

1 Identification of hidden population structure in
2 time-scaled phylogenies

3 Erik M. Volz^{1,*}, Carsten Wiuf², Yonatan H. Grad³, Simon D.W.
4 Frost^{4,5}, Ann M. Dennis⁶, and Xavier Didelot⁷

5 ¹*Department of Infectious Disease Epidemiology and MRC Centre for Global*
6 *Infectious Disease Analysis, Imperial College London*

7 ²*Department of Mathematical Sciences, University of Copenhagen*

8 ³*Department of Immunology and Infectious Diseases, TH Chan School of Public*
9 *Health, Harvard University*

10 ⁴*Department of Veterinary Medicine, University of Cambridge*

11 ⁵*The Alan Turing Institute*

12 ⁶*School of Medicine, University of North Carolina Chapel Hill*

13 ⁷*School of Life Sciences and Department of Statistics, University of Warwick*

14 * *Corresponding author: Norfolk Place, W2 1PG, United Kingdom; E-mail:*
15 *e.volz@imperial.ac.uk*

16

Data: 1) Disjoint sets of tips X and Y
 2) Empirical value of test statistic \hat{R}
 3) Number of simulations n_{sim}
 4) Taxonomic condition E (see Equations 3, 4 or 10)
Result: Two-sided p-value denoted $q = \xi(X, Y, \hat{R})$.
 Initialisation;
 Form a time-ordered sequence of nodes

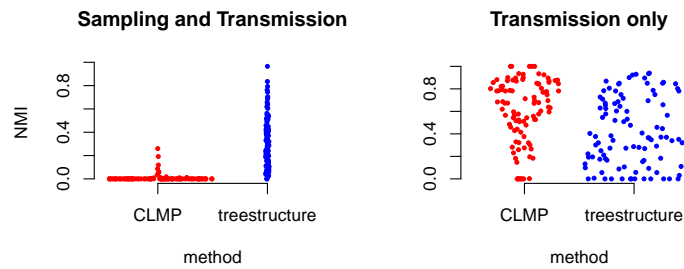
$$U = (u_1, \dots, u_{|D_X|+|D_Y|}) | u_i \in (D_X \cup D_Y), \tau(u_i) \geq \tau(u_{i+1})$$

Form a corresponding numeric sequence:
 $\Upsilon = (v_1, \dots, v_{|D_X|+|D_Y|})$ where

$$v_i = \begin{cases} 1 & \text{if } u_i \in X \\ -1 & \text{if } u_i \in Y \\ 0 & \text{if } u_i \in (D_X \cup D_Y) \cap \mathcal{I} \end{cases}$$

for $k = 1$ **to** n_{sim} **do**
 $z \leftarrow 0$ (simulated lineages through time in clade X) ;
 $w \leftarrow 0$ (simulated lineages through time in clade Y) ;
 $r_{\text{sim}} \leftarrow 0$ (simulated rank-sum statistic) ;
 $c \leftarrow 0$ (number of coalescent events simulated) ;
 for $i = 1$ **to** $|D_X| + |D_Y|$ **do**
 if $v_i = 1$ **then**
 Account for sample in X : $z \leftarrow z + 1$;
 if $v_i = -1$ **then**
 Account for sample in Y : $w \leftarrow w + 1$;
 if $v_i = 0$ **then**
 Increment coalescent counter: $c \leftarrow c + 1$;
 Compute probability $\tilde{p} = \tilde{Q}_E(z, w)$ that next coalescent is in
 D_X or D_Y using Equation 3, 4 or 10;
 Draw a random uniform variable $\omega \leftarrow \text{Unif}(0, 1)$;
 if $\omega < \tilde{p}$ **then**
 $z \leftarrow z - 1$
 $r_{\text{sim}} \leftarrow r_{\text{sim}} + c$
 else
 $w \leftarrow w - 1$
 end
 Record simulated statistic:
 $R_k \leftarrow r_{\text{sim}}$;
end
 Standardize the statistic:
 $\bar{R} \leftarrow (\hat{R} - \langle \{R_k\} \rangle) / \sigma_{R_k}$;
 Return $\min(F(\bar{R}), 1 - F(\bar{R}))$ where F is the standard normal CDF.
Algorithm 1: Algorithm for computing the null distribution and associated p-value of the test-statistic for cladistic outliers.

