

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

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

5 Rodgers MA*¹, Wilkerson E^{2,5}, Vallari A¹, McArthur C³, Sthreshley L⁴, Brennan CA¹,
6 Cloherty G¹, de Oliveira T^{2,5,6}

- 7 1. Infectious Disease Research, Abbott Diagnostics, Abbott Park, IL, USA
8 2. Wellcome Trust – Africa Centre for Population Health, University of KwaZulu-Natal, Durban,
9 4041, Republic of South Africa
10 3. School of Dentistry, University of Missouri-Kansas City, Kansas City, MO, USA
11 4. Presbyterian Church (USA), Kinshasa, DRC
12 5. College of Health Sciences, University of KwaZulu-Natal, Durban 4041, Republic of South Africa
13 6. Research Department of Infection, University College London, London WC1E 6BT, United
14 Kingdom

15
16 *Corresponding Author: Mary A Rodgers, mary.rodgers@abbott.com
17 100 Abbott Park Rd
18 Abbott Park, IL 60064
19 Ph 224-668-8936
20 Fax 224-667-1401

21

22

23 **Abstract**

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

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

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

53

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 lineage 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

85 these viruses is likewise complicated by the extreme genetic diversity observed in HIV variants
86 within the region. Subtype-specific differences in treatment effectiveness, the development of
87 resistance, vaccine coverage, and disease progression make surveillance and accurate
88 classification of HIV-1 strains imperative(14-16). To date, most HIV-1 specimens from the DRC
89 have primarily been classified by phylogenetic analysis of short partial genome sequences (400
90 – 900 bp) in either the *group-specific antigen (gag)*, *polymerase (pol)*, or *envelope (env)* genes
91 of HIV-1(4, 11, 13, 17). However, the full extent of recombination and sequence diversity of a
92 complex genome cannot be completely characterized when partial genome sequences are used
93 for HIV-1 classification. As a region with exceptionally high HIV-1 diversity, this is especially true
94 for specimens from the Congo Basin. Advances in availability and reduction in cost for next
95 generation sequencing (NGS) technologies have enabled complete genome sequences to be
96 used for classification of HIV-1 specimens globally. Despite these advances, only 33 complete
97 HIV-1 genomes are currently available from the DRC in the Los Alamos National Laboratories
98 (LANL) repository(18). In contrast, 1217 complete HIV-1 genomes are available from the United
99 States in the LANL database, 1185 of which are classified as subtype B(18). Therefore, in order
100 to fully characterize the true diversity of circulating HIV-1 strains, additional complete genomes
101 of complex variants from the DRC must be sequenced.

102 As a part of ongoing surveillance efforts in Sub-Saharan Africa, we have previously deposited
103 55 complete HIV-1 genomes from cultured virus isolates and Cameroonian patient specimens
104 into the Genbank database(19-27). Recently, we developed a new HIV-primer specific NGS
105 platform to obtain complete genomes from HIV-1 Group M, N, O, and P as well as HIV-2
106 isolates, called HIV-SMART(19). This method utilizes a set of 6 pan-HIV specific primers fused
107 to the SMART (**S**witching **M**echanism at 5' End of **R**NA **T**emplate, Clontech) sequence to create
108 libraries for NGS on the Illumina MiSeq instrument(19). While this method is excellent for
109 recovering genomes from diverse HIV sequences, it has limited success for low viral load
110 clinical specimens (<5 log copies/ml). HIV-SMART was previously optimized for clinical
111 specimens by adding a benzonase pre-treatment of the sample to digest background human
112 DNA and RNA, and a direct correlation was observed between the sample viral load and
113 resulting genome sequence coverage(19). Increasing the total number or concentration of HIV-
114 SMART primers did not have any benefit to genome coverage, however an increase in the
115 reverse transcription (RT) temperature from 42°C to 47°C improved coverage depth while RT at
116 52°C dramatically reduced genome coverage. Alternative library preparation methods or an RT

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

119 In the present study, we have applied both Sanger-based *env* amplicon sequencing and HIV-
120 SMART NGS to a set of HIV-1 variants from the DRC, resulting in 172 new *env* sequences and
121 14 new complete genomes. Modifications to the HIV-SMART method for low viral load samples
122 improved genome coverage to >90% for clinical samples with viral load >4 log copies/ml by
123 adding an input nucleic acid concentration step and lowering the RT temperature to 42°C.
124 Characterization of the complex *env* and complete genome sequences revealed a high level of
125 HIV-1 diversity and recombination in the DRC, and identified sequences that are outliers to
126 known subtype and CRF sequences. Further analyses of the outlier sequences suggest that
127 they may very well be ancestral to some of the major pandemic subtypes today.

128 **Materials and Methods**

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

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

148 shaker at 300 rpm. Liquid was aspirated using a BioTek 405 TS Magnetic Plate washer, and
149 wells were washed with ~250 ul of PBS-Tween20 wash buffer (BioTek, Shoreline, WA). Plates
150 were incubated with 50 ul of 0.4 ug/ml biotinylated goat anti-human IgG (Sigma, St Louis, MO)
151 for 15 minutes at room temperature in a plate shaker. After washing, plates were incubated with
152 50 ul of 0.4 ug/ml Streptavidin, R-phycoerythrin conjugate, SAPE (Invitrogen, Carlsbad, CA) for
153 10 min at room temperature in a plate shaker. After final washes, beads were re-suspended in
154 150 ul of reading buffer (1% BSA in PBS) and analysed on the Luminex FlexMap3D instrument
155 (Luminex Corp. Austin, TX). For each bead set, ~100 events were read and results were
156 expressed as Median Fluorescence Intensity (MFI) per 100 beads.

157 **RNA extraction** Following the serological classification, nucleic acid was extracted from
158 samples according to the manufacturer's instructions using either: (i) the QIAcube blood and
159 body fluid spin protocol (QIAgen) or (ii) the (research use only) total nucleic acid sample
160 preparation protocol on the *m2000sp* system (Abbott Molecular, Des Plaines, IL). For NGS
161 experiments, plasma was pre-treated with benzonase before nucleic acid extraction. One-tenth
162 volume of 10X benzonase buffer (200 mM Tris-Cl pH 7.5, 100 mM NaCl, 20 mM MgCl₂) and 250
163 units/ml of ultra-pure benzonase (Sigma, St Louis MO) were added to 0.9 volumes of plasma to
164 degrade free DNA and RNA(28, 29). Samples were incubated at 37°C for 3 hours then filtered
165 by centrifugation (5000 rpm) through 0.22 µm spin filters (Millipore, Billerica, MA) before
166 extraction. For low viral load samples post-extraction concentration of 25-50 µl of eluted nucleic
167 acid was performed by using a concentrator column kit (Zymo Research) following the
168 manufacturer's instructions.

