

# Comparing Three HIV-1 Subtyping Tools With a Novel Phylogenetic-based Method

William M. Switzer<sup>1</sup>, Yi Pan<sup>1</sup>, Neeraja Saduvala<sup>1</sup>, Tianchi Zhang<sup>1</sup>, Angela Hernandez<sup>1</sup>, Pieter Libin<sup>2</sup>, Daniel Struck<sup>3</sup>, Tulio de Oliveira<sup>4</sup>, Annemieke Vandamme<sup>2</sup>, Joel Wertheim<sup>5</sup>, Alexandra M. Oster<sup>1</sup>

<sup>1</sup>Division of HIV/AIDS Prevention, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, GA, 30329; <sup>2</sup>Department of Microbiology and Immunology, Rega Institute for Medical Research, Leuven, Belgium; <sup>3</sup>Laboratory of Retrovirology, Val Fleuri, Luxembourg; <sup>4</sup>College of Health Sciences, University of KwaZulu-Natal, Durban, South Africa, <sup>5</sup>Center for AIDS Research, University of California, San Diego, CA 9210

## Background

- HIV-1 evolves rapidly, increasing its genetic diversity and complexity, and is classified into four distinct lineages (groups M, N, O, P), with group M containing >80 subtypes and circulating recombinant forms (CRFs)<sup>1</sup>.
- Subtype determination is important epidemiologically, and subtype can impact pathogenesis, treatment, and vaccine development.
- HIV polymerase (*pol*) sequences are routinely generated for newly diagnosed patients in care for determining optimal antiviral treatment and are available for further analyses.
- Automated HIV subtyping tools are available but can give discordant results, complicating the final interpretation.
- Phylogenetic clustering is a powerful tool to investigate viral diversity.
- Also, some tools and phylogenetic methods can be computationally intensive, a challenge for analyzing large datasets.
- Automated methods have not been fully evaluated in a predominantly subtype B setting or with surveillance data.
- We evaluated three existing tools and a novel method to rapidly and accurately infer subtypes for large datasets.

## Methods

- We determined subtypes for 69,094 *pol* sequences > 500-bp in length reported to the U.S. National HIV Surveillance System between 2001-2014 using three automated tools (REGA V3, SCUEAL (Subtype Classification Using Evolutionary Algorithms), and COMET (Context-based Modeling for Expeditious Typing)).<sup>2-4</sup>
- Sequences with ≥ 5% ambiguities and containing integrase only were excluded from the analyses.
- Sequences were further analyzed phylogenetically using a novel method that combines FastTree maximum likelihood (FML)<sup>5</sup> inference with 2,863 curated reference *pol* sequences from LANL representing 106 unique subtypes and CRFs.<sup>6</sup>
- FML uses the Shimodaira-Hasegawa (SH) test to assess node reliability.
- FML analyses were performed on multiple sequence alignment based on the fast Fourier transform (MAFFT)<sup>7</sup> alignments in batches of 3,000 – 10,000 patient sequences per run to minimize computational intensity.
- To facilitate subtype classification of these very large, visually dense FML trees we used cluster analysis with the program Phylopart<sup>8</sup> (Fig. 1).
- Sequences not clustering with a reference sequence in FML were classified as unique recombinant forms (URFs).
- Trees were manually inspected to confirm Phylopart clusters.
- Selected sequences were analyzed by SimPlot to confirm the URF classification by FML.
- To facilitate comparisons across methods, sub-subtypes were condensed into respective subtypes.
- Subtypes with less than 10 sequences by one or more method were collapsed into “Rare Subtypes”.
- SAS v9.3 was used to:
  - calculate agreement between each pair of subtyping methods (% of sequences with same assignment by both methods)
  - calculate Cohen’s kappa coefficient of agreement for each pair of methods for subtypes with greater than 10 sequences.<sup>9</sup>

