

# HIV-1 Genetic Diversity Among Incident Infections in Mbeya, Tanzania

Erik Billings,<sup>1</sup> Eric Sanders-Buell,<sup>1</sup> Meera Bose,<sup>1</sup> Gustavo H. Kijak,<sup>1</sup> Andrea Bradfield,<sup>1</sup> Jacqueline Crossler,<sup>1</sup> Miguel A. Arroyo,<sup>2,\*</sup> Leonard Maboko,<sup>3</sup> Oliver Hoffmann,<sup>1,3</sup> Steffen Geis,<sup>3,4,†</sup> Deborah L. Bix,<sup>2,‡</sup> Jerome H. Kim,<sup>5</sup> Nelson L. Michael,<sup>2</sup> Merlin L. Robb,<sup>1</sup> Michael Hoelscher,<sup>4,6</sup> and Sodsai Tovanabutra<sup>1</sup>

## Abstract

In preparation for vaccine trials, HIV-1 genetic diversity was surveyed between 2002 and 2006 through the Cohort Development study in the form of a retrospective and prospective observational study in and around the town of Mbeya in Tanzania's Southwest Highlands. This study describes the molecular epidemiology of HIV-1 strains obtained from 97 out of 106 incident HIV-1 infections identified in three subpopulations of participants (one rural, two urban) from the Mbeya area. Near full-genome or half-genome sequencing showed a subtype distribution of 40% C, 17% A1, 1% D, and 42% inter-subtype recombinants. Compared to viral subtyping results previously obtained from the retrospective phase of this study, the overall proportion of incident viral strains did not change greatly during the study course, suggesting maturity of the epidemic. A comparison to a current Phase I-II vaccine being tested in Africa shows ~17% amino acid sequence difference between the gp120 of the vaccine and subtype C incident strains. Phylogenetic and recombinant breakpoint analysis of the incident strains revealed the emergence of CRF41\_CD and many unique recombinants, as well as the presence of six local transmission networks most of which were confined to the rural subpopulation. In the context of vaccine cohort selection, these results suggest distinct infection transmission dynamics within these three geographically close subpopulations. The diversity and genetic sequences of the HIV-1 strains obtained during this study will greatly contribute to the planning, immunogen selection, and analysis of vaccine-induced immune responses observed during HIV-1 vaccine trials in Tanzania and neighboring countries.

**Keywords:** epidemiology, vaccines, HIV transmission, phylogenetics, HIV genetics

## Introduction

**I**N 2014, AN ESTIMATED 1.5 MILLION PEOPLE (5.3% of the 15–49-year-old age group) were living with HIV-1 in Tanzania.<sup>1</sup> In addition to the costs associated with treatment and care, the loss of a significant number of adults during

their most productive years (est. 46,000 deaths among the 15–49-year-old age group in 2014), presents a heavy burden on the people of Tanzania. The national prevalence rates from recent years (7% in 2004, 5.7% in 2008, and 5.1% in 2012) suggest that the HIV-1 epidemic in Tanzania may be stabilizing likely as a result of antiretroviral treatment, prevention,

<sup>1</sup>United States Military HIV Research Program/Henry M. Jackson Foundation, Silver Spring, Maryland.

<sup>2</sup>United States Military HIV Research Program/Walter Reed Army Institute of Research, Silver Spring, Maryland.

<sup>3</sup>NIMR-Mbeya Medical Research Center, Mbeya, Tanzania.

<sup>4</sup>Division of Infectious Diseases and Tropical Medicine, Medical Centre of the University of Munich (LMU), Munich, Germany.

<sup>5</sup>International Vaccine Institute, Seoul, Republic of Korea.

<sup>6</sup>German Centre for Infection Research (DZIF), Munich partner site, Munich, Germany.

\*Current affiliation: Department of Pathology and Area Laboratory Services, Dwight D. Eisenhower Army Medical Center, Fort Gordon, Georgia.

†Current affiliation: London School of Hygiene and Tropical Medicine, London, United Kingdom and Karonga Prevention Study, Chilumba, Malawi.

‡Current affiliation: Office of the Global AIDS Coordinator, United States Department of State, Washington, District of Columbia.

and public awareness programs that have been implemented in various parts of the country.<sup>2-4</sup> However, even at those rates, the high disease burden provides incentive for Tanzania to participate in HIV-1 vaccine trials.

The Cohort Development (CODE) study was conducted in preparation for anticipated vaccine trials in an effort to identify and characterize a regional cohort that had both the need and support for an HIV-1 vaccine trial within a community cohort. The CODE study was conceived and executed in collaboration between the Mbeya Medical Research Programme (MMRP), U.S. Military HIV Research Program (MHRP), Walter Reed Army Institute of Research (WRAIR), and the University of Munich (LMU).<sup>5</sup> CODE was conducted in Mbeya Town and the nearby Itende Village, which are in the Southwest Highlands region of Tanzania.

During recruitment for the CODE study (September 2002–April 2003), the HIV-1 prevalence among the participants was 16.6%.<sup>5</sup> More recent surveys<sup>4</sup> have reported that HIV-1 prevalence in the Mbeya district dropped to 9% in 2012; however, the three neighboring Southern/Southwestern Highlands districts of Njombe (14.8%), Mbeya (9%), and Iringa (9.1%) have had the highest prevalence of HIV-1 in Tanzania for several years.<sup>2-4</sup> The HIV-1 epidemic in Tanzania and East Africa in general is composed of multiple group-M subtypes (mostly A, C, and D) and recombinants between them. Furthermore, the HIV-1 strains that predominate (pure subtype or recombinant) are known to vary from region to region.

In northern Tanzania, a 2013 study of hospitality facility workers and general population participants (samples collected between 2008 and 2010)<sup>6</sup> found subtype distributions of 39% A1, 33% C, 18% D, and 11% recombinant forms. This is in contrast to a study<sup>7</sup> conducted in southern Tanzania, which found subtype distributions of ~8% A1, 41% C, 4% D, and 47% recombinants among high- and low-risk participants (samples collected in 2000). Similarly, other studies have reported varying subtype and recombinant distributions depending on the study location.<sup>8-10</sup> Such variation makes a preliminary survey of potential vaccine cohorts necessary before immunogen selection for vaccine trials and the subsequent analysis of their efficacy.

