Complex Mosaic Composition of Near Full-Length Genomes of Two NED (NIH-ENVA-DOD) Subtype Panel HIV Type 1 Strains, BCF-Dioum and BCF-Kita, Originating from the Democratic Republic of Congo (DRC)

Diana D. Huang,¹ Brian T. Foley,² Catlin A. Tolzmann,¹ Anastasia Ouma,¹ and James W. Bremer¹

Abstract

Sequence characterization of the near full-length genomes of HIV-1 isolates BCF-Dioum and BCF-Kita, originating from the Democratic Republic of Congo (DRC), was continued. These NED panel isolates, contributed by F. Brun-Vezinet (ENVA-France), were first identified as subtypes G and H, respectively. Our earlier analyses of portions of their *pol* genes showed that both were likely to be intersubtype recombinants of different composition. This study analyzed the remainder of each genome, confirming them to be complex recombinants. The BCF-Dioum genome resembles CRF06_cpx strains found in West Africa, composed of subtypes A/G/J/K. The BCF-Kita genome is a unique complex recombinant A-F-G-H-K-U strain. These data support previous observations of the complexity of strains originating from the DRC. BCF-Dioum may be a suitable strain for standards and reagents since it matches a defined circulating recombinant form. Studies and reagents made from BCF-Kita should take into account its complex genome.

THE SEQUENCES OF TWO HIV-1 ISOLATES, BCF-Dioum (Dioum) and BCF-Kita (Kita), contributed by F. Brun-Vezinet (Hôpital Bichat-Claude Bernard, France) to the NED (NIH-European Network for Virology Assurance-DoD, U.S. Military HIV Program) panel, were characterized further. This panel was developed in 1996 to serve as a reference panel of well-characterized but minimally passaged HIV-1 strains, with diverse geographic origins¹ and sequence diversity. The isolates were obtained, for the most part, prior to widespread use of antiretroviral therapy and should therefore be considered to be sensitive to antiretroviral therapies. The BCF-Kita isolate was from an asymptomatic individual living in France who was originally from the Democratic Republic of Congo (DRC) and was believed to be infected in the DRC before arrival in France. BCF-Dioum was isolated in France during a primary infection from a person whose HIV-positive partner was from the DRC, and believed to be infected in the DRC before arrival in France. The isolates were identified as belonging to subtypes G (BCF-Dioum) and H (BCF-Kita) by heteroduplex analysis (HMA) of an envelope (*env*) gene fragment, prior to donation to the panel. Our previous sequence analysis of the protease and partial reverse tran-

scriptase (RT) genes (1190 bases) showed that these strains contain areas suggesting F/G and A/G recombination, respectively.² Further sequencing and evaluation of the near full-length genomes now show that these strains are complex intersubtype recombinants.

A limited expansion of the isolate supernatant was performed in peripheral blood mononuclear cells (PBMCs) upon receipt to make a small master pool of each isolate. The BCF-Dioum and BCF-Kita genomes were sequenced from the viral RNA extracted from cell-free virions, released in these culture supernatants, as well as from the provirus in the PBMCs used to generate these cultures.^{1,3} The PBMCs were collected and frozen upon termination of the culture. All individual reagents for lysis and extraction of the RNA genome were provided with the ViroSeq HIV-1 Genotyping System, v.2 (Celera Diagnostics LLC, Alameda, CA) using kit recommended conditions. Extracted RNA was used in an reverse transcriptase polymerase chain reaction (RT-PCR) according to Rota *et al.*⁴ to generate cDNA and sequencing template using combinations of published and specially designed primers (primer sequences available on request). Two long overlapping products spanning the genome were used to generate

¹Department of Immunology/Microbiology, Rush Medical College, Chicago, Illinois 60612.

