Distinct Human Immunodeficiency Virus Type 1 Subtype A Virus Circulating in West Africa: Sub-Subtype A3

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters.

Citation

Published Version
doi:10.1128/JVI.78.22.12438-12445.2004

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:26881743

Terms of Use
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
Distinct Human Immunodeficiency Virus Type 1 Subtype A Virus Circulating in West Africa: Sub-Subtype A3

Seema Thakore Meloni,1 Bohye Kim,2 Jean-Louis Sankalé,1 Donald J. Hamel,1 Sodsai Tovanabutra,2 Souleymane Mboup,3 Francine E. McCutchan,2 and Phyllis J. Kanki1*

Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, Massachusetts; and Henry M. Jackson Foundation, Rockville, Maryland; and Laboratoire de Bactériologie-Virologie, Université Cheikh Anta Diop, Dakar, Senegal

Received 16 February 2004/Accepted 22 June 2004

Phylogenetic analyses demonstrate significant diversity in worldwide circulating strains of human immunodeficiency virus type 1 (HIV-1). Detailed studies have revealed a complex pattern of intersubtype recombinations, as well as evidence of sub-subtypes circulating in various populations. In this study, we characterized an HIV-1 strain that had previously been identified as a distinct subcluster within the subtype A radiation based on partial sequence data. These viruses were of particular interest given that we recently found that their prevalence was significantly higher in dually infected individuals compared to women who were singly infected with HIV-1. Five viruses isolated from commercial sex workers in Dakar, Senegal, were full-length PCR amplified and sequenced. Phylogenetic analyses indicated that, whereas three of these viruses were closely related and clustered overall within the HIV-1 subtype A radiation, they were distinct from previously characterized sub-subtype A1 and A2 viruses. The clustering pattern was maintained in the individual gag, pol, and env regions of the genome. Distance calculations between these viruses, which we termed A3, and other reference sub-subtype A1 and A2 viruses fell in the range of distances between previously characterized sub-subtype groups. In addition, we found evidence of two A3-containing recombinants in our cohort. These recombinants are mosaics composed of sequence from both sub-subtype A3 and CRF02_AG, the major circulating recombinant form in this West African population. Based on phylogenetic analyses, we propose that the group of viruses found in the Dakar sex worker cohort, previously referred to as HIV-1 A subcluster 2, be referred to as HIV-1 sub-subtype A3.

Subtype nomenclature provides us with an important tool that enables us to monitor the spread of human immunodeficiency virus type 1 (HIV-1), and the creation of an international HIV genetic sequence database has facilitated better classification of globally circulating viral strains. Phylogenetic analyses of the env and gag genes of HIV-1 indicate that there are three distinct groups (M, N, and O) of the virus circulating worldwide. Within the major group M, there are currently nine defined subtypes, or clades (A to D, F to H, J, and K) (3, 6, 18, 19, 23, 24, 34). In addition, there are a number of circulating recombinant forms (CRFs) in the world, which are associated with significant transmission and disease in various populations worldwide (23). Detailed phylogenetic studies have revealed evidence of sub-subtypes, which are characterized as regional subclusters composed of distinct lineages that are closely related to a given subtype (11, 21, 36–38). Sub-subtypes F2 (36) and A2 (11) have been identified and characterized. There have also been suggestions that certain previously identified subtypes could have been categorized as sub-subtypes (19, 26, 36–38). For example, there is evidence suggesting that the previously identified subtype D viruses might have been more appropriately named sub-subtype B2 (26, 38).

Due to the complexity of the nomenclature issues, an international group of experts convened to define methods of classifying HIV viral sequences (23, 26, 27). Emphasis was placed on improving precision with regard to subtype classifications. The group pointed out the importance of recognizing the presence of sub-subtypes. In their recommendations, they mapped out criteria for defining sub-subtypes by genetic distance estimates. Within each of the subtypes and sub-subtypes identified, a range of genetic diversity has been noted (1, 11, 13, 18, 30). Subtypes are genetically equidistant from one another, with the intersubtype nucleotide distances ranging from 15 to 22% in the gag gene and from 20 to 30% in the env gene (7, 11, 12, 26). Intrasubtype distances range from 3 to 10% in the gag gene and 5 to 12% in the env gene (7, 11, 26). Various reports indicate that intersubtype diversity ranges from 7 to 12% in the gag gene from 11 to 16% in the env gene (11, 26, 27).

