Inferring putative transmission clusters with Phydelity

Alvin X. Han[1,2,3]*, Edyth Parker[3,4], Sebastian Maurer-Stroh[1,5] and Colin A. Russell[3]*

[1]Bioinformatics Institute, Agency for Science, Technology and Research (A*STAR), 30 Biopolis Street, Singapore 138671

[2]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore (NUS), 21 Lower Kent Ridge, Singapore 119077

[3]Laboratory of Applied Evolutionary Biology, Department of Medical Microbiology, Academic Medical Centre, Meibergdreef 9, 1105 AZ Amsterdam-Zuidoost, The Netherlands,

[4]Department of Veterinary Medicine, University of Cambridge, Madingley Rd, Cambridge CB3 0ES, United Kingdom

[5]Department of Biological Sciences, National University of Singapore, 16 Science Drive 4, Singapore 117558

*Corresponding authors: hanxc@bii.a-star.edu.sg or c.a.russell@amc.uva.nl

Abstract: Current phylogenetic clustering approaches for identifying pathogen transmission clusters are centrally limited by their dependency on arbitrarily-defined genetic distance thresholds for within-cluster divergence. Incomplete knowledge of a pathogen's underlying dynamics often reduces the choice of distance threshold to an exploratory, ad-hoc exercise that is difficult to standardise across studies. Phydelity is a new tool for the identification of transmission clusters in pathogen phylogenies. It identifies groups of sequences that are more closely-related than the ensemble distribution of the phylogeny under a statistically-principled and phylogeny-informed framework, without the introduction of arbitrary distance thresholds. Relative to other distance threshold-based and model-based methods, Phydelity outputs clusters with higher purity and lower probability of misclassification in simulated phylogenies. Applying Phydelity to empirical datasets of hepatitis B and C virus infections showed that Phydelity identified clusters with better correspondence to individuals that are more likely to be linked by rapid transmission events relative to other widely-used phylogenetic clustering methods without the need for parameter calibration. Phydelity is generalisable to any pathogen and can be used to reliably identify putative direct transmission events. Phydelity is freely available at https://github.com/alvinxhan/Phydelity.

Introduction

Recent advancements in high-throughput sequencing technologies have led to the widespread use of sequence data in infectious disease epidemiology (Gardy and Loman 2017). In particular, phylogenetics is frequently used to infer genetic clusters underlying the structure of transmission networks (Ambrosioni et al. 2012; Bezemer et al. 2015; Matsuo et al. 2017; de Oliveira et al. 2017; Charre et al. 2018). Current phylogenetic approaches for inferring transmission clusters (primarily 'cutpoint-based' methods) are centrally limited by the need to define arbitrary, absolute cluster divergence thresholds (Prosperi et al. 2011; Ragonnet-Cronin et al. 2013). The lack of a consensus definition of a phylogenetic transmission cluster (Grabowski and Redd 2014) coupled with incomplete knowledge of a pathogen's underlying epidemiological dynamics often reduces the choice of cutpoints to an *ad hoc* exploratory exercise resulting in subjective cluster definitions.

Phydelity is a new tool for inferring putative transmission clusters through the identification of groups of sequences that are more closely-related than the ensemble distribution under a statistically-principled framework. Notably, Phydelity only requires a phylogeny as input, negating the need to define arbitrary cluster divergence thresholds, and also only has a single parameter that can either be user defined or determined directly by Phydelity. Phydelity is freely available at http://github.com/alvinxhan/Phydelity.

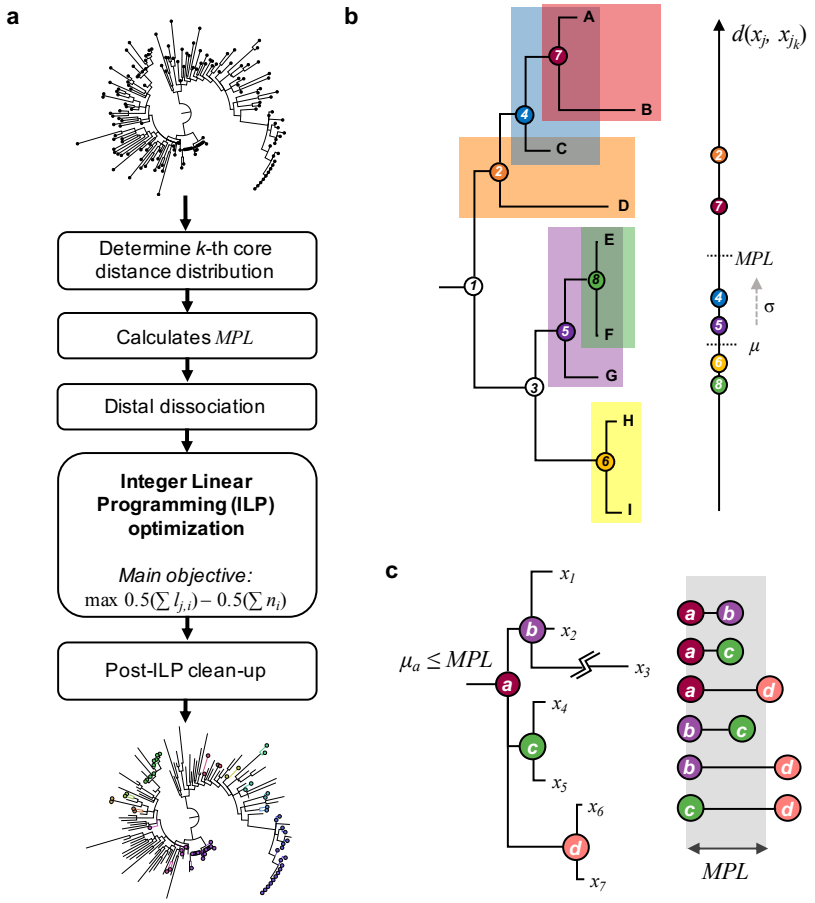


Figure 1. **(a)** Phydelity algorithm pipeline. Phydelity considers the input phylogenetic tree as a collection of putative clusters each defined by an internal node *i* and tips *j* that it subtends. The algorithm first infers the *k*-th core distance distribution ($\mathcal{D}_k$) from the pairwise patristic distances of the closest *k*-neighbouring tips. *k* can be defined by the user or scaled by Phydelity to obtain the supremum $\mathcal{D}_k$ with the lowest divergence. $\mathcal{D}_k$ is then used to compute the maximal patristic distance limit (*MPL*) under which tips are considered to be more closely-related than to the ensemble. Dissociation of distally related subtrees/sequences (Figure 1c) ensues such that both monophyletic and paraphyletic clustering structures can be identified. Phydelity then incorporates the distance and topological information of the remaining nodes and tips into an integer linear programming (ILP) model to be optimised by clustering all tips that satisfy the relatedness constraints within the least number of clusters. Finally, post-ILP steps are implemented to remove any tips that may have been spuriously clustered. **(b)** Determination of the maximal patristic distance limit (*MPL*) using the median (μ) and robust estimator of scale (σ) based on the *k*-th core distance distribution ($\mathcal{D}_k$) of every sequence x_j and its *k*-closest neighbours ($d(x_j, x_{j_k})$; *k*=2 in this case). **(c)** Distal dissociation of a putative cluster subtended by internal node ***a*** where $\mu_a \leq MPL$. Sequence x_3 is dissociated from the putative cluster ***a*** due to its exceedingly long branch length violating the *MPL* threshold (i.e. $d(x_3, x_{3_k}) > MPL$). Additionally, subtree ***d*** is also dissociated from ***a*** as its inter-nodal distance with internal nodes ***b*** and ***c*** exceeds *MPL*.

