Infection, Genetics and Evolution 19 (2013) 337–348

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: Performance evaluation of the new REGA version 3 and seven other tools

Andrea-Clemencia Pineda-Peña [a,b,1,2], Nuno Rodrigues Faria [a,1], Stijn Imbrechts [a,1], Pieter Libin [a,c,1,3], Ana Barroso Abecasis [d,4], Koen Deforche [c,3], Arley Gómez-López [b,2], Ricardo J. Camacho [d,e,4,5], Tulio de Oliveira [f,6], Anne-Mieke Vandamme [a,d,*]

[a] Laboratory for Clinical and Epidemiological Virology, Rega Institute for Medical Research, Department of Microbiology and Immunology, University of Leuven, Belgium
[b] Clinical and Molecular Infectious Diseases Group, Faculty of Sciences and Mathematics, Universidad del Rosario, Bogotá, Colombia
[c] MyBioData, Rotselaar, Belgium
[d] Centro de Malária e Outras Doenças Tropicais and Unidade de Microbiologia, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, Lisboa, Portugal
[e] Laboratory de Biologia Molecular, Centro Hospitalar de Lisboa Ocidental Lisboa, Portugal
[f] Africa Centre for Health and Population Studies, Nelson R Mandela School of Medicine, University of KwaZulu-Natal, South Africa

ARTICLE INFO

Article history:
Available online 7 May 2013

Keywords:
HIV-1
Subtypes
Subtyping
Sensitivity
Phylogenetic analysis
CRF

ABSTRACT

Background: To investigate differences in pathogenesis, diagnosis and resistance pathways between HIV-1 subtypes, an accurate subtyping tool for large datasets is needed. We aimed to evaluate the performance of automated subtyping tools to classify the different subtypes and circulating recombinant forms using *pol*, the most sequenced region in clinical practice. We also present the upgraded version 3 of the Rega HIV subtyping tool (REGAv3).

Methodology: HIV-1 *pol* sequences (PR + RT) for 4674 patients retrieved from the Portuguese HIV Drug Resistance Database, and 1872 *pol* sequences trimmed from full-length genomes retrieved from the Los Alamos database were classified with statistical-based tools such as COMET, jpHMM and STAR; similarity-based tools such as NCBI and Stanford; and phylogenetic-based tools such as REGA version 2 (REGAv2), REGAv3, and SCUEAL. The performance of these tools, for *pol*, and for PR and RT separately, was compared in terms of reproducibility, sensitivity and specificity with respect to the gold standard which was manual phylogenetic analysis of the *pol* region.

Results: The sensitivity and specificity for subtypes B and C was more than 96% for seven tools, but was variable for other subtypes such as A, D, F and G. With regard to the most common circulating recombinant forms (CRFs), the sensitivity and specificity for CRF01_AE was ~99% with statistical-based tools, with phylogenetic-based tools and with Stanford, one of the similarity based tools. CRF02_AG was correctly identified for more than 96% by COMET, REGAv3, Stanford and STAR. All the tools reached a specificity of more than 97% for most of the subtypes and the two main CRFs (CRF01_AE and CRF02_AG). Other CRFs were identified only by COMET, REGAv2, REGAv3, and SCUEAL and with variable sensitivity. When analyzing sequences for PR and RT separately, the performance for PR was generally lower and variable between the tools. Similarity and statistical-based tools were 100% reproducible, but this was lower for phylogenetic-based tools such as REGA (~99%) and SCUEAL (~96%).

Abbreviations: CRFs, Circulating Recombinant Forms; LANL, Los Alamos dataset; *MPhy*, manual phylogenetic analysis; nts, nucleotides; PR, Protease; REGAv2, REGA HIV subtyping tool version 2; REGAv3, REGA HIV subtyping tool version 3; RT, Reverse transcriptase; URFs, Unique recombinant forms.

* Corresponding author. Address: University of Leuven, Rega Institute for Medical Research, Minderbroedersstraat 10, B-3000 Leuven, Belgium. Tel.: +32 16 332160; fax: +32 16 332131.
E-mail addresses: andreapinedap@gmail.com (A.-C. Pineda-Peña), annemie.vandamme@uzleuven.be (A.-M. Vandamme).
[1] Address: Minderbroedersstraat 10, B-3000 Leuven, Belgium.
[2] Address: Calle 63D No. 24-31, Bogotá, Colombia.
[3] Address: Beatrijslaan 93, 3110 Rotselaar, Belgium.
[4] Address: Rua da Junqueira No. 100, Lisboa, Portugal.
[5] Address: Rua da Junqueira No. 126, Lisboa, Portugal.
[6] Address: PO Box 198, Mtubatuba 3935, South Africa.

http://dx.doi.org/10.1016/j.meegid.2013.04.032

Conclusions: REGAv3 had an improved performance for subtype B and CRF02_AG compared to REGAv2 and is now able to also identify all epidemiologically relevant CRFs. In general the best performing tools, in alphabetical order, were COMET, jpHMM, REGAv3, and SCUEAL when analyzing pure subtypes in the *pol* region, and COMET and REGAv3 when analyzing most of the CRFs. Based on this study, we recommend to confirm subtyping with 2 well performing tools, and be cautious with the interpretation of short sequences.

1. Introduction

At the end of 2011 there were 34 million of people living with human immunodeficiency virus (HIV) (UNAIDS-WHO, 2012). Most infections are caused by HIV type 1 group Major (HIV-1 group M), which can be further classified into several clades based on genetic diversity. To date, nine distinct subtypes named A–D,F–H, J, K (Robertson et al., 2000) and 58 Circulating Recombinant Forms (CRFs) (http://www.hiv.lanl.gov/; accessed March 2013) have been identified. While subtype B has been widely studied and is predominant in North America, Europe and Australia, it only causes approximately 10 percent of the infections globally (Hemelaar et al., 2011), while subtype C is causing nearly half of global infections, followed by subtype A with 12% of global infections (Hemelaar et al., 2011). In addition, infections with recombinant forms such as CRFs and unique recombinant forms (URFs) have been increasing over the past decades and are now responsible for a total of 20% of the global infections. The distribution of infections caused by CRFs varies regionally; for example, while CRF02_AG (8% global infections) is mostly prevalent in West and Central Africa, CRF01_AE (5% global infections) is more frequent in South and East Asia (Hemelaar et al., 2011). Additionally, CRF06_cpx has been identified in West Africa and some European countries, BC recombinants such as CRF07_BC are frequent in China, and BF recombinants, particularly CRF12_BF, prevail in Brazil and Argentina (Hemelaar et al., 2011).

Due to the fast pace of evolution and frequent recombination of HIV-1 (Jetzt et al., 2000; Mansky and Temin, 1995), accurate subtyping of the growing arsenal of genetic data arising from epidemiological and antiretroviral resistance studies has become increasingly challenging. Importantly, different HIV-1 clades show differences in pathogenesis and present distinct resistance pathways, which in turn may lead to different clinical outcomes. For example, subtype D seems to be more transmissible and is associated with faster disease progression (Baeten et al., 2007). Moreover, subtypes A, C, F and G have some natural polymorphisms in Protease (PR) and Reverse transcriptase (RT) which contribute to resistance in subtype B (Abecasis et al., 2006; Brenner et al., 2006; Camacho and Vandamme, 2007; Martinez-Cajas et al., 2008; Wainberg and Brenner, 2010). However, it is often difficult to compare epidemiological and clinical impact studies since different subtyping methods are used and the classification of HIV-1 clades frequently seems to differ according to the method employed (Hue et al., 2011).

Although the gold standard for classification of HIV-1 is based on phylogenetic analysis of full-length genome sequences (Robertson et al., 2000), this method is not widely used in clinical settings. Since most available data are derived from genotypic assays for resistance to PR and RT inhibitors and this region has proven to contain enough phylogenetic signal (Snoeck et al., 2002), manual phylogenetic analysis (*MPhy*) on *pol* region can be safely used to identify subtypes (Pasquier et al., 2001; Yahi et al., 2001). However, for large datasets, automated tools are needed since manual subtyping is cumbersome. There are three main types of automated tools based on the method used to assign an HIV-1 clade to a query sequence. First, similarity-based tools include the NCBI subtyping tool (Rozanov et al., 2004), Stanford (Liu and Shafer, 2006), Geno2pheno (Beerenwinkel et al., 2003) and EuResist (http://engine.euresist.org/data_analysis/viral_sequence/new). Second, statistical-based tools use partial matching compression algorithms such as COntext-based Modeling for Expeditious Typing –COMET– (Struck et al., 2010), position-specific scoring matrices plus a statistical model such as STAR (Myers et al., 2005) or jumping profile Hidden Markov Models such as jpHMM (Schultz et al., 2009). Finally, there are phylogenetic-based tools such as REGA (Alcantara et al., 2009; de Oliveira et al., 2005) and SCUEAL (Kosakovsky Pond et al., 2009).

A major objective of this paper was to compare the latest Rega subtyping tool with other available tools. The Rega subtyping tool has as philosophy to use phylogenetic analysis in order to take into account the epidemiological and evolutionary relationships among subtypes, such that it approaches the gold standard to classify subtypes (Robertson et al., 2000). The algorithms used in earlier versions of REGA have been previously described (Abecasis et al., 2010; de Oliveira et al., 2005). REGA subtyping tool version 2 (REGAv2) had a high number of unassigned sequences, in part because of the limited number of CRFs included in the reference dataset (Holguin et al., 2008), and the philosophy to achieve a high specificity at the cost of sensitivity. To overcome these limitations, the new REGA subtyping tool version 3 (REGAv3) uses an improved decision-tree algorithm geared towards increasing the recognition of pure subtypes and recombinants (see further details http://bioafrica.mrc.ac.za:8080/rega-genotype-3.0.2/hiv/typingtool/decisiontrees). The reference dataset has also been improved to include more divergent strains per subtype and to classify up to CRF47_BF (See further details http://bioafrica.mrc.ac.za:8080/rega-genotype-3.0.2/hiv/typingtool/method).

In this paper we aim to determine the performance of REGAv3 in the identification of HIV-1 clades, and to compare its sensitivity, specificity and reproducibility with its previous version REGAv2 and six other publicly available automated subtyping tools (COMET, jpHMM, NCBI, SCUEAL, Stanford and STAR). Another goal of this paper was to give guidance as to which HIV-1 subtyping tool would be better for use in a clinical and a surveillance context.

