Beyond the SNP threshold: identifying outbreak clusters using inferred transmissions

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

September 24, 2018

Abstract

Whole genome sequencing (WGS) is increasingly used to aid the understanding of pathogen transmission. A first step in analysing WGS data is usually to define "transmission clusters", sets of cases that are potentially linked by direct transmission. This is often done by including two cases in the same cluster if they are separated by fewer SNPs than a specified threshold. However, there is little agreement as to what an appropriate SNP threshold should be, or whether such thresholds are appropriate for defining transmission clusters. We propose a probabilistic alternative, suggesting that the key inferential target for transmission clusters is the number of transmissions separating cases. We characterize this by combining the number of SNP differences and the length of time over which those differences have accumulated. Our framework can be modified to consider variable mutation rates across the genome (e.g. SNPs associated with drug resistance) and can incorporate data such as spatial proximity. We use tuberculosis data from Canada and the Republic of Moldova to illustrate our approach, incorporating information about case timing, the molecular clock and transmission processes. Our approach can accommodate selection for antibiotic resistance. Simulation results indicate that our transmission-based method is better at identifying direct transmissions than a SNP threshold. Our results show that it is likely to outperform the SNP threshold where clock rates are variable and sample collection times are spread out. We implement the method in the R package `transcluster`.

Introduction

Whole genome sequencing (WGS) of pathogens has become an essential tool for improving understanding of how infectious diseases spread between hosts, particularly in the case of tuberculosis (TB) [Hatherell et al., 2016]. The phylogeny derived from pathogen genomic data helps us to infer likely transmission events. Typically, samples are taken from patients in the field, the date and other epidemiological data are recorded, and the pathogen's genome is sequenced. A first step is typically to assign cases to clusters; for infectious diseases, a cluster is a group of closely related infections that is usually interpreted as resulting from recent transmission [Poon, 2016]. These clusters are chosen primarily with the aim

of making meaningful subdivisions of the data, with the added benefit of making the amount of data fed into attempts to reconstruct outbreaks and to transmission inference models more tractable. However the assignation method is often somewhat *ad hoc*.

The simplest way to determine sequence relatedness is to count the number of single nucleotide polymorphisms (SNPs) that differ between two sequences. The SNP threshold approach places two individuals in the same putative transmission cluster if there are fewer than a threshold number of SNPs between their sequenced pathogen genomes. Many existing methods to identify outbreak clusters rely on SNP thresholds, as surveyed recently [Hatherell et al., 2016] in the case of TB. Similar methods are also used for other pathogens [Dallman et al., 2015, Octavia et al., 2015]. However, there is little agreement in the literature as to what such a threshold should be – see Table 1 for TB SNP thresholds used in some recent studies. The contexts in which these thresholds are applied differ from study to study, so these numbers are not always strictly comparable, but they do indicate the wide range of values that can reasonably be adopted when determining whether or not cases are closely related. By itself, the number of SNPs does not directly imply a probability of recent transmission. This is implicity recognised in some sources. For example, we have from [Walker et al., 2013]: "We predicted that the maximum number of genetic changes at 3 years would be five SNPs and at 10 years would be ten SNPs". Indeed, other studies directly question the use of SNP thresholds, such as [Guerra-Assunção et al., 2015], [Bergholz et al., 2014] in the context of food-borne pathogens, and [Azarian et al., 2016] in an analysis of the spread of MRSA. Nevertheless, the use of a single SNP threshold is often employed in practice; for example the 12 SNP threshold, used for inferring likely transmission between a pair of TB cases by Public Health England [Walker et al., 2014] amongst others, is perhaps the most common in TB.

The appropriate SNP cut-off for inferring transmission is likely to depend critically on the context. There are many sources of uncertainty. Nucleotide mutation rates vary between pathogens, can vary at different stages of infection, and are subject to the effects of selection pressure. Culture processes (e.g. liquid vs solid culture, single colony picks vs sweeps) may affect the diversity in samples that are sent for sequencing. Furthermore, the process of producing finalised SNP data from patient-derived biological samples is a multi-stage procedure where there are choices to be made - including how stringently quality filtering is applied to raw genomic data - which will in general result in different SNP differences being reported. As such, it is important that during every step of the pipeline from sampling from patients and processing the data, to building the models and drawing conclusions from them, that we are aware of sources of uncertainty and attempt to propagate this uncertainty to any conclusions. It is also important that as WGS is rolled out widely as a tool in infectious disease, we re-calibrate SNP-based methods to accommodate changes in both sequencing technologies and in the bioinformatics pipelines used to call variant SNPs. Clustering methods that use variant SNP calls exclusively will be most sensitive to such changes.

The fundamental logic behind SNP cut-offs is that it takes time to accrue genetic variation; even in organisms where the molecular clock is variable, it seems uncontroversial to assume that two isolates that differ by only a few SNPs are more likely to be a result of recent transmission than isolates that are 50 SNPs

Authors	Lower SNP Threshold	Upper SNP Threshold
[Bryant et al., 2013a]	≤ 6 (relapse)	> 1000 (re-infection)
[Clark et al., 2013]	< 50	> 50
[Guerra-Assunção et al., 2015]	≤ 10 (relapse)	> 100 (re-infection)
[Lee et al., 2015]	< 2	> 40
[Roetzer et al., 2013]	≤ 3	(not specified)
[Walker et al., 2013]	≤ 5	> 12
[Yang et al., 2017]	≤ 12	(not specified)

Table 1: SNP thresholds used in recent TB studies. The lower threshold indicates the number of SNPs below which cases are positively identified as belonging to the same cluster. Where different, the upper threshold indicates the the number of SNPs above which cases are identified as clearly not belonging together. Unless otherwise stated, intermediate values are indeterminate.

apart. However, the rate at which polymorphisms occur varies not only between organisms [Kuo and Ochman, 2009], but also across a genome; it is affected by selection pressure and by horizontal gene transfer (HGT) [Novichkov et al., 2004] (though this is not an issue for TB and there are methods to remove recombination and HGT prior to using SNP cut-offs). As per [Barrick and Lenski, 2013], it is also important to distinguish between the *mutation rate*, the rate at which spontaneous mutations occur, and the *substitution rate*, the rate of accumulation of changes in a lineage; this depends on both the mutation rate and the effects of selection. Here, when we refer to the clock rate, we mean the substitution rate, as we use the rate to interpret variants measured with sequencing technologies.

This distinction is particularly important for diseases like TB, where selection pressure due to antibiotics can be substantial. Whilst the background SNP accumulation rate for *Mycobacterium tuberculosis* has been estimated at 0.5 SNPs/genome/year [Walker et al., 2013], selection pressure and antibiotic resistance can influence this rate considerably. For example, in [Eldholm et al., 2014] we see the observation that "After exclusion of transient mutations in the patient isolates, 4.3 mutations were acquired per year ... or 2.3 mutations per year when excluding resistance mutations." The size of the population of bacteria within a host could also affect the number of SNPs observed between that host and those they infect. Unexplained larger variation is also encountered as documented in [Korhonen et al., 2016], though high SNP numbers could be a result of re-infection or mixed infection rather than in-host evolution. Where we know that selection or high substitution rates are likely to be present and detected, a higher rate is therefore likely to be appropriate for clustering, and this will affect the relationship between SNPs and transmission events.

It should be noted that there are other approaches to clustering, based on molecular (but not WGS) data and including time and geographical data. For example, Kammerer [Kammerer et al., 2013] applies three different statistical tools to spoligotype data and mycobacterial interspersed repetitive units (MIRU) data, together with date and location of cases, to show that these tools can successfully identify TB outbreaks. Donker [Donker et al., 2016] uses variable number tandem repeat (VNTR) data for meticillin-resistant Staphylococcus aureus (MRSA) cases, with time and location data, to identify clusters based on a hierarchical cluster-

ing method [Ypma et al., 2013]. There are also software packages such as *vimes* [Jombart and Cori, 2017], which provides tools allowing users to integrate different types of data and detect outbreaks. These approaches treat each of the underlying variables as independent inputs, without explicitly modelling the connection between time and the accumulation of genetic differences.

We take a slightly different tack here, and jointly use the sample time and genetic distance, together with a model of SNP acquisition over time and transmission events over time, to base putative transmission clusters on the probability that cases are separated by a threshold number of transmission events. This is motivated by a belief that the number of transmission events between two cases is a natural and intuitive measure of how "clustered" they are in the sense of transmission (and how likely they are to be part of the same outbreak). This cannot usually be measured directly and must be inferred from other data. However, it is reasonable to assume that appropriate incorporation of the time over which the accumulation of SNPs occurs, as well as the likely time between transmission events, give a more accurate and nuanced measure of the likelihood that cases are linked by a small number of transmission events. We develop a probabilistic approach which permits variation in the SNP accumulation process, allows for faster SNP accumulation for sites under selection and allows for variation in the speed with which individuals infect their contacts. We aim to provide a principled alternative to SNP cut-offs for clustering pathogen genomes into putative transmission clusters.

New Approaches

Two samples are usually considered to be in the same transmission cluster if the number of SNPs between them is less than or equal to a fixed cut-off, or threshold. This is a quick way to explore relatedness among a group of isolates and gain an approximate understanding of the extent of recent (low-distance) transmission, but it is coarse and embeds a number of strong assumptions.

Our proposed probabilistic transmission approach, in contrast, is based on sample pairs being clustered together if we estimate that there were fewer than a threshold number of transmission events between them, with a given probability. It uses the same genetic (SNP) distance information as the SNP method, but in addition makes use of the sample times, knowledge of the SNP accumulation and transmission processes, and can readily be extended to incorporate other factors: we show in Methods how this can be done for spatial data.

We start by establishing probability distributions for the length of time to the most recent common ancestor of a pair of samples; this depends on the clock process, and helps define the distance between the two samples. We then compute the probability that at least a threshold number of transmissions took place between the two sampled cases. This approach gives the flexibility to incorporate sample time information and other data. The method uses sample dates and aligned sequence data (variant calls) together with models of the clock and transmission processes. For a pair of samples, we use the time difference between their sampling dates and the clock process to write down the probability distribution for when the MRCA of the two sequences existed; this must be before the first

sampled case. Integrating over this unknown time, we can find the probability that a certain number of transmissions separate the two cases.