Recent HIV-1 vaccine development in Tanzania has progressed to the Phase I and II stages and has demonstrated both safety and production of anti-HIV humoral and cellular immune responses using a DNA prime and MVA boost strategy<sup>11,12</sup> that expresses various HIV-1 proteins from main circulating subtypes and circulating recombinant form (CRF). Currently, the HVTN 100 Phase I-II study is assessing the safety and immunogenicity of a subtype C-oriented vaccine within an African cohort. The HVTN 100 vaccine contains a canarypox (ALVAC/vCP2438) vectored *gp120* as part of the prime and protein boost of two recombinant *gp120* proteins, all of which are subtype C. In the event that larger Phase III efficacy trials are conducted, a detailed molecular analysis of the circulating strains is crucial for interpreting vaccine-induced immune responses. Cohort characteristics such as viral subtype distribution and the presence of local transmission networks that may facilitate the spread of closely related strains are important attributes that are taken into consideration during trial planning and analysis.

In this study, we present the molecular epidemiology of incident HIV-1 strains, and the establishment of the circu-

lating recombinant form, CRF41\_CD, among incident HIV-1 infections observed between 2002 and 2006 within a population of initially seronegative participants in the prospective CODE study conducted in Mbeya Town, Tanzania.

## Materials and Methods

### Study participants

In 2002–2003, the CODE study enrolled 3,096 consented participants from Mbeya Town (urban) and neighboring Itende Village (rural) into three subpopulations referred to as Itende, Ghana, and Advert. Door-to-door enrollment garnered 994 participants from Itende and 1,060 participants from the Ghana ward of Mbeya Town. A second enrollment strategy used public advertisement across all wards of Mbeya Town to attain 1,042 participants. Following enrollment and counseling, initial screening of the 3,096 participants revealed 514 prevalent HIV-1 infections.<sup>5</sup> The remaining 2,582 participants were followed for 3.5 years at 6-month intervals to observe new HIV-1 incident infections. Of the 2,582 participants that were initially HIV-1 negative, 106 seroconverted during the course of this study. Plasma, mostly from the first seropositive time point of incident infection samples, was subjected to reverse transcription-polymerase chain reaction (RT-PCR) and sequencing. The sequence names include a visit number (V#), which denotes the clinical visit during which the plasma sample and subsequent HIV-1 sequence were obtained. Average ages described in the text include a standard error of the mean (*SEM*).

### Laboratory procedures

RNA extracted from plasma using the QIAamp Viral RNA Mini Kit was the template for RT-PCR and sequencing. Complementary DNA was synthesized as the complete genome or as two half genomes overlapping by 1.5 kb, using ThermoScript™ RT (Invitrogen Corp., Carlsbad, CA) as instructed by the manufacturer. Primer design, template dilutions, and PCR steps were performed as previously described.<sup>13</sup> PCR products were then purified and sequenced using an ABI 3100 capillary sequencer. DNA sequences were assembled using Sequencher™, version 4.7. The resulting sequences covered HXB2 position ~796–9,496 for the near full-genome sequences and HXB2 position 4,589 or 4,950 to 9,496 for the half-genome sequences.

### Phylogenetic analysis

A multiple alignment of HIV-1 reference and incident infection strains was generated using HIVAlign<sup>14</sup> and refined with MEGA version 5.<sup>15</sup> Neighbor-joining trees were constructed and bootstrap values were calculated with DIVEIN<sup>16</sup> using the estimated GTR+I+G model. Informative site analysis, visual inspection, bootstrap analysis, the jumping profile Hidden Markov Model analytical tool,<sup>17</sup> and bootscanning (300 nt window) were performed to precisely map breakpoints within the final genome structures of inter-subtype recombinants.<sup>18-20</sup> Breakpoint analyses that resulted in discordant subtype assignments were resolved with HIV BLAST ([hiv.lanl.gov](http://hiv.lanl.gov)) analysis of the subgenomic segments. Genomic pairwise distances were calculated with MEGA version 5<sup>15</sup> using the K2P model, 0.5 gamma, pairwise deletion, and bootstrapped for 100 replications. Percent differences

between protein sequences were calculated with MEGA version 5 using the p-distance model and pairwise deletion. The median percent differences listed in the text include the interquartile range (*IQR*).

## Results

### Study participants

The initial findings of the CODE study describing epidemiological data on prevalent HIV-1 infections, coinfections (STDs), sociodemographics, and sampling methods have been previously reported.<sup>5,21</sup> The sequenced incident infections were observed in 22 men and 35 women from the Advert and Ghana groups, and 16 men and 24 women from the Itende group. Within the incident infections, the average age of the infected female participants was Itende: 27.4 years (*SEM* 1.5), Advert: 22.8 years (*SEM* 1.1), and Ghana: 21.7 years (*SEM* 0.9). Among the infected male participants, the average age was Itende: 27.1 years (*SEM* 1.7), Advert: 25.9 years (*SEM* 1.8), and Ghana: 25.3 years (*SEM* 2.0). From the observed incident infections, 95 near full-genome and 3 right half-genome sequences of HIV-1 were obtained from 97 participants (includes two full-genome sequences from one participant who displayed a change in HIV-1 genotype as the study progressed).

### HIV-1 subtype distribution

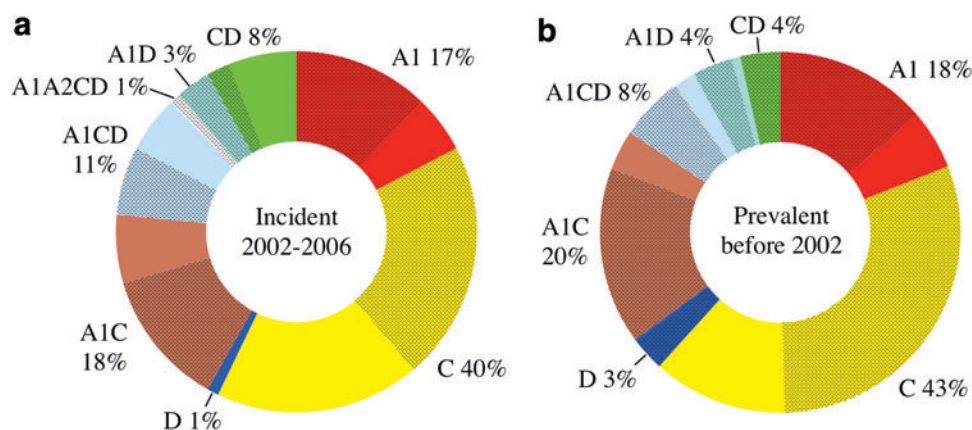
Among the incident infections, the subtype distribution was 40% C, 17% A1, 1% D, and 42% recombinants between parent subtypes A1, A2, C, and D. The overall subtype distribution within the incident infections did not vary substantially from the prevalent infections previously described from this cohort (43% C, 18% A1, 3% D, and 36% recombinant).<sup>5</sup> However, the location of the incident strain types did show a higher proportion of diverse strains from recombinants and subtype C in the Itende (rural) subpopulation when compared to the prevalent infections, Figure 1. Although the differences were not statistically significant, the prevalent infection phase of the study found that ~30% of

