Phylogenetic tree shapes resolve disease transmission patterns

Caroline Colijn
Department of Mathematics, Imperial College London
180 Queen's Gate London SW7 2AZ
c.colijn@imperial.ac.uk; +44 207 594 2647

Jennifer Gardy
Communicable Disease Prevention and Control Services
British Columbia Centre for Disease Control
655 West 12th Avenue, Vancouver, BC, Canada, V5Z 4R4
Jennifer.Gardy@bccdc.ca

Abstract

Whole genome sequencing is becoming popular as a tool for understanding outbreaks of communicable diseases, with phylogenetic trees being used to identify individual transmission events or to characterize outbreak-level overall transmission dynamics. Existing methods to infer transmission dynamics from sequence data rely on well-characterised infectious periods, epidemiological and clinical meta-data which may not always be available, and typically require computationally intensive analysis focussing on the branch lengths in phylogenetic trees. We sought to determine whether the topological structures of phylogenetic trees contain signatures of the overall transmission patterns underyling an outbreak. Here we use simulated outbreaks to train and then test computational classifiers. We test the method on data from two real-world outbreaks. We find that different transmission patterns result in quantitatively different phylogenetic tree shapes. We describe five topological features that summarize a phylogeny's structure and find that computational classifiers based on these are capable of predicting an outbreak's transmission dynamics. The method is robust to variations in the transmission parameters and network types, and recapitulates known epidemiology of previously characterized real-world outbreaks. We conclude that there are simple structural properties of phylogenetic trees which, when combined, can distinguish communicable disease outbreaks with a super-spreader, homogeneous transmission, and chains of transmission. This is possible using genome data alone, and can be done during an outbreak. We discuss the implications for management of outbreaks.

1 Introduction

Whole-genome sequence data contain rich information about a pathogen population from which several evolutionary parameters and events of interest can be inferred. When the population in question comprises pathogen isolates drawn from an outbreak or epidemic of an infectious disease, these inferences may be of epidemiological importance, able to provide actionable insights into disease transmission. Indeed, since 2010, several groups have demonstrated the utility of genome data for revealing pathogen transmission dynamics and identifying individual transmission events in outbreaks [22, 13, 27, 7, 24, 12, 28, 11, 4], with the resulting data now being used to inform public health's outbreak management and prevention strategies. To date, these reconstructions have relied heavily on interpreting genomic data in the context of available epidemiological data, drawing conclusions about transmission events only when they are supported by both sequence data and plausible epidemiological linkages collected through field investigation and patient interviews.

Given the rapidly growing interest in this new field of genomic epidemiology, several recent studies have explored whether transmission events and patterns can be deduced from genomic data alone. These methods rely primarily upon interpreting a phylogenetic tree's branch lengths. Phylogenies derived from whole-genome sequence data can be compared to theoretical models describing how a tree should look under particular processes - for example, predicted branch lengths from sequences modeled using birth-death processes can be compared to viral sequence data to explore transmission patterns [23, 22, 14, 21]. Tools from coalescent theory have also been adapted to the outbreak setting, taking into account the constraints person-to-person transmission puts on pathogen reproduction [25, 26]. These approaches are powerful, but are highly computationally intensive and do not consider another source of information within a phylogeny - *tree shape*.

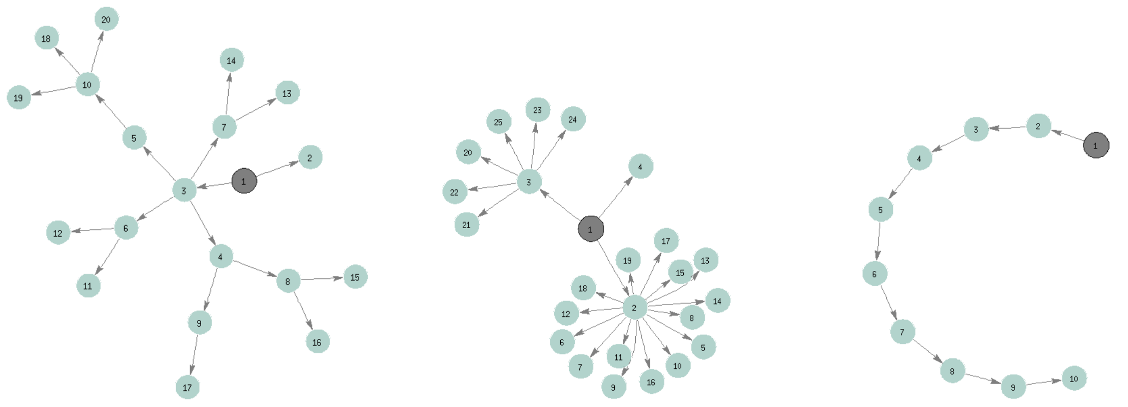
The number of different phylogenetic tree shapes on n leaves is a combinatorially exploding function of n (there are $(2n-3)(2n-5)(2n-7)...(5)(3)(1)$ rooted labelled phylogenetic trees, or approximately 10^{184} trees on 100 tips, compared to approximately 10^{80} atoms in the universe). For the increasingly large outbreak genome datasets being obtained and analysed (390 [27], 616 [3] and recently 1000 [1] bacterial genomes), the numbers of possible tree shapes are effectively infinite. It is therefore likely that in addition to branch lengths, tree shape might also contain substantial information about the underlying transmission network in an outbreak or epidemic dataset. There are already indications that tree shape contains useful information about the evolution of viral pathogens [14, 20, 18, 5], but to date we do not have methods to exploit tree shape in an analysis of pathogen transmission dynamics, built upon simulated data and validated using real-world outbreak data.

