Evolutionary Bioinformatics

SHORT REPORT

PhyloTempo: A Set of R Scripts for Assessing and Visualizing Temporal Clustering in Genealogies Inferred from Serially Sampled Viral Sequences

Melissa M. Norström[1,*], Mattia C.F. Prosperi[2,*], Rebecca R. Gray[3], Annika C. Karlsson[1] and Marco Salemi[2]

[1]Department of Laboratory Medicine, Division of Clinical Microbiology, Karolinska University Hospital, Karolinska Institutet, Stockholm, Sweden. [2]University of Florida, College of Medicine, Department of Pathology, Immunology and Laboratory Medicine and Emerging Pathogens Institute, Gainesville, Florida, USA. [3]Department of Zoology, University of Oxford, Oxford, UK. *These authors contributed equally. Corresponding author email: salemi@pathology.ufl.edu

Abstract: Serially-sampled nucleotide sequences can be used to infer demographic history of evolving viral populations. The shape of a phylogenetic tree often reflects the interplay between evolutionary and ecological processes. Several approaches exist to analyze the topology and traits of a phylogenetic tree, by means of tree balance, branching patterns and comparative properties. The *temporal clustering* (TC) statistic is a new topological measure, based on ancestral character reconstruction, which characterizes the temporal structure of a phylogeny. Here, PhyloTempo is the first implementation of the TC in the R language, integrating several other topological measures in a user-friendly graphical framework. The comparison of the TC statistic with other measures provides multifaceted insights on the dynamic processes shaping the evolution of pathogenic viruses. The features and applicability of PhyloTempo were tested on serially-sampled intra-host human and simian immunodeficiency virus population data sets. PhyloTempo is distributed under the GNU general public license at https://sourceforge.net/projects/phylotempo/.

Keywords: fast evolving viruses, longitudinal samples, phylogenetics, phylodynamics, comparative methods, clustering, software, positive selection, coalescence

Evolutionary Bioinformatics 2012:8 261–269

doi: 10.4137/EBO.S9738

This article is available from http://www.la-press.com.

Introduction

The evolutionary and demographic history of a measurable evolving viral population[1] can be inferred by phylodynamic analysis of longitudinally sampled sequences.[2] In particular, the shape of a phylogenetic tree often reflects the characteristics and the interactions among evolutionary and ecological processes. For example, phylogenies of viruses under continuous positive selection, such as inter-host influenza or intra-host human immunodeficiency virus (HIV), usually exhibit a marked staircase-like topology.[3,4] Pathogens under weak or absent positive selection show a more balanced tree shape, as it happens within the measles serotypes.[5] Trees displaying a star-like topology or increasing root to tip distance with sampling time are typical of an exponentially growing population, whilst the opposite pattern can usually be associated to constant or decreasing population sizes, as for example in Dengue virus inter-host phylogenies.[6] Phylogenetic tree shapes can also be coupled with phenotypic traits (either numeric or categorical) or geographic correlates (such as geographic origin of sampled strains), which can be analyzed *via* comparative methods in terms of evolutionary or spatiotemporal dynamics.[7]

Several statistical approaches exist for determining if and how serially sampled sequences evolve under a strict or relaxed molecular clock.[8,9] However, such methods do not give indications about the phylogenetic tree shape (eg, staircase- or star-like) and the related temporal structure (eg, if sequences sampled at the same time point tend to cluster together and to be direct ancestors of sequences sampled at later time points).

There are various functions for describing the topological features of a phylogenetic tree. Some of these measures consider both topology and branch lengths of a tree, as well as phenotypic tip traits,[10–12] while others evaluate only the tree topology in relation to geographic or phenotypic characters associated with the sampled strains.[13–15] Finally, there are purely topological measures based on tree symmetry/balance.[16–19]

The *temporal clustering* (TC) is a recently developed statistic, which takes into account phylogenetic tree topology and sampling time of the tips.[20] The TC statistic assesses the temporal structure of a phylogenetic tree, by evaluating the order of time changes from internal nodes to tips. It is based on the maximum parsimony reconstruction of ancestral characters implemented in phylogeography,[13] but it has been modified to prevent the estimation of temporally impossible state changes in tip dated trees (ie, an earlier time point emerging from an ancestor assigned to a later time point). It also allows the comparison of phylogenies inferred from data sets including different number of times points and/or sampled sequences per time point. Currently, there is no available software implementation of the TC statistic, although it can be calculated with a series of manual steps using the MacClade program (http://macclade.org/).

In this work, we present PhyloTempo, the first software implementation of the TC statistic. PhyloTempo is written in R, a free software environment for statistical computing and graphics (http://www.r-project.org/). Along with the TC implementation, several other tree topological measures were integrated in PhyloTempo using pre-existing R libraries, in a user-friendly graphical framework. The program was tested on several longitudinally sampled intra-host HIV and simian immunodeficiency virus (SIV) population data sets. The results showed how the comparison of the TC statistic with other topological measures can provide multifaceted insights on the dynamic processes shaping the evolution of pathogenic viruses.