the pure subtype C infections were in the rural population, while in the incident infections nearly half of the subtype C strains (46%) were observed in the rural group. In addition, there were no CD recombinants among the rural prevalent infections and only a quarter (26%) of the A1CD recombinants were found in the rural population. During the incident infection study, 75% of the CD strains and 45% of the A1CD strains were observed in the rural group. While it should be noted that the prevalent infection subtypes were determined using the multi-region hybridization assay (MHA) (based on subtype-specific probe reactivity<sup>5,22</sup>) instead of full-genome sequencing and analysis (as with the incident infections), the potential for inaccurate subtype assignments among the MHA tested prevalent infections is expected to be equally distributed among the 487 prevalent infections and should not be biased toward a single urban or rural subpopulation.

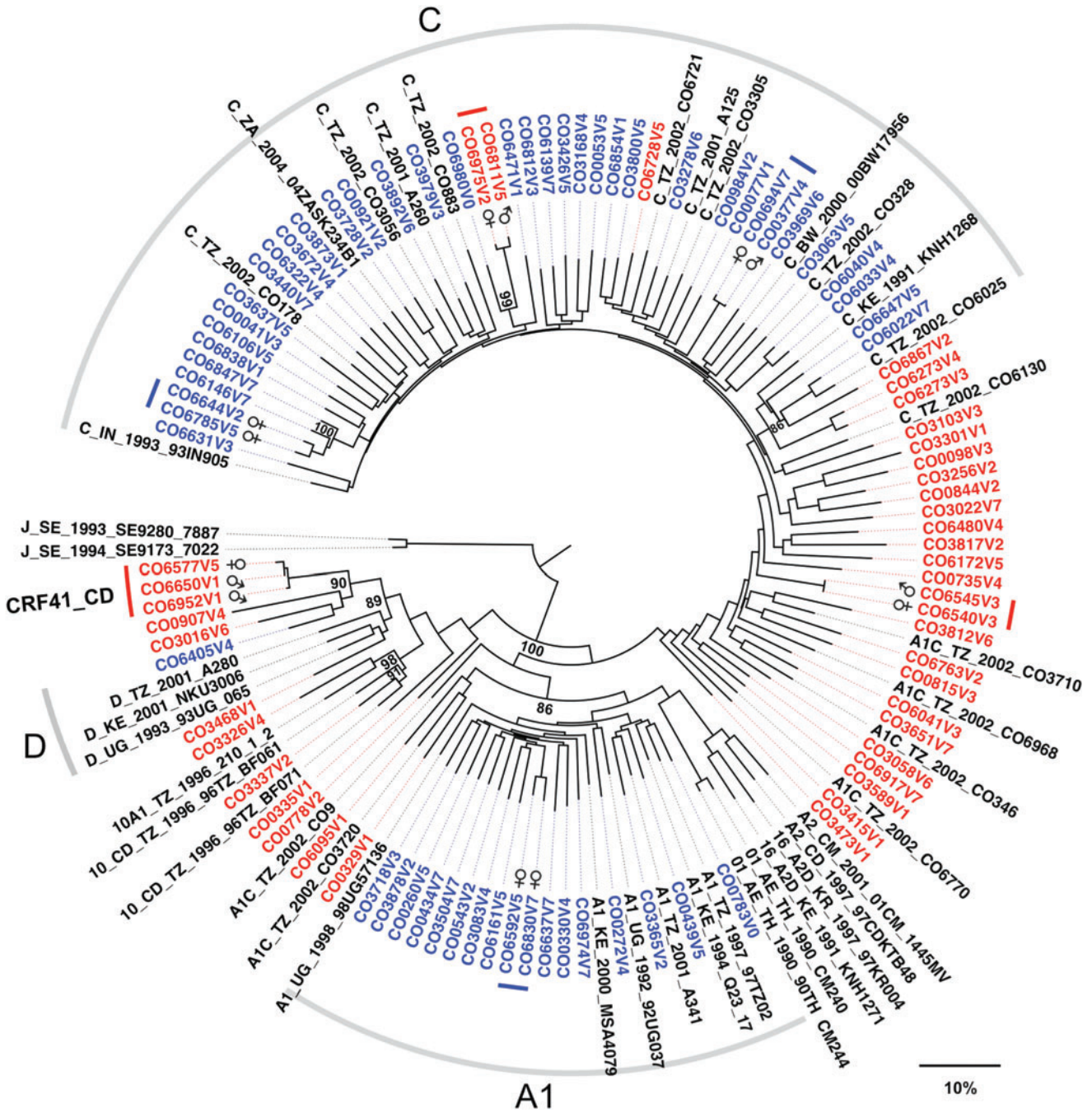
### Phylogenetic relationships between pure and recombinant strains

The neighbor-joining tree in Figure 2 depicts the phylogenetic relationships between the pure and recombinant incident strains from this study as well as previously identified reference strains from East Africa deposited in GenBank. The strains are labeled according to their visit number and subpopulation: Ghana (urban) CO0001- CO2999, Advert (urban) CO3000- CO5999, and Itende (rural) CO6000-CO6999. In general, the pure subtype incident strains cluster with interspersed reference strains from the same region, while the recombinant incident strains mostly cluster between the pure strains according to the relative amounts of genomic contribution from their parent subtypes.