Subtype A viruses are interesting to study not only because they are the second most prevalent strain of HIV-1 circulating globally but also because portions of the subtype are found in a majority of the CRFs, particularly in those of epidemiologic consequence (11, 23, 25, 26). This group of viruses is estimated to account for 27% of global HIV-1 infections, with a majority of the cases represented by the CRF02_AG strains (23). HIV-1 infection in West Africa is principally due to CRF02_AG (2, 4, 6, 19, 22, 23, 25). Recent work conducted in our laboratory revealed that a distinct group of viruses within the HIV-1 subtype radiation is also present in Senegal (30, 32, 35). Phylogenetic analyses indicated that this group of viruses, which we previously referred to as HIV-1A subcluster 2, differed from sub-subtypes A1 and A2 (30, 32). Since the categorization of these viruses was primarily based on sequencing of the
C2-V3 region of the env gene and a small region of the gag gene, further analysis was warranted. In addition, we were particularly interested in further characterization of this virus after we found that it was actually overrepresented in women dually infected with HIV-1 and HIV-2 compared to women infected with HIV-1 alone (32). The goal of this investigation was to obtain additional evidence to determine the genetic relatedness and relationship of this subcluster of subtype A viruses with respect to reference HIV-1 A1 and HIV-1 A2 viruses; we therefore analyzed full-length HIV-1 sequences. Sequences from five different subcluster two viruses were obtained and analyzed. Pairwise genetic distances were evaluated and revealed that three of the five viruses could be designated a distinct sub-subtype within the HIV-1 A radiation, which we have named A3.

**MATERIALS AND METHODS**

**Study population and sample collection.** Since 1985, we have conducted a prospective study of registered female sex workers (FSWs) in Dakar, Senegal; blood samples and questionnaire data were collected after obtaining informed consent. The specific details of the study recruitment procedures and methods have been previously described elsewhere (14, 15). Serostatus was determined by immunoblot on whole virus lysates and recombinant envelope peptides and by diagnostic HIV-1 and/or HIV-2 PCR. Time of infection for women who converted to HIV-positive serostatus while in the study was estimated as the midpoint between their last seronegative and first seropositive samples. For women who appeared to have evidence of the HIV-1 subtype A subcluster 2 virus in the C2-V3 region of the env gene (350 bp). For C2-V3 sequence, DNA samples were amplified by nested PCR with primers and reaction conditions that have been described previously (31, 39). The PCR products were purified and directly sequenced by using the second-round primers. When sequence data could not be obtained through direct sequencing, purified products were cloned into the pCR2.1 vector (T/A Cloning; Invitrogen, San Diego, Calif.), and sequence data were generated from clones.

**DNA amplification and sequencing.** Proviral DNA was extracted from cultured peripheral blood mononuclear cells by using a kit from Qiagen, Inc. (Chatsworth, Calif.). Preliminary subtype and sub-subtype designation was done based on the C2-V3 region of the env gene (500 bp). For C2-V3 sequence, DNA samples were amplified by nested PCR with primers and reaction conditions that have been described previously (31, 39). The PCR products were purified and directly sequenced by using the second-round primers. When sequence data could not be obtained through direct sequencing, purified products were cloned into the pCR2.1 vector (T/A Cloning; Invitrogen, San Diego, Calif.), and sequence data were generated from clones.

**Sequence analysis and statistical methods.** A multiple alignment of the virtually full-length genome with reference sequences of all major subtypes and sub-subtypes (A1, A2, B-D, F1, F2, G, H, J, and K), as well as the full-length sequences we had obtained, was manually generated in a Genetic Data Environment window (33). A maximum-likelihood tree of nearly full-length sequences was produced by using the DNAML program in the PHYLIP package version 3.6 (9) using a transition/transversion ratio of 1.42 (16) and empirical base frequencies. Sequences were gap stripped. Maximum-likelihood trees for gag, pol, and env subregions were also generated. The sequences were also analyzed for hypermutation by using the HYPERMUT program (http://www.hiv.lanl.gov).

Pairwise genetic distances were calculated by using the DNdist program in the PHYLIP version 3.6 package (9), with an F84 model of evolution and a transition/transversion ratio of 1.42. Calculations were done with sequences from the following subtype and sub-subtype categories: A1 (KEQ23, UG037), A2 (94cy013, 97caf0b48), B (HXB2, SF2), C (ETH22230, 93IN905), D (NDK, KE2059), F1 (V1850, F9363), F2 (CAMS3657, MP255), G (NG093, DRC.BLT), H (VI991, CF056), J (SE9713.3, SE9280), and K (EGTB117, MP5535c). The alignments were gap stripped.

