Novel Amplicons based Method for Near-Full-Length Genome (NFLG) Sequencing of HIV-1 Group M Recombinant Forms

Roy Moscona[1,3], Tali Wagner[1], Miranda Geva[1], Efrat Bucris[1], Oran Erster[1], Neta S. Zuckerman[1] and Orna Mor[1,2]

[1]Central Virology Laboratory, Israel Ministry of Health, Sheba Medical Center, Tel – Ha shomer, Israel
[2]Department of Epidemiology and Preventive Medicine, School of Public Health, Faculty of Medicine, Tel-Aviv University, Tel-Aviv, Israel
[3]Division of Infectious Diseases, Cincinnati Children's Hospital, Cincinnati, Ohio, USA

Abstract

Over the years the spread of HIV-1 across the globe resulted in the creation of multiple subtypes and new recombinant forms (RFs). While the pol gene region of the HIV-1 genome is used for resistance mutations analysis and initial detection of recombinants, whole genome sequencing analysis is required to determine recombination events across the viral genome. Here, we present a novel robust near-full length genome (NFLG) sequencing approach for the sequencing of HIV-1 genomes out of clinical whole blood samples. This method has been successfully tested for various HIV-1 subtypes and RFs.

The method is based on an in-house developed set of 32 pan-genotypic primer pairs, divided into two pools, each containing 16 primer pairs covering the entire HIV-1 genome. Two parallel multiplex PCR reactions were used to generate 32 overlapping DNA fragments spanning the HIV-1 genome. Nextera XT protocol was used to obtain barcoded libraries which were run on the Illumina Miseq platform using a V3 kit. A consensus sequence was determined for each sample and was used to define recombination events across the genome. For that aim, a combined analysis of several computational tools including HIV BLAST, phylogenetic analysis, RIP, SimPlot++ and jpHMM was employed. Overall, plasma samples from 33 patients suspected to carry RFs and 2 different, known pure subtypes controls, were included in this study.

Overall coverage varied between different RFs. While the gag and pol genes were highly covered in all samples analyzed, the highly variable env gene region was not fully covered. Yet, these NFLG analyses enabled the identification of recombination events across the viral genome, with the identification of CRFs and RFs such as CRF01_AE, CRF02_AG, BF1, BC, A6B, CA1 and some potential URFs.

In summary, we describe a methodology for HIV-1 NFLG sequencing, which is based on multiple and partially overlapping PCR fragments spanning the genome of HIV-1. This method enables sequencing of 20 different patient samples in a single MiSeq run and was used for the characterization of different HIV-1 RFs and pure subtypes circulating in Israel. Additionally, this newly refined method was compared to HIV-1 NFLG results of PCR-free metagenomic sequencing approach and proved to obtain higher number of covered bases of the HXB2 reference genome.

1. Introduction

The global HIV-1 pandemic can be attributed to HIV-1 group M (major). Over the years, the spread of the virus in the human population generated an increase in viral genetic diversity, creating a split of the group M into several different viral subtypes (A–D, F–H, J, K, L). The leading factors for viral diversity are the high mutation rate of the virus and the recombination events, including intra and inter subtypes, which lead to the creation of new recombinant forms (RFs) [1], [2]. Circulating recombinant forms (CRFs), are RFs identified in at least three epidemiologically nonrelated individuals, while unique recombinant forms (URFs), which are patient specific.

Since sequencing of the pol gene of the HIV-1 genome (mainly the protease {PR}, reverse transcriptase {RT}, and integrase {INT} genes) is still highly recommended before starting anti-retroviral treatment (ART) treatment, many studies utilize these sequences to define RFs [3], [4]. However, regions other than the pol gene are not routinely included. To enable characterization of whole HIV-1 genome recombinants in blood samples, full or near full-length genome (NFLG) sequencing methodology could be utilized. Commonly, PCR amplified DNA fragments, of varying sizes and lengths spanning across the viral genome are being produced. These amplicons serve as input for a selected sequencing approach, either Sanger or Next-Generation Sequencing (NGS). Alternatively, metagenomics, which does not require amplification of the viral genome, could be used. However, the latter demands high resources per sample and is not feasible for large scale clinical settings. Previously, several different methods for HIV-1 NFLG sequencing were developed, some also capture RFs [5]–[8].

Here, we present a novel robust, relatively simple approach, for a pan HIV-1 NFLG sequencing protocol based on multiple short PCR products. This method was developed to capture the clinically relevant HIV-1 group M pure subtypes and CRFs circulating in Israel and in other developed countries. To validate our novel method, we compared our NFLG results to the sequencing results of a PCR-free metagenomic method.

2. Methods

Sample selection

In a previous study by the research team of the National HIV reference lab, HIV-1 CRFs and URFs in Israel were defined, based on HIV-1 PR and RT partial sequences for people living with HIV-1 (PLWH) diagnosed between 2010 and 2018 [3]. Here, samples from 35 patients suspected to represent different CRFs and URFs, were used for the assessment of a new NFLG sequencing method.

Selection of primers for multiplex PCR

The LANAL database (https://www.hiv.lanl.gov/components/sequence/HIV/search/search.comp) was searched for full HIV-1 sequences (sequence length >= 9000 bp) that were sampled between 2000-2019 aiming to generate a diverse input dataset for primer selection . To narrow down the search results, only HIV-1 subtypes and RFs known to be circulating in Israel were selected. These included different group M HIV-1 pure subtypes (B, C, A1, D, G, A6, F1, H, U) and RFs (AE, AG, CRF_206, BC, A6B, A5U) and complex forms (cpx) (06, 45, 27).
The query resulted in 291 HIV-1 genome sequences, of which 118 representative sequences for each HIV-1 pure subtype, RF, cpx (above-mentioned) were selected. This dataset was further divided into two sets of sequences: a training set and a testing set. The training set contained newer isolates, collected between 2003-2019 (n=76), while the testing data set contained older sequences isolated between 2000-2002 (n=42). The training dataset served as template for designing the primers, while the testing dataset was utilized virtually, for in silico PCR analysis, to ensure the newly designed primers were suitable, before synthesizing the primers for in vitro multiplex-PCR.
All HIV-1 sequences were aligned to the HXB2 reference genome by Geneious Prime (Dotmatics, Boston, MA, version 2022.2.2, Build 2022-08-18) default mapper using default settings. The primal-scheme [9] tool (https://primalscheme.com/) was used to identify conserved regions across the HIV-1 genome in a sliding windows of ~400bp (Fig. 1). Primers were constructed to enable amplification of the whole viral genome

in two separate PCR reactions divided into two pools of overlapping fragments, varying in length with an average fragment size of 368bp. The primal-scheme primers design was further edited using Geneious Prime, to incorporate ambiguous nucleotides and adjust the genomic location of each primer, while maintaining an appropriate distance and overlap between the DNA fragments. The two pools of primers and their locations in HXB2 are presented in Supp. Table 1.

HIV-1 amplification and DNA Library preparation for NGS

Following extraction of HIV-1 RNA using Nuclisense protocol (BioMerieux), complementary DNA (cDNA) was synthesized using SuperScript III first strand kit (Thermo Fisher Scientific, Waltham, MA). Two multiplex-PCR reactions were performed in parallel in two test tubes, each containing a different primer pool (pool 1 or pool 2). Next, the two pools were combined in equal molarity and diluted as required according to the NexteraXT protocol (Illumina, San Diego, CA).