Method

Clustering Algorithm

Phydelity considers the input phylogeny as an ensemble of putative clusters, each consisting of an internal node *i* and the leaves it subtends. The within-cluster diversity of node *i* is measured by its mean pairwise patristic distance (μ_i) Sequences subtended by *i* are

considered for clustering if μ_i is less than the maximal patristic distance limit (*MPL*), under which sequences are considered more closely-related to one another than the ensemble distribution (Figure 1).

Phydelity computes the *MPL* by first calculating the pairwise patristic distance distribution (i.e. *k*-th core distance distribution, $\mathcal{D}_k$) of closely-related tips comprising the pairwise patristic distances of sequence x_j to the closest *k*-neighbouring tips (i.e. $d(x_j, x_{j_k}) = d_l$) wherein their closest *k*-neighbours include sequence x_j as well. Additionally, $\mathcal{D}_k$ is incrementally sorted ($d_l \leq d_{l+1}$) and truncated up to d_L if the log difference between d_L and d_{L+1} is more than zero:

$$\mathcal{D}_k = \left\{ d_1, \dots, d_l, d_{l+1}, \dots, d_L \middle| d_l \leq d_{l+1}, \lg\left(\frac{d_{l+1} - d_l}{d_l}\right) \leq 0 \right\}$$

The user can opt to either input the desired *k* parameter or allow Phydelity to automatically scale *k* to the value that yields the supremum *k*-th core distance distribution with the lowest overall divergence. This is done by testing if $\mathcal{D}_{k+1}$ and $\mathcal{D}_k$ are statistically distinct ($p < 0.01$) using the Kuiper's test (see Supplementary Materials). All clustering results of Phydelity presented in this work were performed using the autoscaled value of *k*.

The *MPL* is then calculated by:

$$MPL = \bar{\mu} + \sigma$$

where $\bar{\mu}$ is the median pairwise distance of $\mathcal{D}_k$ and σ is the corresponding robust estimator of scale without assuming symmetry about $\bar{\mu}$ (Figure 1b, see Supplementary Materials).

This is then followed by dissociation of distantly-related descendant subtrees/sequences to all putative nodes where $\mu_i > MPL$, thereby facilitating identification of both monophyletic as well as nested, paraphyletic clusters (Figure 1c; see Supplementary Materials). Phydelity filters outlying tips from putative clusters under the assumptions that viruses infecting individuals in a rapid transmission chain likely coalesce to the same most recent common ancestor (MRCA). Additionally, Phydelity requires any clonal ancestors in between the MRCA and tips of a putative cluster to be as genetically similar to each other as they are to the tips of the cluster. As such, for a putative transmission cluster, the mean pairwise nodal distance between all internal and tip nodes of a cluster must also be $\leq MPL$.

An integer linear programming (ILP) model is implemented and optimised under the objective to assign cluster membership to sequences satisfying the aforementioned relatedness criteria within the least number of clusters. In other words, Phydelity uses ILP optimisation to search for the clustering configuration that favours the designation of larger clusters of closely-related sequences which are likely linked by rapid transmission events. Lastly, topologically outlying singletons that were spuriously clustered are removed. The full algorithm description and mathematical formulation of Phydelity is detailed in Supplementary Materials.

Assessing clustering results of simulated epidemics

Phydelity was evaluated on phylogenetic trees derived from simulated HIV epidemics of a hypothetical men who have sex with men (MSM) sexual contact network (C-type networks in Villandre *et al.*, 2016). The simulated sexual contact network comprised 100 subnetworks (communities) sampled from an empirical distribution obtained from the Swiss HIV Cohort Study. All communities were linked in a chain initially and additional connections were generated at a probability of 0.00075. Subjects in the network can either be in the "susceptible", "infected" or "removed" (i.e. individual is diagnosed and sampled) state. Quick transmission chains (i.e. transmission clusters) were attributed to sexual contact among individuals belonging to the same community.

300 epidemics were simulated for four different weights of inter-community transmission rates (i.e. w = 25%, 50%, 75% or 100% of the within-community rate). Two infected individuals were randomly introduced in any of the 100 communities. Transmission time along an edge followed an exponential distribution with rates directly proportional to the associated weights. Time until removal was based on a shifted exponential distribution with the shift representing the minimum amount of time required for a virus to be transmitted to susceptible neighbours. The simulation ended once 200 individuals were in the "removed" state.

These simulated datasets were tested by Villandre *et al.* (2016) to compare the outputs of four "cutpoint-based" phylogenetic clustering methods where the arbitrary distance threshold defining a transmission cluster (i.e. cutpoint) was computed as the: (i) absolute patristic distance threshold between any two tips (Brenner et al. 2007); (ii) standardised number of nucleotide changes (i.e. ClusterPicker, Ragonnet-Cronin *et al.*, 2013); (iii) percentile of the

phylogeny's pairwise sequence patristic distance distribution (i.e. PhyloPart, Prosperi *et al.*, 2011) and (iv) height of an ultrametric tree obtained using the weighted pair-group method of analysis (WPGMA). For each method, Villandre *et al*. varied the corresponding cutpoint parameter over an equivalent range of thresholds. Comparing the output clusters generated by the four methods at their respective optimal cutpoint by adjusted rand index (see below), it was found that the WPGMA method tended to produce clusters with better correspondence to the underlying sexual contact structure. As such, clustering results from Phydelity were compared to those obtained by Villandre *et al*. using the WPGMA method. Additionally, Phydelity was also compared to the multi-state birth-death (MSBD) method which inferred transmission clusters on the same simulated datasets by detecting significant changes in transmission rates (Barido-Sottani et al. 2018).

To assess and compare the output clusters from Phydelity and the aforementioned clustering methods that had been tested on these networks previously, several metrics were used to measure how well the clustering results corresponded with the known sexual contact network:

i. Adjusted rand index (*ARI*) measures the accuracy of the clustering results by computing the frequencies whereby a pair of sequences of the identical (or distinct) subnetwork(s) was assigned to the same (or different) cluster(s) (Hubert and Arabie 1985). ARI ranges between -1 (matching between output clusters and community labels is worse than random clustering) and 1 (perfect match between output clusters and ground truth).

ii. Modified Gini index (I_G). Gini impurity, commonly used in decision tree learning, refers to the probability of a randomly selected item from a set of classes would be incorrectly labelled if it was randomly labelled by the distribution of occurrences in the class set (Breiman et al. 1984). Here, I_G measures how often a randomly selected sequence from the given network would be incorrectly clustered by the inferred clusters. For a sexual contact network with T communities (i.e. $t \in \{1, 2, \dots, T\}$), I_G is computed as:

$$I_G = \sum_{t=1}^{T} \left[p_t \left(1 - \sum_{c=1}^{C^*} p(c|t) \right) \right]$$

where C^* is the set of clusters defined to have correctly classified sequences attributed to community t (i.e. any cluster that constitutes the largest proportion of sequences from community t at both the cluster and the community label levels), p_t is the

177 probability of sequence from community *t* and $p(c|t)$ refers to the probability that a
178 sequence is clustered under cluster *c* conditional of it being from community *t*. If output
179 clusters perfectly align with the underlying sexual contact network (i.e. one cluster only
180 constitute one class of community), $I_G = 0$. Conversely, if clustering results are
181 completely random, $I_G = 1$.

182 iii. Purity measures the average extent that the output clusters contain only a single class
183 (i.e. a particular sexual contact community; Manning et al. 2008):

184 $$Purity = \sum_{c=1}^{C} \frac{1}{N_c} \left(\frac{\max_t \{N_{c,t}\}}{N_c} \right)$$

185 where N_c is the size of cluster *c*, $N_{c,t}$ is the number of tips from community *t* clustered
186 under cluster *c* and *C* is the set of all output clusters. Note that purity (as well as I_G) can
187 be inflated if the total number of clusters is large (i.e. if each tip is assigned to a unique
188 cluster, purity = 1 and $I_G = 0$).

