	34 (3, 65)	0.500	0.636	18	4	14 (7, 22)	0.882	0.839	yes
ZM247F Ing 2	II	4	2	42 (22, 62)	1.500	0.344	4	28 (18, 38)	2.000	0.769	2	16 (5, 28)	1.000	0.531	0	nd	nd	nd	7	3	14 (0, 28)	0.857	0.803	yes

Phenotype of Transmitted HIV-1 Clade C

	sCD4 nM	HIVIG ug/ml	Clade B Plasma (IC50)	Clade C Plasma (IC50)	17b ug/ml	17b + sCD4 ug/ml	21C ug/ml	21c + sCD4 ug/ml	b12 ug/ml	2G12 ug/ml	2F5 ug/ml	4E10 ug/ml	Z13e1 ug/ml	447-52D ug/ml	447+sCD4 ug/ml	F425-B4e8 ug/ml	F425 + sCD4 ug/ml	T-20 ug/ml	T-1249 ug/ml	antiCD4mab ug/ml	TAK-779 uM	coreceptor usage	Infectivity titer iU/ui
YU-2 FL	13	619	20	20	> 10	> 10	> 10	> 10	7.16	> 50	> 50	> 50	> 50	13.36	1.48	> 25	1.85	0.49	0.094	0.18	0.047	R5	9063
NL4.3 FL	1	60	92	31	0.21	0.53	0.59	0.73	0.01	1.08	0.24	1.05	> 50	0.03	0.01	> 25	> 25	0.47	0.003	0.63	> 10	X4	3400
ADA	105	1000	19	17	> 10	> 10	> 10	> 10	1.31	> 50	33.20	> 50	> 50	> 25	0.03	> 25	0.11	0.15	0.094	0.31	0.307	R5	9479
NL.Bal.ecto	8	113	50	32	> 10	> 10	> 10	> 10	0.11	1.39	12.30	15.10	> 50	0.14	0.04	0.54	0.05	0.09	0.015	0.44	1.040	R5	19200
246F	125	748	13	31	> 10	> 10	> 10	> 10	21.35	> 50	> 50	5.50	> 50	> 25	> 25	> 25	> 25	0.07	0.012	0.17	0.003	R5	3113
247Fv1	155	> 1000	13	18	> 10	> 10	> 10	> 10	> 50	> 50	> 50	11.33	> 50	> 25	> 25	> 25	> 25	0.05	0.011	0.11	0.032	R5	5094
247Fv2	125	> 1000	12	20	> 10	> 10	> 10	> 10	> 50	> 50	> 50	47.11	> 50	> 25	> 25	> 25	> 25	0.08	0.027	0.16	0.034	R5	4038

Reciprocal geometric mean neutralization titers (IC50) of heterologous plasma speciimens from 11 HIV-1 Clade B chronically infected subjects.
Reciprocal geometnc mean neutralization titers (IC50) of heterologous plasma specimens from 22 HIV-1 Clade C chronically infected subjects.

YU-2 FL	0.0583	0.0134
NL4.3 FL	0.0103	0.0091
ADA	0.1000	0.0118
NL.Bal.ecto	0.0184	0.0099
246F	0.1000	0.0104
247Fv1	0.1000	0.0144
247Fv2	0.1000	0.0131

Geometric mean neutralization titers (IC50) of heterologous plasma speciimens from 11 HIV-1 Clade B chronically infected subjects.
Geometric mean neutralization titers (IC50) of heterologous plasma specimens from 22 HIV-1 Clade C chronically infected subjects.

Table S1. Analysis of pn and ps for gag, pol, env and nef.

gag							
	Fiebig Stage	Sd	Sn	ps	pn	pn/ps[a]	pn-ps[b]
04013226-2	II	1	2	0.0005	0.0002	0.5322	-0.0002
04013396-0	IV	0	0	0	0	undef.[c]	0
700010040E	V	0	1	0	0.0001	undef.[c]	*0.0001*
700010058	III	3	1	0.0016	0.0001	0.0891	-0.0014
700010077	V	1	0	0.0005	0	0	-0.0005
SUMA0874	II	1	0	0.0006	0	0	-0.0006
TRJO4551	II	0	0	0	0	undef.[c]	0
WEAU0575	II	0	0	0	0	undef.[c]	0
WITO4160	II	0	3	0	0.0002	undef.[c]	*0.0002*
ZM246F	II	0	0	0	0	undef.[c]	0
ZM247F lng 1	II	1	0	0.0005	0	0	-0.0005
ZM247F lng 2	II	0	1	0	0.0002	undef.[c]	*0.0002*
ZM249M	IV	0	0	0	0	undef.[c]	0
		7	8	0.0003	0.0001	0.3024	-0.0002
pol							
	Fiebig Stage	Sd	Sn	ps	pn	pn/ps[a]	pn-ps[b]
04013226-2	II	2	1	0.0004	0.0001	0.1334	-0.0004
04013396-0	IV	0	2	0	0.0002	undef.[c]	*0.0002*
700010040E	V	2	3	0.0004	0.0002	0.3969	-0.0002
700010058	III	1	3	0.0002	0.0002	0.7927	-5.16E-05
700010077	V	6	6	0.0013	0.0003	0.2641	-0.0010
SUMA0874	II	0	3	0	0.0002	undef.[c]	*0.0002*
TRJO4551	II	2	4	0.0005	0.0003	0.5315	-0.0002
WEAU0575	II	0	1	0	0.0001	undef.[c]	*0.0001*
WITO4160	II	6	2	0.0007	0.0001	0.0884	-0.0007
ZM246F	II	1	3	0.0002	0.0001	0.8015	-3.43E-05
ZM247F lng 1	II	2	2	0.0004	0.0001	0.2633	-0.0003
ZM247F lng 2	II	0	2	0	0.0002	undef.[c]	*0.0002*
ZM249M	IV	0	1	0	0.0001	undef.[c]	*6.57E-05*
		22	33	0.0004	0.0002	0.3978	-0.0002
env							
	Fiebig Stage	Sd	Sn	ps	pn	pn/ps[a]	pn-ps[b]
04013226-2	II	2	8	0.0006	0.0007	*1.0385*	*2.42E-05*
04013396-0	IV	1	0	0.0006	0	0	-0.0006
700010040S	II	3	6	0.0003	0.0002	0.5289	-0.0001
700010040E	V	1	5	0.0003	0.0004	*1.3230*	*0.0001*
700010058	III	0	3	0	0.0003	undef.[c]	*0.0003*
700010077	V	0	2	0	0.0002	undef.[c]	*0.0002*
SUMA0874	II	1	0	0.0004	0	0	-0.0004
TRJO4551	II	2	2	0.0007	0.0002	0.2631	-0.0005
WEAU0575	II	2	4	0.0006	0.0003	0.5160	-0.0003
WITO4160	II	2	2	0.0004	0.0001	0.2630	-0.0003
ZM246F	II	4	5	0.0010	0.0003	0.3311	-0.0007
ZM247F lng 1	II	2	0	0.0006	0	0	-0.0006
ZM247F lng 2	II	0	1	0	0.0002	undef.[c]	*0.0002*
ZM249M	IV	1	3	0.0004	0.0003	0.7815	-0.0001
		18	35	0.0004	0.0002	0.5105	-0.0002
nef							
	Fiebig Stage	Sd	Sn	ps	pn	pn/ps[a]	pn-ps[b]
04013226-2	II	1	2	0.0009	0.0005	0.5541	-0.0004
04013396-0	IV	0	0	0	0	undef.[c]	0
700010040S	II	2	2	0.0006	0.0002	0.2778	-0.0004
700010040E	V	1	4	0.0008	0.0009	*1.1101*	*0.0001*
700010058	III	1	0	0.0011	0	undef.[c]	-0.0011
700010077	V	0	2	0	0.0005	undef.[c]	*0.0005*
SUMA0874	II	0	0	0	0	undef.[c]	0
TRJO4551	II	0	0	0	0	undef.[c]	0
WEAU0575	II	0	3	0	0.0007	undef.[c]	*0.0007*
WITO4160	II	2	0	0.0011	0	undef.[c]	-0.0011
ZM246F	II	0	1	0	0.0002	undef.[c]	*0.0002*
ZM247F lng 1	II	0	2	0	0.0005	undef.[c]	*0.0005*
ZM247F lng 2	II	0	0	0	0	undef.[c]	0
ZM249M	IV	1	0	0.0010	0	0	-0.0010
		6	14	0.0004	0.0003	0.6298	-0.0002

[a]Positive selection by pn/ps (>1) indicated by italics.
[b]Positive selection by pn-ps (>0) indicated by italics.
[c]Undef: undefined due to division by zero.

