Beyond the SNP threshold: identifying outbreak clusters using inferred transmissions

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

May 10, 2018

Abstract

Whole genome sequencing (WGS) is increasingly used to aid in understanding pathogen transmission [1]. Very often the number of single nucleotide polymorphisms (SNPs) separating isolates collected during an epidemiological study are used to identify sets of cases that are potentially linked by direct transmission. However, there is little agreement in the literature as to what an appropriate SNP cut-off threshold should be, or indeed whether a simple SNP threshold is appropriate for identifying sets of isolates to be treated as "transmission clusters". The SNP thresholds that have been adopted for inferring transmission vary widely even for one pathogen. As an alternative to reliance on a strict SNP threshold, we suggest that the key inferential target when studying the spread of an infectious disease is the number of transmission events separating cases. Here we describe a new framework for deciding whether two pathogen genomes should be considered as part of the same transmission cluster, based jointly on the number of SNP differences and the length of time over which those differences have accumulated. Our approach allows us to probabilistically characterize the number of inferred transmission events that separate cases. We show how this framework can be modified to consider variable mutation rates across the genome (e.g. SNPs associated with drug resistance) and we indicate how the methodology can be extended to incorporate epidemiological data such as spatial proximity. We use recent data collected from tuberculosis studies from British Columbia, Canada and the Republic of Moldova to apply and compare our clustering method to the SNP threshold approach. In the British Columbia data, different cases break off from the main clusters as cut-off thresholds are lowered; the transmission-based method obtains slightly different clusters than the SNP cut-offs. For the Moldova data, straightforward application of the methods shows no appreciable difference, but when we take into account the fact that resistance conferring sites likely do not follow the same mutation clock as most sites due to selection, the transmission-based approach differs from the SNP cut-off method. Outbreak simulations confirm that our transmission based method is at least as good at identifying direct transmissions as a SNP cut-off. We conclude that the new method is a promising step towards establishing a more robust identification of outbreaks.

1 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) [1]. 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 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 [2]. 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 [1] in the case of TB. Similar methods are also used for other pathogens [3, 4]. However, there is little agreement in the literature as to what such a threshold should be – see Table 1 for TB SNP cut-off thresholds, either stated explicitly or implied. Whilst the numbers used may well be reasonable in their individual contexts, these thresholds vary substantially across studies even for the same pathogen, ranging from 2 to 50 or more in the case of TB. The variation in SNP threshold occurs for good reason: 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 [5]: "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 [6], [7] in the context of food-borne pathogens, and [8] in an analysis of the spread of MRSA. Nevertheless, the use of a single SNP threshold is often employed; for example the 12 SNP threshold used for inferring likely transmission between a pair of TB cases by Public Health England [9] 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, we account for sources of uncertainty and attempt to propagate this uncertainty into any conclusions. It is also important that

Authors	Year	SNP Threshold
Bryant et al. [10]	2013	6
Clark et al. [11]	2013	50
Guerra-Assunção et al. [6]	2015	10-100
Lee et al. [12]	2015	2
Roetzer et al. [13]	2013	3
Walker et al. [5]	2013	12
Walker et al. [9]	2014	12
Yang et al. [14]	2017	12

Table 1: SNP thresholds used in recent TB studies

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 apart. However, the rate at which polymorphisms occur varies not only between organisms [15], but also across a genome; it is affected by selection pressure and by horizontal gene transfer (HGT) [16] (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 [17], 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 which depends on both the mutation rate and the effects of natural selection. Here, when we refer to the clock rate, we mean the substitution rate.

This distinction is particularly important in the case of diseases like TB, where selection pressure due to antibiotics can be significant. Whilst the background SNP accumulation rate for *Mycobacterium tuberculosis* has been estimated at 0.5 SNPs/genome/year [5], selection pressure and antibiotic resistance can influence this rate considerably. For example, in [18] 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 [19], 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.

Others have introduced statistical frameworks for combining genetic and epidemiological information, e.g. [20], incorporating factors such as sample time, spatial location of hosts, and contact information to allocate cases to putative transmission clusters for onward analysis. There are software packages such as *vimes* [21], which provides tools allowing users to integrate different types of data

and detect outbreaks. However, hosts move in space over the course of infection, particularly for chronic infections like TB, and contact information is often unavailable. 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.

2 Methods

2.1 Clustering

We shall compare two clustering methods. With the commonly used SNP based cut-off, the "SNP method", two samples are considered to be in the same cluster if the SNP distance is less than or equal to than the specified cut-off or threshold (we will use these terms interchangeably). Note that a pair of samples at a greater distance than this may end up in the same cluster because they may be linked by chains of intermediate cases.

Our proposed probabilistic transmission based cut-off, the "transmission method", uses SNP distances, timing information and (optionally) other factors. It is based on sample pairs being clustered together if they are linked, with a given probability, by fewer than a threshold number of transmission events. The transmission method 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 process, and can easily be extended to incorporate other factors.

We start by looking at a single pair of samples, as illustrated in Figure 1. We focus on establishing probability distributions for the length of time back to the most recent common ancestor of a sample pair; this is effectively what defines the distance function between the two samples, which in turn forms the basis of the clustering method. It is by way of this estimated distribution that we have the flexibility to incorporate the sample time information as well as other forms of data.

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 $\delta = \delta(S_1, S_2)$. 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 shall at times refer to h as the "height".

Our method works by first finding a distribution for the amount of time elapsed since the most recent common ancestor of two cases given their sampling dates, the observed number of SNPs between them, and knowledge of the substitution process for the pathogen. The substitution process need not be strictly clock-like (ie not Poisson). Given the time elapsed, we can use a transmission process to estimate the probability that there are more than some threshold number of transmission events in a total time h; we integrate over the unknown h. This transmission process need not be homogeneous.

We make various assumptions in setting up the model. Substitutions and transmissions occur according to a (possibly non-homogeneous) process over time. Unless it is otherwise stated, the population from which the samples are drawn is homogeneous, so transmission is equally likely between hosts irrespective of factors such as location of abode, individual lifestyle etc. (it is straightforward to accommodate this information if it is known, and if the affect on the probability of transmission is known). Also, without loss of generality, we can assume that S_1 is sampled either at the same time as, or before, S_2.