189 iv. Normalised mutual information (*NMI*) trades off the output clustering quality against
190 the number of clusters (Manning et al. 2008):

191 $$NMI = \frac{I(T,C)}{[H(T) + H(C)]/2}$$

192 where *H(T)* and *H(C)* are the respective entropies of the network communities and
193 output clusters, and *I(T,C)* is the mutual information between them. If clustering is
194 random with respect to the network community labels, $I(T,C) = 0$ (i.e. $NMI = 0$). On
195 the other hand, maximum mutual information is achieved (i.e. $I(T,C) = I(T,C)_{max}$)
196 either when the output clusters map the sexual contact network perfectly or all clusters
197 have one member only. Hence, to penalise large cardinalities while normalising $I(T,C)$
198 between 0 and 1, *NMI* is calculated since (a) entropy increases with increasing number
199 of clusters and (b) $[H(T) + H(C)]/2$ is a tight upper bound to $I(T,C)$.

200

201 *Comparison to ClusterPicker and PhyloPart*

202 Phydelity was also tested on two empirical datasets – acute hepatitis C infections among men
203 who have sex with men and hepatitis B viruses collected from members of the same families.
204 Phylogenetic trees for both datasets were reconstructed using RAxML under the
205 GTRGAMMA model (Stamatakis 2014).

206

ClusterPicker (Ragonnet-Cronin et al. 2013) and PhyloPart (Prosperi et al. 2011), two popular phylogenetic clustering tools that are methodologically comparable to Phydelity, were also applied to the same datasets for comparisons. However, other than the phylogenetic tree, both ClusterPicker and PhyloPart also require users to input an arbitrarily-defined genetic distance threshold (as an absolute distance limit for ClusterPicker and percentile of the global pairwise patristic distance for PhyloPart). As such, a range of distance limits (PhyloPart: 0.5-10th percentile; ClusterPicker: 0.005-0.1 nucleotide/site) were applied to both tools. No bootstrap support threshold were implemented for comparability to Phydelity.

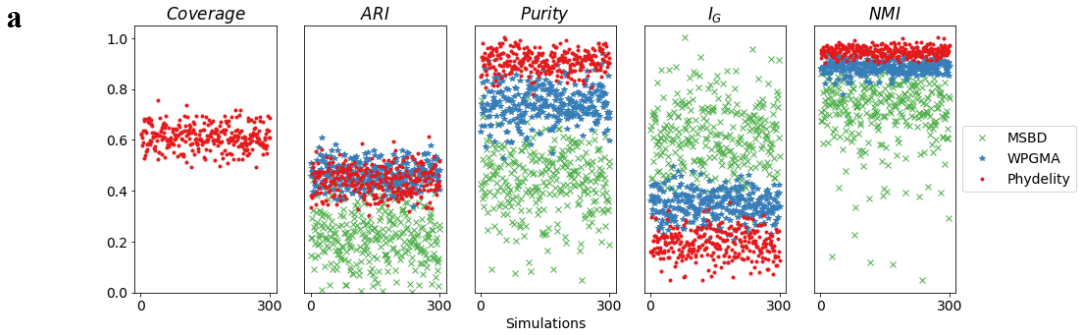
The lowest optimal threshold for the distance range tested was found by maximisation of the mean silhouette index (*SI)* for both ClusterPicker and PhyloPart. The Silohuette index measures how similar an item is to members of its own cluster as opposed to the nearest neighbouring clusters - i.e. a larger mean silhouette index indicates that items of the same cluster are more closely related amongst themselves than to its neighbours (Rousseeuw, 1987). No parameter optimisation was required for Phydelity.

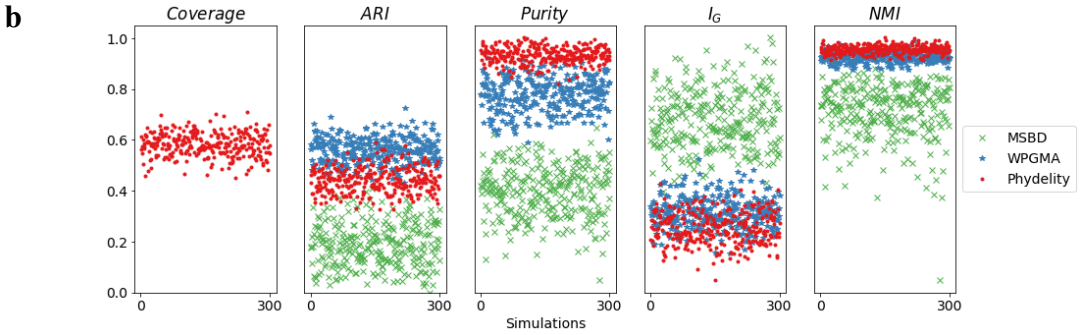
Results

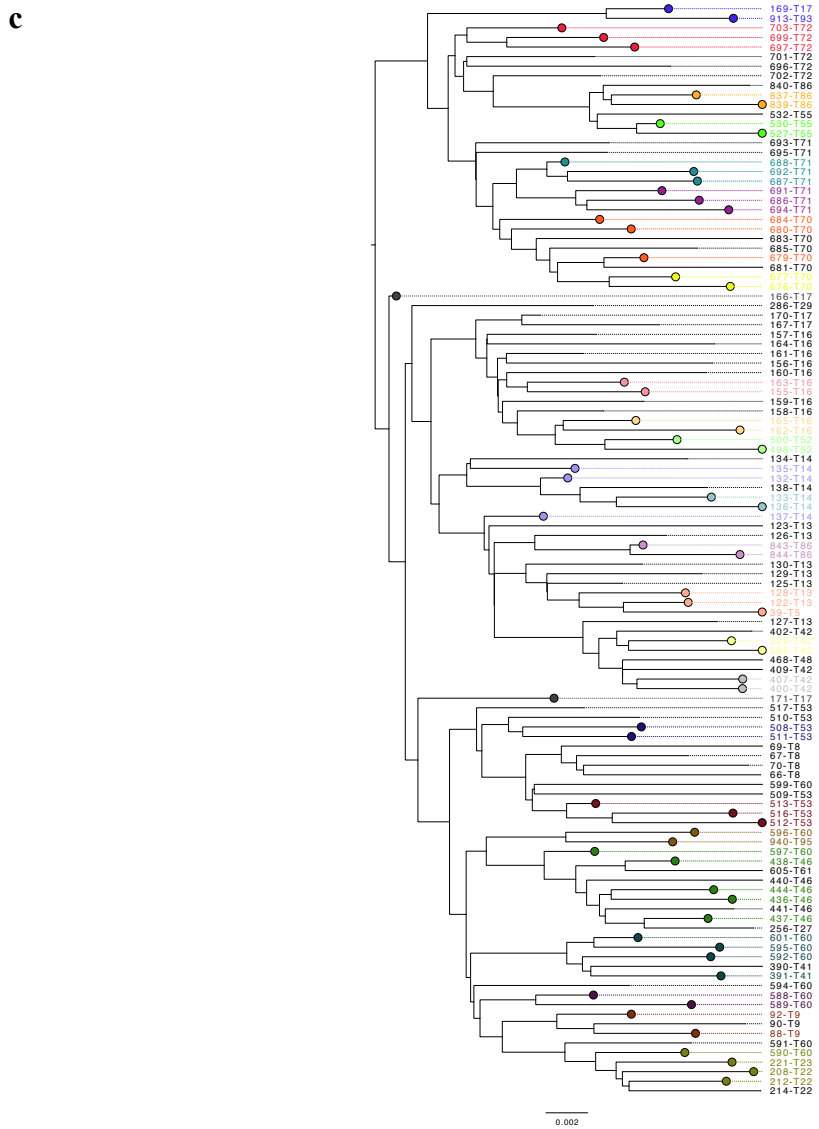
Simulated HIV epidemics

Phydelity was applied to simulated HIV epidemics among men who have sex with men (MSM) belonging to a hypothetical sexual contact network structures where transmission clusters were attributed to quick transmission chains due to sexual contact among individuals belonging to the same subnetwork (see Methods; Villandre *et al.*, 2016). These simulations were originally used to assess the performance of "cutpoint-based" clustering tools, including ClusterPicker, PhyloPart as well as the weighted pair-group method of analysis (WPGMA) which generally attained the highest adjusted rand-index (ARI) score across all simulations when calibrating their respective cutpoint thresholds against the ground-truth. Phylogenetic trees generated from these simulations were also tested by the multi-state birth death (MSBD) method (Barido-Sottani et al. 2018).