Analysis of these incident infections shows the presence of closely related strains, two new CRF10\_CD variant strains, and revealed the emergence of a new circulating recombinant form: CRF41\_CD (see CRF41\_CD section). Since this study did not capture the HIV status of participants' sexual contacts, the closely related strains (denoted in Fig. 2) are proposed as being part of local transmission networks on the basis of having a genetic distance of 1.5% or lower, when



**FIG. 1.** Comparison of pure subtype and recombinant virus distribution between incident and prevalent infection phases of the CODE study. The proportions of HIV-1 strains observed among the three subpopulations within the incident infections are shown in (a) and the prevalent infections in (b). The *crosshatched areas* represent the fraction of each strain type that was found among the urban subpopulations. CODE, cohort development. Color images available online at [www.liebertpub.com/aid](http://www.liebertpub.com/aid)



**FIG. 2.** Phylogenetic relationships and transmission networks between incident infection strains. Neighbor-joining tree showing the phylogenetic relationships between incident infection strains from all three subpopulations was constructed using the full-length genomes and pure subtype and recombinant reference sequences from the prevalent infection phase of this study as well as regionally relevant strains. Pure subtype incident infection strains are shown in blue, recombinant incident strains are shown in red, and reference strains are shown in black. Tree segments corresponding with pure subtype reference strains are demarked (gray bar) and labeled according to their respective subtypes. The incident strains that belong to proposed transmission networks were identified as having a pairwise genetic distance of <1.5% when comparing *prot-RT* (HXB2 2262-3290), and are demarked with colored bars and labeled with the participant’s gender. Meaningful bootstrap values at relevant nodes are shown. The scale bar indicates a genetic distance of 10%. Color images available online at [www.liebertpub.com/aid](http://www.liebertpub.com/aid)

comparing the *prot-RT* region<sup>23,24</sup> between strains (HXB2 position 2262 to 3290). This stringent cutoff represents a genetic distance well below the 99th percentile of pairwise genetic distances for the same subgenomic region among the pure subtype C strains (the largest sample set common to each

subpopulation in this study). General similarity searches (HIVBLAST via the Los Alamos HIV database as of September 1, 2016) did not return any previously published sequences that would indicate these local transmission networks extended beyond the Mbeya area. In addition, closely related



strains were confirmed by extracting, amplifying, and sequencing virus from plasma collected during subsequent clinic visits by each participant. The CODE study design allowed this strategy to be used for all study participants.

The full-length genetic distances between the incident infection subtype C strains: median 10.5% (IQR 9.8%–11.1%) or the subtype A1 strains: median 9.8% (IQR 9.2%–10.4%) are similar to those observed within a prospective study conducted in neighboring Kenya.<sup>13</sup> Interestingly, we obtained two distinct recombinant strains, CO6273V3 an A1C recombinant and CO6273V4 an A1CD recombinant, from one participant who became multiply infected. These two strains have different recombinant structures and a full-length pairwise genetic distance of 2.5%, which is larger than the genetic distance between some of the related strains from different individuals.

#### Sequence similarity to contemporary phase I-II HIV-1 vaccine

The ongoing HVTN 100 vaccine trial (clinicaltrials.gov/show/NCT02404311) delivers a prime dose that contains gp120 from subtype C and parts of gag, pro, and gp41 from subtype B (LA1). The boost dose contains recombinant gp120 from two strains of subtype C HIV-1. All three of the gp120 immunogens were derived from African subtype C strains, 96ZM651 (Zambia), TV1 (South Africa), and 1086-C (Malawi), in an effort to produce relevant immune responses to the most predominant subtype of HIV-1 found in Africa. The percent differences between the amino acid sequences from these gp120 immunogens and the strains observed during this study vary depending on the extent of subtype matching, Table 1. Compared to the incident infection strains, the vaccine immunogens show median percent differences of 16.6%–17.1% to subtype C gp120 and 23.3%–23.8% to non-subtype C gp120 (includes recombinant gp120 containing subtype C). These differences are comparable to the intra-subtype differences 17% (range 4%–30%) and inter-subtype differences 25% (range 20%–36%) observed by Korber *et al.*<sup>25</sup>

#### Genomic structure of inter-subtype recombinants

Recombinant breakpoint analysis of recombinant strains provides a useful tool for characterizing the strain diversity within this cohort. The genomic structures of the 40 recombinant incident infection strains obtained during this study are shown in Figure 3. Analysis of those strains shows that the majority of the parent genomes are derived from subtypes A1, C, and D with a minor contribution from sub-subtype A2 to a single recombinant. The structures contain 2–17 breakpoints coming from as many as four of the parent subtypes, although the majority are A1C recombinants.

As suggested by the extent of their relatedness (Fig. 2), strains CO6540V3–CO6545V3 and CO6811V5–CO6975V2 have nearly identical parental subtype breakpoints within each pair. Strains CO6273V3 and CO6273V4 were obtained from the same participant during consecutive clinic visits and illustrate the result of recombination between two different recombinant strains within a single individual or limited network of individuals. Strains CO3337V2 and CO3326V4 illustrate the result of recombination between the previously identified CRF10\_CD<sup>26</sup> and unidentified unique recombinant forms (URF), see also Supplementary Figures S1–S3 (Supplementary Data are available online at www.liebertpub.com/aid). Strains CO6577V5, CO6650V1, and CO6952V1 have very similar genome structures and comprise the proposed CRF41\_CD cluster shown in Figure 2. The average genetic distance within the CRF41\_CD cluster (0.7%) indicates recombination and dissemination across a limited transmission network.

#### CRF41\_CD genomic structure and phylogenetic dissection

Recombinant breakpoint analysis was used to determine the breakpoints within strains CO6577V5, CO6650V1, and CO6952V1. The subgenomic fragments defined by the recombinant breakpoints were then queried against subtype reference strains to confirm the subtype assignments. Figure 4 depicts the results of phylogenetic analysis of CRF41\_CD and its subgenomic fragments, and shows the bootstrap values that support each subtype assignment. CRF41\_CD contains 9 breakpoints resulting in 10 fragments, which alternate between subtypes C and D. The majority of CRF41\_CD is subtype D with contributions from subtype C in gag, nef, and gp41. The assigned parent subtype of most fragments is supported with a bootstrap value of 93 or higher. Visual inspection identified fragment III as a poor resolution area between subtypes C and D. Informative site analysis and visual inspection identified fragments V and VIII as subtypes D and C, respectively. However, the limited lengths of those fragments prevent the attainment of significant bootstrap values; hence, the subtype assignments of fragments V and VIII were confirmed via BLAST analysis. Fragments II, IV, and VI were analyzed as a contiguous fragment spanning HXB2 nucleotide positions 1203–2162 with masking of fragments III and V.