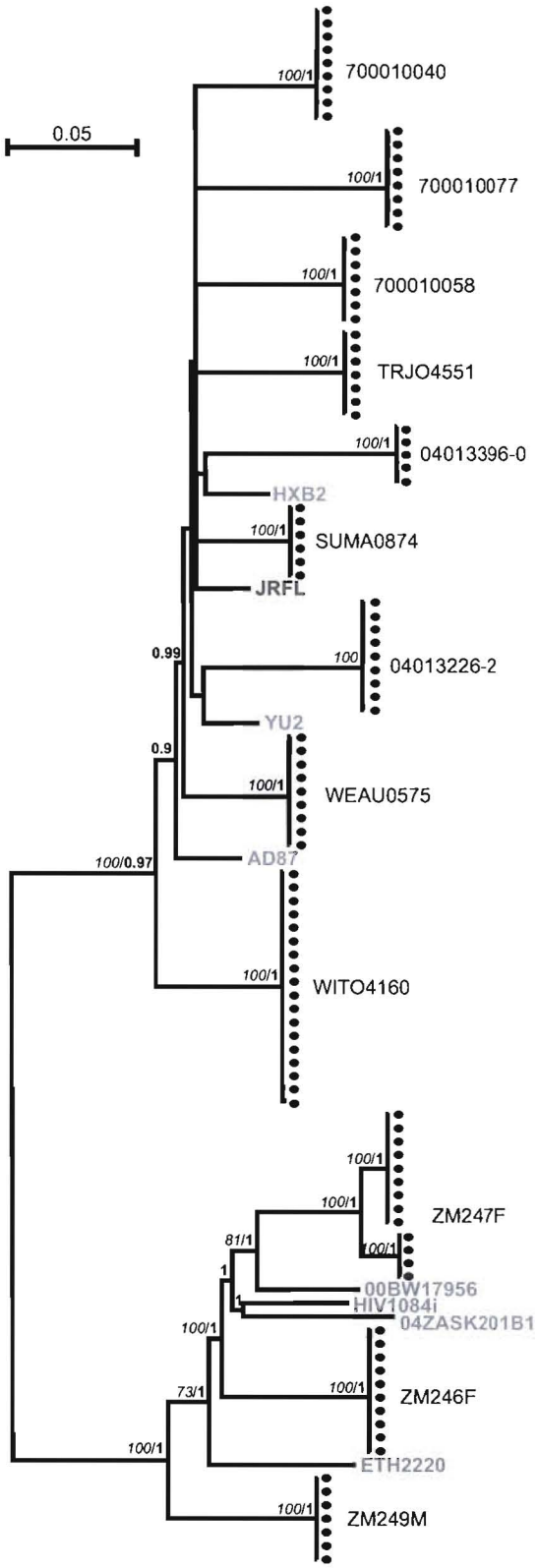
0.05
700010040
700010077
700010058
TRJO4551
04013396-0
HXB2
SUMA0874
JRFL
04013226-2
YU2
WEAU0575
AD87
WITO4160
ZM247F
00BW17956
HIV1084i
04ZASK201B1
ZM246F
ETH2220
ZM249M

Figure 1

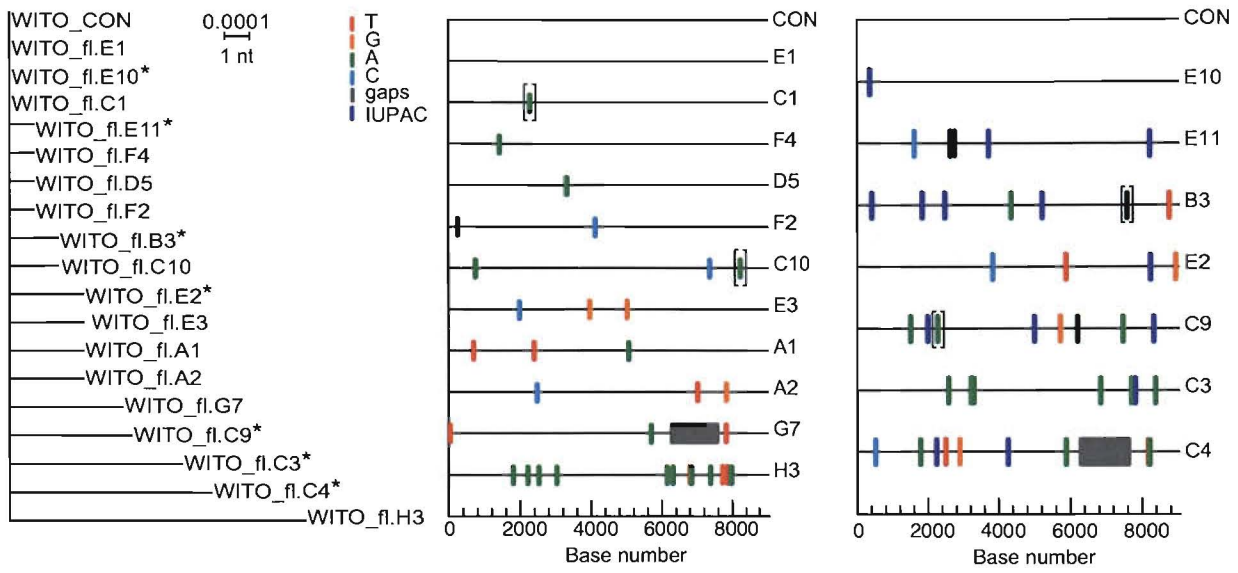
WITO_CON
WITO_fl.E1
WITO_fl.E10*
WITO_fl.C1
WITO_fl.E11*
WITO_fl.F4
WITO_fl.D5
WITO_fl.F2
WITO_fl.B3*
WITO_fl.C10
WITO_fl.E2*
WITO_fl.E3
WITO_fl.A1
WITO_fl.A2
WITO_fl.G7
WITO_fl.C9*
WITO_fl.C3*
WITO_fl.C4*
WITO_fl.H3
0.0001
1 nt
T
G
A
C
gaps
IUPAC
CON
E1
C1
F4
D5
F2
C10
E3
A1
A2
G7
H3
0
2000
4000
6000
8000
Base number
CON
E10
E11
B3
E2
C9
C3
C4
0
2000
4000
6000
8000
Base number

Figure 2

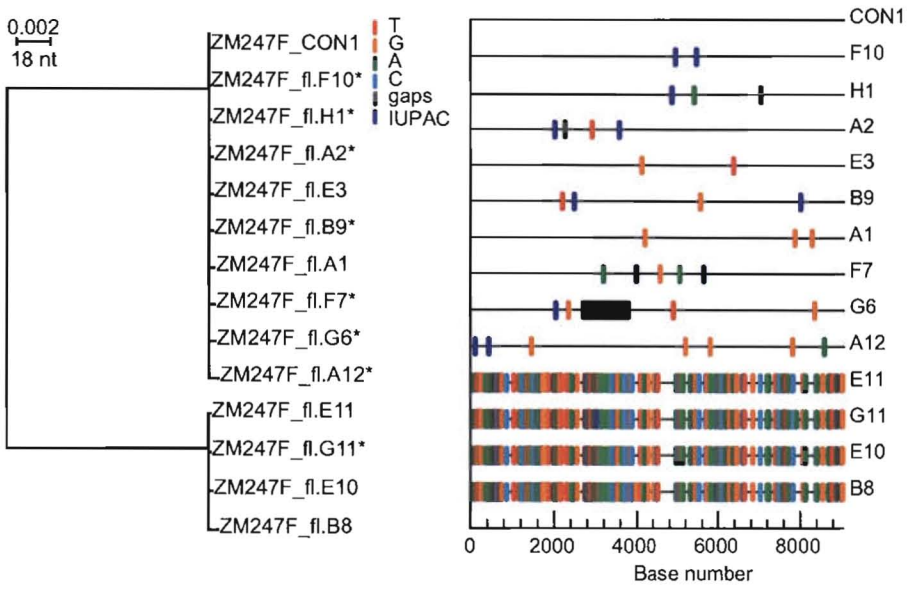
0.002
18 nt
ZM247F_CON1
ZM247F_fl.F10*
ZM247F_fl.H1*
ZM247F_fl.A2*
ZM247F_fl.E3
ZM247F_fl.B9*
ZM247F_fl.A1
ZM247F_fl.F7*
ZM247F_fl.G6*
ZM247F_fl.A12*
ZM247F_fl.E11
ZM247F_fl.G11*
ZM247F_fl.E10
ZM247F_fl.B8
T
G
A
C
gaps
IUPAC
CON1
F10
H1
A2
E3
B9
A1
F7
G6
A12
E11
G11
E10
B8
0
2000
4000
6000
8000
Base number