Clustering results from Phydelity were compared to outputs from the MSBD method and those achieving the best ARI scores based on the WPGMA method. The purity, modified Gini index (I_G) and normalised mutual information (*NMI*) measures were also used to provide a more comprehensive assessment of the clustering results (Figure 2, Supplementary Figure 3 and Supplementary Table 1; see Methods).







241

Figure 2. Clustering results of simulated HIV epidemics in a hypothetical MSM sexual contact network. **(a)** Clustering metrics for clustering algorithms (Phydelity, weighted pair-group method of analysis (WPGMA) and multi-state birth death (MSBD) methods) applied simulated phylogenies with inter-communities transmission rates weighted at half of within-community rates (i.e. $w = 0.5$). Coverage refers to the proportion of tips clustered by Phydelity. Adjusted rand index (*ARI*) measures how accurate the output clusters corresponded with the community labels. *Purity* gives the average extent clusters contain only a single class of community. Modified Gini index (I_G) is the probbility that a randomly selected sequence would be incorrectly clustered. Normalised mutual information (*NMI*) accounts for the tradeoff between clustering quality and number of clusters. **(b)** Results for simulations where inter-communities transmission rates were identical to within-community rates (i.e. $w = 1.0$). **(c)** Sample output clusters of Phydelity for a subtree of an example simulation ($w = 0.5$). Tips that were clustered by Phydelity are distinctly coloured according to their cluster membership. By relaxing the monophyletic assumption, Phydelity is capable of detecting paraphyletic clusters (e.g. transmission pair 166-T17 and 171-T17 and cluster subtending 132-T14, 135-T14 and 137-14).

The phylogenetic trees generated from the simulations had a large number of clusters that were relatively small in size (i.e. percentage of sequences that were part of ground truth clusters with sizes < 8 tips = 33.9% ($w = 25\%$); 55.5% ($w = 100\%$); see Barido-Sottani *et al.* (2018) for more details). Furthermore, these ground truth clusters were not all monophyletic (Figure 2c). As a result, while Phydelity and WPGMA yielded comparable ARI scores (Phydelity: 0.44-0.45 (s.d. = 0.05); WPGMA: 0.44-0.56 (s.d = 0.05-0.05); Supplementary Table 1), the output clusters of Phydelity, which can be paraphyletic (Figure 2c), are purer (mean purity; Phydelity: 0.81-0.88 (s.d. = 0.03); WPGMA: 0.67-0.74 (s.d. = 0.06-0.06)) and have a lower probability of misclassification when compared to WPGMA which assumes clusters are strictly monophyletic (mean I_G; Phydelity: 0.27-0.28 (s.d. = 0.04-0.05); WPGMA: 0.33-0.40 (s.d. = 0.04-0.05)). Coverage of sequences clustered by Phydelity lies between 58.2% and 61.6%.

The clustering results from WPGMA presented in this work were based on the optimal distance threshold derived by calibration against the simulated ground-truth. Notably, Phydelity's auto-scaling mitigates the need for threshold calibration and enables application to empirical datasets where ground truth clustering is unavailable, as is largely the case for epidemiological studies.

Hepatitis B virus transmission between family members

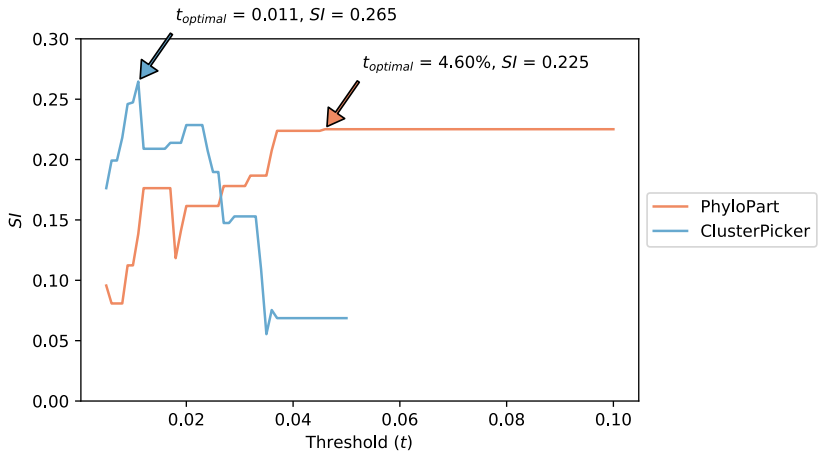
Phydelity was tested on empirical datasets to demonstrate its applicability on real-world data, including hepatitis B viruses (HBV) collected from residents in the Binh Thuan Province of Vietnam. In such highly endemic regions, HBV is commonly transmitted either vertically

from mothers to children during the perinatal period or horizontally between cohabitants of the same household (Matsuo et al. 2017).

As complete genome nucleotide sequences were not available for all individuals, a phylogenetic tree was reconstructed using the viral polymerase sequences collected from 41 patients, of which 12 of them were confirmed to be members of three families (i.e. denoted as F2, F3 and F4) by a family survey as well as mitochondrial analyses. Besides Phydelity, the resulting phylogeny was also implemented in ClusterPicker and PhyloPart for comparison. While WPGMA performed better in the simulations by Villandre *et al.*, ClusterPicker and PhyloPart are arguably the more widely-used phylogenetic clustering tools to date.

Phydelity identified three likely transmission clusters that distinguish between the separate family households (Figure 3). At their respective optimal distance thresholds by mean Silhouette index (see Methods), ClusterPicker and PhyloPart achieved similar clustering results as well. Importantly, however, Phydelity was able to obtain the same optimal clustering results without having to optimise and implement a hard-to-interpret distance parameter.

