High Multiplicity Infection by HIV-1 in Men Who Have Sex with Men

Hui Li1,9, Katharine J. Bar1,9, Shuyi Wang1, Julie M. Decker1, Yalu Chen1, Chuanxi Sun1, Jesus F. Salazar-Gonzalez1, Maria G. Salazar1, Gerald H. Learn1, Charity J. Morgan2, Joseph E. Schumacher1, Peter Hraber3, Elena E. Giorgi3,4, Tanmoy Bhattacharya5,6, Bette T. Korber3,6, Alan S. Perelson3, Joseph J. Eron7, Myron S. Cohen7, Charles B. Hicks8, Barton F. Haynes6, Martin Markowitz9,10, Brandon F. Keele11, Beatrice H. Hahn1,12, George M. Shaw1,12.

1 Department of Medicine, University of Alabama at Birmingham, Birmingham, Alabama, United States of America, 2 Department of Biostatistics, University of Alabama at Birmingham, Birmingham, Alabama, United States of America, 3 Theoretical Biology and Biophysics (T6), Los Alamos National Laboratory, Los Alamos, New Mexico, United States of America, 4 Department of Mathematics and Statistics, University of Massachusetts, Amherst, Massachusetts, United States of America, 5 Nuclear and Particle Physics, Astrophysics and Cosmology (T-2), Los Alamos National Laboratory, Los Alamos, New Mexico, United States of America, 6 Santa Fe Institute, Santa Fe, New Mexico, United States of America, 7 Department of Medicine, University of North Carolina, Chapel Hill, North Carolina, United States of America, 8 Department of Medicine, Duke University, Durham, North Carolina, United States of America, 9 Aaron Diamond AIDS Research Center, New York, New York, United States of America, 10 Rockefeller University, New York, New York, United States of America, 11 SAIC-Frederick, National Cancer Institute, Frederick, Maryland, United States of America, 12 Department of Microbiology, University of Alabama at Birmingham, Birmingham, Alabama, United States of America.

Abstract

Elucidating virus-host interactions responsible for HIV-1 transmission is important for advancing HIV-1 prevention strategies. To this end, single genome amplification (SGA) and sequencing of HIV-1 within the context of a model of random virus evolution has made possible for the first time an unambiguous identification of transmitted/founder viruses and a precise estimation of their numbers. Here, we applied this approach to HIV-1 env analyses in a cohort of acutely infected men who have sex with men (MSM) and found that a high proportion (10 of 28; 36%) had been productively infected by more than one virus. In subjects with multivariate transmission, the minimum number of transmitted viruses ranged from 2 to 10 with viral recombination leading to rapid and extensive genetic shuffling among virus lineages. A combined analysis of these results, together with recently published findings based on identical SGA methods in largely heterosexual (HSX) cohorts, revealed a significantly higher frequency of multivariate transmission in MSM than in HSX [19 of 50 subjects (38%) versus 34 of 175 subjects (19%); Fisher’s exact p = 0.008]. To further evaluate the SGA strategy for identifying transmitted/founder viruses, we analyzed 239 overlapping 5’ and 3’ half genome or env-only sequences from plasma viral RNA (vRNA) and blood mononuclear cell DNA in an MSM subject who had a particularly well-documented virus exposure history 3–6 days before symptom onset and 14–17 days before peak plasma viremia (47,600,000 vRNA molecules/ml). All 239 sequences coalesced to a single transmitted/founder virus genome in a time frame consistent with the clinical history, and a molecular clone of this genome encoded replication competent virus in accord with model predictions. Higher multiplicity of HIV-1 infection in MSM compared with HSX is consistent with the demonstrably higher epidemiological risk of virus acquisition in MSM and could indicate a greater challenge for HIV-1 vaccines than previously recognized.


Editor: Daniel C. Douek, NIH/NIAID, United States of America

Received December 18, 2009; Accepted April 1, 2010; Published May 13, 2010

This is an open-access article distributed under the terms of the Creative Commons Public Domain declaration which stipulates that, once placed in the public domain, this work may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose.

Funding: This work was supported by the Center for HIV/AIDS Vaccine Immunology (CHAVI) and by grants and contracts from the National Institutes of Health (AI67854, AI27767, AI41534, AI28433, CA13148, RR24143, HHSN266200400088C) and the Bill & Melinda Gates Foundation (37874). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: gshaw@uab.edu
9 These authors contributed equally to this work.

Introduction

An effective sterilizing HIV-1 vaccine ideally should target virus in the earliest stages of transmission, prior to dissemination and establishment of persistent infection [1,2,3,4]. To be broadly protective, such a vaccine must defend against a genetically diverse set of viruses transmitted by different sexual practices and risk behaviors. Results from the recently reported ‘Thai Trial’ RV144 of an experimental HIV-1 vaccine showed a decrease in virus acquisition of 31.2% (p = 0.04) based on a modified intention-to-treat analysis and a trend for greater vaccine effectiveness in those subjects identified as practicing lower risk behaviors [5]. These findings suggest that an HIV-1 vaccine might be more efficacious in preventing infection by some exposure routes than others [5,6,7].

Recently, we and others employed SGA, direct sequencing, and a model of random virus evolution to identify those viruses responsible for transmission and productive clinical infection in several largely heterosexual cohorts with acute HIV-1 subtype A, B or C infection [8,9,10,11,12] and in Indian rhesus macaques.
inoculated intra-rectally with SIVmac251 or SIVsmmE660 [13]. This experimental approach allows for the distinction of transmitted/founder viruses that differ by as little as a single nucleotide [10,13]. SGA-direct sequencing also makes possible the identification of transmitted viral sequences in linked transmissions, thereby enabling the unambiguous tracking of viruses from donor to recipient across mucosal surfaces [9,13], and the molecular cloning and analysis of those viruses actually responsible for productive clinical infection as a single exposure event. We go on to molecularly clone the first full-length transmitted/founder subtype B HIV-1 virus and show that it is highly replicative in human CD4+ T-cells but not macrophages. Our study provides the first comparative, quantitative analysis of the multiplicity of HIV-1 infection in the two primary risk groups—MSM and heterosexuals—driving the global pandemic, and we discuss the implications of the findings to HIV-1 vaccine development and prevention research.

**Results**

**Study Subjects**

SGA-direct sequencing was used to identify and enumerate transmitted/founder env sequences in 28 acutely infected MSM subjects who reported homosexual exposure as their primary HIV-1 risk behavior and who denied injection drug use (IDU) (Table 1). At the time of study, 14 subjects were HIV-1 ELISA negative/western immunoblot (WB) negative (Fiebig stage II), 2 were ELISA+/WB− (Fiebig stage III), 6 were ELISA+/WB indeterminate (Fiebig stage IV) and 6 were ELISA+/WB+/p31− (Fiebig stage V) [10,33]. Subjects were identified based on clinical symptoms of an acute retroviral syndrome, routine HIV testing in a health care setting, or contact tracing of an HIV-1 infected index case. Clinical histories of sexually transmitted diseases were not available. Envelope sequences from chronically infected sexual partners of two acutely infected subjects were also evaluated.

