Using nearly full-genome HIV sequence data improves phylogeny reconstruction in a simulated epidemic

Citation for published version:
Yebra, G, Hodcroft, EB, Ragonnet-Cronin, ML, Pillay, D, Brown, AJL & PANGEA_HIV Consortium 2016, 'Using nearly full-genome HIV sequence data improves phylogeny reconstruction in a simulated epidemic: Length of HIV sequence data and phylogeny reconstruction' Scientific Reports, vol. 6, pp. 39489. DOI: 10.1038/srep39489

Digital Object Identifier (DOI):
10.1038/srep39489

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Scientific Reports

Using nearly full-genome HIV sequence data improves phylogeny reconstruction in a simulated epidemic

Gonzalo Yebra[1,*], Emma B. Hodcroft[1], Manon Ragonnet-Cronin[1], Deenan Pillay[2] & Andrew J. Leigh Brown[1] on behalf of the PANGEA_HIV Consortium & the ICONIC Project.

[1] Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK

[2] Wellcome Trust-Africa Centre for Health and Population Studies, University of KwaZulu-Natal, Durban, South Africa

Short title: Length of HIV sequence data and phylogeny reconstruction

*** Corresponding author:**
E-mail: Gonzalo.Yebra@ed.ac.uk (GY)

Abstract

HIV molecular epidemiology studies analyse viral *pol* gene sequences due to their availability, but whole genome sequencing allows to use other genes. We aimed to determine what gene(s) provide(s) the best approximation to the real phylogeny by analysing a simulated epidemic (created as part of the PANGEA_HIV project) with a known transmission tree.

We sub-sampled a simulated dataset of 4662 sequences into different combinations of genes (*gag-pol-env*, *gag-pol*, *gag*, *pol*, *env* and partial *pol*) and sampling depths (100%, 60%, 20% and 5%), generating 100 replicates for each case. We built maximum-likelihood trees for each combination using RAxML (GTR+Γ), and compared their topologies to the corresponding true tree's using CompareTree.

The accuracy of the trees was significantly proportional to the length of the sequences used, with the *gag-pol-env* datasets showing the best performance and *gag* and partial *pol* sequences showing the worst. The lowest sampling depths (20% and 5%) greatly reduced the accuracy of tree reconstruction and showed high variability among replicates, especially when using the shortest gene datasets.

In conclusion, using longer sequences derived from nearly whole genomes will improve the reliability of phylogenetic reconstruction. With low sample coverage, results can be highly variable, particularly when based on short sequences.

Background

Most studies on HIV molecular epidemiology now use the portion of the viral *pol* gene that contains the protease (PR) and reverse transcriptase (RT) coding regions. This is because these partial *pol* sequences (around 1.3Kb long) are routinely sequenced for genotypic resistance testing[1-3]. Although initially the *env* gene was considered to present the strongest phylogenetic signal, it was argued that some *env* fragments were too short and/or variable for a robust analysis[4]. After *pol* was demonstrated to accurately reconstruct HIV transmission[5], its analysis for phylogenetic studies became the standard owing to the very large datasets available for analysis (e.g., the UK[6] and Swiss[7] sequence databases). In the last few years, the increasing availability of HIV whole genome sequences has made possible the analysis of other genetic regions, which has raised discussion about whether full-length genome trees should be used or which viral genes provide the best trees.

A few studies have previously approached this question by analysing HIV transmission networks in which the timing and direction of transmission were known[8-11]. They have suggested that the combination of more than one gene provides the best estimation of the true tree. However, all were limited to very few patients and, in some cases, short nucleotide sequences. The lack of a known, large phylogeny prevents providing a definitive comparison that would answer this question, but simulated data provide an approximation that allows having both the true tree and a recombination-free dataset.