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

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

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

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

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

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

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

211 compared different reverse transcription temperatures of 47°C and 50°C, as well as the Pico
212 SMART cDNA kit (Clontech), a nucleic acid concentrator step (Zymo Research), and a sizing
213 column purification step (Clontech) to the standard HIV-SMART protocol. NGSID10 (viral load
214 5.01 log₁₀ copies/ml) was diluted 1:100 in normal human plasma for the library preparation
215 optimization experiments to create a 3 log₁₀ copies/ml sample. All libraries were fragmented,
216 barcoded, multiplexed, and sequenced on the MiSeq (Illumina) instrument as previously
217 described(19). For samples with multiple NGS datasets from optimization experiments, all reads
218 from conditions with a reverse transcriptional RT temperature of 42°C were used to generate a
219 consensus genome sequence. NGS data was processed as previously described using CLC
220 Genomics Workbench 8.0 software (CLC Bio/QIAGEN) with minor modification(19). Briefly,
221 fastq data files were imported to CLC, trimmed for quality and ambiguity, and the SMART primer
222 adapter sequence was removed. For optimization experiments, reads were only aligned to the
223 HXB2 reference genome for consistency in making comparisons. For building genomes, reads
224 were aligned to 6-10 subtype and CRF HIV reference genomes, and complete genomes were
225 built by aligning the resulting consensus sequences. The references used for read mapping are
226 summarized in **Supplementary Table 1**. For consensus sequences with gaps, contigs
227 generated by *de novo* assembly were kept if they aligned to the consensus genome and were
228 merged with NGS data in Sequencher 5.2.3 software to create a final consensus sequence.
229 Sanger sequencing were used to fill in the remaining small gaps (<200bp) with the use of primer
230 sequences in regions with full NGS coverage flanking the gap. The raw NGS data was re-
231 aligned to the final genome sequence to generate NGS coverage and read mapping statistics.
232 Open reading frames were confirmed and annotated with SeqBuilder (DNASTAR Laservene v
233 11.2) software. The complete genome sequences have been deposited in Genbank under
234 accession references KY392767 to KY392769 (NGSID1 to NGSID3); KY392770 to KY392773
235 (NGSID5 to NGSID8); KY392774 (NGSID10); KY392775 to KY392779 (NGSID12 to NGSID16);
236 and KY392780 (NGSID18).

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

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

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

252 RDP recombination analyses were followed by manually scanning whole genomes in Simplot v
253 3.5.1(34). Briefly, a subset of 120 full-genome HIV-1 sequences was selected from the 485
254 references. This subset broadly covered the global genetic diversity of HIV-1 group M subtypes,
255 while enriching for genomes from the DRC and other sub-Saharan African countries. A list of
256 the 120 HIV-1 strains is presented in **Supplementary Table 2**. Additional references were
257 added for the analyses of some NGSIDs based on preliminary results of previous analyses. For
258 example for NGSID 7 reference sequences belonging to CRF25 were included.

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

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

271 **Results**

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

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

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

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

295 Eighteen isolates were selected for whole genome sequencing based on the subtype
296 assignments in the *gag*, *pol* and *env* data sets. The HIV viral load for the selected specimens
297 ranged from 3.89-5.82 log₁₀ copies/ml (**Table 2**). The HIV-SMART NGS method(19) developed
298 by our group has previously been shown to generate complete genomes from clinical samples
299 with viral loads greater than 5 log₁₀ copies/ml; however, this method has not been applied to
300 clinical specimens with viral load below this cut off. Therefore, the DRC specimens with viral
301 load <5 log₁₀ copies/ml were used in HIV-SMART NGS optimization experiments for low viral
302 load samples.

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

305 primer-binding specificity. Second, a larger amount of input RNA was used to make cDNA
306 libraries either by concentration of nucleic acid extracts on a concentrator filter column (Zymo
307 Research) or by using a Pico SMART cDNA kit (Clontech). Thirdly, a sizing column (Clontech)
308 was used to select larger amplicons from the total SMART cDNA libraries in a clean-up step.
309 Lastly, the 47°C RT condition was combined with the concentrator column. In the RT
310 temperature comparison, n=8 specimens with viral loads ranging from 3.89-5.2 log₁₀ copies/ml
311 were included, and n=7 had the highest read depth at 50°C and highest % HIV reads at 42°C
312 (**Figure 3**). Notably, the genome coverage was approximately 50% lower in the 50°C libraries
313 compared to the 42°C libraries (**Figure 3**). Therefore, an RT temperature of 42°C was selected
314 for low viral load specimens. In the library preparation optimization experiments, n=4 samples
315 covering a viral load range of 3-5 log₁₀ copies/ml were included. All conditions that included the
316 nucleic acid concentrator step had the highest genome coverage and % HIV reads for
317 specimens with viral load >4 log₁₀ copies/ml (**Figure 4**). For these specimens, raising the RT
318 temperature to 47°C did not affect genome coverage when the concentrator was used.
319 Specimens with viral load <4 log₁₀ copies/ml had inconsistent genome coverage, % HIV reads,
320 and read depth for all of the tested conditions. These results indicated that adding the
321 concentrator step greatly improved the quality and coverage of HIV-SMART genome
322 sequences.

323 The HIV-SMART NGS method with an RT temperature of 42°C and a nucleic acid concentrator
324 step for low viral load specimens resulted in complete genome sequences (>99% coverage) for
325 14 specimens and partial genomes for 4 specimens (**Table 2**). All complete genomes had a
326 length of at least 9550 nucleotides and the average coverage depth was >10 reads for all of the
327 HIV-SMART NGS sequenced regions (**Supplementary Figure 1**). Several short gaps (<200bp)
328 were filled in by Sanger sequencing to complete n=5 genomes. The n=4 partial genomes had
329 coverage ranging from 63%-75% with gaps of various sizes throughout the sequences (**Table 2**
330 **and Supplementary Figure 1**). The 14 complete genomes produced by HIV-SMART NGS
331 were subsequently classified by phylogenetic inference and recombinant analysis.

332 Online subtyping tools were initially used to classify the 14 DRC whole genomes (**Table 3**).
333 REGA identified seven pure viral subtypes including one subtype C (NGSID 2), one subtype D
334 (NGSID 3), one subtype F1 (NGSID 5), one subtype J (NGSID 13) and three subtype H (NGSID
335 14 - 16). The remaining seven whole genome DRC sequences were classified as possible
336 recombinants. Subtyping with jpHMM confirmed the classification of pure viral isolates identified
337 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

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

372 Recombination analyses were performed to investigate putative recombinants within our data
373 set. DRP4 analyses of the eight “pure” subtype isolates identified small possible recombination
374 events, although none of the p-values for these classifications were significant. NGSID 6 was
375 classified as a recombinant form between subtypes A1, G and H. Notably, RDP4 indicated that
376 the subtype H fragment of NGSID 6 was more closely related to the homologous subtype H
377 segments of NGSIDs 8 and 10 than to other H references. In the RDP4 phylogenetic tree
378 inference, NGSIDs 6, 8 and 10 clustered basal to the main subtype H clade, while NGSID 14 –
379 16, which are the three pure subtype H whole genome isolates, clustered within the main
380 subtype H clade. RDP4 classified the majority of the viral backbone of NGSID 7 as belonging to
381 CRF25_cpx with a small recombinant fragment in the *env* region corresponding to subtype A1.