2. Material and methods

2.1. Study population and subtyping tools

With the objective of emphasizing the classification of prevalent non B subtypes, we used two datasets (see Fig. 1). The clinical dataset was retrieved from the Portuguese Resistance database and consisted of 4676 *pol* sequences obtained for routine resistance testing and pooled from 22 Portuguese hospitals, (mean length: 1295 bp; min: 993 bp, max: 1311 bp). Sequences were obtained by population sequencing using the ViroSeq 2.0 toolkit (Abbott Laboratories, Abbott Park, IL, USA). (Sequences are available through Euresist http://www.euresist.org). The Los Alamos dataset herein named as LANL dataset, was retrieved using the following search criteria: "subtype" AND genomic region: "complete genome" AND "one sequence per patient" and we excluded CRFs that could not be shown to have epidemiological relevance and with less than 5 full length genome sequences at the time the analyses were initiated (CRF03_AB, CRF04_cpx, CRF05_DF, CRF08_BC, CRF09_cpx, CRF10_CD, CRF11_cpx, CRF13_cpx, and all CRFs later

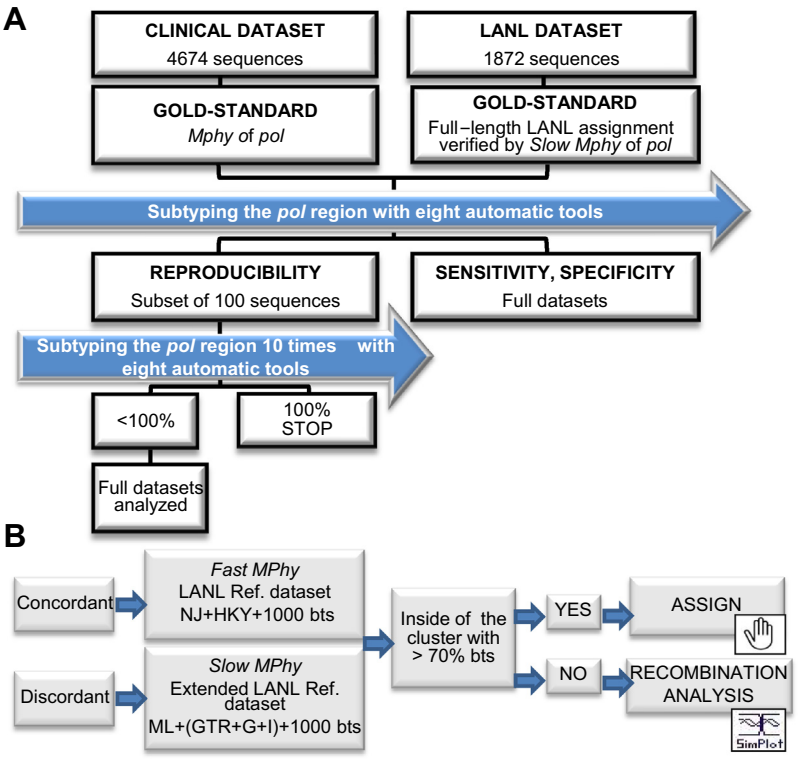

Fig. 1. Methodology of this study. (A) Analyses performed on both datasets as was explained in Section 2. (B) Manual phylogenetic analysis performed on both datasets. *Abbreviations*: *MPhy*: Manual Phylogenetic analysis, LANL: Los Alamos database, NJ: Neighbor joining, HKY: HKY model (Hasegawa, Kishino, Yano), GTR + Γ + I: General time reversible model + Gamma + Proportion of invariant sites, bts: bootstrap, %: percentage.

than CRF15_01B) (Hemelaar et al., 2011). As a result, the LANL dataset included 1872 *pol* sequences (1300 nts), that were trimmed from full genome sequences publicly available in Los Alamos database (http://www.hiv.lanl.gov/; Date of access: October 2011) (accession numbers are shown in the supplementary material number 6). In addition, each sequence of the LANL dataset was divided in PR (mean length: 300 nts) and RT (mean length: 1000 nts) with the objective of evaluating the differences in the performance for identifying PR and RT separately.

Only sequences that passed the quality control check were included: the quality of the sequences was evaluated by using the quality tool of Los Alamos database (available in http://www.hiv.lanl.gov/content/sequence/QC/index.html) and the parameters of Stanford database which are: a maximum number of four for PR and six for RT stop codons + frame-shifts + unpublished AA insertions or deletions + highly ambiguous nucleotides (B,D,H,V,N) (Rhee et al., 2006). As a result, 2 sequences were rejected from the clinical dataset, and the final number of sequences was 4674.

Both datasets were analyzed by the following 8 subtyping tools: COMET version 2 (http://comet.retrovirology.lu), jpHMM (http://jphmm.gobics.de/submission_hiv.html), NCBI subtyping tools using the reference dataset from 2009 (http://www.ncbi.nlm.nih.gov/projects/genotyping/formpage.cgi), Stanford HIVdb version 6.0.10 (http://sierra2.stanford.edu/sierra/servlet/JSierra?action=sequenceInput), SCUEAL (http://www.datamonkey.org/dataupload_scueal.php), STAR (http://www.vgb.ucl.ac.uk/starn.shtml), REGAv2 (http://www.bioafrica.net/rega-genotype/html/subtypinghiv.html) and REGAv3 (http://www.bioafrica.net/typing-v3/hiv).

2.2. Standardization of assignments and manual phylogenetic analysis

Comparison between subtyping tools required standardization of the assignments by different subtyping tools. Thus, 1) sub-subtypes were not taken into account; 2) "A-ancestral" and "A3 subtype" were assigned as A; 3) the assignment "-like" in REGAv3 which is the clustering with a pure subtype outside of the reference cluster with bootstrap >70%, was considered as the subtype or CRF identified by the tool; 4) assignments "complex" or "recombinant" in SCUEAL and REGAv3 were considered recombinants; 5) different subtypes assigned by Stanford to the RT and PR were considered as evidence of recombination.

Each sequence from the clinical and LANL datasets was classified as concordant (all tools agreed on the assignment) or discordant (at least one tool had a different assignment than the other tools) based on the results of the 8 subtyping tools (see Fig. 1). The *MPhy* of concordant sequences was performed by using the 2008 Los Alamos curated subtypes and CRFs reference dataset (available at http://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html), the sequences were aligned with ClustalW (Thompson et al., 1994) and, if needed, the alignment was minimally edited with BioEdit (Hall, 1999). Since assignment for concordant sequences is less problematic than for discordant sequences, and since the dataset is so huge, we opted for a fast Neighbor-joining (NJ) method with 1000 bootstrap replicates and a simple substitution model (HKY85), which we call *fast MPhy*. Such method has been proven useful for subtyping (Gouy et al., 2010; Posada and Crandall, 2001), and it saves computation time. A query sequence was assigned to a particular clade if it clustered monophyletically inside that clade with bootstrap support >70% (Hillis and Bull, 1993; Pasquier et al., 2001; Yahi et al., 2001). Otherwise, the query sequence was considered discordant.

The discordant sequences were further analyzed with the 2008 Los Alamos curated subtypes and CRFs reference dataset complemented with more and curated full-genome sequences available for each subtype in the database using a maximum of 15 sequences per subtype (available in http://www.hiv.lanl.gov/content/sequence/HIV/COMPENDIUM/compendium.html). To optimize subtype classification of the discordant sequences (Kuhner and Felsenstein, 1994; Leitner et al., 1996), we used as gold standard a

slow MPhy using Maximum Likelihood trees with 1000 bootstrap replicates and the best-fitting nucleotide substitution model (in this case GTR + I + Γ) (Posada, 2008; Tamura et al., 2011). If the query sequence clustered monophyletically inside a clade with bootstrap support >70% it was assigned that clade, otherwise the sequence was further screened for recombination using SimPlot with a window size of 300 nts in steps of 20 nts (Lole et al., 1999). For the sequences with no signal for recombination, the sequence was assigned the clade with the highest similarity in SimPlot, and for all such sequences the majority of windows reached >70% bootstrap support. If there was a signal for recombination, the sequence was called unique recombinant form (URF), and the putative recombinant fragments were analyzed separately. A putative recombinant fragment with a phylogenetic signal >0.9 using TREEPUZZLE analysis was assigned a pure subtype or CRF if it clustered inside the respective subtype or CRF clade with >70% bootstrap support (Hillis and Bull, 1993; Schmidt et al., 2002), otherwise the fragment was called unclassified (U) (Robertson et al., 2000).

The analyzed region for CRF01_AE and CRF14_BG is lacking a recombination breakpoint. We considered the *pol* region in the LANL dataset as correctly assigned to these two CRFs, since that assignment is based on the full genome. Such confirmation of breakpoints outside the *pol* region is not available for the clinical dataset, and this can cast doubt on the accurate assignment based on concordance between the tools and confirmed only by *fast MPhy* as described above. Therefore, in addition to *fast MPhy* for concordant sequences, all sequences that were assigned by any of the subtyping tools as either these CRFs or the parent pure subtype (even when concordant) were also analyzed with *slow MPhy* (Guindon et al., 2010), which included all complete genomes of the CRF and parent pure subtype as reference sequences. In order to be considered CRF, the sequence should cluster inside the CRF reference cluster with more than 70% of bootstrap support (Hillis and Bull, 1993; Schmidt et al., 2002), otherwise it was considered the parent pure subtype. Finally, to verify these assignment, all sequences thus assigned subtype G or CRF14_BG were pooled with all full genome CRF14_BG and full genome subtype G sequences from LANL, and a single unrooted tree was constructed using RAxML (Stamatakis, 2006) (supplementary Fig. 3). We found a big discrepancy between the different analyses for CRF14_BG and subtype G, and therefore, for the clinical dataset only, CRF14_BG and subtype G were pooled and analyzed together as a single 'subtype' called "CRF14_BG or G". Tools were considered to correctly assign these sequences when they scored either CRF14_BG or subtype G. We did not encounter problems with CRF01_AE, this was absent in our clinical dataset, and all subtype A sequences were confirmed not to be CRF01_AE.

2.3. Sensitivity, specificity, reproducibility and statistical analysis

The reference standard was *MPhy* of the *pol* region for the clinical dataset and the full-length genome assignment confirmed with *MPhy* of the *pol* region for the LANL dataset (the latter two were 100% concordant). Then we calculated the sensitivity with the formula TP/TP + FN and specificity with the formula TN/TN + FP (Banoo et al., 2010), where TP = true positives, FP = false positives, TN = true negatives, FN = false negatives.