Such data were generated in the context of the PANGEA_HIV Methods Comparison Exercise[12] (http://www.pangea-hiv.org), for which an HIV epidemic in an African village was simulated using an agent-based model in which all sexual contacts were recorded, and those that gave rise to transmissions created a transmission tree which was recorded. Here, we used these HIV datasets to evaluate the effect of utilising viral sequence datasets of different length and from several viral genes and with different sampling depths to reconstruct the known simulated phylogenies.

Results

From the simulated HIV sequence data generated for the PANGEA_HIV project, we produced different combinations of sampling density (100%, 60%, 20% and 5%) and viral gene use (*gag-pol-env*, *gag-pol*, *gag*, *pol*, *env* and partial *pol*). Sixty per cent represents approximately the sampling coverage in the UK HIV Drug Resistance Database[13], whereas 5% represent the range in HIV sequence coverage that is believed to be relevant for cohorts in many African countries. For example, in the region of KwaZulu-Natal, South Africa, the sampling density is estimated to be between 4% and 8%, according to the specific cohort, (Prof. Tulio de Oliveira, pers. comm.). This sub-sampling was randomly replicated 100 times and ML trees were constructed, whose topology was then compared to that of the corresponding true tree. The results of the CompareTree metric (**Figure 1A**) show that the proportion of correct tree splits increased with the length of the sequences used. The genome datasets showed the best performance considering all the sampling coverage levels together (**Table 1**), with an average metric value of 0.965 (95% confidence interval (CI) = 0.964-0.966). It was closely followed by *gag-pol* (0.951 [0.950-0.952]), *pol* (0.934 [0.933-0.935]) and *env* (0.932 [0.930-0.933]) in that order. The smaller *gag* (0.879 [0.877-0.880]) and partial *pol* (0.867 [0.866-0.869]) sequences showed the worst performances.

Thus, the proportion of correct tree splits increased in direct proportion to the length of the sequences used. A linear regression analysis showed a statistically significant positive correlation between the metric and a logarithmic transformation of the sequence length, yielding a correlation value of R^2=0.83 ($p<10^{-16}$; see also **Figure 1B** for the complete formula). This was also true when analysing the sampling coverage levels individually (R^2>0.78 and

$p<0.01$ for all levels; see also **Supplementary Figure 1**). However, when considering specific genes, the analysis of the *env* gene (length=2508bp) was more accurate than that of *pol* (length=3000bp) when reconstructing the true tree in the 100% (point estimation=0.947 versus 0.936), 60% (mean or the replicates=0.946 [95%CI=0.945-0.945] versus 0.935 [0.934-0.935]; Student's t-test $p<10^{-16}$) and 20% (mean of the replicates=0.935 [95%CI=0.934-0.936] versus 0.933 [0.931-0.934]; p=0.01) sampling levels, but it showed more variability and worse results than the *pol* analyses in the replicates with 5% sampling level: mean=0.915 (95%CI=0.912-0.918) in *env* versus mean=0.936 (95%CI=0.933-0.938) in *pol* ($p<10^{-16}$). In general, *env* was the gene that showed the largest difference in the mean estimations across the different sampling coverage levels.

In the subsampled datasets, the 60% sampling coverage dataset performed very similarly to the fully sampled dataset, even showing means significantly higher than the 100% sampling coverage estimates when analysing the *gag*-*pol*-*env* (0.971 [95%CI=0.970-0.971] versus 0.967; $p<10^{-16}$), *gag* (0.880 [0.879-0.881] versus 0.879; $p=6.5 \times 10^{-3}$) and partial *pol* datasets (0.870 [0.869-0.871] versus 0.868; $p=1.6 \times 10^{-4}$).

In the 20% sampling level there was considerable overlap in performance among the larger fragments, but that of the smaller regions was substantially poorer. With 5% sampling coverage levels, the results showed the largest confidence intervals, revealing a substantial variability among the replicates, although some of these replicates outperformed estimations from the levels with higher sampling coverage.