**HIV-1 Env Diversity Analysis**

A total of 1307 full-length env genes encoding gp160 were sequenced from plasma vRNA (median of 40 sequences per subject; range 23–89). In a composite neighbor-joining (NJ) phylogenetic tree (Fig. 1), viral sequences formed distinct patient-specific monophyletic lineages, each with high statistical support. Sequences from known sexual partners, including two acute-to-acute (AD77 to AD75 and AD83 to 04013240) and two chronic-to-acute (LACU9000 to HOBR0961 and AD18 to AD17) transmission pairs, also clustered significantly together (Fig. 1). All sequences were HIV-1 subtype B. Among the 28 acutely infected subjects, maximum within-patient env diversities ranged from 0.12% to 6.82% (Table 2). Sequences from 22 of these individuals had distinctly lower env diversities (<0.75%) compared with env diversities from six others (>1.25%). The latter diversity is inconsistent with single virus transmission within the time frame of acute and early infection (Fiebig stage I–V) [8,9,10,34], while env diversity <0.75% is consistent either with single variant transmission or with transmission of two or more closely related viruses. Phylogenetic and Highlighter analyses of env sequences distinguished between these possibilities for each subject (Fig. 2; see also www.hiv.lanl.gov/content/sequence/HIV/USER_ALIGNMENTS/Li). Fig. 2A shows sequences from a subject (04013440) who was infected by a single virus, Fig. 2B a subject (04013211) infected by two viruses differing by only 4 nucleotides out of 2619 (0.15%), Fig. 2C a subject (04013383) infected by two viruses differing by 65 of 2547 nucleotides (2.55%), and Fig. 2D a subject (04013448) infected by four viruses differing by as many as 47 of 2655 nucleotides (1.79%) with additional sequences showing recombination between the transmitted/founder lineages. Altogether, we determined that 10 of 28 subjects (36%) had been productively infected by more than one virus (Table 2).

**Model Analysis of HIV-1 Diversification**

We next analyzed the env sequences using a mathematical model of random virus evolution that we previously described [10,13,34]. Sequences resulting from multivariant transmission, APOBEC hypermutation, early stochastic mutation, selection by cytotoxic T-cells, or recombination violate model predictions (Table 2) [10,34]. Once these confounders were accounted for, lineage-specific env sequences from each subject conformed to model predictions and coalesced to most recent common ancestor sequences at or near the time of virus transmission estimated from clinical histories and laboratory staging. These results thus corroborated a large body of evidence supporting the SGA-direct sequencing strategy for identifying transmitted/founder viruses [8,9,10,11,12,13,35]. As an
additional test of the model’s validity, we asked in subject AD17 whose history of virus exposure was particularly well-documented (Table 1 and Fig. 3), if plasma vRNA and PBMC viral DNA (vDNA) sequences spanning the complete (9.2 Kb) viral genome coalesced to the same viral sequence as did env-only sequences and if a molecular clone of this viral genome encoded replication competent virus, as would be expected for an authentic transmitted/founder virus. For this analysis, we used SGA-direct sequencing to determine env-only sequences (n = 51) and overlapping 5’ (n = 92) and 3’ (n = 96) half genome sequences (Fig. 4). All 239 vRNA and vDNA sequences coalesced to a single transmitted/founder genome in a time frame consistent with the clinical history of virus exposure as recently as 11 days earlier. The estimated time to a most recent common ancestor (MRCA) sequence for env-only sequences was 8 days (95% CI 5–11) and for all sequences 6–11 days (CI 3–14). MRCA estimates are frequently lower than clinical estimates [10] or experimentally determined intervals between transmission and virus sampling in the rhesus macaque-SIV infection model [13] because of purifying selection or variance in estimated parameters of virus replication [34]. The inferred transmitted/founder viral genome in subject AD17 contained intact LTR-gag-pol-env-vif-vpr-tat-rev-vpu-env-nef-LTR elements, a finding we have replicated for transmitted/founder viruses from 38 other subjects infected by HIV-1 subtypes A, B, C or D ([12] and H.L. and G.M.S., unpublished). A proviral clone (pAD17.1) of the transmitted/founder viral genome from subject AD17 (Fig. 5A), when transfected into 293T cells, produced virions that were infectious and highly replicative in human CD4+ T-cells, but interestingly, not in monocyte-derived macrophages from the same normal donors (Fig. 5B). pAD17.1 virus was CCR5 tropic in JCV3BL-13 cells (Fig. 5C) and in GHOST(3) cells [36], where it infected cells bearing CD4 and CCR5 but not CD4 and CXCR4 (G.M.S., unpublished).

High Multiplicity Infection Followed by Recombination

Extremes in HIV-1 diversity in acute infection could be informative regarding biological events underlying virus transmission. Subject 04013171 had the greatest env diversity (6.82%) (Table 1 and Fig. 3), if plasma vRNA and PBMC viral DNA (vDNA) sequences spanning the complete (9.2 Kb) viral genome coalesced to the same viral sequence as did env-only sequences and if a molecular clone of this viral genome encoded replication competent virus, as would be expected for an authentic transmitted/founder virus. For this analysis, we used SGA-direct sequencing to determine env-only sequences (n = 51) and overlapping 5’ (n = 92) and 3’ (n = 96) half genome sequences (Fig. 4). All 239 vRNA and vDNA sequences coalesced to a single transmitted/founder genome in a time frame consistent with the clinical history of virus exposure as recently as 11 days earlier. The estimated time to a most recent common ancestor (MRCA) sequence for env-only sequences was 8 days (95% CI 5–11) and for all sequences 6–11 days (CI 3–14). MRCA estimates are frequently lower than clinical estimates [10] or experimentally determined intervals between transmission and virus sampling in the rhesus macaque-SIV infection model [13] because of purifying selection or variance in estimated parameters of virus replication [34]. The inferred transmitted/founder viral genome in subject AD17 contained intact LTR-gag-pol-env-vif-vpr-tat-rev-vpu-env-nef-LTR elements, a finding we have replicated for transmitted/founder viruses from 38 other subjects infected by HIV-1 subtypes A, B, C or D ([12] and H.L. and G.M.S., unpublished). A proviral clone (pAD17.1) of the transmitted/founder viral genome from subject AD17 (Fig. 5A), when transfected into 293T cells, produced virions that were infectious and highly replicative in human CD4+ T-cells, but interestingly, not in monocyte-derived macrophages from the same normal donors (Fig. 5B). pAD17.1 virus was CCR5 tropic in JCV3BL-13 cells (Fig. 5C) and in GHOST(3) cells [36], where it infected cells bearing CD4 and CCR5 but not CD4 and CXCR4 (G.M.S., unpublished).