Host contact network structure is one of the most profound influences on the dynamics of an outbreak or epidemic [1], and outbreak management and control strategies depend heavily upon the type of transmission patterns driving an outbreak. It is reasonable to expect that pathogen genomes spreading over different contact network structures - chains, homogenous networks, or networks containing super-spreaders, as illustrated in Figure 1 - would accrue mutations in different patterns, leading to observably different phylogenetic tree shapes. We therefore characterized the structural features of phylogenetic trees arising from the simulated evolution of a bacterial genome as it spreads over multiple types of contact network. We found several simple topological properties of phylogenetic trees that, when combined, can be used to classify trees according to whether the underlying process is chain-like, homogenous, or super-spreading, demonstrating that phylogenetic tree structure is as informative as branch lengths with respect to transmission dynamics. We use these properties as the basis for a computational classifer, which we then use to classify real-world outbreaks. We find that the computational predictions of each outbreak's overall transmission dynamics are consistent with known epidemiology.

2 Materials and Methods

Transmission model

We simulated disease transmission networks with three different underlying transmission patterns: homogeneous transmission, transmission with a super-spreader, and chains of transmission. Each simulation started with a single infectious host who infects a random number of secondary cases over his or her infectious period; each secondary case infects others, and so on,



(a) Homogeneous transmission (b) Super-spreader transmission (c) Chain of transmission

Figure 1: Schematic illustration of different kinds of transmission networks. The index case is marked in grey.

until the desired maximum number of cases is reached. The models share two key parameters: a transmission rate β and a duration of infection parameter D. Our baseline values are $\beta = 0.43$ per month and $D = 3$ months, reflecting a basic reproduction number of 1.3. This is also the mean number of secondary infections for each infectious case.

The homogeneous transmission model assigns each infectious host a number of secondary infections drawn from a Poisson distribution with parameter $R_0 = \beta D$. New infections are seeded uniformly in time over the host's infectious period. In the super-spreader model, one host (at random in the first 5 hosts) seeds 7-24 new infections (uniformly at random), and all other hosts are as in the homogeneous transmission model. In the chain-of-transmission model, almost all hosts infect precisely one other individual. However, 2 (with probability 2/3) or 3 (with probably 1/3) of the hosts infect two other individuals, so that the transmission tree consists of several chains of transmission randomly joined together.

Durations of infection are drawn from a Γ distribution with a shape parameter of 1.5 and a scale parameter of $D/1.5$. To reflect transmission of a chronically-infecting pathogen, such as *Mycobacterium tuberculosis*, cases were infectious for between 2 and 14 months with an average specified by D. The mean infectious period was 4.3 months; a histogram is shown in Figure S2. We simulated 1000 outbreaks containing a super-spreader, 1000 with homogeneous transmission and 1000 chain-like outbreaks. These used a fixed parameter set; we also performed a sensitivity analysis using alternative parameters. To ensure that the size of the outbreak did not affect the tree shape and classification, we simulated outbreaks with 32 hosts - a similar size as the real-world outbreaks we later investigated.

The first infected host in each transmission network was seeded with an initial random genome sequence 1000 base pairs in length. When this individual infected a secondary case, the transmitted sequence was mutated and the number of mutations reflected the time elapsed between the seed time and the transmission time. Subsequent transmission events followed the same rule. Seqeunces were sampled at the end of the infection's duration. Further detail of the genome simulation process is provided in the Supplementary Information.

Genealogies and phylogenies from the process

We extracted the true genealogical relationships as a full rooted binary tree (a "phylogeny"), with tips corresponding to hosts and internal nodes corresponding to transmission events among the hosts, as follows. The outbreak simulations create lists of who infected whom and at what time. Each host also has a recovery time. We sort the times of all of the infection events, and proceed in reverse order. The last infection event must correspond to a "cherry", ie it must have two tip descendants, one corresponding to the infecting host and one to the infectee. For all other infection events proceeding in reverse order through the transmission, we create an internal node, and determine its descendants by determining whether the infector and the infectee went on to infect anyone else subsequently. If not, then the node's descendants are the infector and infectee at the time of sampling. If so, then the descendant represents the infector or infectee at the time of their next transmission. The tree is rooted at the first infection event. Branch lengths correspond to the times between infection events or, for tips, the time between the infection event and the time of sampling.