Although quantitatively small, these differences in accuracy of tree reconstruction are important for identifying transmission clusters. We tested the impact of these differences using a standard methodology to detect transmission networks from the trees generated in this study by comparing the proportion of clusters found in the true tree ("true clusters") that were also found when analysing the ML trees. We did this using the *gag-pol-env* sequence and the partial *pol* sequences (as is the norm in the vast majority of studies) in the 100% sampled dataset, and we discovered that the use of *gag-pol-env* detected a significantly higher proportion of true clusters (778 out of 788 true clusters in *gag-pol-env* (98.73%) versus 774 out of 827 true clusters in partial *pol* (93.59%), chi-square test p =1.95×10^{-7}). Thus, even in the fully sampled dataset, the reconstruction of trees from partial sequences implies a significant and important difference in the outcome.

Discussion

We have used simulated HIV sequence data to show how the use of genes of different lengths can affect the correct reconstruction of the true viral phylogeny. The proportion of correct trees increased in almost direct proportion to the length of the sequences used. Thus, the 7kb *gag-pol-env* nearly full-genome sequences were best at reconstructing the true tree.

The 60% sampling coverage provides the most similar results to the analyses of the complete datasets, which emphasises the superior reliability of studies based on high densely sampled epidemics. In contrast, lower sampling depths (20% and 5%, which resemble the sampling settings found in Africa and developing areas) greatly reduced the accuracy of tree reconstruction –visible in the high variability between the replicates– especially when using the short clinical *pol* dataset.

We presumably obtained values higher than expected in a real-world analysis, particularly because there is a complete fit between the evolutionary model used to simulate the sequence data and the model used for analysing it. In addition, the good performance of the *env* analyses is partly due to the fact that its characteristic insertion/deletion variation was not simulated. Nevertheless the fact that *env* trees can outperform the *pol* trees, suggests that, in principle, the higher evolutionary rate in *env* can improve reconstruction.

Here we used a metric that is proportional to the RF metric –the most widely used method to estimate the distance/similarity between two phylogenetic trees. While this might be a simplistic metric, it is an intuitive and powerful method to compare trees, although its limitation is that it does not provide a means to state that one tree is significantly more similar to the true tree than a second tree is.

Our results demonstrate that the length of the sequence increases the reliability of phylogeny reconstruction in simulated data. In the simulations, different evolutionary rates applied to the *gag-pol* and *env* genes, as seen in real datasets. These were of 1.92 × 10^{-3} for *gag-pol* (or *pol*) and 2.605 × 10^{-3} for *env*, i.e. the evolutionary rate for *env* was 1.4 × that of *gag-pol*. Thus, the amount of variation that we find in *env* (length=2508nt) would be equivalent to an approximately 3401nt-long *gag-pol* sequence. This could explain that, in some replicates, *env* outperforms *pol* (length=3000nt). However, there was no insertion/deletion variation in the simulated sequences and in analysing real datasets this apparent superiority of *env* over more conserved genes is constrained by errors in alignment if hypervariable regions are included.

Although we did not perform a bootstrapping analysis of the reconstructed trees, previous analyses have further demonstrated that support for groupings in the tree is increased when longer sequences are used, and clustering found in full-length datasets can be missed when using sub-genomic regions[14-16]. Given the difficulty in generating and/or handling full genome datasets, our results demonstrate that *gag-pol* provides a dependable approximation; however it should be kept in mind that, at this point and considering we analysed a simulated dataset, the good performance of *gag-pol* could be more attributable to these genes' combined length than to their particular characteristics.

In conclusion, thanks to the more affordable generation of full HIV genomes, as is the goal of the PANGEA_HIV consortium[17], the use of longer genetic regions (such as concatenated *gag*, *pol* and *env* or *gag-pol*) will allow for a more reliable reconstruction of transmission events. The traditional short *pol* sequences generated for resistance testing that are used in most molecular epidemiology studies are substantially less reliable, especially with low sampling depths. An effort to generate highly sampled datasets is also needed to increase our ability to reconstruct real HIV epidemics.

