HIV Surveillance and Parallel Virome Analysis Using Unbiased Next-Generation Sequencing

Ka-Cheung Luk[1], **Michael G. Berg**[1], Beniwende Kabre[2], Samia Nacacche[2], Scot Federman[2], Charlotte Ngansop[3], Dora Mbanya[3], Lazare Kaptué[3], Catherine Brennan[1], Charles Y. Chiu[2], and John Hackett Jr.[1]
[1]Infectious Diseases Research, Abbott Diagnostics; [2]University of California San Francisco, Virus Discovery & Diagnostics Center; [3]Université de Yaoundé, Cameroon

Abstract

Background: Next-generation sequencing (NGS) has the potential to revolutionize strategies for HIV surveillance, genome characterization, and minor variant analysis. Moreover, application of this technology in an unbiased manner offers the promise of simultaneous detection and characterization of viral co-infections. In this study, we examined the utility of high-throughput sequence analysis for characterization of HIV and detection of known and novel viruses.

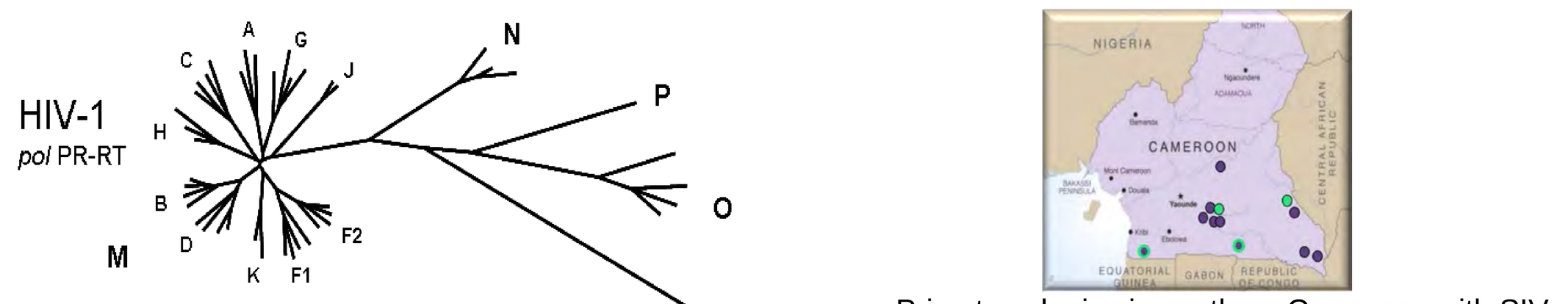
Methodology: Thirty-five plasma specimens from HIV-1-infected Cameroonian blood donors were selected for analysis based on partial genome characterization (*gag* p24, *pol* IN and *env* IDR) using traditional Sanger sequencing. NGS cDNA libraries prepared using nuclease pre-treatment followed by RNA extraction and random-primed reverse transcription were multiplexed (4-5 libraries/run) and run on an Illumina MiSeq^{TM} instrument. Sequence data was analyzed using a customized bioinformatics pipeline for rapid pathogen identification, SURPI (Naccache, et al., *submitted*), to identify and bin all viral sequences. Phylogenetic analysis was performed with PHYLIP and HIV-1 recombinants were characterized using SimPlot.

Results: HIV-1 related sequences ranged from 0.002% to 4.51% of total sequences yielding >90% coverage for 88% of samples. Complete genomes were assembled for 25 strains including 7 CRF11_cpx, 2 CRF13_cpx, 1 CRF18_cpx, and 1 CRF37_cpx, circulating recombinant forms with relatively limited data available as well as 9 unique recombinant forms. Average depth of genome coverage ranged from 121- to 8972-fold. Bioinformatics analysis revealed the presence of numerous other viruses. Notably, GBV-C (HPgV) co-infection was detected in 25% of samples. Seven samples yielded complete GBV-C genomes with read numbers ranging from 0.21% to 5.07%. These GBV-C genomes formed a unique cluster in genotype 1 with Ghanaian sequences distinct from S. African and Ugandan reference strains. Additional detected viruses included hepatitis B virus, adeno-associated virus, hepatitis delta virus, human parvovirus B, human rhinovirus C, and viral sequences with distant homology (<70% aa identity) to small circular DNA viruses (circoviruses, cycloviruses and geminiviruses).

Conclusions: These data demonstrate the utility of unbiased NGS for HIV-1 surveillance, rapid viral genome characterization, and for simultaneous screening of known and potentially novel viruses. Moreover, the extent of HIV-1 genome coverage substantially increases confidence in phylogenetic classification of strains. This technology coupled with powerful bioinformatics represents an unparalleled opportunity to characterize the virome and decipher interrelationships between HIV-1 and other viral infections.

Background

Figure 1. HIV-1 Groups



Primate colonies in southern Cameroon with SIV
SIVcpz ● SIVgor ● both SIVcpz & SIVgor ●

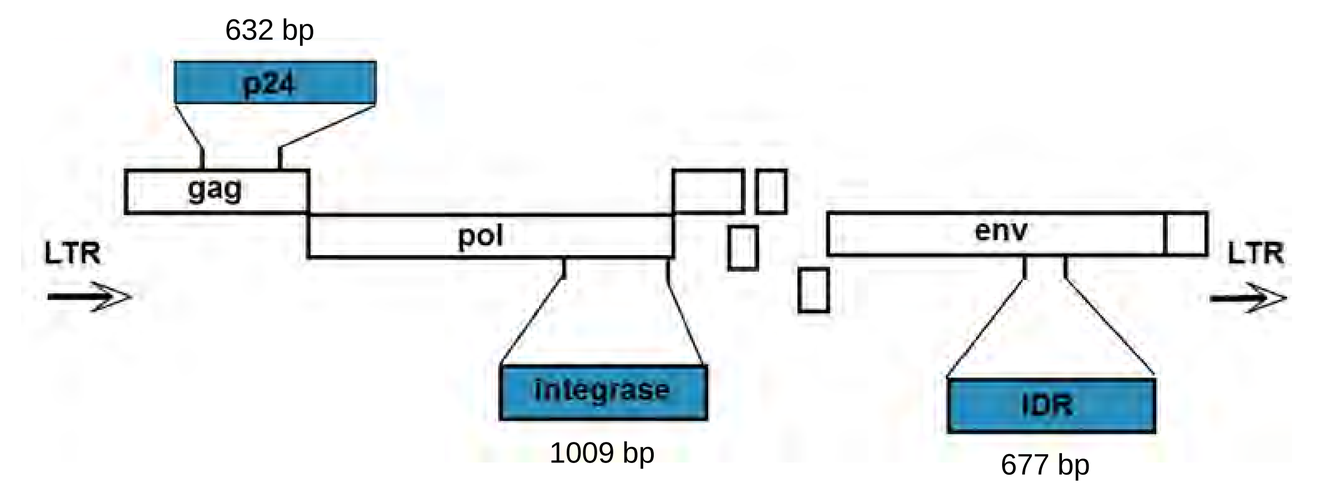
Cameroon : Epicenter of the HIV Epidemic