Figure 3

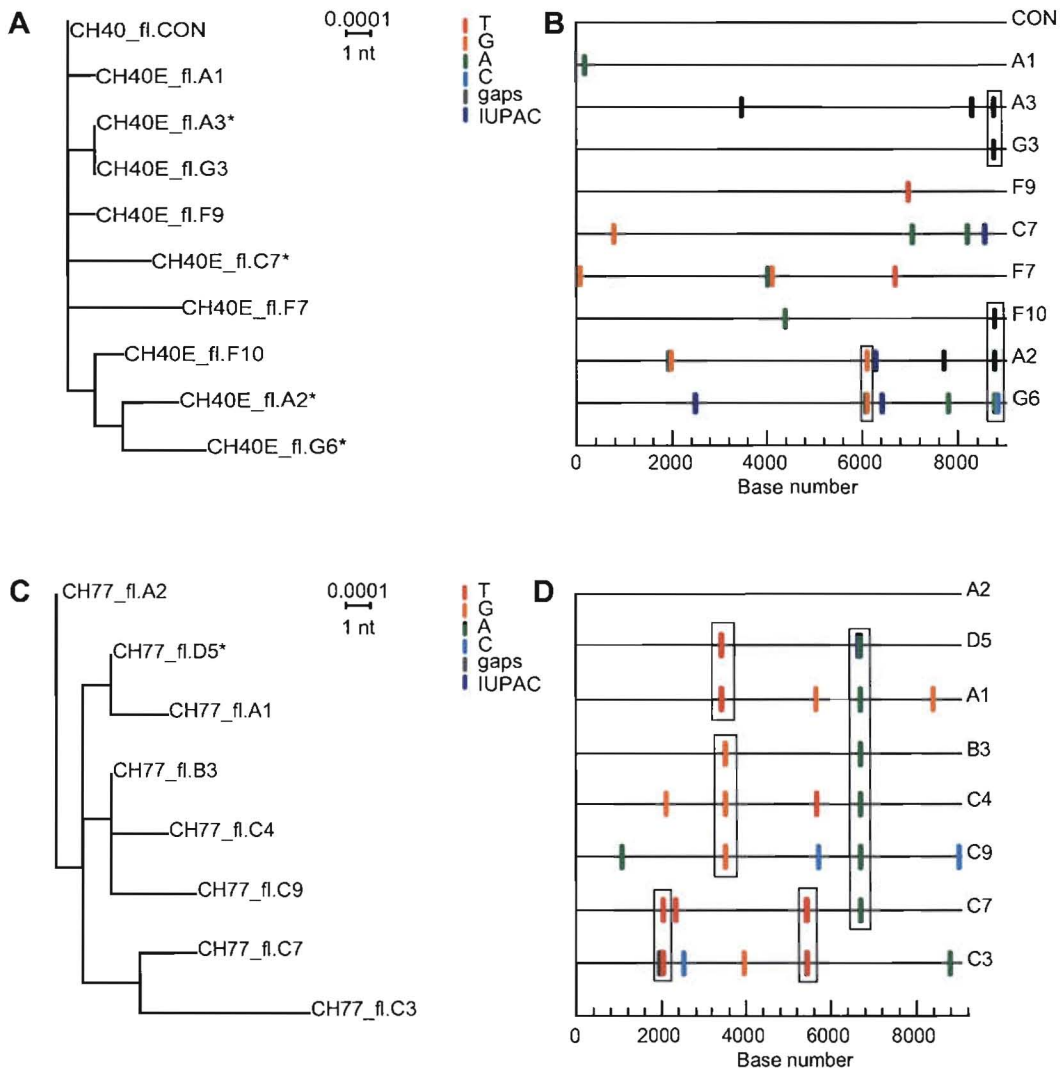
A
CH40_fl.CON
0.0001
1 nt
CH40E_fl.A1
CH40E_fl.A3*
CH40E_fl.G3
CH40E_fl.F9
CH40E_fl.C7*
CH40E_fl.F7
CH40E_fl.F10
CH40E_fl.A2*
CH40E_fl.G6*
T
G
A
C
gaps
IUPAC
B
CON
A1
A3
G3
F9
C7
F7
F10
A2
G6
0
2000
4000
6000
8000
Base number
C
CH77_fl.A2
0.0001
1 nt
CH77_fl.D5*
CH77_fl.A1
CH77_fl.B3
CH77_fl.C4
CH77_fl.C9
CH77_fl.C7
CH77_fl.C3
T
G
A
C
gaps
IUPAC
D
A2
D5
A1
B3
C4
C9
C7
C3
0
2000
4000
6000
8000
Base number

Figure 4

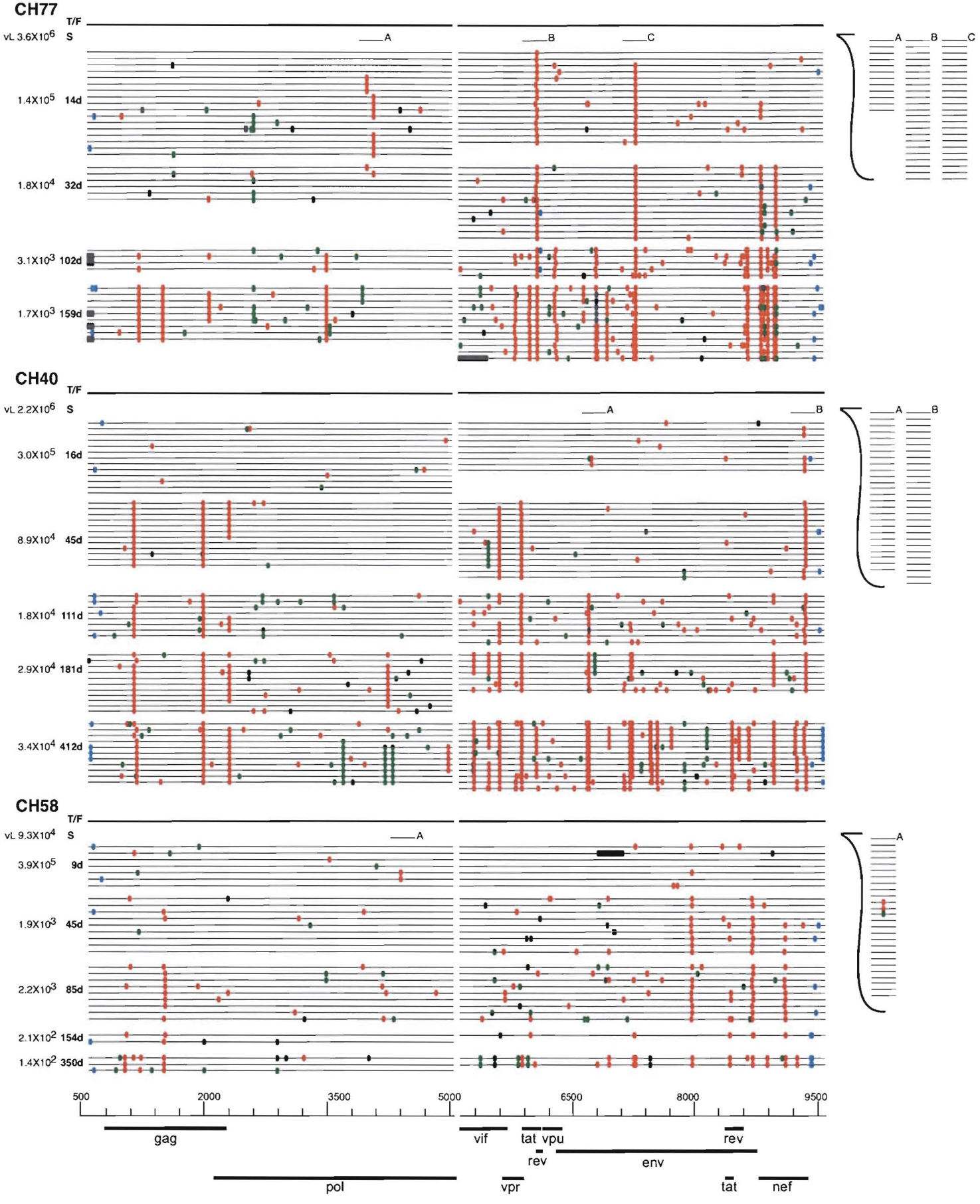
CH77
T/F
vL 3.6X10^6 S
1.4X10^5 14d
1.8X10^4 32d
3.1X10^3 102d
1.7X10^3 159d
A
B
C
CH40
T/F
vL 2.2X10^6 S
3.0X10^5 16d
8.9X10^4 45d
1.8X10^4 111d
2.9X10^4 181d
3.4X10^4 412d
A
B
CH58
T/F
vL 9.3X10^4 S
3.9X10^5 9d
1.9X10^3 45d
2.2X10^3 85d
2.1X10^2 154d
1.4X10^2 350d
A
500
2000
3500
5000
6500
8000
9500
gag
pol
vif
vpr
tat
rev
vpu
rev
env
tat
nef

Figure 5

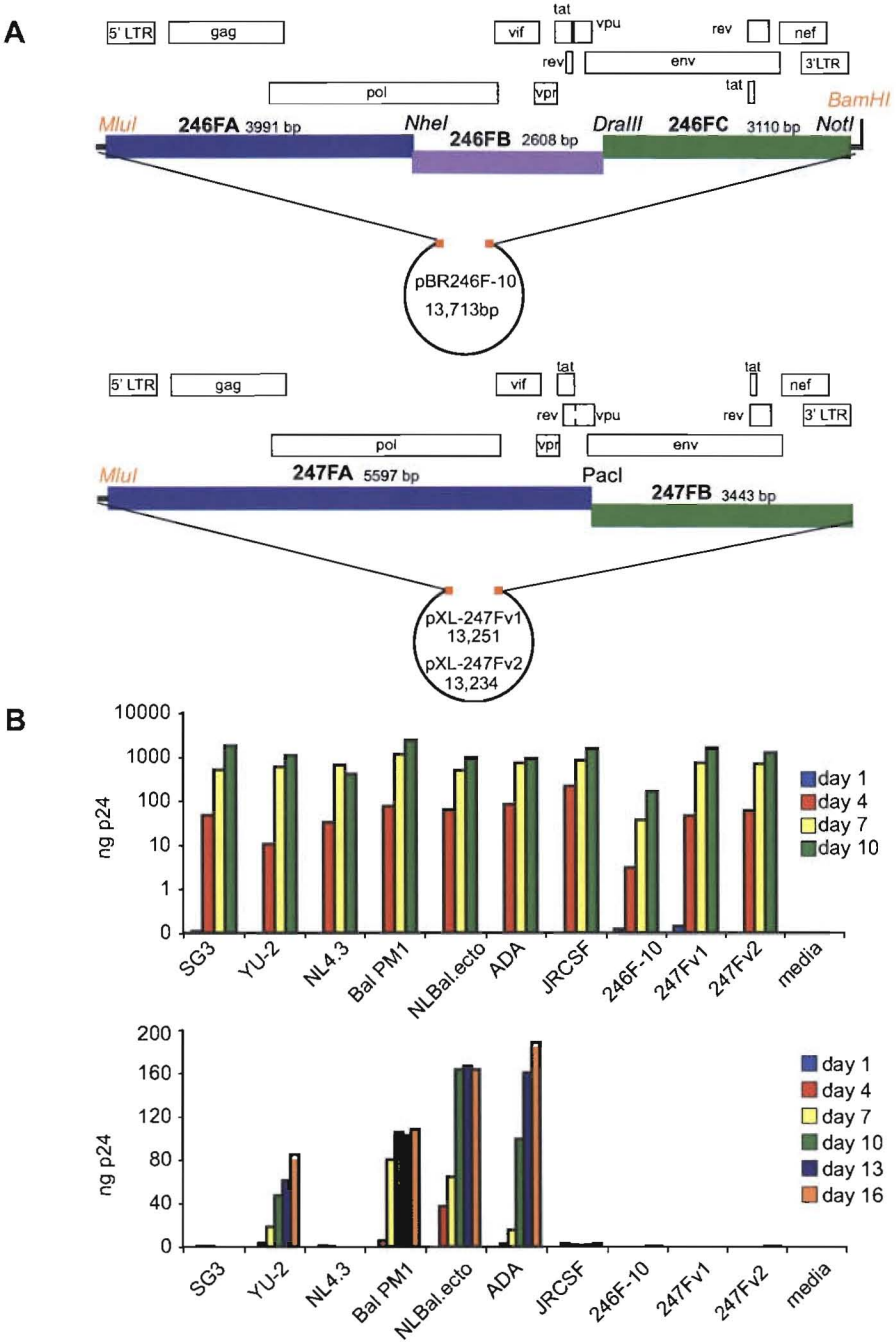
A
5' LTR
gag
pol
vif
tat
vpu
rev
vpr
env
nef
3'LTR
MluI
246FA 3991 bp
NheI
246FB 2608 bp
DraIII
246FC 3110 bp
NotI
BamHI
pBR246F-10
13,713bp
247FA 5597 bp
PacI
247FB 3443 bp
pXL-247Fv1
13,251
pXL-247Fv2
13,234
B
ng p24
day 1
day 4
day 7
day 10
day 13
day 16
SG3
YU-2
NL4.3
Bal PM1
NLBal.ecto
ADA
JRCSF
246F-10
247Fv1
247Fv2
media

