Comparing phylogenetic trees according to tip label categories

MICHELLE KENDALL[1] AND VEGARD ELDHOLM[2] AND CAROLINE COLIJN[1]

[1]*Department of Mathematics, Imperial College London, London, SW7 2AZ, UK* [2]*Department of Bacteriology and Immunology, Norwegian Institute of Public Health, Oslo, Norway*

Corresponding author: Michelle Kendall, Department of Mathematics, Imperial College London, London, SW7 2AZ, UK; E-mail: m.kendall@imperial.ac.uk

Abstract Trees that illustrate patterns of ancestry and evolution are a central tool in many areas of biology. Comparing evolutionary trees to each other has widespread applications in comparing the evolutionary stories told by different sources of data, assessing the quality of inference methods, and highlighting areas where patterns of ancestry are uncertain. While these tasks are complicated by the fact that trees are high-dimensional structures encoding a large amount of information, there are a number of metrics suitable for comparing evolutionary trees whose tips have the same set of unique labels. There are also metrics for comparing trees where there is no relationship between their labels: in 'unlabelled' tree metrics the tree shapes are compared without reference to the tip labels.

In many interesting applications, however, the taxa present in two or more trees are related but not identical, and it is informative to compare the trees whilst retaining information about their tips' relationships. We present methods for comparing trees whose labels belong to a pre-defined set of categories. The methods include a measure of distance between two such trees, and a measure of concordance between one such tree and a hierarchical classification tree of the unique categories. We demonstrate the intuition of our methods with some toy examples before presenting an analysis of *Mycobacterium tuberculosis* trees, in which we use our methods to quantify the differences between trees built from typing versus sequence data.

(Keywords: phylogenetics, tree comparison, Mycobacterium tuberculosis)

Phylogenetic trees are a key tool in understanding the relationships and ancestry in a set of taxa. Although evolution often occurs in a way that does not fit a tree model due to horizontal movement of genes, reassortment, recombination, and of course sexual reproduction, trees are used in applications ranging from macroevolution to infectious disease outbreak reconstruction. It is not straightforward to infer phylogenetic trees from data, and sophisticated software packages have been developed for this task [47, 13, 37]. Tree inference typically requires choices between different computational approaches with different underlying prior models, or even different sources of data. In order to understand posterior uncertainty, determine which sources of data are consistent with each other, and ensure that high-quality trees are obtained, it is useful to directly compare reconstructed phylogenetic trees [4, 2, 7, 21, 25].

Existing tools for comparing trees are strictly limited to comparing trees on identical sets of uniquely-labelled tips [50, 41, 42, 4, 48, 33, 25], or trees whose tips are not labelled at all (which may be considered as unlabelled since their labels are unrelated) [36, 34, 29, 8]. There are many applications, however, in which it is important to compare evolutionary trees which have tips with related but not identical labels.

To our knowledge, the mathematical works on developing metrics to address the problem of comparing trees with non-unique labels are [23, 39, 17]. Huber et al. present extensions of labelled tree metrics that require each tree to have the same multi-set for its tip labels, and thus cannot be calculated for trees with varying tip labels. Ren et al. present an extension to the BHV metric [4] which permits distances to be calculated between trees with differing numbers of tips. This measure can be applied to binary trees with a common set of tip labels but where some trees may have 'additional' or 'missing' taxa. Garba et al. [17] define a probabilistic distance between trees. They note that trees can be compared by pruning them to their set of common taxa, or by an 'augmentation method' which uses the union of their tip sets to define the relevant distributions. Where trees have *entirely* different tip label sets, for example pathogen samples in distinct geographic regions, the former would leave the trees empty and the latter may become unwieldy and fail to draw connections between 'related' tips. Our methods enable the comparison of trees whose tips may be partially or entirely different in numbering and labelling, but where there is a natural relationship between tips by membership of 'categories', as we now explain.

In Table 1 we summarise some of the many possible applications where trees may have related but not identical tip labels; we describe some of these in more detail in the Discussion section. We refer to the 'individuals' which comprise the possible tip labels and the 'categories' to which they belong. These categories may also occur as tip labels in other hierarchical classifications or reference trees.

Comparing trees to each other has a wide range of applications in these domains: testing convergence of Bayesian inference methods [28], quantifying differences across different data sets and locations, and using Approximate Bayesian Computation to infer

Categories	Individuals
Bacterial sub-types e.g. serogroups	Bacterial isolates
Species	Orthologous genes
Hosts	Deep sequencing reads of pathogen from host
Protein families	Individual proteins
Population groups	Individual organisms
Disjoint features or phenotypes	Individual organisms
High-level taxonomy	Individual organisms

Table 1: Examples of applications where individual tree labels may differ but may be comparable according to some broader categories.

parameters of evolution using simulations, for example of the major bacterial subtypes, or of gene and species trees.