Metagenomics Virome sequencing

The same HIV-RNA extractions from four selected samples were used for metagenomics sequencing, with cDNA synthesized using SuperScript III first strand kit (Thermo Fisher Scientific, Waltham, MA). Followed by, SMARTer Stranded RNA-Seq Kit (Takara Bio, San Jose, CA), that was used to generate un-amplified cDNA library.

The resultant libraries were purified with AMPureXP beads (Beckman Coulter, Brea, CA). Library concentration was measured by Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, Waltham, MA). Library validation and mean fragment size (~300bp) was measured by TapeStation4200 (Agilent, Santa Clara, CA) with DNA High-Sensitivity D1000 (Agilent, Santa Clara, CA). Following dilution to 4nM equal molar quantities of the libraries were pooled, denatured and diluted to final concentration of 15pM and sequenced (2X100bp) either on NovaSeq (Illumina, San Diego, CA), for metagenomics, or MiSeq (Illumina, San Diego, CA), for the new, amplicons based, methodology.

Bioinformatics analysis

FastQ sequences were subjected to a general pipeline used to generate a consensus sequence for each sample as described (https://github.com/NetaZuckerman/UPv/blob/main/pipelines/generalPipeline.py).

Analyses of HIV-1 recombination forms with phylogeny, HIV-BLAST, RIP, SimPlot++ and jpHMM

Phylogenetic trees were constructed with neighbor joining with a 1000 bootstrap for branch assessments. Four different phylogenetic analyses were performed. First, a phylogenetic tree based on NFLG sequences was built. This phylogeny approach enabled the characterization of the dominant viral clade of recombination (major parent). Next, an additional three independent phylogenetic trees were constructed, one for each viral gene: gag, pol, env. This approach was chosen to identify specific recombination events in the three different coding regions aiming to identify minor parents. All phylogenetic trees were constructed with known pure subtypes and RFs as reference sequences. This enabled the classification of each sample to an appropriate HIV-1 viral clade.
HIV-BLAST was used to assess the similarity to known public repository sequences. Furthermore, Simplot++ was used to identify genome wide recombination events [10]. Simplot++ is utilizing similarity plots to determine the probability of the recombination event and its breaking points across the HIV-1 genome. Simplot++ was used in BootScan mode with 100 bootstraps on a 400 bp windows with 20 bp step size with the TN93 model for substitutions.
Additionally, jumping profile hidden Markov model (jpHMM) [11] and recombinant identification program (RIP available at www.hiv.lanl.gov/content/sequence/RIP/RIP.html) methods were used with default parameters. Final characterization of each CRF or URF was based on a combinatory analysis of the results of all tools, while considering each tool's advantages and drawbacks in resolving recombination events across the viral genome.

3. Results

Comparison between metagenomics and the novel amplicons method for NFLG sequencing

To compare the sequencing results obtained by the newly developed amplicons method to un-amplified metagenomics virome sequencing, four samples were selected. While the average total number of reads per sample obtained by the metagenomics approach was 66,178,474 (range: 57,842,929-71,825,644) reads, only an average of 59,560 reads (0.09% of total reads) were mapped to the HIV-1 HXB2 reference genome. On the contrary, for the novel amplicons NFLG method, only 2,588,962 (range: 1,254,560-3,099,939) total reads were obtained per sample, however on average 2,191,779 reads (84.6%, of total reads) were mapped to the HIV-1 HXB2 reference genome. Comparison of the coverage (>5 reads per position in both methods) across HIV-1 HXB2 reference genome between the metagenomics and the NFLG amplicons methodology, revealed that the amplicons displayed a higher total coverage when directly compared. The two sequencing methods produced identical results, while surprisingly the amplicons managed to reach greater coverage and depth, sequencing new positions across the HXB2 genome, which were not reported by metagenomics (Fig. 2). Generally, among the different HIV-1 subtypes, metagenomics obtained an average coverage of 85.4% while the newly developed amplicons approach covered 90.6% of the HXB2 reference genome bases. The difference in coverage observed between samples could be attributed to diversity in viral clades. For the coverage of pure subtype B was 97.27% vs 95.38%, amplicons vs metagenomics, respectively, (sample 99666, Fig. 2A), while the coverage of pure subtype C (sample 100545, Fig. 2B) was lower (88.73% vs 81.94%) and with the same range for nonpure subtypes sample 84808 (Fig. 2C, 89.3% vs 74.6%) and sample 79455 (87.65% vs 73.77% Fig. 2D). No discrepancy was noticed between the consensus's sequences of metagenomics and NFLG alignment in all samples.

In silico down sampling scaling to resolve the minimal number of reads required for maximal coverage by the novel amplicons method

The coverage of bases of the HIV-1 reference genome HXB2 varied between different HIV-1 subtypes. Thus, to enable a unify criteria for a pan-HIV-1 settings, a calculation was made to determine the optimal number of reads required for maximal coverage of the 9,719 bases of the HXB2 reference genome, while sequencing the maximal number of samples in parallel on a single run of a Miseq V3 kit. For the down-sampling analysis only 10 samples were sequenced in parallel, and each sample received an even number of reads. We concluded that in the tradeoff between number of reads and coverage of HXB2 bases the required number of paired-end reads is between 500K-750K reads, as there was little increase in coverage beyond 750K reads (Fig. 3). Beyond 750K reads, additional reads did not increase the coverage significantly. Hence, we concluded that one Miseq V3 kit was sufficient to run 20 samples in parallel, while keeping adequate coverage and maximal usage of samples sequenced per kit.

Identification of RFs based in the NFLG sequences

To resolve the recombination events across HIV-1 NFLG sequences for each sample, several analytic tools were practiced. The tools applied here to identify RFs were phylogenic analyses, HIV-BLAST, SIMPLOT++, LANL RIP, jpHMM [3], [8], [10], [11]. The subtyping tools displayed a general strong agreement between themselves, identifying a similar if not identical HIV-1-subtype for every sample. To increase the confidence level that each recombination event reported by each tool is not an artifact, we sought agreement between at least two different tools to resolve a recombination event as valid.