²HIV Sequence Database, Theoretical Biology and Biophysics Group (T-6), Los Alamos National Laboratory, Los Alamos, New Mexico 87545.

multiple smaller templates for sequencing by nested PCR. The overlap between the two fragments was ~2000 nucleotides. The sequencing templates were evaluated by agarose gel electrophoresis to observe that single, robust, predominant bands near the predicted size were produced in each PCR reaction. Sequencing was performed with templates diluted to ≤ 40 ng and in-house designed and published primers (primer sequences available on request) used with the ABI Prism Big Dye Terminator Cycle Sequencing Ready Reaction Kit v. 3.1. The protease-RT region covered by the ViroSeq kit was sequenced using the premixed primers and buffers provided in the kit. The genome was sequenced and analyzed with both forward and reverse overlapping primers using an AB 3100 instrument (Applied Biosystems, Foster City, CA).

Bases were assigned by AB programs Sequencing Analysis v. 3.4.1 and Factura v. 2.2.2. A consensus sequence for each strain was generated using Autoassembler v. 2.0.1 (AB), which assembled and aligned the chromatographs for all the forward and reverse sequence data contiguously. Parameters in the assemblage program were set at ≥ 50 bases for the minimum overlap size with a 70% similarity for the smaller sequenced regions for each piece to be added into the consensus. The editing strategy of the aligned, overlapping sequences required that each called base show the same unambiguous or mixed peak composition in both the forward and reverse direction for inclusion into the overall consensus sequence. Since the sequences were generated by population sequencing, no attempt to resolve mixed bases was made if both bases were represented as comigrating peaks in both directions. Each new major region that was sequenced had to include an unambiguous overlap into an existing region of the current consensus sequence. The final consensus sequences for BCF-Dioum and BCF-Kita were then aligned with the Sequence Locator Tool in the Los Alamos HIV database (<http://hiv.lanl.gov>) to map the genome regions and corresponding proteins against the reference strain HXB2. To generate phylogenetic trees, the consensus genome sequence of either BCF-Dioum or BCF-Kita was added to alignments composed of randomly chosen HIV-1 nonrecombinant reference sequences from the Los Alamos HIV database, including group O sequences. For the BCF-Dioum tree, sequences for seven CRF06_cpx strains were added; for the BCF-Kita tree, two G-K-U recombinant strains⁵ were added. The alignments made within the module were downloaded in FASTA format, checked for quality of alignment, and gapstripped for analysis. These alignments were also examined in SimPlot version 3.5.1.^{6,7} The similarity (SimPlot) and bootscanning (Bootscan) plots were made using the F84 model with a window size of 400 bases.

The similarity plot was the basis used to divide the BCF-Kita alignment without the G-K-U recombinants into regions separated by visible intersubtype recombination breakpoints. The alignment for each demarcated region was saved and converted into a format for PHYLIP v. 3.6.6., then processed first in the DNADIST followed by the NEIGHBOR module,⁸ outgrouping against a group O strain to visualize phylogenetic relationships. Trees were constructed in Treeview⁹ v. 1.6.6. Bootstrap values for these same alignments were determined using PHYLIP modules in the following order: SEQBOOT, DNADIST, NEIGHBOR, and CONSENSE. The National Center for Biotechnology Information (NCBI) Geno-

typing tool (<http://www.ncbi.nlm.nih.gov>)¹⁰ was also used to see if the consensus sequence of the strains matched any other recognized circulating recombinant strain (CRF) by BLAST analysis.

BCF-Dioum: Our sequence for BCF-Dioum is 8689 bases long and maps to HXB2 bases 839–9481, starting from HXB2 base 50 in *gag* to base 396 in the 3' LTR gene. When analyzed using the NCBI genotyping tool against the 2005 pure-type and recombinant list, the sequence aligns with CRF06_cpx reference strains, which are composed of regions of subtypes A, G, J, and K^{11–13} (data not shown). Figure 1A shows BCF-Dioum clustered with reference CRF06_cpx strains in a phylogenetic tree. The *env* gene is composed predominantly of subtype G sequence, characteristic of this clade. The composition of the *env* gene would support the strain being identified as subtype G using HMA. Breakpoints in major gene regions appear to be present in the last half to one-third of p24, the RT gene in at least two positions, integrase and gp120 as previously described for CRF06_cpx.