To assess the reproducibility, we created a random subset of 100 sequences extracted from the clinical and LANL datasets that contained pure subtypes and CRFs and then ran this dataset 10 times with each tool (see Fig. 1). The reproducibility was, by definition, the percentage of times the same results were obtained when a subtyping tool was used 10 times on the same sequence, then the average of these percentages was calculated for the 100 sequences (Banoo et al., 2010). For G and CRF14_BG, only LANL sequences were used. If for a specific tool there were any discordant results between the runs, the entire clinical and LANL datasets were evaluated with that tool. The percentage of reproducibility was then calculated for this entire dataset, but again excluding G and CRF14_BG from the clinical dataset.

We evaluated the performance of the tools in the clinical, LANL, and clinical + LANL datasets (herein named as overall dataset). Statistical significance of the difference between subtyping tools was evaluated with McNemar's test. The statistical analysis was calculated using R version 2.12.1.

3. Results

3.1. Subtype distribution of the datasets

Two sequences were excluded from the clinical dataset according to the quality assessment criteria (Rhee et al., 2006). The distribution of subtypes in the datasets is shown in Tables 1–3 according to the *MPhy* of the pooled clinical and LANL datasets (herein named as overall dataset) (the distribution of subtypes according to the clinical or the LANL dataset separately is shown in supplementary Tables 2 and 3 respectively, phylogenetic trees for the clinical and the LANL datasets are in supplementary material Figs. 1 and 2). With regard to the non-B subtypes and the most common CRFs, despite the inclusion of two datasets, there was a limited number of H, J, K, CRF06_cpx, CRF07_BC, CRF12_BF and CRF14_BG sequences to reliably evaluate the performance of the subtyping tools, yet the results are still listed in the tables. The CRF13_cpx, CRF18_cpx, CRF25_cpx and CRF27_cpx were also found in the clinical dataset (see supplementary Table 2), but these CRFs had a very low prevalence and not enough full genome sequences were found in LANL at the time of the collection of data to reliably assess the performance of the tools for these CRFs, but we included the results in supplementary materials (supplementary Table 2). We did not evaluate the performance of the subtyping tools for CRFs that are not contributing substantially to the epidemic or that are poorly assigned. That is also why REGAv3 does not score all CRFs reported to date (see information about the epidemiological, geographical and recombination information in http://bioafrica.mrc.ac.za/CRFs/CRFs.php). For the clinical dataset, subtype G and CRF14_BG were pooled into a special class "G or CRF14_BG."

3.2. Performance of the subtyping tools for pure subtype assignment

We evaluated the performance on the overall dataset, the clinical separately and the LANL dataset separately. Since we did not find many differences between the results of these datasets, we only show the performance of the overall dataset (see the performance of the tools for the clinical or the LANL dataset separately in supplementary Tables 2 and 3, respectively). However, we found some discrepancies in the results for subtype A when we compared the two results of the clinical and the LANL datasets. For example, subtype A in the LANL dataset was 100% accurately classified by COMET, jpHMM and REGAv2 but the values in the clinical dataset were 76.5%, 86%, and 73%, respectively.

The sensitivity of each of the subtyping tools was more than 96% for subtype B and 98% for subtype C except for NCBI, which had lower values in both datasets (Tables 1–3, see details in supplementary Tables 2 and 3). However the results were variable for other subtypes. For instance, COMET, jpHMM, and REGAv3 had a sensitivity of more than 90% for subtype A. jpHMM and REGAv3 obtained sensitivities of 100% for subtype D and subtype F, respectively. REGAv3, Stanford and STAR classified correctly subtypes H, J, K, although the number of sequences available was limited. Noteworthy, the specificity for all pure subtypes was more than 98% (Tables 1–3).

3.3. Performance of the subtyping tools for Recombinant Forms

COMET, jpHMM, REGAv2, REGAv3, Stanford and STAR had sensitivities and specificities around 99% for the classification of

Table 1
Performance of statistical-based subtyping tools.

Subtype	Total	Statistical-based tools											
		COMET				jpHMM				STAR			
		Sens	95% CI	Spec	95% CI	Sens	95% CI	Spec	95% CI	Sens	95% CI	Spec	95% CI
A	226	91.2	87.4-94.9	100.0	99.9-100	94.7	91.8-97.6	99.8	99.7-99.9	50.4	43.9-57.0	100.0	100-100
B	3023	99.1	98.8-99.5	99.8	99.6-99.9	99.0	98.7-99.4	99.6	99.4-99.8	97.6	97.0-98.1	99.6	99.4-99.8
C	628	99.7	99.2-100	100.0	100-100	100.0	100-100	100.0	100-100	99.8	99.5-100	100.0	100-100
D	69	97.1	93.1-100	100.0	100-100	100.0	100-100	100.0	100-100	89.9	82.7-97.0	100.0	100-100
F	129	89.1	83.8-94.5	100.0	100-100	92.2	87.6-96.9	100.0	100-100	89.1	83.8-94.5	100.0	100-100
LANL G*	34	100.0	85.1-100	99.5	99.0-99.7	97.1	84.7-99.9	99.4	98.9-99.7	73.5	55.6-87.1	99.4	98.9-99.7
H	11	90.9	73.9-100	100.0	100-100	90.9	73.9-100	100.0	100-100	100.0	100-100	100.0	100-100
J	6	50.0	10.0-90.0	100.0	100-100	100.0	100-100	100.0	100-100	100.0	100-100	100.0	100-100
K†	2	100.0	100-100	100.0	100-100	100.0	100-100	100.0	100-100	100.0	100-100	100.0	100-100
CRF01_AE†	169	99.4	98.3-100	100.0	100-100	100.0	100-100	100.0	100-100	98.8	97.2-100	100.0	100-100
CRF02_AG	272	96.3	94.1-98.6	100.0	99.9-100	*NA*	*NA*	*NA*	*NA*	96.0	93.6-98.3	99.7	99.5-99.8
CRF06_cpx	28	50.0	31.5-68.5	100.0	100-100	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
CRF07_BC†	10	90.0	55-100	100.0	100-100	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
CRF12_BF†	5	100.0	36.0-100	100.0	100-100	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
LANL CRF14_BG*	11	81.8	47.8-96.8	99.8	99.6- 99.9	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
Clinical G+CRF14_BG‡	1571	98.7	98.2-99.3	99.9	99.8-100	98.3	97.7-99	99.2	98.9-99.5	95.3	94.2-96.3	99.6	99.4-99.8

The sensitivity (Sens) and specificity (Spec) are reported for statistical-based tools. The values with 100% of performance are highlighted in dark gray; the values with more than 90% of performance are colored in light gray. *The values for G and CRF14_BG are based on the LANL dataset only. †These subtypes of CRFs only were available in the LANL dataset. ‡The 1571 sequences G and CRF14_BG of the clinical dataset were pooled as a single category. *Abbreviations:* n: sample, cpx: complex, LANL: Los Alamos dataset, NA: Not applicable, URF: Unique recombinant form.

Table 2
Performance of similarity-based subtyping tools.

Subtype	Total	Similarity-based tools							
		NCBI				STANFORD			
		Sens	95% CI	Spec	95% CI	Sens	95% CI	Spec	95% CI
A	226	65.5	59.3-71.7	100.0	99.9-100	63.3	57.0-69.6	100.0	99.9-100
B	3023	84.7	83.4-85.9	98.8	98.4-99.1	98.3	97.9-98.8	99.0	98.7-99.4
C	628	92.4	90.3-94.4	100.0	100-100	98.9	98.1-99.7	100.0	99.9-100
D	69	79.7	70.2-89.2	100.0	100-100	91.3	84.7-98.0	100.0	100-100
F	129	87.6	81.9-93.3	99.9	99.9-100	71.3	63.5-79.1	100.0	100-100
LANL G*	34	47.1	29.8-64.9	100.0	99.7-100	97.1	84.7-99.9	99.4	98.9-99.7
H	11	100.0	100-100	100.0	100-100	100.0	100-100	100.0	100-100
J	6	66.7	28.9-100	100.0	100-100	100.0	100-100	100.0	100-100
K†	2	100.0	100-100	100.0	100-100	100.0	100-100	100.0	100-100
CRF01_AE†	169	76.3	69.9-82.7	99.7	99.4-100	100.0	100-100	99.6	99.3-99.9
CRF02_AG	272	48.5	42.6-54.5	99.8	99.7-99.9	98.9	97.7-100	98.0	97.6-98.3
CRF06_cpx	28	82.1	68.0-96.3	99.3	99.1-99.5	*NA*	*NA*	*NA*	*NA*
CRF07_BC†	10	100.0	59.0-100	100.0	100-100	*NA*	*NA*	*NA*	*NA*
CRF12_BF†	5	100.0	36.0-100	100.0	100-100	*NA*	*NA*	*NA*	*NA*
LANL CRF14_BG*	11	100.0	61.5-100	99.5	99.1-99.7	*NA*	*NA*	*NA*	*NA*
Clinical G+CRF14_BG‡	1571	99.7	99.5-100	98.8	98.4-99.2	97.5	96.7-98.3	98.9	98.5-99.2

The sensitivity (Sens) and specificity (Spec) are reported for similarity-based tools. The values with 100% of performance are highlighted in dark gray; the values with more than 90% of performance are colored in light gray. *The values for G and CRF14_BG are based on the LANL dataset. †These subtypes of CRFs only were available in the LANL dataset. ‡The 1571 sequences G and CRF14_BG of the clinical dataset were included in the total. *Abbreviations:* n: sample, cpx: complex, LANL: Los Alamos dataset, NA: Not applicable, URF: Unique recombinant form.

CRF01_AE (see Tables 1–3, supplementary material Table 3). However, the absence of a recombination breakpoint in the *pol* region makes the classification of this CRF challenging by the subtyping tools (see Fig. 2) (Carr et al., 1996). Therefore, COMET and REGAv3 used other assignments; for example, COMET classified one sequence as "01_AE (check for 15_01B)" and REGAv3 identified 99% (168/169) as "HIV Subtype A (CRF01_AE)." On the other hand, SCUEAL and NCBI classified independently CRF15_01B and CRF01_AE, with sensitivity dropping to 84% and 76%, respectively.

The sensitivity and specificity was more than 96% using COMET, REGAv3, Stanford and STAR for CRF02_AG, while NCBI and SCUEAL had values below 50% for sensitivity. In most of the cases, NCBI misclassified some CRF02_AG as CRF30_0206 or CRF36_cpx whereas SCUEAL identified CRF02_AG sequences as "complex".