Methods

HIV epidemic simulation

The PANGEA_HIV phylodynamic Methods Comparison Exercise[12] (http://www.pangea-hiv.org/Projects#phylodynamic) created a simulation resembling an African Village, which was based on high- and low-risk households and a small sex worker group. These simulations made use of the Discrete Spatial Phylo Simulator adapted to HIV-specific components (DSPS-HIV), which is an individual-based stochastic simulator. Using a specifiable contact network, the DSPS-HIV models HIV transmissions and records all sexual contacts. Selecting those which gave rise to transmissions produced the transmission tree. To generate the HIV sequences associated to these transmissions events, viral phylogenies that reflect between- and within-host viral evolution were simulated down the transmission tree using VirusTreeSimulator (https://github.com/PangeaHIV/VirusTreeSimulator).

In order to reconstruct ancestral subtype C sequences to be used as starting point of the simulation, a dataset of Southern African full genome subtype C sequences was downloaded from Los Alamos database (http://www.hiv.lanl.gov/). It included 100 sequences selected to represent a balanced number of sequences per calendar year (1989-2011), and were sampled in South Africa (n=46), Botswana (n=41), Zambia (n=8) and Malawi (n=5). The GenBank accession numbers corresponding for these 100 sequences are provided in the **Supplementary Table 1**. This dataset was separated into *gag*, *pol* and *env* and ancestral sequences for each gene were reconstructed using BEAST v1.8.1[18] applying GTR+4Γ+I as nucleotide substitution model and Bayesian skyride as demographic model.

These ancestral sequences were used as starting point to simulate sequences along these viral phylogenies using πBUSS[19], with substitution rates parameterized from the aforementioned analyses of Southern African sequences. To increase realism, different substitution rates applied to different genes (with a rate twice as high for *env* as for *gag* and *pol*) and different codon positions (1st and 2nd vs 3rd). Finally, the simulations were parameterized to emulate prevalence and incidence estimates from the peak of the African HIV epidemic in the late 1980s-early 1990s[20-22], before treatment roll-out, so the date of the root of the sequences coincides with the subtype C common ancestor in the 1940s[23].

More specific information about the sequence simulation is provided in the following PANGEA_HIV document: https://www.dropbox.com/sh/zlv40u4vnmpvy71/AAC8-yTPJA74OcYzvTCTb-H2a/201502/Village_unblinded/DSPS-Feb15Release-Details.pdf?dl=0.

Analysis dataset

We sampled all HIV simulated sequences corresponding to all infected individuals (one sequence per individual) in a 5-year period –between years 40 and 45 after the simulated epidemic started. From these simulated HIV sequences we created different combinations of sequence sampling depths and genomic regions. The full dataset contained 4662 sequences, and we adopted sub-sampling levels of 60% 20% and 5% sampling density which therefore included, respectively, 2798, 933 and 233 sequences. These sequences were chosen at random from the dataset with 100% sampling coverage. For the 60%, 20% and 5% sampling coverage levels we generated 100 independent sub-samples to test the reproducibility of the analyses.

We split each of these sequence datasets into: 1) "genome" (which represented the concatenation of *gag*, *pol* and *env* (6987bp)), 2) *gag-pol* (4479bp), 3) *gag* (1479bp), 4) complete *pol* (3000bp), 5) *env* (2508bp), and 6) partial *pol* (1302bp, the region commonly generated for PR+RT resistance testing).

The fully-sampled simulated sequence dataset as well as the true transmission tree are available at http://hiv.bio.ed.ac.uk/datasets/Yebra2016_Tree_Comparison_dataset.zip.

Phylogenetic tree comparison

We obtained the top-scoring maximum likelihood (ML) tree for each of these datasets using RAxML v8.2[24] under the GTR+Γ substitution model. For the nearly full genome trees, we applied a partition analysis in RAxML to accommodate for different evolutionary models in *gag-pol* versus *env*.

