JVI Accepted Manuscript Posted Online 11 January 2017
J. Virol. doi:10.1128/JVI.01841-16

Sensitive next generation sequencing method reveals deep genetic diversity of HIV-1 in the Democratic Republic of the Congo

Running title: HIV-1 diversity in new sequences from the Congo Basin

Rodgers MA*[1], Wilkerson E[2,5], Vallari A[1], McArthur C[3], Sthreshley L[4], Brennan CA[1], Cloherty G[1], de Oliveira T[2,5,6]

1. Infectious Disease Research, Abbott Diagnostics, Abbott Park, IL, USA
2. Wellcome Trust – Africa Centre for Population Health, University of KwaZulu-Natal, Durban, 4041, Republic of South Africa
3. School of Dentistry, University of Missouri-Kansas City, Kansas City, MO, USA
4. Presbyterian Church (USA), Kinshasa, DRC
5. College of Health Sciences, University of KwaZulu-Natal, Durban 4041, Republic of South Africa
6. Research Department of Infection, University College London, London WC1E 6BT, United Kingdom

*Corresponding Author: Mary A Rodgers, mary.rodgers@abbott.com
100 Abbott Park Rd
Abbott Park, IL 60064
Ph 224-668-8936
Fax 224-667-1401

Abstract

As the epidemiological epicentre of the human immunodeficiency virus (HIV) pandemic, the Democratic Republic of the Congo (DRC) is a reservoir of circulating HIV strains exhibiting high levels of diversity and recombination. In this study, we characterized HIV specimens collected in two rural areas of the DRC between 2001 and 2003 to identify rare strains of HIV. The *env* gp41 region was sequenced and characterized for 172 HIV-positive specimens. The *env* sequences were predominantly subtype A (43.02%), but 7 other subtypes (33.14%), 20 circulating recombinant forms (CRFs: 11.63%), and 20 unclassified (11.63%) sequences were also found. Of the rare and unclassified subtypes, 18 specimens were selected for next generation sequencing (NGS) by a modified HIV-SMART method to obtain full genome sequences. NGS produced 14 new complete genomes, which included pure subtypes C (n=2), D (n=1), F1 (n=1), H (n=3), and J (n=1). The two Cs and one of the H genomes branched basal to their respective subtype branches but had no evidence of recombination. The remaining 6 genomes were complex recombinants of 2 or more subtypes, including A1, F, G, H, J, K, and unclassified fragments, including one CRF25 isolate, which branched basal to all CRF25 references. Notably, all recombinant H fragments branched basal to the H clade. Spatial-geographical analysis indicated that the diverse sequences identified here did not expand globally. The full- and sub-genomic sequences identified in our study population significantly increase the documented diversity of the continually evolving HIV-1 pandemic.

Importance (150 word limit, nontechnical): Very little is known about the ancestral HIV-1 strains that founded the global pandemic, and very few complete genome sequences are available from patients in the Congo Basin where HIV-1 expanded early in the global pandemic. By sequencing a sub-genomic fragment of the HIV-1 envelope from study participants in the DRC, we identified rare variants for complete genome sequencing. The basal branching of some of the complete genome sequences we recovered suggests that these strains are more closely related to ancestral HIV-1 sequences than to previously reported strains and is evidence that the local diversification of HIV in the DRC continues to outpace the diversity of global strains decades after the emergence of the pandemic.

Key words: full-length genome, HIV-1 surveillance, next generation sequencing, phylogenetic analysis, recombination, genetic diversity

54

55 **Introduction**

56 Multiple independent interspecies transmission events of simian immunodeficiency virus (SIVs)
57 have resulted in the emergence of four major lineages of HIV-1 in humans: groups N, O, P, and
58 the pandemic group M(1). Estimates place the emergence of the group M linage of HIV-1 in the
59 Congo Basin at the beginning of the 20th century(2, 3). For the purpose of this study, the Congo
60 Basin includes the following countries: Angola, Cameroon, the Central African Republic, the
61 Democratic Republic of the Congo, the Republic of the Congo and Gabon. By 1959-1960,
62 considerable HIV-1 diversity was already present in Kinshasa, DRC(4), where HIV-1 group M
63 first emerged and then spread globally(5). The current nomenclature recognizes 9 major
64 subtypes of HIV-1 group M (subtypes A-D, F-H, J and K) and is entirely based on a
65 phylogenetically based classification system. The most prevalent subtypes are A, B, C, D, and
66 G, while subtypes F, H, J, and K collectively comprise only 1% of all infections worldwide(6).
67 Subtypes H, J, and K are primarily found in West, South and Central Africa and only 2-7
68 complete genomes have been reported, making them extremely rare(6-8). A subtype L was
69 suggested as a new classification based on two distinct HIV-1 genomes collected in the DRC in
70 1983 and 1990; however, a third epidemiologically unlinked case has not been reported(9, 10).
71 Many HIV-1 isolates from the Congo Basin, including the two putative subtype L genomes, do
72 not cluster phylogenetically with other known sequences and are considered “unclassified”(9-
73 13).

74 Another important genetic feature of HIV is that it is prone to recombination. High levels of intra-
75 subtype diversity and inter-subtype recombination are found in DRC HIV-1 patient specimens(5,
76 11), which are indicative of an old epidemic. Currently, there are 72 circulating recombinant
77 forms (CRFs) that have each been identified in at least three unlinked HIV-1 individuals while
78 many more unique recombinant forms (URFs) have been described in 1-2 individuals(7).
79 Recent analysis of whole genome HIV-1 sequences from the Congo Basin identified fragments
80 that clustered basal to all major subtypes, suggesting that the parental lineages of these
81 recombinant fragments have not yet been sampled and characterized, or that these strains have
82 gone extinct and are no longer in circulation(12).

83 The diversity present in HIV specimens from the DRC and other countries within the Congo
84 Basin is a unique source for identifying rare and emerging variants; however, classification of

these viruses is likewise complicated by the extreme genetic diversity observed in HIV variants within the region. Subtype-specific differences in treatment effectiveness, the development of resistance, vaccine coverage, and disease progression make surveillance and accurate classification of HIV-1 strains imperative(14-16). To date, most HIV-1 specimens from the DRC have primarily been classified by phylogenetic analysis of short partial genome sequences (400 – 900 bp) in either the *group-specific antigen (gag)*, *polymerase (pol)*, or *envelope (env)* genes of HIV-1(4, 11, 13, 17). However, the full extent of recombination and sequence diversity of a complex genome cannot be completely characterized when partial genome sequences are used for HIV-1 classification. As a region with exceptionally high HIV-1 diversity, this is especially true for specimens from the Congo Basin. Advances in availability and reduction in cost for next generation sequencing (NGS) technologies have enabled complete genome sequences to be used for classification of HIV-1 specimens globally. Despite these advances, only 33 complete HIV-1 genomes are currently available from the DRC in the Los Alamos National Laboratories (LANL) repository(18). In contrast, 1217 complete HIV-1 genomes are available from the United States in the LANL database, 1185 of which are classified as subtype B(18). Therefore, in order to fully characterize the true diversity of circulating HIV-1 strains, additional complete genomes of complex variants from the DRC must be sequenced.

As a part of ongoing surveillance efforts in Sub-Saharan Africa, we have previously deposited 55 complete HIV-1 genomes from cultured virus isolates and Cameroonian patient specimens into the Genbank database(19-27). Recently, we developed a new HIV-primer specific NGS platform to obtain complete genomes from HIV-1 Group M, N, O, and P as well as HIV-2 isolates, called HIV-SMART(19). This method utilizes a set of 6 pan-HIV specific primers fused to the SMART (**S**witching **M**echanism **a**t 5' End of **R**NA **T**emplate, Clontech) sequence to create libraries for NGS on the Illumina MiSeq instrument(19). While this method is excellent for recovering genomes from diverse HIV sequences, it has limited success for low viral load clinical specimens (<5 log copies/ml). HIV-SMART was previously optimized for clinical specimens by adding a benzonase pre-treatment of the sample to digest background human DNA and RNA, and a direct correlation was observed between the sample viral load and resulting genome sequence coverage(19). Increasing the total number or concentration of HIV-SMART primers did not have any benefit to genome coverage, however an increase in the reverse transcription (RT) temperature from 42°C to 47°C improved coverage depth while RT at 52°C dramatically reduced genome coverage. Alternative library preparation methods or an RT

temperature between 47°C and 52°C may improve genome coverage and depth for low viral load clinical samples, although these conditions have not been tested yet.