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

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

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

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

414 Discussion

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

431 Appreciation of the genetic diversity observed in sub-genomic sequences prompted our efforts
432 to sequence the complete genomes of the rare variants identified in the *gag*, *pol*, and *env*
433 regions. Unfortunately, many of the rare variants we selected had low viral load upon
434 quantitation, and failed to produce RT-PCR bands for sequencing in other regions of the HIV-1
435 genome. Therefore, to efficiently sequence the 18 selected samples, we followed an NGS
436 method that was optimized for low viral load samples. Improvement of the HIV-SMART NGS
437 method resulted in complete or near complete genome sequences for samples with viral load
438 below the previous threshold of 5 log₁₀ copies/ml. Despite comparing six different conditions to
439 improve genome coverage and read depth, clinical specimens with viral load lower than 4 log₁₀
440 copies/ml were not consistently sequenced by the modified HIV-SMART method. The reduction
441 in reliable complete genome coverage for these low viral load specimens is likely due to poor
442 HIV-SMART primer annealing, extension or a combination of the two. For samples below this
443 threshold viral load, sequencing libraries of purified RT-PCR amplicons may be an alternative
444 method that could reduce background reads and improve read depth. Of the conditions tested,
445 the lowest RT temperature (42°C) and the addition of a nucleic acid concentrator step before
446 library preparation had the greatest improvements in genome coverage and read depth
447 (**Figures 3 and 4**). While higher temperatures dramatically improved read depth in this study,
448 overall genome coverage was reduced by nearly half, suggesting that the reverse transcriptase
449 enzyme was less processive or that the RNA template was degraded at higher temperatures
450 despite expectations that genome coverage would improve due to increased denaturation of
451 RNA template secondary structures⁴⁶ (**Figure 4**). Although the Pico SMART cDNA synthesis kit
452 accommodated a larger volume of input nucleic acid template for library preparation, it did not
453 improve NGS genome coverage or read depth (**Figure 4**). Removal of short transcripts by the
454 use of a sizing column was expected to improve the signal to background ratio for HIV-1 reads,
455 but this step did not affect the percentage of HIV reads (**Figure 4**). In contrast, concentration
456 of the input nucleic acid greatly improved sequencing results (**Figure 4**), which may be due to both
457 the concentration and the purification functions of the columns. The addition of the concentrator
458 column may also improve read depth for high viral load samples, although this has not been
459 tested. The complex genomes from patient samples that were sequenced by the modified HIV-
460 SMART NGS protocol are an excellent example of the utility of this method for surveillance of
461 diverse HIV-1 sequences. The application of HIV-SMART NGS to larger quantities of samples
462 will ultimately bring viral surveillance to the whole genome scale and improve our understanding
463 of the true diversity and evolution of HIV at inter- and intra-patient level. Ultimately, the
464 optimized HIV-SMART NGS method combined with Sanger sequencing of short gap regions

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

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

482 Although it is clear that HIV-1 originated in the Congo Basin(2, 41) we still know little about the
483 early transmission and dissemination of the strains that left the Congo Basin to cause outside
484 epidemics. In order to test the hypothesis that much of the genetic diversity did not leave the
485 Congo Basin, we analyzed all of the public sequences in the LANL database (date of access: 8
486 April 2016). This dataset included all of the HIV-1 group M subtypes and CRF sequences which
487 were >500 bp. In total, we found 411,194 sequences, 7,158 (1.74%) of which were from the
488 Congo Basin. The Congo Basin still contains most of the diversity of HIV-1 in the world(6, 42).
489 There are subtypes that caused large epidemics outside the Congo Basin that are still prevalent
490 in this region, including subtypes A1, D and G (**Figure 8**). There are also subtypes in the Congo
491 Basin that have not caused significant epidemics outside the region, such as subtypes A2, F2,
492 H, J and K. In total, 1160 sequences of these subtypes have been identified to date and 721
493 (62.15%) of these were identified in the Congo Basin. Furthermore, most of the H, J, and K
494 strains that have been identified outside the Congo Basin are from expatriates or from visitors to
495 the region(43-47).

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

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

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

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

536 **Conclusions**

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

547 **FIGURE LEGENDS**

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

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

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

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

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

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

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

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

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

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

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

590 **Figure 6. Bootscan plots of the six recombinant whole genome sequences from the DRC.** Each
591 bootscan plot was performed under the Kimura-2 model of nucleotide substitution with a window size of
592 500 and a step size of 50. Colour coded key represents different subtypes, sub-subtypes or CRFs of HIV-
593 1. The dotted line represents 70% of permuted trees.

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

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

604 **References**

- 605 1. **Hemelaar J.** 2012. The origin and diversity of the HIV-1 pandemic. *Trends Mol Med*
606 **18:**182-192.
- 607 2. **Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S,**
608 **Bhattacharya T.** 2000. Timing the ancestor of the HIV-1 pandemic strains. *Science*
609 **288:**1789-1796.
- 610 3. **Salemi M, Strimmer K, Hall WW, Duffy M, Delaporte E, Mboup S, Peeters M,**
611 **Vandamme AM.** 2001. Dating the common ancestor of SIVcpz and HIV-1 group M and
612 the origin of HIV-1 subtypes using a new method to uncover clock-like molecular
613 evolution. *FASEB J* **15:**276-278.
- 614 4. **Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K, Bunce M,**
615 **Muyembe JJ, Kabongo JM, Kalengayi RM, Van Marck E, Gilbert MT, Wolinsky SM.**
616 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature*
617 **455:**661-664.
- 618 5. **Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, Tatem AJ, Sousa**
619 **JD, Arinaminpathy N, Pepin J, Posada D, Peeters M, Pybus OG, Lemey P.** 2014.
620 HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human
621 populations. *Science* **346:**56-61.