BCF-Kita: The 7721 base sequence analyzed for BCF-Kita maps to HXB2 bases 612–8350, spanning from base 612 in the 5' LTR to base 2126 of the *env* gene. It covers up to the first 198 amino acids of gp41. Analysis with the NCBI genotyping tool showed no clear association with any single identified subtype or circulating recombinant form (data not shown). Phylogenetic analysis of the entire sequence shows the BCF-Kita did not map with any specific subtype. However, the strain did appear on the same node as two G-K-U recombinant strains, SE 8646 and SE9019,⁵ both isolated in 1995 (Fig. 1B). These genomes also represent complex, but as yet unclassified recombinants originating from Uganda and the DRC, respectively. They also show sizable regions that suggest some relationship with subtypes A1 and H.⁵

A more detailed phylogenetic evaluation of the subtype composition of the BCF-Kita genome was made by splitting the genome alignment into smaller segments of variable size between visible breakpoints, as described above. The portions of the trees containing BCF-Kita made from these alignments are shown in Fig. 2A. Cumulatively, these data indicate that the strain is distinctly a group M intersubtype recombinant and is composed of regions with some similarity to subtypes A–F–G–H–K and some regions which are not significantly related to any one defined subtype and are classified here as U (Untypeable). The two U regions, between bases 4253–4502 in integrase and 7204–7406 near the 3' end of gp120, may be a mix of two or more subtypes that are not clearly distinguishable or derived from an individual subtype that has not yet been described. The map of BCF-Kita generated from these trees using the recombinant HIV-1 drawing tool (<http://hiv.lanl.gov>) is shown in Fig. 2B. The major portion of the *env* gene is mostly like subtype H, which would support the original observation placing this strain into this group by HMA analysis. Breakpoints in major gene regions appear in the p24 gene, protease, RT, which has two clear crossover points, RNase H, integrase, *vpr*, and gp120. The *pol* gene contained breakpoints similar, but not identical to, patterns mapped for several other designated CRF_cpx strains originating from Central and West Africa and Cypress and Greece (<http://hiv.lanl.gov>). Bootscan plots showed similar crossover points (data not shown).

Our data indicate that these two NED panel viruses from the DRC are complex recombinants, consistent with previous

FIG. 1. Phylogenetic trees made with the consensus sequence of near-full length genome. Sequence data of BCF-Dioum (A) and BCF-Kita (B) were used to form trees with reference sequences from the Los Alamos HIV Sequence Database. As described in the text, BCF-Dioum was aligned with randomly selected Los Alamos HIV-1 nonrecombinant reference strains, including group O and seven reference CRF06_cpx strains. Two untyped complex recombinant strains similar in make-up to BCF-Kita were added to the BCF-Kita alignment. The trees were plotted using PHYLIP,^{17,18} which can also be accessed by linking through the Los Alamos HIV database (<http://hiv.lanl.gov>). The roots containing the O strains are not shown in the figure. The positions of BCF-Dioum and BCF-Kita are in a larger font. The subtype or CRF designation of the reference strains in the alignments are represented by the first letter or numbers of the name of the strain.

observations of complex recombinant and untypeable HIV-1 strains originating from the DRC.^{14,15} Vidal *et al.*¹⁴ evaluated 247 samples from the DRC to determine the distribution of subtypes in that country. The subtypes were identified through heteroduplex mapping and sequencing of the *env* and *gag* portions of the genome. Although subtype A was predominant, all other subtypes except subtype B were prevalent in the region, including several complex recombinants. Notably, 29% of the samples showed discordance between *env* and *gag* analysis, reinforcing the high probable incidence of recombinants in the circulating populations in the DRC. Yang *et al.*¹⁵ examined 24 samples collected in 1985 and 83 samples obtained between 1999 and 2000 evaluating the p24 region of *gag* and the C2V3 and gp41 regions of *env*. Their

analysis also indicated extensive subtype mixing occurring during both time periods. Because it has been suggested that areas in central Africa such as the DRC may be the epicenter of HIV infections,¹⁴ the large representation of different subtypes within the genome of BCF-Kita is not unexpected.