- Has high genetic diversity and is endemic for rare groups O, N, & P
- High prevalence of recombinant strains
- Proximity to non-human primates

The HIV Global Surveillance:

- Surveillance traditionally conducted by sequencing 3 regions to determine subtype
- Monitor emergence of new viral strains

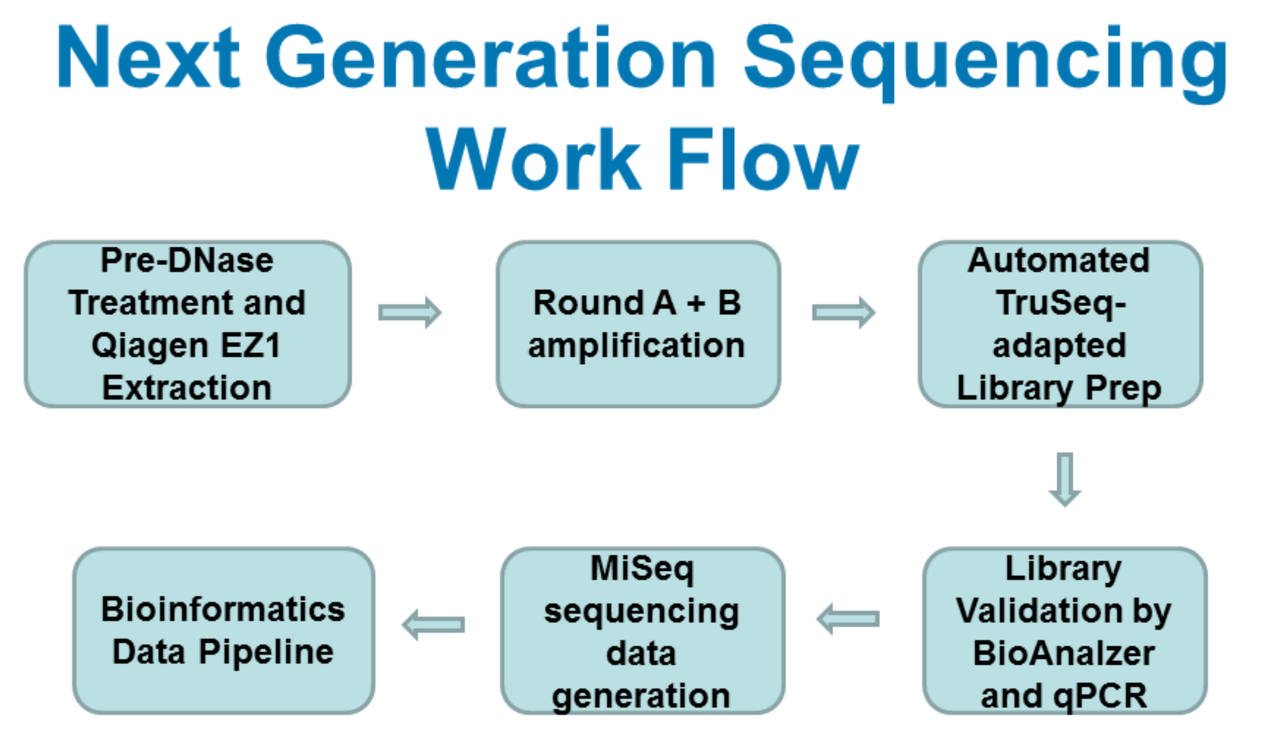
Figure 2. Molecular Surveillance



PCR amplification and sequencing Samples of interest are determined by serological screening. Nucleic acid is extracted from plasma, RNA is reverse transcribed, PCR is performed on 3 regions (blue), and products are Sanger sequenced. Phylogenetic analysis is performed on fragments using PHYLIP.

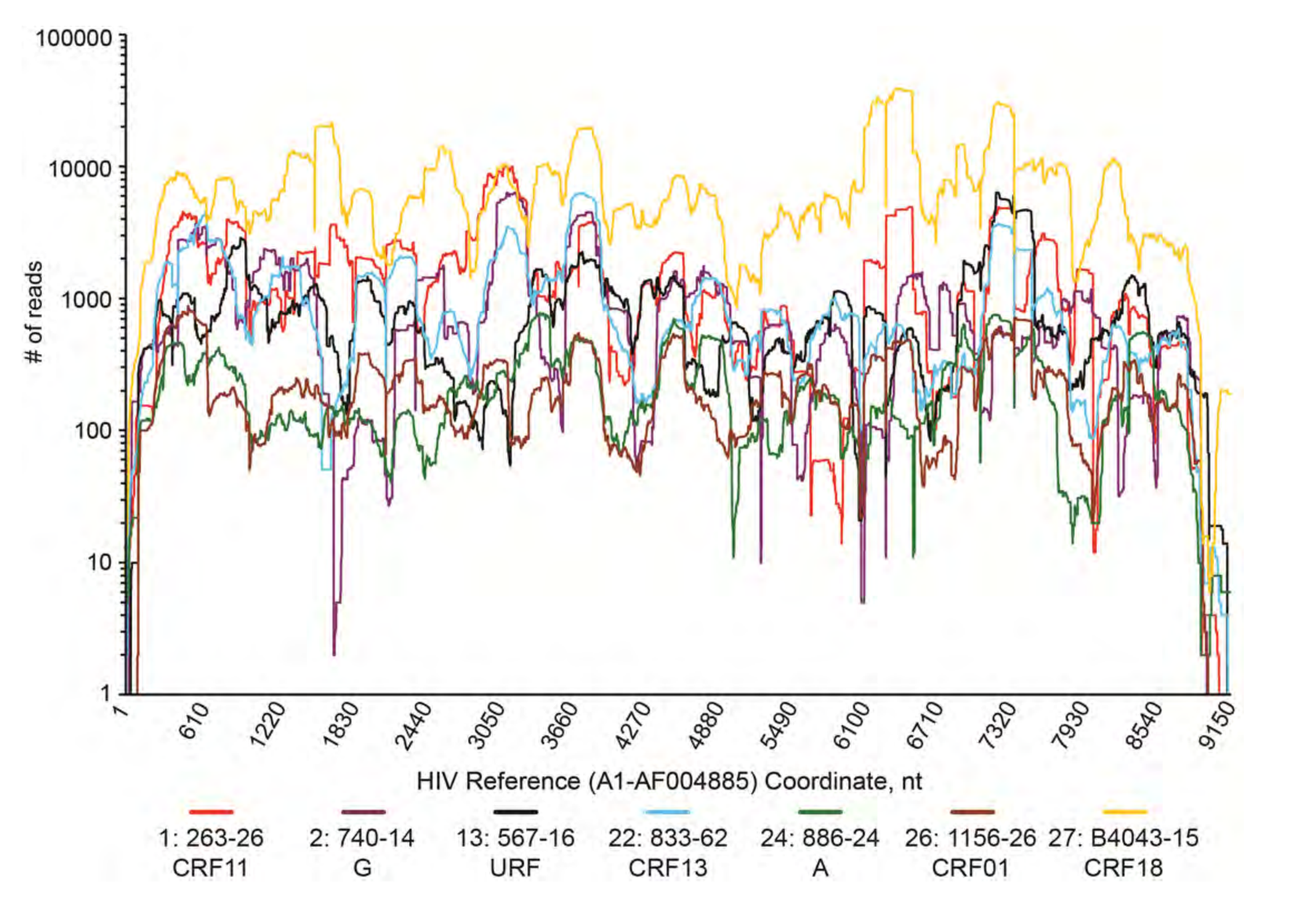
Deep Sequencing of Rare and Emerging HIV Strains

