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. In simulated phylogenies, Phydelity achieves higher rates of correspondence to ground-truth clusters than current model-based methods, and comparable results to parametric methods without the need for parameter calibration.

Availability and implementation: Phydelity is available at http://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; de Oliveira *et al.*, 2017). Current phylogenetic approaches for inferring transmission clusters (primarily 'cutpoint-based' methods) are centrally limited by the need to define arbitrary, absolute cluster divergence thresholds (Ragonnet-Cronin *et al.*, 2013; Prosperi *et al.*, 2011). 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.

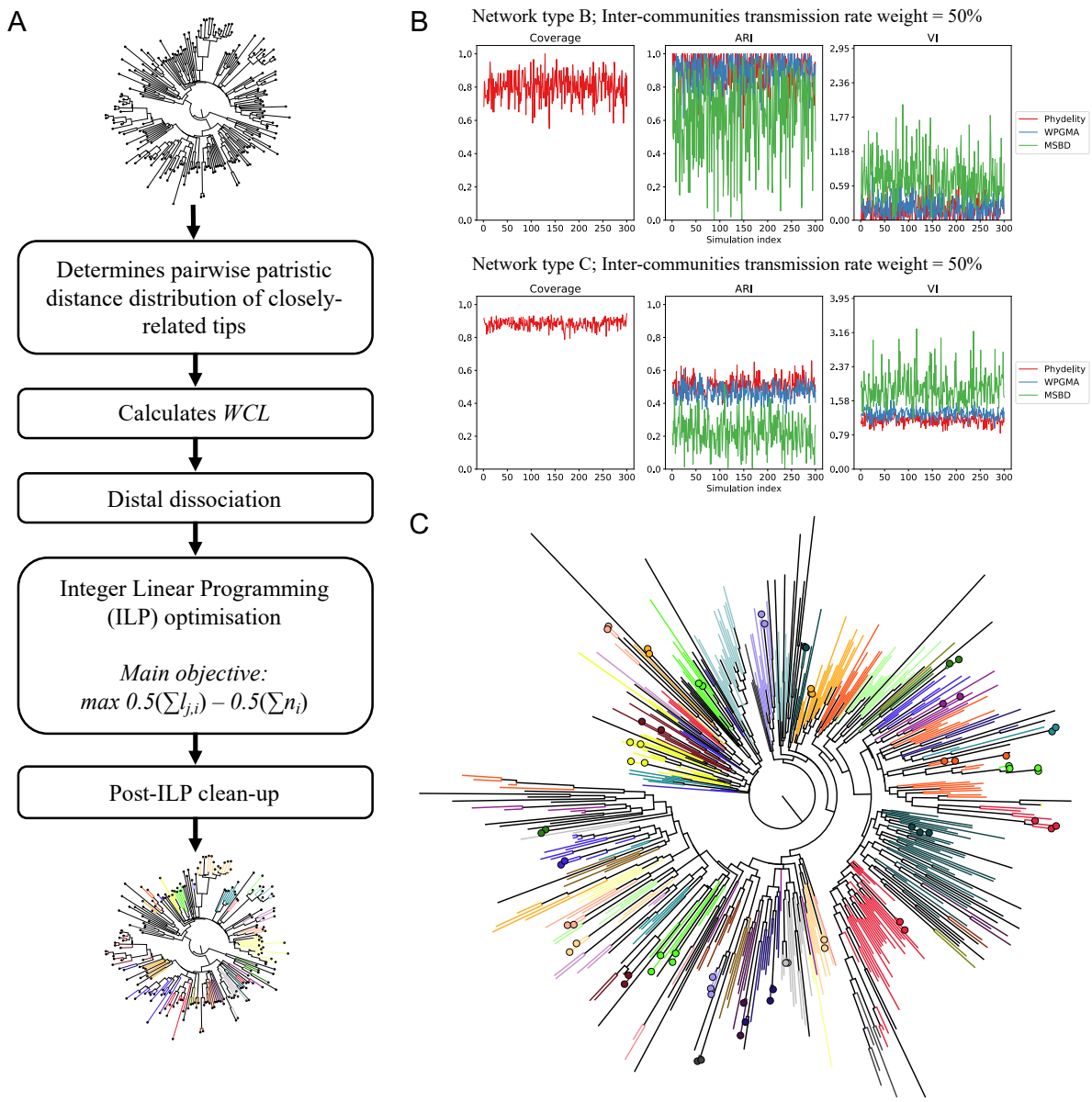

Figure 1. (A) Phydelity algorithm pipeline. (B). Clustering correspondence metrics for clustering algorithms (Phydelity, WPGMA and MSBD) applied to phylogenies generated from simulations of HIV epidemics of hypothetical MSM sexual contact network types B and C, with inter-communities transmission rates weighted at 50% of within-community rate (Villandre *et al.*, 2016). (C) Clustering results of Phydelity on HIV-1 subtype A *env* sequences collected from the Rakai Community Cohort Study (Grabowski *et al.*, 2014). Tips are coloured by Phydelity clusters and those marked with ● were phylogenetic clusters identified by Grabowski *et al.*

Method

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). Phydelity then determines the pairwise patristic distance distribution of closely-related tips, which comprises the pairwise distances of sequence *j* and its closest *k*-neighbouring tips, where the closest *k*-neighbours includes sequence *j*. The user can input the desired *k* parameter or Phydelity can automatically scale *k*

to the value that yields the supremum distribution with the lowest overall divergence. All tests of Phydelity presented in this work were performed using the autoscaled value of *k*.

Regardless of how the distance distribution of closely-related tips is determined, Phydelity uses this distribution to calculate the within-cluster divergence limit (WCL), an upper bound to μ_i of putative clusters:

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

where $\bar{\mu}$ is the median pairwise distance of the closely-related tips distance distribution and σ is the corresponding robust estimator of scale without assuming symmetry about $\bar{\mu}$.

