Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing

Citation for published version:

Digital Object Identifier (DOI):
10.1128/JVI.02660-07

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published in:
Journal of Virology

Publisher Rights Statement:
RoMEO blue

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Deciphering Human Immunodeficiency Virus Type 1 Transmission and Early Envelope Diversification by Single-Genome Amplification and Sequencing

Jesus F. Salazar-Gonzalez, # Elizabeth Bailes, # Kimmy T. Pham, Maria G. Salazar, M. Brad Guffey, Brandon F. Keele, Cynthia A. Derdeyn, Paul Farmer, Eric Hunter, Susan Allen, Olivier Manigart, Joseph Mulenga, Jeffrey A. Anderson, Ronald Swanstrom, Barton F. Haynes, Gayathri S. Athreya, Bette T. M. Korber, Paul M. Sharp, George M. Shaw, and Beatrice H. Hahn*

Departments of Medicine and Microbiology, University of Alabama at Birmingham, Birmingham, Alabama 35294; Institute of Genetics, University of Nottingham, Nottingham NG7 2UH, United Kingdom; Department of Pathology and Laboratory Medicine, Emory University, Atlanta, Georgia 30329; Zambia-Emory HIV Research Group (ZEHRG) and Zambia Blood Transfusion Service, Lusaka, Zambia; Department of Internal Medicine and UNC Center for AIDS Research, University of North Carolina, Chapel Hill, North Carolina 27599; Duke Human Vaccine Institute, Duke University Medical Center, Durham, North Carolina 27710; Los Alamos National Laboratory, Los Alamos, New Mexico 87545; and Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom

Received 14 December 2007/Accepted 29 January 2008

Accurate identification of the transmitted virus and sequences evolving from it could be instrumental in elucidating the transmission of human immunodeficiency virus type 1 (HIV-1) and in developing vaccines, drugs, or microbicides to prevent infection. Here we describe an experimental approach to analyze HIV-1 env genes as intact genetic units amplified from plasma virion RNA by single-genome amplification (SGA), followed by direct sequencing of uncloned DNA amplicons. We show that this strategy precludes in vitro artifacts caused by Taq-induced nucleotide substitutions and template switching, provides an accurate representation of the env quasispecies in vivo, and has an overall error rate (including nucleotide misincorporation, insertion, and deletion) of less than $8 \times 10^{-5}$. Applying this method to the analysis of virus in plasma from 12 Zambian subjects from whom samples were obtained within 3 months of seroconversion, we show that transmitted or early founder viruses can be identified and that molecular pathways and rates of early env diversification can be defined. Specifically, we show that 8 of the 12 subjects were each infected by a single virus, while 4 others acquired more than one virus; that the rate of virus evolution in one subject during an 80-day period spanning seroconversion was $1.7 \times 10^{-5}$ substitutions per site per day; and that evidence of strong immunologic selection can be seen in Env and overlapping Rev sequences based on nonrandom accumulation of nonsynonymous mutations. We also compared the results of the SGA approach with those of more-conventional bulk PCR amplification methods performed on the same patient samples and found that the latter is associated with excessive rates of Taq-induced recombination, nucleotide misincorporation, template resampling, and cloning bias. These findings indicate that HIV-1 env genes, other viral genes, and even full-length viral genomes responsible for productive clinical infection can be identified by SGA analysis of plasma virus sampled at intervals typical in large-scale vaccine trials and that pathways of viral diversification and immune escape can be determined accurately.

The transmission of human immunodeficiency virus type 1 (HIV-1) and the establishment of a productive infection are complex biological processes, the precise mechanisms of which remain to be elucidated. Because most HIV-1 transmissions result from virus exposure at mucosal surfaces and because there is a delay of approximately 1 to 3 weeks between virus infection and the first detection of viral RNA (vRNA) in the blood, it has been difficult to study the molecular details of HIV-1 transmission and early diversification in humans (12, 30, 42). Initial studies of sexually acquired HIV-1 infection suggested that virus populations in the acute phase are generally genetically highly homogeneous, in contrast to the more heterogeneous viral populations found in chronic infections (4, 5, 52, 55–57). This observation, along with the observation that new infections are almost always caused by CCR5-tropic viruses, suggested that HIV-1 infection is associated with a transmission “bottleneck.” More-recent studies, however, have reported heterogeneous virus populations near the time of infection, including in samples from female sex workers from Kenya and South Africa (11, 21, 22, 31) and in samples from men who had sex with men in the United States (32), suggesting that the routes and circumstances of infection may affect the complexity of the transmitted virus (36). In particular, sexually transmitted diseases, high viral loads in the donor, and...
particular routes of transmission increase the transmission efficiency of HIV-1 (28, 29, 48, 49) and may favor the transmission of multiple HIV-1 variants. Finally, it has been suggested that the transmission of multiple variants is common, but that homogenization of viral sequences occurs in the postinfection period (15). Thus, while it is clear that HIV-1 infection can result from the transmission of a single or more than one virus, little else is known about the quantitative or qualitative features of HIV-1 transmission, including the identities and phenotypic properties of the transmitted viruses.

Differing findings concerning the complexity of viruses in the acute and early phases of HIV-1 infection likely result from a combination of factors, including differences in study populations and associated risk behaviors; various clinical and laboratory definitions of acute, early, and chronic infection; and different experimental strategies used to analyze the genetic complexity of the evolving viral quasispecies postinfection. Chief among these are differences in the experimental designs and the methodologies used. A common approach has been to identify subjects within the first several months following infection and to derive viral sequences by bulk or near-limiting-dilution PCR amplification of proviral DNA or plasma RNA in samples obtained from them, followed by cloning, sequencing, and phylogenetic analysis (2, 4, 5, 9, 15, 21, 24, 31, 32, 36, 37, 57). In addition, acute infection cases have been analyzed by using the heteroduplex tracking assay (HTA) (4, 15, 21, 31, 32). While these approaches provide a first approximation of the complexity of transmitted virus(es), they have significant limitations. HTA, for example, interrogates only a fraction of the gene of interest and does not provide sequence information (4, 15, 21, 31, 32). Consequently, HTA allows for only qualitative inferences regarding the genetic complexity of virus populations. Bulk and near-endpoint PCR followed by cloning and sequencing are compromised by Taq polymerase-mediated template switching, which generates recombinants in vitro that do not exist in vivo (6, 26, 35, 43, 54); the introduction of Taq polymerase errors into cloned products (27); and nonproportional representation of target sequences due to template resampling (44, 45) or unequal template amplification and cloning (20, 27). Verified endpoint titration of viral nucleic acids eliminates PCR-induced recombination and ensures the proportional representation of target sequences (43), but if the amplicon is molecularly cloned prior to sequencing (as is generally the case), it will contain Taq-induced nucleotide misincorporations (27).

In this study, we sought to develop an experimental strategy that would allow us to amplify and sequence complete HIV-1 env genes from virion RNA in the plasma of infected individuals while avoiding methodological artifacts inherent in previously used approaches. To this end, we adapted methods described by Palmer and colleagues, who used single-genome sequencing of uncloned pro-pol amplicons to identify genetically linked drug resistance mutations in plasma-derived virus (27). The newly developed methods were applied to characterize the evolving quasispecies in a cohort of heterosexual Zambian subjects with acute or early HIV-1 clade C (or other) infections who were screened at 3-month intervals for evidence of incident infections. We also compared the results obtained by single-genome amplification (SGA)-direct sequencing with those obtained by more-conventional bulk PCR cloning and sequencing methods on the same clinical specimens and developed methods for evaluating the sample misidentification and cross-contamination that invariably occur in the setting of a large field trial. Our results define an effective experimental strategy for the molecular analysis of the transmission of HIV-1 and its early evolution in cohorts with differing risk behaviors and in vaccine trials where the identification of transmitted viruses and molecular pathways of virus immune escape can be instrumental in assessing vaccine efficacy or failure.

MATERIALS AND METHODS

Subjects. The 12 subjects whose infections are characterized in this study were enrolled in a cohort of heterosexual couples with discordant HIV infection status (where one partner is HIV-1 infected and the other is uninfected) which was established in Lusaka, Zambia, in 1994 to provide voluntary HIV-1 testing and counseling, health care, and longitudinal monitoring for incident infections (1, 25). Although the number of new HIV-1 infections in this cohort has been less than the number in the general population, seroconversion still occurs at a frequency of 8.5 per 100 person years, with similar female-to-male and male-to-female transmission rates (7). HIV-1-negative subjects were monitored for antibody seroconversion at 3-month intervals, and all new infections were molecularly characterized to determine whether they had resulted from transmission by the cohabiting partner (“linked” transmission) or a different individual (“unlinked” transmission) (47). Six of the 12 individuals included in the current study had epidemiologically linked infections, while the other six had unlinked infections. Since the primary purpose of this study was to develop and validate a new experimental strategy to infer transmitted HIV-1 env genes, no attempt was made to match the demographic or laboratory characteristics of subjects with linked or unlinked infections.