**TABLE 1. Characteristics of samples from Dakar female sex worker cohort selected for full-length sequencing***

<table>
<thead>
<tr>
<th>Sample identification</th>
<th>HIV-1 case type</th>
<th>Serohistory Yr of HIV-1 seroconversion</th>
<th>Sample yr</th>
<th>env sub-subtype</th>
<th>gag sub-subtype</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDJ360</td>
<td>Incident</td>
<td>HIV-1 only</td>
<td>1991</td>
<td>A3**</td>
<td>A3**</td>
</tr>
<tr>
<td>DDJ369</td>
<td>Incident</td>
<td>HIV-1 only</td>
<td>2001</td>
<td>A3**</td>
<td>A3**</td>
</tr>
<tr>
<td>DDJ362</td>
<td>Prevalent</td>
<td>HIV-1 and HIV-2</td>
<td>1989*</td>
<td>A3**</td>
<td>A3**</td>
</tr>
<tr>
<td>DDJ364</td>
<td>Prevalent</td>
<td>HIV-1 and HIV-2</td>
<td>1996</td>
<td>A3**</td>
<td>A3**</td>
</tr>
</tbody>
</table>

* * * Since these were already seroprevalent at the time they entered the study, the dates correspond to the year of first bleed; +, at the time these samples were first identified for full-length sequencing, we had not yet verified that they could be referred to as sub-subtype A3 (and were referring to them as HIV-1 A subcluster 2 viruses), but for the purposes of this table we are designating the samples as A3.

**FIG. 1.** Maximum-likelihood tree of full-length sequences including three A3 sequences (DDJ369, DDJ360, and DDJ369) and reference sequences from all known subtypes and sub-subtypes (A1, A2, B, C, D, F1, F2, G, H, J, and K; Los Alamos HIV Sequence database [http://hiv-web.lanl.gov/]). The alignments were gap stripped, and a transition/transversion ratio of 1.42 was used. The scale bar indicates 1% nucleotide sequence divergence. The A3 sequences are indicated in boldface.
Sequence alignments were also used to generate bootscan graphs for recombination analyses by using the SIMPLOT version 3.4 beta program (17). Bootscan analyses were performed on neighbor-joining (NJ) trees, using a 400-bp moving window along the aligned sequences in 50-bp increments. A transition/transversion ratio of 1.42 was used with an F84 model of evolution. Gaps were stripped from the alignments. In preliminary analyses, reference sequences from all major subtypes, sub-subtypes, and CRF02_AG were used. Based on recombination patterns from the initial analyses, a subset of sequences was selected for the final bootscan graphs; the subset of reference sequences used included sub-subtype A3 (DDI578, DDJ360, and DDJ369), CRF02_AG (DJ264, IbNG, SE7812, 1MP1211, and MP807), subtype G (G6165, HH8793, NG083, and DRC.BL), and subtype B (HXB2, RF, SF2, and RL42) as the outgroup. The reference sequences were grouped and the final bootscan plot was generated by using the consensus for the reference sequence groups. One hundred bootstrap replicates were analyzed for the plots. Bootstrap values were plotted at the midpoint of each window along the sequence. Breakpoints were determined by using informative site analyses and by maximizing $\chi^2$ values (28). Using the breakpoints, sequence alignments were clipped into appropriate subregions for further analysis. NJ trees for each of the recombined subregions were generated to confirm the results of the bootscan. All NJ trees were generated by the SEQBOOT, DNADIST, NEIGHBOR, and CONSENSE programs in the PHYLIP version 3.6 package (9). For all NJ trees, the F84 model of evolution and a transition/transversion ratio of 1.42 was used.

FIG. 2. Maximum-likelihood subregion trees. (a) gag; (b) pol; (c) env. Sequences maintain unique clustering patterns in individual gene regions. The alignments were gap stripped, and these trees were generated by using the DNAML program in the PHYLIP version 3.6 package (9). A transition/transversion ratio of 1.42 was used for all trees. The scale bar indicates 1% nucleotide sequence divergence. For these trees, reference sequences from all major subtypes and sub-subtypes (A1, A2, B, C, D, F1, F2, G, H, J, and K) were used. The A3 sequences are indicated in boldface.

Sequence information. The full-length sequences have been deposited into the GenBank database under accession numbers AY521629 to AY521633.