## Results

- The majority of sequences were classified as subtype B by each of the 4 methods (Table 1).
- Of all pairwise comparisons, the highest agreement occurred between COMET and FML (Table 2).
- Cohen’s kappa statistic demonstrated substantial to excellent agreement for classification of the major subtypes by each method with the exception of SCUEAL having difficulty with subtypes B, CRF02\_AG, and CRF06, evidenced by poorer agreement with other methods (Table 3).
- Only COMET and FML demonstrated excellent agreement for subtype B.
- Poor agreement was also observed with all methods for URFs and rare subtypes in this population (BF1, BG, H, J, K, O, CRF05, CRF07-19, CRF21-23, CRF25-27, CRF29, CRF31, CRF33, CRF35, CRF37, CRF39, CRF40, CRF42-45, CRF48-50, CRF54, CRF59)
- Rare subtypes and URFs were successfully identified by manual inspection of the FML tree.
- URF classification by FML was confirmed in selected sequences using SimPlot analysis (Fig. 2).

Fig.1. FML Novel HIV-1 subtyping algorithm

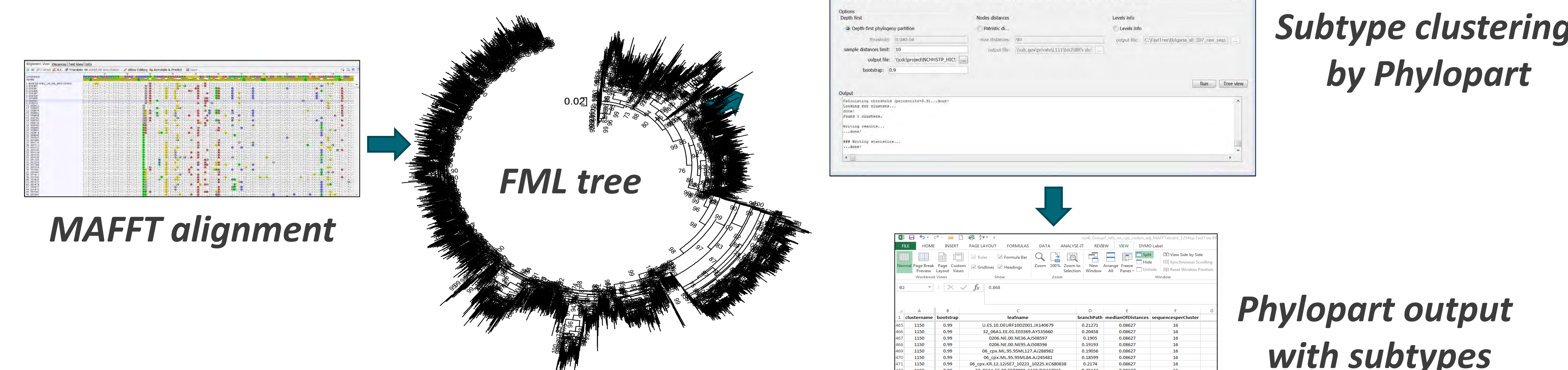


Table 1. Classification (%) of HIV-1 Subtypes in the U.S. by Each Method

HIV subtypes	Marginal Distribution (%)			
	COMET	SCUEAL	REGA	FastTree
A	0.52	0.44	0.57	0.52
B	94.9	91.4	92.4	95.4
C	1.32	1.26	1.35	1.27
CRF01_AE	0.35	0.29	0.34	0.36
CRF02_AG	1.01	0.2	0.97	1
CRF06	0.08	0.04	0.09	0.09
CRF20	0.03	0.02	0.03	0.03
CRF24	0.03	0.02	0.03	0.03
D	0.14	0.11	0.12	0.11
F	0.05	0.05	0.06	0.05
G	0.2	0.27	0.21	0.19
URF	1.24	5.69	3.57	0.62
Rare Subtypes	0.16	0.21	0.24	0.26
Total	100.03	100	99.98	99.93

Table 2. Overall agreement between subtyping methods (%)