Figure 3. NGS library prep workflow



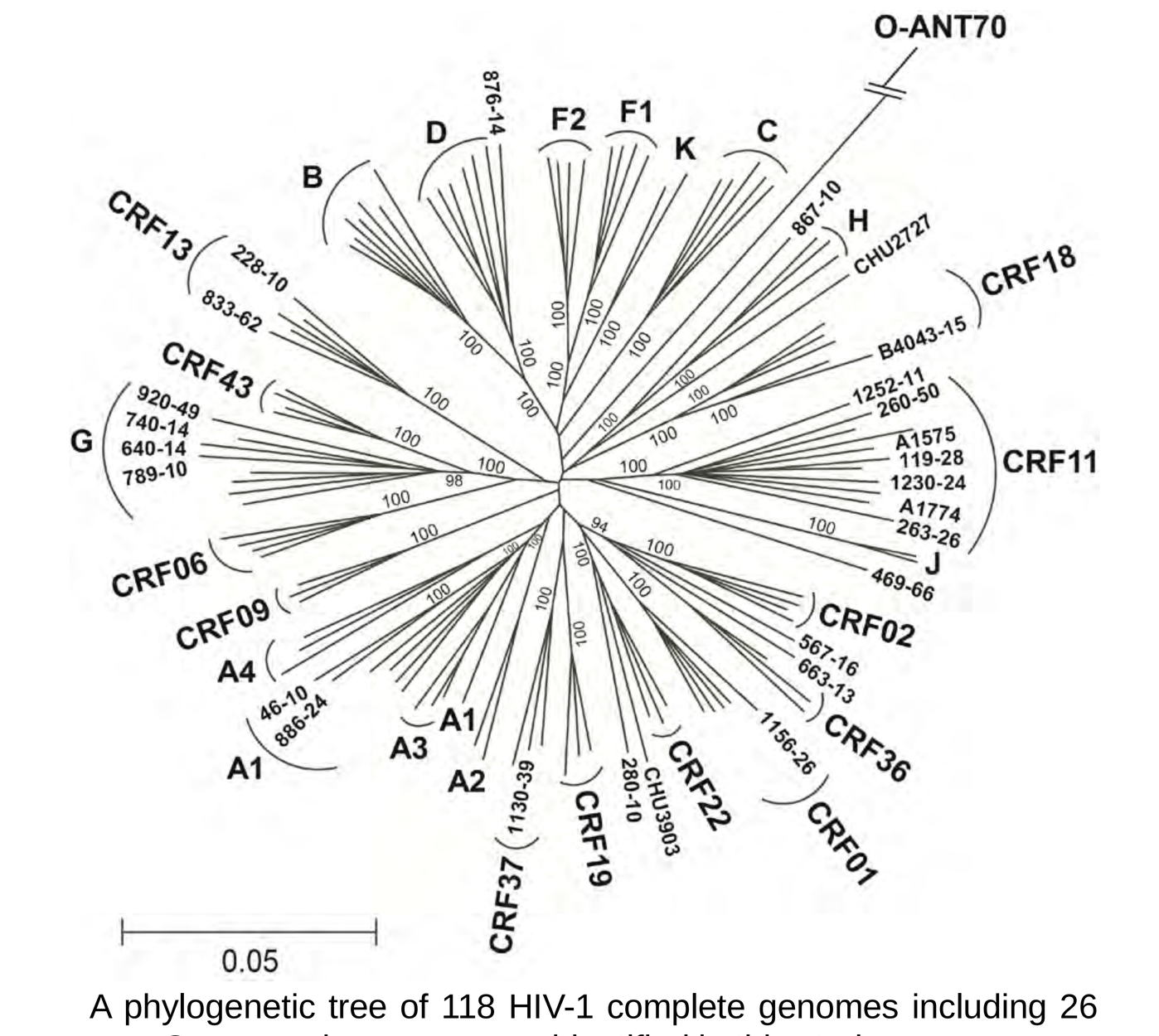
High titer specimens from Cameroon with a diversity of pure subtypes, rare CRFs, and URFs were selected for study. Infected plasma was clarified by centrifugation, pre-treated with a cocktail of nucleases, and extracted robotically using Qiagen EZ1. cDNA was generated by random priming, followed by Round A + B amplification (Chen, EC et al. **2011,** *J Vis Exp*). Deep sequencing Illumina libraries with barcodes for sample multiplexing were made using TruSeq. Libraries were validated by BioAnalyzer and quantified by KAPA qPCR. Five libraries per run were sequenced on a MiSeq and analyzed by a custom bioinformatics pipeline (SURPI: Naccache, et al *submitted*).