In the present study, we have applied both Sanger-based *env* amplicon sequencing and HIV-SMART NGS to a set of HIV-1 variants from the DRC, resulting in 172 new *env* sequences and 14 new complete genomes. Modifications to the HIV-SMART method for low viral load samples improved genome coverage to >90% for clinical samples with viral load >4 log copies/ml by adding an input nucleic acid concentration step and lowering the RT temperature to 42°C. Characterization of the complex *env* and complete genome sequences revealed a high level of HIV-1 diversity and recombination in the DRC, and identified sequences that are outliers to known subtype and CRF sequences. Further analyses of the outlier sequences suggest that they may very well be ancestral to some of the major pandemic subtypes today.

Materials and Methods

Specimens Plasma specimens were collected at the Vanga Hospital, Bandundu Province and The Good Shepard Hospital located 12 kilometres from Kananga, Kasia-Occidental Province in the DRC between 2001 and 2003. The specimens came from voluntary testing participants and pregnant women participating in a prevention of Mother To Child Transmission (pMTCT) program. Samples were acquired according to the 98-041e protocol approved by the University of Missouri Kansas City Research Board. We used a progressive analytical approach to test specimens and to identify rare viral subtypes circulating within our study population. A schematic breakdown of this analytical approach is illustrated in the flow diagram in **Figure 1**.

Serology Briefly, specimens were initially tested using the ARCHITECT HIV Ag/Ab Combo assay (Abbott Diagnostics, Abbott Park, IL) in order to identify HIV infected specimens. The viral load of selected reactive specimens was quantified by the Abbott RealTi*m*e HIV-1 assay (Abbott Molecular, Des Plaines, IL) according to the manufacturer's instructions. HIV reactive specimens were serotyped with a research use only peptide immunoassay (PEIA) in order to classify specimens based on HIV type (HIV-1 or -2) and group (M, N, O or P). Synthetic peptides to the *env* IDR of gp41 and *env* gp120 V3 from HIV-1 groups M, O, N and P, HIV-2, and two strains of SIV CPZ (chimpanzee) and SIV RCM (Red Capped Mangabey) were covalently coupled separately to Luminex MagPlex beads before dilution into buffer (1% BSA in PBS). In each well of a round bottom polystyrene white 96 well plate (Costar), 50 ul of bead mixtures and 50 ul of sample were incubated for 30 minutes at room temperature in a plate

shaker at 300 rpm. Liquid was aspirated using a BioTek 405 TS Magnetic Plate washer, and wells were washed with ~250 ul of PBS-Tween20 wash buffer (BioTek, Shoreline, WA). Plates were incubated with 50 ul of 0.4 ug/ml biotinylated goat anti-human IgG (Sigma, St Louis, MO) for 15 minutes at room temperature in a plate shaker. After washing, plates were incubated with 50 ul of 0.4 ug/ml Streptavidin, R-phycoerythrin conjugate, SAPE (Invitrogen, Carlsbad, CA) for 10 min at room temperature in a plate shaker. After final washes, beads were re-suspended in 150 ul of reading buffer (1% BSA in PBS) and analysed on the Luminex FlexMap3D instrument (Luminex Corp. Austin, TX). For each bead set, ~100 events were read and results were expressed as Median Fluorescence Intensity (MFI) per 100 beads.

RNA extraction Following the serological classification, nucleic acid was extracted from samples according to the manufacturer's instructions using either: (*i*) the QIAcube blood and body fluid spin protocol (QIAgen) or (*ii*) the (research use only) total nucleic acid sample preparation protocol on the *m*2000*sp* system (Abbott Molecular, Des Plaines, IL). For NGS experiments, plasma was pre-treated with benzonase before nucleic acid extraction. One-tenth volume of 10X benzonase buffer (200 mM Tris-Cl pH 7.5, 100 mM NaCl, 20 mM $MgCl_2$) and 250 units/ml of ultra-pure benzonase (Sigma, St Louis MO) were added to 0.9 volumes of plasma to degrade free DNA and RNA(28, 29). Samples were incubated at 37°C for 3 hours then filtered by centrifugation (5000 rpm) through 0.22 µM spin filters (Millipore, Billerica, MA) before extraction. For low viral load samples post-extraction concentration of 25-50 µl of eluted nucleic acid was performed by using a concentrator column kit (Zymo Research) following the manufacturer's instructions.

Sub-genomic sequencing Reverse Transcription Polymerase Chain Reaction (RT-PCR) and Sanger sequencing were used to genotype a 676 base pair (bp) fragment of the *env* gene (immunodominant region of gp41). The detailed RT-PCR and sequencing procedures have previously been described(30). If the *env* RT-PCR failed, alternative primers in *env* were used. For rare subtypes, RT-PCR was also performed for *gag* p24 (468 bp) and *pol* integrase (864 bp). These sub-genomic sequences have been deposited in Genbank under accession numbers KY365010 to KY365181 (*env*); KY365182 to KY365202 (*gag*); and KY365203 to KY365216 (*pol*).

Sub-genomic sequence classification Sub-genomic sequences were phylogenetically subtyped by analysing genotypes against a comprehensive reference data set. This data set includes 480 whole genome references that were used by Tongo and colleagues(12), to

investigate the deep genetic diversity of HIV-1 group M. Five additional whole genome sequences were added to the alignment including three whole genome subtype J and H sequences from Angola(8) and two additional "unclassified" whole genome sequence (KP718920 & KP718929). Briefly, this reference data set includes: (*i*) representative whole genome group M HIV-1 sequences of major subtypes (A-K) and sub-subtypes (i.e. A1/A2 and F1/F2), (*ii*) representative whole genome group M HIV-1 sequences of CRFs 01-72, (*iii*) all whole genome HIV-1 sequences of what the LANL database classifies as "problematic" sequences, (*iv*) and all whole genome HIV-1 sequences that are listed as "unclassified" within the LANL database, including the two whole genome sequences belonging to the putative "subtype L" (last date of access 1st May 2016).

Sub-genomic sequences were aligned against homologous segments of the 485 references in ClustalW. Alignments were manually edited in Geneious 8.0.5 until a perfect codon alignment was achieved for each data set. A Maximum Likelihood (ML) tree topology was inferred for each of the alignments in RAxML v 8.0.0(31) under the General Time Reversible model of nucleotide substitution(32) and estimated gamma shape parameter(33). The fast parametric bootstrap resampling method (n=1,000) was implemented on a 12-core MacPro to infer support for splits. Each tree topology was visualized in FigTree v 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree) and manually annotated.

Genotypes clustering within a subtype or CRF clade and with bootstrap values >70 were classified as belonging to that particular clade. Sequences that were basal to a particular clade were analysed for recombination in Simplot(34) and unique recombinants were classified as URFs. Sequences with different classification in *gag*, *pol*, and *env* were also classified as URF. Sequences that did not branch with any references or which clustered with "unclassified" references were considered "unclassified".

To make an easily visualized tree for Figure 2, a neighbor-joining phylogenetic tree and bootstrap values were inferred using the Phylip 3.5c software package(35).