Noting that δ is fixed by the sampling times in the data, we now 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.

2.1.1 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 [22]) by repeated convolution of densities that the distribution of the Nth arrival time A_N is given by

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

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:

$$\mathcal{L}(h|N) = \frac{\lambda^{N+1} h^N e^{-\lambda h}}{N!} \tag{2}$$

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 [23]. 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

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

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 $b = \lambda$ [24]. In this case the arrival time density for the $(N+1)$-th SNP is given by

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

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 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 $\sum_{i=0}^{N} e^{-\lambda\delta}(\lambda\delta)^i/i!$

$$P(N_\delta|\delta) = \left(\frac{e^{-\lambda\delta}(\lambda\delta)^{N_\delta}}{N_\delta!}\right) \Big/ F(N_\delta) \tag{3}$$

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

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

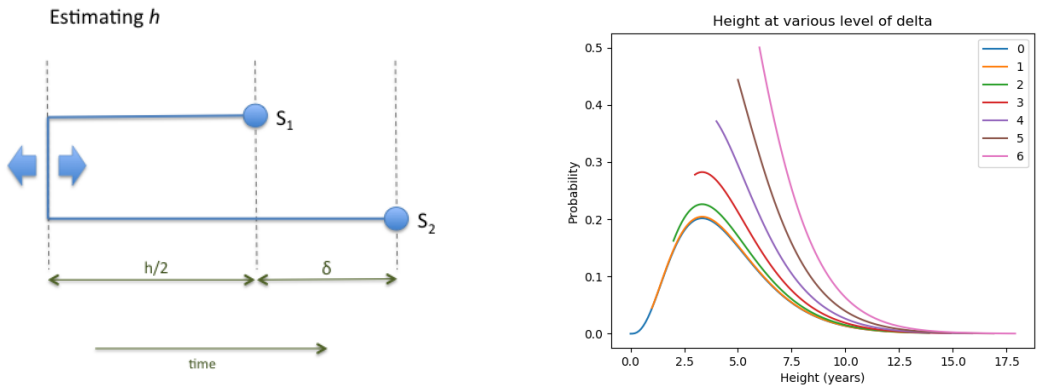

Figure 1: On the left is a schematic illustration of the notation. h is the total time over which SNPs accumulate between two cases before the first sample is taken, whereas the total time over which SNPs can occur is $h + \delta$. δ is known and fixed; h is unknown. On the right is a plot of $h + \delta$, where h is given by Equation (4), for values of δ ranging from 0 through 6, with $N = 3$, $\lambda = 0.9$, and $\beta = 1.2$. Since $h + \delta > \delta$, the lines corresponding to higher values of δ begin above 0.

Substituting into our earlier expression,

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

$$\mathcal{L}(h|N,\delta) = \frac{e^{-\lambda(h+\delta)}\lambda^{N+1}}{F(N)} \sum_{i=0}^{N} \frac{h^i \delta^{N-i}}{i!(N-i)!} \tag{4}$$

An example plot for the equation above is shown in Figure 1.

Modelling transmissions

Now we introduce the transmission process, firstly assuming for simplicity that β is a constant function, and that it is a Poisson process. 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

$$P(k|h,\delta) = \frac{\beta^k (h+\delta)^k e^{-\beta(h+\delta)}}{k!} \tag{5}$$

Integrating over h, we have

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

$$= \frac{e^{-\delta(\lambda+\beta)}\lambda^{N+1}\beta^k}{k!F(N)} \int_{h=0}^{\infty} e^{-h(\lambda+\beta)}(h+\delta)^k \sum_{i=0}^{N} h^i \left(\frac{\delta^{N-i}}{i!(N-i)!}\right) dh \tag{6}$$

This equation expresses the relationship that allows us to translate raw SNP differences and sample time differences into transmission probability distributions - examples are shown in Figure 2. 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.

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 TB disease, our data are on sampled TB cases with active disease, and the time between successive transmissions should roughly reflect the generation time 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. Generation times are often modelled as a gamma distribution [25], which we do here with parameters shape S and scale θ, so that:

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

and the mean value is $S\theta * dt$ 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:

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

$$= \binom{N+S-1}{N} (1 - \frac{\theta\lambda}{\theta\lambda+1})^S (\frac{\theta\lambda}{\theta\lambda+1})^N \tag{7}$$

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

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

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

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

$$P(N|k, \lambda, \beta(t; S, \theta)) = \binom{N+kS-1}{N} (1 - \frac{\theta\lambda}{\theta\lambda+1})^{kS} (\frac{\theta\lambda}{\theta\lambda+1})^N \tag{8}$$

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

$$\begin{aligned} P(N|h) &= P(m|h, \lambda_m)P(n|h, \lambda_n) \\ &= \left(\frac{\lambda_m^m h^m e^{-\lambda_m h}}{m!}\right)\left(\frac{\lambda_n^n h^n e^{-\lambda_n h}}{n!}\right) \\ &= \left(\frac{\lambda_m^m \lambda_n^n}{m!n!}\right) h^m e^{-\lambda_m h} h^n e^{-\lambda_n h} \\ &= \Lambda_{mn} h^{m+n} e^{-(\lambda_m+\lambda_n)h} \end{aligned} \tag{9}$$

where

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

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

$$\mathcal{L}(h|m, n) = \Lambda'_{mn} h^{m+n} e^{-(\lambda_m+\lambda_n)h} \tag{10}$$

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

$$\lambda^* = \frac{n + m}{n/\lambda_n + m/\lambda_m} \tag{11}$$

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.

Further extensions to the model

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 function $\beta = \beta(t, s, x_i)$, where t is time and s is spatial distance and x_i are other covariates.

2.2 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.