Laboratory staging of primary HIV-1 infection. The subjects were classified according to the system of Fiebig et al. (8) that is based on the detection of HIV-1-specific RNA, antigen, and antibody in plasma (Fig. 1). The Fiebig classification includes an eclipse phase that precedes the first detection of vRNA and subsequent stages defined by the orderly appearance of viral markers: stage I (vRNA positive, p24 antigen and antibody negative), stage II (vRNA and p24 antigen positive, antibody negative), stage III (enzyme-linked immunosorbent assay [ELISA] antibody positive, Western blot negative), stage IV (ELISA positive, Western blot indeterminate), stage V (ELISA and Western blot positive, p31 integrase antibody negative), and stage VI (ELISA, Western blot, and p31 integrase antibody positive). Plasma samples from the patients were tested for HIV-1 antibodies by ELISA (Enzymost HIV integral from Dade Behring, Marburg, Germany; Biotest anti-HIV tetra ELISA from Biotest, Dreieich, Germany; or Abbott Murex HIV Ag/Ab combination assay from Murex Biotech Limited, Dartford, United Kingdom) and Western blot analysis (Genetics Systems HIV-1 Western blot from Bio-Rad Laboratories, Hercules, CA). In accordance with diagnostic guidelines, a positive Western blot was defined as reactivity with any two of the following three HIV-1 proteins: the exterior envelope glycoprotein (uncleaved envelope precursor [gp120/gp160]), the transmembrane envelope glycoprotein (gp41), and the major core protein (p24). An indeterminate Western blot was defined as any visible band that did not meet the positivity criteria. A negative Western blot included no visible bands. A subset of samples was also monitored for p24 antigen by using a Beckman Coulter HIV-1 p24 antigen ELISA (Beckman Coulter Inc., Fullerton, CA).

RNA extraction. For plasma specimens containing >10,000 RNA copies/ml, plasma RNA was extracted by using a QIAamp viral RNA mini kit (Qiagen, Valencia, CA). Samples with vRNA loads of <10,000 copies/ml were concentrated by centrifugation at 23,600 × 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 ~50 μl. For SGA and bulk PCR methods, between 1,000 and 20,000 vRNA molecules were typically reverse transcribed.

dDNA synthesis. 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). RNA, deoxynucleoside triphosphates (0.5 mM each), and 0.25 μM primer OFM19 (5′-GGACTGAAAGGGACGACGCTTTTATGTTAGCGCTTA-3′; nucleotides [nt] 3604 to 3653 of the HXB2 sequence) were incubated for 5 min at 65°C to denature secondary structure of the DNA. First-strand cDNA synthesis was carried out in 20- to 100-μl reaction mixtures with 1× reverse transcriptase buffer containing 5 mM dithiothreitol, 2 U/μl of an RNase inhibitor (RNaseOUT), and 10 U/μl Super-
products were analyzed by 1% agarose gel electrophoresis, and products of the nested PCRs were carried out for each specimen (20, 41). The final PCR products were sequenced and found to be representative of the plasma quasispecies, five independent cDNA clones from each patient were sequenced. After restriction enzyme digestion to identify full-length clones, 10 to 37 clones containing full-length clones from each patient were sequenced.

For SGA, the reaction mixture was subjected to the first-round PCR in a volume of 20 μl containing 100 to 1,000 viral templates, followed by 1,000 viral templates) was subjected to first-round PCR in a volume of 20 μl of bulk cDNA (containing known to represent the maximum possible number of days from transmission. According to a Poisson distribution, the cDNA dilution that yields PCR products in no more than 30% of wells contains one amplifiable cDNA template per 96-well plates such that fewer than 29 PCRs yielded an amplification product. First-round PCR was carried out in 1x High Fidelity platinum PCR buffer, 2 mM MgSO4, 0.2 mM of each deoxynucleotide triphosphate, 0.2 μM of primers Vif1 and OFM19, and 0.025 U/μl platinum Taq High Fidelity polymerase (Invitrogen, Carlsbad, CA) in a 20-μl 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 35 cycles of 94°C for 15 s, 55°C for 30 s, and 68°C for 4 min, with a final extension of 68°C for 10 min. Second-round PCR was performed by using 1 μl of the first-round PCR product and primers EnvA (5'-GGGTTTATTACAGGGACAGCAGAG-3'; nt 5954 to 5982) and EnvN (5'-CTGCCAAAGGTTATUTAGCAGGAAC-3'; nt 9145 to 9171) under the same conditions used for the first-round PCR. To ensure that the amplified HIV-1 env sequences were representative of the plasma quasispecies, five independent nested PCRs were carried out for each specimen (20, 41). The final PCR products were analyzed by 1% agarose gel electrophoresis, and products of the predicted size (≈3.2 kb) were ligated into the pcDNA3.1.V5-His TOPO TA vector according to the manufacturer’s instructions (Invitrogen Life Technologies, Carlsbad, CA). The ligated vector was transformed into Stbl2 cells at 42°C for 25 s (Invitrogen Life Technologies, Carlsbad, CA). Transformed reaction mixtures were plated on LB-Amp (ampicillin; 100 μg/ml) plates and cultured overnight at 30°C. Multiple colonies were picked and grown in LB broth (100 μg/ml ampicillin), and plasmid DNA was isolated by using a QIAprep spin miniprep kit (Qiagen, Valencia, CA). The resulting plasmid DNA was subjected to restriction enzyme digestion to identify full-length clones, and 10 to 37 env-containing clones from each patient were sequenced.

**FIG. 1.** Laboratory staging of acute and early HIV-1 infections. (A) Temporal appearance of HIV-1-specific laboratory markers following HIV-1 infection according to the classification system of Fiebig et al. (8). The eclipse phase is defined by the interval between transmission and first detection of vRNA in the plasma and generally lasts about 10 days, with a range of approximately 7 to 21 days (3, 10, 17–19, 40). The mean durations of Fiebig stages I (7 days), II (5 days), III (3 days), IV (6 days), and V/VI (70+ days) are indicated. (B) Time points (x axis) at which plasma samples were obtained for each of the 12 study subjects (y axis). Because subjects were studied at intervals of ~3 months, the symbols are positioned to represent the maximum possible number of days from transmission.
final extension at 68°C for 10 min. The amplicons were sized on precast 1% agarose E-gel 96 (Invitrogen Life Technologies, Carlsbad, CA). All products derived from cDNA dilutions yielding less than 30% PCR positivity were sequenced. A standard operating procedure for SGA derivation of full-length env genes is available upon request.

**Generation of RNA transcripts from HIV-1 molecular clones with T7 polymerase.** Envelope genes corresponding to two transmitted strains of HIV-1 (BORI94.4F8 and BORI94.4F12) obtained from the plasma of an acutely infected individual were directionally cloned into the viral expression vector pcDNA3.1D (Invitrogen Life Technologies, Carlsbad, CA) under the control of the T7 promoter. These plasmids were digested with EcoRV, and linear DNA was recovered by QIAquick gel extraction (Qiagen, Valencia, CA) following electrophoresis in 1% agarose. env RNA transcripts were generated using a Riboprobe in vitro transcription system (Promega, Madison, WI) by incubating 700 ng template DNA with T7 polymerase at 37°C for 80 min and utilizing a standard 100-μl volume transcription protocol under the following conditions: 1× optimized transcription buffer; 10 mM dithiothreitol; 0.5 mM rATP, rGTP, rCTP, and rUTP; 100 U rRNAse inhibitor; and 80 U T7 polymerase. Following transcription, the DNA template was degraded with 1 U per μg of template RQ1 RNase-free DNase (Promega, Madison, WI) for 15 min at 37°C. RNA molecules of more than 200 nucleotides were concentrated with RNAasy MinElute cleanup (Qiagen, Valencia, CA) and stored at −80°C. The RNA’s mass was determined by spectrophotometry (NanoDrop 1000), and the copy number estimated assuming that all transcripts were full-length. RNA transcripts (BORI94.4F8 and BORI94.4F12) were mixed 1:1 at an estimated 100,000 copies per μg. A total of 59 envelope genes representing 145,590 bases were sequenced. A standard operating procedure for SGA derivation of full-length env genes was available upon request.