Figure 6

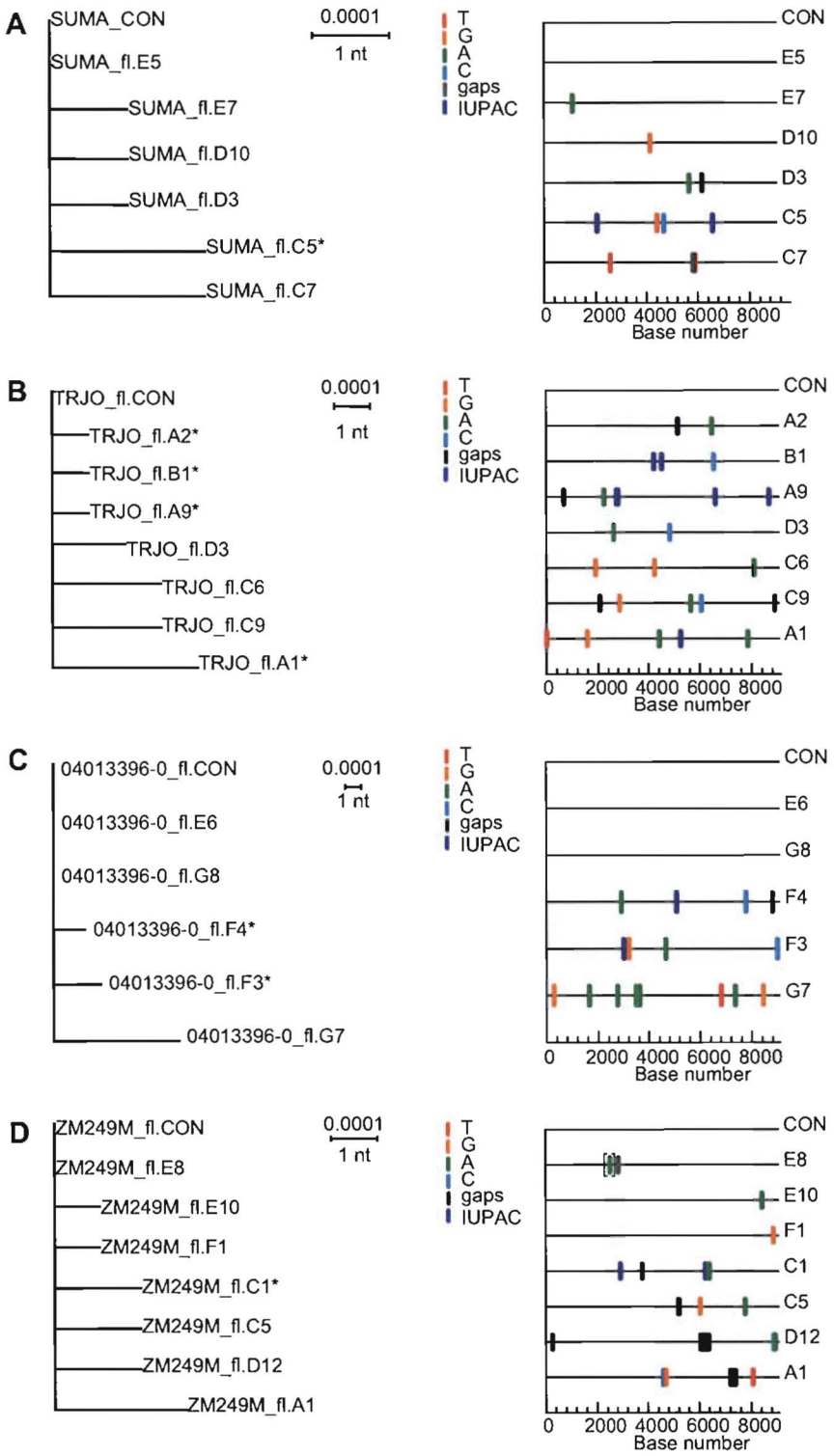
A
SUMA_CON
SUMA_fl.E5
SUMA_fl.E7
SUMA_fl.D10
SUMA_fl.D3
SUMA_fl.C5*
SUMA_fl.C7
0.0001
1 nt
T
G
A
C
gaps
IUPAC
CON
E5
E7
D10
D3
C5
C7
0 2000 4000 6000 8000
Base number
B
TRJO_fl.CON
TRJO_fl.A2*
TRJO_fl.B1*
TRJO_fl.A9*
TRJO_fl.D3
TRJO_fl.C6
TRJO_fl.C9
TRJO_fl.A1*
0.0001
1 nt
T
G
A
C
gaps
IUPAC
CON
A2
B1
A9
D3
C6
C9
A1
0 2000 4000 6000 8000
Base number
C
04013396-0_fl.CON
04013396-0_fl.E6
04013396-0_fl.G8
04013396-0_fl.F4*
04013396-0_fl.F3*
04013396-0_fl.G7
0.0001
1 nt
T
G
A
C
gaps
IUPAC
CON
E6
G8
F4
F3
G7
0 2000 4000 6000 8000
Base number
D
ZM249M_fl.CON
ZM249M_fl.E8
ZM249M_fl.E10
ZM249M_fl.F1
ZM249M_fl.C1*
ZM249M_fl.C5
ZM249M_fl.D12
ZM249M_fl.A1
0.0001
1 nt
T
G
A
C
gaps
IUPAC
CON
E8
E10
F1
C1
C5
D12
A1
0 2000 4000 6000 8000
Base number