Figure 4. NGS sequence coverage and depth



Seven samples representing a range in HIV read depth were aligned to A1-AF004885. Coverage (reads/nucleotide) is plotted along the length of the HIV genome.

Figure 5. Phylogenetic Classification



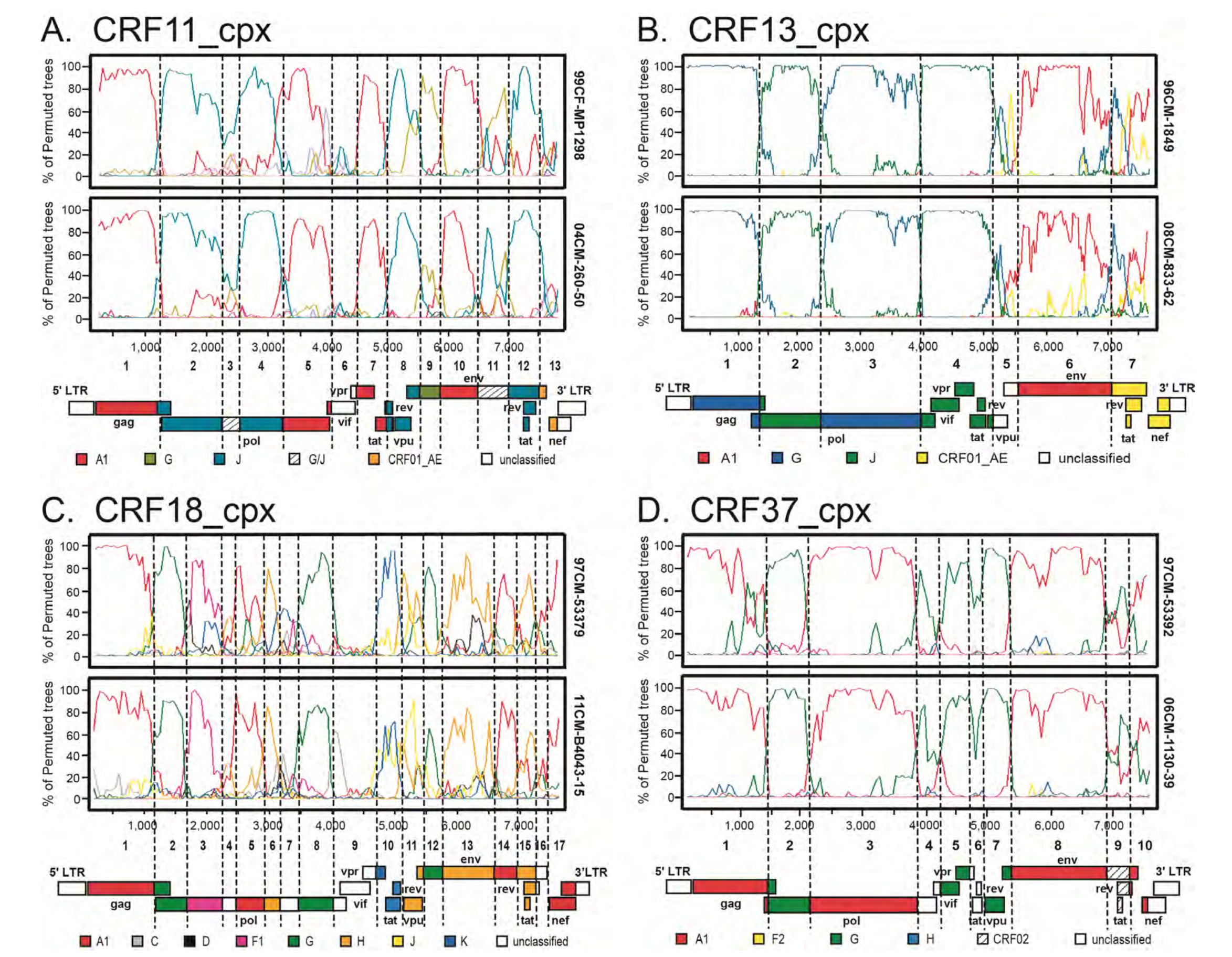
A phylogenetic tree of 118 HIV-1 complete genomes including 26 new Cameroonian sequences identified in this study.