RESULTS

We amplified nearly full-length genome of viruses from five women in our cohort that we had previously identified as having evidence of the HIV-1 subtype A subcluster 2 virus in the C2-V3 region of the env gene and the 3'-p24/5'-p7 region of the gag gene (Table 1); in the table, we referred to these sequences as A3. Three of the women (DDI579, DDJ360, and DDJ369) had seroincident infections and were infected with HIV-1 only. The remaining two women (DDJ362 and DDJ364) we selected for sequencing were seroprevalent upon entry into the cohort and were dually infected with HIV-1 and HIV-2.

Characterization of HIV-1 A subcluster 2 as sub-subtype A3.
To understand the phylogenetic relationships of the newly characterized viruses, we constructed trees from the near full-length genome. Initial phylogenetic tree analyses with se-
quences from all five women revealed that viruses from two of the women (DDJ362 and DDJ364) had branching patterns that suggested they might be recombinants; the details of these analyses are discussed in greater detail below. Figure 1 presents a maximum-likelihood tree with the three remaining sequences (DDJ579, DDJ360, and DDJ369) that contained unique sequence across the length of the entire genome. These sequences clustered overall with HIV-1 subtype A viruses but formed a subcluster that was distinct from A1 sequences as well as from the recently described A2 viruses. It should be noted that because of its unique branching pattern within the subcluster, DDJ369 was also analyzed as a possible recombinant. Bootscan analyses revealed that DDJ369 did not appear to be a recombinant and clustered with the other A3 sequences across the entire length of the genome (data not shown). Finally, for the three sequences with distinct sequence across the length of the genome, all reading frames were open and of complete length. None of the genomes had major insertions or deletions. For sequence DDJ369, the predicted start codon of the env gene had a transition from ATG to ACG, thus changing the amino acid from methionine (M) to a threonine (T); DDJ369 has a methionine three codons downstream, which might serve as the actual start codon for the env. Overall, when the sequences were trimmed into subregions based on the open reading frames for major genes, we found that the sequences maintained their clustering pattern (Fig. 2). Finally, the sequences were analyzed by using the HYPERMUT program (http://www.hiv.lanl.gov), and no evidence of hypermutation was found.

Pairwise distance analysis has previously been used to distinguish between subtypes and sub-subtypes (1, 11, 26, 37). Intersubtype and intrasubtype distances were calculated for the full-length sequences based on reference strains (Fig. 3). These values are useful in determining whether a lineage is more likely a subtype or a sub-subtype (27). By convention, distance values that were less than the distance between different subtypes but greater than the distance within subtype categories were considered to be suggestive of a sub-subtype (4). The genetic distances between sub-subtypes A3 and A1, as well as A3 and A2, fall in the range of 9 to 13%, which is within the range of distances seen between other sub-subtypes (A1:A2 and F1:F2) and subtypes that should have been designated sub-subtypes (B:D; Fig. 3). Overall, all of the sub-subtype genetic distance values fell between the values for the intersubtype and intrasubtype comparisons.

Identification of recombinant genomes. As mentioned above, two of the five samples that we had full-length sequenced revealed distinct clustering patterns when originally put into a phylogenetic tree. We generated two separate maximum-likelihood trees to highlight the clustering pattern of sequences DDJ362 and DDJ364 (Fig. 4a and b). For the first (Fig. 4a), the maximum-likelihood tree was generated by using major subtype and sub-subtype reference sequences, but no CRF02_AG reference sequences. In this tree, DDJ362 and DDJ364 clustered together overall with A3 sequences but formed a small subgroup. However, when CRF02_AG was added to the tree (Fig. 4b), the DDJ364 sequence clustered with CRF02_AG sequences. Although the DDJ362 sequence continued to cluster with the A3 sequences, the branching pattern suggested that it might be a recombinant sequence.
Similarly, the viral sequence from DDJ364 was composed of a mosaic of A3 and CRF02_AG. As shown in Fig. 4d, the sequence clustered with A3 in the V1 to V5 regions of env. In addition, the sequence clustered with A3 in portions of the gag, pol, and nef genes.

**DISCUSSION**

Through phylogenetic analyses, we confirmed that the subcluster of viruses circulating in the Dakar sex worker cohort differs from the previously characterized HIV-1 sub-subtypes A1 and A2. The subjects chosen for more detailed analyses were previously characterized as being infected with HIV-1 subtype A subcluster 2 viruses (30, 32). Examination of full-length sequences revealed that three subjects had sequences that form a distinct subcluster. Furthermore, phylogenetic analyses of subregions verified that these viruses maintained their distinct clustering patterns in the major genes. Genetic distance analysis has previously been identified as a valid method of distinguishing between the subtypes and sub-subtypes of HIV-1. An analysis of nucleotide distances revealed that our A3 sequences differed from previously identified and characterized A1 and A2 sequences by approximately 9 to 13% across the entire genome. This difference was comparable to those found between other sub-subtypes, such as F1:F2, and subtypes B:D. Based on phylogenetic information, we believe that the group of viruses in our cohort previously designated a subcluster within HIV-1 A can be considered a sub-subtype of HIV-1 A. We propose that the subcluster be referred to as HIV-1 sub-subtype A3.