a



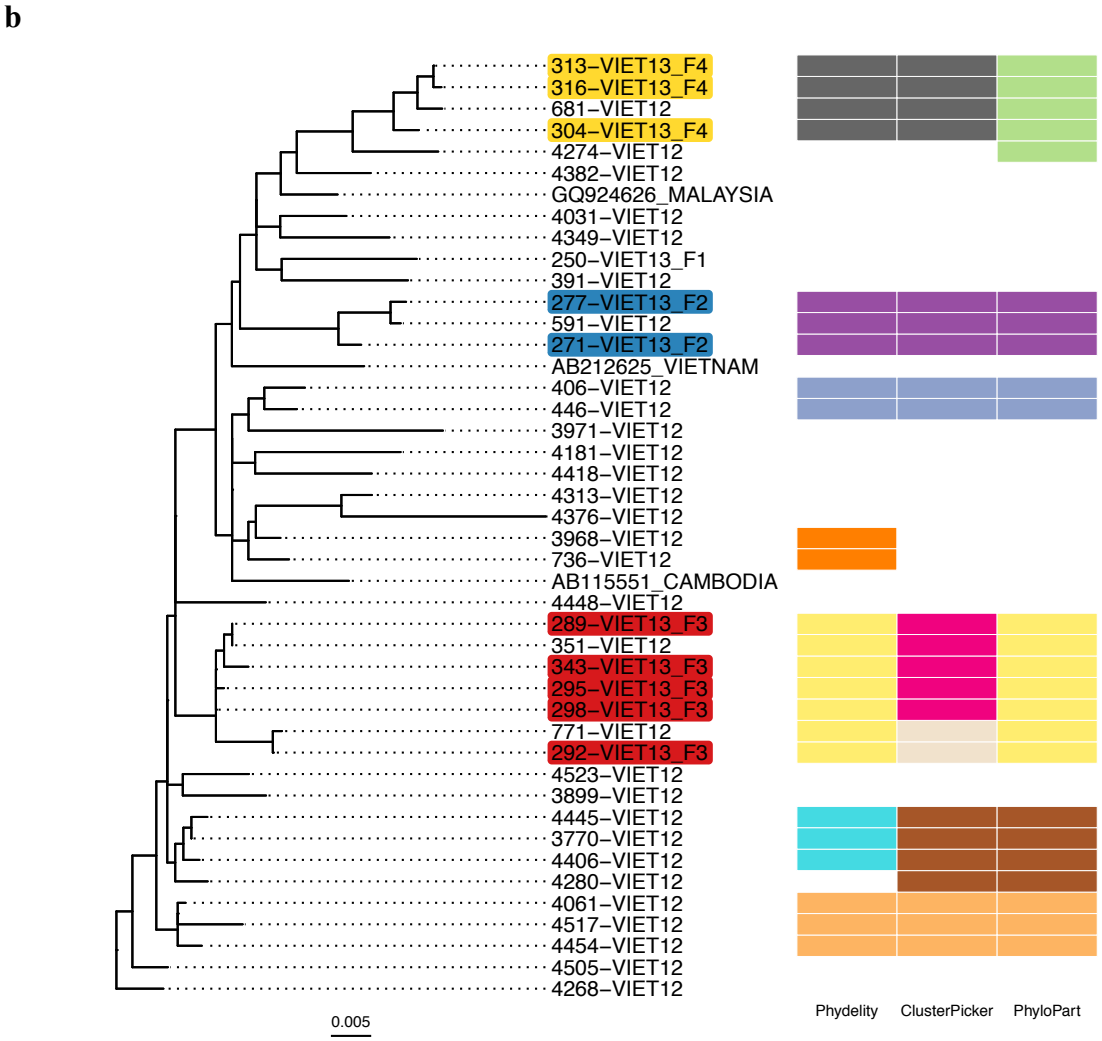


Figure 3. Clustering results of hepatitis B viruses (HBV) collected from residents in the Binh Thuan Province of Vietnam. (**a**) Plots of mean Silhouette index (*SI*) computed for the range of genetic distance thresholds implemented in ClusterPicker and PhyloPart. Clustering results from the lowest optimal distance threshold ($t_{optimal}$) with the highest *SI* value for each method were compared to Phydelity as depicted in **b** (ClusterPicker: $t_{optimal}$ = 0.011 nucleotide/site, *SI* = 0.265; PhyloPart: $t_{optimal}$ = 4.60%, *SI* = 0.225). Plot for ClusterPicker is truncated at ~0.05 nucleotide/site as the entire tree collapsed to single cluster after this threshold. (**b**) Maximum likelihood phylogeny of HBV polymerase sequences derived from viruses collected from 41 patients. 12 patients were confirmed to be members of three separate family households (F2, F3 and F4; tip names shaded with a distinct colour for each family). Clustering results from Phydelity are depicted as a heatmap alongside outputs from ClusterPicker and PhyloPart based on their respective $t_{optimal}$. Each distinct colour of the heatmap cells denotes a different cluster.

Hepatitis C virus transmission among MSM

Incidence of HCV infections among HIV-negative MSM has been relatively limited as compared to their HIV-positive counterparts. However, the recent uptake of pre-exposure prophylaxis (PrEP) among HIV-negative individuals to prevent HIV infection could pose higher risk of sexually transmitted HCV infections (Volk et al. 2015; Charre et al. 2018). In a

study on HIV-positive and HIV-negative MSM patients in Lyon, 108 cases of acute HCV infections (80 primary infections; 28 reinfections) were reported between 2014 and 2017 among 96 MSM (72 HIV-positive; 24 HIV-negative, of which 16 (67%) of them were on PrEP). Separate phylogenetic analyses were performed on a subset of 89 (68 HIV-positive; 21 HIV-negative) HCV isolates belonging to genotypes 1a and 4d based on their NS5B sequences. Additionally, 25 HCV sequences from HIV-infected MSM collected before 2014 were included along with 60 control HCV sequences derived from HIV-negative, non-MSM patients residing in the same geographical area as controls. All sequences collected from MSM patients were given strain names in the format of "MAH(ID)_accession" while control sequences from non-HIV, non-MSM patients were denoted as "NCH(ID)_accession" (Figure 4). Phydelity as well as ClusterPicker and PhyloPart were applied to the reconstructed phylogenies, with the latter calibrated over a range of distance thresholds. Again, only clustering results based on the lowest distance threshold maximising the mean Silhouette index for ClusterPicker and PhyloPart were compared to Phydelity's output clusters (see Methods).

Generally, membership of the MSM transmission clusters and pairs identified by Phydelity across both genotypes were strictly limited to sequences derived from MSM patients. Relaxing the monophyletic assumption by dissociating distantly-related tips from putative monophyletic clusters (see Methods) enables Phydelity to identify likely outlying sequences as evidenced by their relatively longer branch lengths from the cluster ensemble (Table 1 and Figure 4; Genotype 1a: cluster C1 – MAH66 and cluster C3 – MAH31, MAH62 and MAH72; Genotype 4d: cluster C3 – MAH24 and MAH08). In particular, for genotype 1a, even though the mean pairwise distance of MAH72 to members of cluster C3 is within a standard deviation of the latter's within-cluster diversity, its distance to the more distant members (e.g. MAH15 and MAH40, Figure 4) violated the inferred *MPL* (Table 1). Additionally, as a result of distal dissociation, Phydelity distinguishes clusters that are genetically more alike amongst themselves than to those phylogenetically ancestral to it (e.g. cluster C1.1 from C1 for genotype 1a; Figure 4a).