Results

We illustrate how the transmission method compares to the SNP method for a simple toy example. We define the "T cut-off" as the cut-off level for the transmission method, using Equation (10); the samples are clustered together where the implied number of transmissions k is less than or equal to T with a probability of 80%, given some clock rate λ and transmission rate β. We see that the transmission method clusters the cases together in a different order to the SNP method as the cut-off level is incremented. Cases A and B are the closest in SNP distance, but the time elapsed between their sampling dates increases their distance by the transmission distance function relative to cases C and D, which are sampled at the same time as each other. So when we take timing into account, the clustering is altered (also illustrated in Figure 1).

We model the number of intermediate transmissions between two sampled hosts given the total time over which SNPs have likely accumulated. Altering the transmission rate β (by which we mean the rate at which intermediate cases occur in the total time elapsed *between the MRCA of two sampled hosts and the sampling events*; see Methods) alters the absolute transmission cut-off level at which the clusters change – in this example, increasing β to 3.0 transmissions/year gives the same clusters as in Figure 1 but at levels 9, 10 and 12 transmissions rather than 7, 8 and 9 transmissions respectively. This has no impact on the order of the clustering as the level of the cut-off changes.

Label	Sample Date	SNP dist. to A	SNP dist. to B	SNP dist. to C	SNP dist. to D
A	1/1/2018	0	5	7	7
B	1/1/2014	5	0	8	8
C	1/1/2016	7	8	0	6
D	1/1/2016	7	8	6	0

Table 2: Model inputs for toy example data set.

By contrast, altering the clock rate does have a material impact on the way clustering occurs as we increase the transmission threshold. For $\lambda = 0.5$ SNPs/genome/year, cases A and B are closest under the transmission method, just as they are with the SNP method, and so the clustering is the same for both methods. At $\lambda = 1.5$ SNPs/genome/year, cases C and D are closest under the transmission method, and the clustering evolves as shown in Figure 1.

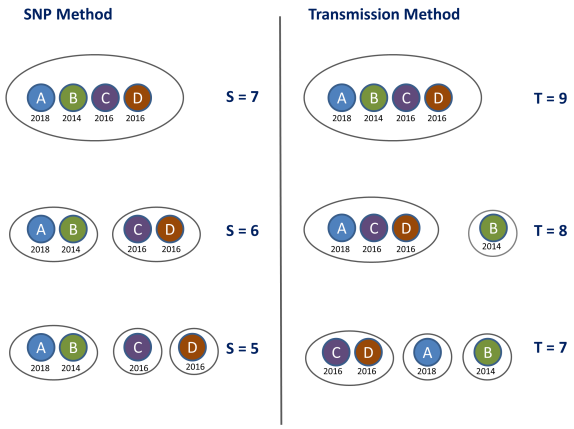

Figure 1: Clustering on the toy example data set provided in Table 2. The left hand panel shows the clusters obtained by applying the SNP method with three different thresholds, with the cut-off level denoted by S; samples are clustered together where the SNP distance is less than or equal to S. The right hand panel shows the clustering obtained by applying the transmission method, using Equation (10), with the cut-off level denoted by T; samples are clustered together where the implied number of transmissions k is less than or equal to T with a probability of 80%, with clock rate $\lambda = 1.5$ SNPs/genome/year and $\beta = 2.3$ transmissions/year.

British Columbia data

We analyse a data set from British Columbia, comparing the SNP method to the transmission method in Equation (10). The dataset comprises 52 samples collected from 51 patients over a 14 year period, and has been pre-filtered with the result that all samples are relatively close - within 25 SNPs. Consequently, using the SNP method with the threshold set to 13 SNPs or higher, all samples are placed in one cluster. When the threshold is 9 SNPs, we obtain a large 42-case cluster, a secondary 8-case cluster and some outliers. As we reduce the threshold further down to 3, the large cluster breaks up but the 8-case cluster persists. We illustrate this in Figure 2.

We used the transmission method, using Equation (10) with $\beta = 2.0$ transmissions/year and two different average clock rates: $\lambda = 0.5$ and 1.5 SNPs/genome/year (1.5 is larger than the typical rate for TB but within other outbreak estimates [Bryant et al., 2013b]). When λ is low we can obtain the same clustering as with the SNP cut-off. When λ is higher, we have one cluster which contains all the samples for $T > 11$. As with the SNP method, at $T = 11$ we have a large 42 case cluster, a secondary 8-case cluster and some outliers. But as we move to $T = 10$, the secondary cluster loses a member, whilst the main cluster stays at size 42. This is because one of the members of the secondary group is very close by the SNP distance to another member of that group, but was sampled more than 10 years before. As with our simple toy example, timing alters the effective distance between samples, and therefore alters the clustering.

Furthermore, the probabilistic nature of the approach means that we can see how strongly we predict cases to be linked; in Figure 3 we use thicker edges to denote a higher probability of being linked by relatively few transmissions. In ad-

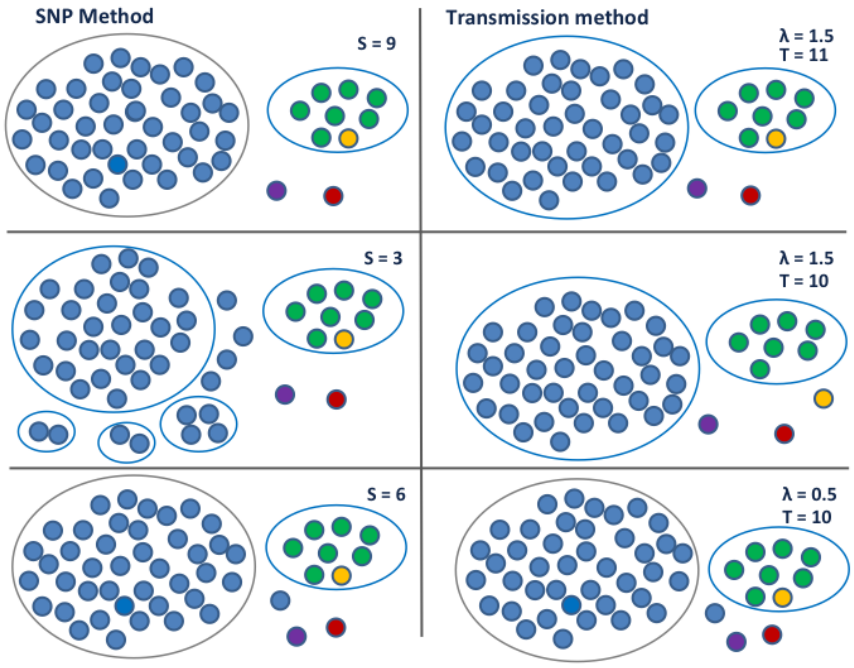

Figure 2: Clustering on the British Columbia data set. The left hand side shows the clusters obtained by applying the SNP method with three different thresholds, with the cut-off level denoted by S. The largest cluster breaks up as the level is lowered whilst the size 8 cluster remains intact. The right hand side shows the clustering obtained by applying the transmission method, using Equation (10); samples are clustered together where the implied number of transmissions k is less than or equal to T with a probability of 80%. As shown in the top two thirds, with clock rate $\lambda = 1.5$ SNPs/genome/year and $\beta = 2.0$ transmissions/year, the size 8 cluster loses a member whilst the largest cluster stays the same as the level is lowered. When λ is low, h is larger, so the MRCA of a cluster gets pushed back further in time. In this case, the value of δ between two cases has a limited impact on the estimated number of transmissions; the SNP difference is dominant, and we recover the same clustering that is obtained with the SNP cut-off. This is shown in the lower third, where the clock rate $\lambda = 0.5$ SNPs/genome/year and $\beta = 1.2$ transmissions/year.

dition, we show the effect of incorporating spatial proximity, using Equation (19); we assign each of the cases into one of six numbered regions. Including a spatial weighting, reflecting the barrier to the infection moving between different regions, and leaving all other parameters unaltered, changes the clustering that is obtained.

Sensitivity to clock rate

An implicit assumption of the SNP method is that each SNP contributes equally towards the SNP distance. This implies that the clock rate or substitution process is constant across the set of isolates and across the genome. When the same threshold is used in different settings and across different pathogen subtypes, the implicit assumption is that the same substitution process holds in these settings. In the transmission method, the effective distance between any two samples is inversely proportional to the assumed mean clock rate. A lower clock rate means

that more time is needed in order for a fixed number of SNPs to be generated; this gives room for more potential intermediate transmission events. A higher clock rate means that the fixed time between samples has a greater effect on the clustering, as the time between samples places a greater constraint on the range of possible heights h; the fixed time "uses up" more of the time available than it would under a low clock rate (because there is less total estimated time available, a higher portion of it is in the time period δ). We show this in Table 3: the transmission clustering method approaches the same results as the SNP clustering method as the assumed clock rate is reduced. We use the variation of information dissimilarity measure given by *clue* [Meilă, 2007] to compare the clusters produced by the two methods.