- 622 6. **Hemelaar J, Gouws E, Ghys PD, Osmanov S, Isolation W-UNfH, Characterisation.**
623 2011. Global trends in molecular epidemiology of HIV-1 during 2000-2007. *AIDS* **25**:679-
624 689.
- 625 7. **Laboratories LAN.**
- 626 8. **Bartolo I, Calado R, Borrego P, Leitner T, Taveira N.** 2016. Rare HIV-1 subtype J
627 genomes and a new H/U/CRF02_AG recombinant genome suggests an ancient origin of
628 HIV-1 in Angola. *AIDS Res Hum Retroviruses* doi:10.1089/AID.2016.0084.
- 629 9. **Gao F, Trask SA, Hui H, Mamaeva O, Chen Y, Theodore TS, Foley BT, Korber BT,**
630 **Shaw GM, Hahn BH.** 2001. Molecular characterization of a highly divergent HIV type 1
631 isolate obtained early in the AIDS epidemic from the Democratic Republic of Congo.
632 *AIDS Res Hum Retroviruses* **17**:1217-1222.
- 633 10. **Mokili JL, Rogers M, Carr JK, Simmonds P, Bopopi JM, Foley BT, Korber BT, Bix**
634 **DL, McCutchan FE.** 2002. Identification of a novel clade of human immunodeficiency
635 virus type 1 in Democratic Republic of Congo. *AIDS Res Hum Retroviruses* **18**:817-823.
- 636 11. **Vidal N, Peeters M, Mulanga-Kabeya C, Nzilambi N, Robertson D, Ilunga W, Sema**
637 **H, Tshimanga K, Bongo B, Delaporte E.** 2000. Unprecedented degree of human
638 immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic
639 Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. *J Virol*
640 **74**:10498-10507.
- 641 12. **Tongo M, Dorfman JR, Martin DP.** 2015. High Degree of HIV-1 Group M (HIV-1M)
642 Genetic Diversity within Circulating Recombinant Forms: Insight into the Early Events of
643 HIV-1M Evolution. *J Virol* **90**:2221-2229.
- 644 13. **Mokili JL, Wade CM, Burns SM, Cutting WA, Bopopi JM, Green SD, Peutherer JF,**
645 **Simmonds P.** 1999. Genetic heterogeneity of HIV type 1 subtypes in Kimpese, rural
646 Democratic Republic of Congo. *AIDS Res Hum Retroviruses* **15**:655-664.
- 647 14. **Gaschen B, Taylor J, Yusim K, Foley B, Gao F, Lang D, Novitsky V, Haynes B,**
648 **Hahn BH, Bhattacharya T, Korber B.** 2002. Diversity considerations in HIV-1 vaccine
649 selection. *Science* **296**:2354-2360.
- 650 15. **Taylor BS, Sobieszczyk ME, McCutchan FE, Hammer SM.** 2008. The challenge of
651 HIV-1 subtype diversity. *N Engl J Med* **358**:1590-1602.
- 652 16. **Longosz AF, Morrison CS, Chen PL, Brand HH, Arts E, Nankya I, Salata RA, Quinn**
653 **TC, Eshleman SH, Laeyendecker O.** 2015. Comparison of antibody responses to HIV
654 infection in Ugandan women infected with HIV subtypes A and D. *AIDS Res Hum*
655 *Retroviruses* **31**:421-427.
- 656 17. **Djoko CF, Rimoin AW, Vidal N, Tamoufe U, Wolfe ND, Butel C, LeBreton M, Tshala**
657 **FM, Kayembe PK, Muyembe JJ, Edidi-Basepeo S, Pike BL, Fair JN, Mbacham WF,**
658 **Saylors KE, Mpoudi-Ngole E, Delaporte E, Grillo M, Peeters M.** 2011. High HIV type
659 1 group M pol diversity and low rate of antiretroviral resistance mutations among the
660 uniformed services in Kinshasa, Democratic Republic of the Congo. *AIDS Res Hum*
661 *Retroviruses* **27**:323-329.
- 662 18. **Anonymous.** on Los Alamos National Laboratory. <http://www.hiv.lanl.gov/>. Accessed
- 663 19. **Berg MG, Yamaguchi J, Alessandri-Gradt E, Tell RW, Plantier JC, Brennan CA.**
664 2015. A Pan-Hiv Strategy for Complete Genome Sequencing. *J Clin Microbiol*
665 doi:10.1128/JCM.02479-15.
- 666 20. **Bodelle P, Vallari A, Coffey R, McArthur CP, Beyeme M, Devare SG, Schochetman**
667 **G, Brennan CA.** 2004. Identification and genomic sequence of an HIV type 1 group N
668 isolate from Cameroon. *AIDS Res Hum Retroviruses* **20**:902-908.
- 669 21. **Brennan CA, Bodelle P, Coffey R, Devare SG, Golden A, Hackett J, Jr., Harris B,**
670 **Holzmayr V, Luk KC, Schochetman G, Swanson P, Yamaguchi J, Vallari A,**
671 **Ndembi N, Ngansop C, Makamche F, Mbanya D, Gurtler LG, Zekeng L, Kaptue L.**

- 672 2008. The prevalence of diverse HIV-1 strains was stable in Cameroonian blood donors
673 from 1996 to 2004. *J Acquir Immune Defic Syndr* **49**:432-439.
- 674 22. **Luk K-C, Berg, M.G., Naccache, S.N., Kabre, B., Federman, S., Mbanya, D., Kaptue, L.,**
675 **Chiu, C.Y., Brennan, C.A. and Hackett, J. Jr.** 2015. Utility of Unbiased Next-Generation
676 Sequencing for HIV Surveillance.
- 677 23. **Luk KC, Holzmayer V, Ndembi N, Swanson P, Brennan CA, Ngansop C, Mbanya D,**
678 **Kaptue L, Gurtler L, Devare SG, Hackett J.** 2008. Near full-length genome
679 characterization of an HIV type 1 CRF25_cpx strain from Cameroon. *AIDS Res Hum*
680 *Retroviruses* **24**:1309-1314.
- 681 24. **Vallari A, Bodelle P, Ngansop C, Makamche F, Ndembi N, Mbanya D, Kaptue L,**
682 **Gurtler LG, McArthur CP, Devare SG, Brennan CA.** 2010. Four new HIV-1 group N
683 isolates from Cameroon: Prevalence continues to be low. *AIDS Res Hum Retroviruses*
684 **26**:109-115.
- 685 25. **Yamaguchi J, Coffey R, Vallari A, Ngansop C, Mbanya D, Ndembi N, Kaptue L,**
686 **Gurtler LG, Bodelle P, Schochetman G, Devare SG, Brennan CA.** 2006. Identification
687 of HIV type 1 group N infections in a husband and wife in Cameroon: viral genome
688 sequences provide evidence for horizontal transmission. *AIDS Res Hum Retroviruses*
689 **22**:83-92.
- 690 26. **Yamaguchi J, Badreddine S, Swanson P, Bodelle P, Devare SG, Brennan CA.** 2008.
691 Identification of new CRF43_02G and CRF25_cpx in Saudi Arabia based on full genome
692 sequence analysis of six HIV type 1 isolates. *AIDS Res Hum Retroviruses* **24**:1327-
693 1335.
- 694 27. **Yamaguchi J, Vallari A, Ndembi N, Coffey R, Ngansop C, Mbanya D, Kaptue L,**
695 **Gurtler LG, Devare SG, Brennan CA.** 2008. HIV type 2 intergroup recombinant
696 identified in Cameroon. *AIDS Res Hum Retroviruses* **24**:86-91.
- 697 28. **Anonymous.** Benzonase endonuclease: The Smart Solution for DNA Removal.
- 698 29. **Law J, Jovel J, Patterson J, Ford G, O'Keefe S, Wang W, Meng B, Song D, Zhang**
699 **Y, Tian Z, Wasilenko ST, Rahbari M, Mitchell T, Jordan T, Carpenter E, Mason AL,**
700 **Wong GK.** 2013. Identification of hepatotropic viruses from plasma using deep
701 sequencing: a next generation diagnostic tool. *PLoS One* **8**:e60595.
- 702 30. **Swanson P, Devare SG, Hackett J, Jr.** 2003. Molecular characterization of 39 HIV
703 isolates representing group M (subtypes A-G) and group O: sequence analysis of gag
704 p24, pol integrase, and env gp41. *AIDS Res Hum Retroviruses* **19**:625-629.
- 705 31. **Stamatakis A.** 2014. RAxML version 8: a tool for phylogenetic analysis and post-
706 analysis of large phylogenies. *Bioinformatics* **30**:1312-1313.
- 707 32. **S T.** 1986. Some Probabilistic and Statistical Problems in the Analysis of DNA
708 Sequences., vol 7. American Mathematical Society.
- 709 33. **Jin L, Nei M.** 1990. Limitations of the evolutionary parsimony method of phylogenetic
710 analysis. *Mol Biol Evol* **7**:82-102.
- 711 34. **Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, Ingersoll**
712 **R, Sheppard HW, Ray SC.** 1999. Full-length human immunodeficiency virus type 1
713 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype
714 recombination. *J Virol* **73**:152-160.
- 715 35. **J F.** 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*
716 **39**:783-791.
- 717 36. **Martin D, Rybicki E.** 2000. RDP: detection of recombination amongst aligned
718 sequences. *Bioinformatics* **16**:562-563.
- 719 37. **Padidam M, Sawyer S, Fauquet CM.** 1999. Possible emergence of new geminiviruses
720 by frequent recombination. *Virology* **265**:218-225.
- 721 38. **Posada D, Crandall KA.** 2001. Evaluation of methods for detecting recombination from
722 DNA sequences: computer simulations. *Proc Natl Acad Sci U S A* **98**:13757-13762.