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 [26]. 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 [27]. Representative strains from other studies in the region, including L4 (LAM, Haarlem, Ural) and L2 [28, 29], were also included. Percy256 (Lineage 7) was included as an outgroup.

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.

3 Results

Figure 3 illustrates 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: the samples are clustered together where the implied number of transmissions 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 4).

Altering the transmission rate β alters the absolute transmission cut-off level at which the clusters change – in this example, increasing β to 3.0 gives the same clusters as in Figure 3 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.

By contrast, altering the clock rate does have a material impact on the way clustering occurs as we increase the transmission threshold. For $\lambda = 0.5$, 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 $\lambda = 1.5$, cases C and D are closest under the transmission method, and the clustering evolves as shown in Figures 3 and 4.

3.1 British Columbia data

We analyse a data set from British Columbia, comparing the SNP method to the transmission method in its simplest form. 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. An illustration is shown in Figure 5. The probabilistic nature of the approach means that we can see how strongly we predict cases to be linked; here we use thicker edges to denote a higher probability of being linked by relatively few transmissions.

We used the transmission method with a transmission rate $\beta = 2.0$ transmissions per year and two different average clock rates: $\lambda = 0.5$ and 1.5 SNPs per genome per year (1.5 is larger than the typical rate for TB but within other outbreak estimates [30]). When λ is low we can obtain the same clustering as with the SNP cutoff. 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 constructed example, timing alters the effective distance between samples, and therefore alters the clustering (Figure 6).

3.1.1 Sensitivity to clock rate

An implicit assumption of the SNP method is that each SNP contributes equally towards the SNP distance. This implicitly assumes 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 2: the transmission clustering method approaches the same results as the SNP clustering method as the assumed clock rate is reduced.

3.2 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 – 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. For this kind of data set, 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.

Method	Lambda	Beta	Threshold	Size of largest cluster	Size of second cluster	Number of size 2 clusters	Number of singletons
SNP			S=6	41	8	0	3
SNP			S=5	37	8	2	3
Transmission	0.5	1.2	T=10	41	8	0	3
Transmission	0.5	1.2	T=9	37	8	2	3
Transmission	1.0	1.2	T=7	41	8	0	3
Transmission	1.0	1.2	T=5	41	7	0	4
Transmission	1.0	1.2	T=4	37	6	2	5
Transmission	1.5	1.2	T=5	42	7	0	3
Transmission	1.5	1.2	T=4	41	7	0	4
Transmission	1.5	1.2	T=3	41	6	0	5
Transmission	2.0	1.2	T=3	42	6	0	4
Transmission	2.0	1.2	T=2	37	6	2	5

Table 2: Effect of varying the clock rate using the British Columbia data. This table shows how the clock rate affects the transmission method, 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 is less than or equal to T with a probability of 80%. For $\lambda = 0.5$ the pattern of clusters at $T = 10$ and $T = 9$ is identical to the SNP method at $S = 6$ and $S = 5$ respectively: as the threshold lowers the largest cluster loses 4 cases while the second largest clusters stays the same. Conversely, for $\lambda = 1.0$, the second cluster loses a case whilst the largest cluster remains the same as the threshold is lowered. As the clock rate increases further it can be seen that the pattern of clustering diverges further. The level of β is effectively just a scale factor and does not affect the pattern of clustering.

3.2.1 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 [31] 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 [18] gives a mutation rate of 4.3 SNPs per genome per year when resistance conferring SNPs are included, in contrast to the 0.5 SNPs per genome per year that is typically estimated for TB [5]. 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 (11), five times higher than for the non-resistance conferring sites.

Overall, the number of 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 7. 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.

3.3 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 methods.

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 [32]. Outbreaks are simulated using *TransPhylo* [33], 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* [34]. Sequences are then generated with *phangorn* [35] 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.

We consider clustering cases based on direct transmission, so that two cases are joined in a cluster if one infected the other, and compare clusters with those generated by the SNP method and 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 simply use the variation of information dissimilarity measure given by *clue* [32] 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 3 show that the transmission method is at least as good at identifying direct transmissions within an outbreak as the SNP method, and typically performs at least slightly better. 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

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 3: 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 transmission rate $\beta = 3.0$, as compared to simulated samples connected by direct transmission. The random column shows the dissimilarity obtained for randomly allocated simulated clusters.

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 existing 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 $\lambda = 1.2$ and $\beta = 1.5$, 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 method, and so we have not chosen to perform extensive simulations to compare the methods.

4 Discussion

A fixed number of SNPs can easily translate into different numbers of transmissions depending on other factors, including the substitution process and timing of transmission, alongside factors we have not explicitly modelled (location, social contacts, host risk factors, pathogen factors). We have seen that sampled cases which are relatively close in terms of genetic distance can nevertheless be separated by large distances in time. In this scenario, a simple SNP cut-off will tend to place the 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. This can make phylogenetic inference challenging because the low genetic variation is hard to reconcile with the large time distance. Because of this, in such scenarios 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 drug resistance or SNPs with a different substitution process, variability in the substitution and transmission processes, and it has the scope to handle extensions to include more epidemiological (such as spatial) 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 can be very robust with respect to changes in these variables, particularly so for the transmission rate, suggesting that the clusters obtained are well supported by the data. As discussed in the results section, varying the value of the constant transmission rate does not have a material impact on the clustering because a re-scaling of the cut-off will compensate. The choice of the function $\beta(t)$ is likely to have an impact on results and is therefore worthy of exploration. In particular we would expect the low probability of very quick transmission – as the pathogen numbers are building up in a new host – to have the biggest impact, compared to the use of a constant transmission rate. One approach is to try to obtain estimates using well annotated data where it is known that most cases have been recorded. Alternatively, values can be chosen conservatively where missing data is suspected. 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 (8) in the Methods section. We do not model within-host diversity, though this is relevant to identifying direct transmission events [36]. However, unless within-host diversity is very high, the portion of the total time separating two cases that lies in a single host but contains the MRCA of the two cases is likely to be small.

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 [37] – 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).