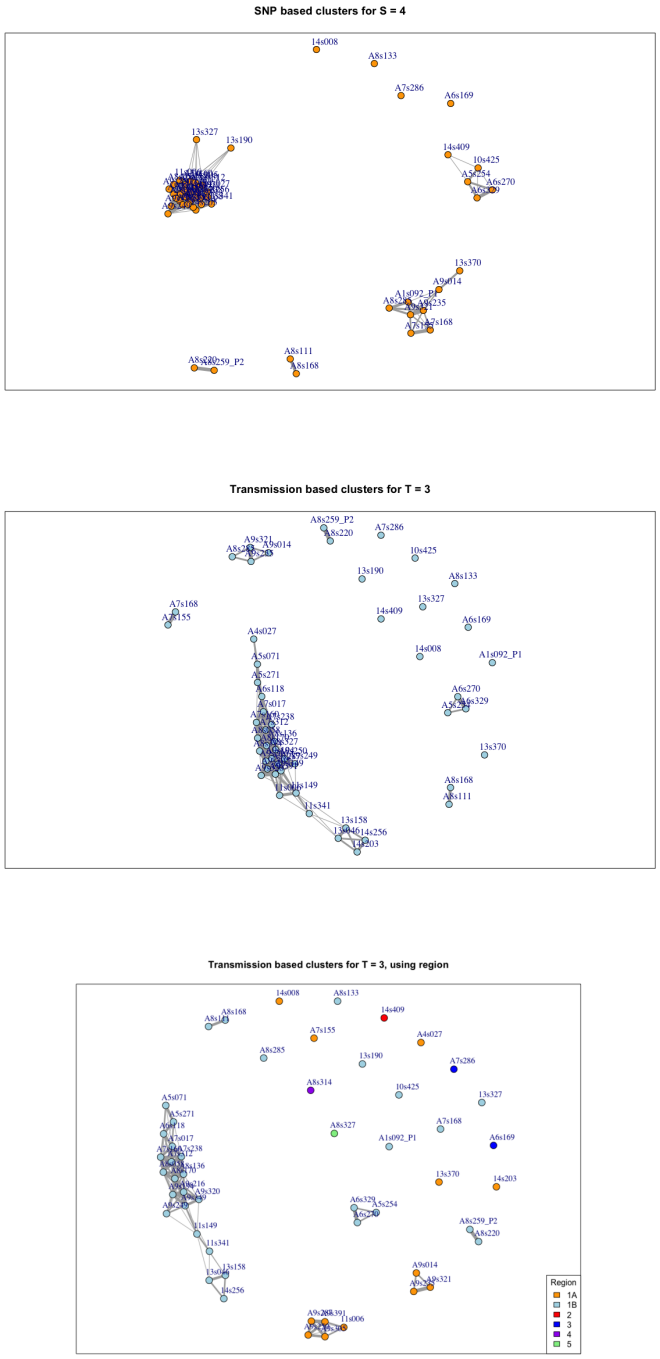

Figure 3: Three views of the same British Columbia TB data illustrating the contrasting effect of implementing the SNP and transmission methods and showing estimates of how close individual cases are to each other. In the top figure, edges between nodes indicate that cases are within 4 SNPs of each other. In the lower figures, edges indicate that cases are 80% likely to be within 3 transmission events of each other, given a clock rate $\lambda = 1.5$ SNPs/genome/year and $\beta = 2.0$ transmissions/year. The middle figure is based on Equation (10), and the bottom figure uses Equation (19), with weighting $w = 20\%$ where two cases are assigned to differing regions. The thicker the edges, the closer the cases are: for the SNP based clusters the thickest edges correspond to no SNP difference, the thinnest to a distance 4 SNPs; for the transmission based clusters the thickest edges correspond to one likely transmission event, the thinnest to 3.

SNP Threshold	Largest cluster	λ	β	Trans. Threshold	Largest cluster	Dissimilarity between SNP and Trans. methods
S=12	50	0.5	1.2	T=22	50	0.000
S=9,10,11	42	0.5	1.2	T=16	42	0.000
S=6,7,8	41	0.5	1.2	T=10	41	0.000
S=5	37	0.5	1.2	T=8	37	0.000
S=4	31	0.5	1.2	T=7	31	0.139
S=3	29	0.5	1.2	T=6	29	0.052
S=2	28	0.5	1.2	T=3	28	0.113
S=12	50	1.0	1.2	T=11	50	0.000
S=9,10,11	42	1.0	1.2	T=8	42	0.000
S=6,7,8	41	1.0	1.2	T=5	41	0.058
S=5	37	1.0	1.2	T=4	37	0.113
S=4	31	1.0	1.2	T=3	29	0.374
S=3	29	1.0	1.2	T=3	29	0.174
S=2	28	1.0	1.2	T=1	28	0.113
S=12	50	1.5	1.2	T=7	52	0.163
S=9,10,11	42	1.5	1.2	T=6	42	0.000
S=6,7,8	41	1.5	1.2	T=4	41	0.058
S=5	37	1.5	1.2	T=2	35	0.289
S=4	31	1.5	1.2	T=2	35	0.400
S=3	29	1.5	1.2	T=1	29	0.240
S=2	28	1.5	1.2	T=1	29	0.290
S=12	50	2.0	1.2	T=6	52	0.163
S=9,10,11	42	2.0	1.2	T=4	42	0.058
S=6,7,8	41	2.0	1.2	T=4	42	0.149
S=5	37	2.0	1.2	T=2	37	0.412
S=4	31	2.0	1.2	T=1	29	0.447
S=3	29	2.0	1.2	T=1	29	0.230
S=2	28	2.0	1.2	T=1	29	0.240

Table 3: Effect of varying the clock rate using the British Columbia data. This table shows how the clock rate affects the transmission method, using Equation (10), keeping the transmission rate β constant. For the SNP method, samples are clustered together where the SNP distance is less than or equal to S. For the transmission method, samples are clustered together where the implied number of transmissions k is less than or equal to T with a probability of 80%. For a clock rate of 0.5 SNPs/genome/year, the transmission method matches all SNP-based clusters for thresholds of 5 SNPs and above. As the clock rate increases, the transmission clustering diverges further from the SNP clustering. As we vary the other parameters, the choice of β is effectively a scale factor and does not affect the pattern of clustering.

SNP Threshold	Largest cluster	λ	β	Trans. Threshold	Largest cluster	Dissimilarity between SNP and Trans. methods
S=12	106	1.0	1.2	T=11	106	0.000
S=10	93	1.0	1.2	T=9	93	0.000
S=8	78	1.0	1.2	T=7	78	0.000
S=7	57	1.0	1.2	T=6	57	0.000
S=6	48	1.0	1.2	T=5	48	0.000
S=5	32	1.0	1.2	T=4	32	0.000
S=4	15	1.0	1.2	T=3	15	0.000
S=11	103	2.0	1.2	T=4	103	0.000
S=9	86	2.0	1.2	T=3	86	0.000
S=7	57	2.0	1.2	T=2	57	0.000
S=4	15	2.0	1.2	T=1	15	0.000
S=10	93	3.0	1.2	T=2	93	0.000
S=7	57	3.0	1.2	T=1	57	0.000

Table 4: Comparison of methods with the Moldova data. This table shows how the clock rate affects the transmission method, using Equation (10), keeping the transmission rate β constant. For the SNP method, samples are clustered together where the SNP distance is less than or equal to S. For the transmission method, samples are clustered together where the implied number of transmissions k is less than or equal to T with a probability of 80%. For $\lambda = 1.0$, 2.0 and 3.0 SNPs/genome/year, the pattern of clusters is identical using either method across a range of thresholds.

Moldova data

This data set comprises 422 samples collected over a period of less than 2 years. For this data – with any reasonable choices of parameters and a fixed substitution rate for all sites – as shown in Table 4, the new transmission method does not differ from the SNP method. This can be explained by two factors that work together: the small distance in time between any two samples and the large SNP differences between cases. There isn't enough variation in the timing information relative to the SNP distances for an appreciable difference to emerge between the two clustering methods.

Use of drug resistance conferring SNPs

We can, however, explore the role drug resistance conferring SNPs on the clustering. Information on the location of resistance conferring sites for TB was obtained using PhyResSE [Feuerriegel et al., 2015] and a resistance conferring SNP distance matrix was computed for the Moldova data by filtering against this information. Selection is likely to lead to resistance conferring SNPs arising more quickly than other SNPs: for example, one TB study [Eldholm et al., 2014] gives a mutation rate of 4.3 SNPs per genome per year when they are included, in contrast to the 0.5 SNPs per genome per year that is typically estimated for TB [Walker et al., 2013]. Resistance acquisition may further increase the rate of acquisition of additional SNPs through multiple resistance, compensatory mutations or other mechanisms. For this analysis we used a clock rate for the drug-resistant

sites, as in equation (18), which is five times higher than for the non-resistance conferring sites.

Overall, resistance conferring SNPs in the Moldova data set form only 0.6% of the total number of SNPs. However, restricting to those sample pairs where the SNP distance is less than or equal to 20, they form 8% of the total. If a high proportion of the SNPs between two cases are resistance conferring SNPs, then this effectively shortens the distance between the cases, making them more likely to be joined together in a transmission cluster. For several sample pairs in this data set, the proportion of resistance conferring SNPs that differ between the two samples is approaching 35%, whilst for some other pairs there are none at all. For this reason we see a difference when we take resistance into consideration, as seen in Figure 4. The largest cluster is not shown in detail in the figure and is more robust with respect to the effect of resistance conferring SNPs than the smaller clusters.

Simulated data

To explore the performance of the clustering methods in a setting where the "ground truth" is known, we simulate data and compare the SNP and transmission (as given by Equation (10)) methods. The "true" clusters are generated from simulated transmission networks produced by *TransPhylo*.

We consider clustering cases based on direct transmission, so that two cases are joined in a cluster if one infected the other, and we compare clusters generated by the SNP method with those generated by the transmission method. In order to compare to the appropriate set of clusters, we find the best match that the method achieves against the true cluster over an appropriately wide range of threshold levels. Then we use the variation of information dissimilarity measure given by *clue* [Meilă, 2007] to compare the results of the two methods to the true clusters. We also compare randomly permuted simulated data to the simulated clusters to provide a yardstick of accuracy. This is achieved by fixing the number of clusters to be the number of the true clusters, and then randomly allocating each sample case to one one of those clusters. The results in Table 5 show that the transmission method is consistently better than the SNP method at identifying direct transmissions within an outbreak. Both methods perform significantly better than the randomly generated data.

Identifying direct transmissions is not the aim of either the SNP cut-off or transmission clustering method; rather, both aim to simply group cases into sets of isolates for onward, more intensive (model-specific, Bayesian for example) outbreak reconstructions. Testing the ability of SNP vs transmission-based methods to accomplish this using simulated data would require an appropriate simulation set-up, which in turn would have a lot of flexibility (and could no doubt be tweaked to ensure that the transmission method performs well, or that the SNP cut-off does). For example, one approach would to simulate the introduction of new cases at that are least 25 SNPs apart from cases in an existing outbreak. The SNP method with a threshold of 12 SNPs will always correctly place such new introductions in a new cluster, and will group their descending infections correctly until one or more of them is more than 12 SNPs away from other sampled cases in the cluster. Conversely, if new introductions were only 12 SNPs from exist-

Clock Rate	Transmission method	SNP method	Random
0.5	0.2501	0.3655	0.8708
1.0	0.2613	0.3655	0.8450
1.5	0.2647	0.3655	0.7951
2.0	0.2821	0.3655	0.8416
2.5	0.2928	0.3655	0.8451
3.0	0.3030	0.3655	0.8476

Table 5: Dissimilarity measure comparing both the SNP and transmission methods against simulated data, averaged over the full set of simulations. Lower numbers indicate sets of clusters that are more similar to the true clusters. An outbreak was simulated 100 times with 10 sampled cases from a total of between 20 and 30 cases, depending on the simulation. The measure is obtained by comparing clusters from a range of thresholds to the known clusters, and picking the one with the lowest score. The clock rate is the rate used by the transmission method to relate the number of SNPs to the time distribution, and thus does not affect the SNP method results. The table shows how the dissimilarity varies as the clock rate varies, for a fixed $\beta = 3.0$ transmissions/year, as compared to simulated samples connected by direct transmission. The random column shows the dissimilarity obtained for randomly allocated simulated clusters.

ing cases the SNP method would mis-classify them as linked to existing clusters. In the transmission method, we can compute the probability that a newly introduced case that is 25 SNPs from existing cases will fall within a certain number of transmission events. This gives us the probability that we would infer an incorrect link to an existing cluster. With $\lambda = 1.2$ SNPs/genome/year and $\beta = 1.5$ transmissions/year, the probability that there are more than 10 transmissions for cases 25 SNPs apart is 99.9%. This falls to 98.3% for more than 15 transmissions. Accordingly, the simulation approach for introducing new clusters will greatly affect the performance of both the SNP and transmission-based methods, and so we have not chosen to perform extensive simulations to compare the methods.

