An Isolate of Human Immunodeficiency Virus Type 1 Originally Classified as Subtype I Represents a Complex Mosaic Comprising Three Different Group M Subtypes (A, G, and I)

FENG GAO,1 DAVID L. ROBERTSON,2 CATHERINE D. CARRUTHERS,1 YINGYING LI,1 ELIZABETH BAILES,3 LEONDIOS G. KOSTRIKIS,4 MIKA O. SALMINEN,5 FREDERIC BIBOLLET-RUCHE,1 MARTINE PEETERS,6 DAVID D. HO,4 GEORGE M. SHAW,7,8 PAUL M. SHARP,9 AND BEATRICE H. HAHN1,10

Department of Medicine and Microbiology,1 Howard Hughes Medical Institute,2 University of Alabama at Birmingham,3 Birmingham, Alabama 35294; Laboratory of Structural and Genetic Information, CNRS-EP 91, Marseilles 13402,4 and Laboratoire Retrovirus, ORSTOM, Montpellier,6 France; Division of Genetics, Queens Medical Centre, University of Nottingham, Nottingham, United Kingdom;8 Aaron Diamond AIDS Research Center, The Rockefeller University, New York, New York 100166; and Department of Infectious Diseases, National Public Health Institute, Helsinki 00300, Finland9

Received 26 May 1998/Accepted 18 August 1998

Full-length reference clones and sequences are currently available for eight human immunodeficiency virus type 1 (HIV-1) group M subtypes (A through H), but none have been reported for subtypes I and J, which have only been identified in a few individuals. Phylogenetic information for subtype I, in particular, is limited since only about 400 bp of env gene sequences have been determined for just two epidemiologically linked viruses infecting a couple who were heterosexual intravenous drug users from Cyprus. To characterize subtype I in greater detail, we employed long-range PCR to clone a full-length provirus (94CY032.3) from an isolate obtained from one of the individuals originally reported to be infected with this subtype. Phylogenetic analysis of C2-V3 env gene sequences confirmed that 94CY032.3 was closely related to sequences previously classified as subtype I. However, analysis of the remainder of its genome revealed various regions in which 94CY032.3 was significantly clustered with either subtype A or subtype G. Only sequences located in vpr and nef, as well as the middle portions of pol and env, formed independent lineages roughly equidistant from all other known subtypes. Since these latter regions most likely have a common origin, we classify them all as subtype I. These results thus indicate that the originally reported prototype subtype I isolate 94CY032 represents a triple recombinant (A/G/I) with at least 11 points of recombination crossover. We also screened HIV-1 recombinants with regions of uncertain subtype assignment for the presence of subtype I sequences. This analysis revealed that two of the earliest mosaics from Africa, Z321B (A/G?) and MAL (A/D/?), contain short segments of sequence which clustered closely with the subtype I domains of 94CY032.3. Since Z321 was isolated in 1976, subtype I as well as subtypes A and G must have existed in Central Africa prior to that date. The discovery of subtype I in HIV-1 hybrids from widely distant geographic locations also suggests a more widespread distribution of this virus subtype, or at least segments of it, than previously recognized.