High Multiplicity Infection Followed by Recombination

Extremes in HIV-1 diversity in acute infection could be informative regarding biological events underlying virus transmission. Subject 04013171 had the greatest env diversity (6.82%) (Table 2). This subject admitted to unprotected receptive anal intercourse with multiple partners over a single eight hour period four weeks before the onset of flu-like symptoms, consistent with his Fiebig IV staging. Fig. 6 shows a NJ tree and Highlighter plot of 86 plasma derived env sequences, which revealed 10 unique...
transmitted/founder virus lineages. In addition, 20 inter-lineage recombinants were identified based on shared polymorphisms in the Highlighter plot with corroboration by Recco analysis [37]. Among these recombinant sequences, the Hudson-Kaplan test [38] indicated a minimum of 44 recombination breakpoints. Interestingly, sequences corresponding to 4 of the 10 virus lineages in subject 04013171 were sampled only once. We could be confident that these represented unique transmitted/founder viruses and not recombinants between two or more predominant virus lineages because of the large number of unique nucleotide changes in each sequence that far exceeded the diversity observed empirically [8,9,10,13] or estimated to occur based on mathe-

Figure 1. Neighbor-joining (NJ) tree of full-length HIV-1 gp160 env sequences from 28 acutely infected subjects and 2 chronically infected sexual partners. Two chronic-to-acute (LACU9000 to HOBR0961 and AD18 to AD17) and two acute-to-acute (AD77 to AD75 and AD83 to 04013240) transmissions were documented, with donor sequences shown in blue and recipient sequences shown in green. Individual sequences with APOBEC G-to-A hypermutation were excluded from the analysis. Bootstrap values (≥70%) are shown for intra-subject clusters, partner pairs, and additional sequences with evidence of epidemiologic linkage. The horizontal scale bar represents 1.0% genetic distance.
doi:10.1371/journal.ppat.1000890.g001
<table>
<thead>
<tr>
<th>Subject</th>
<th>Fiebig stage</th>
<th>No. of SGA envs</th>
<th>Maximum length of env</th>
<th>Maximum Hamming distance (HD)</th>
<th>Maximum diversity %</th>
<th>APOBEC-mediated hyper-mutation</th>
<th>Poisson estimated days since MRCA (95% CI)</th>
<th>Lambda</th>
<th>Goodness of fit P value</th>
<th>HD fit to poisson</th>
<th>Star phylogeny</th>
<th>Explanation for deviation from model</th>
<th>No. of transmitted/founder viruses</th>
<th>Recombinants</th>
</tr>
</thead>
<tbody>
<tr>
<td>04013171</td>
<td>4</td>
<td>23/86</td>
<td>2625</td>
<td>179</td>
<td>6.82%</td>
<td>No</td>
<td>662 (525, 792)</td>
<td>41.370</td>
<td>0.000</td>
<td>no</td>
<td>no</td>
<td>multiple variant transmission</td>
<td>≥10</td>
<td>20/86</td>
</tr>
<tr>
<td>04013211</td>
<td>3</td>
<td>30/30</td>
<td>2619</td>
<td>9</td>
<td>0.34%</td>
<td>No</td>
<td>54 (44, 64)</td>
<td>3.380</td>
<td>0.000</td>
<td>no</td>
<td>no</td>
<td>multiple variant transmission</td>
<td>2</td>
<td>0/30</td>
</tr>
<tr>
<td>04013226</td>
<td>2</td>
<td>33/33</td>
<td>2574</td>
<td>5</td>
<td>0.19%</td>
<td>Yes</td>
<td>15 (9, 21)</td>
<td>0.936</td>
<td>0.825</td>
<td>yes</td>
<td>yes</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>04013240</td>
<td>2</td>
<td>33/66</td>
<td>2568</td>
<td>14</td>
<td>0.55%</td>
<td>No</td>
<td>66 (56, 76)</td>
<td>4.041</td>
<td>0.000</td>
<td>no</td>
<td>no</td>
<td>multiple variant transmission</td>
<td>3</td>
<td>6/66</td>
</tr>
<tr>
<td>04013242</td>
<td>4</td>
<td>37/37</td>
<td>2580</td>
<td>5</td>
<td>0.19%</td>
<td>Yes</td>
<td>16 (10, 22)</td>
<td>0.992</td>
<td>0.902</td>
<td>yes</td>
<td>yes</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>04013291</td>
<td>5</td>
<td>25/25</td>
<td>2559</td>
<td>8</td>
<td>0.31%</td>
<td>No</td>
<td>61 (53, 70)</td>
<td>3.737</td>
<td>0.100</td>
<td>no</td>
<td>no</td>
<td>CTL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>04013296</td>
<td>2</td>
<td>25/25</td>
<td>2640</td>
<td>7</td>
<td>0.27%</td>
<td>No</td>
<td>34 (26, 42)</td>
<td>2.113</td>
<td>0.859</td>
<td>yes</td>
<td>yes</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>04013321</td>
<td>2</td>
<td>49/49</td>
<td>2556</td>
<td>8</td>
<td>0.31%</td>
<td>Yes</td>
<td>39 (33, 45)</td>
<td>2.370</td>
<td>0.212</td>
<td>yes</td>
<td>no</td>
<td>early stochastic mutations</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>04013327</td>
<td>4</td>
<td>24/24</td>
<td>2535</td>
<td>4</td>
<td>0.16%</td>
<td>No</td>
<td>11 (5, 18)</td>
<td>0.667</td>
<td>0.390</td>
<td>yes</td>
<td>yes</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>04013383</td>
<td>2</td>
<td>23/70</td>
<td>2547</td>
<td>70</td>
<td>2.75%</td>
<td>No</td>
<td>554 (534, 572)</td>
<td>33.580</td>
<td>0.000</td>
<td>no</td>
<td>no</td>
<td>multiple variant transmission</td>
<td>2</td>
<td>0/70</td>
</tr>
<tr>
<td>04013396</td>
<td>4</td>
<td>39/39</td>
<td>2577</td>
<td>5</td>
<td>0.19%</td>
<td>No</td>
<td>16 (11, 21)</td>
<td>0.974</td>
<td>0.897</td>
<td>yes</td>
<td>yes</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>04013419</td>
<td>2</td>
<td>27/78</td>
<td>2565</td>
<td>80</td>
<td>3.12%</td>
<td>Yes</td>
<td>548 (506, 591)</td>
<td>33.500</td>
<td>0.000</td>
<td>no</td>
<td>no</td>
<td>multiple variant transmission</td>
<td>3</td>
<td>4/78</td>
</tr>
<tr>
<td>04013440</td>
<td>2</td>
<td>30/30</td>
<td>2580</td>
<td>4</td>
<td>0.16%</td>
<td>No</td>
<td>23 (18, 28)</td>
<td>1.379</td>
<td>0.116</td>
<td>yes</td>
<td>yes</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>04013446</td>
<td>3</td>
<td>23/23</td>
<td>2594</td>
<td>7</td>
<td>0.27%</td>
<td>No</td>
<td>24 (12, 35)</td>
<td>1.450</td>
<td>0.000</td>
<td>early stochastic mutations</td>
<td>no</td>
<td>yes</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>04013448</td>
<td>2</td>
<td>15/54</td>
<td>2625</td>
<td>51</td>
<td>1.94%</td>
<td>No</td>
<td>411 (367, 456)</td>
<td>25.716</td>
<td>0.000</td>
<td>no</td>
<td>no</td>
<td>multiple variant transmission</td>
<td>4</td>
<td>5/54</td>
</tr>
<tr>
<td>AD17</td>
<td>2</td>
<td>51/51</td>
<td>2544</td>
<td>4</td>
<td>0.16%</td>
<td>Yes</td>
<td>8 (5, 11)</td>
<td>0.471</td>
<td>0.162</td>
<td>yes</td>
<td>yes</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AD75</td>
<td>2</td>
<td>54/54</td>
<td>2556</td>
<td>3</td>
<td>0.12%</td>
<td>Yes</td>
<td>9 (6, 13)</td>
<td>0.555</td>
<td>0.774</td>
<td>yes</td>
<td>yes</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AD77</td>
<td>5</td>
<td>40/40</td>
<td>2556</td>
<td>11</td>
<td>0.43%</td>
<td>No</td>
<td>84 (74, 95)</td>
<td>5.127</td>
<td>0.151</td>
<td>no</td>
<td>no</td>
<td>multiple variant transmission</td>
<td>3</td>
<td>nd</td>
</tr>
<tr>
<td>AD83</td>
<td>5</td>
<td>44/44</td>
<td>2568</td>
<td>33</td>
<td>1.29%</td>
<td>Yes</td>
<td>66 (36, 95)</td>
<td>4.000</td>
<td>0.000</td>
<td>no</td>
<td>no</td>
<td>multiple variant transmission</td>
<td>3</td>
<td>3/44</td>
</tr>
<tr>
<td>HOB0R961</td>
<td>2</td>
<td>42/42</td>
<td>2586</td>
<td>4</td>
<td>0.15%</td>
<td>No</td>
<td>17 (13, 22)</td>
<td>1.022</td>
<td>0.503</td>
<td>yes</td>
<td>yes</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INW8632</td>
<td>2</td>
<td>46/46</td>
<td>2580</td>
<td>4</td>
<td>0.16%</td>
<td>No</td>
<td>12 (8, 16)</td>
<td>0.737</td>
<td>0.845</td>
<td>yes</td>
<td>yes</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>701010055</td>
<td>2</td>
<td>28/28</td>
<td>2544</td>
<td>3</td>
<td>0.12%</td>
<td>No</td>
<td>15 (10, 21)</td>
<td>0.923</td>
<td>0.590</td>
<td>yes</td>
<td>yes</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7010100068</td>
<td>4</td>
<td>17/89</td>
<td>2595</td>
<td>115</td>
<td>4.43%</td>
<td>No</td>
<td>688 (583, 792)</td>
<td>42.499</td>
<td>0.000</td>
<td>no</td>
<td>no</td>
<td>multiple variant transmission</td>
<td>7</td>
<td>30/72</td>
</tr>
<tr>
<td>7000101006</td>
<td>2</td>
<td>40/40</td>
<td>2595</td>
<td>5</td>
<td>0.19%</td>
<td>Yes</td>
<td>13 (8, 18)</td>
<td>0.800</td>
<td>0.634</td>
<td>yes</td>
<td>yes</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Diversity and model analysis of full length env sequences from 28 acutely infected subjects.
Table 2. Cont.