- 723 39. **Smith JM.** 1992. Analyzing the mosaic structure of genes. *J Mol Evol* **34**:126-129.
- 724 40. **Gibbs MJ, Armstrong JS, Gibbs AJ.** 2000. Sister-scanning: a Monte Carlo procedure
725 for assessing signals in recombinant sequences. *Bioinformatics* **16**:573-582.
- 726 41. **Sharp PM, Bailes E, Gao F, Beer BE, Hirsch VM, Hahn BH.** 2000. Origins and
727 evolution of AIDS viruses: estimating the time-scale. *Biochem Soc Trans* **28**:275-282.
- 728 42. **Osmanov S, Pattou C, Walker N, Schwardlander B, Esparza J, Isolation W-UNfH,
729 Characterization.** 2002. Estimated global distribution and regional spread of HIV-1
730 genetic subtypes in the year 2000. *J Acquir Immune Defic Syndr* **29**:184-190.
- 731 43. **Cunningham CK, Chaix ML, Rekacewicz C, Britto P, Rouzioux C, Gelber RD,
732 Dorenbaum A, Delfraissy JF, Bazin B, Mofenson L, Sullivan JL.** 2002. Development
733 of resistance mutations in women receiving standard antiretroviral therapy who received
734 intrapartum nevirapine to prevent perinatal human immunodeficiency virus type 1
735 transmission: a substudy of pediatric AIDS clinical trials group protocol 316. *J Infect Dis*
736 **186**:181-188.
- 737 44. **Laukkanen T, Albert J, Liitsola K, Green SD, Carr JK, Leitner T, McCutchan FE,
738 Salminen MO.** 1999. Virtually full-length sequences of HIV type 1 subtype J reference
739 strains. *AIDS Res Hum Retroviruses* **15**:293-297.
- 740 45. **Sullivan PS, Do AN, Ellenberger D, Pau CP, Paul S, Robbins K, Kalish M, Storck C,
741 Schable CA, Wise H, Tetteh C, Jones JL, McFarland J, Yang C, Lal RB, Ward JW.**
742 2000. Human immunodeficiency virus (HIV) subtype surveillance of African-born
743 persons at risk for group O and group N HIV infections in the United States. *J Infect Dis*
744 **181**:463-469.
- 745 46. **Yerly S, Rickenbach M, Popescu M, Taffe P, Craig C, Perrin L, Swiss HIVCS.** 2001.
746 Drug resistance mutations in HIV-1-infected subjects during protease inhibitor-containing
747 highly active antiretroviral therapy with nelfinavir or indinavir. *Antivir Ther* **6**:185-189.
- 748 47. **Holzmayr V, Aitken C, Skinner C, Ryall L, Devare SG, Hackett J, Jr.** 2009.
749 Characterization of genetically diverse HIV type 1 from a London cohort: near full-length
750 genomic analysis of a subtype H strain. *AIDS Res Hum Retroviruses* **25**:721-726.
- 751 48. **Novitsky V, Wang R, Lagakos S, Essex M.** 2010. HIV-1 Subtype C Phylodynamics in
752 the Global Epidemic. *Viruses* **2**:33-54.
- 753 49. **TenoRes Study G.** 2016. Global epidemiology of drug resistance after failure of WHO
754 recommended first-line regimens for adult HIV-1 infection: a multicentre retrospective
755 cohort study. *Lancet Infect Dis* **16**:565-575.
- 756 50. **Lessells RJ, Katzenstein DK, de Oliveira T.** 2012. Are subtype differences important
757 in HIV drug resistance? *Curr Opin Virol* **2**:636-643.

758

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 RealTime 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)	<i>gag</i> subtype	<i>pol</i> subtype	<i>env</i> 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.