Regarding CRF06_cpx, REGAv3 had the highest sensitivity using 17 and 11 sequences in the clinical and LANL datasets respectively. In the clinical dataset low prevalent CRFs were also found, for instance, CRF25_cpx was identified 100% by REGAv3 in 9 sequences (see supplementary Table 2), CRF18_cpx was classified with 100% of sensitivity with NCBI, REGAv3 and SCUEAL in 3 sequences. Only REGAv2 and REGAv3 correctly identified both sequences of CRF13_cpx and both sequences of CRF27_cpx, respectively. On

Table 3
Performance of phylogeny-based subtyping tools.

		Phylogenetic tools											
	Total	REGAv2				REGAv3				SCUEAL			
Subtype		Sens	95% CI	Spec	95% CI	Sens	95% CI	Spec	95% CI	Sens	95% CI	Spec	95% CI
A	226	89.8	85.9-93.8	100.0	99.9-100	95.6	92.9-98.3	100.0	99.9-100	85.8	81.3-90.4	99.9	99.9-100
B	3023	97.3	96.7-97.8	99.8	99.7-99.9	99.2	98.9-99.5	99.3	99.0-99.5	96.3	95.7-97.0	99.9	99.5-99.9
C	628	99.8	99.5-100	100.0	100-100	100.0	100-100	100.0	100-100	99.0	98.3-99.8	100.0	100-100
D	69	84.1	75.4-92.7	100.0	100-100	88.4	80.9-93.0	100.0	100-100	95.7	90.8-100	100.0	100-100
F	129	93.8	89.6-98.0	100.0	100-100	100.0	100-100	100.0	100-100	89.9	84.7-95.1	100.0	100-100
LANL G*	34	100.0	85.1-100	99.4	98.9-99.7	100.0	85.1-100	99.4	98.9-99.7	97.1	84.7-99.9	99.9	99.6-100
H	11	90.9	73.9-100	100.0	100-100	100.0	100-100	100.0	100-100	90.9	73.9-100	100.0	100-100
J	6	100.0	100-100	100.0	100-100	100.0	100-100	100.0	100-100	100.0	100-100	100.0	100-100
K†	2	100.0	100-100	100.0	100-100	100.0	100-100	100.0	100-100	100.0	100-100	100.0	100-100
CRF01_AE†	169	99.4	98.3-100	100.0	100-100	99.4	98.3-100	100.0	100-100	84.0	78.5-89.5	100.0	100-100
CRF02_AG	272	64.7	59.0-70.4	100.0	99.9-100	98.9	97.7-100	100.0	100-100	33.8	28.2-39.4	100.0	100-100
CRF06_cpx	28	78.6	63.4-93.8	99.7	99.6-99.9	96.4	89.6-100	99.5	99.3-99.6	46.4	28.0-64.9	100.0	100-100
CRF07_BC†	10	100.0	59.0-100	100.0	100-100	100.0	59.0-100	100.0	100-100	40.0	9.6-70.4	100.0	100-100
CRF12_BF†	5	80.0	28.0-100	100.0	100-100	100.0	36.0-100	100.0	100-100	80.0	28.0-100	100.0	100-100
LANL CRF14_BG*	11	63.6	30.9-88.9	100.0	99.7-100	72.7	39.1-93.7	99.9	99.7-100	81.8	47.8-96.8	99.9	99.6-99.9
Clinical G+CRF14_BG‡	1571	98.8	98.3-99.3	99.8	99.6-99.9	99.8	99.6-100	98.9	98.6-99.3	97.6	96.9-98.4	99.7	99.5-99.9

The sensitivity (Sens) and specificity (Spec) are reported for phylogenetic-based tools. The values with 100% of performance are highlighted in dark gray; the values with more than 90% of performance are colored in light gray. *The values for G and CRF14_BG are based on the LANL dataset. †These subtypes of CRFs only were available in the LANL dataset. ‡The 1571 sequences G and CRF14_BG of the clinical dataset were included in the total. *Abbreviations:* n: sample, cpx: complex, LANL: Los Alamos dataset, NA: Not applicable, URF: Unique recombinant form, REGAv3: REGA subtyping tool version 3, REGAv2: REGA subtyping tool version 2, URF: Unique recombinant form.

the other hand, the LANL dataset included other prevalent CRFs such as CRF07_BC and CRF12_BF. A 100% of 10 sequences of CRF07_BC were identified correctly using NCBI, REGAv2 and, REGAv3. Similarly, a 100% of 5 sequences of CRF12_BF were correctly classified by COMET, NCBI and REGAv3.

A CRF was never assigned when the sequence clustered significantly with a CRF but outside of the reference clade. In this way, potential CRFs can have been assigned URF; however there is no safe way to assign such a sequence to a CRF in absence of the full genome. To overcome this potential limitation, the sequences assigned in this way as URF were further analyzed with the *slow Mphy* and the CRFs' reference dataset complemented with all curated full genome CRF sequences of Los Alamos database, and with SimPlot. We found that we had not missed any true CRFs and that all sequences that had been assigned URFs were truly URFs with unassigned fragments such as CRF06_cpx/U recombinant (23%, 79/336). Other frequent URFs in the clinical dataset were various B/G recombinants (35%, 118/336), followed by other recombinants with unassigned fragments such as G/U recombinant (9%, 31/336) and B/U recombinant (5%, 16/336). The performance of the subtyping tools was not evaluated for URFs because of the limited number of sequences available and the complexity to evaluate recombinants when tools such as COMET, Stanford and STAR do not show the recombination breakpoints (Liu and Shafer, 2006; Myers et al., 2005; Struck et al., 2010).

3.4. Performance of the subtyping tools for G and CRF14_BG

When confronted with the discrepancy in manual phylogenetic assignment for CRF14_BG and subtype G in the clinical dataset, as described in methods, we decided to only use the LANL dataset to calculate the performance on these two subtypes, because the *pol* region was trimmed from full genomes which is the only way to safely assign CRFs that lack a breakpoint in the here analyzed *pol* region (see Fig. 2) (Delgado et al., 2002). The sensitivity for subtype G was more than 97% for all the tools, except NCBI and STAR in the LANL dataset. The first had a sensitivity of 47% because of misclassification of some subtype G sequences as CRF14_BG or CRF43_02AG (see Fig. 2). STAR had 73% of sensitivity due to "unassigned" sequences. The specificity for subtype G was more than 99% for all the tools. Regarding CRF14_BG, COMET and SCUEAL classified 9 out of 11 sequences correctly followed by REGAv2 and REGAv3, which classified 8 out of 11 sequences as "Subtype G (CRF14_BG)".

To evaluate the discordances between the tools, the *fast* and *slow MPhy* procedures and the manual phylogenetic analysis of subtype G and CRF14_BG sequences in a single tree, as described in methods, we retrieved the envelope (*env*) sequences from the Portuguese resistance database. We found 44 sequences (C2-V5 region of the gp120, mean length: 500 nts) from the same patients whose *pol* sequence was included in this study, 28 had their *pol* sequence classified as G using *slow MPhy*, and 16 as CRF14_BG. 4 subtypes B *env* sequences belonged to patient isolates classified as G in *pol*, similarly 6 subtype G *env* sequences belonged to patient isolates classified as CRF14_BG in *pol*. In addition, we determined how the Portuguese sequences clustered with respect to all full genome G/CRF14_BG sequences available from the Los Alamos database. All Portuguese G or CRF14_BG clinical sequences formed a monophyletic cluster within subtype G including all CRF14_BG full genomes, but the Portuguese clinical sequences, assigned as CRF14_BG and G by *slow MPhy* were paraphyletic with each other, suggesting that there may be a problem with the assignment of CRF14_BG and this CRF may in fact consist of more than one CRF with very similar breakpoints (see supplementary material Fig. 3).

Using *slow Mphy*, 951 sequences were considered subtype G and in more than 96% of the cases these were also classified subtype G by jpHMM, REGAv2, REGAv3 and Stanford (in alphabetical order and see supplementary material Table 2), however, the tools also assigned many of these sequences to CRF14_BG; with the exception of COMET and SCUEAL. Using *slow Mphy*, 620 sequences were considered CRF14_BG, and again, COMET and SCUEAL had the highest agreement with *slow MPhy*, but these values were just 62% and 56%, respectively. Given that we did no longer consider the slow *MPhy* reliable for subtype G and CRF14_BG, these performance statistics are also not reliable. We therefore chose to pool the subtype G and CRF14_BG sequences from the Portuguese clinical database, as they were assigned by *slow MPhy*, and compute the performance of the tools on the combined class "G or CRF14_BG".The sensitivity was above 99% for NCBI and REGAv3,