Methods

The original formulation of the TC statistic[20] requires a phylogenetic tree with n taxa sampled at t different discrete time points. A state (ie, time) transition matrix is then defined, where the cost of going from later to earlier time points is infinite (ie, time-irreversibility). The other costs are usually defined as integer linearly increasing with the time points ordering. Ancestral tree states are inferred using Fitch's parsimony algorithm.[21] A non-normalized TC score is then calculated by summing all the state changes across the tree branches according to the cost matrix weights. A tree with a perfect temporal structure, ie, a tree in which all tips sampled at time point t_i are monophyletic and directly emerge from time point t_{i-1}, would have a non-normalized TC equal to $t-1$. Conversely, a tree with the least temporal structure would have a maximal non-normalized TC, equal to the sum of the n number of taxa multiplied by the corresponding w weights of the cost matrix, ie, $\text{Max} = \sum_{j=1}^{t} n_j w_{1j}$. The normalized TC rescales the

non-normalized TC value in the interval [0,1], by considering a background distribution of TC statistics obtained by shuffling the time points associated to the tree tips (keeping the topology fixed) and re-estimating the ancestral characters. Specifically, the normalized TC statistic is

$$\mathrm{TC} = \max\left\{0, \frac{\ln(S_{avg}) - \ln(S_{obs})}{\ln(S_{\max}) - \ln(\mathrm{Min})}\right\}$$

where the S_{avg} and $S_{\max}$ are -respectively- the average and the maximum non-normalized TC values observed in the randomized trees, while Min is the minimum theoretical non-normalized TC, equal to $t-1$. S_{obs} is the observed non-normalized TC value calculated on the original tree. The numerator represents the deviation from the null hypothesis (ie, no temporal clustering), while the denominator represents the range of possible values for the given number of taxa and time points.

Coupled with the TC statistic, PhyloTempo includes also the following tree topology measures and tests of hypothesis: Aldous' graphical test and lilkelihood ratio test to decide if tree fit the Yule or the uniform models;[22] Colless' and Sackin's shape statistics, both under the Yule or uniform hypotheses;[23] cherry count;[24] Pybus' gamma.[25] In addition, a simple tree statistic called "staircase-ness" is introduced, counting the proportion of sub-trees that are imbalanced (ie, sub-trees where the left child contains more leaves than the right child, or vice-versa) compared against the distribution of such proportions obtained from random trees. See the supplementary material for the properties of this measure.

Implementation

All the code has been written in the R language. Besides the standard core library set of R, the following R libraries have been used (including their dependencies): "ape", "ade4", "phybase", "phylobase", "phangorn", "doBy", "infotheo", "apTreeshape", "diversitree" (http://www.r-phylo.org).

The required input of PhyloTempo is a phylogenetic tree file in "newick" format and a two-column text file in which each tip name present in the phylogenetic tree is associated with its corresponding time of sampling (a numeric value such as days or years).

The input phylogenetic tree is preliminarily checked for polytomies, which are resolved randomly. If present, negative branch lengths are set to zero and then all branch lengths are added a 10^{-5} value. The tree is rooted on the tip that gives the highest linear correlation between the root-to-tip distance and the sampling time of the tip, and finally it is ladderized. The vector of sampling times is then discretized into time intervals by using an equal-frequency binning, where the optimal number of bins is the square root of the vector size. The maximum allowed number of discrete time intervals is nine, and each time bin needs to contain at least two tips.

The TC statistic calculation is made upon the previous theoretical description. However, in this new implementation the ancestral characters are estimated using maximum likelihood[26] rather than parsimony. A major advantage of maximum likelihood is that it also allows for an optimized estimate of the weights of the transition cost matrix.

The number of tip randomizations is set to 300 by default, but the value can be modified by the user. All the other tree statistics are assembled by combining existing R functions.

Both graphical and text output are produced, where figures are plotted in multiple windows, text is printed in the R command-line window and results are saved in a tabulated file. The graphical plots include: the phylogenetic tree with ancestral character state probabilities drawn with pie charts at internal nodes; the TC statistic compared versus the randomized background distribution; a linear correlation plot between the sampling times of tips and root-to-tip distances; a Kruskal-Wallis test comparing distribution of root-to-tip distances with the discretized time points; the staircase-ness, Aldous', Sackin's and Cherry count statistics with the corresponding background randomizations. The text output reports the aforementioned results as well as the *P*-values from the statistical tests. Also, a script that allows the analysis of multiple trees in "nexus" format, from an *a posteriori*, eg, trees as output from MrBayes,[27] or bootstrap analysis has been made available.

PhyloTempo is distributed under the GNU general public license and is available at https://sourceforge.net/projects/phylotempo/ for download.