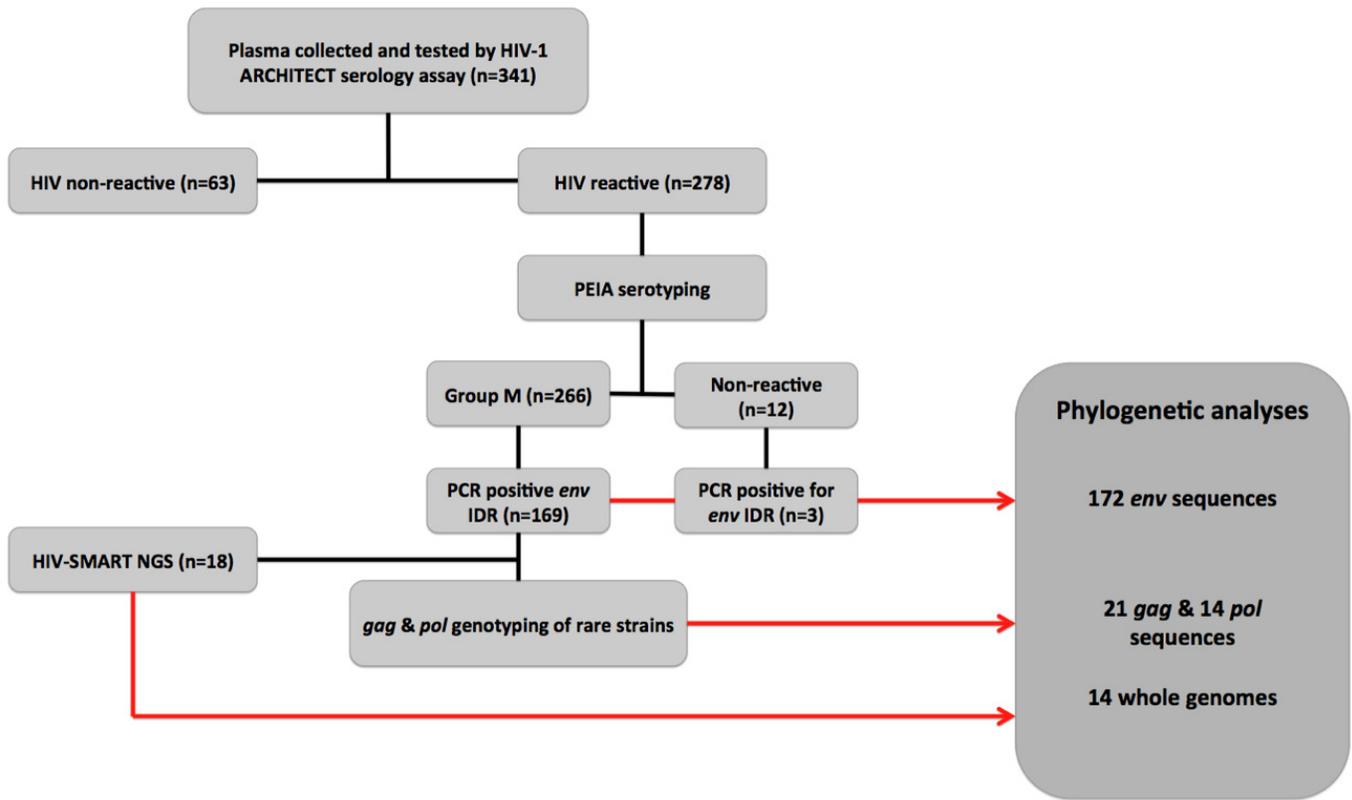
Sample	REGA v 3.0		jpHMM	
	Classification	Support ¹	Classification	Support ²
NGSID_1	Recombinant of C, D	NA	Subtype C	1
NGSID_2	HIV-1 Subtype C	100	Subtype C	1
NGSID_3	HIV-1 Subtype D	100	Subtype D	1
NGSID_5	HIV-1 Subtype F1	100	Subtype F1	1
NGSID_6	Recombinant of G, A1, H	NA	Recombinant of A1, G & H	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

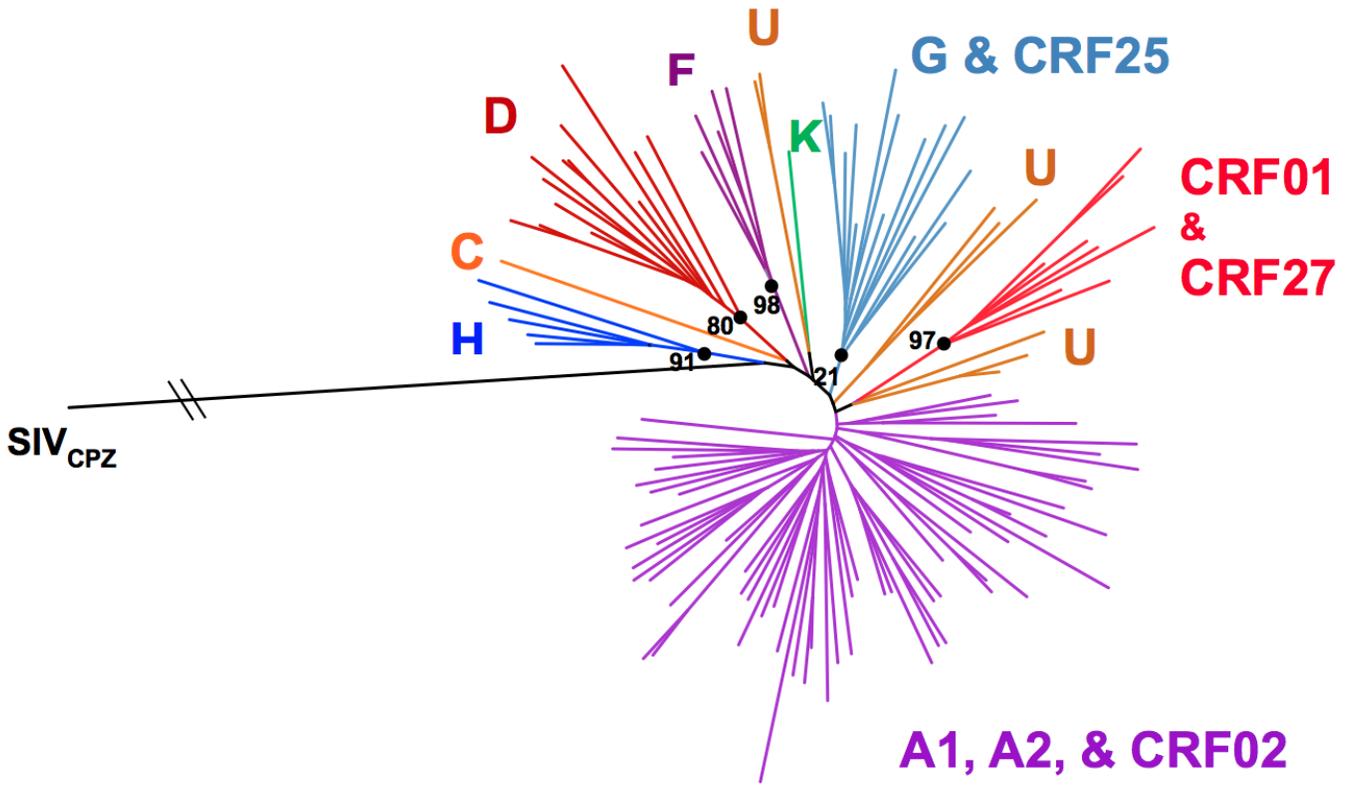
1 – bootstrap support; 2 – posterior support