HIV-SMART NGS Eighteen isolates with remaining volume belonging to rare viral subtypes or isolates that exhibited signs of inter subtype recombination were targeted for whole genome sequencing by the HIV-SMART NGS method. HIV-SMART libraries were prepared, sequenced, and analysed as previously described(19). Briefly, the six-primer HIV-SMART mix was used and reverse transcription reactions were performed at 42°C. HIV-SMART optimization experiments

compared different reverse transcription temperatures of 47°C and 50°C, as well as the Pico SMART cDNA kit (Clontech), a nucleic acid concentrator step (Zymo Research), and a sizing column purification step (Clontech) to the standard HIV-SMART protocol. NGSID10 (viral load 5.01 $\log_{10}$ copies/ml) was diluted 1:100 in normal human plasma for the library preparation optimization experiments to create a 3 $\log_{10}$ copies/ml sample. All libraries were fragmented, barcoded, multiplexed, and sequenced on the MiSeq (Illumina) instrument as previously described(19). For samples with multiple NGS datasets from optimization experiments, all reads from conditions with a reverse transcriptional RT temperature of 42°C were used to generate a consensus genome sequence. NGS data was processed as previously described using CLC Genomics Workbench 8.0 software (CLC Bio/QIAGEN) with minor modification(19). Briefly, fastq data files were imported to CLC, trimmed for quality and ambiguity, and the SMART primer adapter sequence was removed. For optimization experiments, reads were only aligned to the HXB2 reference genome for consistency in making comparisons. For building genomes, reads were aligned to 6-10 subtype and CRF HIV reference genomes, and complete genomes were built by aligning the resulting consensus sequences. The references used for read mapping are summarized in **Supplementary Table 1**. For consensus sequences with gaps, contigs generated by *de novo* assembly were kept if they aligned to the consensus genome and were merged with NGS data in Sequencher 5.2.3 software to create a final consensus sequence. Sanger sequencing were used to fill in the remaining small gaps (<200bp) with the use of primer sequences in regions with full NGS coverage flanking the gap. The raw NGS data was re-aligned to the final genome sequence to generate NGS coverage and read mapping statistics. Open reading frames were confirmed and annotated with SeqBuilder (DNASTAR Laservene v 11.2) software. The complete genome sequences have been deposited in Genbank under accession references KY392767 to KY392769 (NGSID1 to NGSID3); KY392770 to KY392773 (NGSID5 to NGSID8); KY392774 (NGSID10); KY392775 to KY392779 (NGSID12 to NGSID16); and KY392780 (NGSID18).

Genome sequence classification Complete whole genome sequences were initially subtyped with two online subtyping tools: (*i*) the jumping profiles Hidden Markov Models or jpHMM for short (http://jphmm.gobics.de) and (*ii*) the REGA v 3.0 subtyping tool (http://regatools.med.kuleuven.be/typing/v3/hiv/typingtool). Next, we subtyped the complete genomes through manual phylogenetic inference. The complete genomes were aligned against the 485 whole genome reference and manually edited as previously describe. A ML-tree topology was inferred in RAxML v 8.0.0(31) with the implementation of the GTR+GAMMA model

of nucleotide substitution. The multiple or rapid bootstrap resampling method (n = 1,000) was implemented on a 12-core MacPro to infer support for splits. The final tree topology was visualized in FigTree v 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree) and manually annotated.

Recombinant analysis Following the tree inference complete genomes were scanned for recombination in Recombination Detection Programme v 4.0 (RDP4). Six different methods were used to scan for recombination using the default settings in DRP4. These six methods are: Recombination Detection Programme (RDP)(36), GENECONV(37), Chimaera(38), MaxChi(39), Bootscan[41,42] and SiScan(40).

RDP recombination analyses were followed by manually scanning whole genomes in Simplot v 3.5.1(34). Briefly, a subset of 120 full-genome HIV-1 sequences was selected from the 485 references. This subset broadly covered the global genetic diversity of HIV-1 group M subtypes, while enriching for genomes from the DRC and other sub-Saharan African countries. A list of the 120 HIV-1 strains is presented in **Supplementary Table 2**. Additional references were added for the analyses of some NGSIDs based on preliminary results of previous analyses. For example for NGSID 7 reference sequences belonging to CRF25 were included.

Recombinant fragments were extracted based on the recombination breakpoints identified in the bootscan analyses. Recombinant fragments were subtyped through manual phylogenetic inference within the ML framework as previously described. Trees of short recombinant fragments were visualized in FigTree v 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree) and manually annotated.

The mosaic layout of each recombinant genome was annotated. The recombinant-naming scheme suggested by Tongo and colleagues(12) were used for the classification of recombinants. In this scheme recombinant fragments clustering within established subtype or CRF clades of HIV-1 group M are designated with a capital letter (e.g. |A1| for subtype A1), while basal clustering to established subtypes or CRFs are designated by small letters (e.g. |g| for subtype G). Recombinant that was not supported by bootstrap (splits < 70) was designated as unclassified (U).

Results

A total of 341 specimens were collected from study participants between 2001 and 2003 from the two rural study sites. Plasma screening by the ARCHITECT HIV Ag/Ab Combo (Abbott

Diagnostics, Abbott Park, IL) serological assay identified 278 HIV positive specimens (81.53%). Further serotyping with a peptide immunoassay classified 266 of the HIV positives as belonging to group M of HIV-1 (95.68%), while 12 specimens (4.32%) were non-reactive (**Figure 1**).

RT-PCR and Sanger sequencing of the *env* IDR region produced 172 *env* genotypes. Phylogenetic subtyping of env-IDR genotypes identified the presence of eight of the nine major subtypes of HIV-1 group M, including rare subtypes H (n=5), J (n=3), and K (n=1). The majority of isolates were classified as sub-subtype A1 (n=74, 43.02%), while four sub-subtype A2 isolates were also identified. Subtypes D (n=16, 9.30%) and G (n=15, 8.72%) were the second and third most prevalent subtypes identified in the *env* IDR data set (**Table 1**). It is important to note that the use of a sub-genomic region for subtyping, such as the env gp41, may miss several recombinants and it is likely to overestimate the number of pure subtypes identified. In particular, subtype B isolates were not identified in our study population, although four CRFs (CRF01, CRF02, CRF25 and CRF27) were present (n=20). A total of 20 sequences (11.63%) did not branch with references and were unclassified. A representation of the tree topology containing the 172-*env* IDR DRC sequences along with references is presented in **Figure 2**.

Sanger sequencing of the *gag* and *pol* regions of selected specimens produced 20 *gag* sequences and 14 *pol* sequences. Genotypes from *gag*, *pol* and *env* were available for eleven specimens, while an additional eight specimens had at least two genotypes from one of the three sub-genomic regions. Genotyping results of the *gag* and *pol* sub-genomic regions were cross-referenced with *env* IDR genotypes and revealed eight possible recombinants. One isolate clustered basal to the two putative subtype L isolates in all three regions.

Eighteen isolates were selected for whole genome sequencing based on the subtype assignments in the *gag*, *pol* and *env* data sets. The HIV viral load for the selected specimens ranged from 3.89-5.82 $\log_{10}$ copies/ml (**Table 2**). The HIV-SMART NGS method(19) developed by our group has previously been shown to generate complete genomes from clinical samples with viral loads greater than 5 $\log_{10}$ copies/ml; however, this method has not been applied to clinical specimens with viral load below this cut off. Therefore, the DRC specimens with viral load <5 $\log_{10}$ copies/ml were used in HIV-SMART NGS optimization experiments for low viral load samples.

Several conditions were tested to improve the genome coverage and read depth of low viral load clinical specimens. First, the RT temperature was raised to 47°C or 50°C to improve

primer-binding specificity. Second, a larger amount of input RNA was used to make cDNA libraries either by concentration of nucleic acid extracts on a concentrator filter column (Zymo Research) or by using a Pico SMART cDNA kit (Clontech). Thirdly, a sizing column (Clontech) was used to select larger amplicons from the total SMART cDNA libraries in a clean-up step. Lastly, the 47°C RT condition was combined with the concentrator column. In the RT temperature comparison, n=8 specimens with viral loads ranging from 3.89-5.2 $\log_{10}$ copies/ml were included, and n=7 had the highest read depth at 50°C and highest % HIV reads at 42°C (**Figure 3**). Notably, the genome coverage was approximately 50% lower in the 50°C libraries compared to the 42°C libraries (**Figure 3**). Therefore, an RT temperature of 42°C was selected for low viral load specimens. In the library preparation optimization experiments, n=4 samples covering a viral load range of 3-5 $\log_{10}$ copies/ml were included. All conditions that included the nucleic acid concentrator step had the highest genome coverage and % HIV reads for specimens with viral load >4 $\log_{10}$ copies/ml (**Figure 4**). For these specimens, raising the RT temperature to 47°C did not affect genome coverage when the concentrator was used. Specimens with viral load <4 $\log_{10}$ copies/ml had inconsistent genome coverage, % HIV reads, and read depth for all of the tested conditions. These results indicated that adding the concentrator step greatly improved the quality and coverage of HIV-SMART genome sequences.