Data: Time-scaled genealogy \mathcal{G}
Result: Partition of tips of tree, denoted M .
 Initialise ‘active set’ to consist of root node: $\Omega \leftarrow \{\text{root}\}$;
 Initialise partition: $M \leftarrow \emptyset$;
for $u \in \mathcal{I}$ (*internal nodes*) **do**
 | Initialise $\tilde{C}_u \leftarrow C_u$;
end
while $|\Omega| > 0$ **do**
 | Initialise $\Omega' \leftarrow \Omega$;
 for $u \in \Omega$ **do**
 | Find biggest outlier descended from u :
 $v^* \leftarrow \operatorname{argmax}_{v \in C_u} f(v) = \xi(\tilde{C}_u \setminus \tilde{C}_v, \tilde{C}_v)$ (Algorithm 1);
 $q \leftarrow \xi(\tilde{C}_u, \tilde{C}_{v^*})$;
 if $q < \alpha$ **then**
 | $\Omega' \leftarrow \Omega' \cup v^*$;
 $\tilde{C}_u \leftarrow \tilde{C}_u \setminus C_{v^*}$;
 else
 | No significant outliers, so remove u from active sets:
 $\Omega' \leftarrow \Omega' \setminus u$;
 Add the clade descended from u to the partition:
 $M \leftarrow M \cup \{(\mathcal{T} \cap \tilde{C}_u)\}$;
 end
 $\Omega \leftarrow \Omega'$.
end
 Return M .
Algorithm 2: Algorithm for detecting cladistic outliers.



875

876 Figure S1: The normalised mutual information (NMI) for 100 previously
 877 published simulations (McCloskey and Poon 2017). This describes accuracy of
 878 classification of tips into outbreaks using the *treestructure* method and CLMP
 879 (McCloskey and Poon 2017). Results on left were based on simulations where
 880 both transmission and sampling rates varied in the outbreak cluster, whereas
 881 simulations on the right only allowed transmission rates to vary.

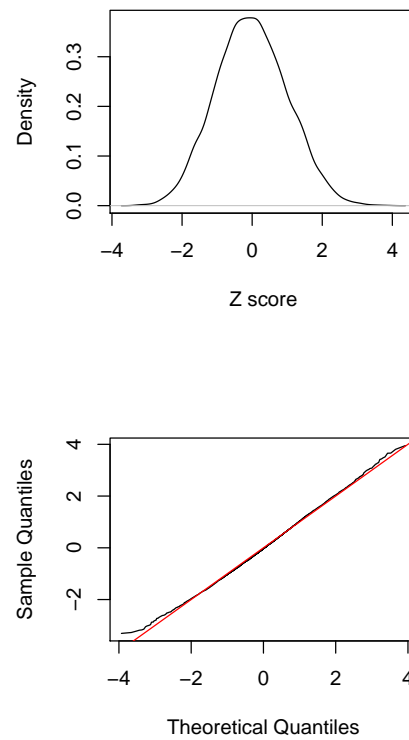


Figure S2: The distribution of the test statistic under the null hypothesis with Kingman coalescent trees simulated with 50 tips. Top: The empirical density of the standardized test statistic (Z score) across internal nodes in 1,000 Kingman coalescent trees. Bottom: A quantile-quantile plot of the Z scores from internal nodes in 1,000 coalescent trees and the standard normal distribution.