This is followed by distal dissociation of distantly-related descendant subtrees/sequences to any ancestral node with $\mu_i > WCL$, thereby facilitating identification of both monophyletic as well as nested, paraphyletic clusters (Han *et al.*, 2018). Phydelity filters outlying tips from putative clusters under the assumption that viruses infecting individuals in a quick transmission chain are ultimately descended from the same source and are highly similar genetically. An outliers is defined by a node-to-tip distance more than three deviations from the median distance. An integer linear programming model is implemented and optimised under a blended objective of equal weights to maximise the number of sequences clustered within the lowest number of clusters. Lastly, clean-up steps are taken to remove any topologically outlying singletons that were spuriously clustered as described in Han *et al.* (2018). The full algorithm description and mathematical formulation of Phydelity is detailed in Supplementary Materials.

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.

Results

Phydelity was evaluated on phylogenetic trees derived from simulated HIV epidemics of two hypothetical MSM sexual contact network types (network types B & C) produced by Villandre *et al.* (2016), wherein quick transmission chains (i.e. transmission clusters) could be attributed to sexual contact among individuals belonging to the same community. Network type B, which corresponded best with the assumption of monophyletic clusters, consisted of a

main contact network of 60 individuals with single linkages to 25 disjoint subnetworks of 20 subjects. Conversely, the more realistic network type C included 100 communities of sizes sampled from an empirical distribution obtained from the Swiss HIV Cohort Study. For both network types, inter-community transmission rates were weighted at 25%, 50%, 75% or 100% of the within-community rate. These simulated datatsets were also tested by Barido-Sottani *et al* using their multi-state birth-death (MSBD) method which infers transmission clusters by detecting significant changes in transmission rates (Barido-Sottani *et al.*, 2018). More information on the simulated epidemics are included in Supplementary Materials.

Clustering results from Phydelity were compared to those generated by the MSBD method and a cutpoint method based on the weighted pair-group method of analysis (WPGMA) (Villandre *et al.*, 2016) (Figure 1B, Supplementary Fig. 1 and Supplementary Table 1). Both the adjusted rand index (ARI) and variation of information (VI) were calculated to quantify the correspondence between the actual network communities and clustering results. Villandre *et al.* (2016) assessed four different commonly used cutpoint-based methods, including arbitrarily varying patristic distance thresholds between any two tips (Brenner *et al.*, 2007), ClusterPicker (varying standardised number of nucleotide changes; Ragonnet-Cronin *et al.*, 2013), PhyloPart (changing arbitrary percentile of pairwise patristic distance distribution; Prosperi *et al.*, 2011) and agglomerative hierarchical clustering methods such as the WPGMA method. WPGMA methods achieved the best overall ARI. As such, only WPGMA clustering results derived from the optimal cutpoint parameter (i.e. maximum ARI) were compared.