<table>
<thead>
<tr>
<th>Subject</th>
<th>No. of transmitted/founder viruses</th>
<th>Poisson days since transmitted/founder</th>
<th>Maximum diversity %</th>
<th>APOBEC-mediated hypermutation</th>
<th>Maximum Hamming distance (90%)</th>
<th>Star phylogeny</th>
</tr>
</thead>
<tbody>
<tr>
<td>701010027 5</td>
<td>27/27 5</td>
<td>2.552 0.977 no yes</td>
<td>39/44 6</td>
<td>19/25 5</td>
<td>0.960 no yes</td>
<td>1.000 no yes</td>
</tr>
<tr>
<td>7010100108 5</td>
<td>35/35 4</td>
<td>0.3% 0.977 no yes</td>
<td>39/44 6</td>
<td>19/25 5</td>
<td>0.960 no yes</td>
<td>1.000 no yes</td>
</tr>
<tr>
<td>701010326 4</td>
<td>45/45 6</td>
<td>0.3% 0.977 no yes</td>
<td>39/44 6</td>
<td>19/25 5</td>
<td>0.960 no yes</td>
<td>1.000 no yes</td>
</tr>
<tr>
<td>701010328 5</td>
<td>38/38 5</td>
<td>0.3% 0.977 no yes</td>
<td>39/44 6</td>
<td>19/25 5</td>
<td>0.960 no yes</td>
<td>1.000 no yes</td>
</tr>
</tbody>
</table>

*HD, Hamming Distance - number of base positions at which two sequences differ.*

- **No. of transmitted/founder viruses**: Number of viruses transmitted/founder in each subject.
- **Poisson days since transmitted/founder**: Poisson distribution of days since transmission.
- **Maximum diversity %**: Maximum diversity percentage among transmitted/founder viruses.
- **APOBEC-mediated hypermutation**: APOBEC-mediated hypermutation percentage.
- **Maximum Hamming distance (90%)**: Maximum Hamming distance (90%) between sequences.
- **Star phylogeny**: Star phylogeny of transmitted/founder viruses.

---

**Explanation for deviation from model**

- **Goodness of fit**
  - **HD fit to poisson lambda**: Goodness of fit P value ($\chi^2$). A P value of 0.05 suggests that the observed distribution of mutations is inconsistent with a Poisson.
  - **MRCA**: Most recent common ancestor (95% CI).
  - **Lamda**: Mean of the best fitting Poisson found through maximum likelihood method.

---

**Table 2**

**Table 3**

**Table 4**

---

**Fig. 6**

**Fig. 7A**

**Fig. 7B**

**Fig. 7C**

---

**Plasmid modeling of the first 35 days of infection (eclipse phase to the end of Fiebig stage IV)** [10]. Power calculations further indicated that with a sample size of 80 sequences, there is a >95% probability of detecting minor sequences representing at least 4% of the population (see Methods). These findings suggest that more extensive sampling might result in the detection of an even greater number of transmitted/founder viruses in this individual. Subject 701010068 had the second highest env diversity (4.45%) among the study subjects (Table 2). He reported a single high risk exposure event involving unprotected receptive anal intercourse with two individuals, one HIV negative and the other HIV positive. He developed flu-like symptoms approximately two weeks later and was studied three weeks after that, again at Fiebig stage IV. Based on the earlier analysis of subject 04013171 (Fig. 6), we were concerned that viral recombination [39,40,41,42], which is sequence length and time (from infection) dependent [10,13,40], could confound the identification of discrete transmitted/founder virus lineages. This could be especially problematic in subjects infected with many different transmitted/founder viruses as opposed to two, since in the former case it is far more likely that doubly or multiply infected cells will spawn heterozygous virus progeny that lead in the next virus generation to recombinant viral genomes [39,40,41]. To test this hypothesis, we amplified and sequenced seventy-two 3’ half genome segments of plasma vRNA from subject 701010068 and then analyzed env gp41 (1035 bp), env gp160 (2630 bp), and 3’ half genome regions (4734 bp) separately. The gp41 sequences (Fig. 7A) revealed discrete low diversity lineages comprised of identical or nearly identical sequences. We interpreted 7 of these sequence clusters as likely to have arisen from distinct transmitted viruses and the remaining sequences to represent inter-lineage recombinants. Clusters of identical or nearly identical sequences were also evident in gp160 sequences (Fig. 7B), but with less clarity due to additional inter-lineage recombination events in the longer sequences. For example, sequences corresponding to lineage 4 in the gp41 sequences (depicted in light blue in Fig. 7A) were dispersed into five widely separated branches in the gp160 tree due entirely to recombination (Fig. 3B). Similarly, sequences comprising lineage 6 in the gp41 sequences (depicted in red in Fig. 7A) were dispersed into three widely separated branches in the gp160 tree, again due entirely to recombination (Fig. 7B). These findings were supported by Hudson-Kaplan analysis [38], which indicated a minimum of 27 recombination breakpoints among the gp160 env sequences. Interspersion of sequences was even more dramatic in the 3’ half genome tree (Fig. 7C). Remarkably, of the 72 3’ half genome sequences depicted in Fig. 7C, 63 (88%) represented overt recombinants between two or more transmitted/founder lineages demonstrable by visual inspection and by computer-assisted algorithms. Only two (dark blue) sequences labeled L7 and L9 at the very top of the tree (Fig. 7C), three (green) sequences labeled B1, L1 and P7 in the middle of the tree, and four (gray) sequences labeled B27, E1, A2 and J1 at the very bottom of the tree showed no evidence of recombination. These findings, together with corroborating data from env-only sequences [8,10,11], lead to the surprising conclusion that by the time of first antibody detection in acute HIV-1 infection (Fiebig stages III/IV), a majority of circulating viruses may be recombinants. This finding is testament to the large number of doubly (or multiply) infected cells in acute and early infection and further evidence of the rapidity with which virus diversifies [43,44], making clear that in order to identify non-recombinant transmitted/founder HIV-1 (or SIV) genomes [10,12,13], it is necessary to characterize viral sequences as close to the transmission event as possible.
Comparisons of Multiplicity of HIV-1 Infection in MSM versus Heterosexuals