Figure S1

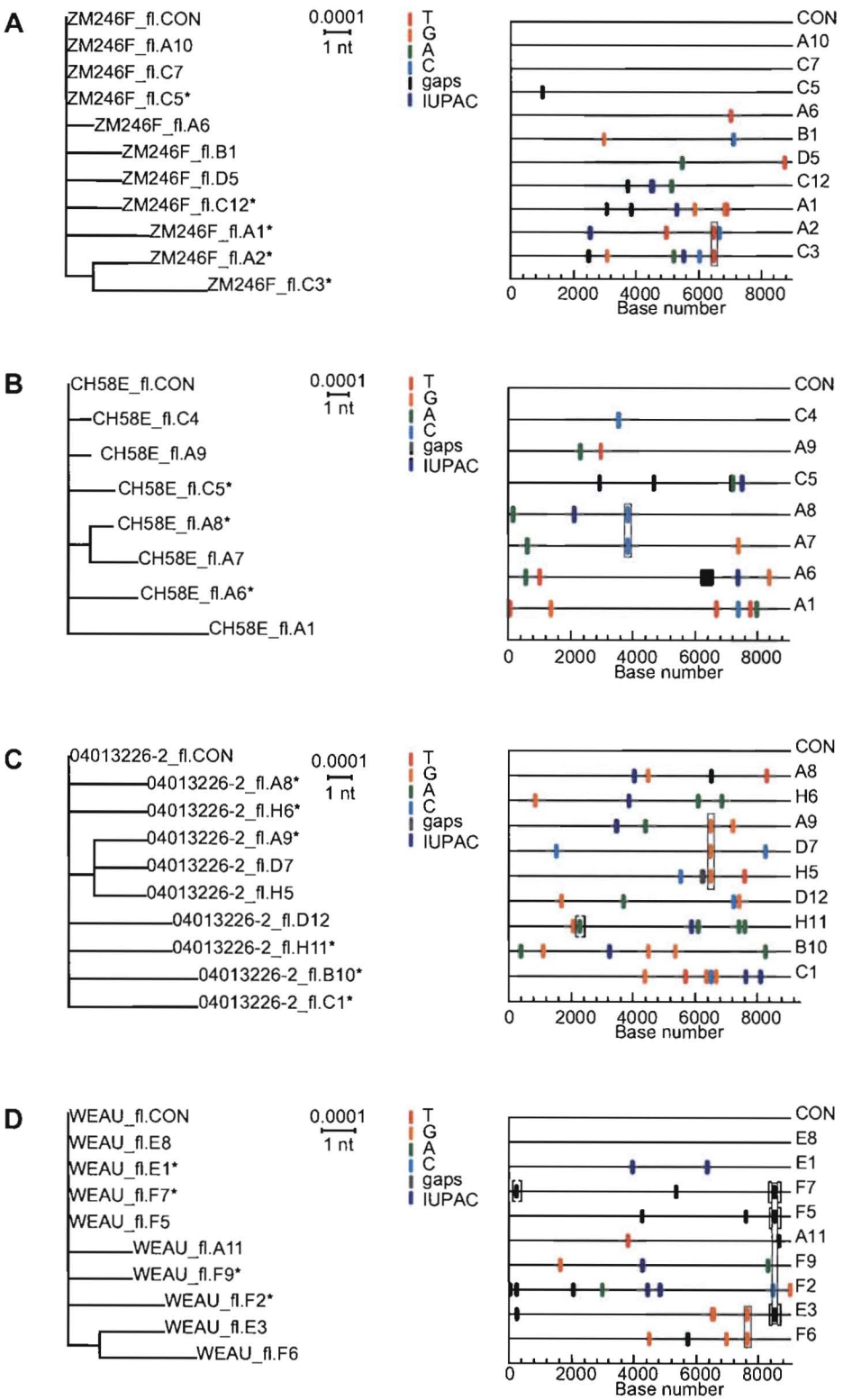
A
ZM246F_fl.CON
ZM246F_fl.A10
ZM246F_fl.C7
ZM246F_fl.C5*
ZM246F_fl.A6
ZM246F_fl.B1
ZM246F_fl.D5
ZM246F_fl.C12*
ZM246F_fl.A1*
ZM246F_fl.A2*
ZM246F_fl.C3*
0.0001
1 nt
T
G
A
C
gaps
IUPAC
CON
A10
C7
C5
A6
B1
D5
C12
A1
A2
C3
0
2000
4000
6000
8000
Base number
B
CH58E_fl.CON
CH58E_fl.C4
CH58E_fl.A9
CH58E_fl.C5*
CH58E_fl.A8*
CH58E_fl.A7
CH58E_fl.A6*
CH58E_fl.A1
0.0001
1 nt
T
G
A
C
gaps
IUPAC
CON
C4
A9
C5
A8
A7
A6
A1
0
2000
4000
6000
8000
Base number
C
04013226-2_fl.CON
04013226-2_fl.A8*
04013226-2_fl.H6*
04013226-2_fl.A9*
04013226-2_fl.D7
04013226-2_fl.H5
04013226-2_fl.D12
04013226-2_fl.H11*
04013226-2_fl.B10*
04013226-2_fl.C1*
0.0001
1 nt
T
G
A
C
gaps
IUPAC
CON
A8
H6
A9
D7
H5
D12
H11
B10
C1
0
2000
4000
6000
8000
Base number
D
WEAU_fl.CON
WEAU_fl.E8
WEAU_fl.E1*
WEAU_fl.F7*
WEAU_fl.F5
WEAU_fl.A11
WEAU_fl.F9*
WEAU_fl.F2*
WEAU_fl.E3
WEAU_fl.F6
0.0001
1 nt
T
G
A
C
gaps
IUPAC
CON
E8
E1
F7
F5
A11
F9
F2
E3
F6
0
2000
4000
6000
8000
Base number

Figure S2

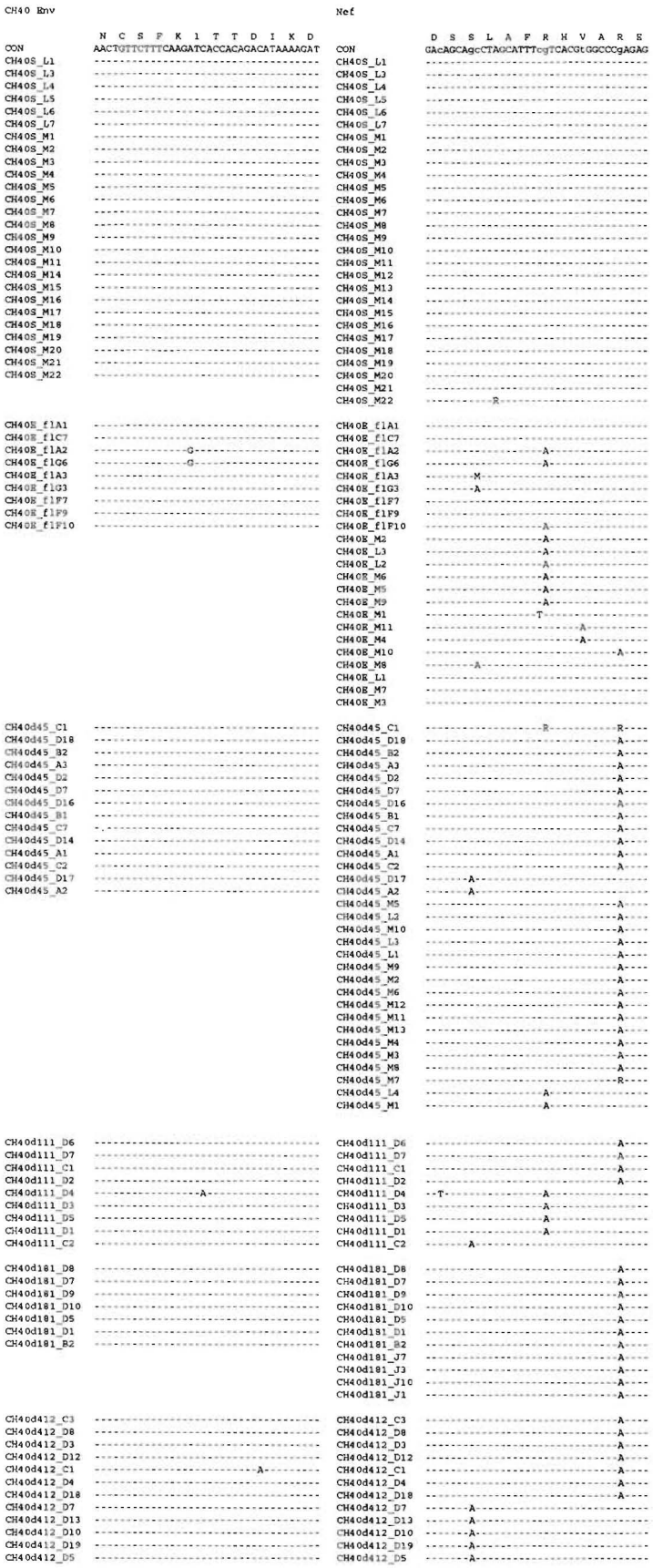
CH40 Env
N C S F K 1 T T D I K D
CON AACTGTTCTTTCAAGATCACCACAGACATAAAAGAT
CH40S_L1
CH40S_L3
CH40S_L4
CH40S_L5
CH40S_L6
CH40S_L7
CH40S_M1
CH40S_M2
CH40S_M3
CH40S_M4
CH40S_M5
CH40S_M6
CH40S_M7
CH40S_M8
CH40S_M9
CH40S_M10
CH40S_M11
CH40S_M14
CH40S_M15
CH40S_M16
CH40S_M17
CH40S_M18
CH40S_M19
CH40S_M20
CH40S_M21
CH40S_M22
CH40E_f1A1
CH40E_f1C7
CH40E_f1A2
CH40E_f1G6
CH40E_f1A3
CH40E_f1G3
CH40E_f1F7
CH40E_f1F9
CH40E_f1F10
CH40d45_C1
CH40d45_D18
CH40d45_B2
CH40d45_A3
CH40d45_D2
CH40d45_D7
CH40d45_D16
CH40d45_B1
CH40d45_C7
CH40d45_D14
CH40d45_A1
CH40d45_C2
CH40d45_D17
CH40d45_A2
CH40d111_D6
CH40d111_D7
CH40d111_C1
CH40d111_D2
CH40d111_D4
CH40d111_D3
CH40d111_D5
CH40d111_D1
CH40d111_C2
CH40d181_D8
CH40d181_D7
CH40d181_D9
CH40d181_D10
CH40d181_D5
CH40d181_D1
CH40d181_B2
CH40d412_C3
CH40d412_D8
CH40d412_D3
CH40d412_D12
CH40d412_C1
CH40d412_D4
CH40d412_D18
CH40d412_D7
CH40d412_D13
CH40d412_D10
CH40d412_D19
CH40d412_D5
Nef
D S S L A F R H V A R E
CON GAcAGCAgcCTAGCATTTcgTCACGtGGCCCgAGAG
CH40S_L1
CH40S_L3
CH40S_L4
CH40S_L5
CH40S_L6
CH40S_L7
CH40S_M1
CH40S_M2
CH40S_M3
CH40S_M4
CH40S_M5
CH40S_M6
CH40S_M7
CH40S_M8
CH40S_M9
CH40S_M10
CH40S_M11
CH40S_M12
CH40S_M13
CH40S_M14
CH40S_M15
CH40S_M16
CH40S_M17
CH40S_M18
CH40S_M19
CH40S_M20
CH40S_M21
CH40S_M22
CH40E_f1A1
CH40E_f1C7
CH40E_f1A2
CH40E_f1G6
CH40E_f1A3
CH40E_f1G3
CH40E_f1F7
CH40E_f1F9
CH40E_f1F10
CH40E_M2
CH40E_L3
CH40E_L2
CH40E_M6
CH40E_M5
CH40E_M9
CH40E_M1
CH40E_M11
CH40E_M4
CH40E_M10
CH40E_M8
CH40E_L1
CH40E_M7
CH40E_M3
CH40d45_C1
CH40d45_D18
CH40d45_B2
CH40d45_A3
CH40d45_D2
CH40d45_D7
CH40d45_D16
CH40d45_B1
CH40d45_C7
CH40d45_D14
CH40d45_A1
CH40d45_C2
CH40d45_D17
CH40d45_A2
CH40d45_M5
CH40d45_L2
CH40d45_M10
CH40d45_L3
CH40d45_L1
CH40d45_M9
CH40d45_M2
CH40d45_M6
CH40d45_M12
CH40d45_M11
CH40d45_M13
CH40d45_M4
CH40d45_M3
CH40d45_M8
CH40d45_M7
CH40d45_L4
CH40d45_M1
CH40d111_D6
CH40d111_D7
CH40d111_C1
CH40d111_D2
CH40d111_D4
CH40d111_D3
CH40d111_D5
CH40d111_D1
CH40d111_C2
CH40d181_D8
CH40d181_D7
CH40d181_D9
CH40d181_D10
CH40d181_D5
CH40d181_D1
CH40d181_B2
CH40d181_J7
CH40d181_J3
CH40d181_J10
CH40d181_J1
CH40d412_C3
CH40d412_D8
CH40d412_D3
CH40d412_D12
CH40d412_C1
CH40d412_D4
CH40d412_D18
CH40d412_D7
CH40d412_D13
CH40d412_D10
CH40d412_D19
CH40d412_D5