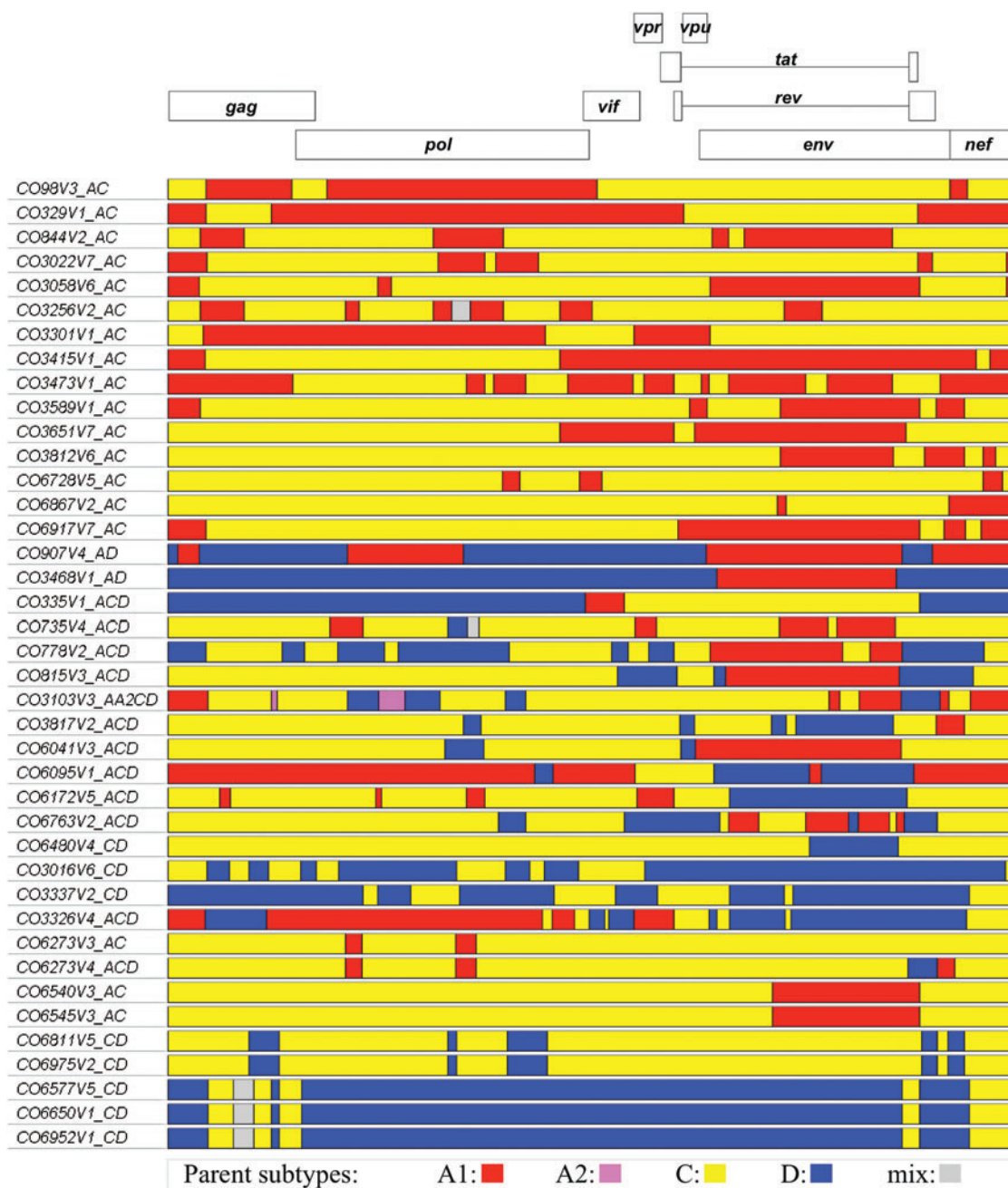
#### Discussion

The CODE study was a closed prospective cohort study involving participants from three subpopulations located in and around Mbeya, Tanzania. The incident infection phase of

TABLE 1. MEDIAN PERCENT DIFFERENCES BETWEEN THE HVTN 100 GP120 IMMUNOGENS AND THE SUBTYPE C AND NON-SUBTYPE C GP120 FROM THE INCIDENT INFECTION STRAINS

Vaccine component	Strain	Incident strain subtype C gp120 median difference (IQR), %	Incident strain non-subtype C gp120 median difference (IQR), %
Prime	96ZM651	16.6 (15.6–17.5)	23.8 (22.1–25.3)
Boost protein 1	TV1	17.1 (15.6–18.3)	23.5 (21.7–24.5)
Boost protein 2	1086-C	16.9 (16.2–17.9)	23.3 (22.0–24.9)

IQR, interquartile range.