Table 1. Cameroonian Blood Donor Sequence Data

No.	Sample ID	HIV-1 Titer (log_{10} cps/ml)	HIV PCR/Sanger Sequencing: *gag* p24	*pol* IN	*env* IDR	Subtype	NGS: HIV Read Count	HIV Read (%)	Coverage (%)	Subtype	GBV-C Read Count	GBV-C Read (%)	Coverage (%)
1	06CM-263-26	5.23	CRF11	CRF11	CRF11	CRF11	75,767	0.60	100	CRF11	602,896	4.78	100
2	06CM-740-14	5.01	G	G	G	G	39,799	0.29	100	G	629,689	4.58	100
3	06CM-876-14	5.38	D	D	D	D	36,919	0.21	100	URF	0	0	0
4	06CM-1130-39	4.91	CRF37	CRF37	CRF37	CRF37	23,055	0.16	100	CRF37	0	0	0
5	06CM-1225-26	5.36	G	G	G	G	162,114	1.61	92	G	0	0	0
6	06CM-1340-10	5.28	F2	F2	F2	F2	47,823	0.35	90	F2	0	0	0
7	06CM-B460-1	5.53	CRF11	CRF11	CRF11	CRF11	86,590	0.81	82	CRF11	18,251	0.17	87
8	07CM-46-10	5.68	A	A	A	A	32,475	0.61	100	A	0	0	0
9	07CM-62-11	5.00	CRF11	CRF11	CRF11	CRF11	1,689	0.01	98	CRF11	35,516	0.26	100
10	07CM-280-10	5.13	CRF22	A	CRF22	URF	5,903	0.04	100	URF	250,766	1.72	100
11	07CM-419-33	5.55	CRF22/U	A/CRF02	CRF22	URF	381	0.003	78	nd	0	0	0
12	07CM-469-66	5.90	CRF36	CRF11	A/CRF02	URF	10,640	0.09	100	URF	6,430	0.06	99
13	07CM-567-16	5.43	CRF22	CRF22	CRF02	URF	45,625	0.55	100	URF	0	0	0
14	07CM-640-14	5.09	G	G	G	G	17,045	0.10	96	G	0	0	0
15	07CM-663-13	5.15	CRF22	CRF22/02	CRF22	URF	9,576	0.04	100	URF	0	0	0
16	07CM-920-49	5.36	CRF43	G	CRF43	URF	108,797	0.62	100	G	799,793	4.57	100
17	07CM-943-11	4.73	CRF22	CRF22	CRF22	CRF22	24,384	0.15	94	CRF22	0	0	0
18	08CM-38-38	4.83	CRF22	CRF22	CRF22	CRF22	682	0.01	66	nd	0	0	0
19	08CM-228-10	4.94	CRF13	CRF13	CRF13	CRF13	38,043	0.27	100	CRF13	0	0	0
20	08CM-669-39	5.17	CRF22	CRF22	CRF22	CRF22	5,334	0.33	92	CRF22	0	0	0
21	08CM-789-10	5.47	G	G	CRF06	URF	23,226	0.14	100	G	0	0	0
22	08CM-833-62	5.58	CRF13	CRF13	CRF13	CRF13	54,772	1.17	100	CRF13	237,565	5.07	100
23	08CM-867-10	5.10	H	H/A	unclassified	URF	2,196	0.01	100	URF	0	0	0
24	08CM-886-24	4.86	A	A	A	A	10,371	0.21	100	A	0	0	0
25	08CM-1252-11	4.83	CRF11	CRF11	CRF11	CRF11	1,724	0.01	100	URF	0	0	0
26	11CM-1156-26	4.48	CRF01	CRF01	CRF01	CRF01	11,168	0.09	100	CRF01	0	0	0
27	11CM-B4043-15	5.30	CRF18	CRF18	CRF18	CRF18	377,128	4.51	100	CRF18	0	0	0
28	11CM-CHU3903	4.61	CRF22	F2/CRF22	CRF22	URF	4,141	0.05	100	URF	0	0	0
29	11CM-CHU2727	5.46	A	A	H	URF	6,151	0.07	100	URF	18,018	0.21	100
30	11CM-CHU2801	4.47	CRF25	CRF25	CRF25	CRF25	140	0.002	53	nd	0	0	0
31	02CM-A1575	4.45	CRF11	CRF11	CRF11	CRF11	6,451	0.04	100	CRF11	0	0	0
32	02CM-A1774	4.53	CRF11	CRF11	CRF11	CRF11	8,682	0.05	100	CRF11	0	0	0
33	04CM-260-50	5.11	CRF11	CRF11	CRF11	CRF11	6,511	0.06	100	CRF11	0	0	0
34	04CM-119-28	5.20	CRF11	CRF11	CRF11	CRF11	10,089	0.13	100	CRF11	0	0	0
35	04CM-1230-24	5.36	CRF11	CRF11	CRF11	CRF11	72,536	0.86	100	CRF11	0	0	0

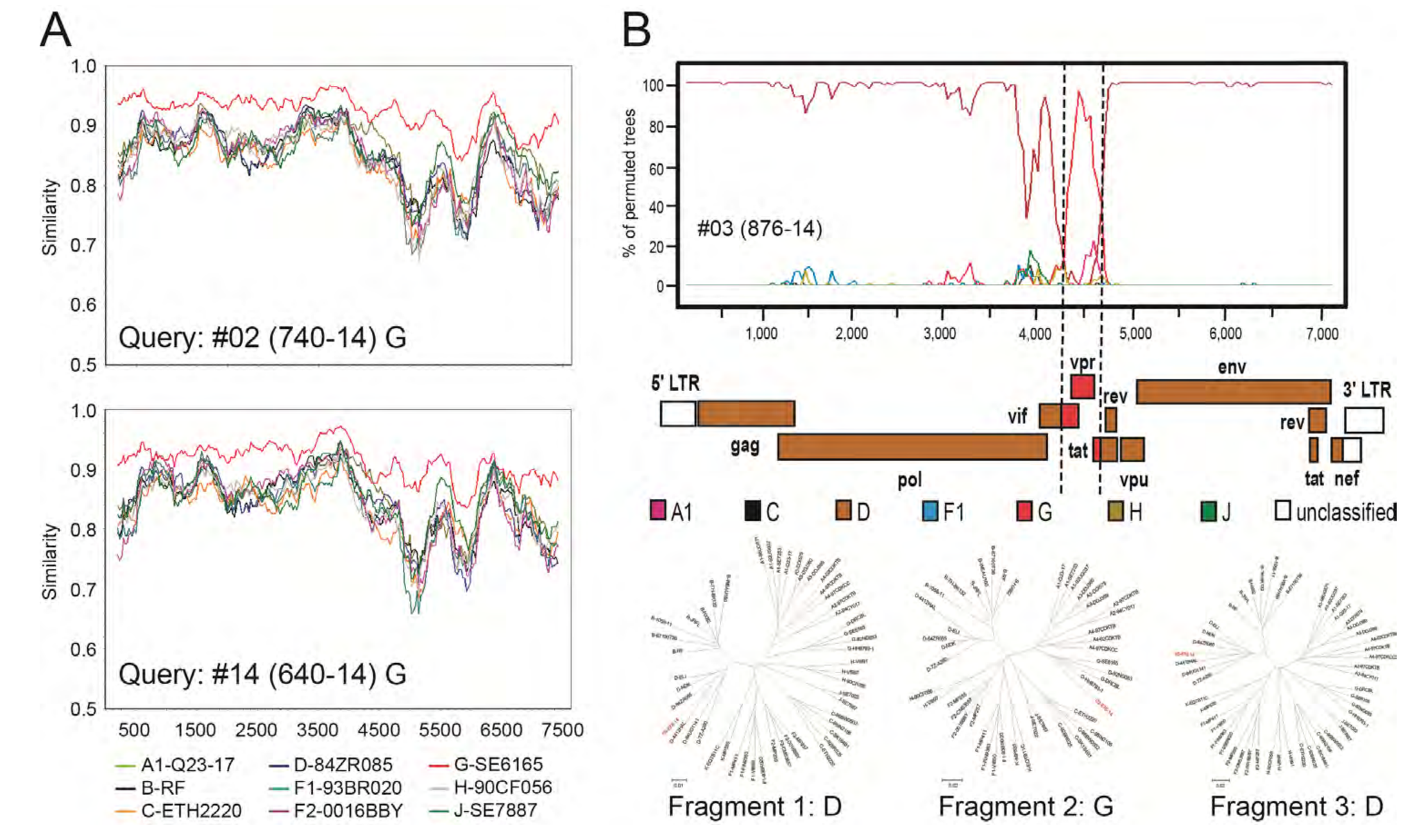
35 Cameroonian blood donors were selected based on viral load and HIV PCR classification for unbiased deep sequencing. Numbers of NGS reads, percentage of total, and sequence coverage are reported for HIV and GBV-C. Discordant HIV subtype classifications are noted in red. nd, not determined