Resolution of recombination events across HIV-1 NFLG sequences

For each sample a combinatory analysis was used to reach a final conclusion to determine which of the HIV-1 subtypes underwent recombination (Table 1). Samples 57030 and 64288 were classified in all four tools as RF AE and resolved to be CRF01_AE. Both samples had of subtype B classification in only one tool: sample 57030 in jpHMM while sample 64288 in the phylogenetic tree, since these

observations were not consistent with at least two different tools, we concluded those as artifacts and resolved these samples as RF AE. Sample 79109 had displayed a minor sequence of RF AE, in addition to subtypes B and F1 recombination. This finding was reported by two different tools (jpHHM and Simplot++), hence it was concluded as true positive and this sample was assigned a subtype F1 and B and CRF01_AE recombination. Samples 65699 was resolved as subtype B and C as at least two tools reported a presence of these subtypes. Similarly, sample 89194 had subtypes B and C finding in addition to subtype D reported by three tools including the phylogenetic analyses of the gag and pol genes, thus it was resolved as subtypes B and CD or 41_CD based on NFLG phylogeny. Samples 69734, 81239, 84808 were resolved as recombination of subtype C and A1, as at least two tools displayed that identification, however the NFLG phylogeny and HIV-BLAST favored subtype C, since it was the major parent in those samples, thus we concluded a subtypes C and A1 recombination. Samples 59068, 74004, 79455 were resolved as either RF AG or CRF02_AG, as all methods (except jpHMM) agreed with that classification. Samples 70264, 80907, 81234, 83456, 85576, 88067, 89514 were resolved as RF AG and subtype A, however all (besides one) were suspected to be cpx 36 or 22_01A1, based on NFLG phylogeny. Only sample 83456 showed a different NFLG phylogeny result with resemblance to cpx09 or subtype A7. The HIV-BLAST results for these samples did not cover enough HXB2 bases (most only covered ~2,000bp) and the classification to subtype A in all of them seemed to match to the minor parent, yet sample 88067 had ~5,000bp matching an HIV-BLAST search with subtype A1 and FR AG matching. Since in most tools AG was the dominant RF, it was concluded as the major parent, while subtype A was the minor parent.

Samples that displayed additional recombination to RF AG were 85779, 86845 with subtypes U and G and RF AG or subtypes A, B and C and RF AG recombination, whether these forms are 06_cpx or CRF01_AE (respectively), based on HIV-BLAST results or URF warrens further investigation. Sample 96338 displayed a smaller coverage across the HXB2 genome compared to the rest of the samples, thus we could not conclude if this is a true recombinant form of subtypes B and F1, so it was resolved as subtype B. On the other hand, samples 82699 and 102880 were resolved as subtypes B and F1 recombinants. Interestingly, the gag gene of sample 102880 was classified as an out group for a cluster containing subtypes K, F2, F1, displaying

the advantages of a gene-by-gene phylogeny analysis and indicating the recombination event happened in the gag gene's location. Sample 70787 was resolved as subtype B and A6 recombination, even though the phylogeny clustered only with subtype A6, RIP and SimPlot++ indicated a presence of subtype B. Therefore, the major parent was resolved as subtype A6, while the minor parent was subtype B. Sample 95741 was resolved as subtype A6, based on phylogeny and HIV BLAST results as the rest of the tools showed discrepancies and were not conclusive. Samples 80743 and 83493 were resolved as a recombination of subtypes B and RF AG, 3 tools (phylogeny, RIP, SimPlott++) identified parts of subtype B and FR AG in the sequence, even though these samples maybe a form of cpx (19/37/56) as they clustered among them on NFLG phylogenetic tree. Additionally, sample 88864 was classified as recombination of subtypes B and G, while sample 99478 showed subtypes B and G and RF AG, recombination. Sample 85576 displayed a complex pattern of different subtypes A recombination among all tools and was suspected to be either 36_cpx or a URF. Similarly, samples 72786 and 82664 displayed a complex pattern of subtypes U, G, J, B, G and RF AG (for sample 72786) and subtypes G, L, U, A1 (for sample 82664), however they may be a form of cpx (06 or 11/13) based on NFLG phylogeny. Samples 80405 and 88192 were resolved as pure subtype B, with no indication of recombination events by any of the tools. Samples 99666 and 100545 served as positive control for pure subtypes B and C, respectively.

4. Discussion

Here we have shown the successful implementation of a novel method for HIV-1 NFLG sequencing. This method was tested with clinical samples of whole blood plasma and shown to be relatively easy to perform in a clinical laboratory setting, while simple and less laborious than previously reported methods for HIV-1 NFLG. Our newly developed method has enabled the identification of HIV-1 RFs including CFRs and potential URFs. However, the method introduced here is lacking full coverage of highly variable regions of the HIV-1 genome like env and repetitive regions like the LTR. The amplicons developed method introduced in this pilot study, in optimal conditions, allocating maximal number of reads per sample, surpassed metagenomics in the percentage of HXB2 bases covered, it achieved higher coverage in all 4 compared samples of different HIV-1 pure subtypes and RFs (Fig. 2). With regards to the env gene classification in all used computational tools, the confidence level of that region classification is low, due to lack of full coverage. Consequentially, its classification was difficult to analyze e.g. if a sample had a different subtype classification present only for a relative short number of bases in the env gene it was ignored, especially if it did not fit with the rest of the genome classification. Additionally, some analytic tools like jpHMM had a bias towards subtype B assignment for the env region. With further regards to jpHMM, this tool does not have RF AG as reference, therefore it classified RF AG as a combination of subtype A (mostly A1) and subtype G, consequently, its classification was ignored with regards to RF AG samples only. On the other hand, HIV-BLAST results should be compared with other methods, as the parts of the HIV genome covered by HIV-BLAST results may have a bias towards the major parent in a RF and parts of the genome unmatched may contain different subtype classification.

Previous studies have utilized HIV-1 NFLG protocols successfully, some of the previously developed methods include Gall et al. 2012 [6], which developed a pan HIV-1 primer set, which generates 4 overlapping PCR fragments to cover the viral genome, followed by Roche/454 (Life Sciences, Branford, CT) sequencing. This method showed high sensitivity and reproducibility in detecting drug resistance and minor variants, and was able to identify CRF01_AE and CRF14_BG. Moreover, Grossman et al. 2015 [12], [13] have developed a two PCR fragments method that covers the entire HIV-1 genome, however this was tested only on subtypes B and C and RFs CRF01_AE and CRF02_AG. Others, such as Hebberecht et al. 2019 [14]

developed a multi-step PCR (RT & Nested) based protocol, generating two amplicons followed by sanger sequencing. Their method managed to identify successfully CRF01_AE and CRF02_AG. Additionally, Lunar et al. 2020 [7] have utilized existing methods for HIV-1 NFLG sequencing to identified 6 new URFs and 3 more potential CRFs, in a country with limited HIV diversity.

Here, we have successfully simplified the procedure for HIV-1 NFLG sequencing, still future maintenance of the primers designed here will be needed, as new HIV-1 NFLG CRFs and URFs sequences will accumulate over the years in publicly available databases. The primers presented here (Supp. table 1) should be tested on newly sequenced HIV-1 genomes and updated accordingly, to insure maximal capture of ongoing HIV-1 viral diversity with this approach. Overall, our newly developed amplicons approach has proven itself as reliable on this pilot study, yet it should be compared on a sample-by-sample basis with other HIV-1 NFLG methods to test its robustness and on a larger scale cohort and by other laboratories to be fully corroborated as a suitable approach for HIV-1 NFLG sequencing.

References

[1] A. F. Santos and M. A. Soares, “HIV genetic diversity and drug resistance,” Viruses, vol. 2, no. 2, pp. 503–531, 2010, doi: 10.3390/v2020503.

[2] S. Mamadou, Y. Hanki, A. R. Ali Maazou, B. Aoula, and S. Diallo, “Genetic Diversity and Drug Resistance Mutations in HIV-1 from Untreated Patients in Niamey, Niger.,” ISRN Microbiol., vol. 2011, p. 797463, 2011, doi: 10.5402/2011/797463.

[3] T. Wagner et al., “HIV-1 Circulating Recombinant Forms (CRFs) and Unique Recombinant Forms (URFs) in Israel, 2010–2018,” Viruses, vol. 14, no. 9, p. 1970, Sep. 2022, doi: 10.3390/v14091970.