It is highly likely that BCF-Kita is related to another reported isolate FR.BCB79, also from Hôpital Bichat-Claude Bernard,¹⁶ potentially from the same infected individual. Duchet *et al.*¹⁶ reported *gag* (accession number Y13196) and *env* (accession number Y13197) sequences for this isolate, showing *gag* to have an A/G subtype composition and *env* to be subtype H, making the strain an A/G/H recombinant. The *gag* and *env* sequences show 97% and 94% DNA sequence identity, respectively, to corresponding areas of the BCF-Kita

FIG. 2. Subtype assignment of regions of BCF-Kita between crossover points in the genome. **(A)** A panel of phylogenetic trees generated from portions of the same alignment for BCF-Kita used to make the tree presented in Fig. 1B. As described in the text, the alignment was analyzed by the similarity plot in SimPlot v.3.5.1. Breakpoints were marked visually within the plot. The alignments corresponding to regions between the crossover points were evaluated further in PHYLIP modules DNADIST and NEIGHBOR. The trees were produced using Treeview. Boot strap values were determined using PHYLIP modules in the following order: SEQBOOT, DNADIST, NEIGHBOR, and CONSENSE. The portions of the total alignment with tree branches containing BCF-Kita are shown. The subtype of the reference strains in the alignments are represented by the first letter of the name of the strain. The regions represented in each alignment portion are designated by HXB2 base pair (bp) numbering at the bottom of each small panel. The position of BCF-Kita in each panel is marked with an arrow. The subtype of group M that best matches the region containing BCF-Kita is shown to the right of each panel. Subtype U designates an untyped region. Only the branch point going into the BCF_Kita node with a bootstrap value >50 is labeled in the trees in blue. Trees with nodes having bootstrap values <50 are not marked. **(B)** The map of BCF-Kita showing the subtype composition as represented in **(A)** was generated with the recombinant HIV-1 drawing tool in the Los Alamos HIV-1 sequence database (<http://hiv.lanl.gov>). The breakpoints are mapped along the genome according to HXB2 numbering. The functional regions of the *gag*, *pol*, and *env* genes are added to the map to show subtype compositions the protein coding region have.

genome. It is also of interest that BCF-Kita branches on the same node as the untyped G-K-U isolates SE8646 and SE9010 (Fig. 1B), displaying, in alignment, overall homologies of 91% and 90%, respectively. Because many more tools as well as reference subtype and CRF strains are now available for sequence analysis, more comprehensive parsing of the subtype source of HIV-1 isolates can be made, especially with a strain such as BCF-Kita, to interpret genomic ancestry in greater detail. Such analyses for BCF-Kita are ongoing.

The complex make-up of the both BCF-Dioum and BCF-Kita was not anticipated at the time that the NED panel was

formed. The strains were contributed to the panel as representatives of subtype G and H, respectively, which was appropriate based on the assay used to determine their subtype. Because the strains were passaged in culture, it is highly likely that any diversity that might have been noted in the original plasma was reduced or lost. This is demonstrated by the low level of ambiguity that appeared during sequencing. Their genomic profiles provide an example that considerations about the entire genome should be made when using these strains or when collecting strains for other subtype "reference" panels for activities such as assay validation or devel-

opment. This is reinforced by the observation that the areas in which breakpoints were noted for these strains, such as p24, RT, and integrase, are targets for diagnostic use and/or therapy. It may be that different levels of characterization of sets of new isolates would need to be performed based on anticipated use of the samples. Studies, assays, and reagents made using BCF-Dioum and BCF-Kita should take into account the complexity noted in their genomes.

The accession number for BCF-Dioum is FJ183725 and for BCF-Kita is FJ183726.