Results

PhyloTempo has been tested on different viral data sets. The first data set included intra-host HIV-1 phylogenetic trees, inferred from serially sampled *envelope* (*env*) C2-V5 sequences, from nine untreated subjects with fast disease progression,[4] named as the "Shankarappa" data set after the first author of the paper. The second data set included intra-host SIV trees, inferred from *env* gp120 sequences sampled longitudinally from four experimentally infected Rhesus macaques that were CD8-depleted before infection and progressed to AIDS within 75–118 days post infection.[28] The third data set included intra-host HIV-1 *gag* p24 trees from six untreated subjects enrolled in the OPTIONS cohort[29] all carrying the HLA-B*5701 allele strongly associated with slower disease progression, that were followed longitudinally from early infection up to seven years.[30]

In Table 1 the text output of PhyloTempo is reported, after running the program on each proof-of-concept data set (note that for simplicity not all indicators output by PhyloTempo are shown). When comparing the former calculation of the TC statistic based on parsimony with the new one based on maximum likelihood, in general we found a high degree of linear ρ correlation ($\rho \approx 0.8$, combining the three data sets, data not shown). On average, the TC exhibited a weak linear ρ correlation with any of the other tree topology measures implemented in PhyloTempo (average $\rho = 0.10$, standard deviation 0.17), including also dN/dS values (estimated via the Nei-Gojobori method averaging across all positions). The maximum value obtained was $\rho = 0.42$, found with respect to the root-to-tip-distance vs. sampling time correlation.

The average TC statistic for the Shankarappa data set was 0.29 (st.dev 0.10), for the SIV data set was 0.11 (st.dev 0.09). The OPTIONS data set, based on the highly conserved HIV-1 *gag* p24 gene, allowed us to evaluate the effect of including or excluding

Table 1. Summary of PhyloTempo output from different proof-of-concept data sets.

Data set	Time range (post-infection)	No. time intervals	No. tips	RTD vs. ST ρ	Staircase-ness	dN/dS	TC
OPTIONS P1 all seqs.	91–1872 days	4	84	0.81	0.75	0.26	0.41
OPTIONS P1 unique seqs.	91–1872 days	4	48	0.89	0.64	0.21	0.35
OPTIONS P2 all seqs.	126–1348 days	3	79	0.88	0.73	0.15	0.31
OPTIONS P2 unique seqs.	126–1348 days	3	65	0.80	0.75	0.17	0.29
OPTIONS P3 all seqs.	91–2234 days	7	186	0.93	0.82	0.17	0.22
OPTIONS P3 unique seqs.	91–2234 days	6	74	0.84	0.70	0.15	0.36
OPTIONS P4 all seqs.	77–2180 days	5	128	0.78	0.80	0.30	0.20
OPTIONS P4 unique seqs.	77–2180 days	5	54	0.66	0.72	0.27	0.33
OPTIONS P5 all seqs.	91–2129 days	5	124	0.94	0.83	0.31	0.37
OPTIONS P5 unique seqs.	91–2129 days	3	55	0.95	0.74	0.18	0.72
OPTIONS P6 all seqs.	70–2602 days	6	140	0.92	0.73	0.24	0.13
OPTIONS P6 unique seqs.	70–2602 days	5	85	0.80	0.65	0.12	0.13
Shankarappa #1	14–133 days	9	137	0.90	0.66	1.00	0.30
Shankarappa #2	14–161 days	9	231	0.93	0.68	1.24	0.20
Shankarappa #3	42–154 days	9	106	0.92	0.72	1.63	0.50
Shankarappa #5	14–567 days	9	236	–0.01	0.68	0.74	0.25
Shankarappa #6	77–154 days	8	130	0.93	0.64	0.89	0.37
Shankarappa #7	35–126 days	7	138	0.92	0.68	1.14	0.22
Shankarappa #8	63–168 days	8	150	0.92	0.69	1.23	0.24
Shankarappa #9	21–357 days	9	120	0.32	0.71	1.71	0.21
Shankarappa #11	35–154 days	6	52	0.94	0.69	0.90	0.32
SIV D03 plasma	22–75 days	3	58	0.50	0.67	0.27	0.03
SIV D04 plasma	22–91 days	3	66	0.52	0.62	0.44	0.22
SIV D05 plasma	22–89 days	3	67	0.35	0.70	0.42	0.05
SIV D06 plasma	22–118 days	3	68	0.52	0.66	0.40	0.13
Correlation with TC		0.05	–0.17	0.42	0.23	0.04	1.00

Abbreviations: Seqs, sequences; RTD, root-to-tip distance; ST, sampling time; ρ, Pearson's linear correlation; SC, staircase-ness; dN/dS, ratio between non-synonymous and synonymous substitutions; TC, temporal clustering statistic.

identical sequences and resulted in a TC value of 0.27 (st.dev 0.11) and 0.36 (st.dev 0.19) when analyzing all sequences or only unique sequences, respectively.

Figures 1 and 2 illustrate the PhyloTempo graphical output. In detail, Figure 1 shows two of the trees analyzed (OPTIONS and the SIV data sets, respectively), and includes the maximum likelihood estimate of the ancestral time states, with state probabilities reported as pie charts at each internal node. Figure 2 reports the placement of the TC statistic, as well as all the other tests, with respect to background random distributions or null hypotheses. In addition, the correlation plot between the root-to-tip distances and the sampling time is shown, along with the boxplots of the root-to-tip distances stratified by the time intervals.

On average, the running time of PhyloTempo on an input phylogenetic tree with 100–150 leaves (3–4 time points and 300 randomizations) takes less than 5 minutes using a standard desktop computer. Running times for trees of 300 or 400 tips increase to half or one hour.