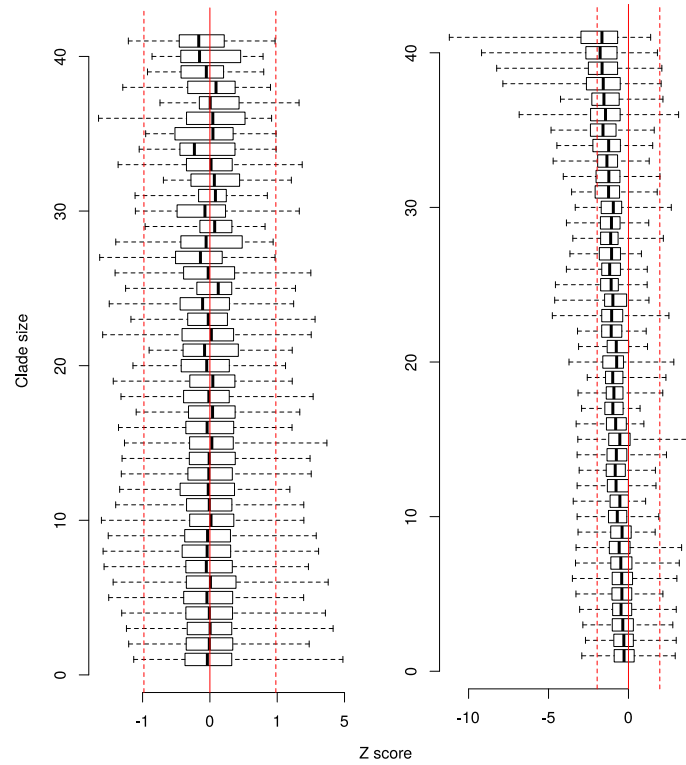


Figure S3: Distribution of the standardized test statistic (Z scores) under the null hypothesis and tabulated by clade size. Each box shows the range (whisker) and interquartile range (box) of Z scores across 1,000 simulated coalescent trees and for a particular clade size (number of tips). The red lines show the interval corresponding to a 95% confidence region. The left part is based on Kingman coalescent trees, while the right part is based on estimated time-scaled phylogenies using simulated sequences as described in the text.

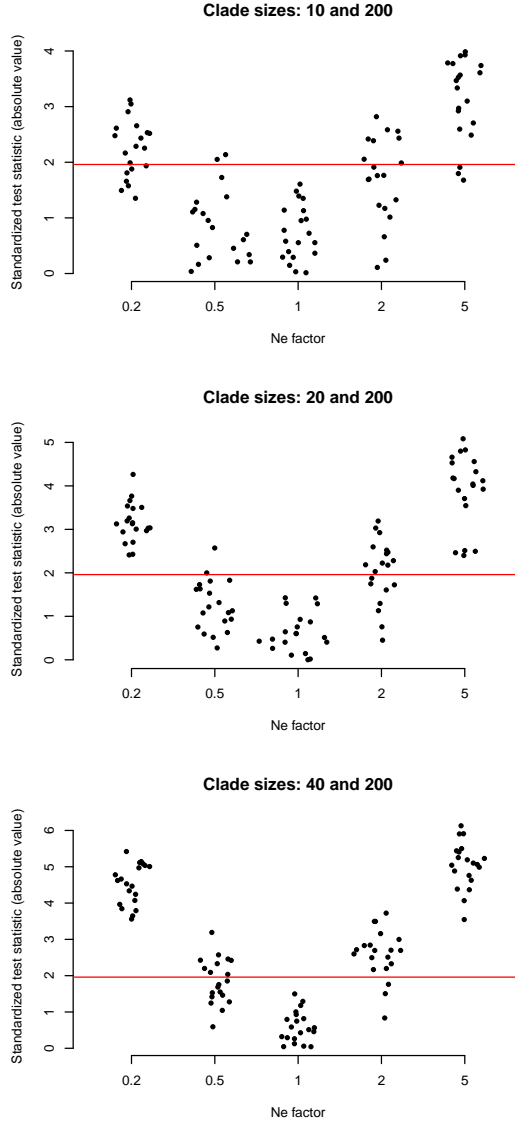
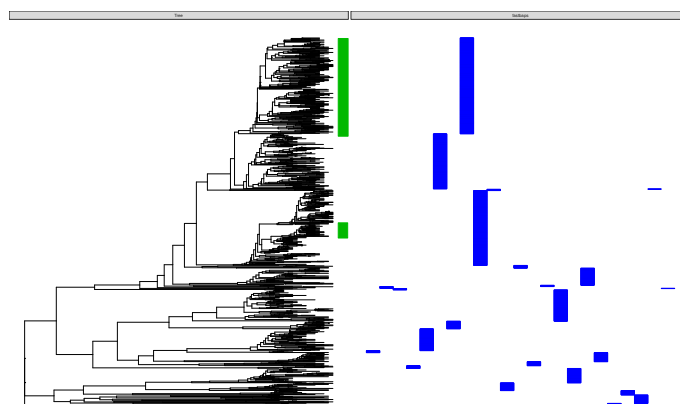
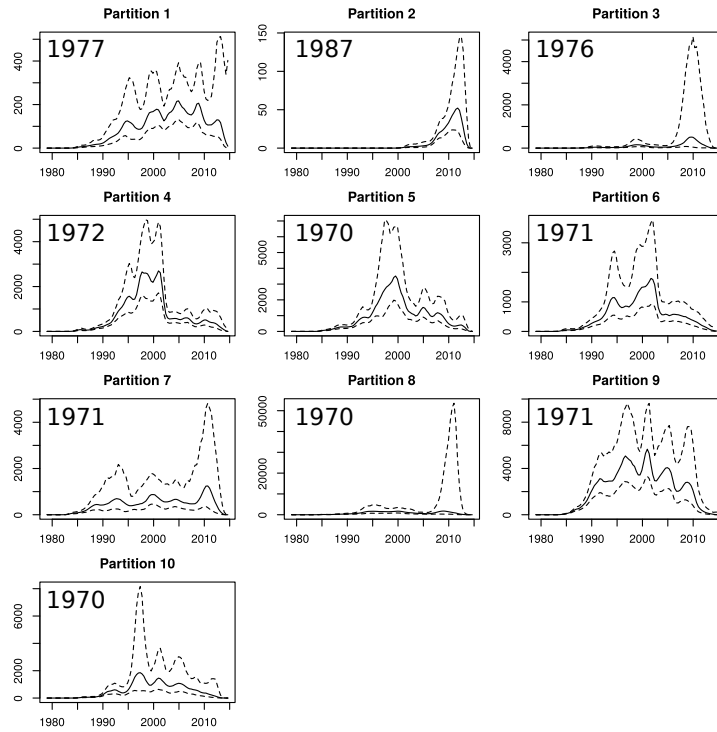