**Viral cDNA synthesis and to identify all transversion, transition, and insertion-deletion mutations.**

Generation of virion RNA by transfection of 293T cells with HIV-1 molecular clones. The infectious molecular clones YUZ and SG3 were independently transfected into 293T cells (Fugene; Roche Diagnostic Corporation, Indianapolis, IN) and cultured for 48 h to generate virion RNA. The virus concentration was determined by using an Amplicor RNA assay, version 1.5 (Roche Diagnostic Corporation, Indianapolis, IN). Equal numbers of RNA molecules from each virus stock were mixed 1:1, and 20,000 vRNA copies were extracted by using a QIAamp viral RNA mini kit (Qiagen, Valencia, CA). Purified vRNA was reverse transcribed into cDNA (using primer Env3out; see below) and diluted to the single-molecule level prior to PCR amplification described above, with the following primer modifications: for first-round PCR, primer Env5in (5′-TATGAGCCATCCTCATTATGGCAGAAG-3′; nt 5957 to 5993) and antisense primer Env3in (5′-TACGGTAGATACTGTCCTACC-3′; nt 8904 to 8882) were used, and for second-round PCR, primers 3′R3BG (5′-CCTATCTGCTCAGCTACGTCG-3′; nt 8510 to 8531) and 5′F3BG (5′-CGACGGAGACCTCTCTCAAGACG-3′; nt 5993 to 6015) were used. A total of 59 envelope genes representing 145,950 bases were sequenced. A total of 59 envelope genes representing 145,950 bases were sequenced and compared to the input env sequences to detect recombination during cDNA synthesis and to identify all transversion, transition, and insertion-deletion mutations.

**DNA sequencing.** Viral env genes were sequenced by using BigDye Terminator chemistry and the protocols recommended by the manufacturer (Applied Biosystems, Foster City, CA). The sequences were determined by using an ABI 3730xl genetic analyzer (Applied Biosystems, Foster City, CA). The sequences were determined by using an ABI 3730xl genetic analyzer (Applied Biosystems, Foster City, CA) and edited by using the Sequencher program, version 4.7 (Gene Codes, Ann Arbor, MI). Both strands of DNA were sequenced. All chromatograms were carefully inspected for sites of ambiguous sequence (double peaks), and those that contained one or more positions of mixed bases were excluded from further analysis.

**Microsatellite analysis.** Samples with heterogeneous infections were examined for potential misidentification by microsatellite analysis using an AmpFSTR Identifiler PCR amplification kit following the manufacturer’s instructions (Applied Biosystems, Foster City, CA). Briefly, DNA was extracted from 200 μl of plasma by using a QIAamp DNA blood minikit (Qiagen, Valencia, CA) and used for the amplification of 15 highly polymorphic microsatellite loci (SSS1179, D2S111, D7S820, CSF1PO, DS1358, TH01, D13S317, D16S539, D2S1338, D19S433, vWA, TPOX, D18S51, D5S818, and FGA) and a portion of the amelogenin gene for gender identification in a single PCR amplification. The amplicons were sized by using an ABI 3130x genetic analyzer. Data analysis and allele designations were carried out by using the GeneMarker program (SoftGenetics LLC). The selected markers are widely accepted tetranucleotide loci for genetic characterization, standardized under the Combined DNA Index System (CODIS).**

**Phylogenetic analyses.** Env protein sequences from each subject were aligned using CLUSTAL W (46), and amino acids were then replaced by codons.

**Intrastrain diversities were calculated by using uncorrected sequence distances. For all other analyses, the alignments were gap stripped and sequences with large deletions were excluded. All trees were constructed by using the neighbor joining method (38) implemented in CLUSTAL W using Kimura’s correction (14).**

**RESULTS**

**Assay validation.** The main objective of this study was to develop and implement an experimental approach that would...
yield an accurate representation of the HIV-1 env genes in plasma vRNA such that meaningful inferences concerning their genetic (and ultimately, biological) properties could be drawn. To this end, we first tested the performance of the SGA-direct sequencing method using in vitro-generated HIV-1 RNA templates. Two different approaches were used to generate RNA templates from different sets of cloned HIV-1 DNA so as to account for (i) differences in fidelities between T7 and human Pol II RNA polymerases and (ii) differences in recombination rates between more- or less-related HIV-1 DNA templates. In the first set of experiments, we used T7 polymerase to synthesize vRNA from two HIV-1 env clones derived from an acutely infected subject (BORI0637) in whom we had observed extensive env gene recombination in vivo (B. F. Keele, unpublished data). The two clones, BORId9.4F12 and BORId9.4F8, are 97.2% identical and represent the progeny of two transmitted variants in this subject. From a 1:1 mixture of the two RNA genomes carried through cDNA synthesis and SGA-direct sequencing, we obtained 59 full-length env sequences, corresponding to 145,000 nucleotides. In none of the sequences did we find evidence for recombination between the two genomes. The combined T7 polymerase and Superscript III nucleotide misincorporation rate was 4.39 × 10⁻⁴, consisting mainly of random nucleotide misincorporations (3.43 × 10⁻⁴) but also of single or double nucleotide insertions or deletions in runs of A or T residues (9.61 × 10⁻⁵). Finally, transcripts from the two molecular clones were recovered in equal proportions, demonstrating that heterogeneous templates are accurately represented using the SGA approach (Table 1).

In a second set of experiments, we transfected 293T cells with infectious molecular clones of HIV-1 (YU2 and SG3) and collected virions from the culture supernatants for analysis. This was done to avoid T7 polymerase contributions to the measured error rate and to more closely replicate in vivo conditions of HIV-1 gene expression and replication. The env genes of YU2 and SG3 are 92.5% identical. Virion RNA preparations from the two strains were mixed 1:1 and subjected to cDNA synthesis and SGA-direct sequencing. Again, the inspection of 50 env sequences corresponding to more than 132,000 nucleotides failed to yield evidence of recombination but provided a nearly equal representation of both input strains (Table 1). The combined error rate using RNA Pol II and Superscript III was 8.28 × 10⁻⁵, consisting of random nucleotide substitutions (6.77 × 10⁻⁵) and single-nucleotide insertions or deletions in runs of A or T residues (1.51 × 10⁻⁵). Of note, Taq polymerase misincorporations do not contribute to this error rate because either they appear as mixed bases in sequence chromatograms (if introduced during the first rounds of amplification) and are thus discarded or they remain undetected because they affect only a minor fraction of the amplified molecules. Thus, in addition to precluding artifactual recombination, SGA-derived sequences are not affected by Taq polymerase-induced errors. These results, along with the rates of transitions, transversions, insertions, and deletions, are summarized in Table 1.

### Laboratory staging of acute and early HIV-1 infection in subjects. In order to evaluate the sequence diversity of HIV-1 in relation to the estimated duration of infection, we adopted the Fiebig classification (8), which categorizes patients based on an orderly and reproducible appearance of HIV-1-specific markers in the plasma (Fig. 1A). The eclipse phase is defined by the interval between virus transmission and the first detection of vRNA in the plasma and is believed to last on average about 10 days, with a range of approximately 7 to 21 days (3, 10, 17–19, 40). The mean durations of Fiebig stages I (7 days), II (5 days), III (3 days), IV (6 days), and V/VI (70+ days) are indicated. The 95% confidence intervals (CI) were calculated and reported elsewhere (8). The timing of patient sampling (Fig. 1B) was determined from the clinical record and by analysis of plasma vRNA, p24 antigen, and antibody, and all were in good agreement (Table 2). Two subjects (ZM246F and ZM247F) were initially studied during Fiebig stage II, when their plasma vRNA levels were extremely high and HIV-1 antibody was undetectable. Subject ZM249M was first studied during Fiebig stage IV, when plasma vRNA was still high and antibodies were detectable by ELISA but the Western blot pattern was indeterminant. Based on estimates of Fiebig stage durations, this sample was still expected to have been obtained within ~31 days of infection. Subjects ZM246F, ZM247F, and ZM249M were each studied at a second time point 4 to 80 days after the first. The other nine subjects were studied on a single