In the main text, we use the true genealogical relationships among the hosts in our outbreak, extracted from the simulations - this reduces phylogenetic noise and it allows us to compare the resulting trees to 100 samples of the BEAST posterior timed phylogenies derived from WGS data from the two real-world outbreaks. To determine how sensitive our approach is to phylogenetic noise, we also classified the outbreaks using neighbour-joining phylogenies derived from simulated gene sequences (Supplementary Information).

Topological summaries of trees Five summary metrics were used to summarise the topology of the trees.

1. **Colless imbalance**. The Colless imbalance is defined as $\frac{1}{(n-1)(n-2)} \sum_{i=1}^{n-1} |T_{ri} - T_{li}|$ where n is the number of tips and T_{ri} and T_{li} are the number of tips descending from the left and right sides at internal node i. It is a normalised measure of the asymmetry of a rooted full binary tree, with a completely asymmetric tree having imbalance of 1 and a symmetric tree having an imbalance of 0 [8].

2. **Ladders** We define the 'ladder length' to be the maximum number of connected internal nodes with a single leaf descendant, and we divide it by the number of leaves in the tree. This measure is not unrelated to tree imbalance but is more local - a long ladder motif may occur in a tree that is otherwise quite balanced. For this reason, ladder length may detect trees in which there has been differential lineage splitting in some clades or lineages but where this occurred too locally or in clades that are too small to have affected traditional approaches to characterising rapid expansion in some lineages. Furthermore, traditional ways of detecting positive selection may not be appropriate in this context because the super-spreader, if present, does not pass any advantageous property to descendant infections.

3. **Maximum width; Maximum width over maximum depth**. The *depth* of a node in a tree is the number of edges between that node and the tree's root. The *width* of a tree at a depth d is defined as the number of nodes with depth d. We calculated the maximum width of each tree divided by its maximum depth ($\max d$, the maximum depth of any leaf in the tree).

4. **Maximum difference in widths** We compared $\Delta w = \max_i \{|w(d_i) - w(d_{i-1})|\}$ in the trees. This summary reflects the maximum absolute difference in widths from one depth to the next, over all depths d_i in the tree.

Outbreak classification We trained a k-nearest-neighbour classifier using matlab's ClassificationKNN.fit function with a Minkowski distance, inverse distance weighting and 100 neighbours. We evaluated the classifier using its resubstitution loss and its performance under cross-validation. We performed cross-validation using crossval in matlab. KNN classification was performed on 1000 trees of each type (homogeneous transmission, super-spreaders, chains).

We used a support vector machine (SVM) to resolve differences between homogeneous transmission vs super-spreader networks. SVMs were constructed using the svmtrain method in matlab with a quadratic kernel function.The training data x_i were the five summary metrics for 300 trees derived from each process. svmtrain returned a set of 128 support vectors s_i and a bias, $b = -0.18$. All training data were from simulations with the baseline set of parameters.

The SVM was tested on the remaining trees using matlab's svmclassify, which computes

$$\text{sign}(y) = \text{sign}(\sum_i \alpha_i k(x_i, x) + b)$$

where α_i are weights, x_i are the support vectors, x is the input to be classified, k is the kernel function and b is the bias. These tests were done separately on the different groups of simulated trees. The svmclassify function was modified to return y (i.e. the degree to which an outbreak could be considered super-spreading) rather than only the sign of y (a binary prediction). Because SVMs are binary classifiers, their quality can be assessed using a receiver operator characteristic (ROC) curve. See [2] for a full discussion of support vector machines and classification.