Figure 6. Rare Circulating Recombinant Forms



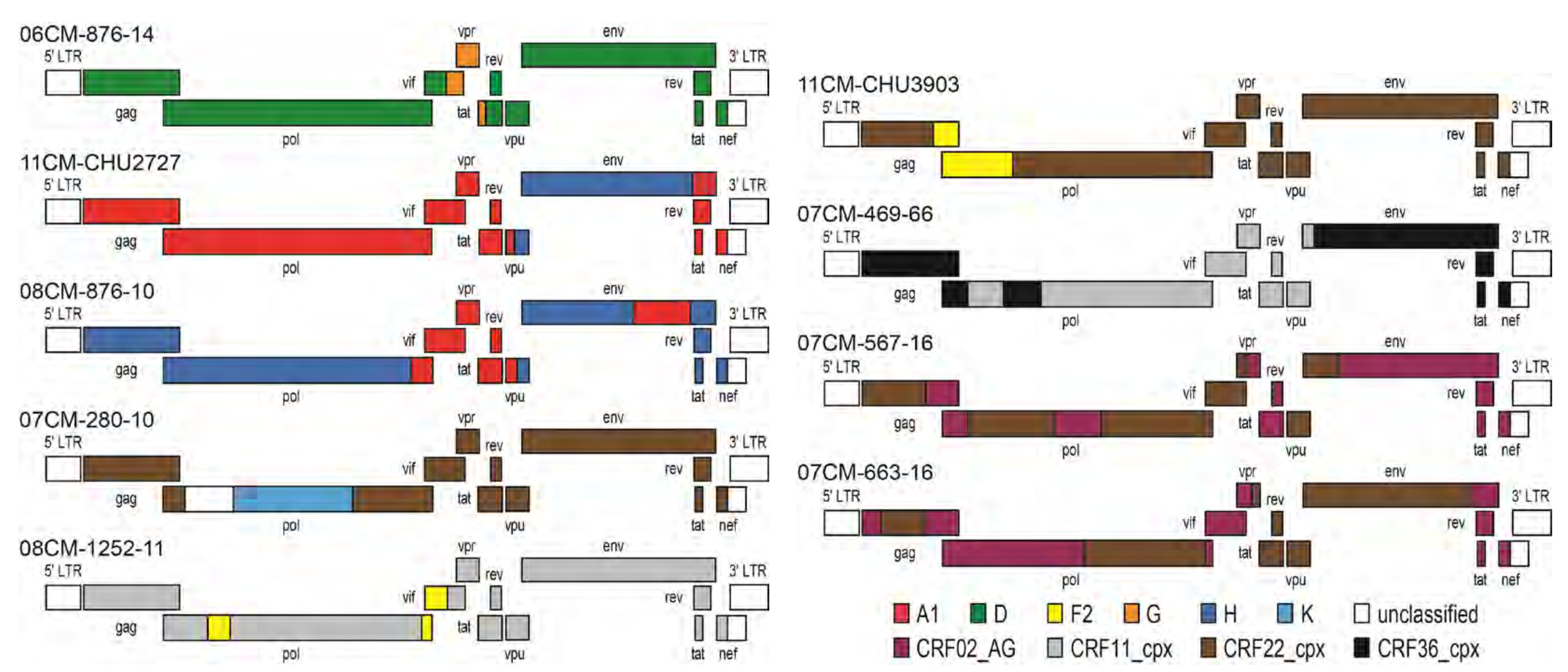
Representative examples of rare CRFs sequenced in this study (lower panels) and reference strains (upper panels) are shown with bootscanning analysis and genomic organizations. Dashed lines indicate recombination break points.

Figure 7. NGS confirms and challenges PCR classification



A. Two pure subtype G samples (red) were confirmed by bootscan analysis in SIMPLOT. **B.** Full genome bootscan of putative pure D reveals recombination in area not probed by PCR. Individual trees confirm presence of subtype G.