Figure S4: Power to discriminate between clades as a function of sample size and difference in effective population size. Each plot shows the absolute value of the standardized test statistic of the MRCA of a minority clade. The minority clade has an effective population size selected to provide various levels of contrast with the majority clade (see text). The x-axis shows $(N_e^1 w)/(N_e^2 z)$ where z and w are the number of tips in the minority and majority clades, and N_e^1 and N_e^2 are the effective population sizes in the minority and majority clades. The red line corresponds to 1.96 which is the 95% quantile of the standard normal distribution. The top, middle and bottom panels are each based on simulations where the minority clade had 10, 20, and 40 tips respectively, whereas the majority clade always had 200 tips.



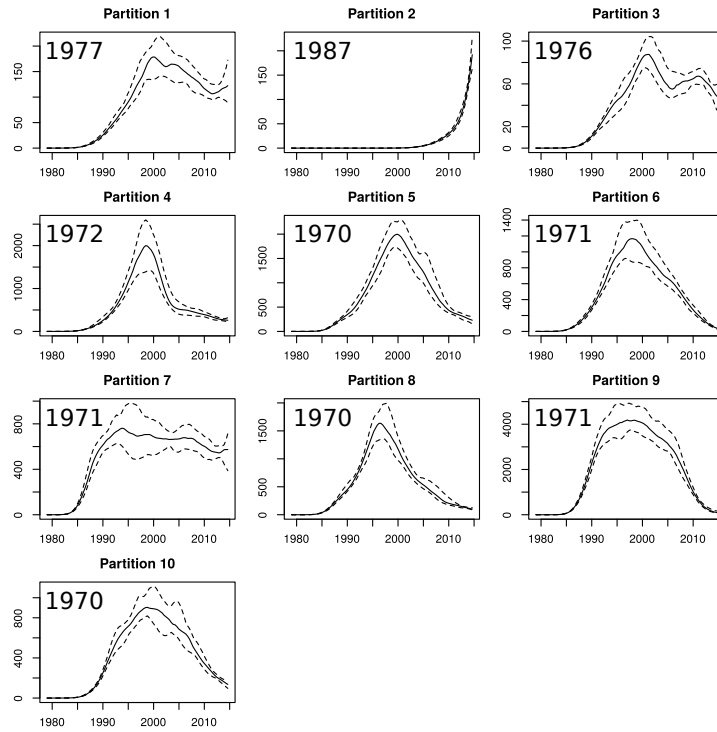
882

883 Figure S5: The output of FastBAPS classification applied to 1102 *N.*
 884 *gonorrhoeae* isolates described in the main text. Clades indicated in green have
 885 CFX resistance.



886

887 Figure S6: Estimated effective population size through time for each partition
 888 in the Tennessee HIV-1 phylogeny. $N_e(t)$ was estimated using the *skygrowth*
 889 method (Volz and Didelot 2018) with precision parameter $\tau = 1$.



890

891 Figure S7: Estimated effective population size through time for each partition
 892 in the Tennessee HIV-1 phylogeny. $N_e(t)$ was estimated using the *skygrowth*
 893 method (Volz and Didelot 2018) with precision parameter $\tau = 100$.