Method	COMET	SCUEAL	REGA	FML
COMET	-	93.7	95.7	98.5
SCUEAL	93.7	-	91.6	94.2
REGA	95.7	91.6	-	96.1
FML	98.5	94.2	96.1	-

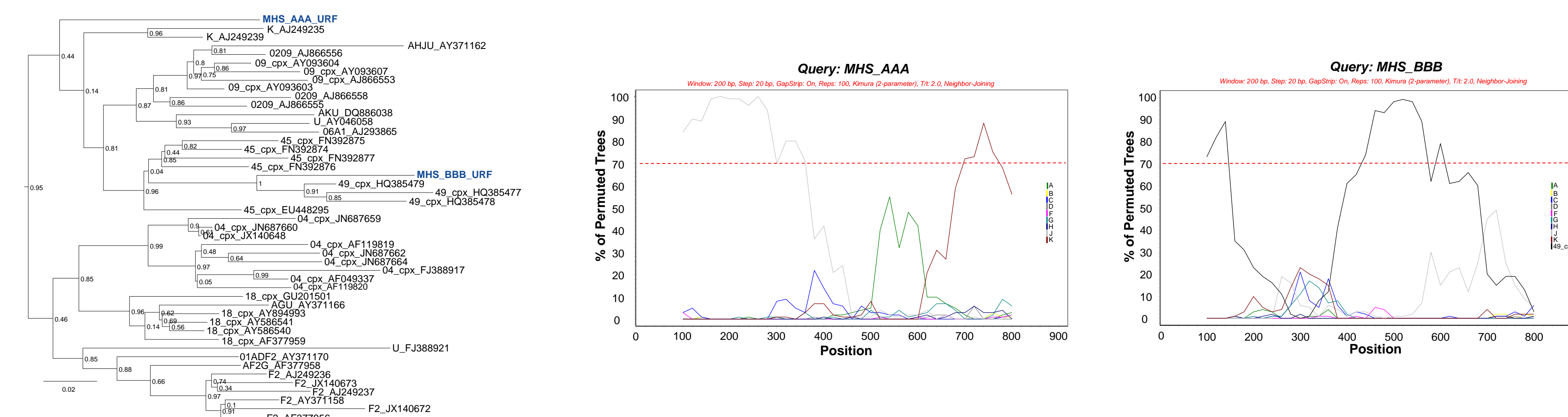
Table 3. Subtype classification agreement between methods using Kappa statistics

Subtype/CRF	COMET vs SCUEAL	COMET vs REGA	COMET vs FML	SCUEAL vs REGA	SCUEAL vs FML	REGA vs FML
A	0.89 (0.87, 0.92)	0.90 (0.88, 0.93)	0.93 (0.91, 0.95)	0.85 (0.83, 0.88)	0.89 (0.87, 0.92)	0.90 (0.88, 0.93)
B	0.62 (0.60, 0.63)	0.67 (0.66, 0.68)	0.89 (0.88, 0.90)	0.52 (0.51, 0.54)	0.64 (0.63, 0.65)	0.70 (0.69, 0.71)
C	0.97 (0.96, 0.98)	0.984 (0.98, 0.99)	0.97 (0.96, 0.98)	0.97 (0.96, 0.97)	0.96 (0.95, 0.97)	0.97 (0.96, 0.97)
CRF01_AE	0.89 (0.86, 0.92)	0.97 (0.96, 0.99)	0.98 (0.97, 0.99)	0.87 (0.84, 0.91)	0.88 (0.85, 0.91)	0.97 (0.96, 0.99)
CRF02_AG	0.32 (0.28, 0.36)	0.95 (0.94, 0.96)	0.95 (0.94, 0.96)	0.33 (0.29, 0.37)	0.31 (0.27, 0.35)	0.95 (0.94, 0.96)
CRF06	0.63 (0.50, 0.75)	0.94 (0.90, 0.98)	0.90 (0.85, 0.96)	0.60 (0.48, 0.73)	0.57 (0.45, 0.69)	0.93 (0.88, 0.97)
CRF20	0.75 (0.58, 0.92)	0.95 (0.87, 1.0)	0.93 (0.84, 1.0)	0.75 (0.58, 0.92)	0.74 (0.58, 0.91)	0.93 (0.84, 1.00)
CRF24	0.83 (0.70, 0.96)	1	1	0.83 (0.70, 0.96)	0.83 (0.70, 0.96)	1
D	0.85 (0.79, 0.91)	0.88 (0.83, 0.93)	0.85 (0.79, 0.91)	0.89 (0.84, 0.94)	0.91 (0.86, 0.96)	0.90 (0.86, 0.95)
F	0.96 (0.91, 1.00)	0.92 (0.86, 0.98)	0.90 (0.82, 0.97)	0.90 (0.84, 0.98)	0.85 (0.76, 0.94)	0.82 (0.73, 0.92)
G	0.78 (0.73, 0.83)	0.91 (0.88, 0.95)	0.95 (0.92, 0.98)	0.79 (0.74, 0.84)	0.79 (0.74, 0.84)	0.95 (0.92, 0.98)
URF	0.11 (0.10, 0.12)	0.12 (0.11, 0.14)	0.28 (0.25, 0.31)	0.09 (0.07, 0.10)	0.09 (0.07, 0.10)	0.11 (0.10, 0.13)
Rare Subtypes	0.30 (0.23, 0.38)	0.54 (0.47, 0.61)	0.49 (0.42, 0.56)	0.28 (0.22, 0.35)	0.35 (0.28, 0.42)	0.43 (0.36, 0.49)