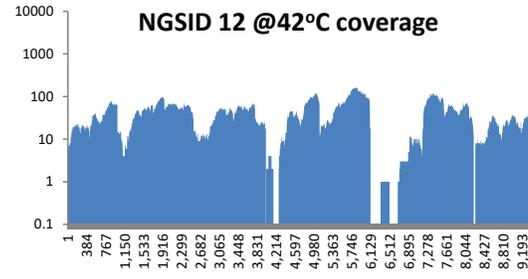
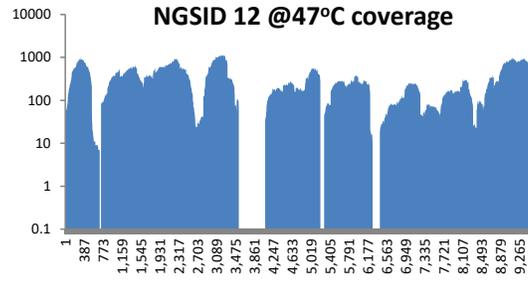
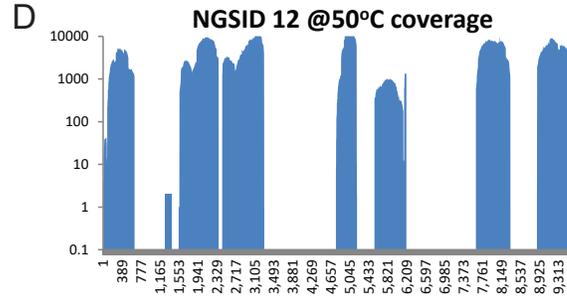
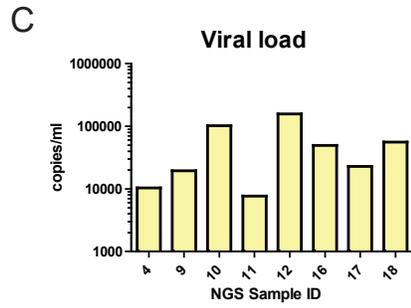
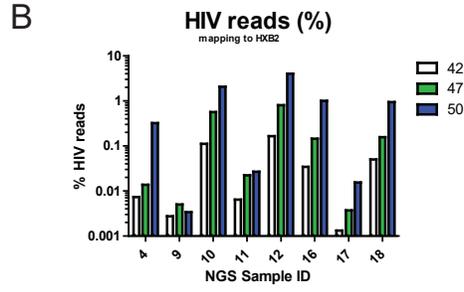
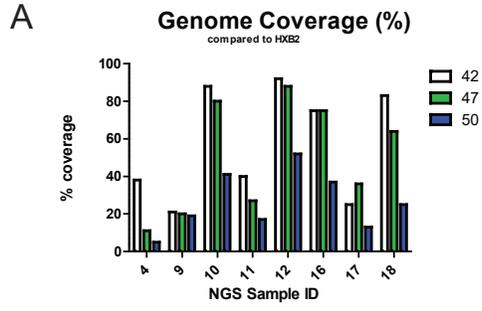
Table 4 - Subtype assignment of the 14 DRC NGS genotypes. This table represents the subtype classification for the 14 NGS genotypes made by Simplot, Bootscan and manual phylogenetic inference.

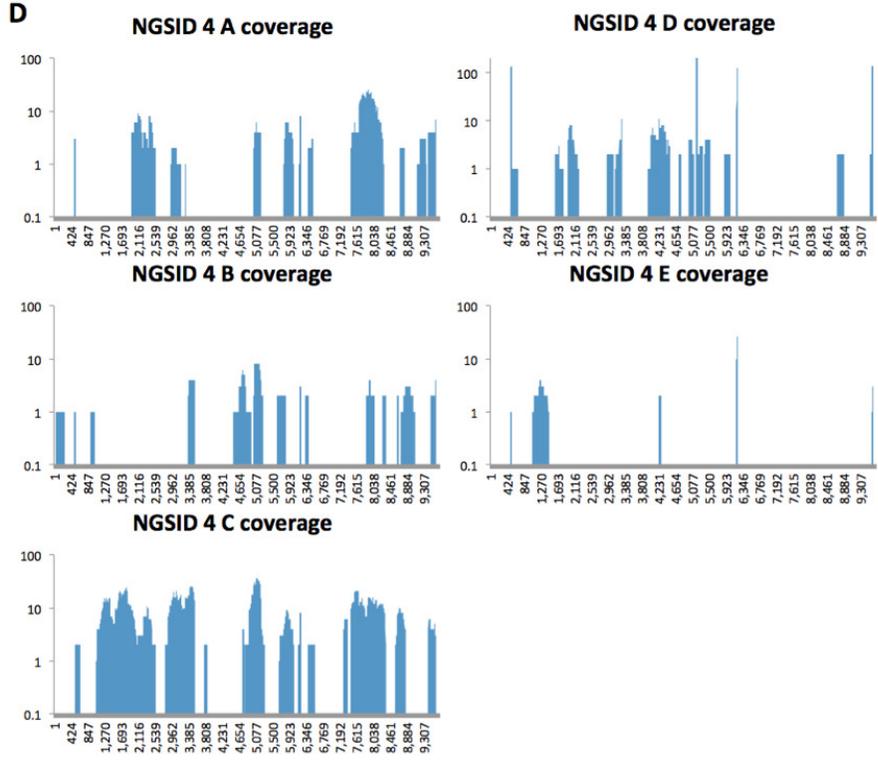
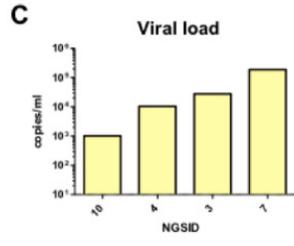
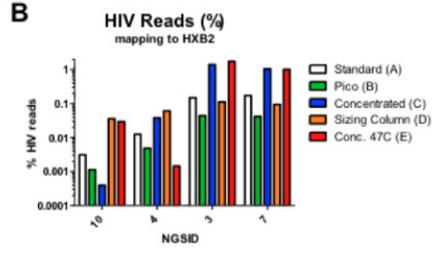
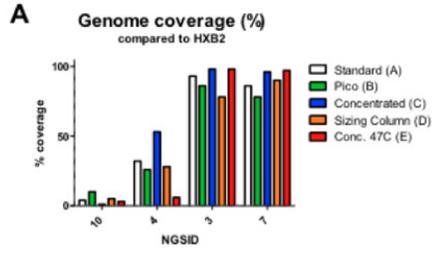
Sample	Simplot		Bootscan		Large phylogeny	
	Classification	Support ¹	Classification	Support ²	Classification	Support ³
NGSID 1	Majority C	0.80-0.94	Majority Sub C	22-100	Outlier Sub C	100
NGSID 2	Majority C	0.83-0.95	Majority Sub C	48-100	Outlier Sub C	100
NGSID 3	Majority D	0.93-0.97	Majority Sub D	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

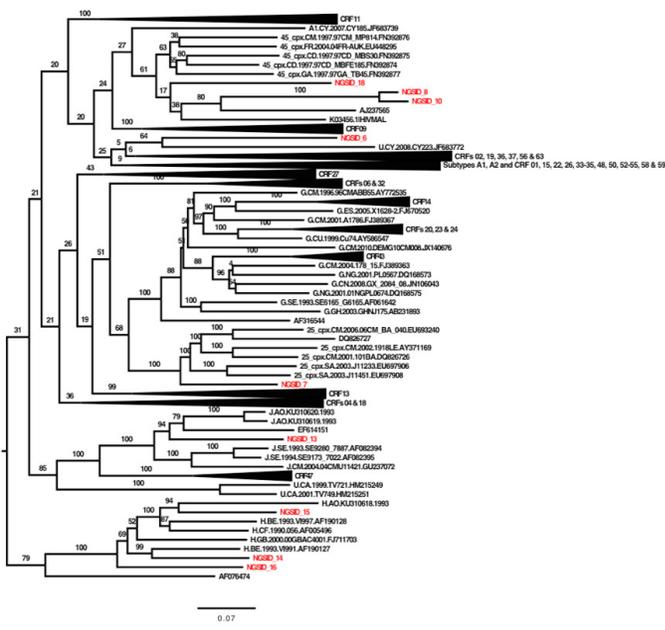
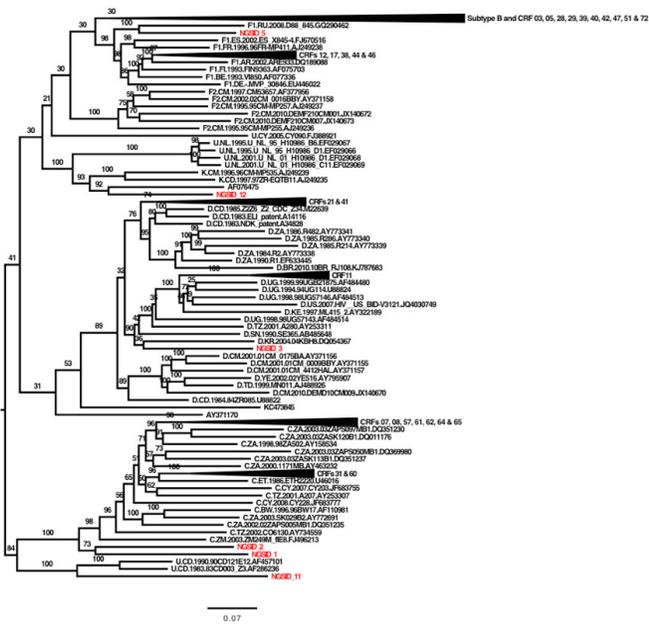
¹ – lowest to highest range of similarity; ² - lowest to highest range of % of permuted trees; ³ – bootstrap support