When comparing samples genetically, we have been looking at SNPs only, without taking into account information that might be extracted from indels, hyper-variable sites and repeating regions. These and other large scale differences between lineages have been filtered out before conducting our analysis. Such information has the potential to be incorporated into the model and could help establish 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.

Basing the clustering on the number of transmission events has been demonstrated to be an achievable aim and, at least in some circumstances, it has been shown to achieve different subdivisions of the data compared to simple SNP-based methods. The methodology put forward aims to improve the way that WGS and epidemiological data are used to cluster cases into subsets which faithfully correspond to outbreaks of an infectious disease. The methods described here provide the foundations of a theoretical framework on which to make decisions about how clustering should be done, allowing the possibility to develop statistically sound methodologies to replace existing *ad hoc* methods.

5 Software availability

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

Supplementary materials

Application of transmission method to timed trees

We can extend the transmission method by applying it to timed phylogenetic trees. Building a timed phylogenetic tree (using, for example, Beast) allows various elements to be factored in. Differing mutation rates at different sites can be specified or estimated, and as the trees are created in the model the shared evolutionary history of cases is taken into account (in contrast to the pairwise application in the main text). The timings of the branches are obtained from the posterior, resulting in a timed tree where the branch lengths are proportional to time. An advantage of this approach is that the clock rate can be estimated from the data, rather than being fixed, though naturally this requires longitudinal data and sufficient variation.

We apply the transmission-based cut-off to the timed tree. The initial tree is subdivided into a resultant set of output trees by cutting the tree along internal branches which exceed the transmission cut-off criterion. Note that we do not cut terminal branches. The cutting of a branch results in a sub-tree being created from the clade descended from this branch, whilst in the initial tree the branch and clade are replaced by a single tip.

In Figure S1 we illustrate the application of this method for a simulated data set containing 22 samples taken over a 10 year period.

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

[1] H.-A. Hatherell, C. Colijn, H. R. Stagg, C. Jackson, J. R. Winter, and I. Abubakar, "Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review," *BMC Medicine*, vol. 14, no. 1, p. 21, 2016.

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

[3] T. J. Dallman, P. M. Ashton, L. Byrne, N. T. Perry, L. Petrovska, R. Ellis, L. Allison, M. Hanson, A. Holmes, G. J. Gunn, *et al.*, "Applying phylogenomics to understand the emergence of Shiga-toxin-producing Escherichia coli O157: H7 strains causing severe human disease in the UK," *Microbial Genomics*, vol. 1, no. 3, 2015.

[4] S. Octavia, Q. Wang, M. M. Tanaka, S. Kaur, V. Sintchenko, and R. Lan, "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*, vol. 53, no. 4, pp. 1063–1071, 2015.

[5] T. M. Walker, C. L. Ip, R. H. Harrell, J. T. Evans, G. Kapatai, M. J. Dedicoat, D. W. Eyre, D. J. Wilson, P. M. Hawkey, D. W. Crook, *et al.*, "Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study," *The Lancet Infectious Diseases*, vol. 13, no. 2, pp. 137–146, 2013.

[6] J. A. Guerra-Assunção, R. Houben, A. C. Crampin, T. Mzembe, K. Mallard, F. Coll, P. Khan, L. Banda, A. Chiwaya, R. Pereira, *et al.*, "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*, vol. 211, no. 7, pp. 1154–63, 2015.

[7] T. M. Bergholz, A. I. M. Switt, and M. Wiedmann, "Omics approaches in food safety: fulfilling the promise?," *Trends in Microbiology*, vol. 22, no. 5, pp. 275–281, 2014.

[8] T. Azarian, N. F. Maraqa, R. L. Cook, J. A. Johnson, C. Bailey, S. Wheeler, D. Nolan, M. H. Rathore, J. G. Morris Jr, and M. Salemi, "Genomic epidemiology of methicillin-resistant Staphylococcus aureus in a neonatal intensive care unit," *PloS One*, vol. 11, no. 10, p. e0164397, 2016.

[9] T. M. Walker, M. K. Lalor, A. Broda, L. S. Ortega, M. Morgan, L. Parker, S. Churchill, K. Bennett, T. Golubchik, A. P. Giess, *et al.*, "Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007–12, with whole pathogen genome sequences: an observational study," *The Lancet Respiratory Medicine*, vol. 2, no. 4, pp. 285–292, 2014.

[10] J. M. Bryant, S. R. Harris, J. Parkhill, R. Dawson, A. H. Diacon, P. van Helden, A. Pym, A. A. Mahayiddin, C. Chuchottaworn, I. M. Sanne, *et al.*, "Whole-genome sequencing to establish relapse or re-infection with Mycobacterium

tuberculosis: a retrospective observational study," *The Lancet Respiratory Medicine*, vol. 1, no. 10, pp. 786–792, 2013.

[11] T. G. Clark, K. Mallard, F. Coll, M. Preston, S. Assefa, D. Harris, S. Ogwang, F. Mumbowa, B. Kirenga, D. M. O'Sullivan, *et al.*, "Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing," *PLoS One*, vol. 8, no. 12, p. e83012, 2013.

[12] R. S. Lee, N. Radomski, J.-F. Proulx, J. Manry, F. McIntosh, F. Desjardins, H. Soualhine, P. Domenech, M. B. Reed, D. Menzies, *et al.*, "Reemergence and amplification of tuberculosis in the Canadian arctic," *The Journal of Infectious Diseases*, vol. 211, no. 12, pp. 1905–1914, 2015.

[13] A. Roetzer, R. Diel, T. A. Kohl, C. Rückert, U. Nübel, J. Blom, T. Wirth, S. Jaenicke, S. Schuback, S. Rüsch-Gerdes, *et al.*, "Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study," *PLoS Medicine*, vol. 10, no. 2, p. e1001387, 2013.

[14] C. Yang, T. Luo, X. Shen, J. Wu, M. Gan, P. Xu, Z. Wu, S. Lin, J. Tian, Q. Liu, *et al.*, "Transmission of multidrug-resistant Mycobacterium tuberculosis in Shanghai, China: a retrospective observational study using whole-genome sequencing and epidemiological investigation," *The Lancet Infectious Diseases*, vol. 17, no. 3, pp. 275–284, 2017.