0.0-0.2	poor agreement
0.2-0.4	fair agreement
0.4-0.6	moderate agreement
0.6-0.8	substantial agreement
0.8-1.0	excellent agreement

Cohen’s kappa compares the observed proportion of in care to the expected proportion of agreement, assuming that the distributions of the observer’s responses are independent\*. When =1, perfect agreement is achieved; and =0 indicates lack of agreement.

Fig.2. SimPlot analysis of URF classification by FML



## Summary and Conclusions

- We compared three automated methods and a novel cluster-based pipeline with fast maximum likelihood tree construction using a large U.S. HIV-1 *pol* sequence dataset.
- Expectedly, the majority of sequences were classified as subtype B, followed by subtypes C, CRF02\_AG, A, CRF01\_AE, D, miscellaneous rare subtypes and URFs.
- Of the automated tools, the COMET algorithm had excellent agreement with the FML method, whereas SCUEAL had the least agreement with all other methods.
- The only two methods with excellent agreement for subtype B were COMET and FML.
- SCUEAL had difficulty identifying CRF02\_AG subtypes and rare subtypes.
- Identification of URFs and rare subtypes had poor or fair agreement among all methods, though manual inspection of the FML trees with SimPlot analysis can resolve the subtype classification.
- REGA and COMET shared the most agreement with rare subtypes followed by COMET and FML, which also had the greatest agreement for identification of URFs.
- We identify a novel bioinformatics pipeline that combines phylogenetic tree construction with cluster analysis for rapid and accurate HIV-1 genotyping of large datasets.

## References

- Hemelaar J. Trends Mol Med. 2012 Mar;18(3):182-92.
- Pineda-Pena A-C, et al. Inf Gen Evol. 2013 Oct; 19:337-48.
- Pond SLK, et al. PLOS Comp Biol 2009 Nov; 5:1-21.
- Struck D, et al. Nucleic Acids Res 2014 Oct; 42(18):e144.
- Price MN, Dehal PS, Arkin AP. PLoS One. 2010 Mar 10;5(3):e9490.
- www.hiv.lanl.gov
- Katoh K, Standley DM, Mol Biol Evol. 2013 Apr;30(4):772-80.
- Prosperi MC, et al. Nat Comm. 2011; 2:321.
- Cohen J. Ed Phys Meas. 1960; 20:37-46.

## CONTACT INFO

Bill Switzer  
Retrovirus Surveillance Activity Lead  
Laboratory Branch  
Division of HIV/AIDS Prevention  
Centers for Disease Control and Prevention  
1600 Clifton Road, Mailstop G-45, Atlanta, GA 30329  
404-639-0219 (Phone), 404-639-0092 (FAX)  
Email:bis3@cdc.gov