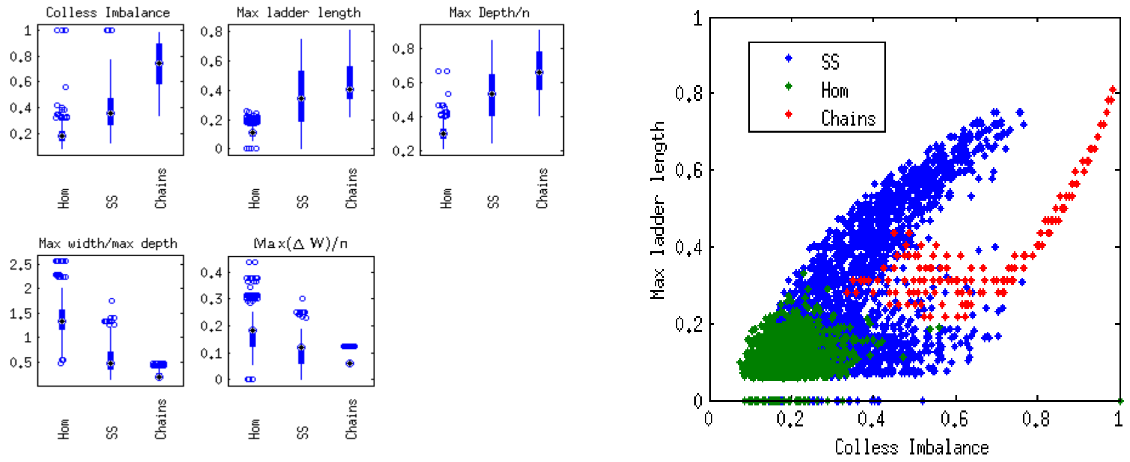
Sensitivity analysis To determine whether the classifier is robust to different choices of model parameters and to sampling, we simulated three groups of 500 homogeneous and super-spreader outbreaks with (i) randomly selected parameters, (ii) a random sampling density, and (iii) both random parameters and random sampling. Group (i) had randomized parameters in which β/D was uniformly distributed between 1.25 and 2.5. Group (ii) had fixed parameters, but the number of cases varied uniformly between 100 and 150, and we sampled only 33 of those cases. The third group had both randomized parameters and random sampling. To determine whether the classifier is relevant to different *kinds* of models, we applied it to simulated phylogenies described in Robinson et al [20]. In that work, dynamic networks of sexual contacts were created based on random graphs with a Poisson distribution, and with a distribution of contacts derived from the National Survey on Sexual Attitudes and Lifestyles (NATSAL) [10]. See Supplementary material for further details.

Classification of outbreaks from published genomic data We used the classifier on phylogenetic trees derived from two real-world tuberculosis outbreak datasets. Outbreak A was previously published [6] and is available in the NCBI Sequence Read Archive under the accession number SRP002589. This dataset comprises 32 *M. tuberculosis* isolates collected in British Columbia over the period 1995-2008 and was sequenced using paired-end 50bp reads on the Illumina Genome Analyzer II. Outbreak B comprises 33 *M. tuberculosis* isolates collected in British Columbia over the period 2006-2011, and was sequenced using paired-end 75bp reads on the Illumina HiSeq. The outbreak, sequences and SNPs are presented in [4].

For both datasets, reads were aligned against the reference genome M. tuberculosis CDC1551 (NC002755) using BWA [16]. Single nucleotide variants were identified using samtools mpileup [17] and were filtered to remove any variant positions within 250bp of each other and any positions for which at least one isolate did not have a genotype quality score of 222. The remaining variants were manually reviewed for accuracy and were used to construct a phylogenetic tree for each outbreak as described above. In the main text we apply the classification methods to 100 samples from the BEAST posterior timed phylogenies estimated from WGS data. In the Supplementary Information we train the classifiers on neighbour-joining phylogenies from simulations, and apply them to neighbour-joining phylogenies from the outbreaks with identical results.

3 Results

Different transmission networks result in quantitatively different tree shapes
To determine whether tree shapes captured information about the underlying disease transmission patterns within an outbreak, we simulated evolution of a bacterial genome over three types of outbreak contact network - homogenous, super-spreading, and chain - and summarized the



(a) Boxplots of five topological summary metrics

(b) Ladder length versus imbalance. Colors correspond to different underlying transmission patterns.

Figure 2: Distribution of five simple summary measures of tree topology

resulting phylogenies with five metrics describing tree shape. Figure 2 shows the distributions of these metrics across the three types of outbreaks, revealing clear differences in tree topology depending on the underlying host contact network. Super-spreader networks gave rise to phylogenies with higher Colless imbalance, longer ladder patterns, lower Δw, and deeper trees than transmission networks with a homogeneous distribution of contacts. Trees derived from chain-like networks were less variable, deeper, more imbalanced, and narrower than the other trees. Other topological summary metrics considered did not resolve the three outbreak types as fully (Supplementary Information).

Topological metrics can be used to computationally classify outbreaks
To evaluate whether the five topological summary metrics could realiably and automatically differentiate between the three types of outbreaks, we trained a series of computational classifiers on the simulated datasets. We first trained a K-nearest-neighbour (KNN) classifier using the five metrics to discern which combinations of features correspond to phylogenies derived from the three underlying transmission processes. The KNN classifier correctly identified nearly all of the underlying transmission dynamics (see Table 1 (a)); overall, the fraction of mis-classification was 0.0036 due to resubstitution loss. To evaluate classifier consistency, we performed 10-fold cross-validation, which resulted in an average error of 7%. In other words, when dividing the test sample into different groups, re-training the model, and then testing on data not used in training, the classifier is wrong on average 7% of the time.