[4] D. E. Bennett et al., “Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update.,” PLoS One, vol. 4, no. 3, p. e4724, 2009, doi: 10.1371/journal.pone.0004724.

[5] S. V. Alampalli et al., “Deep sequencing of near full-length HIV-1 genomes from plasma identifies circulating subtype C and infrequent occurrence of AC recombinant form in Southern India,” PLoS One, vol. 12, no. 12, p. e0188603, Dec. 2017, doi: 10.1371/journal.pone.0188603.

[6] A. Gall et al., “Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes,” J. Clin. Microbiol., vol. 50, no. 12, pp. 3838–3844, Dec. 2012, doi: 10.1128/JCM.01516-12.

[7] M. M. Lunar, J. Mlakar, T. M. Zorec, and M. Poljak, “HIV-1 unique recombinant forms identified in Slovenia and their characterization by near full-length genome sequencing,” Viruses, vol. 12, no. 1, 2020, doi: 10.3390/v12010063.

[8] B. Beamud, M. A. Bracho, and F. González-Candelas, “Characterization of new recombinant forms of HIV-1 from the comunitat Valenciana (Spain) by phylogenetic incongruence,” Front. Microbiol., vol. 10, no. MAY, p. 1006, May 2019, doi: 10.3389/fmicb.2019.01006.

[9] J. Quick et al., “Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples,” Nat. Protoc., vol. 12, no. 6, pp. 1261–1266, May 2017, doi: 10.1038/nprot.2017.066.

[10] S. Samson, É. Lord, and V. Makarenkov, “SimPlot++: a Python application for representing sequence similarity and detecting recombination,” Bioinformatics, vol. 38, no. 11, pp. 3118–3120, 2022, doi: 10.1093/bioinformatics/btac287.

[11] A. K. Schultz et al., “jpHMM: Improving the reliability of recombination prediction in HIV-1,” Nucleic Acids Res., vol. 37, no. SUPPL. 2, pp. 647–651, 2009, doi: 10.1093/nar/gkp371.

[12] S. Grossmann, P. Nowak, and U. Neogi, “Subtype-independent near full-length HIV-1 genome sequencing and assembly to be used in large molecular epidemiological studies and clinical management,” J. Int. AIDS Soc., vol. 18, no. 1, p. 20035, Jun. 2015, doi: 10.7448/IAS.18.1.20035.

[13] S. G. Aralaguppe et al., “Multiplexed next-generation sequencing and de novo assembly to obtain near full-length HIV-1 genome from plasma virus,” J. Virol.

Methods, vol. 236, pp. 98–104, Oct. 2016, doi: 10.1016/J.JVIROMET.2016.07.010.

[14] L. Hebberecht et al., “Single genome sequencing of near full-length HIV-1 RNA using a limiting dilution approach,” J. Virol. Methods, vol. 274, p. 113737, Dec. 2019, doi: 10.1016/j.jviromet.2019.113737.

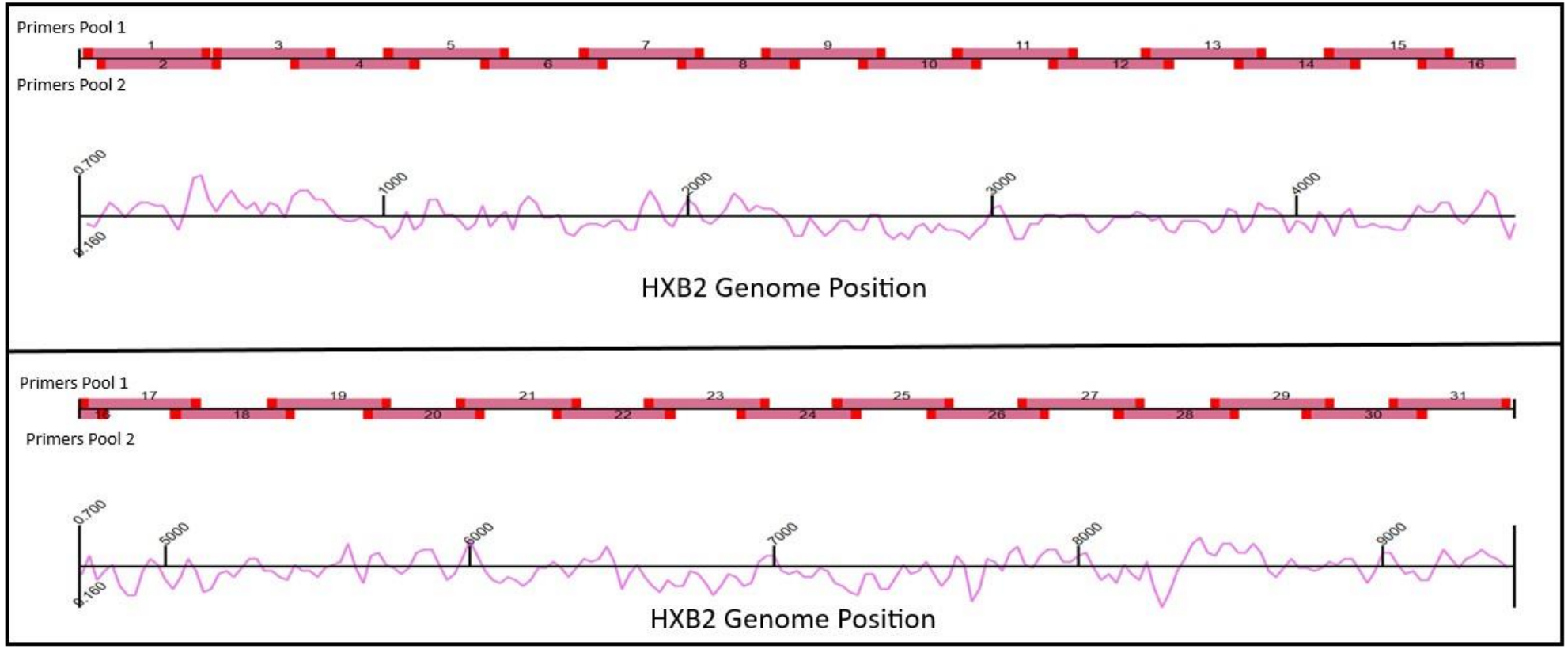

Figure 1. Schematic mapping of the primers covering HIV-1 HXB2 genome.
The general outline of the primers across the HIV-1 reference HXB2 genome designed using primal-scheme [9] is shown. The primer positions are colored in bright red; the amplicons are shown in darker shaded red