Four studies, including the present one, have estimated the numbers of viruses responsible for transmission and productive HIV-1 infection after heterosexual or homosexual exposure using identical SGA-direct env amplicon sequencing methods [8,9,10]. One of these evaluated the frequency of multivariant transmission in a cohort of cohabitating HIV-1 discordant (antiretroviral naïve) heterosexual couples in Zambia and Rwanda followed prospectively for HIV-1 transmission [9]. Remarkably, only 2 of 20 [10%; 95% confidence interval (CI) 1–32%] of the epidemiologically linked infections resulted from multivariant transmission, a finding attributed to the chronicity of infection in the virus positive partner, lower prevalence of comorbid conditions such as untreated tuberculosis or sexually transmitted infections, and the heterosexual route of transmission. Since the frequency of multivariant transmission in HSX in that study was substantially less than what we observed for MSM [2 of 20 HSX (10%, CI 1–32%) versus 10 of 28 MSM (36%, CI 19–56%); Fisher’s exact p = 0.042, odds ratio 4.85, 95% CI 1.1–inf], we performed a combined analysis of data from all four studies, which included 225 patients infected by HIV-1 subtypes A, B or C (Table 3). Again we found that the proportion of MSM subjects infected by more than one virus was substantially higher than for HSX [19 of 50 (38%) versus 34 of 175 (19%); Fisher’s exact p = 0.008, odds ratio 2.5, 95% CI 1.2–5.3]. The MSM subjects were all infected with HIV-1 subtype B; a comparison to only the subset of HSX infections that were subtype B was still significant (Fisher’s exact p = 0.01, odds ratio 2.9, 95% CI 1.2–7.1). The frequency of multiple infections in HSX was not statistically different among subtypes A, B and C nor was it different between males and females. Differences in the frequency of multivariant HIV-1 transmission in MSM versus HSX could not be accounted for by the numbers of sequences analyzed per subject nor by the clinical stage of subjects at the time of study. In the study by Haaland [9], the median number of sequences analyzed was 40, in Keele [10] it was 25, and in Abraham [8] it was 22. In the present study, the median number of sequences that we determined as part of our initial survey was 33 (Table 2). We used this lower number of sequences for statistical comparisons of single and multivariant transmissions in MSM versus HSX subjects in order to allow for comparability among the four studies. When, in this initial

Figure 2. NJ trees and Highlighter plots of HIV-1 env diversity. Full-length gp160 env sequences from four subjects are depicted by NJ tree phylogenies and by Highlighter, a sequence visualization tool that allows tracing of common ancestry between sequences based on individual nucleotide polymorphisms [10]. Sequences from subject 04013440 (A) showed productive infection by a single virus, from subject 04013211 (B) infection by two closely related viruses, from subject 04013383 (C) infection by two distantly related viruses, and from subject 04013448 (D) infection by four viruses with inter-lineage recombinants denoted by orange symbols in the NJ tree. The horizontal scale bar represents genetic distance. doi:10.1371/journal.ppat.1000890.g002
sequence set, we identified samples containing more than one transmitted/founder virus lineage, we went on to obtain additional sequences (as many as 89) in order to estimate more precisely the numbers of transmitted/founder viruses (Table 2). Increasing the numbers of sequences analyzed allowed for greater accuracy and precision in estimating the numbers of viruses transmitted in those individuals with many transmitted viruses (e.g., subjects 04013448, 04013171 and 701010068 in Figs. 2D, 6 and 7), but it did not affect the discrimination between those subjects infected by one virus versus those infected by more than one virus. Finally, we found no significant correlation between the clinical stage of subjects at the time of plasma sampling and the numbers of transmitted/founder viruses identified in those samples: among the four studies, there was a total of 95 antibody negative subjects (Fiebig stages I–II) and 130 antibody positive subjects (Fiebig stages III–VI). Twenty subjects (21%) in the former group and 33 subjects (25%) in the latter group had evidence of productive infection by more than one virus, which was not significantly different (odds ratio 1.27; 95% CI 0.65–2.54; Fisher’s exact p = 0.53).

Discussion

Previous studies used less precise methods for estimating multiplicity of HIV-1 infection in HSX and MSM subjects and reported widely varying results with a trend for higher multiplicities in MSM [15,16,17,18,19,20,21,22,24]. We report here new SGA-based determinations that show significant differences in the multiplicity of virus infection between the two risk groups: MSM were twice as likely as HSX subjects to become infected by more than one virus, with some MSM acquiring as many as 7 to 10 or more viruses. These findings are consistent with the higher epidemiological risk of HIV-1 acquisition in MSM compared with HSX and may be explained in part by the anatomical and immunohistological differences between the male and female genitourinary tracts and the lower intestine.

A limitation of the current study is that it represents a retrospective comparison of multivariant HIV-1 transmission among patient cohorts having different enrollment criteria and different behavioral risk assessments. It must be noted, however, that all study subjects from all cohorts were queried extensively with regard to potential HIV-1 infection risk behaviors. This included acutely infected subjects identified by cross-sectional screening methods [45], subjects enrolled prospectively into HIV-1 discordant couple [9] or Acute Infection Early Disease Research Program cohorts [46], and source plasma donors who became HIV-1 infected during a period of serial plasma collections [10]. The latter subjects, whom we studied anonymously, underwent exhaustive pre-enrollment interrogation for HIV and IDU risk behaviors according to a standardized FDA-approved protocol (http://www.fda.gov/BiologicsBloodVaccines/GuidanceComplianceRegulatoryInformation/Guidances/Blood/ucm073445.htm) that included a written questionnaire and interview inquiring about MSM and IDU activities, sex-for-money, sex with a partner who had sex-for-money, or sex with an individual known to be HIV positive. Source plasma donors also underwent serial laboratory testing for surrogate laboratory markers that could indicate injection drug use (e.g., liver transaminase elevations and hepatitis B or C nucleic acids or antibodies), and these markers were negative among qualified donors. Nonetheless, self-reporting of risk behaviors among paid plasma donors is imperfect [47], and it is possible that some subjects whom we categorized as belonging to the HSX risk group actually had additional risks such as IDU or MSM. However, even if this were the case, it would likely bias the findings in the HSX group toward a greater (not lesser) number of transmitted viruses [48]. Not surprisingly, when we excluded all source plasma donor subjects from our comparative analysis of multivariant HIV-1 transmission, the difference between MSM and HSX groups was still statistically significant [19 of 50 (38%) versus 25 of 119 (21%), respectively; Fisher’s exact p = 0.03, odds ratio 2.3, 95% CI 1.04–5.02]. We conclude that within the limitations of self-reporting and surrogate marker testing, study subjects in the cohorts we examined were correctly assigned to HSX and MSM risk groups, differences in multiplicity of virus transmission between the two groups were significant, and overall study findings were unlikely to have been confounded by injection drug use. Future studies can
HIV-1 in Men Who Have Sex with Men

A

5' half

B

3' half

Base number

0 1000 2000 3000 4000

0.0002
1 nt

0.0002
1 nt

IT

A

C

G

Gaps

Apobec

Base number

0 1000 2000 3000 4000
benefit from a prospective trial design and a common behavioral and medical questionnaire [49,50].

