Supplementary material

Beyond the SNP threshold: identifying outbreak clusters using inferred transmissions

James Stimson, Jennifer Gardy, Barun Mathema,
Valeriu Crudu, Ted Cohen and Caroline Colijn

September 23, 2018

Application of transmission method to timed trees

We can extend the transmission method by applying it to timed phylogenetic trees. Building such a tree (using, for example, Beast2 [Bouckaert et al., 2014]) allows us to consider the joint ancestry of all isolates together, in contrast to the pairwise application in the main text. Tree reconstruction algorithms account for varying mutation rates at different sites (which can be specified or estimated), incorporate evolutionary models that discriminate between transitions and transversions, account for various population models and have other flexibilities. In Bayesian tree reconstruction, the timings of the branches are obtained from the posterior, and timing and sequence data are jointly used to construct a phylogenetic tree in which the branch lengths are in units of time. An advantage of this approach is that the clock rate can be estimated from the data, rather than being a fixed assumption or a range, though naturally this requires longitudinal data and sufficient genetic variation.

To cluster isolates using a timed tree, the timed tree is subdivided into a set of sub-trees by removing internal branches that exceed the transmission cut-off. The cut-off length is obtained using Equation (8) in the Methods section of the main text, which gives the probability of transmissions based on the total time between nodes. In this case, we are considering the time between two internal nodes rather than the total time between two sampled cases (which are tips of the tree), so we replace $h + \delta$ with the branch length between any two internal nodes. Note that we do not cut terminal branches. Cutting a branch results in a sub-tree being created from the clade descended from this branch. In the original tree, the cut branch and its descendant clade are then replaced by a single terminal branch.

In Figure S1 we illustrate the application of this method for a simulated data set containing 22 samples taken over a 10 year period. Note that not all of the clusters obtained are monophyletic clades. Cluster 2 and Cluster 3 are clades; but Cluster 1 is not, Cluster 2 being its phylogenetic descendant. This is a feature which is also obtained by [Barido-Sottani et al., 2018] in the context of HIV transmission clusters, based on a multi-state birth-death model with variable transmission rates.

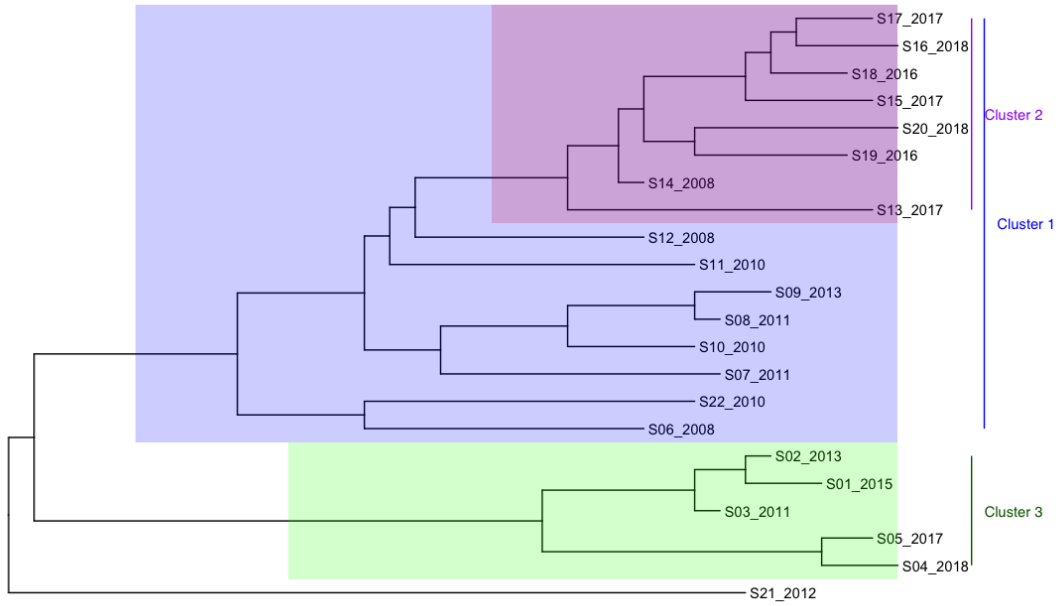

Figure S1: Application of transmission method to a simulated timed tree. Branches are cut where the shading colour changes, which is where there is a greater than 80% probability that more than 10 transmissions have occurred along that branch, with $\beta = 0.9$ transmissions/year. This occurs for any internal branch with length at or in excess of 5.2 years. Three sub-trees are created, corresponding to the identified clusters 1, 2 and 3. Tip labels are suffixed with the year of sampling.