The HIV-SMART NGS method with an RT temperature of 42°C and a nucleic acid concentrator step for low viral load specimens resulted in complete genome sequences (>99% coverage) for 14 specimens and partial genomes for 4 specimens (**Table 2**). All complete genomes had a length of at least 9550 nucleotides and the average coverage depth was >10 reads for all of the HIV-SMART NGS sequenced regions (**Supplementary Figure 1**). Several short gaps (<200bp) were filled in by Sanger sequencing to complete n=5 genomes. The n=4 partial genomes had coverage ranging from 63%-75% with gaps of various sizes throughout the sequences (**Table 2 and Supplementary Figure 1**). The 14 complete genomes produced by HIV-SMART NGS were subsequently classified by phylogenetic inference and recombinant analysis.

Online subtyping tools were initially used to classify the 14 DRC whole genomes (**Table 3**). REGA identified seven pure viral subtypes including one subtype C (NGSID 2), one subtype D (NGSID 3), one subtype F1 (NGSID 5), one subtype J (NGSID 13) and three subtype H (NGSID 14 - 16). The remaining seven whole genome DRC sequences were classified as possible recombinants. Subtyping with jpHMM confirmed the classification of pure viral isolates identified in REGA 3. Additionally, jpHMM classified NGSID 1 as a pure subtype C isolate whereas in

338 REGA this isolates was classified as a recombinant form between subtypes C and D. Broadly
339 similar recombinant profiles were identified between REGA and jpHMM for the six putative
340 recombinant genomes (**Table 3**).

341 Subtyping of the genomes through manual phylogenetic inference supported the classification
342 for the eight NGS isolates that were classified as "pure" subtypes (**Figure 5 and Table 4**).
343 NGSID 1 and 2 clustered basal to the main subtype C clade, containing also major "C-like"
344 CRFs (e.g. CRF07 & 08), with good bootstrap support. Given the clustering of recombinants
345 within the main subtype C clade and the basal clustering of NGSID 1 and 2 to the main subtype
346 C clade there is a good probability that small recombination events might have occurred in
347 these two specimens, which might not have been picked up by the online subtyping tools.
348 NGSID 3 clustered within the main subtype D clade, while NGSID 5 clustered along with a F1
349 isolate from Russia (GQ290462) basal to the main F1 clade. NGSID 6 clustered with one
350 unclassified isolate (JF683772) though the split in the tree topology separating these two
351 isolates from the rest of the tree were not supported (**Figure 5 and Table 4**). These two isolates
352 in turn clustered basal to the main subtype G clade, which are indicative of possible
353 recombination in these two isolates.

354 NGSID 7 clustered basal to a cluster containing isolates belonging to CRF25_cpx with good
355 support. Contained within this CRF25_cpx cluster is one problematic isolate DQ826727. Closer
356 investigation revealed this isolate to be a complex unique recombinant form between CRF02
357 and CRF25. NGSID 8 and 10 clustered basal to a to the clade containing CRF04, though the
358 branch separating these two isolates from the CRF04 clade were very long which are indicative
359 of substantial genetic distance between the DRC isolates and CRF04. NGSID 12 clustered with
360 another unclassified isolate from the reference dataset (AF076475) with 99 bootstrap support
361 for the split separating these two isolates from the rest of the tree topology. AF076475 is a
362 unique recombinant form between subtypes F2 and K along with unclassified regions and was
363 characterized from an individual from Belgium. The basal clustering of NGSID 12 along with
364 AF076475 to the main subtype K clade is indicative of possible recombination in these two
365 isolates. NGSID 13 clustered within the subtype J branch along with the newly characterized
366 subtype J isolates from Angola(8). NGSID 14 and 15 clustered within the main subtype H clade,
367 while NGSID 16 clustered basal to the main subtype H clade. The split separating these three
368 isolates and the subtype H clade from the rest of the tree topology was very well supported.
369 Finally, NGSID 18 clustered basal to a clade containing CRF45_cpx, though the branch of

NGSID 18 was long, which indicates substantial genetic distance between NGSID 18 and CRF45 isolates.

Recombination analyses were performed to investigate putative recombinants within our data set. DRP4 analyses of the eight "pure" subtype isolates identified small possible recombination events, although none of the p-values for these classifications were significant. NGSID 6 was classified as a recombinant form between subtypes A1, G and H. Notably, RDP4 indicated that the subtype H fragment of NGSID 6 was more closely related to the homologous subtype H segments of NGSIDs 8 and 10 than to other H references. In the RDP4 phylogenetic tree inference, NGSIDs 6, 8 and 10 clustered basal to the main subtype H clade, while NGSID 14 – 16, which are the three pure subtype H whole genome isolates, clustered within the main subtype H clade. RDP4 classified the majority of the viral backbone of NGSID 7 as belonging to CRF25_cpx with a small recombinant fragment in the *env* region corresponding to subtype A1.

Similar recombination profiles were uncovered by RDP4 for NGSID 8 and 10. RDP4 classified these two isolates as recombinants between subtypes A1, G and H. RDP4 classified NGSID 12 as a pure subtype K isolate with no sign of recombination. Finally, RDP4 classified NGSID 18 as a recombinant between subtypes A1, K and J.

Manual bootscan analyses that were performed on the 14 whole genome DRC sequences supported the classification of the eight pure viral subtypes. The recombinant breakpoints that were uncovered by bootscan analyses of the six putative recombinant isolates (**Figure 6**) broadly reflected similar trends to the results from the DRP4 analyses as well as those from the online subtype methods.

Phylogenetic inference of recombinant fragments was used to classify each of the six recombinant genomes in our sequence cohort (**Supplementary Figures 2-6**). NGSID 6 was classified as a complex recombinant form between subtypes A1, G and H with the following mosaic structure: A1|g|A1|g|h|H|G (**Supplementary Figure 2**). Analyses of NGSID 7 classified this isolate as a recombinant between CRF25 and sub-subtype A1. Notably, the segment of the genome corresponding to CRF25 clustered basal to the main CRF25 clade, which resulted in the following mosaic recombinant structure for this isolate: crf25|A1|crf25 (**Supplementary Figure 3**). The six recombinant viral genomes and their respective mosaic structures are presented in **Figure 7**.

Our manual phylogenetic inference of recombinant fragments in NGSID 8 and 10 suggests that these two isolates are identical recombinants. The tree inference of recombinant fragments indicates that these two isolates are complex recombinants between subtypes A1, G, H and K with the following mosaic structure: A1|k|a1|h|g|h (**Supplementary Figure 4**). Our tree inference of NGSID 12 classified this isolate as a recombinant form between subtypes K and F. However, the subtype F fragment could not be unambiguously classified as either belonging to sub-subtype F1 or F2 as this isolate clustered basal to the main subtype F clade. An additional two segments, one in the 3'*pol* and *vif* region and the other in the *nef* and 3'LTR region, could not be classified due to poor branch support and were subsequently categorized as unclassified. The mosaic structure of NGSID 12 was assigned as follows: K|f|k|U|k|U (**Supplementary Figure 5**). Finally, NGSID 18 was classified as a complex recombinant between subtypes A1, J and K. The recombinant fragment on the 3'LTR side of this isolate clustered with both subtype A1 and A2 isolates and were subsequently categorized as subtype A. The final mosaic structure for this isolate was called as follows: A1|k|A1|j|A (**Supplementary Figure 6**).