It is noteworthy that while multivariant HIV-1 transmission was twice as common in MSM than in HSX, still more than half of MSM subjects showed evidence of productive infection by just one virus. Moreover, the adjusted median (calculated from subjects with multivariant transmissions only) was 3 in MSM compared with 2 in HSX (Table 3). Even in the Fiebig II subject AD17, where we analyzed a total of 239 sequences (giving us a 95% probability of detecting a second transmitted/founder virus lineage at 1.25% prevalence), all of the sequences coalesced phylogenetically to a single virus, thus providing no evidence for transmission of more than one virus. Elsewhere, we have used 454 deep sequencing to analyze tens of thousands of sequences from three additional Fiebig stage II MSM subjects in whom SGA-direct sequencing suggested transmission and productive clinical infection by a single virus (Will Fischer, B.F.K., G.M.S. B.T.K., unpublished). Even with this greatly enhanced sensitivity of detection of minor sequences, we found no evidence of transmission by more than one virus in these subjects. Considered together, the findings of the present study, previously published studies [8,9,10], and work in progress (Will Fischer, B.F.K., G.M.S., B.T.K), all suggest that a substantial proportion of HSX and MSM patients acquire HIV-1 infection as a consequence of transmission and productive infection by literally one virion or one infected cell. The implication of this finding is that in order for a vaccine, microbicide or other prevention modality to be protective in this fraction of individuals, it need only prevent infection by a single virus or infected cell. Conversely, there is another subset of HSX and MSM subjects in whom the multiplicity of infection is higher. Since the proportion of such multiply infected individuals is far higher than would be expected from a Poisson distribution of independent, low frequency events (see Abrahams [8] for discussion), we suspect that biological events underlying virus...
transmission in these subjects compared with those infected by a single virus are different and that challenges faced by vaccines and microbicides in the higher multiplicity infection group may be higher.

Another interesting observation from the present study relates to viral recombination. Although recombination was not a primary study objective, the identification of two or more transmitted/founder genomes in acutely infected subjects gave us a unique opportunity to examine the dynamics and extent of recombination in primary HIV-1 infection. Five features of our study distinguish it from previous reports of HIV-1 recombination [39,40,41,42].

First, we studied subjects at very early clinical stages following virus transmission (Fiebig stages II–V). Second, we used SGA-direct amplicon sequencing, which provides for a proportional representation of virus present in the plasma, including those that are recombinant [11]. Third, SGA eliminates in vitro recombination artifacts resulting from Taq polymerase-mediated template switching [11]. Fourth, SGA allowed us to identify the exact nucleotide sequences of full-length transmitted/founder virus env genes unambiguously and to distinguish these viruses and their progeny from viruses that contained even short regions of recombinant sequence. Fifth, SGA-direct sequencing of 3’ half genomes allowed us to examine recombination across the boundaries of nef-ltr. Figs 2D, 6 and 7 illustrate examples of recombination and Table 2 summarizes the findings in all multiply infected subjects. Seven of 9 subjects had evidence of recombination within gp160 env (one subject, AD77, could not be analyzed because of excessive virus diversity at a late Fiebig stage). The proportion of recombinants ranged from 0 of 30 sequences in subject 04013211 to 30 of 72 sequences (42%) in subject 701010068. In the latter subject, we amplified a longer fragment of the viral genome so as to include the 3’ half; this allowed us to compare recombination frequencies within gp41 (only), gp160 (only) or the full-length 3’ half genome. The proportion of recombinants in these three regions was 13/72 (18%), 30/72 (42%) and 63/72 (88%), respectively. Recombination breakpoints were more common in sequences flanking gp160 env than within env (Fig. 7C), a finding similar to that reported by Simon-Loriere and colleagues for HIV-1 inter-subtype recombination [42]. In subject 701010068, where 88% of sequences corresponding to only half the viral genome were recombinant, it is likely that nearly all of the full-genome sequences at this time.

Figure 6. NJ tree and Highlighter plot of HIV-1 env diversity in subject 04013171. Sequences emanating from ten transmitted/founder viruses are color-coded and identified as variants 1–10. Inter-lineage recombinants are depicted in orange. The horizontal scale bar represents genetic distance.

doi:10.1371/journal.ppat.1000890.g006
Four other transmitted/founder virus molecular clones have been described ([12] and J.S.G. and G.M.S., unpublished), and all of these correspond to HIV-1 subtype C viruses resulting from heterosexual transmissions. With the addition of the pAD17.1 clone, we now have molecular proviral clones representing male-to-male rectal transmission (pAD17.1), male-to-female vaginal transmission (pZM246F-10; pZM247Fv1; pZM247Fv2), and female-to-male penile transmission (pZM249M-1). All of these viruses are R5 tropic, replicate efficiently in activated human CD4+ T cells, but fail to replicate efficiently in monocyte-derived macrophages. Such molecular clones of transmitted/founder viruses should represent a rich resource for studying the biology of HIV-1 transmission and its prevention.

In summary, the findings presented here provide for the first time a comparative, quantitative view of the HIV-1 transmission event in two patient risk groups that dominate the HIV-1 pandemic. In doing so, they highlight both challenges and opportunities confronting candidate vaccines, microbicides, and other prevention modalities. Elucidation of the biological basis of single versus multivariant transmission in MSM and HSX could help advance prevention strategies [53,56,57], with quantitative analyses of transmitted/founder viruses representing a potentially valuable new endpoint in vaccine and microbicide trial design and assessment [5,6,49,50].

### Materials and Methods

#### Study Subjects

This study was conducted according to the principles expressed in the Declaration of Helsinki. It was approved by the Institutional Review Boards of the University of Alabama at Birmingham, Rockefeller University, Duke University, and the University of North Carolina. All patients provided written informed consent for the collection of samples and subsequent analysis. Blood samples were obtained from 28 subjects with acute HIV-1 infection and from chronically infected sexual partners of two of them. Blood specimens were generally collected in acid citrate dextrose and
plasma separated and stored at −70°C. PBMCs were stored in vapor phase liquid nitrogen.

**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 Or, and Genetic Systems HIV-1 Western Blot Kit. Based on these test results, subjects were staged according to the Fiebig classification system for acute and early HIV-1 infection [33].