Acknowledgments

This work was supported by NIAID contract NO1-AI-85354 and HHSN266200500044C/NO1-AI-50044.

Disclosure Statement

No competing financial interests exist.

References

- Huang D, Polstra D, Giesler T, Scianna S, Bremer J, Robb M, Michael N, Brun-Vezinet F, and Reichelderfer P: Characterization of an international HIV-1 subtype panel for future use as standards and reference strains. Presented at the 6th Conference on Retroviruses and Opportunistic Infections, Chicago, IL, January 31–February 4, 1999.
- Huang DD, Giesler TA, and Bremer JW: Sequence characterization of the protease and partial reverse transcriptase proteins of the NED panel, an international HIV type 1 subtype reference and standards panel. *AIDS Res Hum Retroviruses* 2003;19(4):321–328.
- Division of AIDS, National Institute of Allergy and Infectious Diseases: *Virology Manual for HIV Laboratories*, NIH Publication No. 97–3828, National Institutes of Health, Bethesda, MD, 1997.
- Rota PA, Oberste MS, Monroe SS, *et al.*: Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* 2003;300(5624):1394–1399.
- Wilbe K, Salminen M, Laukkanen T, *et al.*: Characterization of novel recombinant HIV-1 genomes using the branching index. *Virology* 2003;316(1):116–125.
- Lole KS, Bollinger RC, Paranjape RS, *et al.*: Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J Virol* 1999;73(1):152–160.
- Salminen MO, Carr JK, Burke DS, and McCutchan FE: Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS Res Hum Retroviruses* 1995;11(11):1423–1425.
- PHYLP (Phylogeny Inference Package) version 3.572c [computer program]. Department of Genetics, University of Washington; Seattle, WA, 1996.
- Page RD: TreeView: An application to display phylogenetic trees on personal computers. *Comput Appl Biosci* 1996; 12(4):357–358.
- Rozanov M, Plikat U, Chappey C, Kochergin A, and Tatusova T: A web-based genotyping resource for viral sequences. *Nucleic Acids Res* 2004;32(Web Server issue):W654–659.
- Oelrichs RB, Workman C, Laukkanen T, McCutchan FE, and Deacon NJ: A novel subtype A/G/J recombinant full-length HIV type 1 genome from Burkina Faso. *AIDS Res Hum Retroviruses* 1998;14(16):1495–1500.
- Montavon C, Toure-Kane C, Nkengasong JN, *et al.*: CRF06-cpx: A new circulating recombinant form of HIV-1 in West Africa involving subtypes A, G, K, and J. *J Acquir Immune Defic Syndr* 2002;29(5):522–530.
- Triques K, Bourgeois A, Vidal N, *et al.*: Near-full-length genome sequencing of divergent African HIV type 1 subtype F viruses leads to the identification of a new HIV type 1 subtype designated K. *AIDS Res Hum Retroviruses* 2000;16(2):139–151.
- Vidal N, Peeters M, Mulanga-Kabeya C, *et al.*: Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. *J Virol* 2000;74(22):10498–10507.
- Yang C, Li M, Mokili JL, *et al.*: Genetic diversification and recombination of HIV type 1 group M in Kinshasa, Democratic Republic of Congo. *AIDS Res Hum Retroviruses* 2005;21(7):661–666.
- Duchet S, Letourneur F, Loussert-Ajaka I, *et al.*: gag and env sequences of an A/G/H recombinant from a Zairian HIV type 1 isolate. *AIDS Res Hum Retroviruses* 1997;13(15): 1351–1354.
- Guindon S and Gascuel O: A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 2003;52(5):696–704.
- Guindon S, Lethiec F, Duroux P, and Gascuel O: PHYL Online—a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res* 2005;33(Web Server issue):W557–559.

Address correspondence to:

Diana D. Huang
Department of Immunology/Microbiology
Rush Medical College
1653 W. Congress Parkway
Chicago, Illinois 60612

E-mail: diana_huang@rush.edu