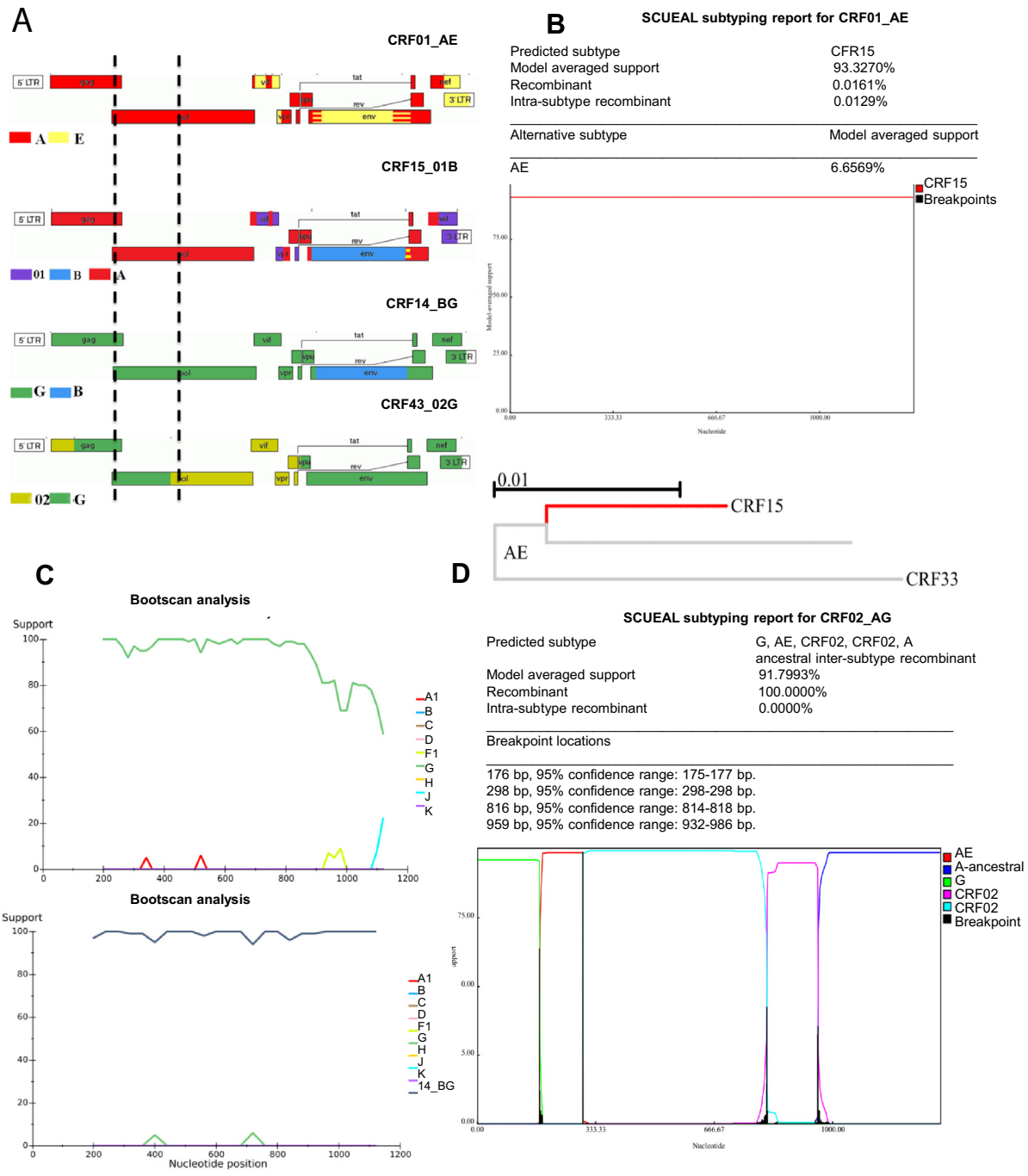

Fig. 2. Frequent problems with the classification of CRFs. (A) Adapted from the Los Alamos database available in http://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html. The dashed lines delineate the region that is used for resistance testing and for the current performance analysis. The CRF01_AE and CRF15_01B are entirely subtype A in the *pol* gene, similarly CRF14_BG and CRF43_02G are entirely subtype G in this region. This complicates the identification since it is difficult to discriminate the parent pure subtypes from the CRFs in geographic areas where the CRFs originated. (B) An example analysis by SCUEAL. The query sequence of the genomic region *pol* has no evidence of recombination and it clusters with CRF15_01B. However, this *pol* gene is from a full-length genome sequence assigned as CRF01_AE. It is possible that it concerns here a CRF01_AE that was very closely related to the founder of CRF15_01B. (C) An example analysis by REGAv3. The query sequence is the *pol* region trimmed from a full-length genome sequence assigned as CRF14_BG. In the pure subtype analysis it has a high support with subtype G and in the CRF analysis it has a high support with CRF14_BG. The algorithm classified it as G. This might be due to the fact that the sequence did not cluster reliably within the CRF14_BG clade. (D) Example of the assignment "complex" by SCUEAL. The assignment of the full genome is CRF02_AG but the tool identified it as a G, CRF01_AE, CRF02_AG, A ancestral recombinant.

followed by COMET, jpHMM, and REGAv2 with more than 98% while specificity was around 99% for all the tools.

3.5. Performance of the subtyping tools for PR and RT separately

The performance of the tools on PR and RT separately was only evaluated on the LANL dataset and the results are shown in supplementary data (supplementary Tables 4 and 5). When analyzing PR separately, COMET and jpHMM showed similar performance as for *pol* with regard to pure subtypes and the CRF01_AE from the LANL dataset. However, the sensitivities varied for the other tools. For instance, Stanford and STAR had a better sensitivity for subtype A, but NCBI a worse sensitivity for subtypes A, C, F, and G. Stanford and STAR had similar sensitivities for CRF02_AG but other tools

had a lower performance like COMET, NCBI, REGA and SCUEAL. With regard to RT sequences, in general, the performance of the subtyping tools was the same as for the *pol* sequences, again using the LANL dataset only. The exceptions included tools with improved sensitivity such as NCBI for subtypes B, D and G; REGAv3 for CRF02_AG, Stanford for subtypes A, F and STAR for subtype G. However, SCUEAL had decreased sensitivity for subtype CRF02_AG (supplementary Table 4).

3.6. Performance of REGAv3 versus the previous version REGAv2

The performance of REGAv3 was better than REGAv2 for subtypes B ($p = 0.01$), and CRF02_AG ($p = 0.001$) in the genomic region *pol* in both datasets (see Tables 1–3 and supplementary Tables 2 and 3). In the case of CRF02_AG, for instance, the changes in the decision tree for REGAv3 improved sensitivity by properly assigning CRF02_AG sequences that were classified as "check the bootscan" by REGAv2.

We compared the new term "-like" of REGAv3 with *MPhy*. "Subtype B-like" corresponded to subtype B in 28 sequences analyzed with the *MPhy*, while 8 were B/U recombinants and 2 were B/G recombinants. In the case of "subtype G-like", the *MPhy* showed 1 subtype G, 3 recombinants G and one CRF06_cpx. "Subtype A1-like" and "Subtype F1-like" were identified in 2 and 4 sequences, respectively, but in both cases the *MPhy* showed these were A and F subtypes.

We also evaluated the performance for PR and RT separately but only using the LANL dataset. In the analysis of RT, REGAv3 had higher sensitivity than REGAv2 for subtype B (98.8 versus 91.9). However, the performance in PR was variable (see supplementary Table 5) because the REGA subtyping tool algorithm is different for sequences shorter than 800 nts. In short sequences, the criteria are based on clustering only and potential recombination is not analyzed. This is because the window size for recombination analysis in REGA is chosen as 400 nts to avoid losing too much phylogenetic signal (Strimmer and von Haeseler, 1997) and recombination is scanned in steps of 50 nts, such that no meaningful recombination signal can be obtained for such short sequences. REGAv3 correctly identified only 54% of the subtype A sequences, which is better than the 45% with REGAv2, but the sensitivity for subtype B decreased from 86% to 73% when comparing REGAv2 versus REGAv3. There was no difference between the tools for subtype C and G. The lack of phylogenetic signal in the short fragment of PR significantly reduced the sensitivity of the tool, for instance REGAv2 had a sensitivity of 27% and REGAv3 had 24% for subtype F; while none of the subtypes D, CRF01_AE and CRF02_AG sequences analyzed were properly identified.

3.7. Reproducibility of HIV-1 subtyping tools

COMET, jpHMM, NCBI, Stanford, and STAR were 100% reproducible. Subtyping tools based on phylogenetic methods such as REGAv2, REGAv3 and SCUEAL were reproducible with values of 99.2% (95% CI: 99.10–99.26), 99.2% (95% CI: 99.15–99.30) and 96.4% (95% CI: 96.27–96.60), respectively. When the clinical and LANL datasets were independently analyzed, the reproducibility did not change significantly; for instance, the reproducibility for REGAv2 was 99.1% and 99.5%, for REGAv3 98.8% and 99.7%, and for SCUEAL 95.4% and 98.1%, respectively.

For REGAv2 and REGAv3, most of the non-reproducible results were related to subtype B. For example, when the sequence was subtype B, the tool might classify it as subtype B, or as "check the report" or "B/D recombinant" (for details see also Supplementary Table 1). In this paper we considered "subtype-like" as belonging to the subtype (or CRF) identified by the tool for REGAv3. If "subtype-like" would be considered discordant, then the sensitivity of REGAv3 would go down (from 99.2% to 97.8% for subtype B, the assignment with the highest number of –like assignments), and the reproducibility would go up (from 98.9% to 99.2%). For SCUEAL, the non-reproducible results were related mainly to subtype B and CRF02_AG. For instance, subtype B was sometimes classified as "B/D recombinant" by the tool, and CRF02_AG was frequently assigned as "complex".

4. Discussion

In the present study, we compared and described the performance of the phylogenetic based automated HIV-1 subtyping tool REGAv3, its previous version REGAv2, and six other commonly used automated HIV-1 subtyping tools: one other phylogenetic based tool, SCUEAL; three statistical-based tools, COMET, jpHMM and STAR; and two similarity based tools, NCBI and Stanford. We used only the *pol* (PR + RT) region that is usually sequenced for drug resistance testing (Thompson et al., 2012; Vandamme et al., 2011), since this generates the largest datasets for which these tools are designed. This restriction was also made since tools such as SCUEAL and Stanford cannot assign sequences outside this region (Kosakovsky Pond et al., 2009; Liu and Shafer, 2006). In addition, we analyzed with phylogenetic analysis two datasets; one dataset was derived from clinical samples and another from the Los Alamos database (available in http://www.hiv.lanl.gov/). We used the clinical data from Portugal because it is one of the European countries with the highest proportion of non B-subtypes (Abecasis et al., 2008, 2013). Since phylogenetic analysis of full-length genomes is the gold standard to define the current subtypes (Robertson et al., 2000), we also used the assignment of the *pol* region confirmed with *MPhy* and trimmed from full-length genomes of pure subtypes and the most common CRFs from the Los Alamos database, with the aim to better evaluate the performance of automated subtyping tools on all epidemic subtypes and CRFs.

Our primary aim was to evaluate the new subtyping tool REGAv3 versus other available tools. REGAv3 identified subtype B, most of the non-B pure subtypes and the most frequent CRFs with a sensitivity and specificity of more than 96% in the *pol* region. The classification of REGAv3 for subtype B and CRF02_AG has improved compared to REGAv2; additionally, with an updated algorithm and reference dataset, REGAv3 is designed to identify most of the epidemic CRFs. Consequently, REGAv3 performs equally well as other tools, such as COMET and jpHMM, which also had high sensitivity and specificity for classifying most of the pure subtypes and such as COMET, Stanford and STAR for classifying CRF01_AE and CRF02_AG.

Concerning some previous reports that suggested a low performance of REGAv2 compared with other subtyping tools (Holguin et al., 2008; Yebra et al., 2010b), it is pertinent to add that these discrepancies were almost exclusively due to unassigned reports, and not due to wrong assignments (Yebra et al., 2010b). The number of unassigned sequences was reduced in REGAv3 compared to REGAv2, by introducing the term "like". Although 74% of the sequences assigned as "like" were classified as a pure subtype by the *MPhy* in our analysis, 26% of the samples showed evidence of recombination; therefore, this terminology "like" is indeed useful, as it alerts the user to further verify these sequences. This helps to reduce the number of inaccurate assignments, while also reducing the number of unassigned sequences. Thus, REGAv3 uses the "subtype-like" assignment to indicate the most likely subtype for a particular strain, and at the same time to caution for potential discrepancies, thereby increasing the usefulness of the tool both for epidemiological statistical purposes where it is important to have as few as possible unassigned sequences, and for situations where correct assignment is more important.