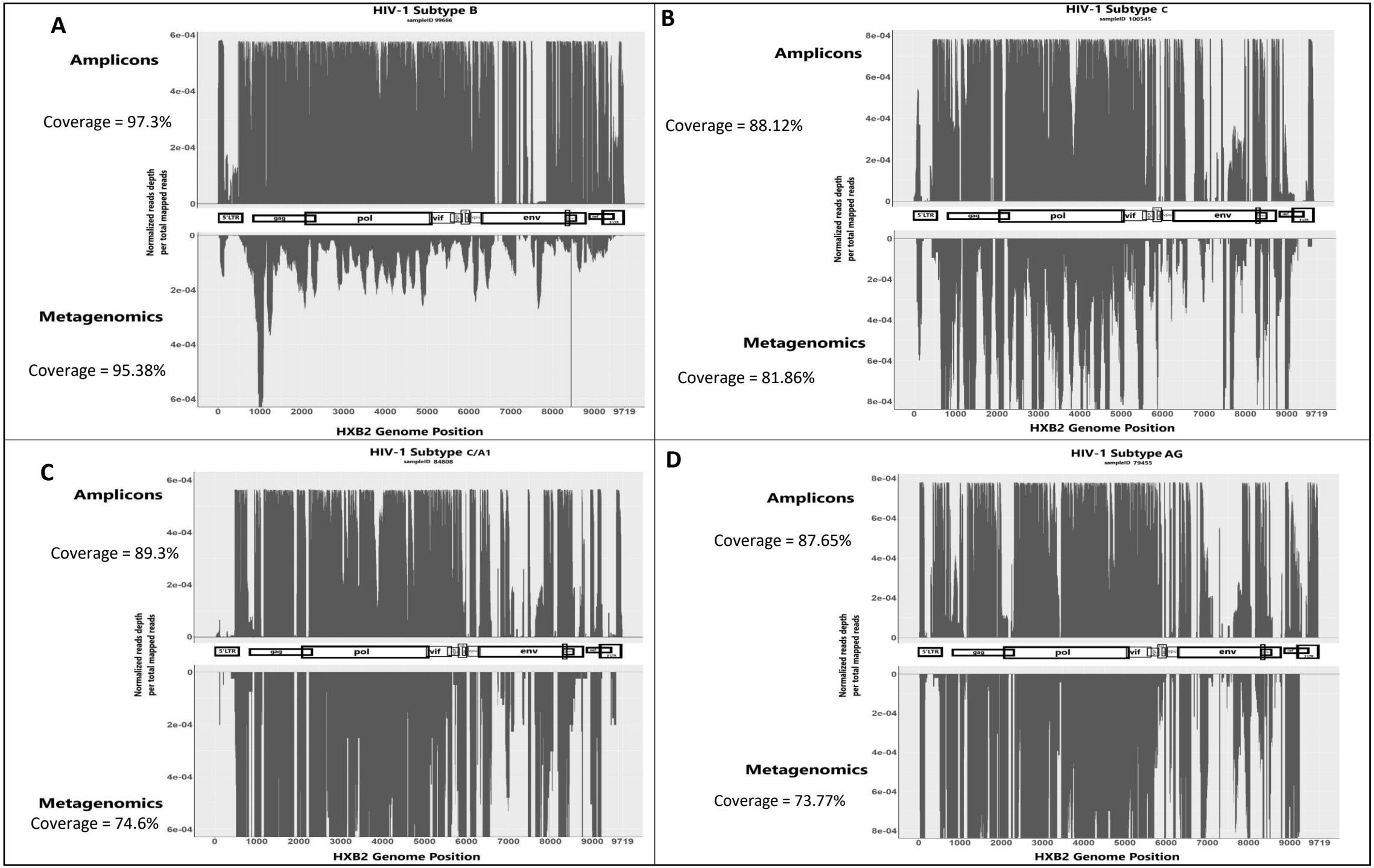
A
HIV-1 Subtype B
sampleID 99666
Amplicons
Coverage = 97.3%
Metagenomics
Coverage = 95.38%
Normalized reads depth per total mapped reads
5'LTR
gag
pol
vif
vpr
vpu
env
HXB2 Genome Position
B
HIV-1 Subtype C
sampleID 100545
Amplicons
Coverage = 88.12%
Metagenomics
Coverage = 81.86%
C
HIV-1 Subtype C/A1
sampleID 84808
Amplicons
Coverage = 89.3%
Metagenomics
Coverage = 74.6%
D
HIV-1 Subtype AG
sampleID 79455
Amplicons
Coverage = 87.65%
Metagenomics
Coverage = 73.77%

Figure 2. Comparison of HIV-1 HXB2 genome coverage between the novel amplicons method and metagenomics.

Values (Y-axis) represent normalized number of reads mapped to the corresponding base position in the HXB2 genome (X-axis). Y-axis directionality (+) for Amplicons and (-) for Metagenomics has no mathematical meaning and was plotted for comparison of each genomic position between the two methods.

In the middle of the plot the HXB2 genome map is presented.

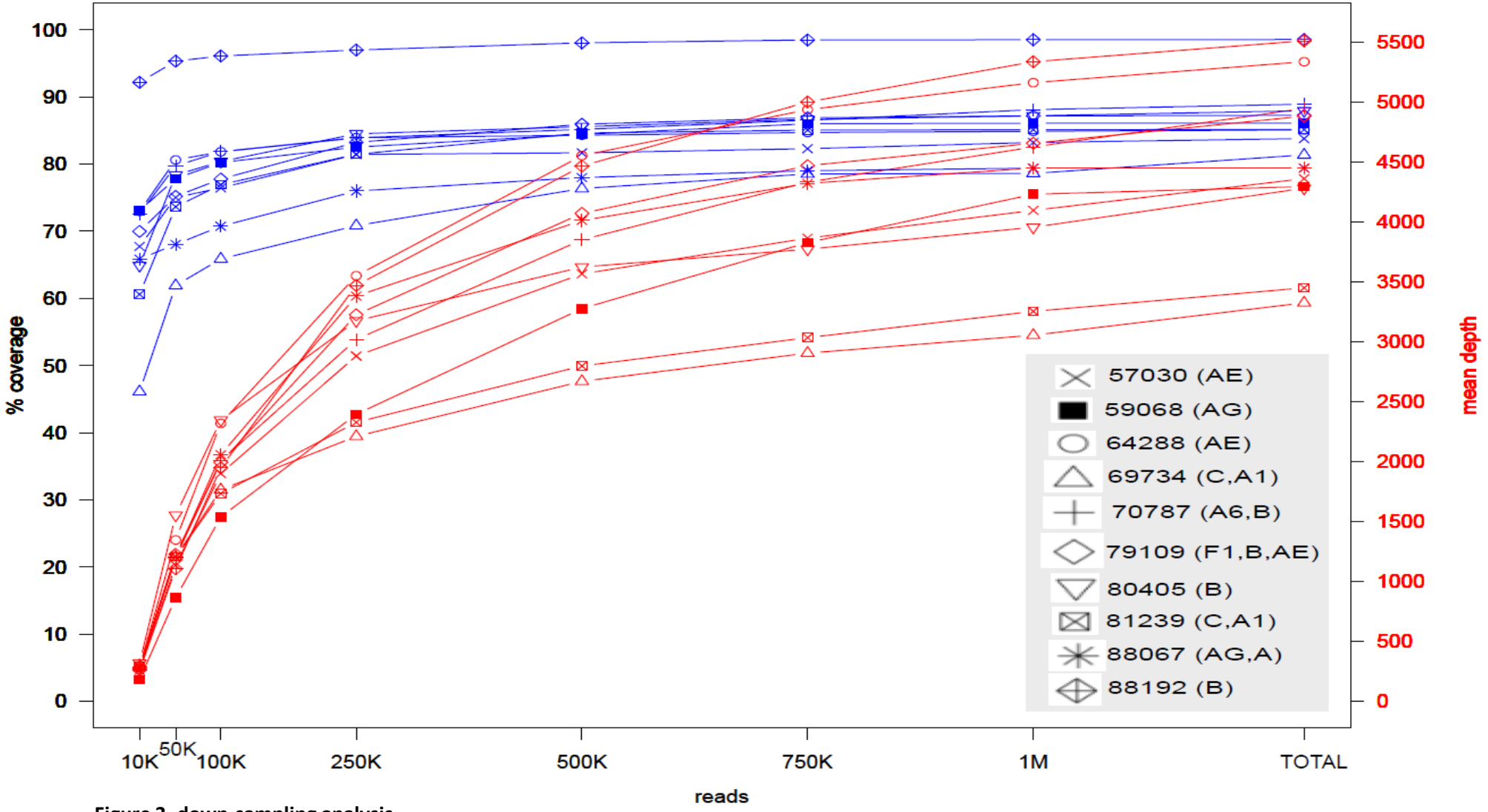

Figure 3. down-sampling analysis.