### Table 1. Determination of SGA-direct sequencing error rates using in vitro-synthesized HIV-1 RNA templates

<table>
<thead>
<tr>
<th>Strain</th>
<th>Genomes sequenced</th>
<th>Genomes with no errors</th>
<th>Nucleotides sequenced</th>
<th>Transitions</th>
<th>Transversions</th>
<th>Insertions</th>
<th>Deletions</th>
<th>Total no. of errors (%)</th>
<th>Number of recombinants (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BORId9.4F12</td>
<td>29</td>
<td>12</td>
<td>71,340</td>
<td>15</td>
<td>7</td>
<td>2</td>
<td>3</td>
<td>27 (0.038)</td>
<td>0 (&lt;1)</td>
</tr>
<tr>
<td>BORId9.4F8</td>
<td>30</td>
<td>9</td>
<td>74,250</td>
<td>21</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>37 (0.050)</td>
<td>0 (&lt;1)</td>
</tr>
<tr>
<td>Total for both strains</td>
<td>59</td>
<td>21</td>
<td>145,590</td>
<td>36 (0.0247)</td>
<td>14 (0.0096)</td>
<td>9 (0.0062)</td>
<td>5 (0.0034)</td>
<td>64 (0.044)</td>
<td>0 (&lt;1)</td>
</tr>
<tr>
<td>YU2</td>
<td>27</td>
<td>24</td>
<td>71,508</td>
<td>2 (0.0028)</td>
<td>0</td>
<td>1 (0.0014)</td>
<td>0</td>
<td>3 (0.0042)</td>
<td>0 (&lt;1)</td>
</tr>
<tr>
<td>SG3</td>
<td>23</td>
<td>16</td>
<td>61,331</td>
<td>7 (0.0114)</td>
<td>0</td>
<td>0</td>
<td>1 (0.0016)</td>
<td>8 (0.013)</td>
<td>0 (&lt;1)</td>
</tr>
<tr>
<td>Total for both strains</td>
<td>50</td>
<td>40</td>
<td>132,839</td>
<td>9 (0.0068)</td>
<td>0</td>
<td>1 (0.0008)</td>
<td>1 (0.0008)</td>
<td>11 (0.0083)</td>
<td>0 (&lt;1)</td>
</tr>
</tbody>
</table>

* Insertion or deletion of 1 or 2 A or T nucleotides to a preexisting run of A or T residues.
* Total errors including transitions, transversions, insertions, and deletions.
* T7-generated HIV-1 RNA transcripts were mixed 1:1 and subjected to SGA analysis as described in the text.
* HIV-1 vRNA from transfected human 293T cell culture supernatants was mixed 1:1 and subjected to SGA analysis as described in the text.
TABLE 2. Epidemiological and laboratory information for primary infection subjects

<table>
<thead>
<tr>
<th>Subject</th>
<th>Direction of transmission</th>
<th>Epidemiological linkage</th>
<th>Genetic subtype</th>
<th>Time of sample collection relative to seroconversion</th>
<th>Date (mo/day/yr) of last seronegative sample</th>
<th>Date (mo/day/yr) sample was analyzed</th>
<th>Days since last seronegative sample</th>
<th>Type of specimen</th>
<th>Plasma viral load (RNA copies/ml)</th>
<th>e p24</th>
<th>EIA</th>
<th>WB</th>
<th>Feibig stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZM184F</td>
<td>M to F</td>
<td>Unlinked</td>
<td>C</td>
<td>Post</td>
<td>04/11/03</td>
<td>07/10/03</td>
<td>90</td>
<td>Plasma</td>
<td>89,340</td>
<td></td>
<td></td>
<td></td>
<td>V</td>
</tr>
<tr>
<td>ZM214M</td>
<td>F to M</td>
<td>Unlinked</td>
<td>C</td>
<td>Pre</td>
<td>04/05/03</td>
<td>07/01/02</td>
<td>186</td>
<td>Plasma</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
<td>VB</td>
</tr>
<tr>
<td>ZM215F</td>
<td>M to F</td>
<td>Unlinked</td>
<td>C</td>
<td>Pre</td>
<td>07/04/02</td>
<td>07/04/02</td>
<td>0</td>
<td>Serum</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
<td>VB</td>
</tr>
<tr>
<td>ZM246F</td>
<td>M to F</td>
<td>Unlinked</td>
<td>C</td>
<td>Pre</td>
<td>01/14/03</td>
<td>01/14/03</td>
<td>0</td>
<td>Plasma</td>
<td>10,013,800</td>
<td></td>
<td></td>
<td></td>
<td>VB</td>
</tr>
<tr>
<td>ZM247F</td>
<td>M to F</td>
<td>Unlinked</td>
<td>C</td>
<td>Pre</td>
<td>10/28/03</td>
<td>10/01/03</td>
<td>4</td>
<td>Plasma</td>
<td>10,823,500</td>
<td></td>
<td></td>
<td></td>
<td>VB</td>
</tr>
<tr>
<td>ZM249M</td>
<td>F to M</td>
<td>Unlinked</td>
<td>C</td>
<td>Pre</td>
<td>08/05/03</td>
<td>08/05/03</td>
<td>0</td>
<td>Plasma</td>
<td>2,000,000</td>
<td></td>
<td></td>
<td></td>
<td>VB</td>
</tr>
<tr>
<td>ZM178F</td>
<td>M to F</td>
<td>Linked</td>
<td>C</td>
<td>Post</td>
<td>10/01/02</td>
<td>01/18/03</td>
<td>109</td>
<td>Plasma</td>
<td>15,620</td>
<td></td>
<td></td>
<td></td>
<td>VB</td>
</tr>
<tr>
<td>ZM180M</td>
<td>F to M</td>
<td>Linked</td>
<td>C</td>
<td>Post</td>
<td>04/09/02</td>
<td>07/13/02</td>
<td>129</td>
<td>Plasma</td>
<td>87,480</td>
<td></td>
<td></td>
<td></td>
<td>VB</td>
</tr>
<tr>
<td>ZM229M</td>
<td>F to M or F</td>
<td>Linked</td>
<td>C</td>
<td>Pre</td>
<td>07/17/02</td>
<td>07/17/02</td>
<td>0</td>
<td>Plasma</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
<td>VB</td>
</tr>
<tr>
<td>ZM231F</td>
<td>M to F</td>
<td>Linked</td>
<td>U</td>
<td>Post</td>
<td>09/06/02</td>
<td>12/17/02</td>
<td>102</td>
<td>Plasma</td>
<td>87,300</td>
<td></td>
<td></td>
<td></td>
<td>VB</td>
</tr>
<tr>
<td>ZM235F</td>
<td>M to F</td>
<td>Linked</td>
<td>C</td>
<td>Post</td>
<td>09/20/02</td>
<td>12/20/02</td>
<td>91</td>
<td>Plasma</td>
<td>227,000</td>
<td></td>
<td></td>
<td></td>
<td>VB</td>
</tr>
</tbody>
</table>

**Notes:**
- M, male; F, female; Pre, before; Post, after; Pos, positive; Neg, negative; NA, not available; Ind, indeterminant; U, HIV-1 (group M) subtype unclassified.
- Subjects were enrolled in a cohort of cohabitating couples with discordant HIV-1 infection status.
- Viral linkage was determined by diagnostic PCR of donor and recipient virus as described previously (47).
- The date of seroconversion was determined in accordance with the system of Fiebig et al. (8).
- Plasma was obtained 186 days prior to the last seronegative sample.
occasion no more than 99 to 151 days after infection, according to their clinical records (Table 2).

**env gene diversity in acute and early infections.** A total of 417 SGA-derived full-length *env* gene sequences were obtained from the 12 subjects, with a range of 24 to 49 sequences per subject. A phylogenetic tree of these *env* sequences is shown in Fig. 2. Eleven subjects were infected with HIV-1 subtype C; one individual (ZM231F) harbored a strain that fell outside
a tree was constructed using 41 sequences from both time
(Fig. 1B). First, a tree was constructed using the 27 sequences
were obtained 80 days apart, during Fiebig stages II and VI
individual (ZM246F) who had been infected by a single virus
spanning seroconversion, we examined sequences from one
other four individuals, including subject ZM247F, whose sam-
infection, as indicated by their Fiebig stages (V and VI). The
subjects (ZM178F, ZM180M, ZM184F, ZM206F, ZM231F,
sequences from six other
subjects analyzed during the earliest stages of infection
ZM249M, each had been infected by a single virus. Conversely,
subjects ZM214M, ZM215F, and ZM229M all had env
TABLE 3. Diversity of SGA-direct sequence-derived complete env
genes in subjects with primary HIV-1 infections