Discussion

In this paper we presented PhyloTempo, a set of scripts in the R language that calculates the TC clustering statistics and other measures of phylogenetic tree shape, with a comprehensive text and graphical output. The choice of the R software environment gives to the tool the advantage to be available for many platforms (Microsoft Windows, Mac, or Linux) and, since R features a *plethora* of libraries both for phylogenetic analysis and graphics, to be ready for the inclusion of other functions related to the analysis of phylogenetic tree shape and comparative statistics.

Although other programs that calculate tree shape statistics are available, such as the java application TreeStat (http://tree.bio.ed.ac.uk/software/treestat/), and Path-O-Gen (http://tree.bio.ed.ac.uk/software/pathogen/), as well as several command-line functions in R, this is the first that implements the TC statistic merging in a user-friendly interface both graphical and text outputs. In addition, PhyloTempo is capable of generating an *a posteriori* TC statistic, reading a tree ensemble in "nexus" format, such as the output by MrBayes (http://mrbayes.sourceforge.net/). As a future perspective in the context of Bayesian analysis, a theoretical approach to derive an analytical formula for the TC statistic is advisable, allowing the avoidance of the time-consuming tree randomization for each tree.

Two interesting biological insights are evident from the present analysis. First, TC does not correlate with any previously described topological tree

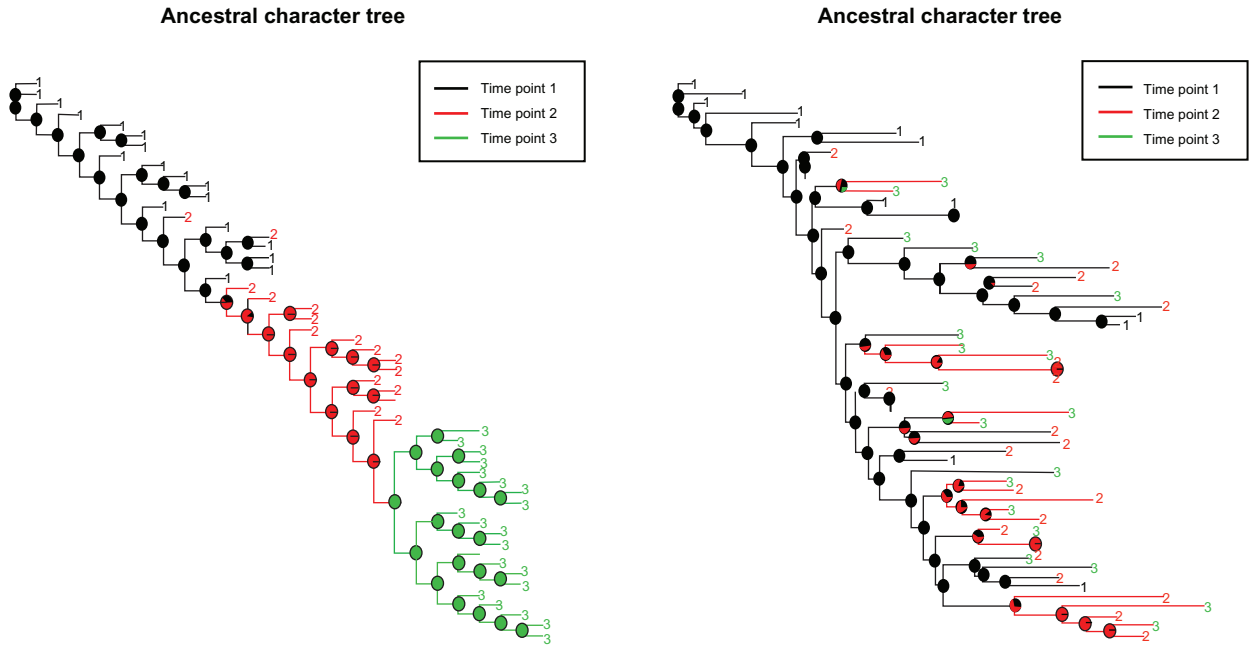

Figure 1. PhyloTempo graphical output showing the ancestral character estimation on an input phylogenetic tree.
Notes: Pie charts in the internal nodes of the tree represent probabilities of ancestral states. Left panel shows a tree from the OPTIONS data set (patient P5, unique sequences) with a high TC statistic (0.7); right panel shows a tree from the SIV data set (subject D03) with a poor TC statistic (0.1).