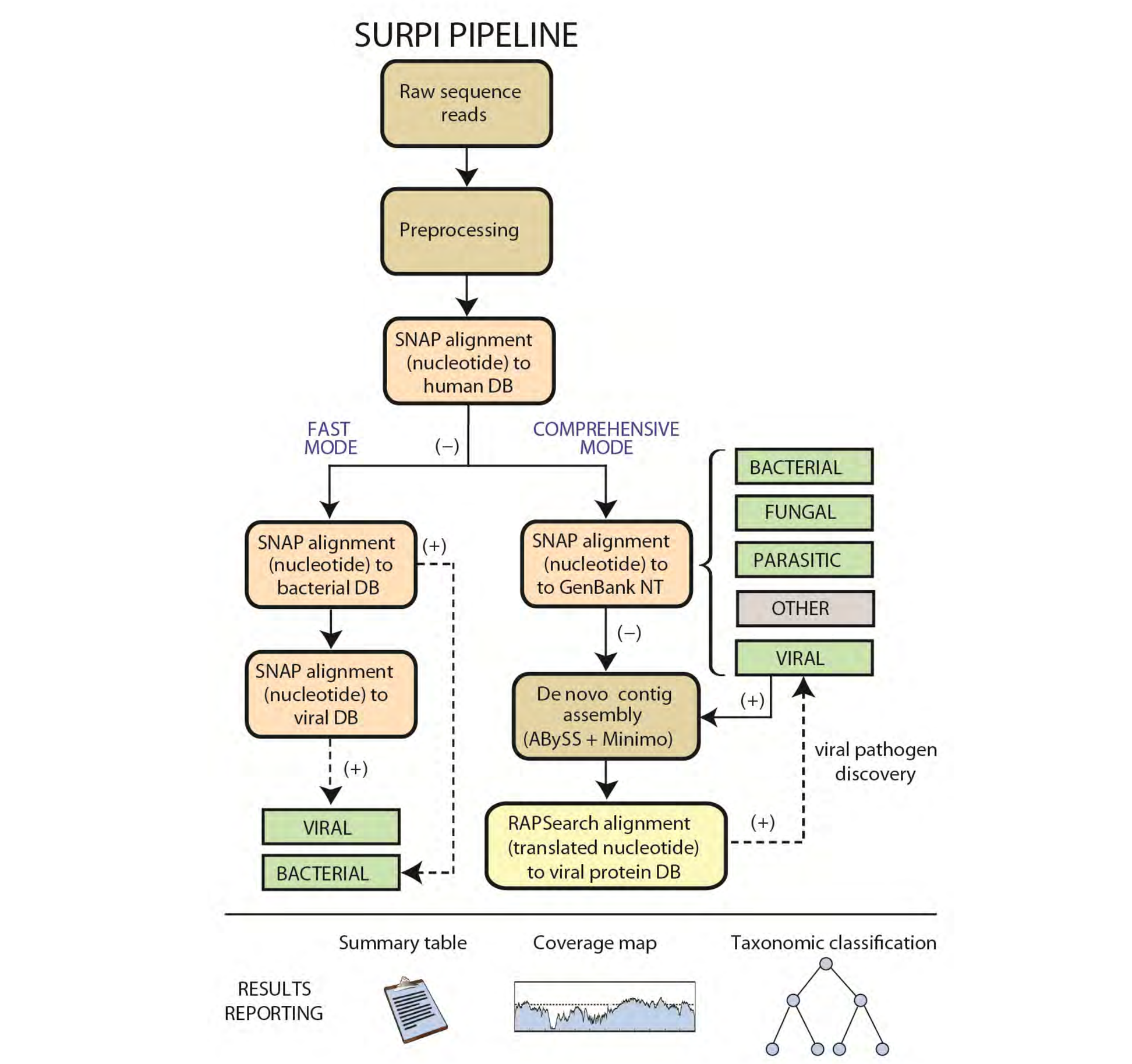
Figure 8. Unique Recombinant Forms



The genomic organization of nine new URFs from Cameroon are reported. Recombination breakpoints were determined in SIMPLOT and fragments were confirmed by individual phylogenetic trees (not shown).

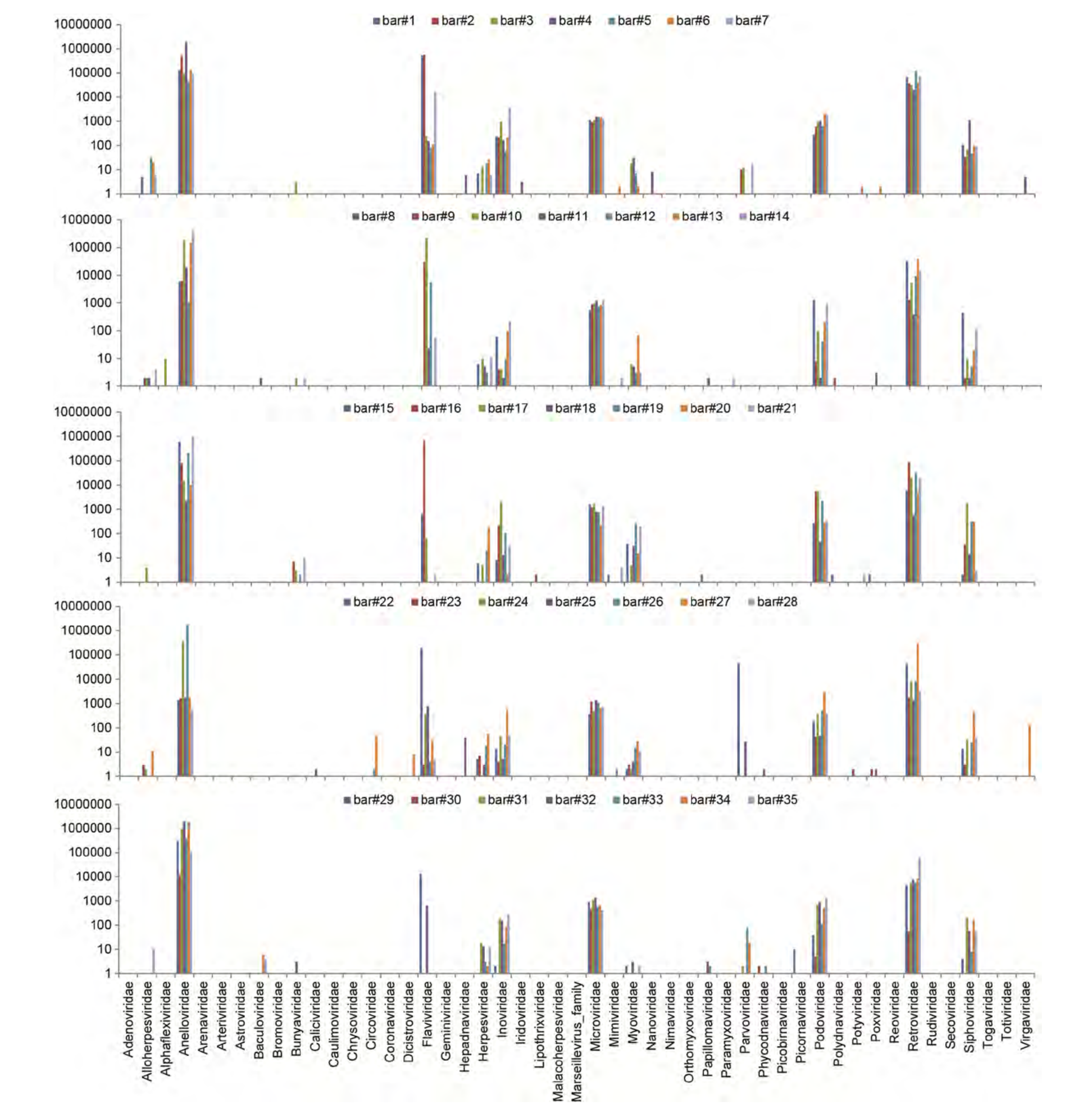
SURPI reveals co-infections with numerous viruses