Discussion

The complete and sub-genomic HIV-1 sequences characterized in the present study have considerably expanded the known genetic pool of HIV-1 strains from the DRC and improved our understanding of the epidemic within the Congo Basin. The characterization of 172 *env*-IDR genotypes identified a wide pool of genetic variants, with subtypes A1, D and G being the most frequently identified within the cohort. Small numbers of rare viral subtypes were also identified, including A2 (n=4), H (n=5), J (n=3) and K (n=1). A small number of CRFs have also been identified, including CRF01_AE (n=6), CRF02_AG (n=10), CRF25_cpx (n=2) and CRF27_cpx (n=2) (**Table 1** and **Figure 2**). Initial subtyping was inferred by analyzing env gp41 sequences rather than whole genome sequences. This approach may fail to detect recombinants and overestimate the number of pure subtypes. Of the 172 specimens sequenced in the *env*-IDR region, 19 were also sequenced in either the *gag* or *pol* regions. One isolate clustered basal in all three regions to two unclassified isolates which were previously suggested as a new subtype of HIV-1 group M (Subtype L). Cross-referencing of *gag* and *pol* with *env*-IDR genotypes identified eight possible recombinants. However, the true number of recombinants in our study cohort is most likely much higher, given the small number of patients sequenced in more than one region of the HIV-1 genome.

Appreciation of the genetic diversity observed in sub-genomic sequences prompted our efforts to sequence the complete genomes of the rare variants identified in the *gag*, *pol*, and *env* regions. Unfortunately, many of the rare variants we selected had low viral load upon quantitation, and failed to produce RT-PCR bands for sequencing in other regions of the HIV-1 genome. Therefore, to efficiently sequence the 18 selected samples, we followed an NGS method that was optimized for low viral load samples. Improvement of the HIV-SMART NGS method resulted in complete or near complete genome sequences for samples with viral load below the previous threshold of 5 $\log_{10}$ copies/ml. Despite comparing six different conditions to improve genome coverage and read depth, clinical specimens with viral load lower than 4 $\log_{10}$ copies/ml were not consistently sequenced by the modified HIV-SMART method. The reduction in reliable complete genome coverage for these low viral load specimens is likely due to poor HIV-SMART primer annealing, extension or a combination of the two. For samples below this threshold viral load, sequencing libraries of purified RT-PCR amplicons may be an alternative method that could reduce background reads and improve read depth. Of the conditions tested, the lowest RT temperature (42^{o}C) and the addition of a nucleic acid concentrator step before library preparation had the greatest improvements in genome coverage and read depth (**Figures 3** and **4**). While higher temperatures dramatically improved read depth in this study, overall genome coverage was reduced by nearly half, suggesting that the reverse transcriptase enzyme was less processive or that the RNA template was degraded at higher temperatures despite expectations that genome coverage would improve due to increased denaturation of RNA template secondary structures[46] (**Figure 4**). Although the Pico SMART cDNA synthesis kit accommodated a larger volume of input nucleic acid template for library preparation, it did not improve NGS genome coverage or read depth (**Figure 4**). Removal of short transcripts by the use of a sizing column was expected to improve the signal to background ratio for HIV-1 reads, but this step did not affect the percentage of HIV reads (**Figure 4**). In contrast, concentration of the input nucleic acid greatly improved sequencing results (**Figure 4**), which may be due to both the concentration and the purification functions of the columns. The addition of the concentrator column may also improve read depth for high viral load samples, although this has not been tested. The complex genomes from patient samples that were sequenced by the modified HIV-SMART NGS protocol are an excellent example of the utility of this method for surveillance of diverse HIV-1 sequences. The application of HIV-SMART NGS to larger quantities of samples will ultimately bring viral surveillance to the whole genome scale and improve our understanding of the true diversity and evolution of HIVat inter- and intra-patient level. Ultimately, the optimized HIV-SMART NGS method combined with Sanger sequencing of short gap regions

resulted in complete genome sequences for 14 rare variant HIV-1 specimens and 4 partial genomes (**Table 2**). We plan to use this important dataset in the future to identify the date of origin of subtypes C, H and J. Future incorporation of primer ID sequences into the HIV-SMART method to allow intra-patient diversity analysis for rare HIV specimens from the Congo Basin.

The characterization and classification of the 14 whole genome isolates from the DRC were complicated by the extreme genetic diversity observed within our sample cohort and sequence dataset. Particularly, several of the pure viral isolates and recombinant fragments clustered basal to major HIV-1 group M clades. In the majority of cases, this basal clustering is due to the small number of viral isolates to compare new genotypes against. For example, any analysis against subtype K is limited to only two whole genome reference sequences. This has resulted in basal clustering of any of the subtype K recombinant fragments identified in NGSIDs 8, 10, 12 and 18. Similar basal clustering of recombinant fragments corresponding to subtype A1, G, J, H and CRF25 were observed in the recombinants that are described here. Additionally, one of the subtype H isolates and the two subtype C isolates described here also clustered phylogenetically basal to their respective subtype clades. This basal clustering of sequences from the DRC underscores the deep genetic diversity of the global pandemic that is still circulating in relatively high numbers in areas within the Congo Basin.

Although it is clear that HIV-1 originated in the Congo Basin(2, 41) we still know little about the early transmission and dissemination of the strains that left the Congo Basin to cause outside epidemics. In order to test the hypothesis that much of the genetic diversity did not leave the Congo Basin, we analyzed all of the public sequences in the LANL database (date of access: 8 April 2016). This dataset included all of the HIV-1 group M subtypes and CRF sequences which were >500 bp. In total, we found 411,194 sequences, 7,158 (1.74%) of which were from the Congo Basin. The Congo Basin still contains most of the diversity of HIV-1 in the world(6, 42). There are subtypes that caused large epidemics outside the Congo Basin that are still prevalent in this region, including subtypes A1, D and G (**Figure 8**).There are also subtypes in the Congo Basin that have not caused significant epidemics outside the region, such as subtypes A2, F2, H, J and K. In total, 1160 sequences of these subtypes have been identified to date and 721 (62.15%) of these were identified in the Congo Basin. Furthermore, most of the H, J, and K strains that have been identified outside the Congo Basin are from expatriates or from visitors to the region(43-47).

Interestingly, the two main epidemiologically important subtypes in the world originated in the Congo Basin but are now not commonly found in the region(6, 42). The LANL public dataset contained only 18 subtype B sequences and 14 subtype C sequences from this region. We also did not identify any subtype B isolates in our samples. This supports the hypothesis that the subtype B ancestral strain left the Congo Basin at an early stage of the HIV pandemic(4). However, we were able to identify five samples from subtype C and managed to sequence the first two whole genomes of subtype C isolates from the Congo Basin. In our phylogenetic reconstruction, these two whole genome subtype C isolates clustered basal to all known subtype C sequences (**Figure 5**). Their basal clustering suggests that these sequences are ancestors of the global subtype C pandemic. Our results are supported by other epidemiological and phylogenetic studies that used sub-genomic regions of subtype C(5, 48).

The current nomenclature system for HIV-1 needs to be updated in order to track epidemiologically important strains. The current nomenclature recognizes nine subtypes (A, B, C, D, F, G, H, J and K), four sub-subtypes (A1, A2, F1 and F2) and almost 100 CRFs (74 CRFs at the time of writing this paper). The majority of the subtypes, sub-subtypes and CRFs have very limited epidemiological importance(6). For example, as shown in this manuscript, subtypes A2, F2, H, J and K are mostly restricted to the Congo Basin (**Figure 8**). We have also estimated that only ten of the 74 CRFs in the LANL public dataset seem to be epidemiologically important by plotting their distribution over time (i.e. more than 50 sequences sampled over 5 years). These include CRF01_AE, which is currently spreading in South East Asia, CRF02_AG in West Africa, CRF07_BC in China, CRFs 18_cpx and 19_cpx in Cuba, CRF35_AD in Afghanistan and CRF63_02A1 in Russia.

We suggest that in the future the HIV-1 nomenclature system annotate only strains of epidemiological importance. This would be especially valuable for new CRFs, as otherwise, high-throughput NGS methods, such as the one described in this manuscript, will end up identifying 100s of new CRFs, which will further complicate the current nomenclature system. Focusing on the most epidemiologically important strains may facilitate the development of more effective HIV drugs and vaccines. For example, focused research in subtype C, which accounts for over 50% of the global infections, is crucial. Recent results suggest that the K65R mutation, one of the main mutations causing resistance to first line antiretroviral drug Tenofovir, emerges rapidly in subtype C(49, 50).