The blue line corresponds to the Y-axis on the left, representing of HXB2 genome precent coverage. The red line corresponds to the Y-axis on the right corresponding to the mean sequence depth. The X-axis are different number of reads randomly sampled.

A shape was assigned to each sample, in the legend in parenthesis next to sample number is the assigned HIV-1 subtype.

The top blue line with diamond shape corresponds to sample 88192 of HIV-1 subtype B, while bottom blue line with triangle shape corresponds to sample 69734 of HIV-1 subtype C and A1.

In the plot X-axis, the TOTAL average number of mapped paired-end reads was 1,095,890, and varied by samples and subtypes.

SAMPLE ID/ANALYSIS	PHYLOGENETIC TREE (NJ). (*gag,pol,env*)	RIP	jpHMM	SimPlot ++	HIV BLAST {HXB2 bases/similarity}	NFLG PHYLOGENETIC TREE (NJ)	FINAL DECISION
57030	AE,AE,AE	AE	AE, B	AE	01_AE : AB220946 , {5371/5673 (95%)}	AE	**CRF01_AE**
59068	AG,AG,AG	AG	G, A1,	AG	Subtype 02_AG : AB485634 , { 5885/6612 (89%)}	AG	**CRF02_AG**
64288	AE,AE,AE	AE	AE,	AE	Subtype 01_AE : AB220946 , {5681/6298 (90%)}	01B/AE	**CRF01_AE**
65699	C,C,C	B,C	B,C	C,B	Subtype 07_BC : KF250372, {5641/5777 (98%)}	BC	**BC**
69734	(U,G),C,C	C,A1,B	C, G, A1	C,A1	Subtype C : AB485645, { 5509/6331 (87%)}	C	**C, A1**
70264	AG,(A7,A4),(A4,A1,A)	AG,A1,A3, B	A2, A1	AG,A3,A 4,A7	Subtype 02A3 : AY521633, {3726/4040 (92%)}	36_cpx/22_01A1	**AG, A**
70787	A6,A6,A6	A6,B	A1, B, G	A6,B	Subtype A6 : EF589044 , { 5763/6383 (90%)}	A6	**A6, B**
72786	U,(U,G),G	J,G,B	C, G, A1, J, B	AG,J,U,G	Subtype 06_cpx : AJ288982 , {5279/5676 (93%)}	U/32_06A6/06_cpx	**U, G, J**
74004	AG,(A7,A4),(A4,A1,A)	AG,A1	A1, C	AG,A4,A 7	Subtype CRF_30_0206 (AG derived) or A1: AJ508595, OP039481, {4973/5505 (90%)}	36_cpx/22_01A1	**AG**
79109	F1,F1,F1	F1,B	F1, B, AE	F1,B,AE	Subtype B : KT427797 , 5298/6078 (87%)	B/F1	**F1, B, CRF01_AE**
79455	AG,AG,AG	AG	AE, A1, G, B	AG,B	Subtype 02_AG : AB485635 , {6250/7114 (88%)}	95_02B OR 19/37/56 cpx	**AG**
80405	B,B,B	B	B	B	Subtype B : AY835781 , { 5754/6164 (93%)}	B	**B**
80743	AG,AG,AG	AG,B	F2	AG,B,C	Subtype 02B : ON421476 , { 4126/4605 (90%) }	95_02B OR 19/37/56 cpx	**AG, B**
80907	AG,(A7,A4),(A4,A1,A)	AG,A3,B	A1, B	AG,A4,A 7,A1	Subtype 02A or 02A3 : KX228819 or AY521633 , { 2913/3377 (86%), 2318/2523 (92%)}	36_cpx/22_01A1	**AG, A**
81234	AG,(A7,A4),(A4,A1,A)	AG,A3	A1, G	AG,A7,A 1	Subtype 02A3 : AY521633 , { 4036/4370 (92%)}	36_cpx/22_01A1	**AG, A**
81239	C,C,C	C,A1	C, A1	C,A	Subtype BC or C : HM100716 or DQ396370 , {5107/5612 (91%)}	C	**C, A1**
82664	G,(U,G),(L,J)	G,H,J,A7, A1	G, C, A1, J	G,H,AE,J, L,U	Subtype G or 25_cpx : , MH705134 or KY392772, {5023/5757 (87%)}	27_cpx/13_cpx	**G, H, J**
82699	F1,F1,F1	F1,B	F1, B	F1,B,AE	Subtype 12_BF : , AF385934 {5847/6255 (93%)}	BF1	**BF1**
83456	A7,AG,A7	AG	A2	G,AG,A3, A4,A7	Subtype 02_AG : MF109709 , {4516/5152 (88%)}	09_cpx/A7	**AG, A**
83493	AG,B,B	B,AG	B, A1	B,AG,G	Subtype B : MF109558 , { 7459/8541 (87%)}	03_A6B/94_cpx/132_cpx	**AG, B**