Support vector machine improves classification accuracy
To better resolve the separation between super-spreader-type outbreaks and those with homogeneous transmission, we trained a support vector machine (SVM) classifier to distinguish between those two types of outbreaks alone. Figure 3(a) shows the receiver-operator characteristic curve (ROC) for the SVM classification (ROC curves are the most natural way to assess the quality of a binary classifier). The area under the curve (AUC) is 0.97, reflecting a very good classifier performance; the theoretical maximum AUC is 1, and 0.5 corresponds to random guessing.

Outbreak classification is robust to variable parameters and model choice, but not to sampling

Truth	Hom	SS	Ch
Classification			
Hom	995	5	0
SS	5	995	0
Ch	0	0	500

(a) Baseline

Truth	Hom	SS	Ch
Classification			
Hom	455	37	0
SS	45	462	0
Ch	0	1	500

(b) Varied transmission

Truth	Hom	SS	Ch
Classification			
Hom	447	159	0
SS	52	296	0
Ch	1	45	500

(c) Sampling

Truth	Hom	SS	Ch
Classification			
Hom	454	184	0
SS	46	262	0
Ch	0	54	500

(d) Varied transmission and sampling

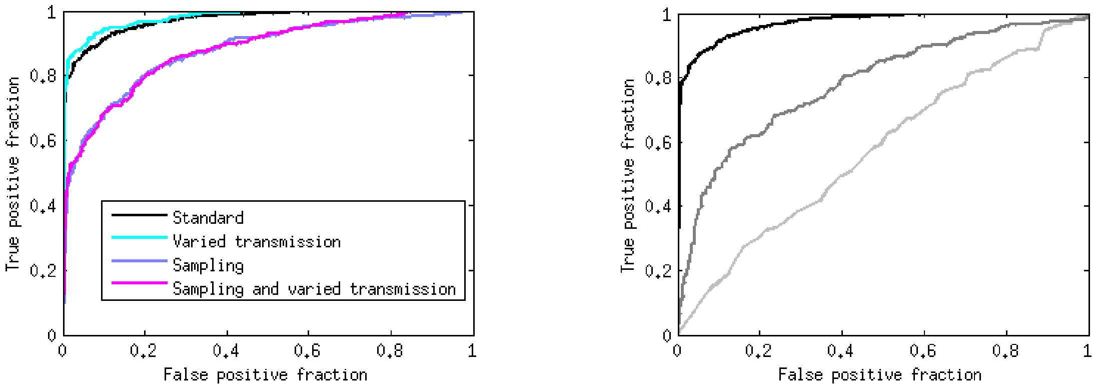
Truth	Hom	SS	Ch
Classification			
Hom	290	160	32
SS	177	326	0
Ch	33	14	468

(e) 10 isolates

Truth	Hom	SS	Ch
Classification			
Hom	447	45	0
SS	53	454	0
Ch	0	1	500

(f) 20 isolates

Table 1: K-nearest neighbour classification results matrix. True outbreak type is shown in columns; predicted outbreak types is shown across rows. Table (a) shows the results for KNN classification with baseline parameters, (b) with variable transmission patterns, (c) with variable sampling, (d) with both variable transmission parameters and sampling, (e) for classification based only on the first 10 isolates sampled in an outbreak, (f) with only the first 20 isolates.



(a) ROC with sensitivity analysis groups

(b) ROC with early isolates

Figure 3: Receiver-operator characteristic for the SVM classifier based on five summary metrics. ROC curves are a visual way to assess the classifier's quality – a perfect classifier will obtain all the true positives and will have no false positives, giving an AUC of 1. Guessing yields an AUC of 0.5. In Figure 3(a), different lines correspond to the different groups of simulations in the SVM sensitivity analysis. Figure 3(b) shows the SVM classifier's performance when only the earliest outbreak isolates are sampled.

To explore how robustly phylogenetic structure captures variation in transmission processes, we performed sensitivity analyses in which we explored the effect of varying the transmission parameters β/D, sampling, and both the parameters and sampling together.

Using the KNN classifier applied to the three outbreak types, we found that the overall classifier error increased to 10% when the transmission rate varied up to a factor of 2 (Table 1(b)). The effect of reduced sampling density was greater, increasing the KNN classifier error to 26%. Varying both the sampling and the parameters resulted in an error rate of 28% (Table 1(c,d)) We also evaluated the sensitivity of SVM classification to different transmission model parameters by training and testing an SVM on a further 500 simulated super-spreading and homogenous networks, with variable transmission parameters β/D. As with the baseline parameter networks, the SVM returned an AUC of 0.97 for the variable β/D groups. However, the SVM's performance declined with decreased sampling density (AUC of 0.88) and decreased sampling with variable transmission parameters (Figure 3(a)) Notably, the decline in performance was less with the SVM method than the KNN method - 12% versus 28%.