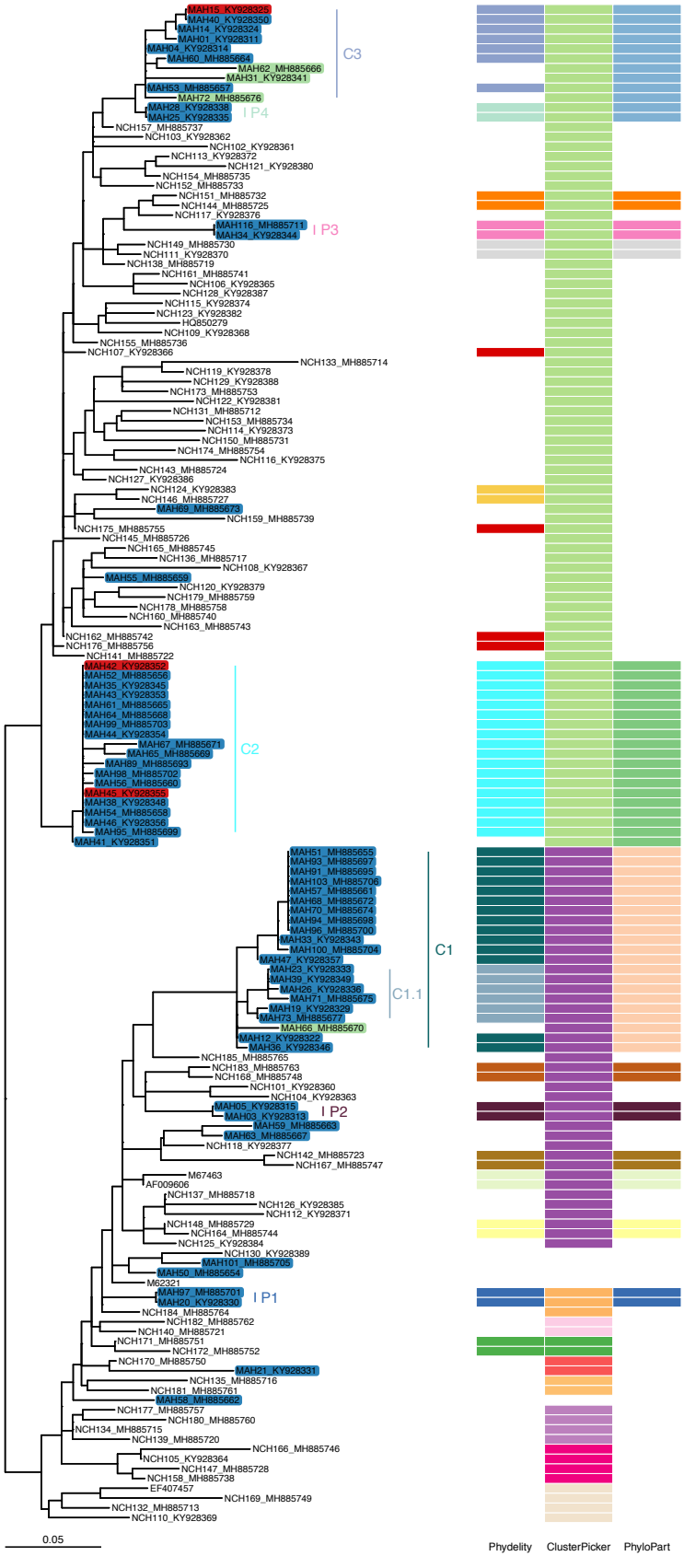
For both genotypes, Phydelity found multiple clusters that included both HIV-positive and HIV-negative MSM patients (i.e. Genotype 1a: clusters C2 and C3; Genotype 4d: clusters C2 and C2.2, as well as pair P2). While it is not clear which of the HIV-negative patients were on PrEP (information is not given in the original paper), the clustering results from Phydelity

348 were in line with the findings by Charre *et al*. that acute HCV infections among HIV-
349 negative MSM were likely sourced from their HIV-positive counterparts.
350
351 While ClusterPicker did manage to consolidate all of the MSM genotype 4d sequences into a
352 single monophyletic cluster, its clustering of genotype 1a was clearly problematic as a large
353 number of non-MSM control sequences were clustered together with those from MSM
354 patients. PhyloPart's optimal clustering output was consistent Phydelity's for genotype 1a.
355 However, the larger number of identical sequences in the genotype 4d tree skewed the
356 optimal distance parameter (expressed as *x*-th percentile of the pairwise patristic distribution
357 of the entire phylogeny) to only cluster these identical sequences.
358

Genotype	*MPL*	Cluster	Mean pairwise patristic distance of cluster (σ)	Outlier	Mean pairwise patristic distance of outliers to cluster members (σ)
1a	0.029	C1	0.011 (0.012)	MAH66	0.043 (0.009)
		C3	0.016 (0.009)	MAH62	0.045 (0.027)
				MAH31	0.041 (0.025)
				MAH72	0.022 (0.015)
4d	0.010	C1	0.006 (0.004)	MAH24	0.019 (0.006)
				MAH08	0.009 (0.005)

359 **Table 1**: Comparing the genetic distance between outlying tips and the clusters they coalescence to with the
360 genetic diversity of those clusters.

a.



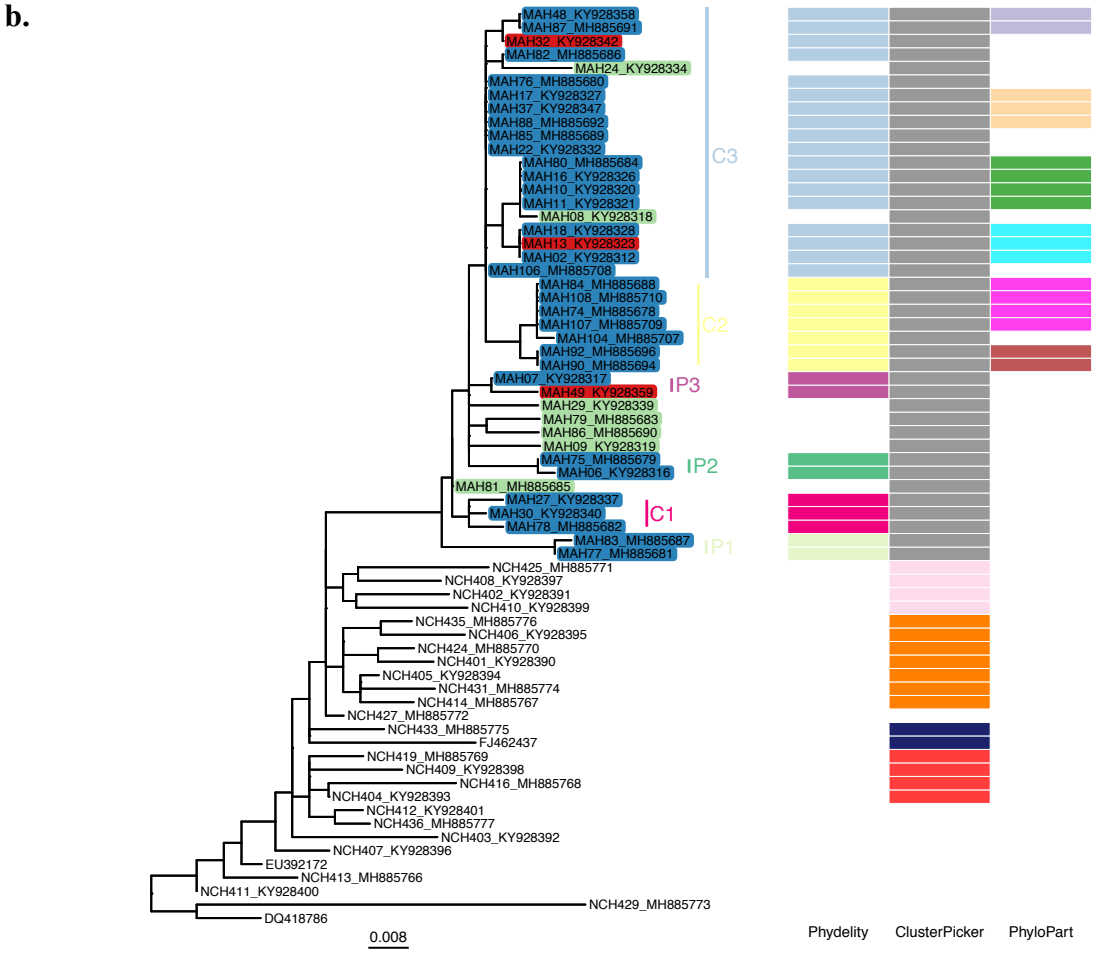


Figure 4. Maximum likelihood phylogeny and clustering results of hepatitis C viruses (HCV) obtained from men who have sex with men (MSM) in Lyon, France. All highlighted tip names denoted in the format "MAH(ID)_accession" were samples from MSM patients (blue: HIV-positive, red: HIV-negative, green: HIV-positive and considered as outlying sequences by Phydelity). Non-highlighted tips were collected from non-HIV, non-MSM patients residing in the same geographic region and time period. Clustering results from Phydelity, ClusterPicker and PhyloPart are depicted as a heatmap. Each distinct colour refers to a different cluster. Similar to the Vietnamese hepatitis B empirical viral datasets (Figure 3a and Supplementary Figure 4), mean Silhouette index was used as the optimality criterion to determine the optimal absolute distance threshold for ClusterPicker and PhyloPart. Only results based on the optimised thresholds are shown here for ClusterPicker and PhyloPart. No parameter optimisation is required for Phydelity. **(a)** Genotype 1a. **(b)** Genotype 4d.