[15] C.-H. Kuo and H. Ochman, "Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria," *Biology Direct*, vol. 4, no. 1, p. 35, 2009.

[16] P. S. Novichkov, M. V. Omelchenko, M. S. Gelfand, A. A. Mironov, Y. I. Wolf, and E. V. Koonin, "Genome-wide molecular clock and horizontal gene transfer in bacterial evolution," *Journal of Bacteriology*, vol. 186, no. 19, pp. 6575–6585, 2004.

[17] J. E. Barrick and R. E. Lenski, "Genome dynamics during experimental evolution," *Nature Reviews. Genetics*, vol. 14, no. 12, p. 827, 2013.

[18] V. Eldholm, G. Norheim, B. von der Lippe, W. Kinander, U. R. Dahle, D. A. Caugant, T. Mannsåker, A. T. Mengshoel, A. M. Dyrhol-Riise, and F. Balloux, "Evolution of extensively drug-resistant Mycobacterium tuberculosis from a susceptible ancestor in a single patient," *Genome Biology*, vol. 15, no. 11, p. 490, 2014.

[19] V. Korhonen, P. Smit, M. Haanperä, N. Casali, P. Ruutu, T. Vasankari, and H. Soini, "Whole genome analysis of Mycobacterium tuberculosis isolates from recurrent episodes of tuberculosis, Finland, 1995–2013," *Clinical Microbiology and Infection*, vol. 22, no. 6, pp. 549–554, 2016.

[20] S. P. Velsko, J. E. Allen, and C. T. Cunningham, "A statistical framework for microbial source attribution," tech. rep., Lawrence Livermore National Laboratory (LLNL), Livermore, CA, 2009.

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

[22] R. G. Gallagher, *Stochastic processes: theory for applications*. Cambridge University Press, 2013.

[23] Å. Svensson, "A note on generation times in epidemic models," *Mathematical biosciences*, vol. 208, no. 1, pp. 300–311, 2007.

[24] A. C. Cameron and P. K. Trivedi, *Regression analysis of count data*, vol. 53. Cambridge university press, 2013.

[25] O. Diekmann, H. Heesterbeek, and T. Britton, *Mathematical tools for understanding infectious disease dynamics*. Princeton University Press, 2012.

[26] V. Eldholm, J. Monteserin, A. Rieux, B. Lopez, B. Sobkowiak, V. Ritacco, and F. Balloux, "Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain," *Nature Communications*, vol. 6, p. 7119, 2015.

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

[28] N. Casali, V. Nikolayevskyy, Y. Balabanova, S. R. Harris, O. Ignatyeva, I. Kontsevaya, J. Corander, J. Bryant, J. Parkhill, S. Nejentsev, *et al.*, "Evolution and transmission of drug-resistant tuberculosis in a Russian population," *Nature Genetics*, vol. 46, no. 3, p. 279, 2014.

[29] M. Merker, C. Blin, S. Mona, N. Duforet-Frebourg, S. Lecher, E. Willery, M. G. Blum, S. Rüsch-Gerdes, I. Mokrousov, E. Aleksic, *et al.*, "Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage," *Nature Genetics*, vol. 47, no. 3, p. 242, 2015.

[30] J. M. Bryant, A. C. Schürch, H. van Deutekom, S. R. Harris, J. L. de Beer, V. de Jager, K. Kremer, S. A. van Hijum, R. J. Siezen, M. Borgdorff, *et al.*, "Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data," *BMC Infectious Diseases*, vol. 13, no. 1, p. 110, 2013.

[31] S. Feuerriegel, V. Schleusener, P. Beckert, T. A. Kohl, P. Miotto, D. M. Cirillo, A. M. Cabibbe, S. Niemann, and K. Fellenberg, "PhyResSE: a web tool delineating Mycobacterium tuberculosis antibiotic resistance and lineage from whole-genome sequencing data," *Journal of Clinical Microbiology*, vol. 53, no. 6, pp. 1908–1914, 2015.

[32] M. Meilă, "Comparing clusterings—an information based distance," *Journal of Multivariate Analysis*, vol. 98, no. 5, pp. 873–895, 2007.

[33] X. Didelot, C. Fraser, J. Gardy, and C. Colijn, "Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks," *Molecular Biology and Evolution*, vol. 34, no. 4, pp. 997–1007, 2017.

[34] M. Kendall, M. Boyd, and C. Colijn, "phyloTop." https://CRAN.R-project.org/package=phyloTop, 2016.

[35] K. P. Schliep, "phangorn: phylogenetic analysis in R," *Bioinformatics*, vol. 27, no. 4, p. 592, 2011.

[36] C. J. Worby, M. Lipsitch, and W. P. Hanage, "Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data," *PLoS Computational Biology*, vol. 10, no. 3, p. e1003549, 2014.

[37] L. S. Katz, A. Petkau, J. Beaulaurier, S. Tyler, E. S. Antonova, M. A. Turnsek, Y. Guo, S. Wang, E. E. Paxinos, F. Orata, *et al.*, "Evolutionary dynamics of Vibrio cholerae O1 following a single-source introduction to Haiti," *MBio*, vol. 4, no. 4, pp. e00398–13, 2013.