Figure S3

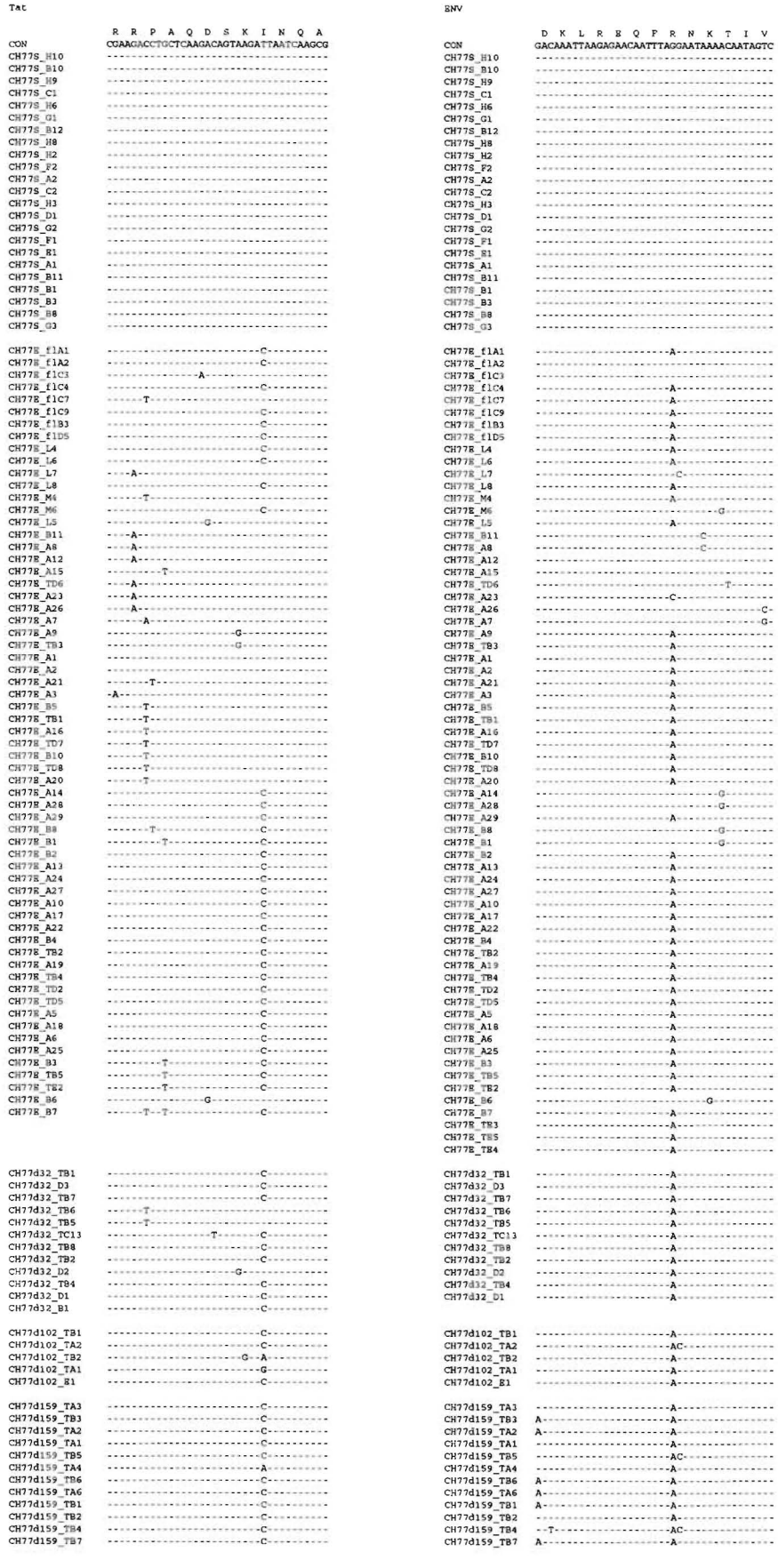
Tat
R R P A Q D S K I N Q A
CON CGAAGACCTGCTCAAGACAGTAAGATTAATCAAGCG
CH77S_H10
CH77S_B10
CH77S_H9
CH77S_C1
CH77S_H6
CH77S_G1
CH77S_B12
CH77S_H8
CH77S_H2
CH77S_F2
CH77S_A2
CH77S_C2
CH77S_H3
CH77S_D1
CH77S_G2
CH77S_F1
CH77S_E1
CH77S_A1
CH77S_B11
CH77S_B1
CH77S_B3
CH77S_B8
CH77S_G3
CH77E_f1A1
CH77E_f1A2
CH77E_f1C3
CH77E_f1C4
CH77E_f1C7
CH77E_f1C9
CH77E_f1B3
CH77E_f1D5
CH77E_L4
CH77E_L6
CH77E_L7
CH77E_L8
CH77E_M4
CH77E_M6
CH77E_L5
CH77E_B11
CH77E_A8
CH77E_A12
CH77E_A15
CH77E_TD6
CH77E_A23
CH77E_A26
CH77E_A7
CH77E_A9
CH77E_TB3
CH77E_A1
CH77E_A2
CH77E_A21
CH77E_A3
CH77E_B5
CH77E_TB1
CH77E_A16
CH77E_TD7
CH77E_B10
CH77E_TD8
CH77E_A20
CH77E_A14
CH77E_A28
CH77E_A29
CH77E_B8
CH77E_B1
CH77E_B2
CH77E_A13
CH77E_A24
CH77E_A27
CH77E_A10
CH77E_A17
CH77E_A22
CH77E_B4
CH77E_TB2
CH77E_A19
CH77E_TB4
CH77E_TD2
CH77E_TD5
CH77E_A5
CH77E_A18
CH77E_A6
CH77E_A25
CH77E_B3
CH77E_TB5
CH77E_TB2
CH77E_B6
CH77E_B7
CH77d32_TB1
CH77d32_D3
CH77d32_TB7
CH77d32_TB6
CH77d32_TB5
CH77d32_TC13
CH77d32_TB8
CH77d32_TB2
CH77d32_D2
CH77d32_TB4
CH77d32_D1
CH77d32_B1
CH77d102_TB1
CH77d102_TA2
CH77d102_TB2
CH77d102_TA1
CH77d102_E1
CH77d159_TA3
CH77d159_TB3
CH77d159_TA2
CH77d159_TA1
CH77d159_TB5
CH77d159_TA4
CH77d159_TB6
CH77d159_TA6
CH77d159_TB1
CH77d159_TB2
CH77d159_TB4
CH77d159_TB7
ENV
D K L R E Q F R N K T I V
CON GACAAATTAAGAGAACAATTTAGGAATAAAACAATAGTC
CH77S_H10
CH77S_B10
CH77S_H9
CH77S_C1
CH77S_H6
CH77S_G1
CH77S_B12
CH77S_H8
CH77S_H2
CH77S_F2
CH77S_A2
CH77S_C2
CH77S_H3
CH77S_D1
CH77S_G2
CH77S_F1
CH77S_E1
CH77S_A1
CH77S_B11
CH77S_B1
CH77S_B3
CH77S_B8
CH77S_G3
CH77E_f1A1
CH77E_f1A2
CH77E_f1C3
CH77E_f1C4
CH77E_f1C7
CH77E_f1C9
CH77E_f1B3
CH77E_f1D5
CH77E_L4
CH77E_L6
CH77E_L7
CH77E_L8
CH77E_M4
CH77E_M6
CH77E_L5
CH77E_B11
CH77E_A8
CH77E_A12
CH77E_A15
CH77E_TD6
CH77E_A23
CH77E_A26
CH77E_A7
CH77E_A9
CH77E_TB3
CH77E_A1
CH77E_A2
CH77E_A21
CH77E_A3
CH77E_B5
CH77E_TB1
CH77E_A16
CH77E_TD7
CH77E_B10
CH77E_TD8
CH77E_A20
CH77E_A14
CH77E_A28
CH77E_A29
CH77E_B8
CH77E_B1
CH77E_B2
CH77E_A13
CH77E_A24
CH77E_A27
CH77E_A10
CH77E_A17
CH77E_A22
CH77E_B4
CH77E_TB2
CH77E_A19
CH77E_TB4
CH77E_TD2
CH77E_TD5
CH77E_A5
CH77E_A18
CH77E_A6
CH77E_A25
CH77E_B3
CH77E_TB5
CH77E_TB2
CH77E_B6
CH77E_B7
CH77E_TB3
CH77E_TB5
CH77E_TB4
CH77d32_TB1
CH77d32_D3
CH77d32_TB7
CH77d32_TB6
CH77d32_TB5
CH77d32_TC13
CH77d32_TB8
CH77d32_TB2
CH77d32_D2
CH77d32_TB4
CH77d32_D1
CH77d102_TB1
CH77d102_TA2
CH77d102_TB2
CH77d102_TA1
CH77d102_E1
CH77d159_TA3
CH77d159_TB3
CH77d159_TA2
CH77d159_TA1
CH77d159_TB5
CH77d159_TA4
CH77d159_TB6
CH77d159_TA6
CH77d159_TB1
CH77d159_TB2
CH77d159_TB4
CH77d159_TB7

Figure S4

Two nt positions with shared mutations in pol gene.