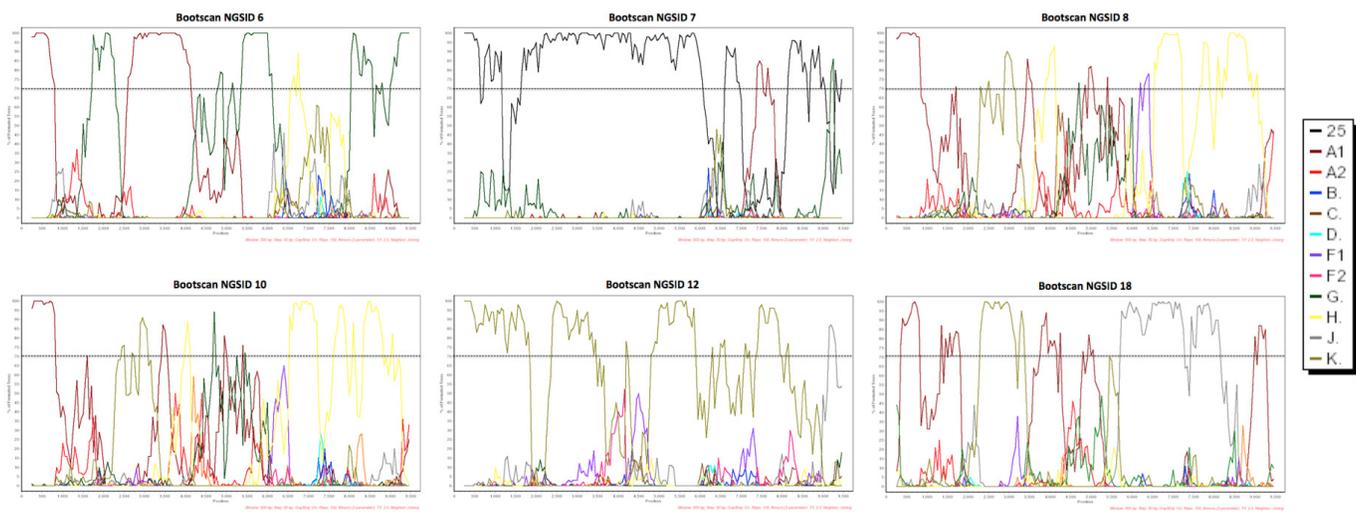


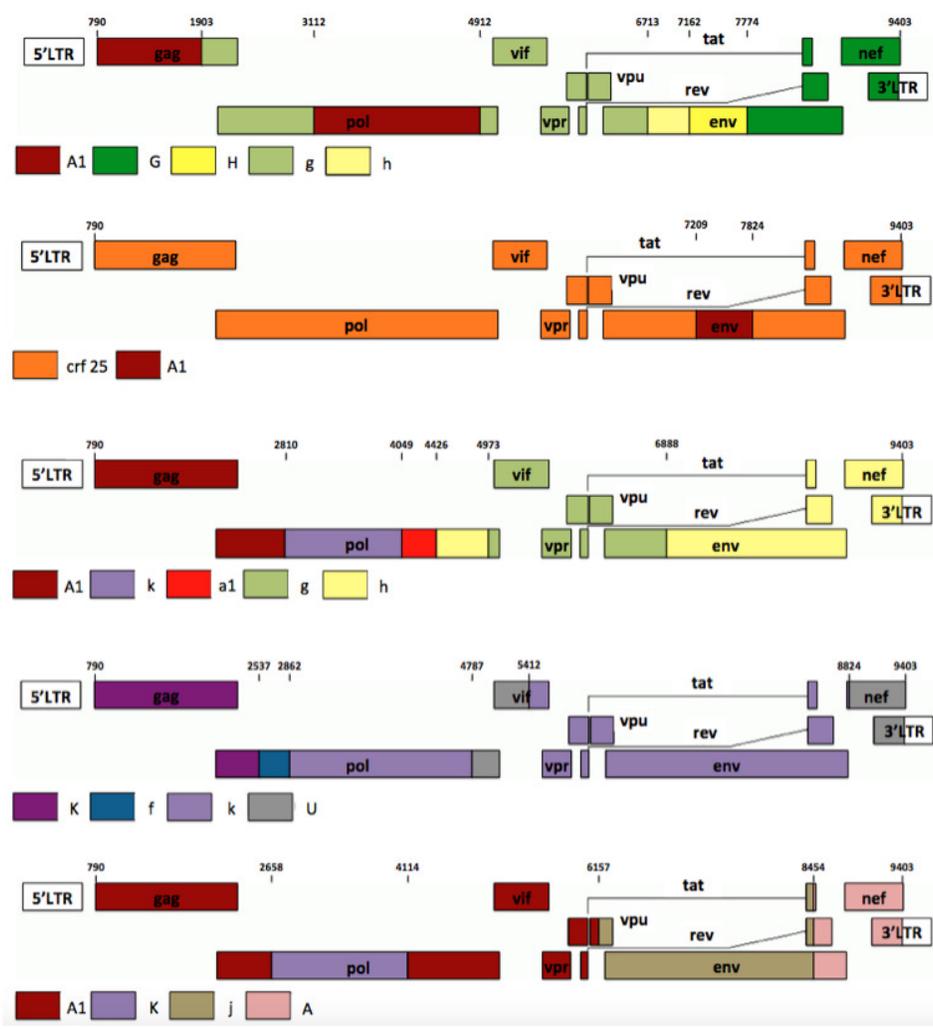












NGSID 6 - A1|g|A1|g|h|H|G

NGSID 7 - crf25|A1|crf25

NGSIDs 8 and 10 - A1|k|a1|h|g|h

NGSID 12 - K|f|k|U|k|U

NGSID 18 - A1|k|A1|j|A