The Robinson-Foulds (RF)[25] metric is the most widely used measure of phylogenetic tree similarity. Given two phylogenetic trees, this metric counts the number of splits or clades induced by one of the trees but not the other. Here, we use an approximation to the RF metric implemented in the CompareTree program (http://meta.microbesonline.org/fasttree/treecmp.html), which also calculates the fraction of splits in the query tree (i.e., the reconstructed trees) that are shared with the reference one (i.e., the true trees). Unlike the RF metric, this value represents a proportion (therefore it ranges from 0 to 1), providing a metric that is more intuitive and easier to interpret and compare. We use the proportion of shared splits as an indicator of the fidelity in reconstructing the corresponding, sub-sampled true tree.

Finally, in order to evaluate the implications of the topology differences, a phylogenetic cluster comparison analysis was performed in the fully sampled dataset using the Cluster Picker and Cluster Matcher programs[26].

Statistical analyses

We compared the results from different genes and/or sampling coverage levels by using a two-sample Student's t-test. When comparing to the fully sampled datasets (100% sampling coverage), for which only point estimations were obtained because replicates cannot be produced, a one-sample t-test was performed to test whether the corresponding mean distribution was significantly different than the point estimation of the 100% sampling coverage level. Finally, we applied a linear regression analysis to explore the relationship between the results and the sequence length. All this calculations were produced in R[27] version 3.1.2.

References

1 Dolling, D. *et al.* Time trends in drug resistant HIV-1 infections in the United Kingdom up to 2009: multicentre observational study. *Brit. Med. J.* **345**, e5253 (2012).

2 Wheeler, W. H. *et al.* Prevalence of transmitted drug resistance associated mutations and HIV-1 subtypes in new HIV-1 diagnoses, US-2006. *AIDS* **24**, 1203-1212 (2010).

3 Frentz, D. *et al.* Increase in transmitted resistance to non-nucleoside reverse transcriptase inhibitors among newly diagnosed HIV-1 infections in Europe. *BMC Infect. Dis.* **14** (2014).

4 DeBry, R. W. *et al.* Dental HIV transmission? *Nature.* **361**, 691 (1993).

5 Hué, S., Clewley, J. P., Cane, P. A. & Pillay, D. HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS* **18**, 719-728 (2004).

6 Ragonnet-Cronin, M. *et al.* Transmission of non-B HIV subtypes in the United Kingdom is increasingly driven by large non-heterosexual transmission clusters. *J. Infect. Dis.* **213**, 1410-1418 (2016).

7 Shilaih, M. *et al.* Genotypic resistance tests sequences reveal the role of marginalized populations in HIV-1 transmission in Switzerland. *Sci. Rep.* **6**, 27580 (2016).

8 Leitner, T., Escanilla, D., Franzen, C., Uhlen, M. & Albert, J. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 10864-10869 (1996).

9 Mikhail, M. *et al.* Full-length HIV type 1 genome analysis showing evidence for HIV type 1 transmission from a nonprogressor to two recipients who progressed to AIDS. *AIDS Res. Hum. Retroviruses* **21**, 575-579 (2005).

10 Paraskevis, D. *et al.* Phylogenetic reconstruction of a known HIV-1 CRF04_cpx transmission network using maximum likelihood and Bayesian methods. *J. Mol. Evol.* **59**, 709-717 (2004).

11 Rachinger, A., Groeneveld, P. H., van Assen, S., Lemey, P. & Schuitemaker, H. Time-measured phylogenies of gag, pol and env sequence data reveal the direction and time interval of HIV-1 transmission. *AIDS* **25**, 1035-1039 (2011).

12 Ratmann, O. *et al.* Phylogenetic Tools for Generalized HIV-1 Epidemics: Findings from the PANGEA-HIV Methods Comparison. *Mol. Biol. Evol.* (2016).

13 Leigh Brown, A. J. *et al.* Transmission network parameters estimated from HIV sequences for a nationwide epidemic. *J. Infect. Dis.* **204**, 1463-1469 (2011).

14 Lemey, P. & Vandamme, A. M. Exploring full-genome sequences for phylogenetic support of HIV-1 transmission events. *AIDS* **19**, 1551-1552 (2005).