A molecular epidemiological survey of human immunodeficiency virus type 1 (HIV-1) conducted in Cyprus in 1994 revealed a surprising degree of genetic variation among locally circulating strains (14). Kostritis and colleagues screened 25 HIV-1-infected individuals attending an outpatient clinic in Nicosia by performing phylogenetic analyses of C2-V3 env gene sequences amplified from their uncultured peripheral blood mononuclear cells (PBMCs). This identified representatives of several different HIV-1 group M clades, including subtype A, C, and F strains, that are not commonly found in European populations. Moreover, both members of a heterosexual couple with a history of intravenous drug use and documented travel outside of Cyprus were found to be infected with similar viruses that could not be assigned to any of the previously defined HIV-1 subtypes. These viruses formed an independent lineage roughly equidistant from all other group M subtypes, and so it was proposed to classify them as members of a new clade, termed subtype I (14). At about the same time as this initial description of subtype I, it was realized that numerous HIV-1 strains are mosaics of sequences from more than one clade (18, 19). Subsequent confirmation of the widespread occurrence of such hybrid viruses, often with evidence of multiple recombination crossovers along the genome (3, 4, 6, 9, 10, 17, 23), indicated that classification and definition of new subtypes should be based on complete genomic sequences (9, 16, 17). This is particularly important for viruses originating from geographic regions where multiple subtypes cocirculate since these have a high probability of being recombinant. To characterize subtype I in greater detail, we thus cloned a full-length provirus from a short-term cultured, primary isolate established from one of the two individuals (H032) infected with this subtype (14). Using primers corresponding to the RNA primer binding site (5′-TCTCTAaggGCGGCGCCGAA//AGACGGGCACACACTACIT-3′), we used 10234 Fonds Document 1.

1 Corresponding author. Mailing address: Department of Medicine and Microbiology, University of Alabama at Birmingham, 701 S. 19th St., LEHR 613, Birmingham, AL 35294. Phone: (205) 934-0412. Fax: (205) 934-1389. E-mail: bhahn@cordelia.dom.uab.edu.
to amplify nearly full-length genomic fragments (9, 10, 21) that contained all coding and regulatory regions except for 102 bp of 5' unique LTR sequences (U5). Amplification products were subcloned into a plasmid vector and mapped by restriction enzyme digestion. One clone, termed 94CY032.3, was selected for further analysis. A 694-bp fragment spanning the remainder of the LTR was amplified separately using a semi-nested approach (10).

The complete sequence of 94CY032.3 was determined by the primer-walking approach. Examination of potential coding regions revealed the expected reading frames for gag, pol, vif, vpr, tat, rev, vpu, env, and nef (data not shown). None of the genes contained major deletions, insertions, or rearrangements. However, both env and vif genes contained single in-frame stop codons. There was also a single-base-pair insertion at position 5199 which caused a frameshift and altered six amino acid residues at the C terminus of the Vpr protein. All other protein domains of known function as well as major regulatory sequences, including the primer binding site, the packaging signal, and major splice sites, appeared to be intact. Similarly, the number, position, and consensus sequences of promoter and enhancer elements in the 94CY032.3 LTR were indistinguishable from those of most other HIV-1 strains, except for the presence of an unusual TATA sequence (TAAAA), thus far only found in subtype E (A/E) viruses from Thailand and the Central African Republic (4, 10).

To compare 94CY032.3 to previously reported subtype I sequences, we constructed a phylogenetic tree from C2-V3 sequences, including representatives of all 10 known group M subtypes (Fig. 1). As expected, 94CY032.3 clustered most closely with CYH0321 and CYH0322, sequences amplified from uncultured PBMC DNA of the same individual (H032) from whom the 94CY032 isolate was derived. 94CY032.3 also clustered very closely with CYH0311, a sequence derived from the sexual partner of HO32 (14), strongly suggesting that the two infections were epidemiologically linked. Finally, as observed in the past (14), all subtype I sequences clustered independently, forming a distinct lineage roughly equidistant from all other subtypes, including subtype J (15). These findings thus confirmed the authenticity of the 94CY032.3 clone and validated it as a representative of subtype I in the C2-V3 region of the viral envelope.