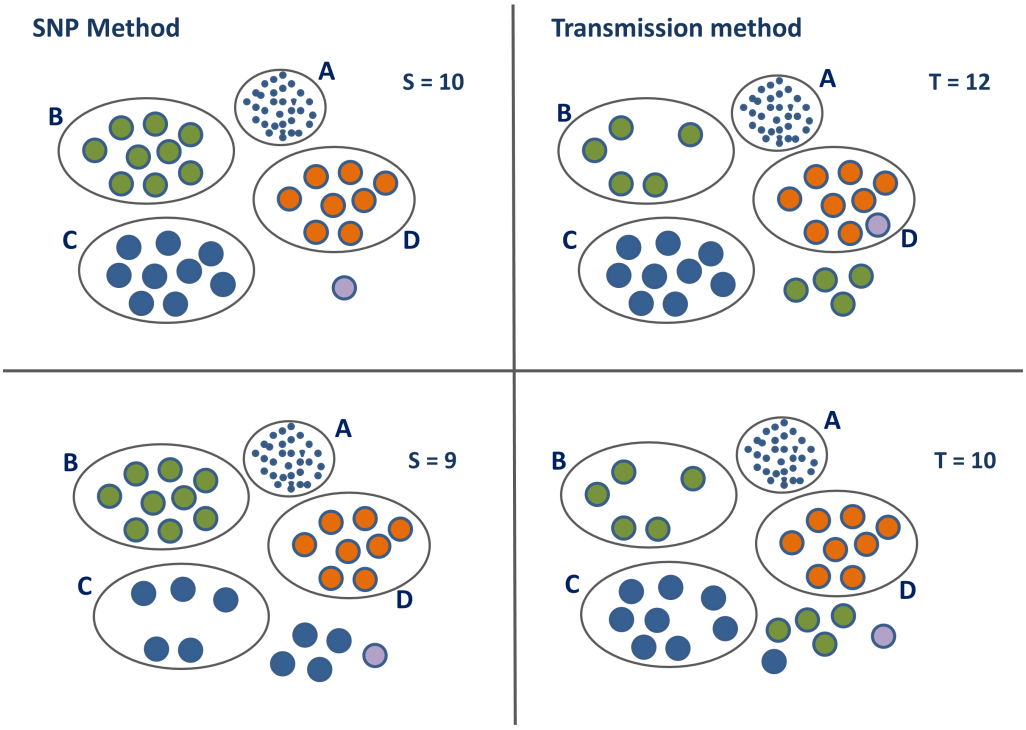

Figure 4: Clusters in the Moldova data set illustrating the effect of accounting for resistance conferring SNPs in the transmission method, using Equation (18). Clusters B, C and D are the second to fourth largest clusters in the Moldova data using the SNP method. The largest cluster A, with 93 members for $S = 10$ is shown for completeness. Isolated cases are shown with no enclosing oval. Colours are chosen to enable identification of the same cases in the four different scenarios. The left hand panel shows the clusters obtained by applying the SNP method with two different thresholds, with the cut-off level denoted by S; samples are clustered together where the SNP distance is less than or equal to S. The right hand panel shows the clustering obtained by applying the transmission method, using Equation (10), with the cut-off level denoted by T; samples are clustered together where the implied number of transmissions k is less than or equal to T with a probability of 80%, with clock rate $\lambda = 1.5$ SNPs/genome/year and $\beta = 2.0$ transmissions/year.

Discussion

A fixed number of SNPs can arise from different numbers of transmissions depending on other factors, including the timing of transmission, selection for resistance, the substitution process, location and factors we have not explicitly modelled (social contacts, host risk factors, pathogen factors). We have seen that sampled cases which are relatively close in genetic distance can nevertheless be separated by large distances in time. In this scenario, a simple SNP cut-off may place samples too close together for outbreak clustering purposes. In contrast, our new method is robust with respect to outlying cases which have been sampled at very different times compared to the majority of cases. These cases can make inference of timed phylogenetic trees challenging because the low genetic variation is hard to reconcile with the large time distance. Furthermore, true transmission clusters need not be clades in phylogenetic trees, because one cluster could descend from another but be separated by a long time or a large genetic distance (due to sampling effects). Accordingly, the clusters obtained by our method do not necessarily correspond to phylogenetic clades.

Our probabilistic transmission method has certain advantages. It is relatively simple, requiring only the implementation of fast-running algorithms to estimate the time distributions; the heavy machinery to run large simulation methodologies (like MCMC) is not required. The amount of information required for the model is limited and consists of as little as the SNP distances, the timing data and a knowledge about the substitution and transmission processes. Nevertheless it has the flexibility to be able to handle SNPs under selection, SNPs with a different substitution process and variability in the substitution and transmission processes, and it has the scope for extensions to include more epidemiological data. Even in data sets where there is not much timing information to work with, we have seen that the integration of information on resistance conferring sites can be used within our framework to fine tune the clustering.

There are some limitations. Prior knowledge of the substitution and transmission processes is required, and there is some uncertainty in choosing appropriate values. However, the model is typically robust with respect to changes in these variables; in particular, varying the transmission rate does not have a material impact on the clustering because a re-scaling of the cut-off will compensate. The choice of a time-varying transmission function $\beta(t)$ is, however, likely to have an impact on results. In particular we would expect a low probability of very quick transmission – as the pathogen numbers are building up in a new host – to have a significant impact, compared to the use of a constant transmission rate, as would a fast rate early diminishing to a much lower rate later. Note also that the parameter t in our model represents the total time since infection to both the sample dates: so we are not modelling the variation of transmission rates in calendar time.

In some diseases, such as TB, there is considerable variation in the latency period, during which the mutation rate may be lower than it is during active disease. This variability can be incorporated into the negative binomial model as expressed in equation (14). We do not explicitly model within-host diversity, though this is relevant to identifying direct transmission events [Worby et al., 2014, Hall et al., 2015, Hall et al., 2016, Didelot et al., 2014, Didelot et al., 2017]. Cases

of direct transmission will be clustered together with high probability in our method despite slight inaccuracy in the timing due to both branches of the pair's two-case tree spending time in the same host. Pairs of cases for which the clustering decision is ambiguous are likely to have several intermediate cases between them, with a larger tree height, and so the contribution of in-host diversity in either sampled case will be small. In-host diversity in unsampled cases would not affect our estimates unless it contributed to changes in the molecular clock rate.

WGS data has been noted to be helpful in ruling out transmission but insufficient, on its own, to resolve transmission events [Casali et al., 2016, Campbell et al., 2018]. If the primary use of WGS data is only to refute transmission, one might ask why clustering matters. We would argue that the transmissions that are not refuted by WGS are then presumably considered to be possible recent, or direct, or clustered transmissions. Even if the primary use of WGS data is to refute direct transmission, there is a trade-off between the strength of that refutation and the possibility of mistakenly refuting genuine recent transmission events. This is more likely, using SNP cut-offs, where selection (say for antibiotic resistance) has led to higher SNP differences than expected. In addition, in practice WGS data are not only used to refute direct transmission, but to produce clusters that inform onward analyses, reports on the extent of recent transmission, outbreak analysis and reconstruction and even public health policy; see [Guthrie et al., 2018] for one example.

We have accommodated the possibility of low substitution rates in latency with a non-Poisson model for the clock process, λ, in Equation (5) (though we have not implemented this) and to some extent with the option of a non-constant transmission rate. However, we have not modelled the possibility of a direct relationship between low SNP accumulation and low probability transmission. If this relationship exists – for example if latent cases both do not transmit and do not accumulate SNPs [Colangeli et al., 2014] – then low SNP differences could correspond to fewer intermediate hosts despite long elapsed times. This is an implicit assumption of a SNP-only method; while it may be correct it is a strong assumption, and there is evidence that mutation rates in latency are not reduced compared to active disease [Ford et al., 2011, Lillebaek et al., 2016].

We have not used the probability of sampling in forming our clusters, in contrast to other tools including the `vimes` package [Jombart and Cori, 2017]. For example, if it is known that surveillance is strong, then it would be less likely for 10 intermediate cases to be unsampled than 5, and this could be built in to a clustering method. Our rationale for not taking this into account is to provide a clustering approach that is as parallel as possible to the SNP cut-offs currently in widespread use while taking additional information on timing, molecular evolution and transmission into account. It is often the case that the true sampling rate is not known and may change over time, and particularly for TB, in high-resource settings cases can be missed because they are hard to identify (perhaps being at higher risk of TB due to homelessness or other factors, as in [Casali et al., 2016]). In many settings the sampling probability may be uncertain. We have taken the approach of defining the clusters themselves without explicit reference to the sampling probability, with the view that the clusters are central inputs to other analyses which will take sampling into account (as is done for example in TransPhylo [Didelot et al., 2017]). However, in our approach, changes

in the sampling probability would likely be apparent in changes in the temporal and genetic distance between cases over time.

We have also not modelled changes in the transmission process over time in a community (eg due to depletion of susceptible individuals, improved infection control, etc). As with including sampling, this may best be done in a more nuanced analysis after the initial clustering rather than as part of the clustering itself, but in principle, changes to the transmission function over calendar time could be incorporated into the mathematics behind Equation (8). However, this would raise interpretation challenges because of the fact that our transmission process reflects the rate of the pathogen moving between hosts where it is known that there is an infected host at the "end" of the chain (since each pair consists of two sampled hosts, whose pathogen was sequenced and who were therefore certainly infected). We do not model the number of contacts over which transmission could have occurred.