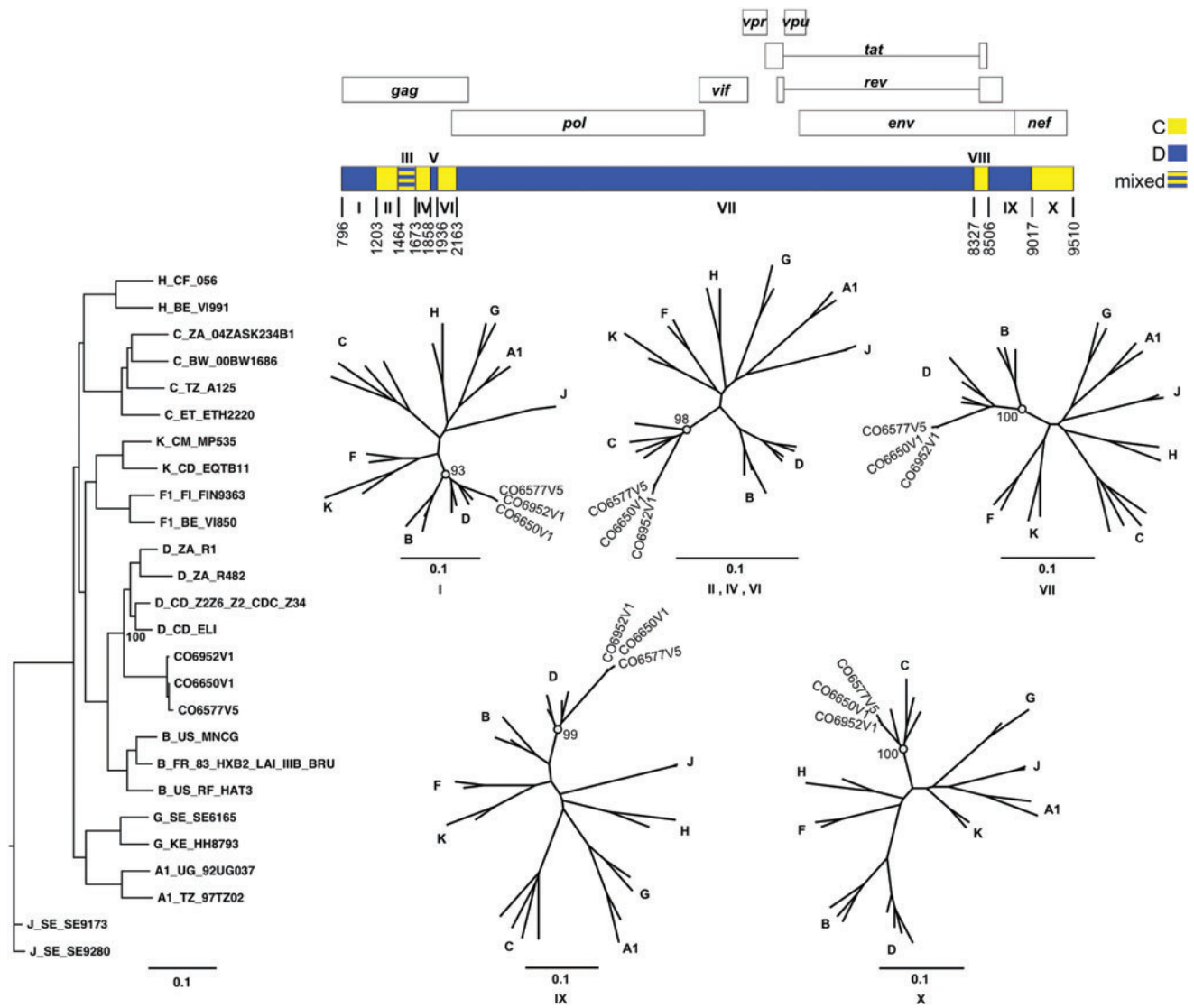


**FIG. 3.** Genome structures of the 40 HIV-1 recombinant strains identified during the CODE incident infection study. Incident recombinant genomes are depicted in relation to the HXB2 reference strain. The colored segments represent the predicted parent subtype based on recombinant breakpoint analysis. Subtype A1 is shown in red, A2 in pink, C in yellow, and D in blue. Color images available online at [www.liebertpub.com/aid](http://www.liebertpub.com/aid)

the study observed 106 new infections, from which 95 near full-length and 3 right half-length HIV-1 sequences were characterized. The incident infection strains identified during the CODE study reveal the presence of pure HIV-1 subtypes A1, C, D, a large fraction of recombinants (42%), several distinct local transmission networks, and a new CRF. Among the pure strains, the majority are subtype C (39%) followed by subtype A1 (18%) with only a single subtype D. The general pattern of pure subtype distribution within the incident infections (collected 2002–2006) is relatively unchanged compared to the subtype distribution observed during the prevalent in-

fection phase of this study (2002), Figure 1. Likewise, the parental nature of the incident recombinant strains also resembles that seen among the prevalent infections, with the majority of the genomic contributions coming from subtype C, followed by subtypes A1 and D, with a small fraction of subtype A2 within a single recombinant.

Phylogenetic analysis of the incident strains confirms that most of them are similar to previously published reference strains from the same region (Fig. 2), which is suggestive of a mature epidemic, in which the incident infections originate from endemic strains within the same populace. Since the



**FIG. 4.** CRF41\_CD genome structure and phylogenetic analysis of subgenome fragments. The recombinant genome of CRF41\_CD is depicted in relation to the HXB2 reference strain (*top*). Subgenomic fragments corresponding to parent subtype C are shown in *yellow* and subtype D in *blue*. The phylogenetic relationship between the full-length CR41\_CD strains and subtype reference strains is shown (*left*). Confirmation of the parent subtypes of the subgenomic fragments is shown using neighbor-joining trees (*center*) constructed from the corresponding numbered fragments of CRF41\_CD (*top*) and the reference strains (*left*). Meaningful bootstrap values at relevant nodes are shown. The scale bar indicates a genetic distance of 10%. Color images available online at [www.liebertpub.com/aid](http://www.liebertpub.com/aid)

recombinant incident infection strains are predominantly a mix between parent subtypes A1, C, and D, they cluster between the pure subtypes that contribute the majority of the recombinant's genetic materials, Figure 2. The few recombinant strains that do cluster among the pure subtype C strains are primarily composed of subtype C and have genetic distances similar to the pure strains. While we do not view the current phylogenetic analysis and its resulting trees as a means for reproducing transmission histories or contact networks,<sup>27,28</sup> the viral phylogenies did indicate the presence of closely related strain clusters worthy of further analysis via intracluster pairwise genetic distances.

Based on pairwise genetic distance analysis, with a cutoff of less than 1.5% in *pol*, six local transmission networks were identified. The nature of HIV evolution<sup>23</sup> (both within and between hosts) and the lack of an external reference<sup>29</sup> (root)

in distance defined networks, prevents determination of the direction of transmission or whether transmission events proceeded directly between participants or via a third party. However, distance defined transmission networks do allow us to recognize closely related strains and the subpopulation(s) within which they are observed,<sup>30</sup> Figure 2.