We also suggest that recombinant fragments of the genome are named according to the scheme introduced by Tongo and colleagues(12). This nomenclature system uses lowercase letters to classify recombinant fragments derived from a virus that branches basal to a given subtype in a phylogenetic tree. The use of capital letters represents clustering within the currently known subtype diversity. For example, when we use this system to annotate our recombinants, NGSID8 was classified as A1-k-A1-j-A. This system may be particularly useful to discover and characterize more of the highly divergent lineages that exist in the Congo Basin. This may shed light on specific viral genetic factors that enabled HIV-1 group M strains to leave the Congo Basin and cause major worldwide epidemics.

Conclusions

The advance of next generation sequencing methods, such as the one presented in this study, can be used to sequence rare and diverse HIV-1 samples. Here, we have identified new subtypes and recombinants that expand the genetic diversity of HIV-1 in the Congo Basin, which is the region where HIV-1 originated. The basal branching of some of the subtypes and recombinant segments we recovered show that these strains are more closely related to ancestral HIV-1 sequences than to previously reported strains. It is evidence that the local diversification of HIV-1 in the Congo Basin continues to outpace the diversity of global strains. With an improved understanding of HIV-1 genetic diversity, we will be better able to assess the risks of the emergence of future outbreaks, to track the evolution of the global pandemic and to develop new drugs and vaccine targets.

FIGURE LEGENDS

Table 1 - Subtype assignment of 172 *env* IDR sequences. The total number of sequences (n) and percentage of the total of all sequences (%) for each subtype are listed for the *env* region sequences. Unclassified sequences did not branch with references with bootstrap support > 70.

Table 2 - Summary of the eighteen isolates that were chosen for whole genome sequencing using the HIV-SMART method. The viral load was quantified by the HIV RealTi*m*e assay (Abbott Molecular Diagnostics). Subtyping of *gag*, *pol*, and *env* IDR sequences was performed through Maximum Likelihood phylogenetic inference of a 468 bp region of *gag*, an 864 bp region of *pol*, and a 676 bp region of *env*, respectively. The whole genome coverage

and genome length were calculated in CLC Bio for the final consensus genome sequences that was generated using the HIV-SMART sequencing method.

Table 3 - Results of the online subtyping methods. Fourteen whole genome sequences were subtyped with the REGA v 3.0 and jpHMM online subtyping methods.

Table 4 - Subtype assignment of the 14 DRC NGS genotypes. This table represents the subtype classification for the 14 NGS genotypes made by Simplot, Bootscan and manual phylogenetic inference.

Figure 1 - Specimen testing and genotyping workflow. The specimen processing and testing steps in this study are summarized with the results in each box in the flow chart. The number (n) of specimens included in each step is indicated. Red arrows indicate genotypes used for phylogenetic analyses (dark grey box).

Figure 2. Neighbor Joining tree of 117 *env* IDR sequences. The tree was inferred by Phylip. The tree was limited to a subset of 117 specimen sequences representing the major identified classifications for better visualization. Bootstrap values are shown for major branchpoints indicated by black dots.

Figure 3. Reverse transcriptase temperature optimization. The trimmed NGS reads from HIV-SMART libraries prepared at the indicated reverse transcriptase temperatures (42°C, 47°C, or 50°C) were mapped to the HXB2 reference genome. The genome coverage (A) and %HIV reads (B) were calculated for this alignment in CLC Bio. The viral load for each sample tested is plotted on a log scale in (C). The genome coverage plots for each position of the genome are shown for NGSID12 (D), which was representative of the trend seen in all other samples tested.

Figure 4. Library preparation optimization. The trimmed NGS reads from HIV-SMART libraries prepared following the indicated protocols A-E (A Standard protocol, B Pico SMART cDNA kit, C nucleic acid concentrator, D sizing column, and E condition C with 47°C reverse transcription) were mapped to the HXB2 reference genome. The genome coverage (A) and %HIV reads (B) were calculated for this alignment in CLC Bio. The viral load for each sample tested is plotted on a log scale in (C). The genome coverage plots for each position of the genome are shown for NGSID4 (D), which was representative of the trend seen in all other samples tested.

Figure 5. ML-phylogenetic tree of 14 whole genome DRC sequences (highlighted in red) and a representative number (n=485) of HIV-1 reference strains. The tree was constructed in RAxML v 8.0.0 with the GTR+G model and 1000 bootstrap replicates. The bar at the bottom represents

the genetic distance along branches. This tree was midpoint rooted and branches for major clades that did not cluster with DRC isolates were collapsed. Bootstrap values for major splits in the tree topology are shown in individual clades (blank circle).

Figure 6. Bootscan plots of the six recombinant whole genome sequences from the DRC. Each bootscan plot was performed under the Kimura-2 model of nucleotide substitution with a window size of 500 and a step size of 50. Colour coded key represents different subtypes, sub-subtypes or CRFs of HIV-1. The dotted line represents 70% of permutated trees.

Figure 7. Recombinant mosaic profiles of the six recombinant whole genome sequences that were generated in the course of this study. The numbers at the top of each genome represents the recombinant breakpoints relative to the HXB2 reference strain. The recombinant profile of each isolate is presented on the right hand side with capital lettering referring to clustering within established clades and small letters representing basal clustering to that particular clade.

Figure 8. Frequency of sampling of major subtypes of HIV-1. Red indicates sampling within the Congo Basin and blue represents sampling outside the Congo Basin. The y-axes are not proportional. The y-axes represents the number of genotypes (>500bp) while the x-axes represents years in calendar time.

References

1. **Hemelaar J.** 2012. The origin and diversity of the HIV-1 pandemic. Trends Mol Med **18:**182-192.
2. **Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S, Bhattacharya T.** 2000. Timing the ancestor of the HIV-1 pandemic strains. Science **288:**1789-1796.
3. **Salemi M, Strimmer K, Hall WW, Duffy M, Delaporte E, Mboup S, Peeters M, Vandamme AM.** 2001. Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes using a new method to uncover clock-like molecular evolution. FASEB J **15:**276-278.
4. **Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K, Bunce M, Muyembe JJ, Kabongo JM, Kalengayi RM, Van Marck E, Gilbert MT, Wolinsky SM.** 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature **455:**661-664.
5. **Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, Tatem AJ, Sousa JD, Arinaminpathy N, Pepin J, Posada D, Peeters M, Pybus OG, Lemey P.** 2014. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. Science **346:**56-61.