The choice of a particular SNP cut-off also takes no account of the inevitable uncertainties involved in the gathering and processing of raw read data, and does not allow for the modelling of this uncertainty. Different bioinformatics pipelines - and different parameters used within those pipelines - can have a substantial effect on the number of SNP differences reported between cases. It is usual for SNP differences to be taken as given and, although sometimes details are provided – see for example [Katz et al., 2013] – it is important to recognise that there can be considerable variation between SNPs reported using different pipelines and parameters. For example, the level of quality scores and read depth cut-offs used will generally have a high impact, as will the precise way in which hyper-variable sites and repeat regions are handled (or excluded). As technology improves we may begin to capture variation in repeat regions, or types of variation (eg insertions/deletions) that are currently masked, and in that new pipeline 12 SNPs may not carry the interpretation it does today. The model could easily incorporate more genomic information, resulting in a more sophisticated version of the distance function. In particular, large scale genomic features can readily help to establish that cases belong to separate and therefore distantly related lineages. As variation-calling pipelines evolve, our method could be used to relate each pipeline to numbers of transmissions or to estimated divergence time; this would form an approach to compare bioinformatics pipelines and data sources, and to curate their use in defining distances between isolates.

TB has distinct phylogeographic lineages which have been reported to have different mutation rates, with lineage 2 (the East Asian and Beijing lineage) having higher mutation rates than lineage 4 (Euro-American) [Ford et al., 2013]. Our approach could unify clustering despite such differences, as the same transmission and probability settings could be used under different SNP accumulation rates. This would provide a consistent approach to clustering in areas where multiple lineages co-circulate, and allow comparison of TB clustering patterns in different settings. The same would be true for adapting to differing natural histories across different pathogen lineages or sub-populations: the choice of β could reflect transmission differences while the other settings remained the same.

The long-term aim of changing how cases are assigned to clusters is to improve the way that WGS and epidemiological data are used and to best capture clusters that correspond to transmissions of an infectious disease. We have found

that basing clusters on the number of transmission events, with a probabilistic cut-off, is feasible, can integrate timing and other data, and compares favourably to clustering based on SNP cut-offs.

Materials and Methods

Data

In this paper we focus on TB, but our approach is applicable to other pathogens for which whole genome sequencing can be carried out and where it is appropriate to use SNPs to compare closely-related isolates (naturally, parameters will vary). TB provides a convenient model as it avoids the complications associated with horizontal gene transfer, it is an important pathogen worldwide, it has very diverse epidemiological settings and WGS tools are increasingly used for public health purposes.

British Columbia The British Columbia Centre for Disease Control (BCCDC)'s Public Health Laboratory (BCPHL) receives all *Mycobacterium tuberculosis* (Mtb) cultures for the province and performs routine MIRU-VNTR genotyping on all Mtb isolates. Mtb isolates belonging to MIRU-VNTR cluster MClust-012 were revived from archived stocks, DNA extracted, and sequenced using 125bp paired-end reads on the Illumina HiSeqX platform at the Michael Smith Genome Sciences Centre (Vancouver, BC). The resulting fastq files were analyzed using a pipeline developed by Oxford University and Public Health England. Reads were aligned to the Mtb H37Rv reference genome (GenBank ID: NC000962.2), with an average of 92% of the reference genome covered. Single nucleotide variants (SNVs) were identified across all mapped non-repetitive sites. Fastq files for all genomes are available at NCBI under BioProject PRJNA413593.

Republic of Moldova *Sample collection and epidemiological data*: The study population included patients diagnosed with culture positive tuberculosis at the Municipal hospital from October 2013 – December 2014 in the Republic of Moldova. All epidemiological and laboratory data from TB patients are routinely entered into a country-wide web-based TB electronic medical record (EMR) database. Epidemiological data including age, sex, previous TB history, results of chest radiograph, history of incarceration, and place of residence were collected. Laboratory data, including mycobacterial smear grade, culture and drug-susceptibility testing to first and second line anti-tuberculosis agents, were extracted from the EMR. As part of this study, all *M. tuberculosis* patient isolates were subcultured and frozen for genomic analysis.

Variant calling and phylogenetic analysis: DNA was extracted from M. tuberculosis grown on Lowenstein-Jensen slants as described previously. Paired-end (250 base pair) sequences were generated on the Illumina MiSeq platform. Raw fastq reads were filtered for length and trimmed for low-quality trailing base pairs using Trim Galore, aligned to the H37Rv NC000962.3 reference genome using BWA, with duplicate reads removed using PicardTools. The mpileup function in samtools was used for single-isolate variant calling. Isolates with a high proportion of apparent mixed or heterozygous single nucleotide polymorphism

(SNP) calls (i.e. those with >25% reads supporting the reference allele) were excluded from analysis. SNPs within 15 base pairs of insertions or deletions (indels) or with variant quality scores < 100 were excluded. SNPs in or within 50 base pairs of hypervariable PPE/PE gene families, repeat regions, and mobile elements were excluded [Eldholm et al., 2015]. A phylogenetic tree was constructed in RAxML (GTR-gamma for nucleotide substitution and correcting for SNP ascertainment bias) and annotated with DST results and drug-resistance associated variants from Mykrobe Predictor [Bradley et al., 2015]. Representative strains from other studies in the region, including L4 (LAM, Haarlem, Ural) and L2 [Casali et al., 2014, Merker et al., 2015], were also included. Percy256 (Lineage 7) was included as an outgroup. Fastq files for all genomes are available at NCBI under Accession number SRP156366.

Clustering approach

The overall approach is to use the SNPs and case timing to derive a distribution for the time to the MRCA of each pair of samples, condition on that time to write the probability that the samples were separated by some number of transmission events, and then integrate out the unknown time to the pairs' MRCA. The first step makes use of the molecular clock process and depends on the clock rate and on the numbers of SNPs under a form of selection (like antibiotic resistance). The second step using information about the transmission process and the natural history of the pathogen.

For each sample, we start with the date on which the sample was taken and the aligned nucleotide sequence for the set of variable sites in our set of samples. For any two samples S_1 and S_2, we have the SNP distance $N = N(S_1, S_2)$ which is equal to the Hamming distance between their respective nucleotide sequences. We also have the sampling time difference $\delta = \delta(S_1, S_2)$. Without loss of generality, we can assume that S_1 is sampled either at the same time as, or before, S_2. What we do not know *a priori*, and therefore we have to estimate, is the total amount of time h over which the SNPs have accumulated (on both branches in total) since the date of the MCRA of S_1 and S_2. We also refer to h as the "height".

Given the time h, we can use a transmission process to estimate the probability that there are more than some threshold number of transmission events T in a total time h; we integrate over the unknown h. This transmission process need not be homogeneous.

Symbol	Meaning	Units
λ	Clock rate	SNPs/genome/year
β	Transmission rate	Transmissions/year
h	Total time over which SNPs occur	years
δ	Time between two sample dates	years
N	No. SNP differences between two cases	
N_δ	No. SNP differences occurring after first sample	
k	Number of transmissions	
S	Cut-off threshold for SNP method	
T	Cut-off threshold for transmission method	

Table 6: Symbols used in the model and their meaning.

We make various assumptions in setting up the model. Both substitutions and transmissions occur according to (possibly non-homogeneous) Poisson processes over time. Unless it is otherwise stated, the population from which the samples are drawn is homogeneous, so transmission is random and equally likely between hosts irrespective of factors such as location of abode, individual lifestyle etc. (We show in Methods how to accommodate location information). We do not assume that all infected cases are reported and sequenced: where we do have sequence data, we assume that it is correct and complete. Reported cases may be sampled more than once. We do not explicitly model reporting and sampling rates. If these change over time, then this would be reflected in the time and genetic distance between nearby cases and consequently in the estimated number of intermediate transmissions between reported cases. Once infected, we assume that a patient becomes infectious immediately, either with a constant probability of infection per unit of time, or in a process yielding a gamma-distributed time to the next infection. This "natural history" model is assumed not to change with calendar time, such that the course of infectiousness proceeds in the same manner from infection to infecting others independent of the calendar time of infection. Our approach is intended to group sequences into clusters, and does not model reported cases for whom there is no sequence data.

Noting that δ is fixed by the sampling times in the data, we estimate the distribution of the time $h/2$ over which the SNPs have had to accumulate before the sample date of S_1. This is equivalent to estimating the date of the MRCA of S_1 and S_2. Because both branches are free to evolve over this time, $h/2 + h/2 = h$ is the effective overall time between the MRCA and S_1, and $\delta + h$ is therefore the total evolutionary time separating the two cases.

Estimate of the height where sample dates are the same

The simplest model for the number of SNPs per unit time is a Poisson process with a constant rate λ; we can also accommodate overdispersion, reflecting a more variable SNP accumulation process suitable for pathogens whose substitutions are not as clock-like (see below). The standard Poisson distribution with parameter λh gives the probability density of the number of SNPs on a given time interval h:

$$P(N|h) = \frac{e^{-\lambda h}(\lambda h)^N}{N!} \tag{1}$$

However we are interested in the likelihood of the time h as a function of the specified number of SNPs.

We know by standard theory that the arrival time density – that is, the time density until the next SNP – can be modelled by the exponential density function $\lambda e^{-\lambda h}$. Furthermore, the waiting time until the N-th SNP is also a Poisson process, as the arrivals are assumed to be independent and identically Poisson distributed. It can be shown (for example in Chapter 2 of [Gallagher, 2013]) by repeated convolution of densities that the distribution of the Nth arrival time A_N is given by

$$\mathrm{PDF}_{A_N} = \frac{\lambda^N h^{N-1} e^{-\lambda h}}{(N-1)!} \tag{2}$$

for $N > 0$. This is the Erlang distribution, with mean $= N/\lambda$, as expected.

We know that exactly N SNPs have already occurred on a time interval of uncertain length h, and we are interested in the likelihood of h given the data N. Since we already have N SNPs and are waiting for the $(N+1)$-th, this is given by the arrival time density for the $(N+1)$-th SNP; by replacing N with $N+1$ in the above and interpreting it as a function of h, we have:

$$\mathcal{L}(h|N) = \frac{\lambda^{N+1} h^N e^{-\lambda h}}{N!} \tag{3}$$

Note that when $N = 0$, this reduces to $\lambda e^{-\lambda h}$.

Alternatively, we can generalise the arrival time density to a gamma distribution, where the extra parameter allows us to fix the mean but change the variance. This allows us to be more flexible with respect to dispersion than with using the exponential distribution. The gamma density, with two parameters a and b, is

$$f(h; a, b) = \frac{b^a h^{a-1} e^{-bh}}{\Gamma(a)} \tag{4}$$

The mean is a/b and the variance a/b^2. Note that we can recover the Poisson model result by setting $a = 1$ and $b = \lambda$ [Cameron and Trivedi, 2013]. In this case the arrival time density for the $(N+1)$-th SNP is given by

$$f(h; a(N+1), b) = \frac{b^{a(N+1)} h^{a(N+1)-1} e^{-bh}}{\Gamma(a(N+1))} \tag{5}$$

by standard properties of the Gamma distribution.

