213

Complete Genomes Reveal Complex HIV-1 Diversity in the Democratic Republic of Congo

Rodgers MA[1]*, Vallari A[1], Harris B[1], McArthur C[2], Sthreshley L[3], Brennan CA[1]

[1]Infectious Diseases Research, Abbott Diagnostics, Abbott Park IL, [2]School of Dentistry, University of Missouri-Kansas City, Kansas City, MO, [3]Presbyterian Church (USA), Kinshasa, DRC

Contact Information:
Mary Rodgers
Abbott Diagnostics
100 Abbott Park Rd
Dept 09NG/Building AP20
Abbott Park IL 60064-6015
Email: mary.rodgers@abbott.com

Abstract

Background: Surveillance of human immunodeficiency virus-1 (HIV-1) strain diversity is fundamental to collective efforts towards the prevention, diagnosis, and treatment of HIV-1 infections globally. However, limited surveillance has been conducted in the Democratic Republic of Congo (DRC), where a high level of strain diversity and intersubtype recombination have been reported. Since partial sequences may underrepresent recombination and diversity, complete genome sequencing is essential to improving HIV surveillance in the DRC. In this study we have sequenced the complete genomes of rare variant HIV-1 specimens from the DRC to meet neglected surveillance needs and to examine HIV-1 diversity.

Methods: HIV-1 specimens were collected at two rural hospitals in the DRC from 2001-2003. Out of 172 HIV-1 specimens classified by phylogenetic analysis of the envelope immunodominant region sequence, 18 rare and unclassified subtypes were selected for next generation sequencing (NGS) using an established HIV-specific primer approach. The viral load for 9 of these specimens was less than 5 $\log_{10}$ copies/ml, and NGS library preparation conditions were further optimized for these specimens. Genomes were assembled by aligning reads to references and *de novo* assembly using CLC Bio software. Strain classification was determined by phylogenetic and recombination analysis.

Results: For low viral load specimens, the highest genome coverage was obtained by concentrating the total nucleic acid extract before reverse transcription at 42ºC using a SMART cDNA kit (Clontech). Genome sequences with >99.5% genome coverage, an average read depth of >10, and a sequence length of >9500 nucleotides were obtained for 14 HIV-1 specimens. The remaining four genomes had coverage of >60%. Phylogenetic and recombination analysis of the 14 complete genomes identified pure subtypes D (n=1), H (n=3), and CRF25 (n=1). The remaining genomes were simple recombinants of a single subtype with unclassified or CRF sequence (n=6), or complex recombinants of 4 or more subtypes, including A, G, H, K, and unclassified (n=3). Two of these complex recombinants shared 97% sequence identity.

Conclusions: The sequence complexity of the reported genomes demonstrates a high level of diversity in HIV-1 strains circulating in the DRC. These complete genomes are a valuable contribution towards HIV-1 surveillance, and the complex recombinants will be useful for modeling the history of HIV-1 in the DRC.

Study Population

Specimens were collected at the Vanga Hospital, Bandundu Province and The Good Shepard Hospital located 12 kilometers from Kananga, Kasia-Occidental Province in the DRC from 2001-2003. The specimens came from voluntary testing and pregnant women participating in a Prevention of Mother To Child Transmission (PMTCT) program. Of a total of 341 specimens, 278 were determined to be HIV-infected and 172 with remaining volume were selected for molecular characterization. Sanger sequencing of the *env* IDR region identified nine different subtypes and 5 CRFs in the population. Subtype A was the most prevalent strain. Rare subtypes H, J, K and L were present at a low frequency.

Figure 1: Map of the Democratic Republic of Congo and study sites



Red circles show regions where the specimens were collected for this study

Rare HIV-1 variants

Since partial sequences may underrepresent recombination and diversity within a population, complete genome sequencing is essential to improving HIV surveillance in the DRC. Therefore, 18 specimens with discordant subtypes between *gag*, *pol*, and *env* sequences, as well as specimens with rare subtypes were selected for further characterization by next generation sequencing to provide full genome sequences.