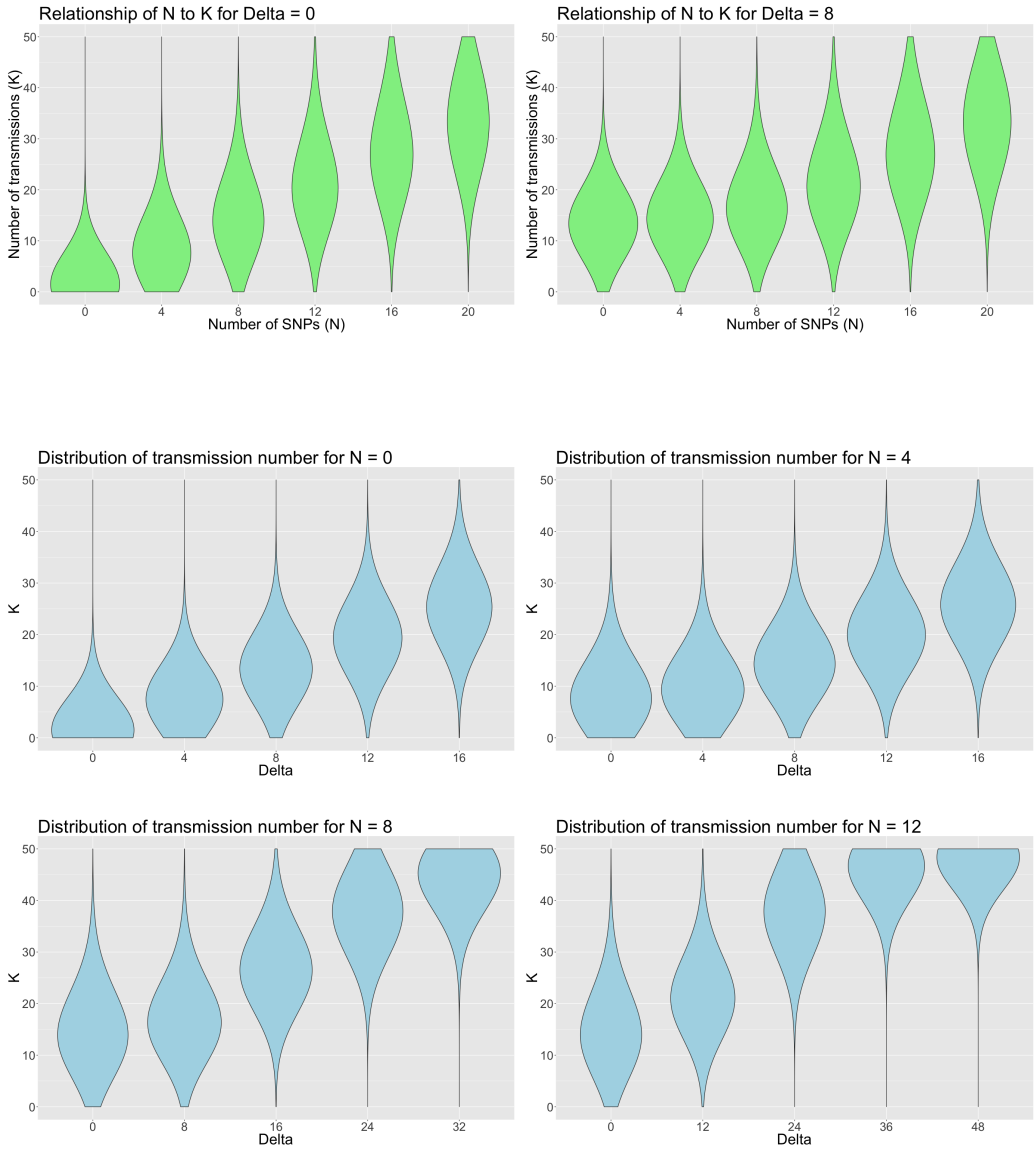

Figure 2: Probability density for the number of transmissions, with clock rate $\lambda = 0.9$ and transmission rate $\beta = 1.5$. The upper two panels show the densities for delta values of 0 and 4 years for a range of SNP distance between 0 and 20. The lower four panels show the densities for 0, 4, 8 and 12 SNP distance respectively, for $\delta = 0, 4, 8, 12, 16$ years.

INPUTS

Sample dates

A	1 Jan 2018
B	1 Jan 2014
C	1 Jan 2016
D	1 Jan 2016

SNP distance matrix

	A	B	C	D
A	0	5	7	7
B	5	0	8	8
C	7	8	0	6
D	7	8	6	0

OUTPUTS SNP method

Cut-off	Clusters		
5	AB	C	D
6	AB	CD	
7	ABCD		

OUTPUTS transmission method

T cut-off	Clusters		
7	CD	A	B
8	ACD	B	
9	ABCD		

Figure 3: Model inputs and outputs for a simple example data set. For the SNP method, samples are clustered together where the SNP distance is less than or equal to the cut-off. For the transmission method, samples are clustered together where the implied number of transmissions is less than or equal to the T cut-off with a probability of 80% with clock rate $\lambda = 1.5$ and transmission rate $\beta = 2.3$.

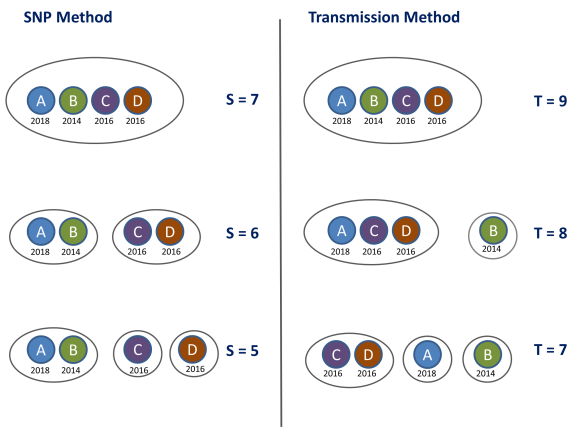

Figure 4: Clustering on the simple example data set. The left hand panel shows the clusters obtained by applying the SNP method with three different thresholds, with the cut-off level denoted by S; samples are clustered together where the SNP distance is less than or equal to S. The right hand panel shows the clustering obtained by applying the transmission method, with the cut-off level denoted by T; samples are clustered together where the implied number of transmissions is less than or equal to T with a probability of 80%, with clock rate $\lambda = 1.5$ and transmission rate $\beta = 2.3$.