Computational performance

For computational performance, Phydelity can process a phylogeny of 1000 tips, on an Ubuntu 16.04 LTS operating system with an Intel Core i7-4790 3.60 GHz CPU, in ~3 minutes using a single CPU core and 253 MB of peak memory usage.

Discussion

Phydelity is a statistically-principled tool capable of identifying putative transmission clusters from pathogen phylogenies without the need to introduce arbitrary distance thresholds.

Instead, Phydelity infers the maximal patristic distance limit (*MPL*) for cluster designation using the pairwise patristic distance distribution of closely-related tips in the input phylogenetic tree. Additionally, unlike other cutpoint-based methods, Phydelity does not assume clusters are strictly monophyletic and can identify paraphyletic clustering owing to its distal dissociation approach. For datasets that span extended periods of time, multiple introductions within the same contact network and concurrent onward transmissions to other communities can result in "nested" introduction events that would go undetected by monophyletic clustering (Barido-Sottani et al. 2018). By relaxing this assumption, not only can Phydelity pick up these "nested" events, it tends to produce clusters that are purer with a lower chance of misclassification while excluding putative outlying tips that are exceedingly distant from the inferred cluster.

There are algorithmic overlaps between Phydelity and PhyCLIP, which is also a statistically-principled phylogenetic clustering algorithm based on integer linear programming optimisation (Han et al. 2019). However, the two clustering tools have substantially different approaches that recover clusters with distinctly different interpretations. PhyCLIP was developed to identify statistically-supported subpopulations in pathogen phylogenies that putatively capture variant ecological, evolutionary or epidemiological processes that could underlie sub-species nomenclature development. As such, PhyCLIP's designated clusters should not be interpreted as sequences linked by rapid transmission events. For instance, when applied to the HCV genotype 1a NS5B dataset, PhyCLIP clustered 131 of the 155 input sequences into seven clades, all of which encompasses genetically similar viruses of both MSM and non-MSM origins that were endemic in Lyon during a specific period in time. In contrast, Phydelity assigned 73 sequences into 12 transmission pairs and 5 transmission clusters that distinguished the underlying MSM transmission events from non-MSM ones (Supplementary Figure 1). A detailed comparison between Phydelity and PhyCLIP can be found in Supplementary Materials.

There are two key assumptions underlying Phydelity's clustering algorithm. First, Phydelity expects transmissions to be linked by events rapid enough such that molecular evolution between the transmitted pathogens is minimal, and thus genetically more similar amongst themselves than to the given ensemble. At least for rapidly evolving pathogens such as RNA viruses, genetic changes between sequences sampled from transmission pairs were found to be generally low (Campbell et al. 2018).

Phydelity also assumes that the transmitted pathogens coalesce to the same most recent common ancestor (MRCA) and that the pairwise genetic distance of internal nodes found between the MRCA and the tips of the cluster to be bounded below *MPL*. Even though Phydelity does not explicitly equate the inferred phylogeny to a transmission tree, imposing a distance threshold between the internal nodes within a phylogenetic cluster may be construed as an implicit assumption that the internal nodes are representative of transmission events. There are central differences in the interpretation of phylogenetic and transmission trees respectively. The former depicts the shared ancestry between the sampled tips while the latter represents the true transmission history between the transmitted pathogens (Pybus and Rambaut 2009; Ypma et al. 2013). It should be noted that Phydelity does not attribute any interpretation of transmission events to the internal nodes and does not relate branch lengths of the phylogenetic tree, which correlates with the timing of coalescence, to transmission times. Restricting the distances between internal nodes below the *MPL* is strictly meant to increase conservatism in identifying clusters that are as closely-related as possible. Any tips clustered within the same cluster should be interpreted as a network of undirected transmission pairs.

Furthermore, incorporating the aforementioned assumptions also means that Phydelity is not exempt from well-documented pitfalls associated with other non-parametric, phylogenetic clustering methods. Firstly, these clustering algorithms largely operate with conservatively low thresholds. Resultantly, cluster identification is biased towards recent infections as opposed to detecting differences in transmission rates between subpopulations. This bias could be further worsened if oversampling occurs (Poon 2016; Dearlove et al. 2017; Le Vu et al. 2018). While this caveat limits the interpretation of phylogenetic clusters, it does not render phylogenetic clustering tools obsolete. As demonstrated by the HBV and HCV empirical studies above, with meta-data associated with the individuals clustered, phylogenetic clustering can still be used to identify infection trends as well as potential risk factors and/or target subpopulations in retrospective studies.

There are a few more limitations that should be noted when using Phydelity. As the *MPL* is wholly informed by the phylogenetic tree, clustering results will consequently be sensitive to the diversity of closely-related tips within the input phylogeny. Specifically, the closely-related sequences that constitute the *k*-th core patristic distance distribution ($\mathcal{D}_k$) must be

homogenous (i.e. similar difference between consecutive distances when $\mathcal{D}_k$ is sorted; see Methods) but sufficiently distinct from the background diversity of the phylogeny. Two scenarios can arise if this is not the case: 1) few to no tips will be clustered if $\mathcal{D}_k$ is not homogenous. This may arise if sampling is so scattered such that few to no transmission pairs are sampled; 2) potentially erroneous clustering of distantly-related tips may be obtained if $\mathcal{D}_k$ has a similar distance distribution relative to the entire tree. This could be possible if sampling rate is too low relative to the mutation rate of the pathogen.

Additionally, constructing a phylogenetic tree can be a computational bottleneck for large sequence datasets. As an alternative, genetic distance-based clustering algorithms such as HIV-TRACE (Kosakovsky Pond et al. 2018) which negate the need to build a phylogenetic tree have becoming increasingly popular. However, HIV-TRACE still requires users to specify an arbitrary absolute distance threshold. Additionally, while it performed better than other phylogenetic clustering method, HIV-TRACE did not preclude problems with bias towards higher sampling rates (Poon 2016).

Despite the limitations discussed above, clustering results generated by Phydelity for the simulation and empirical datasets in this study demonstrate its superior performance over current widely used phylogenetic clustering methods. Importantly, Phydelity obviates the need for users to define or optimise non-biologically-informed distance thresholds. Phydelity is fast, generalisable, and freely available at https://github.com/alvinxhan/Phydelity.

Acknowledgements

We would like to thank Frits Scholer for his help in writing the program, as well as Jelle Koopsen and Velislava Petrova for their key intellectual contributions.

Data availability

Phydelity is freely available on https://github.com/alvinxhan/Phydelity. All simulated datasets were downloaded from Villandre et al. (2016). Genbank accession numbers of HBV polymerase sequences: AB212625, GQ924626, AB115551, LC57377-LC57378, LC60789-LC60790, LC63767, LC64366-LC64378, LC64380-LC64381, LC80779-LC80783, LC80785, LC80787-LC80800, and LC80802-LC80804. Genbank accession numbers of HCV NS5B sequences: AF9606, EF407457, HQ850279, EU392172, FJ462437, DQ418786,

M62321, MH885654-MH885777, and KY928311-KY928401. Jupyter notebooks used to analyse both simulated and empirical datasets can be found in the same aforementioned github repository.