Here we describe a distance measure approach to comparing trees whose labels are related at the 'category-level'. It is based on capturing trees as vectors which summarise the extent of shared ancestry between taxa. This also leads to a natural measure of 'concordance' between trees, which we define. We demonstrate the methods using small example trees to illustrate the intuition. Next we apply the methods to trees inferred from *Mycobacterium tuberculosis* data, demonstrating the power of the method to quantify and identify the differences between trees inferred using MIRU versus whole genome sequence data. Finally, we discuss extensions and further applications of the methods.

Materials and Methods

Definitions

In common with much of the literature concerning phylogenetic trees, we will use the following terms:

Definition 1. *A* path *in a tree T is a non-repeating traversal of consecutive edges in the tree. The* length *of a path is the number of edges on the path. Since there are no cycles in a tree, we know that for each tip i there is a unique path p_i from the root to that tip in the tree.*

The depth *of a node is the number of edges on the path from the root to that node. The root has depth zero. We require throughout this paper that the root has at least two direct descendants (outdegree 2) in order to avoid spurious apparent dissimilarities between trees which simply have leading edges before the first node with outdegree greater than 1; we prune such edges from the tree.*

For a pair of tips i and j, their common ancestors *are the nodes which lie on both paths p_i and p_j from the root to i and j respectively. The* most recent common ancestor (MRCA) *of a set of tips S is the node with the greatest depth amongst their set of common ancestors. For convenience, we maintain this terminology so that the MRCA of tip i and tip i is the tip i itself. A set of tips S forms a* monophyletic clade *if the MRCA of the tips in S has no descendants that are not in S.*

We propose measures to compare trees where there is a meaningful correspondence between tip labels at the category level but not necessarily at the individual level. Explicitly, let $\mathcal{C} = \{C_1, \ldots, C_n\}$ be a set of categories of tip labels, as described in Table 1. For example, each C_j could represent a human patient, and the individual tip labels could be various deep sequencing reads of a pathogen common to all the patients. We will denote the set of all possible individual tip labels as $\mathcal{S}$. In this example, $\mathcal{S}$ could be { `Patient_C1_read_1`, ..., `Patient_C1_read_100`, `Patient_C2_read_1`, ..., `Patient_Cn_read_100` }.

Let $\sigma : \mathcal{S} \to \mathcal{C}$ be the function which associates a tip label with its category, so that if $i \in \mathcal{S}$ belongs to category C_j then $\sigma(i) = C_j$. In practice, σ can be thought of as a table or data frame, such as:

Category	Individuals
Patient A	Patient A read 1
	Patient A read 2
	$\vdots$
	Patient A read 100
Patient B	Patient B read 1
	$\vdots$
	Patient B read 100
$\vdots$	$\vdots$
Patient N	Patient N read 1
	$\vdots$
	Patient N read 100

Table 2: Example of a set of categories $\mathcal{C}$ and all possible individuals $\mathcal{S}$, as linked by the function σ.

Definition 2. *Let $\mathcal{C}$ be a set of categories, $\mathcal{S}$ a set of individuals, and $\sigma : \mathcal{S} \to \mathcal{C}$ a well-defined (many-one) function which associates each individual with a category. We will refer to a* comparable set of trees $\mathcal{T}$ *as a set of rooted trees with associated sets $\mathcal{C}, \mathcal{S}$ and function σ, where each tree $T \in \mathcal{T}$ has at least one individual tip label associated with each category in $\mathcal{C}$. That is, for each category C_j, T has at least one tip i such that $\sigma(i) = C_j$.*

We note that alternative definitions of tree distances and concordance are possible if this condition is relaxed, but the natural variations we explored placed undesirable weight on some small changes between trees.

We now define a *tip-collapsing* operation $*$ as follows.

Definition 3. *For $T \in \mathcal{T}$, we define the* collapsed *tree T^* to be the tree obtained from T by:*

1. *replacing each tip label i with the label $\sigma(i) = C_j$*

2. *wherever there is a monophyletic clade of tips labelled C_j (tips from the same category) we prune away the whole clade leaving only its MRCA node. This node (which is now a tip) is given the label of the pruned category C_j.*

We demonstrate the tip-collapsing operation $*$ with an example in Figure 1.