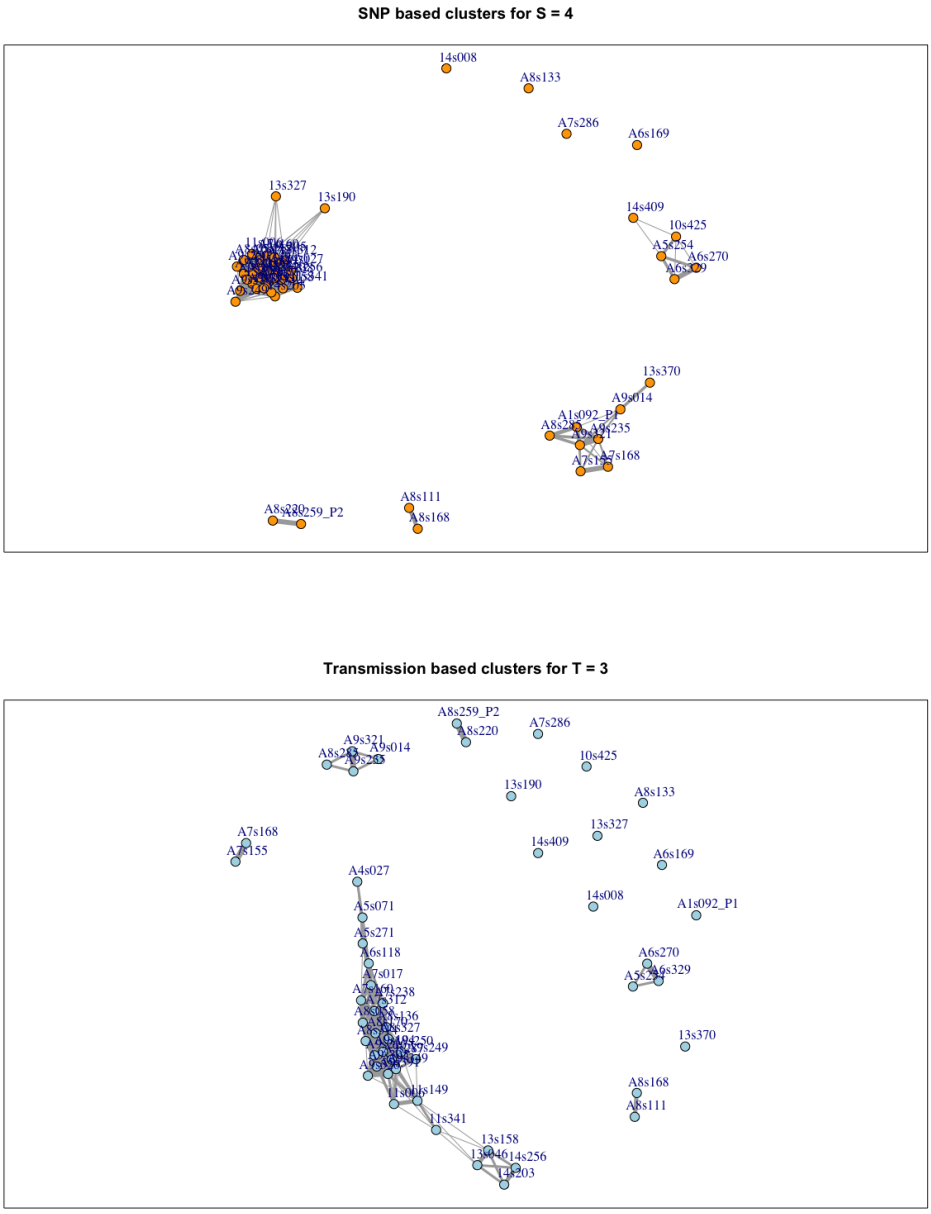

Figure 5: Two views of the same British Columbia TB data illustrating the contrasting effect of implementing the SNP and transmission methods, and showing estimates of how close individual cases are to each other. In the top figure, edges between nodes indicate that cases are within 4 SNPs of each other. In the bottom figure, edges indicate that cases are 80% likely to be within 3 transmission events of each other, given a clock rate $\lambda = 1.5$ and transmission rate $\beta = 2.0$. 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.

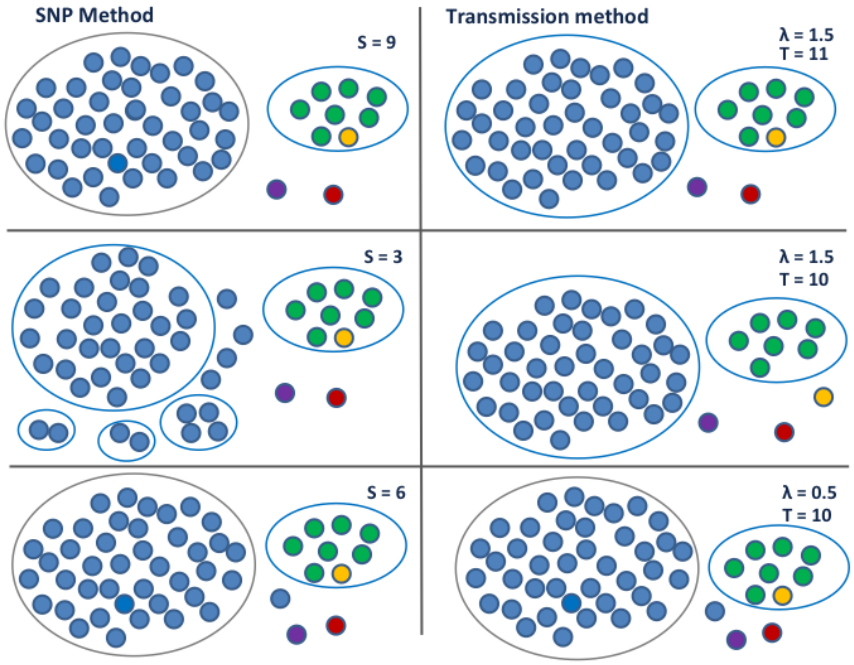

Figure 6: Clustering on the British Columbia data set. The left hand side shows the clusters obtained by applying the SNP method with three different thresholds, with the cut-off level denoted by S. The largest cluster breaks up as the level is lowered whilst the size 8 cluster remains intact. The right hand side shows the clustering obtained by applying the transmission method; samples are clustered together where the implied number of transmissions is less than or equal to T with a probability of 80%. As shown in the top two thirds, with clock rate $\lambda = 1.5$ and transmission rate $\beta = 2.0$, 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 root of the mini-tree 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 cutoff. This is shown in the lower third, where the clock rate $\lambda = 1.5$ and the transmission rate is $\beta = 1.2$.