Table 1: Summary of rare HIV-1 specimen panel

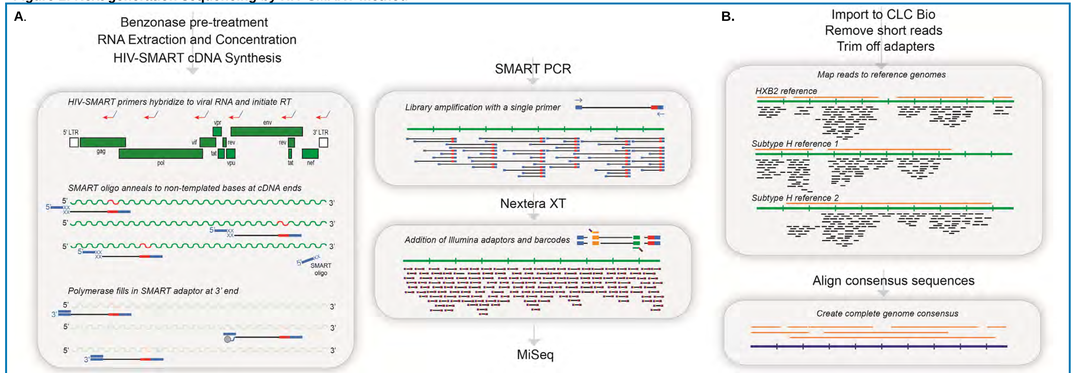
NGS Sample ID	Viral load, (copies/ml)	*gag* Subtype	*pol* Subtype	*env* IDR Subtype	Genome Coverage	Genome Subtype
NGSID1	1.81E+05	-	-	C	100	URF_CU
NGSID2	2.22E+04	C	C	C	100	URF_CU
NGSID3	2.80E+04	-	-	D	100	D
NGSID4	1.04E+04	-	-	F1	75	URF_F1CU
NGSID5	4.44E+04	-	-	F1	100	URF_F1U
NGSID6	2.38E+05	-	-	U	100	URF_AGHU
NGSID7	1.89E+05	-	-	CRF25	100	CRF_25
NGSID8	1.57E+05	PCR Neg	PCR Neg	U	100	URF_AKHU
NGSID9	1.94E+04	A	A	H	72	URF_AH
NGSID10	1.02E+05	A	PCR Neg	H	100	URF_AKHU
NGSID11	7.71E+03	L	L	L	63	L
NGSID12	1.57E+05	K	K	K	99.75	URF_KU
NGSID13	2.61E+05	J	J	J	100	URF_JU
NGSID14	6.56E+05	H	H	H	100	H
NGSID15	1.73E+05	H	H	H	100	H
NGSID16	4.98E+04	H	H	H	100	H
NGSID17	2.28E+04	A	-	H	67	URF_AGUHC
NGSID18	5.59E+04	A3	PCR Neg	J	100	URF_45J

Specimens were selected by sample volume and subtyping of *gag*, *pol*, and *env* IDR sequences generated by Sanger sequencing. The HIV-1 viral load was quantified by the RealTime HIV-1 Assay on the *m*2000 instrument (Abbott Molecular). Complete genome sequencing and subtyping are described in the following sections.

Molecular characterization by next generation sequencing

Next generation sequencing (NGS) was completed by using 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. Libraries were barcoded by a Nextera kit (Illumina) and sequenced as a pooled superlibrary by a MiSeq sequencer (Illumina). Reads were trimmed to remove primer and adapter sequences and aligned to reference HIV genomes using CLC Bio software. Complete genomes were built by aligning 6-10 reference alignment consensus sequences. Protocol optimization indicated that RNA concentration and reverse transcription at 42ºC gave the best genome coverage for low viral load samples. Small gaps were filled in with Sanger sequencing.