SAMPLE ID/ANALYSIS	PHYLOGENETIC TREE (NJ). (*gag,pol,env*)	RIP	jpHMM	SimPlot ++	HIV BLAST {HXB2 bases/similarity}	NFLG PHYLOGENETIC TREE (NJ)	FINAL DECISION
84808	C,C,C	A1, C	C, A1, G	C,B,A,AE	Subtype C : MF109413 , {6955/8229 (85%)}	C	**C,A1**
85576	AG,(A7,A4),(A4,A1,A)	A3,AG,A1,B	A1	AG,A3,AE,A7	Subtype 01_AE or 02A: U51189 or KX228819 , {4951/5534 (89%)}	36_cpx/22_01A1	**AG, A**
85779	U,(U,G),G	G,K,AG	A2	AG,J,G,U	Subtype 06_cpx or A1DK : FJ183725 or X04415 , 4640/5309 (87%) or 4597/5299 (87%)	A6/06_cpx	**U, G, AG**
86845	AG,AG,AG	AG,C,B	A1,C , G, B	AG,C,B,AE,A,G	Subtype CRF01_AE : ON421476 {5023/5923 (85%}, VI26F2885 {4795/5418 (89%)}	95_02B/56 cpx	**A, B, C , AG**
88067	AG,(A7,A4),(A4,A1,A)	AG,A1	A1, B	A3,AG	Subtype A1 or 02_AG : OP039479 or AB286862 , 4942/5564 (89%) or 4936/5555 (89%)	36_cpx/22_01A1	**AG, A**
88192	B,B,B	B	B	B	Subtype B : AY835754 , 8769/9728 (90%)	B	**B**
88864	G,(D,B),G	B,G	G, A2, B , F1	G,B,D,U	Subtype 20_BG : AY586544 , 2421/2547 (95%) , 2554/2874 (89%)	BG	**B, G**
89194	D,D,C	D,C	D, C,G	C,B	Subtype D : AJ320484 , 6064/6623 (92%)	41_CD/D	**D, C, B**
89514	AG,(A7,A4),AG	AG,A3	A1	B,AG,A,A7	Subtype 02_AG or A1 : AB485634 or OP039481 , 5005/5538 (90%)	36_cpx/22_01A1	**AG, A**
95741	A6,A6,A6	AG	A2	A6,B,A	Subtype A6 : EF589044 , 4873/5275 (92%) AND 1510/1805 (84%)	A6	**A6**
96338	B,B,B	B	B,A2,F1	B	Subtype B: KJ704791 , 2588/2793 (93%) AND 2488/2721 (91%)	BF1/01B	**B**
99478	G,(D,B),AG	B,G	G, A2, B	G,AG,B	Subtype 20_BG OR 24_BG : AY586545 OR AY900574 , 2383/2512 (95%) AND 2432/2745 (89%)	BG	**G, AG, B**
99666	B,B,B	B	B	B	Subtype B : JA806687/HH978650 {8697/9746 (89%)}	B	**B**
100545	C,C,C	C	C, K ,B	C	Subtype C : AB254141, 02ZM108, {5812/6587 (88%)}	C	**C**
102880	(K,F1,F2),B,F1	B,F1	B, F1, A1	F1,B,A	Subtype BF1 : MN485983 , 5562/6067 (92%) and 1383/1737 (80%)	DF/F1/BF	**BF1**
84808	C,C,C	A1, C	C, A1, G	C,B,A,AE	Subtype C : MF109413 , {6955/8229 (85%)}	C	**C,A1**
85576	AG,(A7,A4),(A4,A1,A)	A3,AG,A1,B	A1	AG,A3,AE,A7	Subtype 01_AE or 02A: U51189 or KX228819 , {4951/5534 (89%)}	36_cpx/22_01A1	**AG, A**
85779	U,(U,G),G	G,K,AG	A2	AG,J,G,U	Subtype 06_cpx or A1DK : FJ183725 or X04415 , 4640/5309 (87%) or 4597/5299 (87%)	A6/06_cpx	**U, G, AG**
86845	AG,AG,AG	AG,C,B	A1,C , G, B	AG,C,B,AE,A,G	Subtype CRF01_AE : ON421476 {5023/5923 (85%}, VI26F2885 {4795/5418 (89%)}	95_02B/56 cpx	**A, B, C , AG**
88067	AG,(A7,A4),(A4,A1,A)	AG,A1	A1, B	A3,AG	Subtype A1 or 02_AG : OP039479 or AB286862 , 4942/5564 (89%) or 4936/5555 (89%)	36_cpx/22_01A1	**AG, A**

Table 1. Combinatory analyses for the resolution of HIV-1 subtypes recombination across the study samples.
All analyzed samples in the study HIV-1 subtyping results of all computational tools used for classifying each Sample to an HIV-1 subtype. On the second column of the phylogeny analysis, the HIV-1 genome was broken to the three main viral genes : *gag*, *pol*, *env*. Each HIV-1 subtype separated by a comma (,) in each row corresponds to each gene, respectively. For the columns of RIP, jpHMM, SimPlot++ the major parent HIV-1 subtype is written first, followed by minor parent/s. On the final decision column (shaded in red), where there are two different HIV-1 subtype written; the first indicates the major parent, while the second indicates the minor parent of the recombination event.

NAME	POOL	SEQUENCE	START (HXB2)	END (HXB2)	SIZE (bp)	% GC	TM (°C)	STRAND	PCR PRODUCT SIZE (bp)
SC_CON_3_471_F	1	CCAGATCTGAGCCTGGGA	471	487	18	61.1	57.6	+	319
SC_CON_3_789_R	1	GACGCTCTCGCACCCATC	789	806	18	67	60.6	-	
SC_CON_5_997_F	1	ACAGGAWCAGARGARMTTARATCATTAT	997	1024	28	29.2	55.9 - 61.5	+	359
SC_CON_5_1369_R	1	CATRGCTGCYTGATGTCCCC	1364	1383	20	61.1	58.4 - 63.3	-	
SC_CON_7_1630_F	1	AGCATTYTGGAYATAARACAAGGRC	1630	1654	25	38.1	56.4 - 63.1	+	385
SC_CON_7_2019_R	1	CCACATTTCCAACAGCCYYTT	2019	2039	21	47.4	57.5 - 61.1	-	
SC_CON_9_2251_F	1	TCCCTCARATCACTCTTTGGCA	2251	2272	22	47.6	59.9 - 61.1	+	363
SC_CON_9_2614_R	1	TCTTCTGTYAATGGCCATTGTT	2614	2635	21	38.1	56.2 - 58.5	-	
SC_CON_11_2862_F	1	ACAGTRCTRGATGTRGGDGAT	2868	2889	22	50	53.7 - 64.6	+	378
SC_CON_11_3245_R	1	TCCATTTRTCAGGATGGAGYTCA	3245	3267	23	42.9	57.0 - 61.7	-	
SC_CON_13_3514_F	1	CAAAAGACTTAATAGCAGAAATACAGAAACA	3514	3544	31	29	59.5	+	342
SC_CON_13_3865_R	1	GCTGCYCCATCTACATAGAARG	3865	3886	22	50	56.4 - 60.3	-	
SC_CON_15_4050_F	1	TATGCAYTAGGRATYATTCARGCAC	4050	4074	25	38.1	56.1 - 62.5	+	430
SC_CON_15_4482_R	1	TCTGCTGGRATAACYTCTGCTTC	4482	4504	23	47.6	58.7 - 62.0	-	
SC_CON_17_4749_F	1	ACAGCAGTACARATGGCAGTATTC	4749	4772	24	43.5	59.6 - 60.7	+	331
SC_CON_17_5077_R	1	CCATGTKYTAATCYTCATCCTGTCTAC	5077	5103	27	41.7	58.4 - 62.8	-	