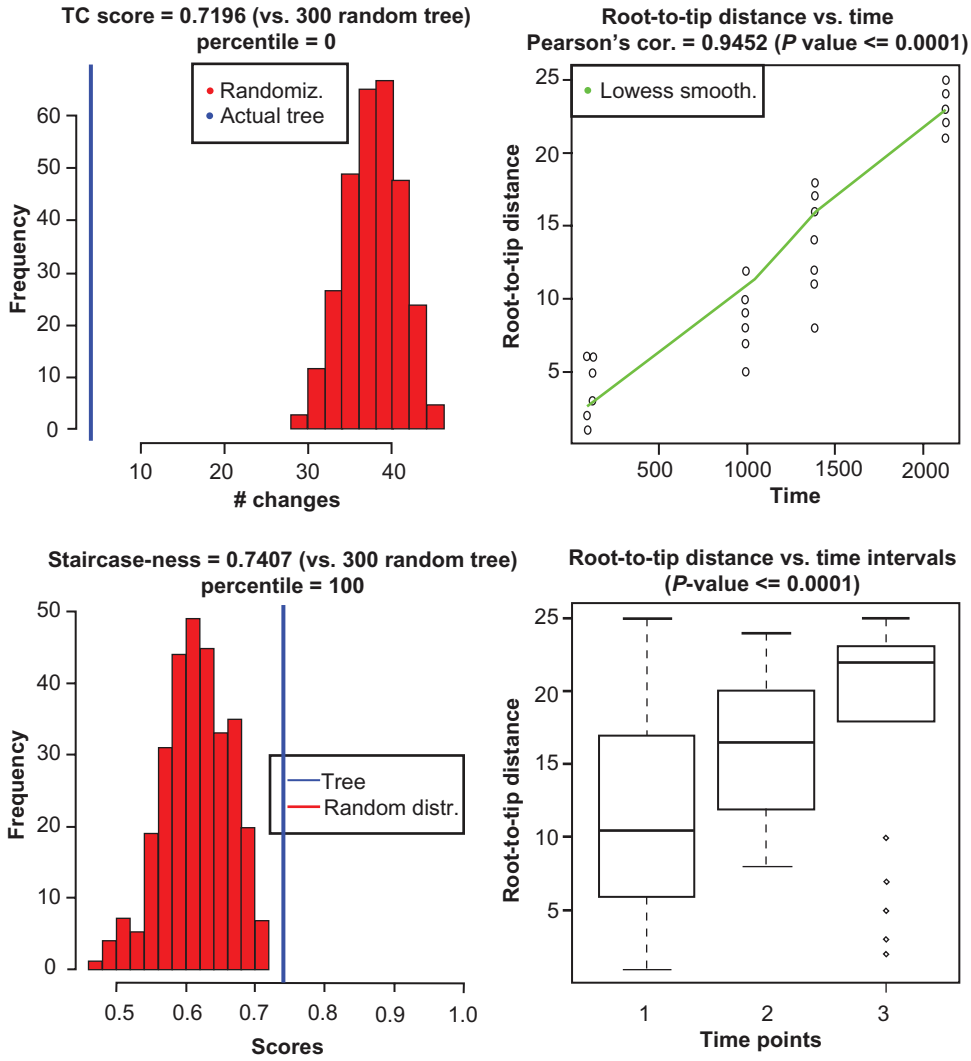

Figure 2. PhyloTempo graphical output summarizing phylogenetic tree shape statistics.
Note: A tree from the OPTIONS data set (patient p5, unique sequences) with a high TC statistic (0.7) was used.

measure implying that the new statistic evaluates aspects of the evolutionary process not captured by other methods. Second, TC does not correlate with estimated dN/dS ratios in different data sets. Several studies have interpreted temporally structured phylogenies as evidence of sequential viral population bottlenecks driven by continuous selection pressure.[2,4,31,32] The trees inferred from the OPTIONS data sets include HIV-1 sequences from patients with the HLA-B*5701 allele that has been associated with slower disease progression, possibly due to strong positive or purifying selection driving viral escape from cytotoxic T lymphocyte recognition.[33]

Interestingly, the TC calculated for the OPTIONS data sets are not significantly different ($P = 0.21$ from a *t*-test) from those calculated for the Shankarappa data sets. The finding suggests that temporally structured genealogies may reflect intra-host evolutionary processes that are similar in two groups of patients characterized by different rates of disease progression and that may not be related to selection pressure. However, it is important to point out that the subjects in the Shankarappa data set were followed for a shorter period of time than the OPTIONS subjects, and that the intervals between longitudinal samples were overall shorter in the first data set (data not shown). The low TC values for the Shankarappa data set may simply reflect an incomplete turnover of the viral quasispecies, which can require up to 22 months,[34] causing an intermix of sequences sampled at different time points. Moreover, archival viral strains expressed in cellular reservoirs would decrease the temporal