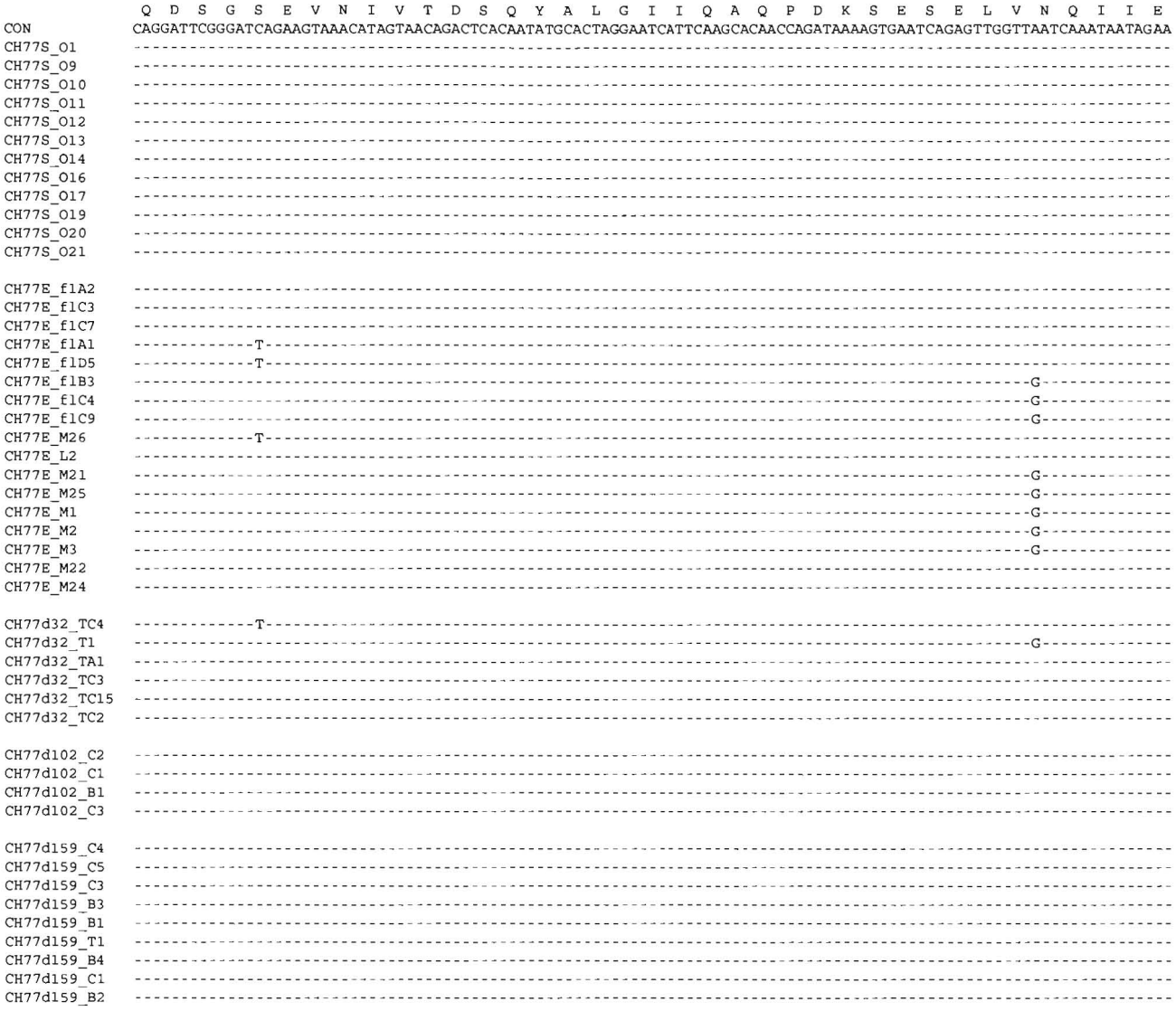

Figure S5

One nt position with a shared mutation in pol.

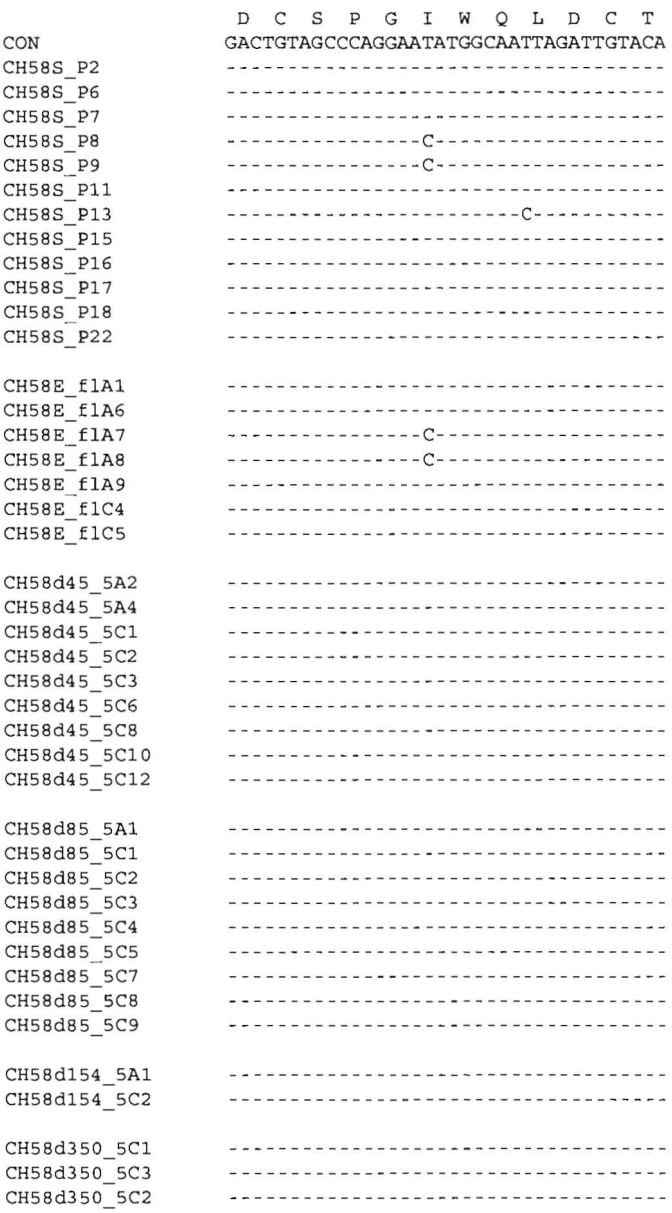

Figure S6

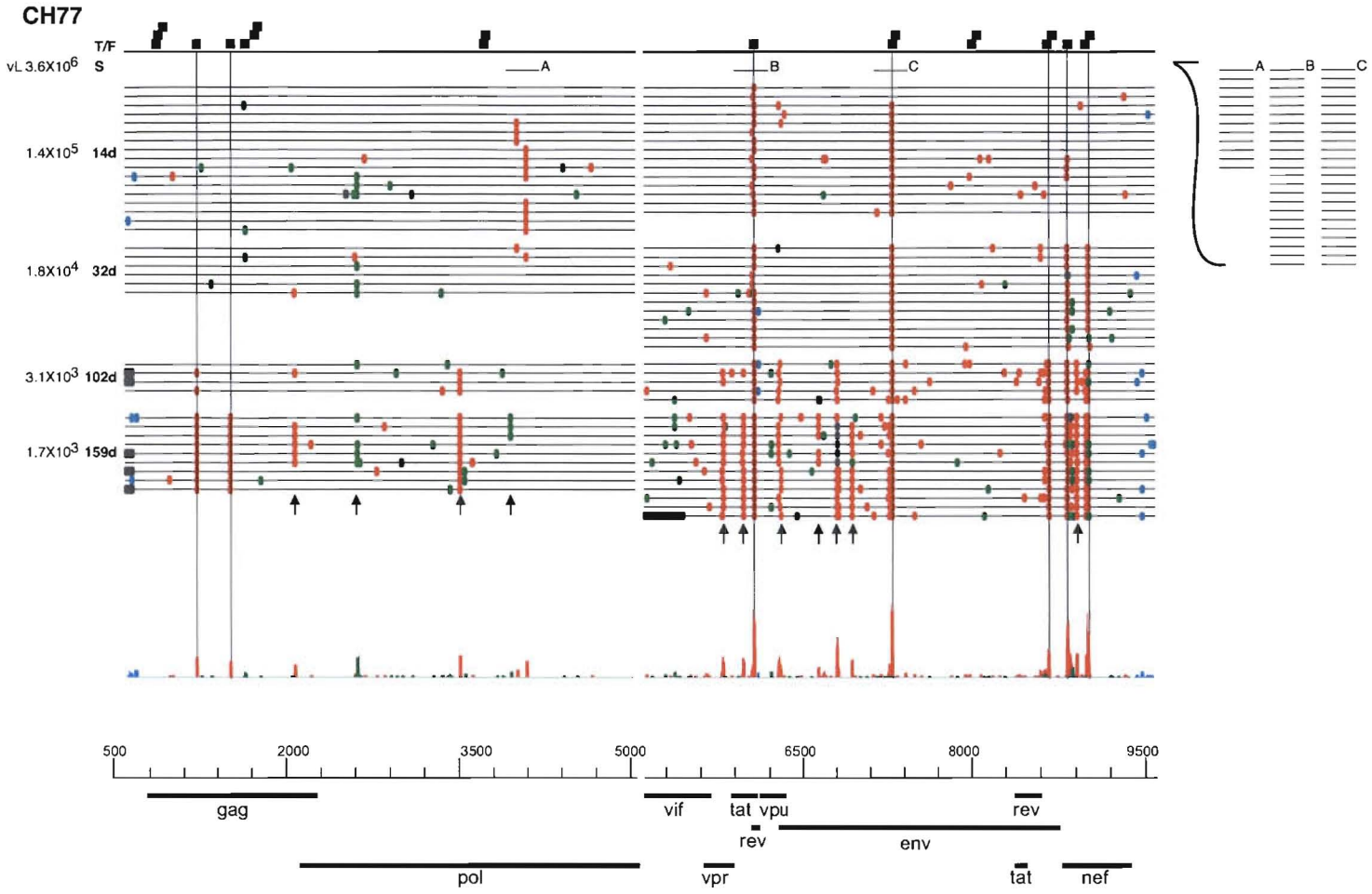
CH77
T/F
vL 3.6X10^6
S
1.4X10^5 14d
1.8X10^4 32d
3.1X10^3 102d
1.7X10^3 159d
A
B
C
500
2000
3500
5000
6500
8000
9500
gag
pol
vif
vpr
tat
rev
vpu
env
rev
tat
nef