When looking at our three sequences that clustered together as sub-subtype A3, we found that the virus from DDJ369 was more distant from those from DDJ359 and DDJ360, respectively, than they were from each other. We conducted bootscan analyses and ruled out that the sequence from subject DDJ369 was a recombinant. This finding suggests that either diversity
FIG. 4—Continued.
has accrued in the A3 sequences in Senegal over the years or that there might have been more than one introduction of A3 into Senegal, such that the virus diversified prior to its importation into Senegal.

Viral recombination is generally seen in areas that have several subtypes or recombinant forms of the virus circulating in the population (5). In areas where multiple subtypes coinfection, it is estimated that intersubtype recombinant sequences (ISRs) account for between 8 and 24% of infections (23). Senegal has a high prevalence of CRF02_AG infections (4, 24, 25, 30, 32) and recombination with this form of the virus could have been predicted. Given that subtype A viruses have been identified in a majority of the circulating intersubtype recombinants, as well as both intergroup (MO) recombinants (10, 20, 24), it is also not surprising to find that the HIV-1 A3 viruses are involved in recombination events. In the present study, we found two viruses with recombinant genomes, where the HIV-1 A3 sequence was maintained in the env gene. Interestingly, the two subjects that had recombinant viruses also happened to be dually infected, with HIV-1 and HIV-2. However, given that both subjects DDJ362 and DDJ364 were seroregressive with both HIV-1 and HIV-2 upon entry into the cohort, it is difficult to speculate on the implications of infection with these HIV-1 recombinants. Another point to note is that these women were infected with A3-containing recombinant viruses as early as 1989. That suggests that A3 might have been in the cohort prior to 1989 for recombination events to take place. Alternatively, the virus might have recombined in another country prior to 1989 and then been imported into the Dakar FSW cohort already a recombinant. However, given we have no information regarding the actual time of infection for these two women, we cannot make assumptions as to when A3 was first introduced into the population. Ultimately, the discovery that the viruses from these two women were recombinants reemphasizes the importance of full-length sequencing in subtype and sub-subtype characterization; had we based our sub-subtype designations on env and gag results only, these two women would have originally been characterized as infected with A3.

A clearer picture of sequence diversity in a given population will lead to better understanding and insights into the biologic consequences of viral diversity. Identification of the different subtypes and sub-subtypes of HIV-1 circulating in a given region will aid in molecular surveillance and development of relevant vaccine candidates (4). Since sequence changes are thought to represent, in part, escape from selective pressure, genetic analyses and molecular surveillance are critical to understanding host immune response (21). In addition, as subtype-specific vaccines are being produced and tested, it is imperative that we understand the scope of diversity. Phylogenetic analyses have revealed a significant amount of sequence divergence within each of the subtypes, and it may well be that understanding regional differences is requisite for subtype-specific vaccines to be effective (8).

In summary, our data indicate that the women previously identified as being infected by HIV-1 subtype A subcluster 2 viruses can be classified as infected with HIV-1 sub-subtype A3. The phylogenetic analyses of the full-length genomes corroborate the findings from the original studies (30, 32). Continued monitoring will indicate the future role of this subtype in the epidemic and will demonstrate the contribution of HIV-1 sub-subtype A3 to recombinant viruses. Further molecular epidemiologic studies will require additional discrimination of this particular sub-subtype from other circulating subtype A viruses to identify its role in the epidemic.

ACKNOWLEDGMENTS

We thank Abdoulaye Dieng Sarr, Christopher Mullins, Aissatou Guéye-Ndiaye, Ibrahim Traoré, Mamadou Cité Dia, and Ousmane Diouf for their continued assistance and support on this project. We are extremely grateful for all of the assistance provided by the Mucutchan lab in generating the full-length sequences. This study was supported in part by U.S. Army grant DAMD 17-95-C-5005 and by NIH grants AI43879 and AI467274.

REFERENCES

virus type 1 populations with a known transmission history. J. Virol. 71:4761–4770.