6. **Hemelaar J, Gouws E, Ghys PD, Osmanov S, Isolation W-UNfH, Characterisation.** 2011. Global trends in molecular epidemiology of HIV-1 during 2000-2007. AIDS **25:**679-689.
7. **Laboratories LAN.**
8. **Bartolo I, Calado R, Borrego P, Leitner T, Taveira N.** 2016. Rare HIV-1 subtype J genomes and a new H/U/CRF02_AG recombinant genome suggests an ancient origin of HIV-1 in Angola. AIDS Res Hum Retroviruses doi:10.1089/AID.2016.0084.
9. **Gao F, Trask SA, Hui H, Mamaeva O, Chen Y, Theodore TS, Foley BT, Korber BT, Shaw GM, Hahn BH.** 2001. Molecular characterization of a highly divergent HIV type 1 isolate obtained early in the AIDS epidemic from the Democratic Republic of Congo. AIDS Res Hum Retroviruses **17:**1217-1222.
10. **Mokili JL, Rogers M, Carr JK, Simmonds P, Bopopi JM, Foley BT, Korber BT, Birx DL, McCutchan FE.** 2002. Identification of a novel clade of human immunodeficiency virus type 1 in Democratic Republic of Congo. AIDS Res Hum Retroviruses **18:**817-823.
11. **Vidal N, Peeters M, Mulanga-Kabeya C, Nzilambi N, Robertson D, Ilunga W, Sema H, Tshimanga K, Bongo B, Delaporte E.** 2000. 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 **74:**10498-10507.
12. **Tongo M, Dorfman JR, Martin DP.** 2015. High Degree of HIV-1 Group M (HIV-1M) Genetic Diversity within Circulating Recombinant Forms: Insight into the Early Events of HIV-1M Evolution. J Virol **90:**2221-2229.
13. **Mokili JL, Wade CM, Burns SM, Cutting WA, Bopopi JM, Green SD, Peutherer JF, Simmonds P.** 1999. Genetic heterogeneity of HIV type 1 subtypes in Kimpese, rural Democratic Republic of Congo. AIDS Res Hum Retroviruses **15:**655-664.
14. **Gaschen B, Taylor J, Yusim K, Foley B, Gao F, Lang D, Novitsky V, Haynes B, Hahn BH, Bhattacharya T, Korber B.** 2002. Diversity considerations in HIV-1 vaccine selection. Science **296:**2354-2360.
15. **Taylor BS, Sobieszczyk ME, McCutchan FE, Hammer SM.** 2008. The challenge of HIV-1 subtype diversity. N Engl J Med **358:**1590-1602.
16. **Longosz AF, Morrison CS, Chen PL, Brand HH, Arts E, Nankya I, Salata RA, Quinn TC, Eshleman SH, Laeyendecker O.** 2015. Comparison of antibody responses to HIV infection in Ugandan women infected with HIV subtypes A and D. AIDS Res Hum Retroviruses **31:**421-427.
17. **Djoko CF, Rimoin AW, Vidal N, Tamoufe U, Wolfe ND, Butel C, LeBreton M, Tshala FM, Kayembe PK, Muyembe JJ, Edidi-Basepeo S, Pike BL, Fair JN, Mbacham WF, Saylors KE, Mpoudi-Ngole E, Delaporte E, Grillo M, Peeters M.** 2011. High HIV type 1 group M pol diversity and low rate of antiretroviral resistance mutations among the uniformed services in Kinshasa, Democratic Republic of the Congo. AIDS Res Hum Retroviruses **27:**323-329.
18. **Anonymous.** *on* Los Alamos National Laboratory. http://www.hiv.lanl.gov/. Accessed
19. **Berg MG, Yamaguchi J, Alessandri-Gradt E, Tell RW, Plantier JC, Brennan CA.** 2015. A Pan-Hiv Strategy for Complete Genome Sequencing. J Clin Microbiol doi:10.1128/JCM.02479-15.
20. **Bodelle P, Vallari A, Coffey R, McArthur CP, Beyeme M, Devare SG, Schochetman G, Brennan CA.** 2004. Identification and genomic sequence of an HIV type 1 group N isolate from Cameroon. AIDS Res Hum Retroviruses **20:**902-908.
21. **Brennan CA, Bodelle P, Coffey R, Devare SG, Golden A, Hackett J, Jr., Harris B, Holzmayer V, Luk KC, Schochetman G, Swanson P, Yamaguchi J, Vallari A, Ndembi N, Ngansop C, Makamche F, Mbanya D, Gurtler LG, Zekeng L, Kaptue L.**

2008. The prevalence of diverse HIV-1 strains was stable in Cameroonian blood donors from 1996 to 2004. J Acquir Immune Defic Syndr **49:**432-439.

22. **Luk K-C, Berg,M.G., Naccache,S.N., Kabre,B., Federman,S., Mbanya,D., Kaptue,L., Chiu,C.Y., Brennan,C.A. and Hackett,J. Jr.** 2015. Utility of Unbiased Next-Generation Sequencing for HIV Surveillance.
23. **Luk KC, Holzmayer V, Ndembi N, Swanson P, Brennan CA, Ngansop C, Mbanya D, Kaptue L, Gurtler L, Devare SG, Hackett J.** 2008. Near full-length genome characterization of an HIV type 1 CRF25_cpx strain from Cameroon. AIDS Res Hum Retroviruses **24:**1309-1314.
24. **Vallari A, Bodelle P, Ngansop C, Makamche F, Ndembi N, Mbanya D, Kaptue L, Gurtler LG, McArthur CP, Devare SG, Brennan CA.** 2010. Four new HIV-1 group N isolates from Cameroon: Prevalence continues to be low. AIDS Res Hum Retroviruses **26:**109-115.
25. **Yamaguchi J, Coffey R, Vallari A, Ngansop C, Mbanya D, Ndembi N, Kaptue L, Gurtler LG, Bodelle P, Schochetman G, Devare SG, Brennan CA.** 2006. Identification of HIV type 1 group N infections in a husband and wife in Cameroon: viral genome sequences provide evidence for horizontal transmission. AIDS Res Hum Retroviruses **22:**83-92.
26. **Yamaguchi J, Badreddine S, Swanson P, Bodelle P, Devare SG, Brennan CA.** 2008. Identification of new CRF43_02G and CRF25_cpx in Saudi Arabia based on full genome sequence analysis of six HIV type 1 isolates. AIDS Res Hum Retroviruses **24:**1327-1335.
27. **Yamaguchi J, Vallari A, Ndembi N, Coffey R, Ngansop C, Mbanya D, Kaptue L, Gurtler LG, Devare SG, Brennan CA.** 2008. HIV type 2 intergroup recombinant identified in Cameroon. AIDS Res Hum Retroviruses **24:**86-91.
28. **Anonymous.** Benzonase endonuclease: The Smart Solution for DNA Removal.
29. **Law J, Jovel J, Patterson J, Ford G, O'Keefe S, Wang W, Meng B, Song D, Zhang Y, Tian Z, Wasilenko ST, Rahbari M, Mitchell T, Jordan T, Carpenter E, Mason AL, Wong GK.** 2013. Identification of hepatotropic viruses from plasma using deep sequencing: a next generation diagnostic tool. PLoS One **8:**e60595.
30. **Swanson P, Devare SG, Hackett J, Jr.** 2003. Molecular characterization of 39 HIV isolates representing group M (subtypes A-G) and group O: sequence analysis of gag p24, pol integrase, and env gp41. AIDS Res Hum Retroviruses **19:**625-629.
31. **Stamatakis A.** 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30:**1312-1313.
32. **S T.** 1986. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences., vol 7. American Mathematical Society.
33. **Jin L, Nei M.** 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol Biol Evol **7:**82-102.
34. **Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, Ingersoll R, Sheppard HW, Ray SC.** 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J Virol **73:**152-160.
35. **J F.** 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39:**783-791.
36. **Martin D, Rybicki E.** 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics **16:**562-563.
37. **Padidam M, Sawyer S, Fauquet CM.** 1999. Possible emergence of new geminiviruses by frequent recombination. Virology **265:**218-225.
38. **Posada D, Crandall KA.** 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc Natl Acad Sci U S A **98:**13757-13762.