Estimate of the height where sampling times differ

In this case, we account for the fact that some of the SNPs may have occured in the the fixed time interval of length δ between the two sample dates. Again, we begin with the simple model in which the number of SNPs occurring in this time is given by a Poisson distribution, in this case with parameter $\lambda\delta$. We write $N = N_h + N_\delta$, where N_δ is Poisson distributed with parameter $\lambda\delta$.

The number of SNPs N_δ accumulated on the fixed interval of length δ is somewhere between 0 and N inclusive; $0 \leq N_\delta \leq N$. Unconstrained, N_δ is Poisson

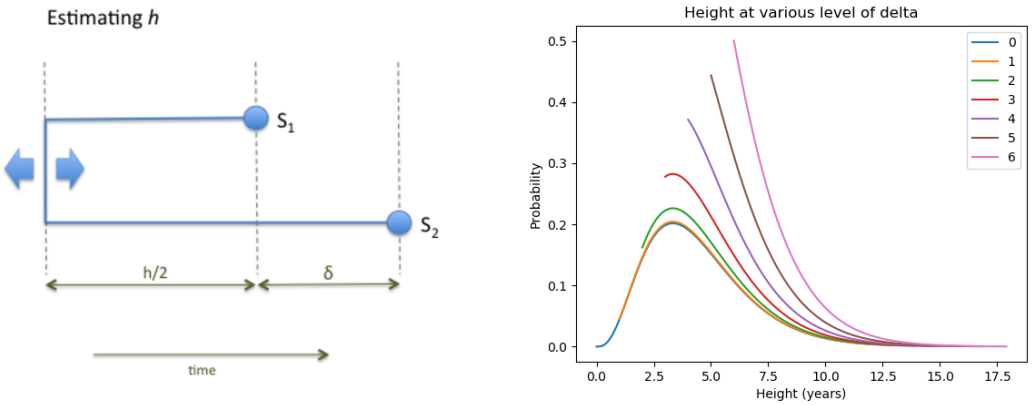

Figure 5: On the left is a schematic illustration of the notation. h is the total time in years over which SNPs accumulate between two cases before the first sample is taken, whereas the total time over which SNPs can occur is $h + \delta$ years. δ is known and fixed; h is unknown. On the right is a plot of $h + \delta$, where h is given by Equation (4), for values of δ ranging from 0 through 6 years, with $N = 3$, $\lambda = 0.9$ SNPs/genome/year, and $\beta = 1.2$ transmissions/year. Since $h + \delta > \delta$, the lines corresponding to higher values of δ begin above 0.

distributed with parameter $\lambda\delta$. Conditioning on the probability that N_δ does not exceed N gives us the probability density

$$P(N_\delta|\delta) = \left(\frac{e^{-\lambda\delta}(\lambda\delta)^{N_\delta}}{N_\delta!}\right) \Big/ \sum_{i=0}^{N} e^{-\lambda\delta}(\lambda\delta)^i/i!$$

and writing $F(N_\delta)$ for $\sum_{i=0}^{N} e^{-\lambda\delta}(\lambda\delta)^i/i!$

$$P(N_\delta|\delta) = \left(\frac{e^{-\lambda\delta}(\lambda\delta)^{N_\delta}}{N_\delta!}\right) \Big/ F(N_\delta) \tag{6}$$

To obtain the expression for $\mathcal{L}(h|N,\delta)$, we sum over all the possible values of N_δ, giving

$$\mathcal{L}(h|N,\delta) = \sum_{i=0}^{N} \mathcal{L}(h|i,\delta)P((N-i)|\delta)$$

Substituting into our earlier expression,

$$\mathcal{L}(h|N,\delta) = \sum_{i=0}^{N} \left(\frac{\lambda^{i+1}h^i e^{-\lambda h}}{i!}\right)\left(\frac{e^{-\lambda\delta}(\lambda\delta)^{(N-i)}}{(N-i)!}\right) \Big/ F(N)$$

$$\mathcal{L}(h|N,\delta) = \frac{e^{-\lambda(h+\delta)}\lambda^{N+1}}{F(N)} \sum_{i=0}^{N} \frac{h^i\delta^{N-i}}{i!(N-i)!} \tag{7}$$

An example plot for the equation above is shown in Figure 5.

Modelling transmissions

We connect SNP distances to transmissions using a model for the number of transmissions likely to have occurred over a given total time period, *conditional*

on the two cases being infected at or before the sampling times. This means that unlike a transmission rate in a population-level epidemic model, which typically describes the rate of transmission per unit time given contact between a susceptible and an infectious individual, our transmission process is better described in terms of the rate at which a pathogen lineage will jump to a new host. This is, of course, distinct from the rate at which new transmissions occur in a community and the per-contact rate of transmission of infection between two individuals. We first assume for simplicity that β is a constant function, and that it is a Poisson process; we allow a more general model later. The amount of time over which transmissions can occur between our two cases is $h + \delta$, and the expected number of transmissions is $\beta(h + \delta)$. The number of transmission events k is therefore given by

$$P(k|h, \delta) = \frac{\beta^k (h+\delta)^k e^{-\beta(h+\delta)}}{k!} \tag{8}$$

Integrating over h, we have

$$P(k|N, \delta) = \int_{h=0}^{\infty} \mathcal{L}(h|N, \delta)\ P(k|h)\ dh \tag{9}$$

$$= \frac{e^{-\delta(\lambda+\beta)} \lambda^{N+1} \beta^k}{k! F(N)} \int_{h=0}^{\infty} e^{-h(\lambda+\beta)} (h+\delta)^k \sum_{i=0}^{N} h^i \left(\frac{\delta^{N-i}}{i!(N-i)!} \right) dh \tag{10}$$

This equation expresses the key relationship that allows us to translate raw SNP differences and sample time differences into transmission probability distributions - examples are shown in Figure 6. As the sample time between cases increases, it can be seen that this factor makes an increasingly important contribution, relative to the SNP distance, to the distance between cases.

Unless stated otherwise, Equation (10) is used to generate the data presented in the Results Section.

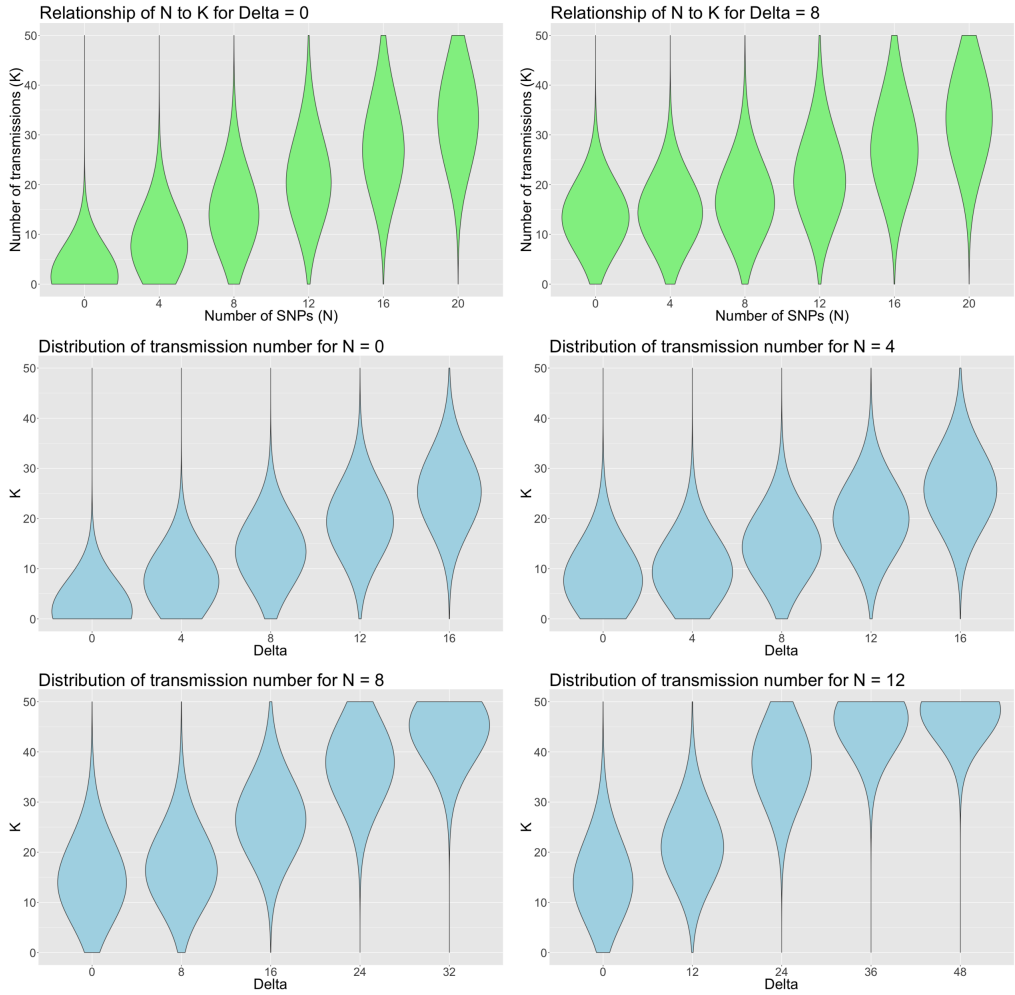

Figure 6: Probability density for the number of transmissions, given by Equation (10), with clock rate $\lambda = 0.9$ SNPs/genome/year and $\beta = 1.5$ transmissions/year. The upper two panels show the densities for delta values of 0 and 4 years for a range of SNP distance between 0 and 20. The lower four panels show the densities for 0, 4, 8 and 12 SNP distance respectively, for $\delta = 0, 4, 8, 12, 16$ years.

Time varying transmissions

In our context, a transmission event should be understood as an event in which a pathogen is transferred to a new host, ultimately causing a secondary case in that host. While there may be undetected transmission events in which the secondary cases never develop disease, our data are on sampled cases with active disease, and the time between successive transmissions should approximately reflect the serial interval between cases with active disease. We allow the number of transmissions $\beta = \beta(t)$ to be a function of time since infection, allowing for a variable risk of infecting others during the course of infection. Once a host is infected, the details of the natural history of the pathogen affect the generation time - the exact form of the function $\beta(t)$ allows us to incorporate the varying rates of progression from infection to active disease, and then on to transmission. This illustrates that our framework has the flexibility to include more detailed and accurate modelling of the underlying disease dynamics. As stated in [Didelot et al., 2017], the generation time distribution can take

any form [Fine, 2003, Wallinga and Lipsitch, 2007], but Gamma distributions are often used, as for example in [Conlan et al., 2010]. We apply the Gamma distribution with parameters shape α and scale θ, so that:

$$\beta(t) = \beta(t;\alpha,\theta) = \frac{1}{\Gamma(\alpha)\theta^{\alpha}} t^{\alpha-1} e^{-t/\theta} \tag{11}$$

and the mean value is $\alpha\theta * dt$ in a given time interval dt.