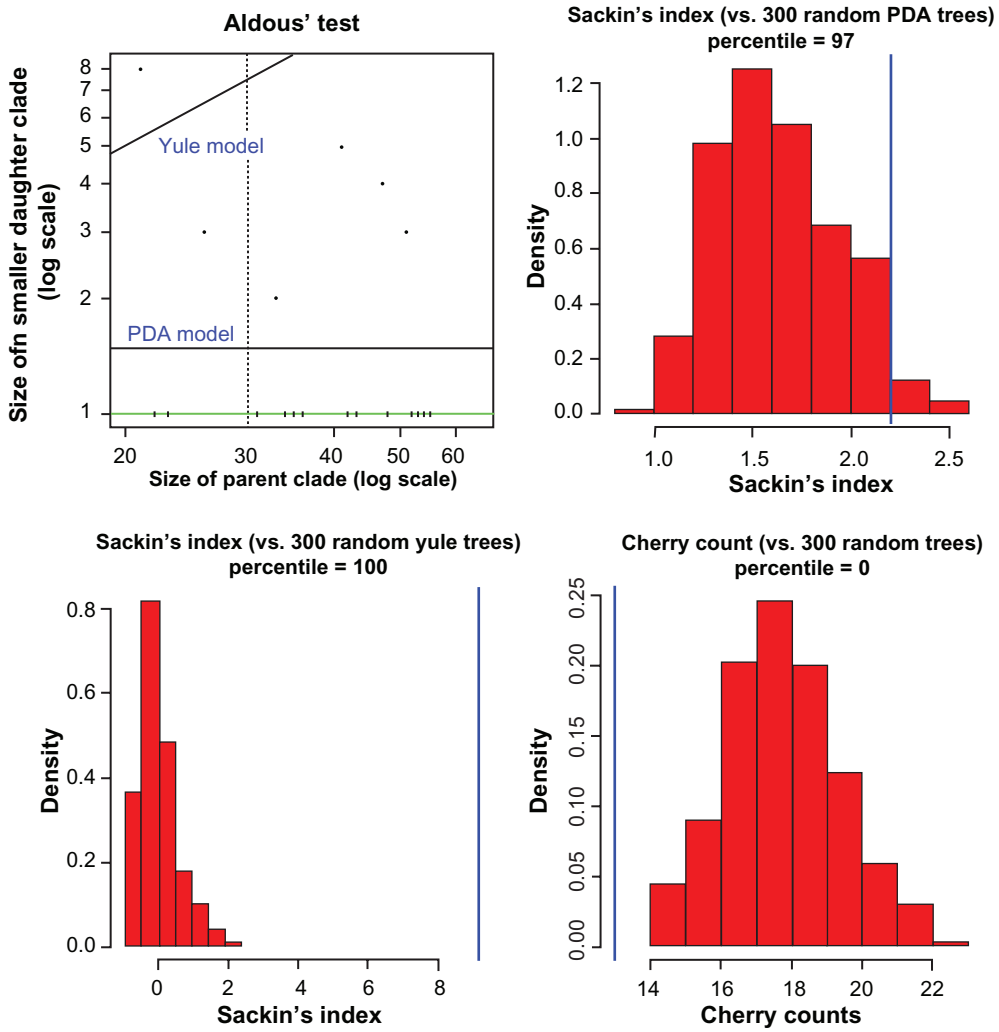

Figure 2. (*Continued*)

structure of serially sampled genealogies, because sequences from later time points may share a most recent common ancestor with sequences collected much earlier in infection.[35] Therefore, the TC statistic could be a powerful tool to investigate the extent and impact of latent viral reservoirs in intra-host HIV-1 evolution. Finally, it is interesting to note that the SIV data sets show the lowest TC values. This may be the result of the relatively short time of infection in these animals, as well as a consequence of the depletion of $CD8^+$ T cells right after infection.[28]

In conclusion, the present work describes a practical and user-friendly implementation of a novel statistic to evaluate the shape of phylogenetic trees, inferred from longitudinal samples of measurably evolving viral populations, which can provide significant insights on underlying evolutionary processes linked to infection dynamics and pathogenesis.

Author Contributions

Conceived and designed the experiments: MMN, AK. Analysed the data: MCFP. Wrote the first draft of the manuscript: MMN, MCFP. Contributed to the writing of the manuscript: MS. Agree with manuscript results and conclusions: MS, RRG, AK. Jointly developed the structure and arguments for the paper: MCFP, MMN, MS. Made critical revisions and approved final version: MCFP, MMN, RRG, AK, MS. All authors reviewed and approved of the final manuscript.

Funding

MCFP and MS were supported by the University of Florida award UL1 RR02989 and the 2012 EPIG grant, and by the NIH/NINDS R01 grant NS063897-01A2. MMN and ACK were supported by grants from the Swedish Research Council (K2010-56X-20345-04-3), Karolinska Institutet, Åke Wibergs Foundation

(40418186), and the Swedish Society of Medicine (SLS-101021).

Disclosures and Ethics

In regards to the OPTIONS cohort, the University of California, San Francisco (UCSF) Committee on Human Research and the Regional Ethical Council in Stockholm, Sweden (2008/1099-31), approved this study and all patients provided written informed consent.