Figure 2: Next generation sequencing by HIV-SMART method

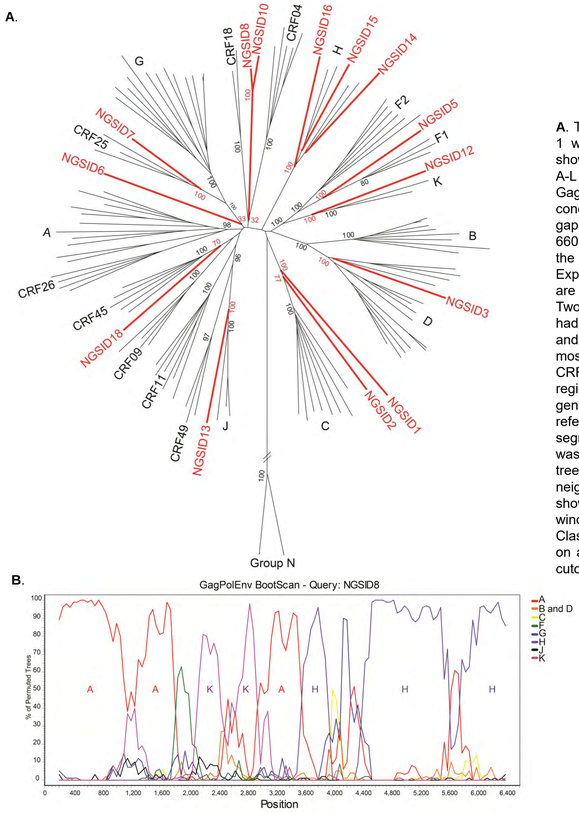


A. Sequencing workflow: HIV genomes from groups M, N, O and P were aligned and conserved regions spaced 1.5-2kb apart were identified. Primers at these locations (red arrows), fused to a common adaptor sequence (SMART, blue bar), were used to reverse transcribe RNA. The SMART sequence was also added to the 3' end of the cDNA for PCR amplification of libraries. Nextera XT was used for multiplexing and sequencing on an Illumina MiSeq. **B**. Data analysis workflow: Illumina paired read data was imported into CLC Bio for read mapping alignment to references. Consensus sequences were aligned to generate complete genomes.
Adapted from *Berg MG et al. J Clin Microbiology. 2015 Dec 23. pii: JCM.02479-15.*

Subtype classification of complete genomes

The *gag*, *pol*, and *env* ORF regions were concatenated for genomes with >99% coverage and were aligned to 322 reference strain *gag/pol/env* sequences, including subtypes A-L and CRF01-72, using the MUSCLE alignment tool in Sequencher. All gaps were removed in BioEdit and a neighbor joining phylogenetic tree was created using MEGA V6.06.

Figure 5: NGS genome phylogenetic tree and recombination analysis



A. The 14 HIV-1 genomes in Table 1 with NGS coverage >99% are shown with references to Group M A-L and closely related CRFs. Gag/Pol/Env ORFs were concatenated, aligned, and gapstripped to a final length of 6603nt and the tree was rooted to the Group N branch in Tree Explorer. Key branchpoint nodes are labeled with bootstrap values. Two of the specimens, 8 and 10, had high similarity to each other and form a unique branch that is most closely related to CRF18 and CRF04 sequences. **B**. To identify regions of recombination, all genomes were compared to reference strains in Simplot. Each segment subtype classification was confirmed by phylogenetic tree analysis. An example neighbor-joining bootscan plot is shown for sample NGSID 8 with windows of 400bp and 50bp steps. Classifications were made based on a bootstrap value of 70 as the cutoff for significance.

Conclusions

- Complete genome sequencing of 18 specimens with viral loads ranging from $7.71x10^3$-$6.56x10^5$ (copies/ml) by the HIV-SMART-NGS method provided 14 new genome sequences for rare variants in the DRC with >99% genome coverage.
- Two of the URF genomes were closely related and formed a unique branch in a neighbor joining phylogenetic tree.
- The sequence complexity of the complete genomes in this study demonstrate the high level of diversity of HIV viruses circulating in the DRC.