Putting this together with our Poisson model for the number of SNPs on a time interval (Equation (1)), we obtain:

$$P(N|\lambda,\beta(t;\alpha,\theta)) = \int_{t=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^N}{N!} \Big(\frac{1}{\Gamma(\alpha)\theta^{\alpha}} t^{\alpha-1} e^{-t/\theta}\Big) dt \tag{12}$$

$$= \binom{N+\alpha-1}{N} \big(1 - \frac{\theta\lambda}{\theta\lambda+1}\big)^{\alpha} \big(\frac{\theta\lambda}{\theta\lambda+1}\big)^{N} \tag{13}$$

This is a negative binomial (denoted NB) distribution for the number of SNPs for one transmission generation,

$$P(N|\lambda,\beta(t;\alpha,\theta)) \sim \mathrm{NB}\big(\alpha; \frac{\theta\lambda}{\theta\lambda+1}\big)$$

Assuming transmission events are independent of each other, it then follows (by standard properties of the negative binomial) that the probability of N given k transmissions is also distributed as a negative binomial, with

$$P(N|k,\lambda,\beta(t;\alpha,\theta)) \sim \mathrm{NB}\big(k\alpha; \frac{\theta\lambda}{\theta\lambda+1}\big)$$

$$P(N|k,\lambda,\beta(t;\alpha,\theta)) = \binom{N+k\alpha-1}{N} \big(1 - \frac{\theta\lambda}{\theta\lambda+1}\big)^{k\alpha} \big(\frac{\theta\lambda}{\theta\lambda+1}\big)^{N} \tag{14}$$

Modelling resistance conferring SNPs

Suppose that we know that there are resistance conferring SNPs in our sample population, or perhaps other SNPs at sites known to be under selection or simply to have a different rate of substitution. Let us assume they account for a certain fixed proportion of the observed SNP differences. Given N SNPs, assume that m are not resistance conferring and n are, so $N = m + n$. Their respective mutation rates are given by λ_m and λ_n, where $\lambda_n > \lambda_m$. Assuming independence, on a given time interval of length h we have

$$\begin{aligned} P(N|h) &= P(m|h,\lambda_m)P(n|h,\lambda_n) \\ &= \Big(\frac{\lambda_m^m h^m e^{-\lambda_m h}}{m!}\Big)\Big(\frac{\lambda_n^n h^n e^{-\lambda_n h}}{n!}\Big) \\ &= \Big(\frac{\lambda_m^m \lambda_n^n}{m!n!}\Big) h^m e^{-\lambda_m h} h^n e^{-\lambda_n h} \\ &= \Lambda_{mn} h^{m+n} e^{-(\lambda_m+\lambda_n)h} \end{aligned} \tag{15}$$

where

$$\Lambda_{mn} = \frac{\lambda_m^m \lambda_n^n}{m!n!} \tag{16}$$

Compare this to Equation (1), which we can recover by setting $m = N$, $n = 0$, $\lambda = \lambda_m$, and $\lambda_n = 0$.

We have a Poisson process which is the sum of two independent Poisson processes with $\lambda = \lambda_m + \lambda_n$. As before, we can derive expressions for $\mathcal{L}(h|N = n + m)$ and $P(k|N = n + m)$, so

$$\mathcal{L}(h|m, n) = \Lambda'_{mn} h^{m+n} e^{-(\lambda_m + \lambda_n)h} \tag{17}$$

where

$$\Lambda'_{mn} = \frac{\lambda_m^{m+1} \lambda_n^{n+1}}{m!n!}$$

A way to illustrate the effect of including resistance-conferring SNPs is to consider the expected value of h. Recall that under Equation (1), the mean is given by N/λ. Thinking of our resistance and non-resistance conferring SNP processes, they have means respectively of n/λ_n and m/λ_m. Thus the combined process has mean $n/\lambda_n + m/\lambda_m$, and we can write the rate parameter λ^* of the combined process as

$$\lambda^* = \frac{n + m}{n/\lambda_n + m/\lambda_m} \tag{18}$$

Note that for large λ_m, λ^* tends to $\lambda_n * (n + m)/n$. The larger the value of λ_m as compared to λ_n, the smaller the contribution that the resistance conferring SNPs make to the value of h - accordingly, 4 SNPs likely to have arisen due to inappropriate treatment or another selection process should not contribute as strongly towards separating two cases into different transmission clusters as 4 "neutral" SNPs. Ideally, the value of λ_m should be estimated from data. Once resistance SNPs have occurred in an individual, they are likely to be transmitted onwards when the individual infects others. These secondary cases share the resistance SNPs with each other ($n = 0$ in these pairs) and they are likely to be placed in the same cluster. Between each secondary case and the infecting case, $n > 0$; our method allows the resistance SNPs to "count for" less time than other SNPs, and the index case is likely clustered with the onward cases.

Spatial proximity and other individual data

Other factors that affect the likelihood of transmission, such as spatial proximity or other covariate data including contact tracing, demographics or other host factors, can be built into the model.

To incorporate spatial proximity, we assign each of the cases into one of a number of regions R_i where i is the region index. For the British Columbia data set, there are six regions defined, as shown in 3. For any pair of cases, a probability weighting is assigned which is equal to 1 in the case that both cases belong to the same region, and a value below 1 for cases which belong to different regions. This weighting w is then applied to the probability of obtaining k transmissions

given N SNPs, giving us a modified version of Equation (10)

$$P(k|N,\delta,w) = w \int_{h=0}^{\infty} \mathcal{L}(h|N,\delta)\ P(k|h)\ dh \tag{19}$$

Simulations

We generate simulated outbreaks and compare the SNP and transmission methods on them with a technique that measures the similarity of clusters using an information-theoretic approach [Meilă, 2007]. Outbreaks are simulated using *TransPhylo* [Didelot et al., 2017], which generates a dated transmission network for each simulation, containing both sampled and unsampled cases. From these, and for all the cases, phylogenetic trees are extracted using *phyloTop* [Kendall et al., 2016]. Sequences are then generated with *phangorn* [Schliep, 2011] and output as *fasta* format files. For the sampled, and therefore "known", cases we generate sets of clusters using the SNP and transmission methods for a range of cut-off levels. We also generate the "true" clustering of the sampled cases implied by the simulated *TransPhylo* transmission networks.

Software availability

The methods presented here are available as R functions in the `transcluster` package, available at `https://github.com/JamesStimson/transcluster`.

Acknowledgements

CC and JS are supported by EPSRC grant EP/K026003/1 and CC is additionally supported by EPSRC grant EP/N014529/1. TC is supported in part by U54GM088558 from the National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute Of General Medical Sciences or the National Institutes of Health. We would like to thank Tyler S. Brown for contributions to the bioinformatics and analysis of the data from Moldova.

References

[Azarian et al., 2016] Azarian, T., Maraqa, N. F., Cook, R. L., Johnson, J. A., Bailey, C., Wheeler, S., Nolan, D., Rathore, M. H., Morris Jr, J. G., and Salemi, M. (2016). Genomic epidemiology of methicillin-resistant Staphylococcus aureus in a neonatal intensive care unit. *PloS One*, 11(10):e0164397.

[Barrick and Lenski, 2013] Barrick, J. E. and Lenski, R. E. (2013). Genome dynamics during experimental evolution. *Nature Reviews Genetics*, 14(12):827.

[Bergholz et al., 2014] Bergholz, T. M., Switt, A. I. M., and Wiedmann, M. (2014). Omics approaches in food safety: fulfilling the promise? *Trends in Microbiology*, 22(5):275–281.

[Bradley et al., 2015] Bradley, P., Gordon, N. C., Walker, T. M., Dunn, L., Heys, S., Huang, B., Earle, S., Pankhurst, L. J., Anson, L., De Cesare, M., et al. (2015). Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nature Communications*, 6:10063.

[Bryant et al., 2013a] Bryant, J. M., Harris, S. R., Parkhill, J., Dawson, R., Diacon, A. H., van Helden, P., Pym, A., Mahayiddin, A. A., Chuchottaworn, C., Sanne, I. M., et al. (2013a). Whole-genome sequencing to establish relapse or re-infection with Mycobacterium tuberculosis: a retrospective observational study. *The Lancet Respiratory Medicine*, 1(10):786–792.

[Bryant et al., 2013b] Bryant, J. M., Schürch, A. C., van Deutekom, H., Harris, S. R., de Beer, J. L., de Jager, V., Kremer, K., van Hijum, S. A., Siezen, R. J., Borgdorff, M., et al. (2013b). Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data. *BMC Infectious Diseases*, 13(1):110.

[Cameron and Trivedi, 2013] Cameron, A. C. and Trivedi, P. K. (2013). *Regression Analysis of Count Data*, volume 53. Cambridge University Press.

[Campbell et al., 2018] Campbell, F., Strang, C., Ferguson, N., Cori, A., and Jombart, T. (2018). When are pathogen genome sequences informative of transmission events? *PLoS Pathog.*, 14(2):e1006885.

[Casali et al., 2016] Casali, N., Broda, A., Harris, S. R., Parkhill, J., Brown, T., and Drobniewski, F. (2016). Whole Genome Sequence Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective Observational Study. *PLoS Med.*, 13(10):e1002137.

[Casali et al., 2014] Casali, N., Nikolayevskyy, V., Balabanova, Y., Harris, S. R., Ignatyeva, O., Kontsevaya, I., Corander, J., Bryant, J., Parkhill, J., Nejentsev, S., et al. (2014). Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nature Genetics*, 46(3):279.

[Clark et al., 2013] Clark, T. G., Mallard, K., Coll, F., Preston, M., Assefa, S., Harris, D., Ogwang, S., Mumbowa, F., Kirenga, B., O'Sullivan, D. M., et al. (2013). Elucidating emergence and transmission of multidrug-resistant tuberculosis

in treatment experienced patients by whole genome sequencing. *PLoS One*, 8(12):e83012.