References

1. Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. Measurably evolving populations. *Trends Ecol Evol*. Sep 2003;18(9):481–8.
2. Grenfell BT, Pybus OG, Gog JR, et al. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*. Jan 16, 2004;303(5656):327–32.
3. Lin JH, Chiu SC, Cheng JC, et al. Phylodynamics and molecular evolution of influenza A virus nucleoprotein genes in Taiwan between 1979 and 2009. *PLoS One*. 2011;6(8):e23454.
4. Shankarappa R, Margolick JB, Gange SJ, et al. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J Virol*. Dec 1999;73(12):10489–502.
5. Rima BK, Earle JA, Yeo RP, et al. Temporal and geographical distribution of measles virus genotypes. *J Gen Virol*. May 1995;76(Pt 5):1173–80.
6. Holmes EC, Twiddy SS. The origin, emergence and evolutionary genetics of dengue virus. *Infect Genet Evol*. May 2003;3(1):19–28.
7. Harvey PHaM, RM, editor. *The Comparative Method in Evolutionary Biology*. Oxford University Press; 1991. Oxford Monographs in Ecology and Evolution.
8. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol*. May 2005;22(5):1185–92.
9. Drummond A, Pybus OG, Rambaut A. Inference of viral evolutionary rates from molecular sequences. *Adv Parasitol*. 2003;54:331–58.
10. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol*. Dec 2005;71(12):8228–35.
11. Chang Q, Luan Y, Sun F. Variance adjusted weighted UniFrac: a powerful beta diversity measure for comparing communities based on phylogeny. *BMC Bioinformatics*. 2011;12:118.
12. Webb CO. Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *Am Nat*. Aug 2000;156(2):145–55.
13. Slatkin M, Maddison WP. A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics*. Nov 1989;123(3):603–13.
14. Parker J, Rambaut A, Pybus OG. Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infect Genet Evol*. May 2008;8(3):239–46.
15. Wang TH, Donaldson YK, Brettle RP, Bell JE, Simmonds P. Identification of shared populations of human immunodeficiency virus type 1 infecting microglia and tissue macrophages outside the central nervous system. *J Virol*. Dec 2001;75(23):11686–99.
16. Shao KT, Sokal RR. Tree balance. *Systematic Zoology*. Sep 1990;39(3):266–76.
17. Fusco G, Cronk QCB. A new method for evaluating the shape of large phylogenies. *J Theor Biol*. Jul 21, 1995;175(2):235–43.
18. Agapow PM, Purvis A. Power of eight tree shape statistics to detect nonrandom diversification: a comparison by simulation of two models of cladogenesis. *Syst Biol*. Dec 2002;51(6):866–72.
19. Purvis A, Katzourakis A, Agapow PM. Evaluating phylogenetic tree shape: Two modifications to Fusco & Cronk's method. *J Theor Biol*. Jan 7 2002;214(1):99–103.
20. Gray RR, Pybus OG, Salemi M. Measuring the temporal structure in serially-sampled phylogenies. *Methods Ecol Evol*. Oct 2011;2(5):437–45.
21. Fitch WM. Toward Defining Course of Evolution—Minimum Change for a Specific Tree Topology. *Systematic Zoology*. 1971;20(4):406-16.
22. Aldous DJ. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Stat Sci*. Feb 2001;16(1):23–34.
23. Mooers AO, Heard SB. Evolutionary process from phylogenetic tree shape. *Q Rev Biol*. Mar 1997;72(1):31–54.
24. McKenzie A, Steel M. Distributions of cherries for two models of trees. *Math Biosci*. Mar 2000;164(1):81–92.
25. Pybus OG, Harvey PH. Testing macro-evolutionary models using incomplete molecular phylogenies. *P Roy Soc Lond B Bio*. Nov 22, 2000;267(1459):2267–72.
26. Pagel M. Detecting Correlated Evolution on phylogenies—a general-method for the comparative-analysis of discrete characters. *P Roy Soc Lond B Bio*. Jan 22, 1994;255(1342):37–45.
27. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*. Aug, 12, 2003;19(12):1572–4.
28. Strickland SL, Gray RR, Lamers S, et al. Efficient Transmission and Persistence of Low Frequency SIVmac251 Variants in CD8-depleted Rhesus Macaques with Different Neuropathology. *J Gen Virol*. Feb 1, 2012.
29. Hecht FM, Busch MP, Rawal B, et al. Use of laboratory tests and clinical symptoms for identification of primary HIV infection. *AIDS*. May 24, 2002;16(8):1119–29.
30. Norström M, Buggert M, Huang W, et al. Multi-level Analysis of HIV-1 Phylodynamic and Immunological Patterns in HLA-B*570+ Subjects Explains Different Rates of Disease Progression. *19th Conference on Retroviruses and Opportunistic Infections (CROI)*. Washington State Convention Center, Seattle, WA, US 2012.
31. Salemi M, Burkhardt BR, Gray RR, Ghaffari G, Sleasman JW, Goodenow MM. Phylodynamics of HIV-1 in Lymphoid and Non-Lymphoid Tissues Reveals a Central Role for the Thymus in Emergence of CXCR4-Using Quasispecies. *PLoS One*. Sep 26, 2007;2(9).
32. Rambaut A, Posada D, Crandall KA, Holmes EC. The causes and consequences of HIV evolution. *Nature Reviews Genetics*. Jan 2004;5(1):52–61.
33. Leslie AJ, Pfafferott KJ, Chetty P, et al. HIV evolution: CTL escape mutation and reversion after transmission. *Nat Med*. Mar 2004;10(3):282–9.
34. Achaz G, Palmer S, Kearney M, et al. A robust measure of HIV-1 population turnover within chronically infected individuals. *Molecular Biology and Evolution*. Oct 2004;21(10):1902–12.
35. Blankson J, Persaud D, Siliciano RF. Latent reservoirs for HIV-1. *Curr Opin Infect Dis*. Feb 1999;12(1):5–11.