15 Novitsky, V., Moyo, S., Lei, Q., DeGruttola, V. & Essex, M. Importance of Viral Sequence Length and Number of Variable and Informative Sites in Analysis of HIV Clustering. *AIDS Res. Hum. Retroviruses* **31**, 531-542 (2015).

16 Amogne, W. *et al.* Phylogenetic analysis of Ethiopian HIV-1 subtype C near full-length genomes reveals high intrasubtype diversity and a strong geographical cluster. *AIDS Res. Hum. Retroviruses* **32**, 471-474 (2016).

17 Pillay, D. *et al.* PANGEA-HIV: phylogenetics for generalised epidemics in Africa. *Lancet Infect. Dis.* **15**, 259-261 (2015).

18 Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969-1973 (2012).

19 Bielejec, F. *et al.* piBUSS: a parallel BEAST/BEAGLE utility for sequence simulation under complex evolutionary scenarios. *BMC bioinformatics* **15** (2014).

20 Serwadda, D. *et al.* HIV risk-factors in three geographic strata of rural Rakai District, Uganda. *AIDS* **6**, 983-989 (1992).

21 Wawer, M. J. *et al.* Incidence of HIV-1 infection in a rural region of Uganda. *Brit. Med. J.* **308**, 171-173 (1994).

22 Muller, O. *et al.* HIV prevalence, attitudes and behavior in clients of a confidential HIV testing and counseling-center in Uganda. *AIDS* **6**, 869-874 (1992).

23 Faria, N. R. *et al.* HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. *Science* **346**, 56-61 (2014).

24 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).

25 Robinson, D. F. & Foulds, L. R. Comparison of Phylogenetic Trees. *Math Biosci* **53**, 131-147 (1981).

26 Ragonnet-Cronin, M. *et al.* Automated analysis of phylogenetic clusters. *BMC bioinformatics* **14**, 317 (2013).

27 R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria, 2010). Retrieved from: https://www.r-project.org.

PANGEA_HIV Consortium members

Christophe Fraser[3], Paul Kellam[4], Tulio de Oliveira[2], Ann Dennis[5], Anne Hoppe[6], Cissy Kityo[7], Dan Frampton[6], Deogratius Ssemwanga[8], Frank Tanser[2], Jagoda Keshani[6], Jairam Lingappa[9], Joshua Herbeck[9], Maria Wawer[10], Max Essex[11], Myron S. Cohen[5], Nicholas Paton[12], Oliver Ratmann[3], Pontiano Kaleebu[8], Richard Hayes[13], Sarah Fidler[14], Thomas Quinn[10] & Vladimir Novitsky[11]

[3] Department of Infectious Disease Epidemiology, Imperial College London, London, UK

[4] Wellcome Trust Sanger Institute, Hinxton, UK

[5] University of North Carolina at Chapel Hill, University of North Carolina, Chapel Hill, USA

[6] Farr Institute of Health Informatics Research, University College London, London, UK

[7] Joint Clinical Research Centre, Kampala, Uganda.

[8] MRC/UVRI, Uganda Research Unit on AIDS, Entebbe, Uganda

[9] Department of Global Health, University of Washington, Seattle, WA, USA

[10] Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

[11] Harvard T.H. Chan School of Public Health, Boston, MA, USA

[12] MRC Clinical Trials Unit, University College London Hospital, London, UK

[13] Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK

[14] Department of Medicine, Imperial College London, London, UK

ICONIC Project members

Andrew Haywards[6], Eleni Nastouli[15], Steven Morris[16], Duncan Clark[17] & Zisis Kozlakidis[18]

[15] Department of Virology, University College London Hospital, London, UK

[16] Department of Health Economics, University College London, London, UK

[17] Department of Virology, Barts Health NHS Trust, London, UK

[18] Division of Infection and Immunity, University College London, London, UK

Acknowledgements

We would like to thank the four anonymous refereees for providing very constructive comments that improved the original manuscript. This work was supported by the PANGEA_HIV Consortium (with support provided by the Bill & Melinda Gates Foundation), by the ICONIC project and by NIH GM110749. This publication presents independent research supported by the Health Innovation Challenge Fund T5-344 (ICONIC), a parallel funding partnership between the Department of Health and Wellcome Trust. The views expressed in this publication are those of the authors and not necessarily those of the Department of Health or Wellcome Trust.