Of the six identified transmission networks, five of them are from the Itende (rural) subpopulation and one is from the Ghana ward (urban) subpopulation. This bears some consideration since both the Ghana and Itende subpopulations were recruited via a door-to-door campaign (instead of the city-wide advertisement campaign), and since none of the transmission networks involved study participants who bridged separate subpopulations. Inherent limitations of this study are the lack of comprehensive enrollment and its focus on capturing incident infections. As a result, we are likely

underreporting the number of incident infections and missing individuals who may be associated with these transmission networks, as well as selecting for participants with known risk factors for HIV acquisition. Indeed, hazard analysis<sup>31</sup> of this incident infection study did identify several traditional risk factors (age, education, alcohol use, and number of lifetime sexual partners) associated with HIV infection. The hazard analysis also identified young females from Itende as an adjusted multivariate risk factor; however, residence location was not an independent risk factor and there was no significant difference in the number of observed incidence infections between the two recruitment strategies used in Mbeya Town.<sup>31</sup>

The presence of several transmission networks within two of the subpopulations, but lack of transmission networks between the subpopulations, suggests that during its 3.5-year duration, the CODE study was able to access distinct transmission dynamics within the subpopulations even though they are neighboring areas of the same township. This was unexpected since all three subpopulations are within walking distance to the urban center of Mbeya Town. In that regard, the primary differences between the transmission networks would be the locale (urban vs. rural) and recruitment strategy (advertisement vs. door-to-door canvassing).<sup>5,31</sup>

In both the prevalent and incident phases of this study, the majority of the recombinant strains are URF, Figure 3. However, within the incident phase, there were three participants who became infected with structurally and phylogenetically similar strains that were identified as the new circulating recombinant form termed CRF41\_CD, Figure 4. In addition, there were two incident strains that appeared to be recombinants deriving from CRF10\_CD (Fig. 2 and Supplementary Figs. S1–S3) and there were two other pairs of URFs that may lead to new CRFs in the future, Figure 3. The substantial representation of recombinant strains within the CODE cohort is similar to that seen in another prospective cohort study conducted in neighboring Kenya.<sup>13</sup> It should be noted that while the number of URFs identified in East Africa are abundant,<sup>5,13,32–35</sup> very few of them have been subsequently observed as CRF. While this might lead to an interpretation of reduced viral fitness, our study reconfirms that URFs continue to represent a significant fraction of new infections within the HIV-1 epidemic of East Africa.

In the context of vaccine development, the new CRF41\_CD, CRF10\_CD variants, and the multitude of URFs represent additional variables that may challenge the expected benefits of a single subtype matched vaccine. Furthermore, the CRF41\_CD strains (and possibly the CRF10\_CD variants) represent known fit and infectious recombinant strains that have the ability to transmit.

The results of the prevalent infection phase of this study (Fig. 1) indicated that the majority of the pre-existing subtype C and D containing strains (pure subtypes and recombinants) were concentrated in the urban populations. The incident infection results show an increase in the representation of subtype D and CD containing recombinants among the rural population, which will further complicate the therapeutic and prophylactic vaccine options available to that group. In addition, although the three subpopulations in this study are within walking distance to each other, the presence of separated transmission networks suggests that the transmission dynamics within these groups may be distinct. In the event of

large-scale (phase III) vaccine trials within Mbeya Town, the probability of transmission between subpopulations of participants and the nature of the circulating strains within each group of participants will need to be considered both during participant selection and vaccine design.

### Acknowledgments

The authors are grateful to the dedicated and committed efforts of all study participants. They thank the researchers and the field personnel for their invaluable assistance and cooperation during the conduct of this research study. They are grateful to Dr. Paul Scott and Mr. Mark Milazzo for their help in providing demographic and risk characteristics of the participants; they also thank Dr. Mary Marovich, Dr. Barton Haynes, and Dr. Sanjay Phogat for their consultations regarding vaccine design and the components of the ALVAC/vCP2438 with bivalent C protein boost vaccine referenced herein. Disclaimer: The views expressed are those of the authors and should not be construed to represent the positions of the U.S. Army or the Department of Defense.

### Sequence Data

The 98 HIV-1 sequences obtained from this general population cohort incident infection study have been submitted to GenBank and are available under accession numbers KX907336–KX907433.

### Disclosure Statement

The material has been reviewed by the Walter Reed Army Institute of Research and there is no objection to its presentation and/or publication. The Walter Reed Army Institute of Research Institutional Research Board human use protocol #W923 (RV143). “Infectious Disease Surveillance (HIV, TB, and Malaria) and CODE Among Urban and Rural Adults in Mbeya Municipality, Tanzania,” is funded through the U.S. Military HIV Research Program (the Walter Reed Army Institute of Research and the Henry M. Jackson Foundation for the Advancement of Military Medicine). This work was supported by cooperative agreements (DAMD17-98-2-7007, W81XWH-04-2-0005, W81XWH-07-2-0067, and W81XWH-11-2-0174) between the Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc. and the U.S. Department of Defense (DOD). No competing financial interests exist.

### References

1. UNAIDS: United Republic of Tanzania—HIV and AIDS estimates. [www.unaids.org/en/regionscountries/countries/unitedrepublicoftanzania](http://www.unaids.org/en/regionscountries/countries/unitedrepublicoftanzania) (2015) accessed September 23, 2015.
2. Tanzania Commission for AIDS (TACAIDS) NBoSN, and ORC Macro. *Tanzania HIV/AIDS Indicator Survey 2003–04*. Calverton, Maryland, TACAIDS, NBS, and ORC Macro, 2005.
3. Tanzania Commission for AIDS (TACAIDS) ZACZ, National Bureau of Statistics (NBS), Office of the Chief Government Statistician (OCGS), and Macro International Inc. *Tanzania HIV/AIDS and Malaria Indicator Survey 2007–08*. Dar es Salaam, Tanzania, TACAIDS, ZAC, NBS, OCGS, and Macro International, Inc., 2008.
4. Tanzania Commission for AIDS (TACAIDS) ZACZ, National Bureau of Statistics (NBS), Office of the Chief