NAME	POOL	SEQUENCE	START (HXB2)	END (HXB2)	SIZE (bp)	% GC	TM (°C)	STRAND	PCR PRODUCT SIZE (bp)
SC_CON_19_5326_F	1	ACACAARTAGAYCCTRRCCTRGC	5326	5349	23	50	55.6 - 66.5	+	379
SC_CON_19_5683_R	1	ACYCCTGCCCAAGTATCYCC	5709	5728	20	61.1	58.5 - 63.5	-	
SC_CON_21_5970_F	1	ATGGCAGGAAGAAGMGGARAC	5970	5990	21	52.6	57.9 - 62.5	+	365
SC_CON_21_6333_R	1	ACAGGYACCCCATARTADACWGT	6333	6355	23	45	56.0 - 62.8	-	
SC_CON_23_6579_F	1	TGTGTAAARYTRACCCCWCTYTGYGT	6579	6604	26	42.9	59.8 - 68.1	+	369
SC_CON_23_6953_R	1	ATYCCATGTGTRCAYTGTACTGWR	6953	6973	24	40	55.9 - 63.1	-	
SC_CON_25_7193_F	1	AGGARAYATAAGRMAAGCAYATT	7193	7215	23	27.8	49.7 - 60.3	+	325
SC_CON_25_7521_R	1	ATGGGAGGGGCATAYATTG	7522	7540	20	50	53.6 - 56.8	-	
SC_CON_27_7834_F	1	TGACGGTACAGGCYARACA	7834	7852	19	52.9	56.0 - 60.5	+	391
SC_CON_27_8221_R	1	AAACCAAYTCCACARAYTTKYCC	8221	8243	23	38.9	55.0 - 63.4	-	
SC_CON_29_8510_F	1	CCTGTGCCTYTTCAGCTACC	8510	8529	20	57.9	58.8 - 60.1	+	245
SC_CON_29_8749_R	1	GCCYTGTCTKATTCTTSTAGGTATRT	8749	8774	26	39.1	56.2 - 62.3	-	
SC_CON_31_8797_F	1	ATGGGDRGCAARTGGTCAAAA	8797	8817	21	44.4	56.0 - 62.5	+	269
SC_CON_31_9064_R	1	AGTCCCCCCTTTTCTTTTAAAAA	9064	9086	23	34.8	56.4	-	
SC_CON_33_9080_F	1	GGGGACTGGAWGGGYTARTT	9080	9099	20	55.6	56.8 - 61.0	+	318
SC_CON_33_9394_R	1	TCAKCAGTYYTTGWARWACTCCGG	9394	9417	24	45	58.4 - 64.4	-	

NAME	POOL	SEQUENCE	START (HXB2)	END (HXB2)	SIZE (bp)	% GC	TM (°C)	STRAND	PCR PRODUCT SIZE (bp)
SC_CON_2_63_F	2	CAAGGYTWCTTCCCWGATTGGC	63	84	22	54.5	61.3 - 61.8	+	412
SC_CON_2_481_R	2	CAGAGAGCTCCCRGGC	481	496	16	73.3	55.6 - 58.7	-	
SC_CON_4_672_F	2	CCAGAGRAGHTCTCTCGACGC	672	692	22	63.2	60.0 - 64.3	+	402
SC_CON_4_1070_R	2	TCTARRGCTTCYTTGGTGTCTYKTA	1070	1094	25	40	56.9 - 64.6	-	
SC_CON_6_1307_F	2	CAGCATTATCAGAARGAGCCAC	1307	1328	22	47.6	57.5 - 59.1	+	354
SC_CON_6_1655_R	2	TCTACATARTCTCTRAARGGTTCYT	1658	1682	25	33.3	53.5 - 60.0	-	
SC_CON_8_1963_F	2	TGTTTCAAYTGTGGCAARGARG	1963	1984	22	42.1	56.3 - 61.3	+	372
SC_CON_8_2331_R	2	TCTTCTAATACTGTRTCATCTGCTCC	2331	2356	26	40	57.9 - 60.2	-	
SC_CON_10_2562_F	2	ATTGARACTGTACCAGTAAMATTAAA	2562	2589	28	28	56.0 - 59.9	+	429
SC_CON_10_2988_R	2	GTGAYCCTTTCCATCCYTGTGG	2997	3018	22	55	59.0 - 62.9	-	
SC_CON_12_3198_F	2	CCAGAYAARAARCATCAGAAAGAACC	3198	3223	26	39.1	57.3 - 62.2	+	362
SC_CON_12_3557_R	2	ATRGYTCYTGRTWAATTTGATATGTCCAY	3557	3585	29	29.2	56.1 - 63.5	-	
SC_CON_14_3789_F	2	CCTGARTGGGARTTTGTYAATACY	3783	3806	24	40	55.1 - 61.1	+	389
SC_CON_14_4171_R	2	TCTACYTGTTCATTTCCTCCAATYC	4171	4195	25	39.1	57.3 - 60.3	-	
SC_CON_16_4389_F	2	GTAGAYTGYAGTCCAGGRATATGG	4389	4412	24	47.6	56.2 - 62.1	+	386
SC_CON_16_4771_R	2	CCCCCTTTTCTTTTAAAATTGTGRATGA	4771	4798	28	33.3	59.5 - 60.9	-	

NAME	POOL	SEQUENCE	START (HXB2)	END (HXB2)	SIZE (bp)	% GC	TM (°C)	STRAND	PCR PRODUCT SIZE (bp)
SC_CON_18_4956_F	2	TGGAAAGGTGAAGGGGCAGT	4956	4975	20	55	61.4	+	481
SC_CON_18_5434_R	2	TGTCCTGCTKGATADTYACAMCT	5434	5456	23	39.1	55.9 - 63.0	-	
SC_CON_20_5643_F	2	GAAGCWGTYAGACAYTTTCCTAGR	5643	5666	24	45.83	56.5 - 60.9	+	324
SC_CON_20_5970_R	2	GYYTCCKCTTCTTCCTGCCAT	5970	5990	21	55.6	57.9 - 65.1	-	
SC_CON_22_6206_F	2	AGAGCAGAAGAYAGTGGMAATGA	6206	6228	23	50	57.1 - 62.1	+	355
SC_CON_22_6560_R	2	ACACATGGYTTTAGRCTTTSATCC	6560	6583	24	40.9	57.4 - 61.7	-	
SC_CON_24_6858_F	2	CCAATTCCYATACATTAYTGTRCYCC	6858	6883	26	40.9	55.5 - 63.8	+	454
SC_CON_24_7316_R	2	AAYTTCTRRGTCCCCTCCTGAG	7316	7337	22	52.6	56.9 - 63.1	-	
SC_CON_26_7482_F	2	ATAAARCAAATTRTAARHATGTGGC	7482	7506	27	23.8	51.0 - 59.4	+	433
SC_CON_26_7921_R	2	CCCCAGACYGTGAGYTKCA	7921	7939	19	62.5	56.6 - 64.5	-	
SC_CON_28_8016_F	2	TGCTCTGGAAAACWCATYTGC	8016	8036	21	45	57.4 - 59.1	+	508
SC_CON_28_8522_R	2	AGTCTCTCAAKYGGTGGTAGCTG	8522	8544	23	52.4	60.1 - 64.1	-	
SC_CON_30_8751_F	2	ATACCYASAAGRATMAGACARGGC	8751	8774	24	45	55.8 - 64.0	+	318
SC_CON_30_9074_R	2	CCWTCCAGTCCCCCCTTTT	9074	9092	19	57.9	58.8 - 58.9	-	
SC_CON_32_9390_F	2	CAAGGYTWCTTCCCWGATTGGC	9147	9167	20	54.5	61.3 - 61.8	+	243
SC_CON_32_9390_R	2	TYYTTGWAGWACTCCGGATGY	9390	9410	21	44.4	54.6 - 60.1	-	

Supplement table 1

All primer pairs designed in the study. Primers in pool 1 starts with primer-pair #3. Primers were design to cover bases 488-9389 of the HXB2 reference genome. Primer #2 & #32 Forward have the same sequence.