Author Contributions Statement

A.J.L.B and D.P conceived the study. G.Y and M.R.C performed the analyses. E.B.H designed and generated the HIV simulation. G.Y wrote the first draft. All authors reviewed, contributed to, and approved the final version of the manuscript. The PANGEA_HIV Consortium and the ICONIC project provided funding and resources and their members approved the final version of the manuscript.

Competing financial interests

The authors declare no competing financial interests.

Figure Legends

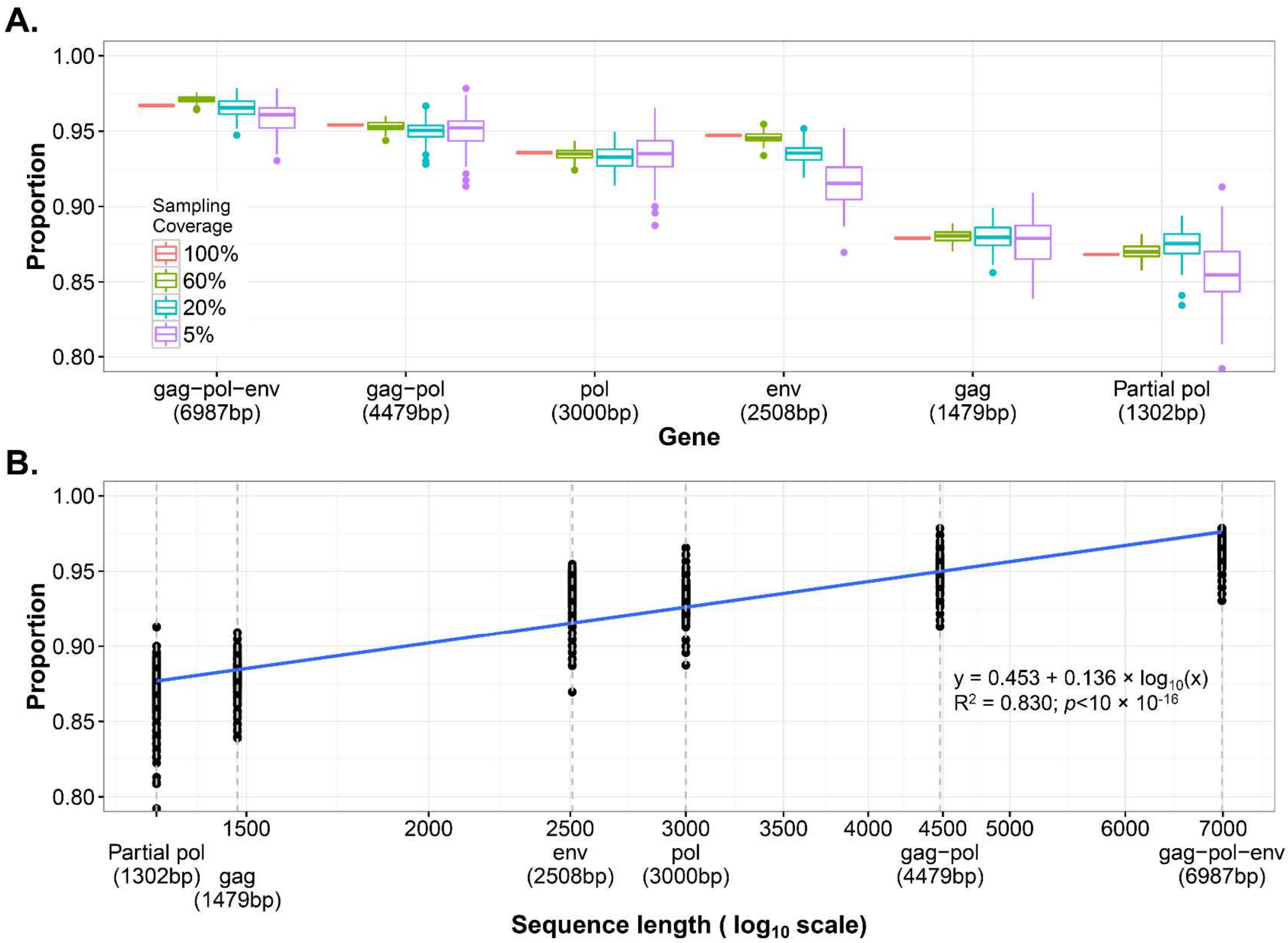
Figure 1:

A) Proportion of the maximum likelihood trees splits shared with the true tree for each gene and sampling coverage level. Genes are sorted according to length. The top and bottom limits of the boxes represent, respectively, the first and third quartiles (the distance between them represents the inter-quartile range, IQR). The lines (whiskers) include the highest and lowest values that lie within the 1.5 × IQR distance from the first and third quartiles, respectively. Data points outside this range are outliers. **B) Proportion of the maximum likelihood trees splits shared with the true tree according to gene length.** All sampling coverage levels were considered together. The regression line is shown in blue, for which the formula, the correlation coefficient (R^2) and the p-value are presented. The shaded area shows the regression line's confidence intervals. The grey, dotted vertical lines show the length of each gene considered.

Table 1. Proportion of the maximum likelihood trees splits shared with the true tree according to gene and sampling coverage level.

Gene	**Length (bp)**	**Sampling coverage level (mean [95% confidence interval])**				
		All	**100%**	**60%**	**20%**	**5%**
gag-pol-env	6987	0.965 (0.964-0.966)	0.967	0.971 (0.970-0.971)	0.965 (0.964-0.966)	0.959 (0.957-0.961)
gag-pol	4479	0.951 (0.950-0.952)	0.954	0.953 (0.953-0.954)	0.950 (0.948-0.951)	0.950 (0.948-0.953)
pol	3000	0.934 (0.933-0.935)	0.936	0.935 (0.934-0.935)	0.933 (0.931-0.934)	0.936 (0.933-0.938)
env	2508	0.932 (0.930-0.934)	0.947	0.946 (0.945-0.946)	0.935 (0.934-0.936)	0.915 (0.912-0.918)
gag	1479	0.879 (0.877-0.880)	0.879	0.880 (0.879-0.881)	0.880 (0.878-0.881)	0.877 (0.873-0.880)
Partial *pol*	1302	0.867 (0.866-0.869)	0.868	0.870 (0.869-0.871)	0.875 (0.873-0.877)	0.857 (0.853-0.861)

The table shows the mean value and its 95% confidence intervals for the 100 replicates performed in each case. Note that for the full dataset (100% sampling coverage) only one estimation is shown because no replicates can be performed. The genes are ordered in descending order of sequence length.


A.
1.00
0.95
0.90
0.85
0.80
Proportion
Sampling Coverage
100%
60%
20%
5%
gag−pol−env (6987bp)
gag−pol (4479bp)
pol (3000bp)
env (2508bp)
gag (1479bp)
Partial pol (1302bp)
Gene
B.
1.00
0.95
0.90
0.85
0.80
Proportion
$y = 0.453 + 0.136 \times \log_{10}(x)$
$R^2 = 0.830$; $p<10 \times 10^{-16}$
1500
2000
2500
3000
3500
4000
4500
5000
6000
7000
Partial pol (1302bp)
gag (1479bp)
env (2508bp)
pol (3000bp)
gag−pol (4479bp)
gag−pol−env (6987bp)
Sequence length ($\log_{10}$ scale)