- Government Statistician (OCGS), and ICF International. *Tanzania HIV/AIDS and Malaria Indicator Survey 2011–12*. Dar es Salaam, Tanzania, TACAIDS, ZAC, NBS, OCGS, and ICF International, 2013.
5. Arroyo MA, Hoelscher M, Sateren W, *et al.*: HIV-1 diversity and prevalence differ between urban and rural areas in the Mbeya region of Tanzania. *AIDS* 2005;19:1517–1524.
  6. Njai HF, Ewings FM, Lyimo E, *et al.*: Deciphering the complex distribution of human immunodeficiency virus type 1 subtypes among different cohorts in Northern Tanzania. *PLoS One* 2013;8:e81848.
  7. Herbinger KH, Gerhardt M, Piyasirisilp S, *et al.*: Frequency of HIV type 1 dual infection and HIV diversity: Analysis of low- and high-risk populations in Mbeya Region, Tanzania. *AIDS Res Hum Retroviruses* 2006;22:599–606.
  8. Nyombi BM, Kristiansen KI, Bjune G, Muller F, Holm-Hansen C: Diversity of human immunodeficiency virus type 1 subtypes in Kagera and Kilimanjaro regions, Tanzania. *AIDS Res Hum Retroviruses* 2008;24:761–769.
  9. Kiwelu IE, Novitsky V, Margolin L, *et al.*: HIV-1 subtypes and recombinants in Northern Tanzania: Distribution of viral quasispecies. *PLoS One* 2012;7:e47605.
  10. Arroyo MA, Hoelscher M, Sanders-Buell E, *et al.*: HIV type 1 subtypes among blood donors in the Mbeya region of southwest Tanzania. *AIDS Res Hum Retroviruses* 2004;20:895–901.
  11. Bakari M, Aboud S, Nilsson C, *et al.*: Broad and potent immune responses to a low dose intradermal HIV-1 DNA boosted with HIV-1 recombinant MVA among healthy adults in Tanzania. *Vaccine* 2011;29:8417–8428.
  12. Munseri PJ, Kroidl A, Nilsson C, *et al.*: Priming with a simplified intradermal HIV-1 DNA vaccine regimen followed by boosting with recombinant HIV-1 MVA vaccine is safe and immunogenic: A phase IIa randomized clinical trial. *PLoS One* 2015;10:e0119629.
  13. Billings E, Sanders-Buell E, Bose M, *et al.*: The number and complexity of pure and recombinant HIV-1 strains observed within incident infections during the HIV and Malaria Cohort Study Conducted in Kericho, Kenya, from 2003 to 2006. *PLoS One* 2015;10:e0135124.
  14. Gaschen B, Kuiken C, Korber B, Foley B: Retrieval and on-the-fly alignment of sequence fragments from the HIV database. *Bioinformatics* 2001;17:415–418.
  15. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 2011;28:2731–2739.
  16. Deng W, Maust BS, Nickle DC, *et al.*: DIVEIN: A web server to analyze phylogenies, sequence divergence, diversity, and informative sites. *Biotechniques* 2010;48:405–408.
  17. Schultz AK, Zhang M, Leitner T, *et al.*: A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes. *BMC Bioinformatics* 2006;7:265.
  18. Salminen MO, Carr JK, Burke DS, McCutchan FE: Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS Res Hum Retroviruses* 1995;11:1423–1425.
  19. Tovanabutra S, Watanaveeradej V, Viputtikul K, *et al.*: A new circulating recombinant form, CRF15\_01B, reinforces the linkage between IDU and heterosexual epidemics in Thailand. *AIDS Res Hum Retroviruses* 2003;19:561–567.
  20. Kijak GH, Tovanabutra S, Beyrer C, *et al.*: RecDraw: A software package for the representation of HIV-1 recombinant structures. *AIDS Res Hum Retroviruses* 2010;26:1317–1321.
  21. Saathoff E, Pritsch M, Geldmacher C, *et al.*: Viral and host factors associated with the HIV-1 viral load setpoint in adults from Mbeya Region, Tanzania. *J Acquir Immune Defic Syndr* 2010;54:324–330.
  22. Hoelscher M, Dowling WE, Sanders-Buell E, *et al.*: Detection of HIV-1 subtypes, recombinants, and dual infections in east Africa by a multi-region hybridization assay. *AIDS* 2002;16:2055–2064.
  23. Leigh Brown AJ, Lycett SJ, Weinert L, *et al.*: Transmission network parameters estimated from HIV sequences for a nationwide epidemic. *J Infect Dis* 2011;204:1463–1469.
  24. Pacha LA, Hakre S, Myles O, *et al.*: Centralized HIV Program Oversight. *Medicine (Baltimore)* 2015;94:e2093.
  25. Korber B, Gaschen B, Yusim K, Thakallapally R, Kesmir C, Detours V: Evolutionary and immunological implications of contemporary HIV-1 variation. *Br Med Bull* 2001; 58:19–42.
  26. Koulinska IN, Ndung'u T, Mwakagile D, *et al.*: A new human immunodeficiency virus type 1 circulating recombinant form from Tanzania. *AIDS Res Hum Retroviruses* 2001;17:423–431.
  27. Lewis F, Hughes GJ, Rambaut A, Pozniak A, Leigh Brown AJ: Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS Med* 2008;5:e50.
  28. Poon AF, Joy JB, Woods CK, *et al.*: The impact of clinical, demographic and risk factors on rates of HIV transmission: A population-based phylogenetic analysis in British Columbia, Canada. *J Infect Dis* 2015;211:926–935.
  29. Romero-Severson EO, Bulla I, Leitner T: Phylogenetically resolving epidemiologic linkage. *Proc Natl Acad Sci U S A* 2016;113:2690–2695.
  30. Wertheim JO, Leigh Brown AJ, Hepler NL, *et al.*: The global transmission network of HIV-1. *J Infect Dis* 2014; 209:304–313.
  31. Geis S, Maboko L, Saathoff E, *et al.*: Risk factors for HIV-1 infection in a longitudinal, prospective cohort of adults from the Mbeya Region, Tanzania. *J Acquir Immune Defic Syndr* 2011;56:453–459.
  32. Dowling WE, Kim B, Mason CJ, *et al.*: Forty-one near full-length HIV-1 sequences from Kenya reveal an epidemic of subtype A and A-containing recombinants. *AIDS* 2002;16: 1809–1820.
  33. Harris ME, Serwadda D, Sewankambo N, *et al.*: Among 46 near full length HIV type 1 genome sequences from Rakai District, Uganda, subtype D and AD recombinants predominate. *AIDS Res Hum Retroviruses* 2002;18:1281–1290.
  34. Arroyo MA, Sateren WB, Serwadda D, *et al.*: Higher HIV-1 incidence and genetic complexity along main roads in Rakai District, Uganda. *J Acquir Immune Defic Syndr* 2006;43: 440–445.
  35. Hoelscher M, Kim B, Maboko L, *et al.*: High proportion of unrelated HIV-1 intersubtype recombinants in the Mbeya region of southwest Tanzania. *AIDS* 2001;15:1461–1470.

Address correspondence to:

Sodsai Tovanabutra  
 U.S. Military HIV Research Program  
 Henry M. Jackson Foundation  
 Walter Reed Army Institute of Research  
 503 Robert Grant Avenue, Room 2N25  
 Silver Spring 20910  
 Maryland

E-mail: stovanabutra@hivresearch.org