Owing to Phydelity's definition of a closely-related neighbourhood, its distal dissociation approach and outlier detection, its mean coverage of sequences clustered ranges from 70.8–80.0% for B networks and 87.5–87.9% for C networks. Phydelity (mean ARI = 0.90-0.91, mean VI = 0.17-0.18) consistently performed as well as optimised WPGMA (mean ARI = 0.88-0.96, mean VI = 0.09-0.25) for B networks. Phydelity was the best performing method for C networks (Phydelity: mean ARI = 0.49-0.59, mean VI = 0.95-1.18; WPGMA: mean ARI = 0.44-0.56, mean VI = 1.07-1.33; MSBD: mean ARI = 0.19-0.22, mean VI = 1.80-2.02; Supplementary Table 1). Notably, even though results generated by WPGMA are comparable to those from Phydelity, this was only possible for WPGMA when the optimal cutpoint could be determined by calibration with the simulated ground-truth. However, ground truth

clustering is largely unavailable in epidemiological studies. Phydelity, on the other hand, does not require this calibration step.

Phydelity was also tested on an empirical dataset of HIV-1 subtype A *env* sequences obtained from the Rakai Community Cohort Study (Grabowski *et al.*, 2014). Grabowski *et al.* identified 35 clusters by phylogenetic analysis, of which 18 constituted individuals of the same community and 12 clusters were confirmed to be from the same household. As community and household information was blinded for privacy reasons, the available sequence data could not be matched for the exact clusters identified by Grabowski *et al*. However, the stringent cluster definitions used by Grabowski et al. (≥90% bootstrap support (1000 iterations), median genetic pairwise distance ≤ 2.6%) restricts most of these clusters to pairs, with non-pair clusters made up of no more than 5 individuals. As such, we found the same number of clusters of similar size distribution through visual inspection when we recapitulated the phylogeny using the same methods (GTR+I+G substitution model, Garli) and cluster definition as Grabowski et al. Phydelity identified 33 out of the 35 epidemiologically-identified clusters as distinct transmission clusters (Figure 1C).

Conclusion

Phydelity is a statistically-principled and phylogeny-informed tool capable of identifying putative transmission clusters in pathogen phylogenies without the introduction of arbitrary distance thresholds. It is fast, generalizable, 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.

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. *et al.* (2012) Impact of highly active antiretroviral therapy on the molecular epidemiology of newly diagnosed HIV infections. *AIDS*, **26**.

Barido-Sottani,J. *et al.* (2018) Detection of HIV transmission clusters from phylogenetic trees using a multi-state birth–death model. *J. R. Soc. Interface*, **15**.

Bezemer,D. *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.*, **12**, e1001898.

Brenner,B.G. *et al.* (2007) High Rates of Forward Transmission Events after Acute/Early HIV-1 Infection. *J. Infect. Dis.*, **195**, 951–959.

Gardy,J.L. and Loman,N.J. (2017) Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat. Rev. Genet.*, **19**, 9–20.

Grabowski,M.K. *et al.* (2014) The Role of Viral Introductions in Sustaining Community-Based HIV Epidemics in Rural Uganda: Evidence from Spatial Clustering, Phylogenetics, and Egocentric Transmission Models. *PLOS Med.*, **11**, e1001610.

Grabowski,M.K. and Redd,A.D. (2014) Molecular tools for studying HIV transmission in sexual networks. *Curr. Opin. HIV AIDS*, **9**.

Han,A.X. *et al.* (2018) Phylogenetic Clustering by Linear Integer Programming (PhyCLIP). *bioRxiv*.

de Oliveira,T. *et al.* (2017) Transmission networks and risk of HIV infection in KwaZulu-Natal, South Africa: a community-wide phylogenetic study. *Lancet HIV*, **4**, e41–e50.

Prosperi,M.C.F. *et al.* (2011) A novel methodology for large-scale phylogeny partition. *Nat. Commun.*, **2**, 321.

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

Villandre,L. *et al.* (2016) Assessment of Overlap of Phylogenetic Transmission Clusters and Communities in Simple Sexual Contact Networks: Applications to HIV-1. *PLoS One*, **11**, e0148459.