Figure S7

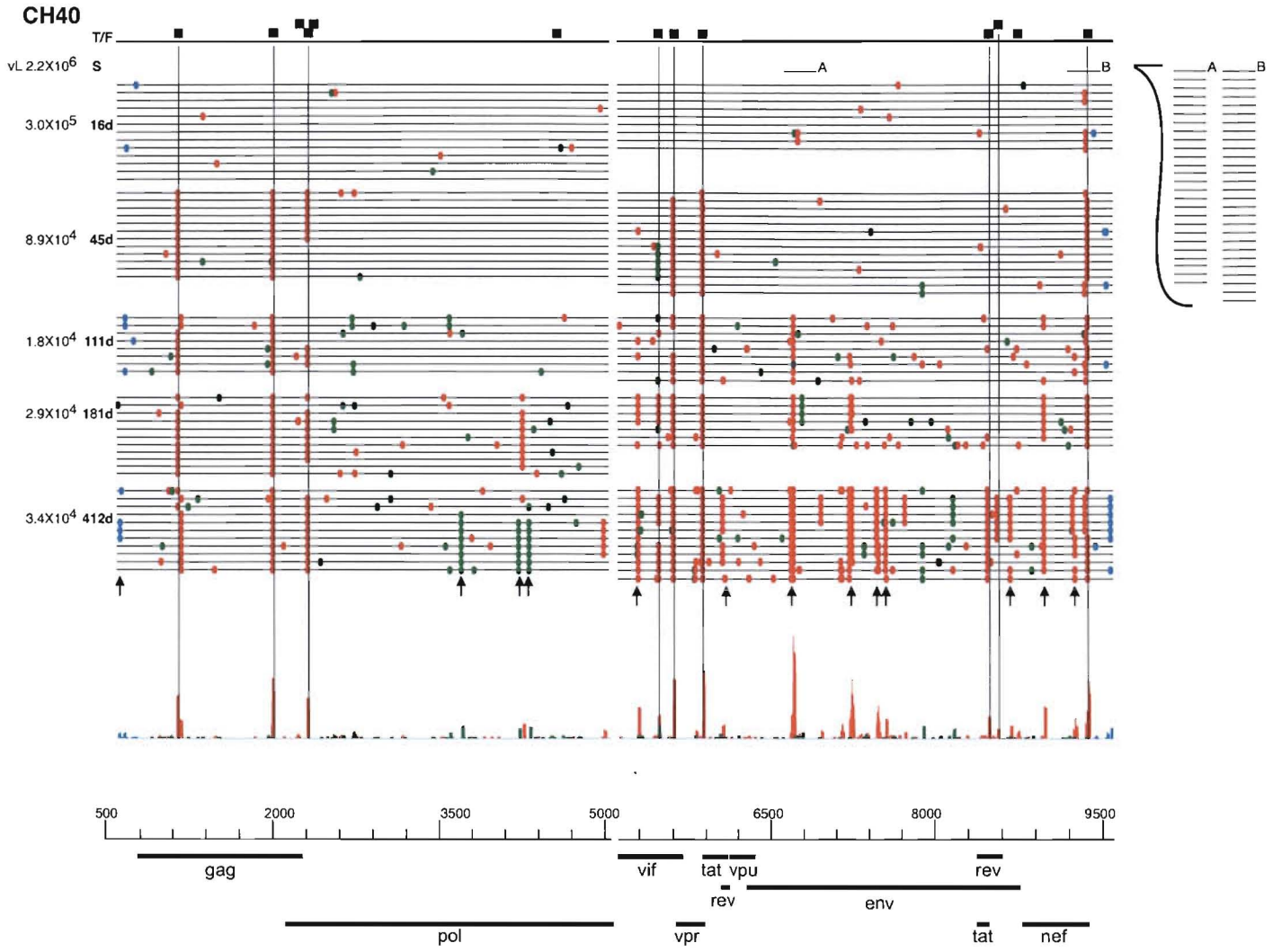
CH40
T/F
vL 2.2X10^6 S
3.0X10^5 16d
8.9X10^4 45d
1.8X10^4 111d
2.9X10^4 181d
3.4X10^4 412d
A
B
500
2000
3500
5000
6500
8000
9500
gag
pol
vif
vpr
tat
rev
vpu
env
rev
tat
nef

Figure S8

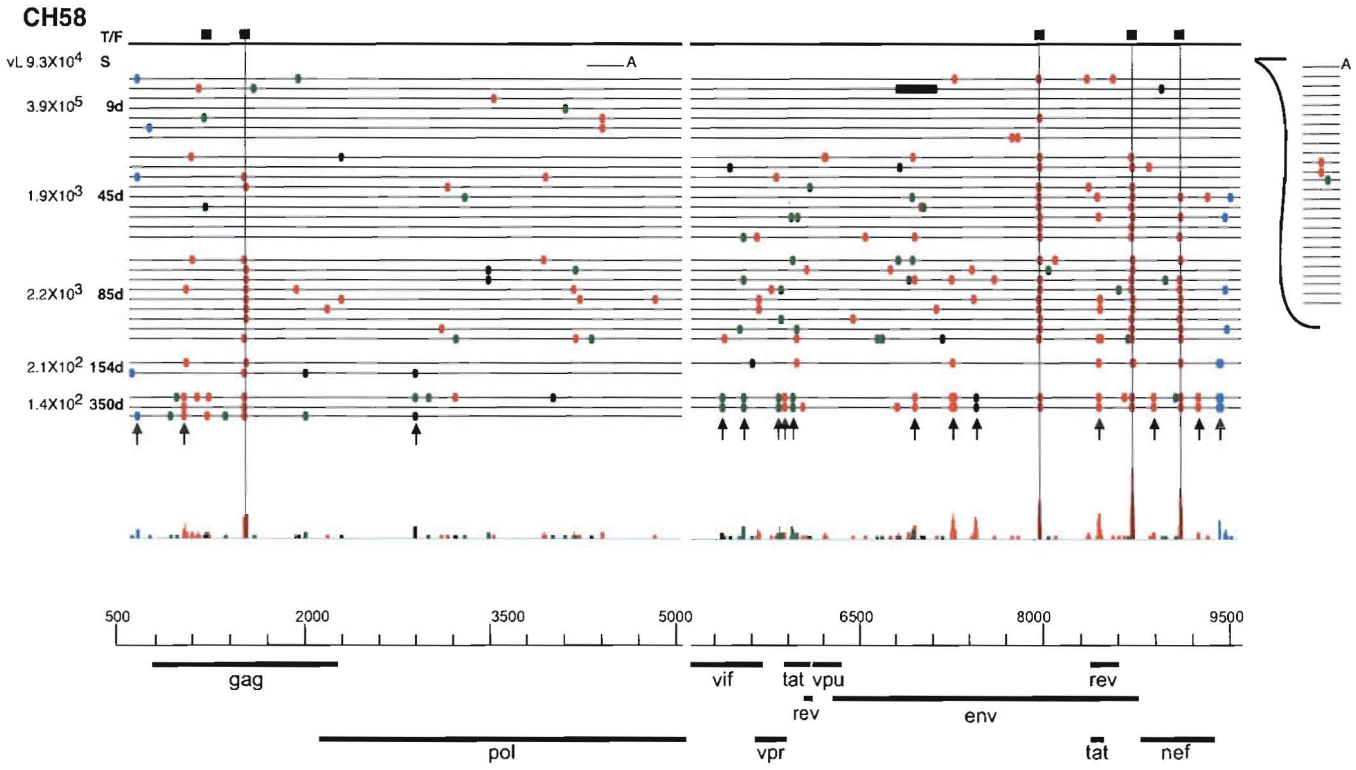
CH58
T/F
vL 9.3X10^4 S
3.9X10^5 9d
1.9X10^3 45d
2.2X10^3 85d
2.1X10^2 154d
1.4X10^2 350d
A
A
500
2000
3500
5000
6500
8000
9500
gag
pol
vif
vpr
tat
rev
vpu
env
rev
tat
nef

Figure S9