One of the aims of the study was to give guidance as to which tool would perform well in a context of HIV-1 surveillance activities where the overall prevalence and spread of the epidemic is important to estimate (Hemelaar et al., 2011). We showed that no subtyping tool is able to classify all HIV-1 clades with a 100% accuracy, and we highlighted the difference in performance of the tools according to the subtype (or CRF) or the region analyzed such as PR, RT or PR + RT, such that our results can be directly compared with other studies (Holguin et al., 2008; Loveday et al., 2006; Yebra et al., 2010b). In general, our findings corroborate that phylogenetic-based, statistical-based tools and the similarity-based tool Stanford perform well for the most frequent subtypes worldwide such as B and C (Gifford et al., 2006; Hemelaar et al., 2011; Kosakovsky Pond et al., 2009; Myers et al., 2005; Schultz et al., 2009; Struck et al., 2010). However for other important clades such as A, D, F, G, CRF01_AE, CRF02_AG, COMET and REGAv3 correctly identified most strains while the remaining tools failed much more often.

Since we only had the *pol* region of the clinical dataset, we cannot exclude that other regions of the genome belong to other subtypes or CRFs, as has been reported in other cohorts (Abecasis et al., 2011; Njouom et al., 2003). However, overall, the performance of the tools was similar for the clinical and the LANL dataset, suggesting that for most subtypes and CRFs, our performance evaluation is valid. The exception is the assignment of subtype A and CRF06_cpx. For example, COMET correctly identified subtype A in 76% of the sequences from the clinical dataset and 100% of the LANL dataset, similarly this tool also identified only 18% of the CRF06_cpx in the clinical dataset and 100% in the LANL dataset (see supplementary Tables 2 and 3). The reason for this discrepancy has not been further investigated, but we suspect that by not taking into account the evolutionary relationship of the sequences, statistical-based tools like COMET are prone to overfitting on the training dataset (LANL).

Another reason for variation in the performance of subtyping tools is the analysis of PR and RT separately (Holguin et al., 2008; Loveday et al., 2006). Most of the tools had similar performance for the RT and the *pol* region. However, with regard to the PR region separately, statistical-based tools such as COMET and jpHMM had a higher performance than the other tools. These disagreements, for instance with REGA, occurred because short sequences with low phylogenetic signal were frequently reported as "unassigned" (Holguin et al., 2008). The philosophy of REGA is to avoid assigning sequences with low phylogenetic signal, with the aim to avoid false conclusions about evolutionary relationships (Revell et al., 2008; Strimmer and von Haeseler, 1997). Consequently, the user must be aware that short sequences require further examination to be correctly classified.

Current definitions of some subtypes and CRFs contribute to problems with the performance of automated tools. For example, the relationship between CRF02_AG and its parental strain A and G is still a matter of debate (Abecasis et al., 2007; Kosakovsky Pond et al., 2009; Zhang et al., 2010). For other CRFs, such as CRF02_AG, CRF07_BC, CRF08_BC, CRF12_BF, and CRF17_BF, atypical breakpoints were found in the sequences assigned to these CRFs in the Los Alamos database (Zhang et al., 2010), suggesting perhaps a wrong assignment in this database. In fact, the proper assignment of several CRFs can be disputed, for example when the recombinant region is so small that there is not sufficient phylogenetic signal for classification (e.g. around 100 nts for CRF12_BF, CRF20_BG, CRF35_A1D, CRF41_CD). Finally, several CRFs are so closely related to each other (e.g. CRF20_BG, CRF23_BG and CRF24_BG) that automated tools have great difficulty to discriminate between them. A thorough re-analysis of all CRFs is therefore urgently needed.

The absence of breakpoints in the region of study was another frequent cause of misclassifications; for instance CRF01_AE was classified as CRF15_01B (Tovanabutra et al., 2003) (see Fig. 2) and G was classified as CRF14_BG. There was no problem with the evaluation of the CRF01_AE *pol* sequences, but we had to question either the value of the manual phylogenetic analysis as gold standard for CRF14_BG, or the definition of CRF14_BG itself. The clinical dataset is derived from the country where this CRF originated, and the prevalence of the parent subtype G and of CRF14_BG as defined by manual phylogenetic analysis was very high. However we had access to *env* sequences from several patient isolates that were included in the *pol* dataset, and for several of these, there was discordance with the *pol* assignment, as has been reported before (Abecasis et al., 2011). In a joined phylogenetic analysis of all G and CRF14_BG sequences, the CRF14_BG sequences were not monophyletic but were spread among the subtype G sequences, and this was also the case for the LANL full genome CRF14_BG sequences. As a result, and only for the clinical dataset, we pooled

Table 4
Operational characteristics of subtyping tools.

Characteristics	Tools							
	Phylogenetic			Similarity		Statistical		
	REGAv3	REGAv2	SCUEAL	NCBI	STANF	COMET	jpHMM	STAR
Analysis of full genomes	+	+	–	+	–	+	+	+
Exact recombination breakpoints	+	–	+	*	–	–	+	–
Intra-subtype recombination	–	–	+	–	–	–	–	–
Latest CRF that can be analyzed†	CRF47_BF	CRF14_BG	CRF43_02G	CRF43_02G	CRF02_AG	CRF49_cpx	CRF01_AE	CRF02_AG
Batch analysis online	1000	1000	500	1	>100‡	10 Mb §	5 ‖	500
Waiting job queue	–	–	+	–	–	–	–	–
Average time for 500 sequences¶	~5 h	~4 h	~3 h	–	~5 min	sec	~2–7 h ‖	~15 min
Average time for 1 sequence**	~min	~min	~min	sec	sec	sec	sec-min	sec
Part of resistance analysis	–	–	–	–	+	–	–	–
Phylogenetic signal analysis	+	+	–	–	–	–	–	–
Summary table report	+	–	+	–	–	–	–	–
Graphical visualization of results	+	+	+	+	–	–	+	+
Position of sequence according to HXB2 reference	+	+	–	–	–	–	+	–
Download additional files (csv, txt, fasta, etc.)	+	+	+	+	+	+	+	+

*NCBI report shows an approximation of the breakpoints. †Some CRFs are excluded from the analysis, such as those with a limited number of strains available in LANL, or where the pol region cannot be discriminated from other CRFs (see http://bioafrica.mrc.ac.za/CRFs/CRFs.php). ‡Stanford is able to analyze up to 100 sequences at a time (character limit: 600,000). However there are other options to analyze more sequences. § COMET accepts files with a maximum size of 10 Mb (around 8000 sequences PR+RT). ‖jpHMM has the option to download a command line program without limit of batch analysis. In addition, there is an option to speed up the program. ¶We ran 500 sequences with pure subtypes and CRFs five times in different days. Then we calculated the average time. **We ran 5 times a recombinant, the average time for analysis of 1 sequence is about 1 min for phylogenetic-based tools and seconds for the other tools. *Abbreviations*: (+) characteristic available in the tool, (–) characteristic not available, (h) hours, (min) minutes, (sec) seconds.

CRF14_BG and subtype G in a single, but separate, classification "CRF14_BG or G".

Phylogenetic-based subtyping tools had lower reproducibility than similarity-based tools or statistical-based tools and the sequences causing this problem were not consistent across tools. The bootstrapping procedure in the tree-based algorithms is responsible for this lower reproducibility. Bootstrapping is a random process of resampling (Abecasis et al., 2010; Hillis and Bull, 1993), and each bootstrap sample is a different sample. For a more robust assignment, it would be needed to perform 1000 bootstrap samples, but this would cost too much computer time, and the phylogenetic-based tools are already slow. Other causes for the lower reproducibility were the introduction of new thresholds in the bootscan support of REGAv3 and the term "complex" in SCUEAL. This led sometimes to misidentification especially of subtype B as B/D recombinant because of the high similarity between these subtypes in the *pol* region (Leitner et al., 1995; Robertson et al., 2000).

Other considerations that influences the widespread use of subtyping tools are the operational characteristics (see Table 4). For example, the analysis of a bulk of sequences usually takes more time with phylogenetic-based tools than statistical-based tools (Abecasis et al., 2010; Kosakovsky Pond et al., 2009; Struck et al., 2010), and this is important for any application in a context of a large dataset. On the other hand, for the surveillance of the HIV-1 epidemic it is sometimes important to have information on recombination breakpoints, which are only shown in phylogenetic-based tools and jpHMM (Kosakovsky Pond et al., 2009; Schultz et al., 2009).

Although we included prevalent epidemiological non-B subtypes and CRFs, we acknowledge the limited number of samples available for subtypes H, J, K, CRF06_cpx, CRF07_BC and CRF12_BF, which prevents us from drawing firm conclusions for these subtypes and CRFs. We also did not evaluate all available tools, since many are based on similarity, and some have as their main objective the evaluation of antiretroviral resistance rather than subtyping (Beerenwinkel et al., 2003). We included the two most commonly used similarity-based tools, NCBI and Stanford (Rhee et al., 2006; Rozanov et al., 2004).

Other factors, that influence the performance of subtyping tools, are the high recombination rate of HIV-1 (Mansky and Temin, 1995) and human migration as determinants of global HIV dynamics (Rambaut et al., 2004). HIV-1 recombination increases the complexity and frequency of recombinant forms (Zhang et al., 2010), while migration has driven the dissemination of subtypes to new regions and established new epidemics (Pybus and Rambaut, 2009). As a consequence the subtyping tools should be regularly updated, especially tools which do not consider the intrinsic biologically relevant evolutionary relationships like statistical or similarity-based tools. The analysis of an epidemic where many subtypes or new CRFs are prevalent must be identified with COMET or phylogenetic-based tools that have an updated reference dataset (de Oliveira et al., 2005; Kosakovsky Pond et al., 2009; Struck et al., 2010).

5. Conclusions and recommendations

To our knowledge, this is the first study with an extensive comparison between subtyping tools, and manual phylogenetic analysis in PR, RT, PR + RT in two large datasets: a clinical dataset and a LANL dataset in which *pol* region was trimmed from full-length genomes. The performance of the new REGAv3 to identify subtype B and CRF02_AG in the *pol* region was much better than with REGAv2. REGAv3 had a very good performance in classifying pure subtypes, similar to that of COMET, jpHMM and SCUEAL, and it was also very good at identifying CRFs in the *pol* region, comparable to the best other tool, COMET. REGAv3 and COMET are currently the best available tools to automatically subtype HIV-1 sequences, however recombination breakpoint analysis is not possible with COMET. The performance of jpHMM is comparable but this tool has the big disadvantage that it does not classify CRFs, except for CRF01_AE.