<table>
<thead>
<tr>
<th>Subject</th>
<th>Fiebig stage</th>
<th>Date sample collected (mo/day/yr)</th>
<th>No. of SGA env amplicons</th>
<th>Nucleotide sequence length (amplicons)</th>
<th>Nucleotide sequence diversity Mean</th>
<th>Range</th>
<th>Amino acid sequence diversity Mean</th>
<th>Range</th>
<th>Time to MRCAa (days)</th>
<th>Viral complexity</th>
<th>No. of variants transmitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZM184F V</td>
<td>10/12/03</td>
<td>26</td>
<td>2560–2613</td>
<td>0.41</td>
<td>0.00–0.67</td>
<td>0.99</td>
<td>0.00–1.85</td>
<td>112</td>
<td>Homogeneous</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ZM214M VI</td>
<td>07/02/03</td>
<td>25</td>
<td>2538–2560</td>
<td>0.25</td>
<td>0.04–0.47</td>
<td>0.61</td>
<td>0.00–1.29</td>
<td>73</td>
<td>Homogeneous</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ZM229M V</td>
<td>01/18/03</td>
<td>24</td>
<td>2562–2563</td>
<td>0.07</td>
<td>0.00–0.27</td>
<td>0.16</td>
<td>0.00–0.82</td>
<td>31</td>
<td>Homogeneous</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ZM247F I</td>
<td>11/10/03</td>
<td>19</td>
<td>2592–2595</td>
<td>0.56</td>
<td>0.00–2.70</td>
<td>0.90</td>
<td>0.00–4.29</td>
<td>774</td>
<td>Heterogeneous</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>ZM249M IV</td>
<td>12/02/03</td>
<td>28</td>
<td>2244–2565</td>
<td>0.16</td>
<td>0.00–0.39</td>
<td>0.34</td>
<td>0.00–0.82</td>
<td>54</td>
<td>Homogeneous</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

a The time to the MRCA was estimated by using the average branch length to root of midpoint-rooted neighbor-joining trees (inferred after excluding gaps and with Kimura's two-parameter correction) assuming a rate of $1.7 \times 10^{-3}$ substitutions per site per day (estimated from the extent of env sequence divergence in subject ZM246F).

b The number of transmitted variants could not be determined.

c The time to the MRCA for sequences within each of the two discrete transmitted virus lineages in subject ZM247F was estimated to be between 10 and 31 days.

The number of transmitted variants could not be determined.

To estimate the rate of env evolution in the critical period spanning seroconversion, we examined sequences from one individual (ZM246F) who had been infected by a single virus (Fig. 2, Table 3) and from whom sequential blood samples were obtained 80 days apart, during Fiebig stages II and VI (Fig. 1B). First, a tree was constructed using the 27 sequences from the earliest time point and was midpoint rooted. Second, a tree was constructed using 41 sequences from both time points and rooted to match the first tree, assuming a molecular clock. The distance from the tip of each branch to the root of the tree was calculated by using Branchlength (http://www.hiv.lanl.gov/content/sequence/BRANCHLENGTH/branchlength.html) and plotted against the time in days (0 days for sequences from the 14 January 2003 time point and 80 days for sequences from the 4 April 2003 time point). This yielded a rate of evolution of $1.7 \times 10^{-3}$ substitutions per site per day ($6 \times 10^{-3}$ substitutions per site per year), which is very close to a rate previously obtained for HIV-1 env gene sequences (16) and similar to the mutation rate of HIV-1 in cell culture ($2.4 \times 10^{-3}$ mutations per site per replication cycle, excluding insertions and deletions), assuming a replication time between 1 and 2 days (23). To estimate the time to the most-recent common ancestor (MRCA) for sequences from all 12 primary infection subjects, this rate was applied to the average branch length of their respective env sequences to the root of the tree. For subjects ZM247F and ZM246F, from whom samples were obtained very early (Fig. 1B), estimates of the time to the MRCA of their transmitted viruses ranged between 10 and 31 days (Table 3), well within the upper bounds of the Fiebig eclipse phase and stages I and II (note that this is true for each of the two variants acquired by subject ZM247F). For subject ZM249M, the same estimate was 31 days, which corresponds to the approximate beginning date for Fiebig stage IV (Fig. 1B). For the other six subjects with homogeneous virus populations, from whom samples were obtained no later than 99 to 151 days postinfection according to clinical and serological data, the env diversities implied a time to the MRCA ranging from 43 to 73 days for stage V subjects and 73 to 112 days for stage VI subjects, again within the bounds of the corresponding Fiebig stages. These findings supported the conclusion that subjects ZM178F, ZM180M, ZM184F, ZM206F, ZM231F, and ZM235F, like subjects ZM246F and ZM249M, each had been infected by a single virus. Conversely, subjects ZM214M, ZM215F, and ZM229M all had env
sequences whose diversities led to time estimates for an MRCA virus that far exceeded plausible durations for a single virus source given their Fiebig stage, and this was also true for the two variants identified in subject ZM247F (Table 3). These subjects had therefore clearly been infected by two or more viral variants. Detailed recombination analysis of the viruses from subjects ZM229M and ZM215F indicated at least four different transmitted variants (see below), while the number of transmitted variants for subject ZM214M could not be determined with certainty (Table 3).

Identification of transmitted or early “founder” viruses. The estimates of time to the MRCA for all subjects with homogeneous infections suggested that viral sequences were coalescing at or near the transmission event. To evaluate this further, we selected for analysis three subjects (ZM249M, ZM247F, and ZM246F) from whom we had very early samples. Figure 3 depicts the phylogenetic trees of the SGA-derived env sequences from two consecutive plasma samples from each of these subjects 4 to 80 days apart. The sequences were also examined by using a novel analytical tool, Highlighter (http://www.hiv.lanl.gov/content/sequence/HIGHLIGHT/highlighter.html), which provides a visually informative representation comparing each env sequence to a selected reference sequence. For subject ZM249M and ZM246F were infected by a single virus and ZM247F by two viruses, identified as variant 1 and variant 2 in the tree (panel B). The tick in parentheses in sequence 080503_A5 (panel A) indicates a single-nucleotide insertion. The boxed region in panel C indicates clustered mutations in V3. Interestingly, these changes were limited to position 8 (Ile to Thr, Arg, or Lys) and position 25 (Asp to Asn, Glu, or Gly).

FIG. 3. Identification of transmitted or early founder env genes in subjects with acute HIV-1 infection. SGA-derived env sequences from pre- and postseroconversion plasma samples from subjects ZM249M (A), ZM247F (B), and ZM246F (C) were examined by phylogenetic tree construction (left panels) and Highlighter analysis (right panels). Sequences from the earlier time points are in bold face. Trees are midpoint rooted. The scale bars represent 1 (A and C) or 10 (B) nt substitutions per site. The corresponding Highlighter diagrams denote the locations of nucleotide sequence substitutions in each env sequence in comparison to a reference sequence listed at the top; the positions of these substitutions within the env sequence are indicated at the bottom. Nucleotide substitutions and gaps are color coded. For each subject, sets of identical and nearly identical env sequences form consensus sequences that coalesce to transmitted or early founder viral env sequences. Subjects ZM249M and ZM246F were infected by a single virus and ZM247F by two viruses, identified as variant 1 and variant 2 in the tree (panel B). The tick in parentheses in sequence 080503_A5 (panel A) indicates a single-nucleotide insertion. The boxed region in panel C indicates clustered mutations in V3. Interestingly, these changes were limited to position 8 (Ile to Thr, Arg, or Lys) and position 25 (Asp to Asn, Glu, or Gly).
by as many as 66 nucleotides (2.7% maximum diversity), the sequences from within a lineage differed by no more than 4 nucleotides from each other and by no more than 2 nucleotides from the lineage consensus (<0.2% maximum diversity) (Fig. 3B). Finally, for subject ZM246F, a set of 26 identical or nearly identical sequences that differed from the consensus by 3 or fewer nucleotides and from each other by a maximum of 5 nucleotides (<0.2% maximum diversity) was identified for the first time point (Fig. 3C). Fourteen env sequences obtained 80 days later were still closely related to the original consensus sequence, although none was identical to an earlier sequence. Nonetheless, their consensus was still identical to the consensus of the earlier time point. Mathematical models of virus replication and diversification that assume exponential growth in the absence of immune selection predict a high proportion of identical sequences in acute and very early infections (B. F. Keele, unpublished data). For subject ZM247F, lineages 1 and 2, and subjects ZM246F and ZM249M, the proportions of identical sequences in the earliest samples were 5/7 (71%), 11/18 (61%), 16/27 (59%), and 11/24 (45%), consistent with the model expectations. Thus, we conclude that these consensus sequences correspond to the transmitted or early “founder” viruses that initiated productive clinical infection in these subjects.