Supplementary Material

On the properties of the staircase-ness measure

The staircase-ness measure counts the (i) proportion of sub-trees that are imbalanced (ie, sub-trees where the left child contains more leaves than the right child, or vice-versa). An alternative formulation (ii) is to make the average of all the min(l,r)/max(l,r) values of each sub-tree, where l and r are the number of leaves in the left and right children of a sub-tree. In this work we compared the staircase-ness values against the distribution of such proportions obtained from random trees. However, there are also a few properties of this measure that are worth to be analyzed analytically. First of all, the staircase-ness of perfectly balanced binary trees is always zero, whichever formulation is used. On the other hand, the staircase-ness of perfectly imbalanced trees (ie, ladder-like trees) is always one when counting the proportions (ie, formulation i), whilst depends on the number of leaves when performing the average (ie, formulation ii). Specifically, the staircase-ness values for perfectly imbalanced trees using formulation (ii) is $S_n = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{i}$, where n is the number of sub-trees. This formula tends to zero as n increases since the limit of the series $\lim_{n\to\infty} S_n = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{i}$ converges to zero, as previously demonstrated by E. Cesàro (http://en.wikipedia.org/wiki/Ces%C3%A0ro_mean.

The distribution of the staircase-ness values obtained by simulating random trees (function rtree(number_of_tips, branch_length = runif(1)) of the R library "ape") does not pass the Shapiro-Wilk normality test ($P<<0.0001$, even by considering only trees with a number of tips >300), neither resembles a Gamma distribution, whose parameters had been fit on the actual data ($P<<0.0001$, using a Kruskal/Wallis test on simulations). However, the average values of both definitions values look stable across all the tree sizes (Fig. S1), while the standard deviation seems to decrease by increasing the tree size. The limits of the average staircase-ness values for formulation (i) and (ii) are close to 0.61 and 0.64, respectively.

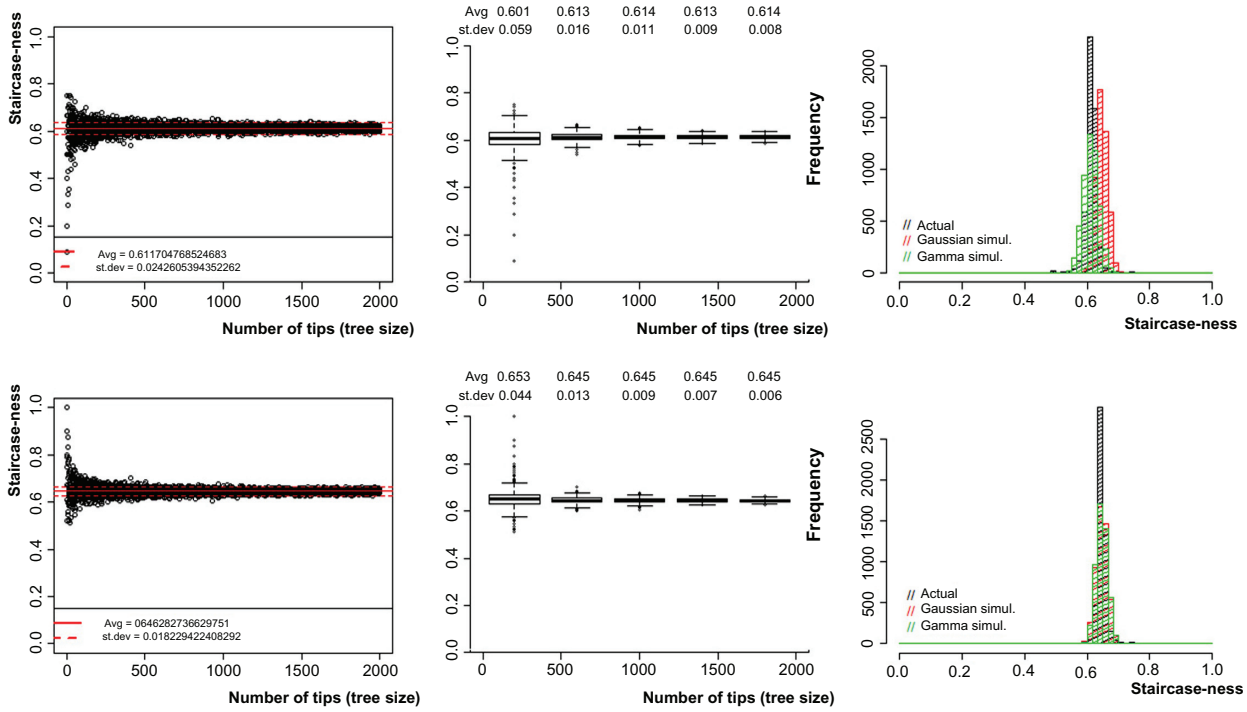

Figure S1. Distribution of staircase-ness values from random trees (5,000 simulations), by varying tree size (from 3 to 2,000 leaves).
Notes: Upper panels show results for formulation (i), whilst lower panels for formulation (ii). Left panels represent the scatterplot of all staircase-ness values depending on the tree size, with a global average and standard deviation indicated in red. Central panels show the boxplots of staircase-ness values by stratifying for tree sizes (5 equal-width intervals spanning tree sizes between 3 and 2,000), with the corresponding stratified average and standard deviation. The right panels show the histograms for all staircase-ness values, and compares them with simulated distributions whose parameters have been fit on the empirical data (Gaussian and Gamma functions).