We tested the SVM classifier to determine whether it could distinguish between phylogenetic trees derived from simulated sequence transmission on very different contact networks, namely dynamical models of sexual contact networks over a 5-year simulated time period [20]. The performance was excellent when sampling was done over time, such that cases infected early in the simulation were likely to be sampled. When sampling was done at one time, years after seeding the simulated infection, neither classifier detected differences between the two types of contact network. Details are presented in the Supplementary Information.

Outbreak classification is possible using early isolates only

To determine whether classification of an outbreak is possible early in an outbreak - information that could potentially inform real-time deployment of a specific public health response - we evaluated the KNN and SVM classifiers' performance when only the first 10 and first 20 genomes of the outbreak were sampled. The KNN had an average error of 28% using a phylogeny built on the first 10 isolates and an avergae error of 7% after 20 isolates had been sampled. The SVM had AUC values of 0.53 and 0.71 after 10 and 20 isolates were detected, respectively (see Table 1 and Figure 3). These data suggest that reasonable classification of an outbreak's transmission dynamics is possible after at early points within the outbreak.

Topological metric-based classification recapitulates known epidemiogology of real-world outbreaks

Finally, to evaluate the classifiers' performance on real-world outbreaks with known epidemiology, we applied the classifier to genome sequence data from two tuberculosis outbreaks whose underlying transmission dynamics have been described through comprehensive field and genomic epidemiology. Outbreak A [6]was reported to arise from super-spreading activity, while Outbreak B displayed multiple waves of transmission, resulting in a somewhat more homogenous network.

We found that our classification results agreed with the empirical characterisations of the two outbreaks' underlying transmission dynamics. In the KNN classification, Outbreak A was grouped with super-spreader outbreaks most often (54%), with 45% of the posterior trees grouping with homogeneous outbreaks and only 1% with chains. 83% of the trees from Outbreak B were classed as homogeneous, with the other 17% classed with super-spreader outbreaks. The SVM classifier returned a mean classification over the BEAST posterior trees of $y = 0.09(-1.7, 2.7)$ for Outbreak A and $1.5(-0.5, 4.6)$ for Outbreak B (numbers in parenthesis are 95% CIs). The optimal cutoff was 0.3 (based on matlab's perfcurve function), indicating that posterior trees from Outbreak A group more with super-spreader outbreaks and those from Outbreak B group with homogeneous transmission.

4 Discussion

We have found that there are simple topological properties of phylogenetic trees which, when combined, are informative as to the underlying transmission patterns at work in an outbreak. Tree structures can be used as the basis of a classification system, able to describe an outbreak's dynamics from genomic data alone. These topological signatures are robust to variation in the transmissibility, and to the nature and structure of the model, but sampling has a detrimental effect on the strength of the signal. Signs of the underlying transmission dynamics are present within the first 20 genomes sampled from an outbreak, and the classifiers are able to recapitulate known, real-world epidemiology from actual outbreak datasets.

The relationship between host contact heterogeneity and pathogen phylogenies is complex. In large datasets, phylogenetic branch lengths can reveal heterogeneous contact numbers [23], but distributions of branch lengths are not a suitable tool for small outbreaks of a chronically-infecting and slowly mutating organism like TB. Early work made the assumption that heterogeneous contact numbers would yield heterogeneous cluster sizes in viral phylogenies [15]. But cluster sizes also depend on the pathogen population dynamics [20] and the epidemic dynamics [5]. The relationship between heterogeneous contact numbers and tree imbalance [14] is not robust to the dynamics of a contact network [20], sampling [20, 5] or the epidemic model used [5]. It is clear from this body of work that increased heterogeneity in contact numbers will not always lead to a simple increase or decrease of some measure, like imbalance, of tree structure. However, we have found that in small outbreaks, several simple topological features, taken together, can distinguish between outbreaks with high heterogeneity (a super-spreader) and low heterogeneity.

In any modelling endeavor, when a model reproduces features of real data – whether those are tree structures, branch lengths, or other data such as prevalence and incidence of an infection, locations of cases and so on – it remains possible that there are processes not included in the model that are the real origin of the observations. When we use models to interpret data, we use formal or informal priors to weigh the likelihoods of the assumptions behind the model compared to other processes that could drive the same phenomena. Here, one aspect of the complex relationship between contact heterogeneity and phylogeny structure is illustrated by the fact that genealogies from a long chain of transmission can look similar to genealogies derived from a super-spreader. Indeed, if one individual infects 10 others over a long period, and none of those infects anyone else, the genealogy among isolates would look the same as a genealogy in which each case infected precisely one other. However, it is very unlikely that such a chain of cases would occur, with no one *ever* infecting two others rather than one. Similarly, it is unlikely that one host could infect everyone in an outbreak, with no onward transmission by anyone else. In our simulations, once the occasional person in a long chain can infect two others, and if non-super-spreader individuals infect others homogeneously, we find that simple topological structures are well able to resolve the differences between chains and super-spreader outbreaks.