**env gene evolution postseroconversion.** Although it is preferable to collect sequential plasma samples before and after seroconversion for the analysis of env evolution, this is not always feasible, especially in large-scale natural history studies or vaccine trials. We thus sought to determine what inferences could be drawn concerning the transmitted virus and its early evolution from the analysis of SGA-derived sequences from later time points, without the benefit of earlier samples. Figure 4 illustrates two examples. Samples were obtained from subject ZM180M during Fiebig stage V, no more than 124 days after infection, and from subject ZM206F during Fiebig stage VI, no more than 151 days after infection. Based on the extent of maximum env diversity in these subjects (0.39% and 0.47%, respectively) as well as the env phylogenetic trees, we determined that each had likely been productively infected by a single virus. Inspection of the Highlighter plot of sequences from subject ZM180M provided evidence of strong selection in the env region that overlaps the second exon of rev (Fig. 4A, left panel). When compared to the consensus sequence, each of 24 env/rev sequences contained a mutation within a region spanning 24 nucleotides; one of the sequences contained two mutations (Fig. 4A, middle panel). These 25 differences from the consensus comprised seven different mutations, each present in one to eight different sequences. All of these mutations were nonsynonymous in the rev reading frame, yielding seven different amino acid replacements within a single Rev 9-mer (Fig. 4A, right panel). We note that 9 amino acids is the length of a typical T-cell epitope. Given that there were only eight different mutations within the rev exon, we estimated the probability that seven would fall within a single 9-mer as less than $10^{-6}$. There was equally strong evidence of selection in the sample from subject ZM226F, where each of 35 env sequences contained at least one mutation or a deletion within a span of 9 codons (Fig. 4B, left panel). The 41 nucleotide changes comprised 16 different mutations (Fig. 4B, middle panel), all of which were nonsynonymous with respect to the consensus sequence (Fig. 4B, right panel). Given that there were 54 different nonsynonymous mutations within the env gene, we estimated the probability that 16 would fall within a single 9-mer as less than $10^{-12}$ (these probability calculations are conservative in that they assume that each of the observed mutations arose only once and that the multiple occurrences of some of the mutations reflect common ancestry).

**Contribution of viral recombination to env gene diversity.** Recombination can accelerate virus evolution and adaptation by reassortment of large sequence fragments (13, 33, 53). Subjects ZM229M and ZM215F exhibited the greatest within-patient env diversity, and we examined sequences from each for evidence of recombination by using the SGA-direct sequencing approach. Both subjects had evidence of infection by two or more viruses, based on their env diversity and estimates of timing to the MRCA (Table 2). Sequences from both subjects also contained numerous recombinants comprised of two or more transmitted lineages. The majority of these recombinants were likely generated in the newly infected individuals and not in the transmitting partner, since most of the mosaic sequences were unique and very closely related to only two transmitted parental strains. For example, sequence C15 from subject ZM229M contained seven putative recombination crossover sites, but all fragments between these crossover points were identical (or nearly identical) to those of one of the two transmitted variants, C2 or D19 (Fig. 5B). Sequences D1 and D10 were very closely related to C15 (Fig. 5A) and shared identical recombination breakpoints; this is most readily explained by assuming recombination in the newly infected subject followed by a clonal expansion of the mosaic virus. For subject ZM215F, there was an even greater number of recombinants which were derived from the two major transmitted variants (Fig. 5D). However, three mosaic sequences (D4, B8, and F17) were of particular interest since they contained short regions that were equally distant from the two principal lineages, suggesting that these viruses had recombined with additional viruses not represented in the phylogenetic tree (two examples are shown in Fig. 5E and F). Based on these and additional phylogenetic analyses (not shown), we estimated that subjects ZM229M and ZM215F each had acquired at least four different variants which had undergone extensive recombination. Thus, recombination accelerates sequence diversification in newly infected individuals.

**Comparison of SGA and bulk amplification methods for analyzing HIV-1 diversity.** Because previous studies of viral diversity in acute HIV-1 infection have generally utilized non-SGA methods, we sought to compare the two strategies side-by-side on the same clinical samples. Specifically, we asked whether SGA and bulk PCR methods were equivalent in providing an accurate representation of virus identity and diversity. The results showed that this was not the case: bulk amplification was prone to introducing a variety of methodological artifacts. Figure 6 depicts the phylogenies of 19 SGA-derived (Fig. 6A) and 34 bulk-amplified env sequences (Fig. 6B), both obtained from the sample obtained on 1 November 2003 from subject ZM247F. What is immediately apparent is that the terminal branches of the bulk-amplified env sequences are consistently longer than those of SGA-derived sequences. This is most obvious for the more abundant of the two trans-
FIG. 4. Evidence of immune selection in two subjects sampled at later Fiebig stages. (A) Highlighter analysis of SGA-derived env sequences (left panel) from subject ZM180M (Fiebig stage VI), and corresponding nucleotide (middle panel) and amino acid (right panel) sequence alignments from the overlapping rev gene. Boxes indicate mutations that cluster within a 27-nt region that corresponds to a 9-mer in the Rev protein sequence. (B) Highlighter analysis of SGA-derived env sequences (left panel) from subject ZM206F (Fiebig stage VI), and corresponding nucleotide (middle panel) and amino acid (right panel) sequence alignments of the V1 region. Boxes indicate mutations that cluster within a 27-nt region that corresponds to a 9-mer of the V1 loop. In Highlighter analyses, nucleotide substitutions and gaps are color coded. The consensus sequence is at the top of each alignment; lowercase letters show residues where mutations occurred. Dashes in the alignments indicate sequence identity to the consensus; dots indicate deletions.
FIG. 5. In vivo recombination in multiply-infected subjects. (A) Neighbor-joining tree of SGA-derived env sequences from subject ZM229M depicting two major transmitted variants (orange and purple), as well as five in vivo-generated recombinants (black). Asterisks at nodes indicate 90% or higher bootstrap values. The scale bars represent 0.005 nucleotide substitutions per site. (B-C) Diversity plots of two representative recombinants identified in panel A. The sequence distances of ZM229M_C15 (panel B) and ZM229M_D1 (panel C) are compared to those of representatives of the two parental lineages (orange and purple, respectively). The two recombinants contain between one and seven crossovers; schematic representations of their putative mosaic structures are shown below. (D) Neighbor-joining tree of SGA-derived env sequences from subject ZM215F depicting two major transmitted variants (orange and purple), as well as 19 different recombinants (black). Asterisks and scale bar are as described for panel A. (E-F) Diversity plots of two ZM215F recombinants containing “extraneous” sequences. The sequence distances of ZM215F_D4 (panel B) and ZM215F_F17 (panel C) are compared to those of representatives of the two parental lineages (orange and purple, respectively). Shaded areas indicate regions where ZM215F_D4 and ZM215F_F17 are equidistant from the two parental lineages, suggesting recombination with additional variants. Schematic representations of their putative mosaic structures are shown below.
mitted variants: in comparing 17 SGA-derived and 25 bulk-derived env sequences to their corresponding consensus sequence, we found average distances of 0.028% (0.73 substitutions) and 0.094% (2.4 substitutions) for SGA- and bulk PCR-derived sequences, respectively. This difference is highly significant ($t$-test; $P < 0.0001$) and can be attributed to Taq-induced misincorporations; such Taq-induced errors are evident in sequences of individual env clones derived from bulk PCR, but not in sequences of uncloned SGA amplicons since they represent only a very minor population. Similar results were obtained for the sample from a second subject (ZM249M) when 25 SGA-derived and 12 bulk-derived sequences were compared to their corresponding consensus env sequence (not shown). The average distance of the SGA sequences was 0.027% (0.7 substitutions), compared to 0.073% (1.9 substitutions) for the bulk sequences. Again, these differences were statistically significant ($P < 0.002$) and attributable to Taq-induced errors.

Bulk amplification also generated a surprisingly large number of recombinants compared to the SGA approach, where none were evident. This is exemplified in Fig. 6B, showing the results of bulk amplification of plasma from subject ZM247F (sample obtained 1 November 2003) depicting two transmitted variants (green and blue; sequences of the more-abundant variant are in green) and no recombinants. (B) Neighbor-joining tree of bulk PCR-derived env sequences amplified from the same specimen depicting two transmitted variants (green and blue; sequences of the more-abundant variant are in green) and eight recombinants (red). Asterisks at nodes indicate 90% or higher bootstrap values. The scale bars show 0.002 substitutions per site. (C and D) Diversity plots of two representative recombinants in panel B. The two recombinants contain one or more crossovers, the approximate positions of which are indicated by nucleotide position ($x$ axis) and shown schematically below the panels.

Comparison of SGA and bulk amplification methods for detection of specimen cross contamination. Quality control and assurance are always a concern in large-scale clinical trials, since specimens can be misidentified, cross-contaminated, or otherwise compromised. In the course of the present study, we obtained one plasma sample from subject ZM246F in which phylogenetic analyses suggested specimen cross contamination. This questionable sample revealed three SGA-derived env lineages (Fig. 7A), in contrast to the presence of a single viral lineage in this individual’s plasma 3 months earlier (Fig. 3C). We considered the possibility that subject ZM246F had become super-infected in the intervening 80 days, but the extent of sequence diversity within the blue lineages made this explanation extremely unlikely. When bulk PCR was performed on the questionable specimen (Fig. 7B), we obtained an even more complicated phylogenetic tree, this time revealing three lineages plus two additional recombinant viruses (ZM246F/M_BULK_41 and ZM246F/M_BULK_60). We resolved this quandary by reviewing clinic records and subjecting
the questionable plasma sample to DNA microsatellite analysis. The comparison of 15 polymorphic loci and one gender marker revealed three rather than two alleles in the questionable sample, but not in samples from three control subjects with heterogeneous HIV-1 infections (Table 4). This finding indicated that genetic material from two different individuals had been mixed in the questionable ZM246F sample. Analysis of peripheral blood mononuclear cell DNA from the partner (ZM246M) of subject ZM246F provided an explanation: the contaminating alleles in the questionable specimen were his (Table 4). Thus, during the processing of the couple’s blood on the 4 April 2003 clinic visit, the two plasma samples were inadvertently mixed. This inadvertent sample contamination was informative in that the two recombinant viruses whose diversity plots are shown in Fig. 7C and D could only have been generated in vitro as an artifact of Taq-induced template switching, since the two viral lineages from which they were derived never coexisted in the same individual. Moreover, a substantial number of the bulk-derived (17 out of 25), but not SGA-derived, sequences of ZM246M were identical or nearly identical, most likely because of target resampling and/or differences in cloning efficiency among different env sequences (20).