**Viral RNA Extraction and cDNA Synthesis**

For each sample, approximately 20,000 viral RNA copies were extracted using the Qiagen BioRobot EZ1 Workstation with EZ1 Virus Mini Kit v2.0 (Qiagen, Valencia, CA). RNA was eluted in 60 μl of elution buffer and subjected to first strand cDNA synthesis immediately by using the SuperScript III Reverse Transcriptase according to manufacturer’s instructions (Invitrogen Life Technologies). Each first strand synthesis reaction included ~10,000 or fewer vRNA molecules, 1X reverse transcription buffer, 0.5 mM of each dNTP, 5 mM DTT, 2 units/μl of RNaseOUT, 10 units/μl of SuperScript III reverse transcriptase and 0.25 μM of antisense primer. The cDNA syntheses were performed using antisense primers located at different genomic regions. The primers for synthesizing the cDNA of env, 5’ half genome (U5, gag and pol) and 3’ half genome (vif, vpr, tat, rev, spv, env, nef, U3 and R) were env3out 5’T-TTGCATCTTGGTATCGTCATG-3’, 1.int.R1 5’-CTTGCCACACATCTCAACCTGCCC-3’, 2nd round sense primer 2.U5.F2 5’-TTTGCTACTTGTGATTGCTCCATG-3’, and 2nd round antisense primer 2.int.R2 5’-CAATCATTACGCCGATCGTTTTCACATA-3’, (3’) 3’ half genome: 1st round sense primer 1.F1 5’-ACAGGAGTA-CAATGGCAGTTATT-3’, 1’ round antisense primer 1.R3.B3R 5’-ACTCTGTGAACTCATGCCAAGGTATTGTTCAG-3’, 2nd round sense primer 2.int.F2 5’-TGGAAAGTGAAGGGGCTAGTGGTACTAC-3’, and 2nd round antisense primer 2.R3.B6R 5’-TGAAATCCAGGGAACATTTATAAGGCC-3’. PCR parameters were as follows: 94°C for 2 min, followed by 35 cycles of 94°C for 15 s, 58°C for 30 s, and 68°C for 4 min (env) or 5 min (5’ or 3’ half genomes), followed by a final extension of 68°C for 10 min. The product of the first-round PCR was subsequently used as a template in the second-round PCR under same conditions but with a total of 45 cycles. The amplicons were analyzed on a 1% agarose gel and purified using the QIAquick Gel Extraction Kit (Qiagen, Valencia, CA). Amplified products were sequenced directly using BigDye Terminator chemistry and protocols recommended by the manufacturer (Applied Biosystems, Foster City, CA). Sequence 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 4.8 (Gene Codes; Ann Arbor, MI). All chromatograms were inspected for sites of mixed bases (double peaks), which would be evidence of priming from more than one template or the introduction of PCR error in early cycles. Any sequence with evidence of double peaks was excluded from further analysis.

**Proviral DNA Extraction**

Blood was collected from subject AD17 14–17 days following exposure to HIV-1 at Fiebig stage II. Genomic DNA was extracted from 1.3 million PBMCs using Qiagen Tissue DNA Extraction kit according to manufacturer’s instructions.

**Single Genome Amplification**

cDNA or genomic DNA was serially diluted and distributed in replicates of 8 PCR reactions in MicroAmp 96-well plates (Applied Biosystems, Foster City, CA) so as to identify a dilution where PCR positive wells constituted less than 30% of total number of the reactions. At this dilution, most wells contain amplicons derived from a single cDNA molecule. Additional PCR amplifications were performed using this dilution in 96-well reaction plates. PCR amplification was carried out in presence of 1X High Fidelity Platinum Taq PCR buffer, 2 mM MgSO4, 0.2 mM each deoxynucleoside triphosphate, 0.2 μM each primer, and 0.025 units/μl of Platinum Taq High Fidelity polymerase in a 20-ul reaction (Invitrogen, Carlsbad, CA). The nested primers for generating different genomic segments included: (1) full length env: 1st round sense primer env3out 5’-TAGGCCCTGGAG-CATCCAGGAAG-3’, 1st round antisense primer env3out 5’T-TTGCATCTTGGTATCGTCATG-3’, 2nd round sense primer env3in 5’-TTAGCCATCTCCTATGGCAGGAAGA-3’, and 2nd round antisense primer env3in 5’-GTCTCGAGATACTGTGCCCACCC-3’, (2) 5’ half genome: 1st round sense primer 1.U5.F1 5’-CCTTGAGTGCTTCAAGTAG-TGTGGCCCGCCGTCTGT-3’, 1st round antisense primer 1.int.R1 5’-CTTGCCACACATCTCAACCTGCCC-3’, 2nd round sense primer 2.U5.F2 5’-TTAGCTGTGGCCGGCGTCTGT-TGTGGTACTC-3’, and 2nd round antisense primer 2.int.R2 5’-CAATCATTACGCCGATCGTTTTCACATA-3’; (3) 3’ half genome: 1st round sense primer 1.F1 5’-ACAGGAGTA-CAATGGCAGTTATT-3’, 1’ round antisense primer 1.R3.B3R 5’-ACTCTGTGAACTCATGCCAAGGTATTGTTCAG-3’, 2nd round sense primer 2.int.F2 5’-TGGAAAGTGAAGGGGCTAGTGGTACTAC-3’, and 2nd round antisense primer 2.R3.B6R 5’-TGAAATCCAGGGAACATTTATAAGGCC-3’. PCR parameters were as follows: 94°C for 2 min, followed by 35 cycles of 94°C for 15 s, 58°C for 30 s, and 68°C for 4 min (env) or 5 min (5’ or 3’ half genomes), followed by a final extension of 68°C for 10 min. The product of the first-round PCR was subsequently used as a template in the second-round PCR under same conditions but with a total of 45 cycles. The amplicons were analyzed on a 1% agarose gel and purified using the QIAquick Gel Extraction Kit (Qiagen, Valencia, CA). Amplified products were sequenced directly using BigDye Terminator chemistry and protocols recommended by the manufacturer (Applied Biosystems, Foster City, CA). Sequence 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 4.8 (Gene Codes; Ann Arbor, MI). All chromatograms were inspected for sites of mixed bases (double peaks), which would be evidence of priming from more than one template or the introduction of PCR error in early cycles. Any sequence with evidence of double peaks was excluded from further analysis.

**Sequence Alignments**

All the sequence alignments were initially made with ClustalW and then hand-checked using MacClade 4.0 to improve the alignments according to the codon translation. Consensus sequences were generated for each individual. The full sequence alignment is available as a supplemental data file (www.hiv.lanl.gov/content/sequence/HIV/User_ALIGNMENTS/). Each consensus sequence is deposited in GenBank (accession numbers: GU1930247–GU1931770).

**Sequence Diversity Analysis**

Complete env sequences (n = 1307) were derived from 30 individuals, and 5’ (U5, gag and pol) and 3’ (vif, vpr, tat, rev, env, nef, U3, and R) half genome sequences (n = 185) were derived from PBMC and plasma at two different time points from subject AD17. We analyzed sequences for maximum sequence diversity and then visually inspected each set of sequences using neighbor-joining (NJ) and Highlighter tools (www.hiv.lanl.gov). Phylogenetic trees were generated by the neighbor-joining method with ClustalW or PAUP.

**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.
HIV-1 in Men Who Have Sex with Men

Enrichment for mutations with APOBEC3G/F signatures was assessed using Hypermut 2.0 (www.hiv.lanl.gov). Sequences that yielded a p-value of 0.05 or lower were considered significantly hypermutated and excluded from subsequent analyses.