39. **Smith JM.** 1992. Analyzing the mosaic structure of genes. J Mol Evol **34:**126-129.
40. **Gibbs MJ, Armstrong JS, Gibbs AJ.** 2000. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics **16:**573-582.
41. **Sharp PM, Bailes E, Gao F, Beer BE, Hirsch VM, Hahn BH.** 2000. Origins and evolution of AIDS viruses: estimating the time-scale. Biochem Soc Trans **28:**275-282.
42. **Osmanov S, Pattou C, Walker N, Schwardlander B, Esparza J, Isolation W-UNfH, Characterization.** 2002. Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000. J Acquir Immune Defic Syndr **29:**184-190.
43. **Cunningham CK, Chaix ML, Rekacewicz C, Britto P, Rouzioux C, Gelber RD, Dorenbaum A, Delfraissy JF, Bazin B, Mofenson L, Sullivan JL.** 2002. Development of resistance mutations in women receiving standard antiretroviral therapy who received intrapartum nevirapine to prevent perinatal human immunodeficiency virus type 1 transmission: a substudy of pediatric AIDS clinical trials group protocol 316. J Infect Dis **186:**181-188.
44. **Laukkanen T, Albert J, Liitsola K, Green SD, Carr JK, Leitner T, McCutchan FE, Salminen MO.** 1999. Virtually full-length sequences of HIV type 1 subtype J reference strains. AIDS Res Hum Retroviruses **15:**293-297.
45. **Sullivan PS, Do AN, Ellenberger D, Pau CP, Paul S, Robbins K, Kalish M, Storck C, Schable CA, Wise H, Tetteh C, Jones JL, McFarland J, Yang C, Lal RB, Ward JW.** 2000. Human immunodeficiency virus (HIV) subtype surveillance of African-born persons at risk for group O and group N HIV infections in the United States. J Infect Dis **181:**463-469.
46. **Yerly S, Rickenbach M, Popescu M, Taffe P, Craig C, Perrin L, Swiss HIVCS.** 2001. Drug resistance mutations in HIV-1-infected subjects during protease inhibitor-containing highly active antiretroviral therapy with nelfinavir or indinavir. Antivir Ther **6:**185-189.
47. **Holzmayer V, Aitken C, Skinner C, Ryall L, Devare SG, Hackett J, Jr.** 2009. Characterization of genetically diverse HIV type 1 from a London cohort: near full-length genomic analysis of a subtype H strain. AIDS Res Hum Retroviruses **25:**721-726.
48. **Novitsky V, Wang R, Lagakos S, Essex M.** 2010. HIV-1 Subtype C Phylodynamics in the Global Epidemic. Viruses **2:**33-54.
49. **TenoRes Study G.** 2016. Global epidemiology of drug resistance after failure of WHO recommended first-line regimens for adult HIV-1 infection: a multicentre retrospective cohort study. Lancet Infect Dis **16:**565-575.
50. **Lessells RJ, Katzenstein DK, de Oliveira T.** 2012. Are subtype differences important in HIV drug resistance? Curr Opin Virol **2:**636-643.

Table 1 - Subtype assignment of 172 *env* IDR sequences. The total number of sequences (n) and percentage of the total of all sequences (%) for each subtype are listed for the *env* region sequences. Unclassified sequences did not branch with references with bootstrap support > 70.

Subtype/CRF	Number (n)	Percentage (%)
Subtype A1	74	43.023
Subtype A2	4	2.326
Subtype B	0	0
Subtype C	5	2.907
Subtype D	16	9.302
Subtype F1	8	4.651
Subtype F2	0	0
Subtype G	15	8.721
Subtype H	5	2.907
Subtype J	3	1.744
Subtype K	1	0.581
Subtype L	1	0.581
Unclassified	20	11.628
CRF01	6	3.488
CRF02	10	5.814
CRF25	2	1.163
CRF27	2	1.163
Total	**172**	100

Table 2 - Summary of the eighteen isolates that were chosen for whole genome sequencing using the HIV-SMART method. The viral load was quantified by the HIV RealTi*m*e assay (Abbott Molecular Diagnostics). Subtyping of *gag*, *pol*, and *env* IDR sequences was performed through Maximum Likelihood phylogenetic inference of a 468 bp region of *gag*, an 864 bp region of *pol*, and a 676 bp region of *env*, respectively. The whole genome coverage and genome length were calculated in CLC Bio for the final consensus genome sequences that was generated using the HIV-SMART sequencing method.

NGSID	Viral load, (Log_{10} copies/ml)	*gag* subtype	*pol* subtype	*env* IDR subtype	Genome coverage	Genome length
NGSID 1	5.26	-	-	C	100	9692
NGSID 2	4.35	C	C	C	100	9723
NGSID 3	4.45	-	-	D	100	9751
NGSID 4	4.02	-	-	F1	75	9459
NGSID 5	4.65	-	-	F1	100	9743
NGSID 6	5.38	-	-	U	100	9660
NGSID 7	5.28	-	-	CRF25	100	9764
NGSID 8	5.2	PCR neg	PCR neg	U	100	9556
NGSID 9	4.29	A1	A1	H	72	9633
NGSID 10	5.01	U	PCR neg	U	100	9621
NGSID 11	3.89	L	L	L	63	8872
NGSID 12	5.2	K	K	K	99.75	9579
NGSID 13	5.42	J	J	J	100	9688
NGSID 14	5.82	H	H	H	100	9716
NGSID 15	5.24	H	H	H	100	9734
NGSID 16	4.7	H	H	H	100	9658
NGSID 17	4.36	A1	-	H	67	8649
NGSID 18	4.75	A1	PCR neg	J	100	9622

PCR – Polymerase chain reaction, neg - negative

Table 3 - Results of the online subtyping methods. Fourteen whole genome sequences were subtyped with the REGA v 3.0 and jpHMM online subtyping methods.

Sample	REGA v 3.0		jpHMM	
	Classification	Support[1]	Classification	Support[2]
NGSID_1	Recombinant of C, D	NA	Subtype C	1
NGSID_2	HIV-1 Subtype C	100	Subtype C	1
NGSID_3	HIV-1 Subtype D	100	Subtype D	1
NGSID_5	HIV-1 Subtype F1	100	Subtype F1	1
NGSID_6	Recombinant of G, A1, H	NA	Recombinant of A1, G & H	0.8 - 1.0
NGSID_7	Recombinant of 25_cpx, A1, G	NA	Recombinant of A1 & G	0.7 - 1.0
NGSID_8	Recombinant of H, A1, 04_cpx, G, K	NA	Recombinant of A1, H & K	0.9 - 1.0
NGSID_10	Recombinant of H, A1, 04_cpx, G, K	NA	Recombinant of A1, H & K	0.9 - 1.0
NGSID_12	Recombinant of K, J, F1	NA	Recombinant of C, F1 & K	0.6 - 1.0
NGSID_13	HIV-1 Subtype J	100	Subtype J	1
NGSID_14	HIV-1 Subtype H	100	Subtype H	1
NGSID_15	HIV-1 Subtype H	100	Subtype H	1
NGSID_16	HIV-1 Subtype H	100	Subtype H	1
NGSID_18	Recombinant of A1, J, K, G	NA	Recombinant of A1, J & K	0.6 - 1.0

1 – bootstrap support; 2 – posterior support

Table 4 - Subtype assignment of the 14 DRC NGS genotypes. This table represents the subtype classification for the 14 NGS genotypes made by Simplot, Bootscan and manual phylogenetic inference.

Sample	Simplot		Bootscan		Large phylogeny	
	Classification	Support[1]	Classification	Support[2]	Classification	Support[3]
NGSID 1	Majority C	0.80-0.94	Majority Sub C	22-100	Outlier Sub C	100
NGSID 2	Majority C	0.83-0.95	Majority Sub C	48-100	Outlier Sub C	100
NGSID 3	Majority D	0.93-0.97	Majority Sub D	42-100	Sub D	89
NGSID 5	Majority F1	0.87-0.98	Majority Sub F1	74-100	Sub F1	100
NGSID 6	A1/G/A1/G/H/G	N/A	A1/G/A1/G/H/G	N/A	Unclassified URF	64
NGSID 7	A1/G/A1/G/A1/G	N/A	A1/G/A1/G/A1/G	N/A	Basal to 25_cpx	100
NGSID 8	A1/K/A1/H/A1	N/A	A1/K/A1/H/A1	N/A	Basal to 45_cpx	61
NGSID 10	A1/K/A1/H/A1	N/A	A1/K/A1/H/A1	N/A	Basal to 45_cpx	61
NGSID 12	K/F1/K/F1/K	N/A	K/F1/K/F1/K/G	N/A	Outlier Sub K	93
NGSID 13	Majority J	0.82-0.96	Majority Sub J	84-100	Outlier Sub J	100
NGSID 14	Majority H	0.79-0.98	Majority Sub H	82-100	Sub H	100
NGSID 15	Majority H	0.86-0.99	Majority Sub H	74-100	Sub H	100
NGSID 16	Majority H	0.86-0.96	Majority Sub H	96-100	Outlier Sub H	100
NGSID 18	A1/K/A1/J/A1	N/A	A1/K/A1/J/A1	N/A	Basal to 45_cpx	61

1 – lowest to highest range of similarity; 2 - lowest to highest range of % of permutated trees; 3 – bootstrap support