Figure 9. SURPI Pipeline for Virus Discovery



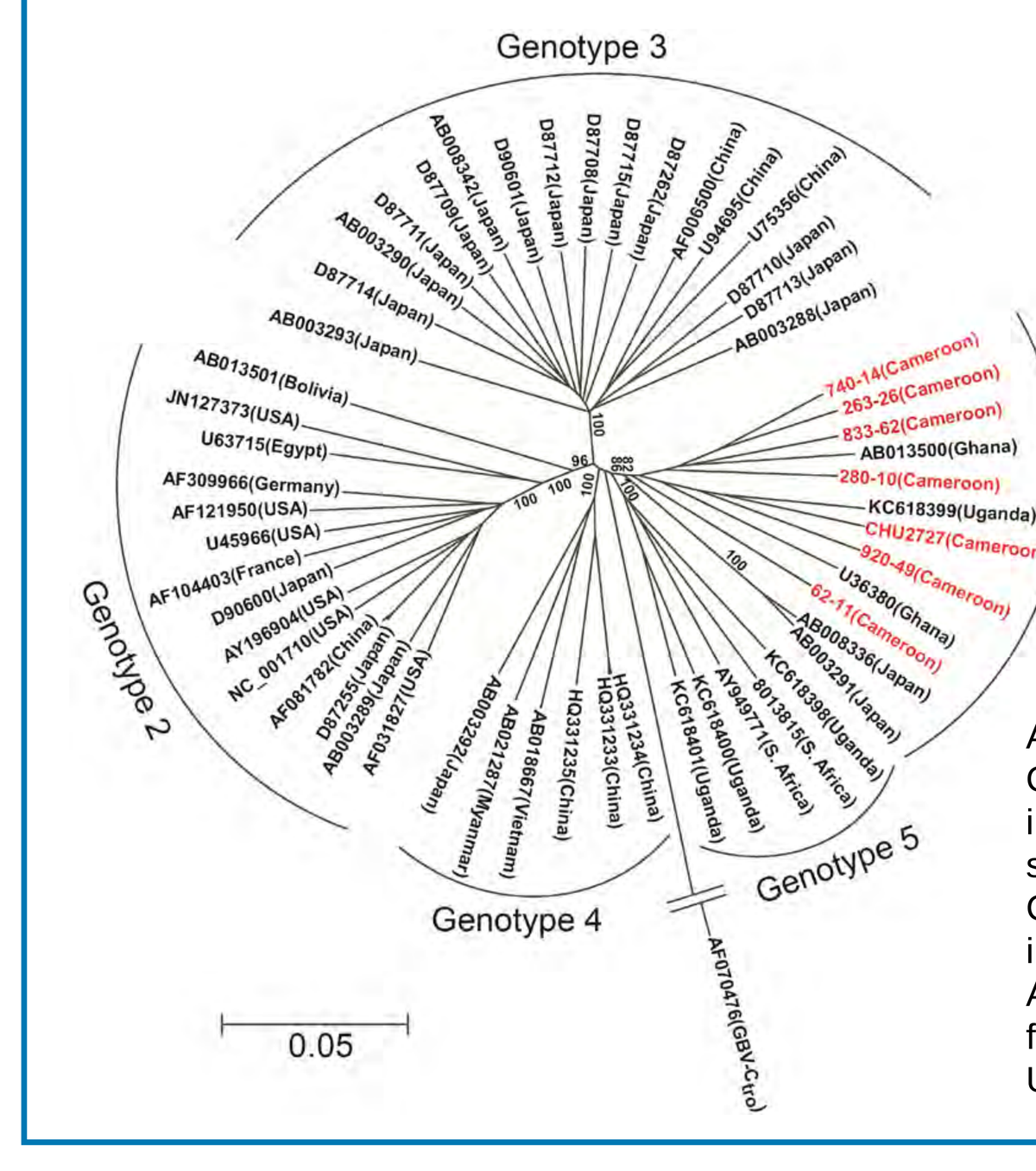
Sequence data was analyzed using a newly developed bioinformatics pipeline for rapid pathogen identification, SURPI. During preprocessing, raw reads have adapters trimmed and low-quality and low-complexity reads are removed. Human reads are computationally subtracted using the SNAP nucleotide aligner. Unmatched reads are then mapped using SNAP in either fast mode (against bacterial and viral databases) or comprehensive mode (against all of GenBank NT), followed by taxonomic categorization of each read. Viral and unmatched reads are *de novo* assembled and aligned to all viral protein sequences using RAPSearch.

Figure 10. SNAP Alignment to Viruses



SURPI results for SNAP alignment to viral sequences in the GenBank NT database. Total reads per taxonomic family are shown. Highly represented families include *Retroviridae* (e.g HIV), *Anelloviridae* (e.g. TTV), phages of the *Inoviridae*, *Microviridae*, *Podoviridae*, and *Siphoviridae* families presumably corresponding to environmental or laboratory bacterial contaminants, and *Herpesviridae*, common, often asymptomatic infections. Of particular interest is *Flaviviridae*, for which all were GB virus-C (GBV-C; see Table 1). Additional detected viruses included hepatitis B virus, adeno-associated virus, hepatitis delta virus, human parvovirus B, human rhinovirus C (not expected in blood), and viral sequences with distant homology (<70% aa identity) to small circular DNA viruses (circoviruses, cycloviruses, and geminiviruses).

Figure 11. Cameroonian GBV-C groups with Genotype 1



A phylogenetic tree of 55 GBV-C complete genome sequences including 8 Cameroonian strains sequenced by NGS (red). Cameroonian sequences group in genotype 1 with other West African references, and away from South African and Ugandan strains.

Conclusions

- Our workflow allowed for HIV whole-genome coverage for the majority of specimens sequenced.
- NGS confirmed 89% of subtype classifications determined by 3-region PCR and revealed additional recombination events.
- Cameroon remains a hotbed for HIV diversity. 23 full genomes reported include rare CRFs and URFs.
- Deep sequencing of infected samples provides a wealth of information: full genome coverage, the ability to look for quasi-species, minor variants, and dual infections of HIV, while also probing for the presence of other viruses.
- The SURPI bioinformatics pipeline can rapidly characterize the virome present in clinical samples. While the vast majority of viruses identified are non-pathogenic, some samples were co-infected with known disease-causing agents. RAPSearch and *de novo* assembly permits new virus discovery.
- 25% of samples were co-infected with GBV-C. Phylogenetic analysis of new full genomes indicates high diversity among these isolates and places them in genotype 1.
- The data support the use of NGS as a new, comprehensive mode for conducting viral surveillance.