DISCUSSION

The primary objective of this study was to develop, validate, and implement a new experimental strategy for analyzing complete HIV-1 env genes and, eventually, complete HIV-1 genomes from plasma RNA in a manner that would accurately reflect their identity and composition in vivo. To this end, we adapted methods previously described by other investigators (27, 43–45) and tested them using in vitro-synthesized HIV-1 RNA transcripts of known sequence identity and plasma specimens from subjects with acute and early infections. Using an equal mixture of T7-synthesized RNA transcripts from two related but distinct HIV-1 env clones (BORId9.4F12 and 4F8), we carried out SGA-direct sequencing to estimate the rates of nucleotide misincorporation and recombination from this method. We observed 3.4 assay-related errors per 10,000 nucleotides, indicating a misincorporation rate of 0.034%. In addition, we observed a 0.01% rate of elongation errors within runs of the same nucleotide. We attribute these rates to a combination of T7 polymerase and Superscript III reverse transcriptase errors and note that Palmer and coworkers reported very similar values (0.011% nucleotide substitution errors and 0.022% elongation errors) for T7 RNA transcripts of FIG. 7. Bulk PCR-induced in vitro recombination in a mixture of two plasma samples. (A) Neighbor-joining tree of SGA-derived env sequences derived from a mixture of plasma from two infected subjects, ZM246F (green) and ZM246M (blue). Subject ZM246M was chronically infected with at least two major viral lineages (dark and light blue) that differed in their env sequences by approximately 6%. ZM246F was acutely infected with a virus from an unrelated individual which differed from the ZM246M env sequences by approximately 10%. (B) Neighbor-joining tree of bulk PCR-amplified env sequences from the same mixed-plasma specimen. In addition to viral lineages representing ZM246F (green) and ZM246M (blue), two additional recombinants (ZM246F/M_BULK_41 and ZM246F/M_BULK_60) are apparent (red). Asterisks at nodes indicate 90% or higher bootstrap values. The scale bars show 0.01 nucleotide substitutions per site. (C and D) Diversity plots of the two recombinants in panel B. The approximate position of recombination crossovers is indicated by nucleotide position (x axis) and schematically shown below the panels.

3966 SALAZAR-GONZALEZ ET AL. J. VIROL.
Nevertheless, we suspect that 8 of uncloned amplicons, since it can lead to single-nucleotide polymerase errors are avoided altogether (by direct sequencing). III errors. This is not a negligible error rate even when Taq method as described here for the analysis of HIV-1.

Thus, the total error rate of the SGA-direct sequencing script III error of nucleotides altogether; this yields an inferred rate of Superscript III error in our analyses, since in some pa-

found as many as 48 of 52 (92%) plasma virion env sequences to be identical, with the remaining four varying by only five nucleotides altogether; this yields an inferred rate of Superscript III error of $<4 \times 10^{-5}$ (G. M. Shaw and B. H. Hahn, unpublished data). For certain applications, such as identifying transmitted or early founder sequences, infrequent nucleotide misincorporations are of no consequence since many independently generated sequences are analyzed together and all coalesce to a single consensus. However, if SGA-derived env genes from chronically infected subjects, in whom most circulating viruses are unique (27), are analyzed, then single-base-pair misincorporations due to Superscript III error can be a confounding variable. Finally, we examined whether cDNA synthesis using Superscript III generated recombinant viral sequences in vitro. Among 109 complete env sequences, corresponding to 278,000 nucleotides, we observed no recombinants. We also observed no instances of intragenic recombination or of insertion, deletion, or duplication. Our findings are thus in agreement with those of Palmer and coworkers, who also found no evidence of Superscript III-mediated recombination, insertion, deletion, or duplication in any of 50 genomes (66,000 nucleotides) analyzed (27). Although we did not formally evaluate the substitution rates and template-switching frequencies for Taq polymerase when using the bulk amplification method, we frequently encountered Taq-induced recombinants and/or misincorporations in bulk-amplified sequences (Fig. 6 and 7). Moreover, in a separate study where we obtained functional env clones from SGA-derived amplicons (B. F. Keele, unpublished data), we identified (and discarded) numerous clones that contained Taq-induced errors and thus did not correspond to the env consensus sequences. These clones were excluded from subsequent biological analyses since they did not represent viruses present in the patient but rather in vitro artifacts. We thus conclude that only SGA-based strategies can unequivocally identify genetically linked mutations and that assay-related nucleotide misincorporations and recombination frequencies are much lower for SGA approaches than with other strategies.

A second objective of this study was to evaluate in a field trial setting the ability of SGA-direct sequencing strategies to decipher transmitted clade C (or other non-clade B) viruses and their early evolution in a time frame typical of vaccine trial follow-up schedules (every 3 months). In a companion study of acute and early subtype B infections (B. F. Keele, unpublished data), we analyzed 15 tetranucleotide repeat loci and the amelogenin gender-determining marker in a single PCR amplification; all identifiable alleles per locus are shown.

### Table 4. Microsatellite analysis of plasma samples with heterogeneous viral sequences

<table>
<thead>
<tr>
<th>Locus</th>
<th>Allele(s) identified in sample obtained (mo/yr) from indicated subject</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ZM229M</td>
</tr>
<tr>
<td>D8S1179</td>
<td>14, 16</td>
</tr>
<tr>
<td>D211S11</td>
<td>30, 32, 31</td>
</tr>
<tr>
<td>D7S820</td>
<td>11, 12</td>
</tr>
<tr>
<td>CSF1PO</td>
<td>7, 12</td>
</tr>
<tr>
<td>D3S1358</td>
<td>15, 17</td>
</tr>
<tr>
<td>TH01</td>
<td>6, 8</td>
</tr>
<tr>
<td>D113S317</td>
<td>12, 13</td>
</tr>
<tr>
<td>D165S39</td>
<td>12, 13</td>
</tr>
<tr>
<td>D251S38</td>
<td>22, 23</td>
</tr>
<tr>
<td>D18S51</td>
<td>14, 16, 16</td>
</tr>
<tr>
<td>vWA</td>
<td>13, 20</td>
</tr>
<tr>
<td>TPOX</td>
<td>9, 11</td>
</tr>
<tr>
<td>D185S11</td>
<td>14, 16, 16</td>
</tr>
<tr>
<td>D55S18</td>
<td>12</td>
</tr>
<tr>
<td>FGA</td>
<td>20, 22</td>
</tr>
<tr>
<td>Amelogenin Y</td>
<td>Y</td>
</tr>
<tr>
<td>Amelogenin X</td>
<td>Y</td>
</tr>
</tbody>
</table>

*a* Microsatellite analyses were performed using DNA extracted from plasma samples, except for subject ZM246M for whom only peripheral blood mononuclear cell DNA was available. Loci were amplified using an AmpFSTR Identifiler kit (Applied Biosystems, Foster City, CA). This PCR amplification kit is a short tandem repeat multiplex assay that amplifies 15 tetranucleotide repeat loci and the amelogenin gender-determining marker in a single PCR amplification; all identifiable alleles per locus are shown.

*b* This sample was identified as an inadvertent mixture of plasma from two different individuals (ZM246F and ZM246M).