Proviral DNA Cloning
To obtain an infectious molecular clone of the transmitted/founder virus of subject AD17, we amplified overlapping 5’ and 3’ half genomes from proviral DNA of earliest sample (day 14; 6/11/99) by single round PCR using Phusion Hot Start High-Fidelity DNA polymerase (Biola). Both fragments contained a complete LTR element and an overlap of 170 base pairs encompassing a unique SalI restriction site. For the 5’ half genome, U3-R-U5, gag, pol, vif, vpr and tat1 was amplified. For the 3’ half genome, tat1, rev1, vpu, env, nef, tat2, rev2 and U3-R-U5 was amplified. The primers were designed to complement the confirmed transmitted/founder sequence as determined by SGA-direct amplicon sequencing. The recognition sequences of MluI and SalI restriction enzymes were appended to the 5’ ends of the sense and antisense primers, respectively. Single round bulk PCR amplifications were carried out in the presence of 1X Phusion Hot Start HiFi buffer, 0.2 mM of each deoxynucleoside triphosphate, 0.5 μM of each primer, 3% final concentration of DMSO, and 0.02 units/μl of Phusion Hot Start High Fidelity polymerase in 50 μl reactions. The PCR product of each half genome was subjected to MluI and SalI digestion and gel purification and then independently cloned into the MluI-SalI site of TOPO XL vector (Invitrogen). The ligation mixture was transformed into XLI2 Blue MRF competent cells and plated onto LB agar plates supplemented with 50 μg/ml of kanamycin and grown overnight at 30°C. Single colonies were selected and grown overnight in LB medium with the same concentration of kanamycin at 30°C with constant shaking. Plasmid DNA was isolated and sequenced to confirm the identity of transmitted/founder sequences. The 5’ genome half was excised and cloned into 3’ TOPO XL vector by utilizing the MluI and SalI restriction sites thereby generating the full length clone of the transmitted/founder provirus.

Phenotypic Analyses
Replication competency of the full length molecular proviral clone AD17.1 was assessed using 293T cells, JC53BL-13 cells (NIH AIDS Research and Reference Reagent Program catalogue #8129, TZM-bl), activated primary human CD4+ lymphocytes, and monocyte-derived macrophages. Infectious virus stock generation, Env-pseudotyped virus stocks, titrations, cell infections and virus neutralization assays were performed according to methods previously described [12]. Virus controls (replication competent or Env-pseudotyped) included the HIV-1 macrophage-tropic strains YU2 and BaL, the non-macrophage tropic T-cell line-adapted strain NL4.3, the dual R5/X4 tropic strain WEAU1.60, and the xenotropic MuLV env [12]. The coreceptor inhibitors TAK779 and AMD3100 were obtained from the NIH AIDS Research and Reference Reagent Program (4983 and 8128). R5 and X4 tropism was assessed in both JC53BL-13 cells and in GHOST[3] cells that stably express CD4 along with CCR3 or CXCR4 or both or neither coreceptor.

Recombination Analyses
Recombination was evaluated using GARD [38] and Recco [37] and by visual inspection of Highlighter plots. The minimum number of recombination events required to explain sequence datasets was estimated using the four-gamete method of Hudson and Kaplan [38] as implemented in DNASP v5.00.07 [59]. Recombinant sequences reported in Table 2 were identified by Highlighter analysis and confirmed by Hudson-Kaplan, GARD and Recco analyses.

Mathematical Model
The model employed in the present study has been described [10,12,34] as have measured parameters of early virus expansion [52,53,54]. Under this model, with no selection pressure and fast expansion, one can expect small samples from homogeneous virus populations to have evolved from a founder strain in a star-like phylogeny with all sequences coalescing at the founder [10,13]. Occasional deviations from a star phylogeny are, however, expected. The sampling of 10 sequences, for example, from a later generation of an exponentially growing population with six-fold growth per generation (R0 = 6) has about 3% chance of including at least one pair sharing the first four generations, a 19% chance of including sequences that share three, and a 73% chance of sharing two. Using a point mutation rate of about 1 per 5 generations for the full-length 9 kb HIV-1 genome [10,34], there is about 75% chance of finding among ten sequences two that share one mutation, about 20% chance of finding two sequences that share a pair of mutations, and <2% chance of sharing more than that. These probabilities are slightly enhanced by early stochastic events that can lead to the virus producing less than six descendants in some generations but are diminished by the chances that mutations cause a fitness disadvantage that results in early purifying selection, as previously observed [10,60]. We found examples of such early stochastic mutations leading to deviations from star phylogeny in several subjects (Table 2).

Statistical Analyses and Power Calculations
We calculated the statistical significance of differences in rates of single versus multivariant HIV-1 transmission using Fisher’s exact test. Differences were considered statistically significant at a value of p≤0.05. To estimate the likelihood of missing infrequently represented transmitted variants, we described previously a power study that estimated the probability of sampling low frequency plasma viral sequences [10]. In a sample set of at least n = 20, there is a 95% probability that a given missed variant comprises less than 14% of the virus population. For a sample set of 30, there is a 95% probability not to miss a variant that comprises at least 10% of the total viral population. And for a sample set of 80, there is a 95% probability not to miss a variant that comprises at least 4% of the total viral population.

We also considered the possibility that the number of transmitted/founder viruses detected could be influenced by the clinical stage (Fiebig stage) of the subjects at the time of virus sampling, because differences in virus replication rates could lead to increasing differences in virus frequencies with time. If this were the case and some viruses were outcompeted, the prediction would be that at later Fiebig stages the numbers of transmitted/founder virus lineages would be less than at earlier Fiebig stages. Our model (which is based on previously estimated parameters of an HIV-1 generation time of 2 days, a reproductive ratio [R0] of 6, and a reverse transcriptase error rate error rate of 2.16×10⁻³ and assumes that the initial virus replicates exponentially infecting R0 new cells at each generation and diversifies under a model of evolution that assumes no selection) predicts that descendants of a transmitted virus at 45% replicative disadvantage compared to another transmitted virus, still have more than a 5% chance of occurring in a sample size of 20, ten generations (~20 days) later. In humans, the eclipse period, defined as the time between HIV-1 transmission and first detection of virus in the plasma, has been estimated to be approximately 10–14 days, and the eclipse period plus Fiebig stages I and II, approximately 22–26 days [10,33]. In the present
study, the numbers of subjects in Fiebig stages I/II, III, IV and V were 14, 2, 6 and 6, respectively, and this relative distribution was similar in the three other studies included in the combined analysis \([9,10]\). As described in the main text, there was no significant correlation between clinical stage and multiplicity of infection (Fisher’s exact \(p = 0.53\)).

Acknowledgments

We thank the study participants; clinical staff at UNC, UAB, Duke and Rockefeller Universities; core research facilities of the Centers for HIV/

References


AIDS Research at Duke, UNC and UAB; and Jamie White for manuscript preparation.

Author Contributions

Conceived and designed the experiments: HL, KJB, BTK, JJE, MSC, CBH, GMS. Performed the experiments: HL, KJB, SW, JMD, VC, CS, FSG, MGS. Analyzed the data: HL, KJB, GMS, JES, PH, IE, G, ASB, FSG, BHH, GMS. Wrote the paper: HL, KJB, BHH.