The function *clusterTimedTree*, available in the R package `transcluster`, was used to partition the example in Figure S1.

References

[Barido-Sottani et al., 2018] Barido-Sottani, J., Vaughan, T. G., and Stadler, T. (2018). Detection of HIV transmission clusters from phylogenetic trees using a multi-state birth–death model. *Journal of The Royal Society Interface*, 15(146):20180512.

[Bouckaert et al., 2014] Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., and Drummond, A. J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Computational Biology*, 10(4):e1003537.

Supporting information for British Columbia clustering

SNP S=9		SNP S=3		Transmission T=11		Transmission T=10		Transmission T=3	
10s425	42	11s006	29	10s425	42	10s425	42	11s006	29
11s006		11s149		11s006		11s006		11s149	
11s149		11s341		11s149		11s149		11s341	
11s341		13s046		11s341		11s341		13s046	
13s046		13s158		13s046		13s046		13s158	
13s158		14s203		13s158		13s158		14s203	
13s190		14s256		13s190		13s190		14s256	
13s327		A4s027		13s327		13s327		A4s027	
14s008		A5s071		14s008		14s008		A5s071	
14s203		A5s271		14s203		14s203		A5s271	
14s256		A6s118		14s256		14s256		A6s118	
14s409		A7s017		14s409		14s409		A7s017	
A4s027		A7s155		A4s027		A4s027		A7s155	
A5s071		A7s238		A5s071		A5s071		A7s238	
A5s254		A7s312		A5s254		A5s254		A7s312	
A5s271		A8s058		A5s271		A5s271		A8s058	
A6s118		A8s136		A6s118		A6s118		A8s136	
A6s270		A8s170		A6s270		A6s270		A8s170	
A6s329		A8s314		A6s329		A6s329		A8s314	
A7s017		A8s327		A7s017		A7s017		A8s327	
A7s155		A9s194		A7s155		A7s155		A9s194	
A7s238		A9s216		A7s238		A7s238		A9s216	
A7s312		A9s249		A7s312		A7s312		A9s249	
A8s058		A9s250		A8s058		A8s058		A9s250	
A8s111		A9s287		A8s111		A8s111		A9s287	
A8s133		A9s305		A8s133		A8s133		A9s305	
A8s136		A9s320		A8s136		A8s136		A9s320	
A8s168		A9s349		A8s168		A8s168		A9s349	
A8s170		A9s391		A8s170		A8s170		A9s391	
A8s220		13s370	8	A8s220		A8s220		A8s285	4
A8s259_P2		A1s092_P1		A8s259_P2		A8s259_P2		A9s014	
A8s314		A7s160		A8s314		A8s314		A9s235	
A8s327		A7s168		A8s327		A8s327		A9s321	
A9s194		A8s285		A9s194		A9s194		A5s254	3
A9s216		A9s014		A9s216		A9s216		A6s270	
A9s249		A9s235		A9s249		A9s249		A6s329	
A9s250		A9s321		A9s250		A9s250		A7s160	2
A9s287		10s425	4	A9s287		A9s287		A7s168	
A9s305		A5s254		A9s305		A9s305		A8s111	2
A9s320		A6s270		A9s320		A9s320		A8s168	
A9s349		A6s329		A9s349		A9s349		A8s220	2
A9s391		A8s111	2	A9s391		A9s391		A8s259_P2	
13s370	8	A8s168		13s370	8	13s370	7	10s425	1
A1s092_P1		A8s220	2	A1s092_P1		A7s160		13s190	1
A7s160		A8s259_P2		A7s160		A7s168		13s327	1
A7s168		13s190	1	A7s168		A8s285		13s370	1
A8s285		13s327	1	A8s285		A9s014		14s008	1
A9s014		14s008	1	A9s014		A9s235		14s409	1
A9s235		14s409	1	A9s235		A9s321		A1s092_P1	1
A9s321		A6s169	1	A9s321		A1s092_P1	1	A6s169	1
A6s169	1	A7s286	1	A6s169	1	A6s169	1	A7s286	1
A7s286	1	A8s133	1	A7s286	1	A7s286	1	A8s133	1

Figure S2: Supporting information for British Columbia results. Data labels are shown arranged into clusters for various threshold levels for both methods, with size of cluster indicated to the right of each cluster.