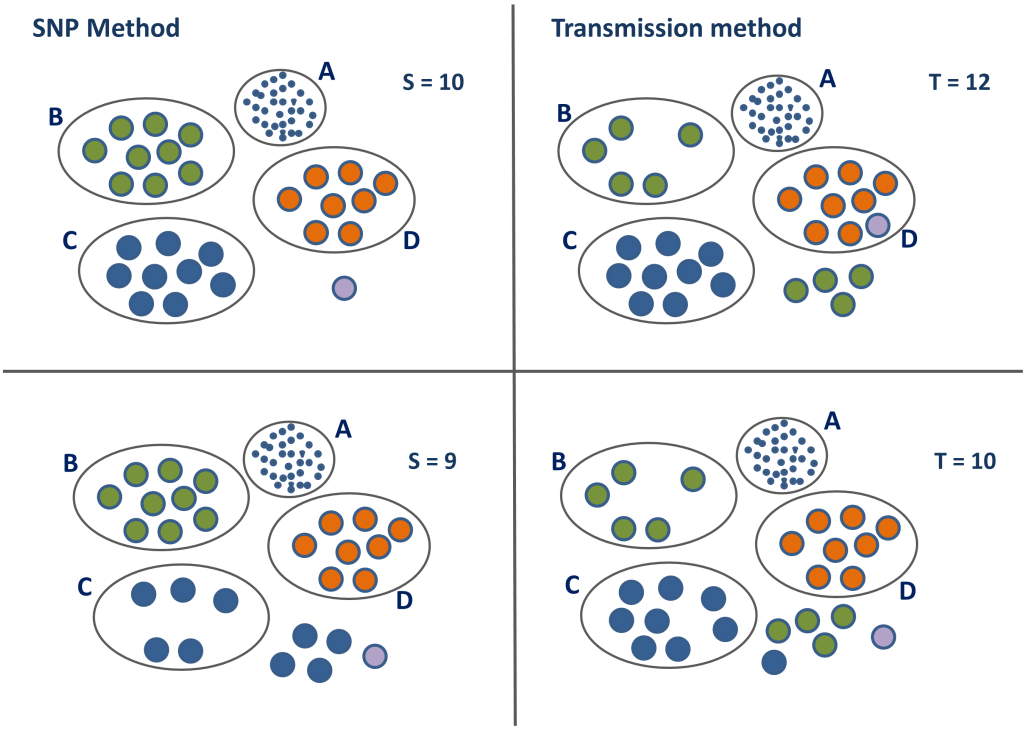

Figure 7: Some medium sized clusters illustrating the effect of accounting for resistance conferring SNPs in the transmission method. Clusters B, C and D are the second to fourth largest clusters in the Moldova data using the SNP method. The largest cluster A, with 93 members for $S = 10$ is shown for 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, with the cut-off level denoted by T; samples are clustered together where the implied number of transmissions is less than or equal to T with a probability of 80%, with clock rate $\lambda = 1.5$ and transmission rate $\beta = 2.0$. Date information is not shown – for this data set, the sample dates have low variation and do not significantly impact the results.

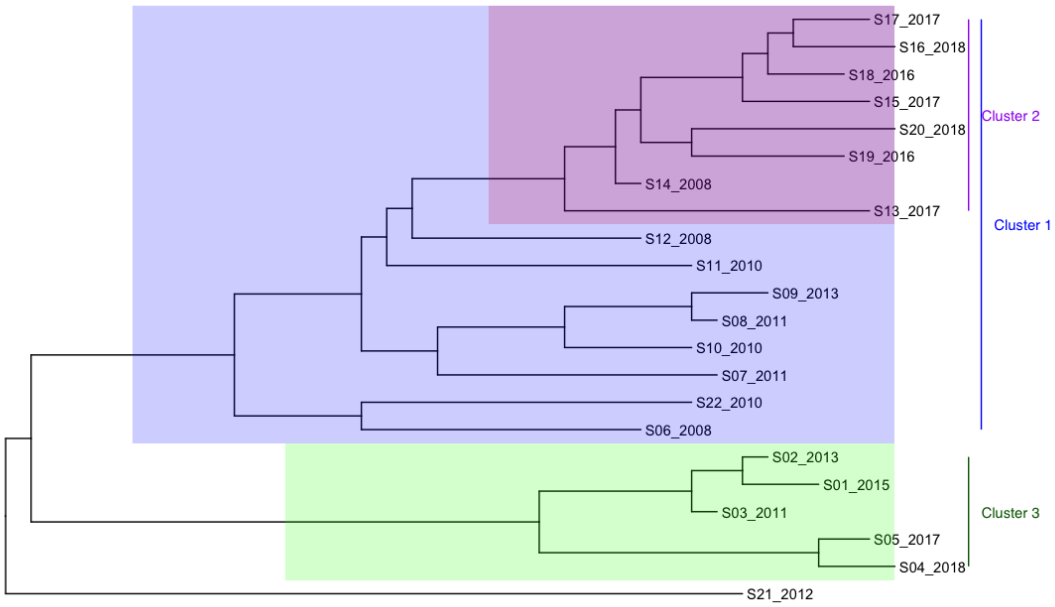

Figure S1: Application of transmission method to a simulated timed tree. Branches are cut where the shading colour changes, which is where there is a greater than 80% probability that more than 10 transmissions have occurred along that branch. Three sub-trees are created, corresponding to the identified clusters 1, 2 and 3. Tip labels are suffixed with the year of sampling.

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

Table S1: Supporting information for Figure 6. 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.