We could draw some general recommendations from this analysis to use in future surveys of HIV-1 genetic diversity. First, automated tools might be useful for subtyping large *pol* datasets that are used in clinical and surveillance settings (Gifford et al., 2006; Kosakovsky Pond et al., 2009). Nevertheless, if accuracy is important, for example in individual patient follow-up or in detailed epidemiological analyses, it is necessary to use at least two subtyping tools whose overall performance is high in the genetic region analyzed such as COMET and REGAv3. This methodology has been previously used in different studies that required stringent analyses of large datasets (Faria et al., 2011; Hue et al., 2011; Jacobs et al., 2009; Yebra et al., 2010a). This comes at the cost of speed, which is determined by the slower of the two tools, the phylogenetic-based tool. The discordant sequences between the two tools can then be analyzed using manual phylogenetic analysis, still the gold standard. Second, for very short sequences such as PR, tools like COMET are recommended given that REGAv3 will give a considerable number of unassigned sequences, but only if accuracy is not a big issue. Therefore, we insist that for short sequences with low phylogenetic signal, such as PR, manual phylogenetic analysis is still needed (Strimmer and von Haeseler, 1997). Third, subtyping is often done in the context of an individual patient follow-up, using PR + RT sequences that are available from resistance genotyping. Thus it is often the case that resistance and subtyping are analyzed together in the clinical settings. Stanford does provide this information; however, the main goal of Stanford is to provide an accurate algorithm of resistance rather than subtyping (http://hivdb.stanford.edu/DR/asi/releaseNotes/index.html#hivdb_subtyping). If this tool is used, analysis of subtypes A, F and CRFs should be complemented with other statistical-based or phylogenetic-based tools. For example, REGAv3 is also on the same website. Fourth, the use of at least two automated tools to classify subtypes in the patient follow-up could also be useful, for example, for clinical collaborators that have little experience with manual analysis. However, if superinfection is suspected, phylogenetic analysis should be carried out (Ntemgwa et al., 2008).

Funding

Pineda-Peña AC was supported by a Doctoral Research Training Program "Francisco Jose de Caldas" by the Departamento Administrativo de Ciencia, Tecnología e Innovación, COLCIENCIAS, Republic of Colombia and ERACOL academic exchange in medicine and the health sciences, between Europe and Latin America. The research was supported in part by the European Community's Seventh Framework Programme (FP7/2007–2013) under the project "Collaborative HIV and Anti-HIV Drug Resistance Network (CHAIN)" grant agreement n° 223131; and by the Fonds voor Wetenschappelijk Onderzoek – Flanders (FWO) grant G.0611.09. Faria NR was supported by the Fundação para a Ciência e a Tecnologia (grant no. SFRH/BD/64530/2009). Abecasis A was supported by a post-doc fellowship by the Fundação para a Ciência e a Tecnologia (SFRH/BPD/65605/2009).

Acknowledgments

We would like to thank Jurgen Vercauteren, Joke Snoeck and the anonymous referees for their comments to this work and,

Soraya Menezes and Li Guangdi for their technical support. We acknowledge the Portuguese HIV-1 Resistance Study Group for providing us the clinical dataset. Part of the results of this paper has been presented at the 17th International Bioinformatics Workshop on Virus Evolution and Molecular Epidemiology, Belgrade, Serbia 2012 (http://regaweb.med.kuleuven.be/workshop) and the Second BREACH symposium (Belgian Research on AIDS and HIV Consortium), Brussels, Belgium 2012.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.meegid.2013.04.032.

References

Abecasis, A., Wensing, A.M., Vercauteren, J., Paraskevis, D., Van de Vijver, D., Albert, J., Asjo, B., Balotta, C., Bruckova, M., Camacho, R., Coughlan, S., Grossman, Z., Hamouda, O., Hatzakis, A., Horban, A., Korn, K., Kostrikis, L., Kucherer, C., Nielsen, C., Poljak, M., Puchhammer-Stockl, E., Riva, C., Ruiz, L., Salminen, M., Schmit, J.C., Schuurman, R., Sonnerborg, A., Stanekova, D., Stanojevic, M., Struck, D., Boucher, C.A., Vandamme, A.M., on behalf of the SPREAD-programme, 2008. HIV-1 genetic diversity in Europe and its demographic predictors. Demographic determinants of HIV-1 subtype distribution in Europe, 6th European HIV Drug Resistance Workshop, Budapest, Hungary.

Abecasis, A.B., Deforche, K., Bacheler, L.T., McKenna, P., Carvalho, A.P., Gomes, P., Vandamme, A.M., Camacho, R.J., 2006. Investigation of baseline susceptibility to protease inhibitors in HIV-1 subtypes C, F, G and CRF02_AG. Antiviral Therapy 11, 581–589.

Abecasis, A.B., Lemey, P., Vidal, N., de Oliveira, T., Peeters, M., Camacho, R., Shapiro, B., Rambaut, A., Vandamme, A.M., 2007. Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: subtype G is a circulating recombinant form. Journal of Virology 81, 8543–8551.

Abecasis, A.B., Martins, A., Costa, I., Carvalho, A.P., Diogo, I., Gomes, P., Camacho, R.J., 2011. Molecular epidemiological analysis of paired *pol*/*env* sequences from Portuguese HIV type 1 patients. AIDS Research and Human Retroviruses 27, 803–805.

Abecasis, A.B., Wang, Y., Libin, P., Imbrechts, S., de Oliveira, T., Camacho, R.J., Vandamme, A.M., 2010. Comparative performance of the REGA subtyping tool version 2 versus version 1. Infection, Genetics and Evolution 10, 380–385.

Abecasis, A.B., Wensing, A.M., Paraskevis, D., Vercauteren, J., Theys, K., Van de Vijver, D.A., Albert, J., Asjo, B., Balotta, C., Beshkov, D., Camacho, R.J., Clotet, B., De Gascun, C., Griskevicius, A., Grossman, Z., Hamouda, O., Horban, A., Kolupajeva, T., Korn, K., Kostrikis, L.G., Kucherer, C., Liitsola, K., Linka, M., Nielsen, C., Otelea, D., Paredes, R., Poljak, M., Puchhammer-Stockl, E., Schmit, J.C., Sonnerborg, A., Stanekova, D., Stanojevic, M., Struck, D., Boucher, C.A., Vandamme, A.M., 2013. HIV-1 subtype distribution and its demographic determinants in newly diagnosed patients in Europe suggest highly compartmentalized epidemics. Retrovirology 10, 7.

Alcantara, L.C., Cassol, S., Libin, P., Deforche, K., Pybus, O.G., Van Ranst, M., Galvao-Castro, B., Vandamme, A.M., de Oliveira, T., 2009. A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. Nucleic Acids Research 37, W634–642.

Baeten, J.M., Chohan, B., Lavreys, L., Chohan, V., McClelland, R.S., Certain, L., Mandaliya, K., Jaoko, W., Overbaugh, J., 2007. HIV-1 subtype D infection is associated with faster disease progression than subtype A in spite of similar plasma HIV-1 loads. The Journal of Infectious Diseases 195, 1177–1180.

Banoo, S., Bell, D., Bossuyt, P., Herring, A., Mabey, D., Poole, F., Smith, P.G., Sriram, N., Wongsrichanalai, C., Linke, R., O'Brien, R., Perkins, M., Cunningham, J., Matsoso, P., Nathanson, C.M., Olliaro, P., Peeling, R.W., Ramsay, A., 2010. Evaluation of diagnostic tests for infectious diseases: general principles. Nature Reviews Microbiology 8, S17–29.

Beerenwinkel, N., Daumer, M., Oette, M., Korn, K., Hoffmann, D., Kaiser, R., Lengauer, T., Selbig, J., Walter, H., 2003. Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes. Nucleic Acids Research 31, 3850–3855.

Brenner, B.G., Oliveira, M., Doualla-Bell, F., Moisi, D.D., Ntemgwa, M., Frankel, F., Essex, M., Wainberg, M.A., 2006. HIV-1 subtype C viruses rapidly develop K65R resistance to tenofovir in cell culture. AIDS (London, England) 20, F9–F13.

Camacho, R.J., Vandamme, A.M., 2007. Antiretroviral resistance in different HIV-1 subtypes: impact on therapy outcomes and resistance testing interpretation. Current Opinion in HIV and AIDS 2, 123–129.

Carr, J.K., Salminen, M.O., Koch, C., Gotte, D., Artenstein, A.W., Hegerich, P.A., St Louis, D., Burke, D.S., McCutchan, F.E., 1996. Full-length sequence and mosaic structure of a human immunodeficiency virus type 1 isolate from Thailand. Journal of Virology 70, 5935–5943.

de Oliveira, T., Deforche, K., Cassol, S., Salminen, M., Paraskevis, D., Seebregts, C., Snoeck, J., van Rensburg, E.J., Wensing, A.M., van de Vijver, D.A., Boucher, C.A., Camacho, R., Vandamme, A.M., 2005. An automated genotyping system for analysis of HIV-1 and other microbial sequences. Bioinformatics (Oxford, England) 21, 3797–3800.

Delgado, E., Thomson, M.M., Villahermosa, M.L., Sierra, M., Ocampo, A., Miralles, C., Rodriguez-Perez, R., Diz-Aren, J., Ojea-de Castro, R., Losada, E., Cuevas, M.T., Vazquez-de Parga, E., Carmona, R., Perez-Alvarez, L., Medrano, L., Cuevas, L., Taboada, J.A., Najera, R., 2002. Identification of a newly characterized HIV-1 BG intersubtype circulating recombinant form in Galicia, Spain, which exhibits a pseudotype-like virion structure. Journal of Acquired Immune Deficiency Syndromes 29, 536–543.

Faria, N.R., Suchard, M.A., Abecasis, A., Sousa, J.D., Ndembi, N., Bonfim, I., Camacho, R.J., Vandamme, A.M., Lemey, P., 2011. Phylodynamics of the HIV-1 CRF02_AG clade in Cameroon. Infection, Genetics and Evolution 12, 453–460.

Gifford, R., de Oliveira, T., Rambaut, A., Myers, R.E., Gale, C.V., Dunn, D., Shafer, R., Vandamme, A.M., Kellam, P., Pillay, D., 2006. Assessment of automated genotyping protocols as tools for surveillance of HIV-1 genetic diversity. AIDS (London, England) 20, 1521–1529.

Gouy, M., Guindon, S., Gascuel, O., 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular Biology and Evolution 27, 221–224.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology 59, 307–321.

Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41, 95–98.

Hemelaar, J., Gouws, E., Ghys, P.D., Osmanov, S., 2011. Global trends in molecular epidemiology of HIV-1 during 2000–2007. AIDS (London, England) 25, 679–689.

Hillis, D.M., Bull, J.J., 1993. An empirical-test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Systematic Biology 42, 182–192.

Holguin, A., Lopez, M., Soriano, V., 2008. Reliability of rapid subtyping tools compared to that of phylogenetic analysis for characterization of human immunodeficiency virus type 1 non-B subtypes and recombinant forms. Journal of Clinical Microbiology 46, 3896–3899.

Hue, S., Hassan, A.S., Nabwera, H., Sanders, E.J., Pillay, D., Berkley, J.A., Cane, P.A., 2011. HIV type 1 in a rural coastal town in kenya shows multiple introductions with many subtypes and much recombination. AIDS Research and Human Retroviruses 27.

Jacobs, G.B., Loxton, A.G., Laten, A., Robson, B., van Rensburg, E.J., Engelbrecht, S., 2009. Emergence and diversity of different HIV-1 subtypes in South Africa, 2000–2001. Journal of Medical Virology 81, 1852–1859.

Jetzt, A.E., Yu, H., Klarmann, G.J., Ron, Y., Preston, B.D., Dougherty, J.P., 2000. High rate of recombination throughout the human immunodeficiency virus type 1 genome. Journal of Virology 74, 1234–1240.

Kosakovsky Pond, S.L., Posada, D., Stawiski, E., Chappey, C., Poon, A.F., Hughes, G., Fearnhill, E., Gravenor, M.B., Leigh Brown, A.J., Frost, S.D., 2009. An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1. PLoS Computational Biology 5, e1000581.

Kuhner, M.K., Felsenstein, J., 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Molecular Biology and Evolution 11, 459–468.

Leitner, T., Escanilla, D., Franzen, C., Uhlen, M., Albert, J., 1996. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proceedings of the National Academy of Sciences of the United States of America 93, 10864–10869.

Leitner, T., Escanilla, D., Marquina, S., Wahlberg, J., Brostrom, C., Hansson, H.B., Uhlen, M., Albert, J., 1995. Biological and molecular characterization of subtype D, G, and A/D recombinant HIV-1 transmissions in Sweden. Virology 209, 136–146.

Liu, T.F., Shafer, R.W., 2006. Web resources for HIV type 1 genotypic-resistance test interpretation. Clinical Infectious Diseases 42, 1608–1618.

Lole, K.S., Bollinger, R.C., Paranjape, R.S., Gadkari, D., Kulkarni, S.S., Novak, N.G., Ingersoll, R., Sheppard, H.W., Ray, S.C., 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. Journal of Virology 73, 152–160.

Loveday, C., MacRae, E., on behalf of the ICVC Clinical Collaborative Research Group, 2006. Limitations in using online tool to determine HIV-1 subtype in clinical patients: a comparison of 5 tools, XV International HIV Drug Resistance Workshop: Basic Principles & Clinical Implications. Antiviral Therapy, Sitges, Spain, p. S129.

Mansky, L.M., Temin, H.M., 1995. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. Journal of Virology 69, 5087–5094.

Martinez-Cajas, J.L., Pant-Pai, N., Klein, M.B., Wainberg, M.A., 2008. Role of genetic diversity amongst HIV-1 non-B subtypes in drug resistance: a systematic review of virologic and biochemical evidence. AIDS Reviews 10, 212–223.

Myers, R.E., Gale, C.V., Harrison, A., Takeuchi, Y., Kellam, P., 2005. A statistical model for HIV-1 sequence classification using the subtype analyser (STAR). Bioinformatics (Oxford, England) 21, 3535–3540.

Njouom, R., Pasquier, C., Sandres-Saune, K., Harter, A., Souyris, C., Izopet, J., 2003. Assessment of HIV-1 subtyping for Cameroon strains using phylogenetic analysis of pol gene sequences. Journal of Virology Methods 110, 1–8.

Ntemgwa, M., Gill, M.J., Brenner, B.G., Moisi, D., Wainberg, M.A., 2008. Discrepancies in assignment of subtype/recombinant forms by genotyping programs for HIV type 1 drug resistance testing may falsely predict superinfection. AIDS Research and Human Retroviruses 24, 995–1002.

Pasquier, C., Millot, N., Njouom, R., Sandres, K., Cazabat, M., Puel, J., Izopet, J., 2001. HIV-1 subtyping using phylogenetic analysis of pol gene sequences. Journal of Virology Methods 94, 45–54.

Posada, D., 2008. JModelTest: phylogenetic model averaging. Molecular Biology and Evolution 25, 1253–1256.

Posada, D., Crandall, K.A., 2001. Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1). Molecular Biology and Evolution 18, 897–906.

Pybus, O.G., Rambaut, A., 2009. Evolutionary analysis of the dynamics of viral infectious disease. Nature Reviews Genetics 10, 540–550.

Rambaut, A., Posada, D., Crandall, K.A., Holmes, E.C., 2004. The causes and consequences of HIV evolution. Nature Reviews Genetics 5, 52–61.

Revell, L.J., Harmon, L.J., Collar, D.C., 2008. Phylogenetic signal, evolutionary process, and rate. Systematic Biology 57, 591–601.

Rhee, S.Y., Kantor, R., Katzenstein, D.A., Camacho, R., Morris, L., Sirivichayakul, S., Jorgensen, L., Brigido, L.F., Schapiro, J.M., Shafer, R.W., 2006. HIV-1 pol mutation frequency by subtype and treatment experience: extension of the HIVseq program to seven non-B subtypes. AIDS (London, England) 20, 643–651.

Robertson, D.L., Anderson, J.P., Bradac, J.A., Carr, J.K., Foley, B., Funkhouser, R.K., Gao, F., Hahn, B.H., Kalish, M.L., Kuiken, C., Learn, G.H., Leitner, T., McCutchan, F., Osmanov, S., Peeters, M., Pieniazek, D., Salminen, M., Sharp, P.M., Wolinsky, S., Korber, B., 2000. HIV-1 nomenclature proposal. Science (New York, NY) 288, 55–56.

Rozanov, M., Plikat, U., Chappey, C., Kochergin, A., Tatusova, T., 2004. A web-based genotyping resource for viral sequences. Nucleic Acids Research 32, W654–659.

Schmidt, H.A., Strimmer, K., Vingron, M., von Haeseler, A., 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics (Oxford, England) 18, 502–504.

Schultz, A.K., Zhang, M., Bulla, I., Leitner, T., Korber, B., Morgenstern, B., Stanke, M., 2009. JpHMM: improving the reliability of recombination prediction in HIV-1. Nucleic Acids Research 37, W647–W651.

Snoeck, J., Van Dooren, S., Van Laethem, K., Derdelinckx, I., Van Wijngaerden, E., De Clercq, E., Vandamme, A.M., 2002. Prevalence and origin of HIV-1 group M subtypes among patients attending a Belgian hospital in 1999. Virus Research 85, 95–107.

Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics (Oxford, England) 22, 2688–2690.

Strimmer, K., von Haeseler, A., 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proceedings of the National Academy of Sciences of the United States of America 94, 6815–6819.

Struck, D., Perez-Bercoff, D., Devaux, C., Schmit, J.C., 2010. COMET: a novel approach to HIV-1 subtype prediction, 8th European HIV Drug Resistance Workshop, Sorrento, Italy.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution 28, 2731–2739.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22, 4673–4680.

Thompson, M.A., Aberg, J.A., Hoy, J.F., Telenti, A., Benson, C., Cahn, P., Eron, J.J., Gunthard, H.F., Hammer, S.M., Reiss, P., Richman, D.D., Rizzardini, G., Thomas, D.L., Jacobsen, D.M., Volberding, P.A., 2012. Antiretroviral treatment of adult HIV infection: 2012 recommendations of the International Antiviral Society-USA panel. JAMA 308, 387–402.

Tovanabutra, S., Watanaveeradej, V., Viputtikul, K., De Souza, M., Razak, M.H., Suriyanon, V., Jittiwutikarn, J., Sriplienchan, S., Nitayaphan, S., Benenson, M.W., Sirisopana, N., Renzullo, P.O., Brown, A.E., Robb, M.L., Beyrer, C., Celentano, D.D., McNeil, J.G., Birx, D.L., Carr, J.K., McCutchan, F.E., 2003. A new circulating recombinant form, CRF15_01B, reinforces the linkage between IDU and heterosexual epidemics in Thailand. AIDS Research and Human Retroviruses 19, 561–567.

UNAIDS-WHO, 2012. Global Report: UNAIDS report on the global AIDS epidemic, Geneva.

Vandamme, A.M., Camacho, R.J., Ceccherini-Silberstein, F., de Luca, A., Palmisano, L., Paraskevis, D., Paredes, R., Poljak, M., Schmit, J.C., Soriano, V., Walter, H., Sonnerborg, A., 2011. European recommendations for the clinical use of HIV drug resistance testing: 2011 update. AIDS Reviews 13, 77–108.

Wainberg, M.A., Brenner, B.G., 2010. Role of HIV subtype diversity in the development of resistance to antiviral drugs. Viruses 2, 2493–2508.

Yahi, N., Fantini, J., Tourres, C., Tivoli, N., Koch, N., Tamalet, C., 2001. Use of drug resistance sequence data for the systematic detection of non-B human immunodeficiency virus type 1 (HIV-1) subtypes: how to create a sentinel site for monitoring the genetic diversity of HIV-1 at a country scale. The Journal of Infectious Diseases 183, 1311–1317.

Yebra, G., de Mulder, M., del Romero, J., Rodriguez, C., Holguin, A., 2010a. HIV-1 non-B subtypes: High transmitted NNRTI-resistance in Spain and impaired genotypic resistance interpretation due to variability. Antiviral Research 85, 409–417.

Yebra, G., de Mulder, M., Martin, L., Perez-Cachafeiro, S., Rodriguez, C., Labarga, P., Garcia, F., Tural, C., Jaen, A., Navarro, G., Holguin, A., 2010b. Sensitivity of seven HIV subtyping tools differs among subtypes/recombinants in the Spanish cohort of naive HIV-infected patients (CoRIS). Antiviral Research 89, 19–25.

Zhang, M., Foley, B., Schultz, A.K., Macke, J.P., Bulla, I., Stanke, M., Morgenstern, B., Korber, B., Leitner, T., 2010. The role of recombination in the emergence of a complex and dynamic HIV epidemic. Retrovirology 7, 25.