*c* Subject ZM246M is the chronically infected partner of subject ZM246F but not the individual who transmitted virus to subject ZM246F (see Fig. 7 and associated text for explanation).
data), we studied 51 subjects in Fiebig stages I/II and 26 subjects in Fiebig stages III/IV; with such early sampling we found that we could infer transmitted or early founder env sequences in most patients, including those infected by more than one virus. A mathematical model of early HIV-1 replication and diversification described in that study provided the theoretical basis for identifying transmitted or early founder viral genomes. Here, we were less certain whether this approach would be applicable since the frequency of patient sampling was less, samples were obtained from the majority of subjects (9/12) weeks to months after infection (Fiebig stage V or VI), and the genetic subtypes analyzed were non-clade B. Nonetheless, we show here for three subjects (ZM249M, ZM247F, and ZM246F) studied prior to seroconversion that the phylogenetic trees and Highlighter analyses allow for an unambiguous identification of the transmitted or early founder virus(es). For six homogeneous-transmission cases studied later in the infection process (ZM178F, ZM180M, ZM184F, ZM206F, ZM231F, and ZM235), the env sequences also coalesced in a time frame consistent with transmitted or early founder viruses (Fig. 3, Table 3). However, this was not the case for individuals who were infected by more than one virus and were sampled for the first time at later time points (e.g., Fiebig stages V/VI); in these instances, identification of the transmitted viruses was precluded by more-extensive nucleotide substitutions, as well as in vivo recombination. This limitation notwithstanding, our findings for primary clade C infections mirror data obtained for primary clade B infections (B. F. Keele, unpublished data): sequences of transmitted or early founder env genes can be readily inferred from SGA-derived sequences if subjects are sampled sufficiently early (Fiebig stages I to IV) and, in some cases, also at later time points (Fiebig stages V to VI) but only if the infection was initiated by a single virus.

A third study objective was to determine if insights into selection pressures on virus replication could be inferred from SGA-derived sequences from single time points distant from the transmission event. We show two examples of this. In Fig. 4A, the results of Highlighter analysis of 24 env sequences from a subject at Fiebig stage V (ZM180M) are shown, illustrating a heavy concentration of nucleotide substitutions in the region of env that overlaps the second exon of rev. The actual nucleotide substitutions are shown in the middle panel. Each of the 24 sequences was found to contain one or more of seven different mutations when compared to the consensus sequence. Because of the large number of different mutations, it was possible to infer the consensus sequence in this region and across the entire env gene. Moreover, all of the nucleotide substitutions concentrated within this 9-codon stretch of the Rev open reading frame were nonsynonymous (Fig. 4A, right panel). Statistical analysis ruled out the possibility that this cluster of mutations arose by chance, and the observation most likely reflects selection for sequences with amino acid differences. Although viably frozen lymphocytes were not available for cytotoxic T-lymphocyte studies, this subject’s HLA profile was typed as A*2901, A*3002, B*1510, B*4201, Cw*0304, Cw*17(01-03). The Rev sequence under selection pressure is LAEVYPLPLLPLNLGID, with the variable region underlined. There are several HLA-B42 and HLA-C motifs that overlap this region of interest, where potential second-position and C-terminal anchor motifs are indicated as follows: XPXXXXXXXX (B*4201); XPXXXXXXXX (B*4201); and XAXAXXXAXXX (Cw*1701, Cw*1702, and Cw*0304). The potential B*4201 epitopes are embedded directly in the region that is the focus of the mutations, while the potential HLA-C epitope is slightly offset. People who carry Cw*03 tend to have a reaction to the peptide that spans this region more often than people without Cw*03, suggesting that a Cw*03 epitope is present and recognized in many subtype C infections (B. T. Korber, unpublished data). A similar pattern of mutations in a Rev 9-mer was recently identified in a subtype B-infected subject, and in this individual, HLA-restricted cytotoxic T-lymphocyte reactivity was confirmed by enzyme-linked immunospot assay and gamma interferon induction (G. M. Shaw and P. Borrow, unpublished data). In the sample from subject ZM206, obtained in stage VI, there was equally strong evidence of selection, again within a 9-amino-acid fragment but this time within the variable loop 1 (V1) region of Env (Fig. 4B). Remarkably, 33 out of 35 sequences had 1 or more of 16 different point mutations within this region, while the other 2 sequences had deletions. This again allowed for the identification of a consensus sequence that likely corresponds to the transmitted or early founder sequence. Again, these changes meant that every sequence sampled encoded a different amino acid sequence when compared to the consensus, and again, the likelihood of such a concentration of mutations occurring by chance was estimated to be extremely low. The HLA profile of this subject was A*0202, A*2301; B*1510, B*180101; Cw*0501, Cw*1601, and the region of extreme selection is GSSKANDNNVNITSD. There are no obvious anchor motifs for the relevant HLA s in this sequence, although KANDNVNI could fit an A*0202 binding pocket (P. Goulder, personal communication). Alternatively, the observed cluster of V1 mutations could be the result of neutralizing-antibody escape (34). Taken together, these results indicate that molecular patterns of virus adaptation can be inferred even in samples obtained several months after transmission from subjects for whom earlier specimens are not available for comparison.

The SGA-direct sequencing approach is ideally suited to the evaluation of genetic linkages, as described by Palmer et al. for the analysis of drug resistance mutations in the pro-pol genes (27). We sought to determine if SGA-direct sequencing might reveal env gene recombination in subjects acutely infected by more than one virus and then to compare recombination frequencies between SGA and bulk amplification methods. Figure 5 illustrates multiple examples of viral recombination in the two subjects at Fiebig stage VI, ZM229M and ZM215F. Interestingly, in subject ZM215F, the recombination involved not only two principal transmitted virus lineages but also additional sequences not otherwise represented in the sequence set. Exhaustive phylogenetic analyses indicated that subjects ZM215F and ZM229M had each been infected by four or more viruses (Table 3). Thus, in these two heterogeneous infections, viral diversification was accelerated by extensive recombination.

Although viral recombination assessed by SGA methods can be complicated and nearly indecipherable in multiply-infected individuals at later time points, this problem is magnified if bulk PCR methods are used. In Fig. 6, we show results for a subject at Fiebig stage II who was infected by two variants. The analysis of a total of 44 SGA-derived env sequences reveals no
evidence of recombination, but 8 of 34 bulk PCR-derived sequences are mosaic, each exhibiting different breakpoint patterns. Artificial Taq-mediated template switching was also demonstrated in an example of a cross-contaminated plasma specimen from two subjects (ZM246F and ZM246M) who were infected by unrelated viruses (Fig. 7). In the contaminated specimen, the SGA method clearly distinguished a single ZM246F lineage from two ZM246M lineages with no recombinant viruses among them. Conversely, the bulk method generated two mosaic sequences in vitro that did not exist in vivo. From the results of our analyses and those of other reports (27, 43–45, 54), we conclude that SGA methods do not generate in vitro recombinants, whereas bulk methods commonly do. Bulk amplification-cloning-sequencing strategies are also susceptible to Taq-induced nucleotide misincorporation, template resampling, and cloning bias. These limitations may not be problematic for certain applications. However, if a goal is to obtain sequences of appreciable length that correspond to HIV-1 genomes that exist in vivo, then SGA-direct sequencing approaches have distinct advantages.

The results of the present study, together with those of Palmer et al. (27) and B. F. Kellee (unpublished), illustrate new scientific avenues for deciphering HIV-1 transmission and patterns of early virus diversification. Based on the data reported here, it is likely that these new approaches will help to clarify the genetic and biological complexity of viruses transmitted by different routes and under various clinical circumstances, all factors that may be important in the design and assessment of candidate vaccines, antiretroviral drugs, or microbicides. In samples from clinical trial participants, it may be possible to use SGA-based methods to generate not only transmitted and evolved env genes but all HIV-1 genes of interest. Recently, we have shown for seven clade B- or C-infected subjects that complete (9 kb) HIV-1 genomes corresponding to the transmitted or early founder virus can be identified by SGA-direct sequencing methods (39). Such approaches may be useful in mapping linked mutations conferring escape from cellular and humoral immune responses in naive or vaccinated individuals.

ACKNOWLEDGMENTS

We thank John Coffin and Sarah Palmer for technical advice and assistance in establishing high-throughput SGA methodologies; James Theiler and Tammy Bhattacharya for statistical support; Jianming Tang and Richard Kaslow for HLA analyses; Philip Goulder for helpful discussions; the participants, staff, and project management group of the Lusaka HIV-1 transmission cohort for providing patient samples; the participants, staff, and project management group of the International AIDS Vaccine Initiative for cohort recruitment and maintenance; and the UAB Medical Genomics Laboratory for performing microsatellite analyses. We thank David McPherson, Yalu Chen, and Barry Cochran for expert technical consultation and assistance and Jamie White for artwork and manuscript preparation.

This work was supported by grants from the Bill and Melinda Gates Foundation Grand Challenges Program (grant 37874), the National Institutes of Health (grants R01 AI58706, R01 AI51231, P01 AI061734, P30 AI27767, and P30 AI05410), the Center for HIV/AIDS Vaccine Immunology (grant U01 AI067854), an internally directed research grant for vaccine design at the Los Alamos National Laboratory, and the International AIDS Vaccine Initiative for cohort recruitment and maintenance.

REFERENCES