Funding

A.X.H. was supported by the A*STAR Graduate Scholarship programme from A*STAR to carry out his PhD work via collaboration between Bioinformatics Institute (A*STAR) and NUS Graduate School for Integrative Sciences and Engineering from the National University of Singapore. E.P. was funded by the Gates Cambridge Trust (Grant numer OPP1144). S.M.S. was supported by the A*STAR HEIDI programme (Grant number: H1699f0013) and Bioinformatics Institute (A*STAR).

References

Ambrosioni J, Junier T, Delhumeau C, Calmy A, Hirschel B, Zdobnov E, Kaiser L, Yerly S, Study the SHIVC. 2012. Impact of highly active antiretroviral therapy on the molecular epidemiology of newly diagnosed HIV infections. AIDS [Internet] 26. Available from: https://journals.lww.com/aidsonline/Fulltext/2012/10230/Impact_of_highly_active_antiretroviral_therapy_on.10.aspx

Barido-Sottani J, Vaughan TG, Stadler T. 2018. Detection of HIV transmission clusters from phylogenetic trees using a multi-state birth–death model. J. R. Soc. Interface [Internet] 15. Available from: http://rsif.royalsocietypublishing.org/content/15/146/20180512.abstract

Bezemer D, Cori A, Ratmann O, van Sighem A, Hermanides HS, Dutilh BE, Gras L, Rodrigues Faria N, van den Hengel R, Duits AJ, et al. 2015. Dispersion of the HIV-1 Epidemic in Men Who Have Sex with Men in the Netherlands: A Combined Mathematical Model and Phylogenetic Analysis. PLOS Med. [Internet] 12:e1001898. Available from: https://doi.org/10.1371/journal.pmed.1001898

Breiman L, Friedman J, Stone CJ, Olshen RA. 1984. Classification and regression trees. Florida.

Brenner BG, Roger M, Routy J-P, Moisi D, Ntemgwa M, Matte C, Baril J-G, Thomas R, Rouleau D, Bruneau J, et al. 2007. High Rates of Forward Transmission Events after Acute/Early HIV-1 Infection. J. Infect. Dis. [Internet] 195:951–959. Available from: http://dx.doi.org/10.1086/512088

Campbell F, Strang C, Ferguson N, Cori A, Jombart T. 2018. When are pathogen genome sequences informative of transmission events?Parrish C, editor. PLOS Pathog. [Internet] 14:e1006885. Available from: https://dx.plos.org/10.1371/journal.ppat.1006885

Charre C, Cotte L, Kramer R, Miailhes P, Godinot M, Koffi J, Scholtès C, Ramière C. 2018. Hepatitis C virus spread from HIV-positive to HIV-negative men who have sex with men.Shoukry NH, editor. PLoS One [Internet] 13:e0190340. Available from:

https://dx.plos.org/10.1371/journal.pone.0190340

Dearlove BL, Xiang F, Frost SDW. 2017. Biased phylodynamic inferences from analysing clusters of viral sequences. Virus Evol. [Internet] 3:vex020. Available from: +

Gardy JL, Loman NJ. 2017. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat. Rev. Genet. 19:9–20.

Grabowski MK, Redd AD. 2014. Molecular tools for studying HIV transmission in sexual networks. Curr. Opin. HIV AIDS [Internet] 9. Available from: https://journals.lww.com/co-hivandaids/Fulltext/2014/03000/Molecular_tools_for_studying_HIV_transmission_in.6.aspx

Han AX, Parker E, Scholer F, Maurer-Stroh S, Russell CA. 2019. Phylogenetic Clustering by Linear Integer Programming (PhyCLIP).Shapiro B, editor. Mol. Biol. Evol. [Internet]. Available from: https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msz053/5373046

Hubert L, Arabie P. 1985. Comparing partitions. J. Classif. [Internet] 2:193–218. Available from: http://link.springer.com/10.1007/BF01908075

Kosakovsky Pond SL, Weaver S, Leigh Brown AJ, Wertheim JO. 2018. HIV-TRACE (TRAnsmission Cluster Engine): A tool for large scale molecular epidemiology of HIV-1 and other rapidly evolving pathogens.Shapiro B, editor. Mol. Biol. Evol. 35:1812–1819.

Manning CD, Raghavan P, Schütze H. 2008. Introduction to Information Retrieval. New York, NY, USA: Cambridge University Press

Matsuo J, Do SH, Yamamoto C, Nagashima S, Chuon C, Katayama K, Takahashi K, Tanaka J. 2017. Clustering infection of hepatitis B virus genotype B4 among residents in Vietnam, and its genomic characters both intra- and extra-family.Chemin I, editor. PLoS One [Internet] 12:e0177248. Available from: https://dx.plos.org/10.1371/journal.pone.0177248

de Oliveira T, Kharsany ABM, Gräf T, Cawood C, Khanyile D, Grobler A, Puren A, Madurai S, Baxter C, Karim QA, et al. 2017. Transmission networks and risk of HIV infection in KwaZulu-Natal, South Africa: a community-wide phylogenetic study. Lancet HIV [Internet] 4:e41–e50. Available from: https://doi.org/10.1016/S2352-3018(16)30186-2

Poon AFY. 2016. Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks. Virus Evol. 2:vew031.

Prosperi MCFF, Ciccozzi M, Fanti I, Saladini F, Pecorari M, Borghi V, Di Giambenedetto S, Bruzzone B, Capetti A, Vivarelli A, et al. 2011. A novel methodology for large-scale phylogeny partition. Nat. Commun. [Internet] 2:321. Available from: https://doi.org/10.1038/ncomms1325

Pybus OG, Rambaut A. 2009. Evolutionary analysis of the dynamics of viral infectious disease. Nat. Rev. Genet. [Internet] 10:540–550. Available from: http://www.nature.com/articles/nrg2583

Ragonnet-Cronin M, Hodcroft E, Hué S, Fearnhill E, Delpech V, Brown AJLAAL, Lycett SS, Holmes E, Nee S, Rambaut A, et al. 2013. Automated analysis of phylogenetic clusters. BMC Bioinformatics [Internet] 14:317. Available from: https://doi.org/10.1186/1471-2105-

559 14-317

560 Rousseeuw PJ. 1987. Silhouettes: A graphical aid to the interpretation and validation of
561 cluster analysis. J. Comput. Appl. Math. 20:53–65.

562 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
563 large phylogenies. Bioinformatics [Internet] 30:1312–1313. Available from:
564 http://dx.doi.org/10.1093/bioinformatics/btu033

565 Villandre L, Stephens DA, Labbe A, Günthard HF, Kouyos R, Stadler T, Study TSHIVC.
566 2016. Assessment of Overlap of Phylogenetic Transmission Clusters and Communities in
567 Simple Sexual Contact Networks: Applications to HIV-1. PLoS One [Internet] 11:e0148459.
568 Available from: https://doi.org/10.1371/journal.pone.0148459

569 Volk JE, Marcus JL, Phengrasamy T, Hare CB. 2015. Incident Hepatitis C Virus Infections
570 Among Users of HIV Preexposure Prophylaxis in a Clinical Practice Setting. Clin. Infect.
571 Dis. [Internet] 60:1728–1729. Available from: https://academic.oup.com/cid/article-
572 lookup/doi/10.1093/cid/civ129

573 Le Vu S, Ratmann O, Delpech V, Brown AE, Gill ON, Tostevin A, Fraser C, Volz EM.
574 2018. Comparison of cluster-based and source-attribution methods for estimating
575 transmission risk using large HIV sequence databases. Epidemics [Internet] 23:1–10.
576 Available from:
577 https://www.sciencedirect.com/science/article/pii/S1755436517301159#bib0030

578 Ypma RJF, van Ballegooijen WM, Wallinga J. 2013. Relating phylogenetic trees to
579 transmission trees of infectious disease outbreaks. Genetics [Internet] 195:1055–1062.
580 Available from: http://www.ncbi.nlm.nih.gov/pubmed/24037268

581