We have used five coarse and simple summaries of tree topology, though we have explored a number of others and each time, found the same qualitative story (Supplementary Information). However, any small set of a few summary statistics cannot capture the topology with much resolution. In contrast, most methods to compare phylogenies in fine detail are suited only for phylogenies on the same sets of tips [19], and so cannot be used to compare different outbreaks or to compare simulations to data. Finding the correct balance to summarize trees sufficiently that they can be compared across different tree sizes, different outbreaks and different settings, without summarizing them so much as to remove the most useful information is a challenge, and a number of methods will likely be developed, beginning with viral pathogens as in the recent work on Poon et al [18]. Indeed, while we feel that the measures we have used are demonstrative that tree structure is revealing, they are not intended to be comprehensive or exhaustive descriptions of tree topology. The fact that a few simple topological summaries can reveal underlying transmission patterns is a proof-of-principle that tree shape is informative.

Tree shapes from the real outbreaks were inferred using BEAST, and tree shapes depend on the prior and on any inherent shape bias in the inference method [9]. In our case the classifier performance was robust to genealogies vs neighbour-joining and worked on maximum-likelihood trees from the dynamic sexual contact networks, suggesting that phylogenetic noise did not play a large role. However, shapes of the posterior outbreak trees varied widely. Ultimately, an understanding of tree shape will allow phylogenetic inference tools such as BEAST to use priors that take shape explicitly into account (whereas currently the shape is incorporated implicitly through the quality of the match of a tree to the genomic dataset).

The classification method we have developed provides not only an important empirical quantification of the degree to which genomic data is informative in the absence of epidemiological information, but is also a useful tool that can be used to describe outbreaks both retrospectively and prospectively. The ability to situate an outbreak on the spectrum from homogeneous transmission to super-spreading and to do so within the earliest stages of an outbreak when neither a large number of specimens nor detailed epidemiological information is available represents an important opportunity for public health investigations. Situating an outbreak on this spectrum does not require pinning down individual transmission events, but relies more on characterizing summary features of the outbreak and/or its phylogeny. If the data point towards a significant role for super-spreading in an outbreak, a containment strategy will require intensive screening of the super-spreader's contacts. In an outbreak where onward transmission is occurring in chains, a focus on active case finding around multiple individuals will be needed instead. Ultimately, investigation of any outbreak of a communicable disease will involve the collation of multiple sources of information, including epidemiological, clinical, and genomic data. The approach described here represents one part of this toolbox, and has the advantages of being robust to the unique nature of complex chronic infection, providing useful information even when epidemiological information is incomplete, and being informative within the earliest stages of an outbreak.

Ackowledgments

We thank the Vancouver Island Health Authority and the Interior Health Authority for their contributions to the real-world outbreak field investigations, and Dr. Patrick Tang, Dr. James Johnston, and Dr. Mabel Rodrigues for their contributions to the genome sequencing work. CC is supported by the EPSRC (EP/K026003/1).

References

[1] Nicola Casali, Vladyslav Nikolayevskyy, Yanina Balabanova, Simon R Harris, Olga Ignatyeva, Irina Kontsevaya, Jukka Corander, Josephine Bryant, Julian Parkhill, Sergey Nejentsev, et al. Evolution and transmission of drug-resistant tuberculosis in a russian population. *Nature genetics*, 2014.

[2] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[3] Nicholas J Croucher, Jonathan A Finkelstein, Stephen I Pelton, Patrick K Mitchell, Grace M Lee, Julian Parkhill, Stephen D Bentley, William P Hanage, and Marc Lipsitch. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nature genetics*, 45(6):656–663, 2013.

[4] Xavier Didelot, Jennifer Gardy, and Caroline Colijn. Bayesian inference of infectious disease transmission from whole genome sequence data. *bioRxiv*, 2013.

[5] Simon DW Frost and Erik M Volz. Modelling tree shape and structure in viral phylodynamics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1614), 2013.

[6] J.L. Gardy, J.C. Johnston, S.J.H. Sui, V.J. Cook, L. Shah, E. Brodkin, S. Rempel, R. Moore, Y. Zhao, R. Holt, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *New England Journal of Medicine*, 364(8):730–739, 2011.

[7] Yonatan H Grad, Marc Lipsitch, Michael Feldgarden, Harindra M Arachchi, Gustavo C Cerqueira, Michael FitzGerald, Paul Godfrey, Brian J Haas, Cheryl I Murphy, Carsten Russ, et al. Genomic epidemiology of the escherichia coli o104: H4 outbreaks in europe, 2011. *Proceedings of the National Academy of Sciences*, 109(8):3065–3070, 2012.

[8] Stephen B. Heard. Patterns in Tree Balance among Cladistic, Phenetic, and Randomly Generated Phylogenetic Trees. *Evolution*, 46(6):1818–1826, 1992.