FIG. 1. Confirmation of 94CY032.3 (highlighted) as a subtype I representative in the C2-V3 envelope region. Reference sequences for all known group M subtypes were obtained from the Los Alamos sequence database (15), aligned using CLUSTAL W (26), and adjusted manually using the alignment editor MASE (7). Sites for which there was a gap in any of the sequences were excluded from all sequence comparisons. Pairwise evolutionary distances were estimated by using Kimura's two-parameter method to correct for superimposed substitutions (12). Phylogenetic trees were constructed by the neighbor-joining method (20), and the reliability of topologies was estimated by performing bootstrap analysis using 1,000 replicates (6). Values at nodes indicate the percentage of bootstrap replicates in which the cluster to the right was found (bootstrap values of 85% and higher are shown). Branch lengths are drawn to scale.
FIG. 2. Exploratory tree analysis. Neighbor-joining trees were constructed for a 400-bp window moved in increments of 10 bp along the multiple genome alignment. Trees depicting discordant branching orders for 94CY032.3 (highlighted) are shown. The position of each tree in the alignment is indicated; subtypes are identified by brackets. Numbers at nodes indicate the percentage of bootstrap values with which the adjacent cluster is supported (only values above 80% are shown). Branch lengths are drawn to scale.

To characterize the remainder of the 94CY032.3 genome, we next performed pairwise sequence comparisons with recently reported nonmosaic reference sequences for subtypes A through H (9, 16) as well as selected intersubtype recombinants (17). We have used this approach in the past to screen newly derived sequences for regions of unusual sequence similarity or dissimilarity that might indicate recombination (9, 10). Briefly, 94CY032.3 was added to a multiple genome align-
ment (available upon request) which included a total of 28 sequences from the database (13), representing subtypes A (U455 and 92UG037.1), B (LA1, RF, OYI, MN, and SP2), C (C2220 and 92BR025.8), D (NDK, Z223, ELI, 84ZR085.1, and 94UG114.1), F (93BR020.1), and H (90CF056.1) as well as A/C (ZAM184 and 92RW009.6), A/G (92NG083.2, 92NG003.1, Z321, and IBNG), A/D (MAL), A/E (93TH253.3, CM240, and 90CF402.1), and B/F (93BR020.1) recombinants. Simian immunodeficiency virus SIVcpzGAB was included in the alignment to provide an outgroup. AU sites with a gap in any of the comparisons were made across the same sites. The percent identity involving sequences from subtypes A or G segments along the 94CY032.3 genome (Fig. 3). Bearing in mind the window size of 400 nucleotides and considering only peaks of significant bootstrap values (>80%), we identi-
FIG. 4. Recombination breakpoint analysis of 94CY032.3 in the vif/vpr region. (A) Diversity plots comparing 94CY032.3 to reference sequences from the database (U455, MN, and ELI) as well as to a recently identified (partial) nonrecombinant subtype G sequence (95ML045) (2). Sequences are color coded. Distance values were calculated for a window of 300 bp moved in steps of 10 nucleotides. The x axis indicates the nucleotide positions along the alignment (gaps were stripped and removed from the alignment). The positions of the start codons of the vif, vpr, and tat genes are shown. The y axis denotes the distance between the viruses compared (0.05 = 5% difference). (B) Neighbor-joining trees depicting the position of 94CY032.3 in regions flanking the breakpoints identified in panel A. Trees were constructed from the genomic regions indicated. Subtypes are identified by brackets. Four sequences from Mali represent subtype G; these are the only available subtype G sequences in this region, since all other characterized subtype G viruses contain A fragments (2, 9, 16, 17). Numbers at nodes indicate the percentage of bootstrap values with which the adjacent cluster is supported; only values above 80% are shown. Branch lengths are drawn to scale.

fied two A/G crossovers around positions 1200 and 5600 and one G/A crossover around position 4100. The bootstrap plots also outlined regions with no peaks or peaks below 80%, which coincided with segments that clustered independently in the exploratory tree analysis and thus were likely in subtype I. Delineating the boundaries of these regions suggested five additional breakpoint positions: G/I at 1500, I/G at 3800, G/I at 6000, I/A at 6900, and A/I at 7200. Because full-length, nonmosaic reference sequences for the parental lineages (G and I) were not available, most of the breakpoints could not be

FIG. 5. Inferred structure of the mosaic 94CY032.3 genome. Regions of different subtype origin are color coded. All breakpoints are tentative, since full-length nonrecombinant subtype G and I reference sequences are not available. LTR sequences were not analyzed and are thus shown as hatched boxes.
mapped with certainty. However, the A/G breakpoints at positions 1200 and 5000 were confirmed by informative site analysis (data not shown). The recombinant nature of 92NG083.2 prohibited reliable breakpoint analysis between positions 4200 and 4800 (9, 16) (highlighted in Fig. 3).