[Colangeli et al., 2014] Colangeli, R., Arcus, V. L., Cursons, R. T., Ruthe, A., Karalus, N., Coley, K., Manning, S. D., Kim, S., Marchiano, E., and Alland, D. (2014). Whole genome sequencing of Mycobacterium tuberculosis reveals slow growth and low mutation rates during latent infections in humans. *PLoS One*, 9(3):e91024.

[Conlan et al., 2010] Conlan, A. J., Rohani, P., Lloyd, A. L., Keeling, M., and Grenfell, B. T. (2010). Resolving the impact of waiting time distributions on the persistence of measles. *Journal of the Royal Society Interface*, 7(45):623–640.

[Dallman et al., 2015] Dallman, T. J., Ashton, P. M., Byrne, L., Perry, N. T., Petrovska, L., Ellis, R., Allison, L., Hanson, M., Holmes, A., Gunn, G. J., et al. (2015). Applying phylogenomics to understand the emergence of Shiga-toxin-producing Escherichia coli O157: H7 strains causing severe human disease in the UK. *Microbial Genomics*, 1(3).

[Didelot et al., 2017] Didelot, X., Fraser, C., Gardy, J., and Colijn, C. (2017). Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular Biology and Evolution*, 34(4):997–1007.

[Didelot et al., 2014] Didelot, X., Gardy, J., and Colijn, C. (2014). Bayesian inference of infectious disease transmission from whole-genome sequence data. *Molecular Biology and Evolution*, 31(7):1869–1879.

[Donker et al., 2016] Donker, T., Bosch, T., Ypma, R., Haenen, A., van Ballegooijen, W., Heck, M., Schouls, L., Wallinga, J., and Grundmann, H. (2016). Monitoring the spread of meticillin-resistant Staphylococcus aureus in The Netherlands from a reference laboratory perspective. *Journal of Hospital Infection*, 93(4):366–374.

[Eldholm et al., 2015] Eldholm, V., Monteserin, J., Rieux, A., Lopez, B., Sobkowiak, B., Ritacco, V., and Balloux, F. (2015). Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. *Nature Communications*, 6:7119.

[Eldholm et al., 2014] Eldholm, V., Norheim, G., von der Lippe, B., Kinander, W., Dahle, U. R., Caugant, D. A., Mannsåker, T., Mengshoel, A. T., Dyrhol-Riise, A. M., and Balloux, F. (2014). Evolution of extensively drug-resistant Mycobacterium tuberculosis from a susceptible ancestor in a single patient. *Genome Biology*, 15(11):490.

[Feuerriegel et al., 2015] Feuerriegel, S., Schleusener, V., Beckert, P., Kohl, T. A., Miotto, P., Cirillo, D. M., Cabibbe, A. M., Niemann, S., and Fellenberg, K. (2015). PhyResSE: a web tool delineating Mycobacterium tuberculosis antibiotic resistance and lineage from whole-genome sequencing data. *Journal of Clinical Microbiology*, 53(6):1908–1914.

[Fine, 2003] Fine, P. E. (2003). The interval between successive cases of an infectious disease. *American Journal of Epidemiology*, 158(11):1039–1047.

[Ford et al., 2011] Ford, C. B., Lin, P. L., Chase, M. R., Shah, R. R., Iartchouk, O., Galagan, J., Mohaideen, N., Ioerger, T. R., Sacchettini, J. C., Lipsitch, M., Flynn, J. L., and Fortune, S. M. (2011). Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. *Nature Genetics*, 43(5):482–486.

[Ford et al., 2013] Ford, C. B., Shah, R. R., Maeda, M. K., Gagneux, S., Murray, M. B., Cohen, T., Johnston, J. C., Gardy, J., Lipsitch, M., and Fortune, S. M. (2013). Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nature Genetics*, 45(7):784–790.

[Gallagher, 2013] Gallagher, R. G. (2013). *Stochastic Processes: Theory for Applications*. Cambridge University Press.

[Guerra-Assunção et al., 2015] Guerra-Assunção, J. A., Houben, R., Crampin, A. C., Mzembe, T., Mallard, K., Coll, F., Khan, P., Banda, L., Chiwaya, A., Pereira, R., et al. (2015). Relapse or reinfection with tuberculosis: a whole genome sequencing approach in a large population-based cohort with high HIV prevalence and active follow-up. *The Journal of Infectious Diseases*, 211(7):1154–63.

[Guthrie et al., 2018] Guthrie, J. L., Delli Pizzi, A., Roth, D., Kong, C., Jorgensen, D., Rodrigues, M., Tang, P., Cook, V. J., Johnston, J., and Gardy, J. L. (2018). Genotyping and Whole-Genome Sequencing to Identify Tuberculosis Transmission to Pediatric Patients in British Columbia, Canada, 2005–2014. *J. Infect. Dis.*, 40:1–9.

[Hall et al., 2015] Hall, M., Woolhouse, M., and Rambaut, A. (2015). Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set. *PLoS Computational Biology*, 11(12):e1004613.

[Hall et al., 2016] Hall, M. D., Woolhouse, M. E. J., and Rambaut, A. (2016). Using genomics data to reconstruct transmission trees during disease outbreaks. *Rev. Sci. Tech.*, 35(1):287–296.

[Hatherell et al., 2016] Hatherell, H.-A., Colijn, C., Stagg, H. R., Jackson, C., Winter, J. R., and Abubakar, I. (2016). Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. *BMC Medicine*, 14(1):21.

[Jombart and Cori, 2017] Jombart, T. and Cori, A. (2017). vimes. `http://www.repidemicsconsortium.org/vimes/articles/distributions.html`.

[Kammerer et al., 2013] Kammerer, J. S., Shang, N., Althomsons, S. P., Haddad, M. B., Grant, J., and Navin, T. R. (2013). Using statistical methods and genotyping to detect tuberculosis outbreaks. *International Journal of Health Geographics*, 12(1):15.

[Katz et al., 2013] Katz, L. S., Petkau, A., Beaulaurier, J., Tyler, S., Antonova, E. S., Turnsek, M. A., Guo, Y., Wang, S., Paxinos, E. E., Orata, F., et al. (2013). Evolutionary dynamics of Vibrio cholerae O1 following a single-source introduction to Haiti. *MBio*, 4(4):e00398–13.

[Kendall et al., 2016] Kendall, M., Boyd, M., and Colijn, C. (2016). phyloTop. `https://CRAN.R-project.org/package=phyloTop`.

[Korhonen et al., 2016] Korhonen, V., Smit, P., Haanperä, M., Casali, N., Ruutu, P., Vasankari, T., and Soini, H. (2016). Whole genome analysis of Mycobacterium tuberculosis isolates from recurrent episodes of tuberculosis, Finland, 1995–2013. *Clinical Microbiology and Infection*, 22(6):549–554.

[Kuo and Ochman, 2009] Kuo, C.-H. and Ochman, H. (2009). Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria. *Biology Direct*, 4(1):35.

[Lee et al., 2015] Lee, R. S., Radomski, N., Proulx, J.-F., Manry, J., McIntosh, F., Desjardins, F., Soualhine, H., Domenech, P., Reed, M. B., Menzies, D., et al. (2015). Reemergence and amplification of tuberculosis in the Canadian arctic. *The Journal of Infectious Diseases*, 211(12):1905–1914.

[Lillebaek et al., 2016] Lillebaek, T., Norman, A., Rasmussen, E. M., Marvig, R. L., Folkvardsen, D. B., Andersen, Å. B., and Jelsbak, L. (2016). Substantial molecular evolution and mutation rates in prolonged latent Mycobacterium tuberculosis infection in humans. *International Journal of Medical Microbiology*, 306(7):580–585.

[Meilă, 2007] Meilă, M. (2007). Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895.

[Merker et al., 2015] Merker, M., Blin, C., Mona, S., Duforet-Frebourg, N., Lecher, S., Willery, E., Blum, M. G., Rüsch-Gerdes, S., Mokrousov, I., Aleksic, E., et al. (2015). Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. *Nature Genetics*, 47(3):242.

[Novichkov et al., 2004] Novichkov, P. S., Omelchenko, M. V., Gelfand, M. S., Mironov, A. A., Wolf, Y. I., and Koonin, E. V. (2004). Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. *Journal of Bacteriology*, 186(19):6575–6585.

[Octavia et al., 2015] Octavia, S., Wang, Q., Tanaka, M. M., Kaur, S., Sintchenko, V., and Lan, R. (2015). Delineating community outbreaks of Salmonella enterica serovar Typhimurium by use of whole-genome sequencing: insights into genomic variability within an outbreak. *Journal of Clinical Microbiology*, 53(4):1063–1071.

[Poon, 2016] Poon, A. F. (2016). Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks. *Virus Evolution*, 2(2):vew031.

[Roetzer et al., 2013] Roetzer, A., Diel, R., Kohl, T. A., Rückert, C., Nübel, U., Blom, J., Wirth, T., Jaenicke, S., Schuback, S., Rüsch-Gerdes, S., et al. (2013). Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. *PLoS Medicine*, 10(2):e1001387.

[Schliep, 2011] Schliep, K. P. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4):592.

[Walker et al., 2013] Walker, T. M., Ip, C. L., Harrell, R. H., Evans, J. T., Kapatai, G., Dedicoat, M. J., Eyre, D. W., Wilson, D. J., Hawkey, P. M., Crook, D. W., et al. (2013). Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet Infectious Diseases*, 13(2):137–146.

[Walker et al., 2014] Walker, T. M., Lalor, M. K., Broda, A., Ortega, L. S., Morgan, M., Parker, L., Churchill, S., Bennett, K., Golubchik, T., Giess, A. P., et al. (2014). Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007–12, with whole pathogen genome sequences: an observational study. *The Lancet Respiratory Medicine*, 2(4):285–292.

[Wallinga and Lipsitch, 2007] Wallinga, J. and Lipsitch, M. (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society of London B: Biological Sciences*, 274(1609):599–604.

[Worby et al., 2014] Worby, C. J., Lipsitch, M., and Hanage, W. P. (2014). Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. *PLoS Computational Biology*, 10(3):e1003549.

[Yang et al., 2017] Yang, C., Luo, T., Shen, X., Wu, J., Gan, M., Xu, P., Wu, Z., Lin, S., Tian, J., Liu, Q., et al. (2017). Transmission of multidrug-resistant Mycobacterium tuberculosis in Shanghai, China: a retrospective observational study using whole-genome sequencing and epidemiological investigation. *The Lancet Infectious Diseases*, 17(3):275–284.

[Ypma et al., 2013] Ypma, R. J., Donker, T., van Ballegooijen, W. M., and Wallinga, J. (2013). Finding evidence for local transmission of contagious disease in molecular epidemiological datasets. *PloS One*, 8(7):e69875.