[9] John P Huelsenbeck and Mark Kirkpatrick. Do phylogenetic methods produce trees with biased shapes? *Evolution*, pages 1418–1424, 1996.

[10] A. M. Johnson, C. H. Mercer, B. Erens, S. Copas, A. J. an d McManus, K. Wellings, K. A. Fenton, C. Korovessis, W. Macdowal l, K. Nanchahal, S. Purdon, and J. Field. Sexual behaviour in britain: partnerships, practices, and hiv risk behaviours. *Lancet*, 358(9296):1835–1842, December 2001.

[11] Thibaut Jombart, Anne Cori, Xavier Didelot, Simon Cauchemez, Christophe Fraser, and Neil Ferguson. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS computational biology*, 10(1):e1003457, 2014.

[12] Midori Kato-Maeda, Christine Ho, Ben Passarelli, Niaz Banaei, Jennifer Grinsdale, Laura Flores, Jillian Anderson, Megan Murray, Graham Rose, L Masae Kawamura, et al. Use of whole genome sequencing to determine the microevolution of mycobacterium tuberculosis during an outbreak. *PloS one*, 8(3):e58235, 2013.

[13] Claudio U Köser, Matthew TG Holden, Matthew J Ellington, Edward JP Cartwright, Nicholas M Brown, Amanda L Ogilvy-Stuart, Li Yang Hsu, Claire Chewapreecha, Nicholas J Croucher, Simon R Harris, et al. Rapid whole-genome sequencing for investigation of a neonatal mrsa outbreak. *New England Journal of Medicine*, 366(24):2267–2275, 2012.

[14] G.E. Leventhal, R. Kouyos, T. Stadler, V. von Wyl, S. Yerly, J. Böni, C. Cellerai, T. Klimkait, H.F. Günthard, and S. Bonhoeffer. Inferring epidemic contact structure from phylogenetic trees. *PLoS Computational Biology*, 8(3):e1002413, 2012.

[15] F. Lewis, G.J. Hughes, A. Rambaut, A. Pozniak, and A.J.L. Brown. Episodic sexual transmission of hiv revealed by molecular phylodynamics. *PLoS medicine*, 5(3):e50, 2008.

[16] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[17] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et al. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[18] Art FY Poon, Lorne W Walker, Heather Murray, Rosemary M McCloskey, P Richard Harrigan, and Richard H Liang. Mapping the shapes of phylogenetic trees from human and zoonotic rna viruses. *PloS one*, 8(11):e78122, 2013.

[19] DF Robinson and Leslie R Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131–147, 1981.

[20] Katy Robinson, Nick Fyson, Ted Cohen, Christophe Fraser, and Caroline Colijn. How the dynamics and structure of sexual contact networks shape pathogen phylogenies. *PLoS computational biology*, 9(6):e1003105, 2013.

[21] T. Stadler. Inferring epidemiological parameters on the basis of allele frequencies. *Genetics*, 188(3):663–672, 2011.

[22] T. Stadler, R. Kouyos, V. Von Wyl, S. Yerly, J. Böni, P. Bürgisser, T. Klimkait, B. Joos, P. Rieder, D. Xie, et al. Estimating the basic reproductive number from viral sequence data. *Molecular Biology and Evolution*, 29(1):347–357, 2012.

[23] Tanja Stadler and Sebastian Bonhoeffer. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1614), 2013.

[24] M Estée Török, Sandra Reuter, Josephine Bryant, Claudio U Köser, Sian V Stinchcombe, Bernadette Nazareth, Matthew J Ellington, Stephen D Bentley, Geoffrey P Smith, Julian Parkhill, et al. Rapid whole-genome sequencing for investigation of a suspected tuberculosis outbreak. *Journal of clinical microbiology*, 51(2):611–614, 2013.

[25] E.M. Volz. Complex population dynamics and the coalescent under neutrality. *Genetics*, 190(1):187–201, 2012.

[26] E.M. Volz, J.S. Koopman, M.J. Ward, A.L. Brown, and S.D.W. Frost. Simple epidemiological dynamics explain phylogenetic clustering of hiv from patients with recent infection. *PLoS Computational Biology*, 8(6):e1002552, 2012.

[27] Timothy M Walker, Camilla LC Ip, Ruth H Harrell, Jason T Evans, Georgia Kapatai, Martin J Dedicoat, David W Eyre, Daniel J Wilson, Peter M Hawkey, Derrick W Crook, et al. Whole-genome sequencing to delineate¡ i¿ mycobacterium tuberculosis¡/i¿ outbreaks: a retrospective observational study. *The Lancet infectious diseases*, 2012.

[28] Rolf JF Ypma, W Marijn van Ballegooijen, and Jacco Wallinga. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics*, 195(3):1055–1062, 2013.