To map the recombination breakpoints in this remaining region, we used four recently reported, partial but nonmosaic subtype G sequences from Mali which spanned the vif/vpr region and thus bridged the subtype A gap of 92NG083.2 (2). Figure 4A illustrates a set of distance plots that compare 94CY032.3 to one of these newly derived G sequences (95ML045) as well as representatives of subtype A (U455), B (MN), and D (ELI), respectively. Consistent with the results from the exploratory tree analysis (Fig. 2), 94CY032.3 was disproportionately more closely related to U455 in the 5' and 3' thirds of this fragment, suggesting the presence of subtype A-like segments. However, in the middle of the fragment, 94CY032.3 was clearly equidistant from U455 and the other subtypes, suggesting an independent position. Thus, noting the points at which the "A" distance increased and decreased relative to the other distances allowed us to map the two remaining breakpoints, one near position 4650 and the other near position 5000. Trees constructed from sequences surrounding these two breakpoints (Fig. 4B) confirmed that 94CY032.3 switched position from subtype A (panel 4255-4650) to subtype I (4651-5000) and back to subtype A (5001-

FIG. 6. Screening of intersubtype recombinants containing regions of uncertain subtype assignment for the presence of subtype I sequences. (A) Schematic representation of the recombinant genomes of MAL and Z321. Fragments of subtypes A, D, and G origin are color coded red, blue, and green, respectively; segments of uncertain subtype assignment are shown in white. Arrows denote regions of unusual sequence similarity with 94CY032.3 as determined by diversity plot analysis (data not shown). (B and C) Phylogenetic analysis of the putative subtype I domains of MAL and Z321 delineated in panel A. Neighbor-joining trees of partial pol (A) and vif/vpr (B) sequences. Horizontal branch lengths are drawn to scale; the scale bar represents 0.1 nucleotide substitutions per site. Vertical separation is for clarity only. Values at the nodes indicate the percent bootstraps in which the cluster to the right was supported; bootstrap values of 80% and higher are shown. Brackets at the right represent the major sequence subtypes of HIV-1 group M. Subtype I clusters are highlighted; trees were midpoint rooted.
Figure 5 summarizes the results from all phylogenetic analyses, depicting a schematic representation of the mosaic genome structure of 94CY032.3. Segments of different subtype origin are color coded, and there are a total of 10 recombination breakpoints between positions 4255 and 5300.

Having identified several fragments of subtype I in 94CY032.3, we next determined whether there was any evidence for its mosaic structure of 94CY032.3. Segments of different subtype origin are color coded, and there are a total of 10 recombination events, either in the same or different genomes. Two known mosaics, MAL (1, 19) and 2321 (17, 19), are represented subtype I, we again performed distance plot analysis, presented subtype I domains of 94CY032.3. These findings support the recent molecular characterization of a virus isolated from a serum sample collected in 1976 in Thailand (2321) as an indication of relatively greater sequence variability and geographic distribution of subtype I-containing viruses.

Nucleotide sequence accession numbers. The complete sequence of 94CY032.3 has been deposited at GenBank under accession no. AF049337 and AF049338.

We thank F. E. McCutchan for unpublished sequence information, Y.-L. Chen and B. Jian for excellent technical assistance, and W. L. Allsbro for artwork and preparation of the manuscript.

This work was supported by grants from the National Institutes of Health (ROI AI 35170 and ROI AI 25291), by shared facilities of the UAB Center for AIDS Research (DNA Sequencing Core; P30 AI27760), and by the Birmingham Veterans Administration Medical Center.

REFERENCES


sequences. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, N.M.