(a) Original tree T_1 (b) Collapsed tree T_1^*

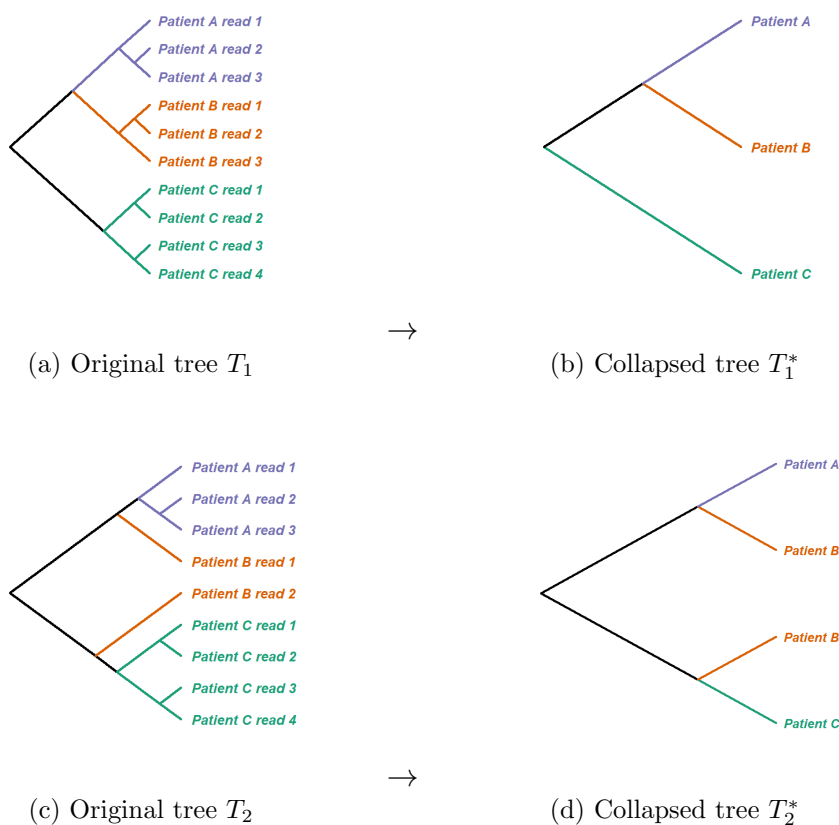(c) Original tree T_2 (d) Collapsed tree T_2^*

Figure 1: An illustration of the tip collapsing operation. In (b) and (d), each tip from (a) and (c) respectively has been relabelled with its category name (here, the patient ID, also indicated by colour), and each monophyletic clade of same-category tips has been reduced to a single tip with that category label.

Pairwise tree distance measure

In order to define a distance between a pair of trees we first need to define what it means for two trees to be equivalent.

Definition 4. *We will consider two trees T_1, T_2 in a comparable set of trees $\mathcal{T}$ to be $*$-isomorphic if T_1^* and T_2^* have the same labelled shape (topology) [19, 42], which is equivalent to saying that there is a tip-label-preserving isomorphism from T_1^* to T_2^*.*

Because the labels of T_1^* and T_2^* may be multisets (some labels repeated as with 'species B' in Figure 1), the label-preserving isomorphism may permute multiply-occurring labels. We define a distance measure on the trees $\mathcal{T}^*$, and by extension, on $\mathcal{T}$. In a method similar to those proposed in [6, 25], we characterise each collapsed tree T^* by a vector v, and then the distance between two trees will be given by the Euclidean distance between their collapsed tree vectors. This distance function is in general not a metric (see below).

Definition 5. *Suppose we have a comparable set of trees $\mathcal{T}$, with tip labels from $\mathcal{S}$ and associated $\mathcal{C}$ and σ. For each tree $T \in \mathcal{T}$, let $P(T)$ be the set of all tip pairs in T from different categories. That is, if i, j are tips in T and $\sigma(i) \neq \sigma(j)$ then $(i, j) \in P(T)$. Note that if T has exactly one tip associated with each category in $\mathcal{C}$ then $|P(T)| = \binom{|C|}{2}$.*

For a tip pair $(i, j) \in P(T)$, let $w_{i,j}(T)$ be the depth of the MRCA of i and j in T. (Note that in the case where $\mathcal{T}^*$ has exactly one label $\sigma(i)$ and exactly one label $\sigma(j)$ then $w_{i,j}(T)$ is equivalent to the depth of the MRCA of the category labels $\sigma(i)$ and $\sigma(j)$ in T^*.) For categories C_x and C_y with $C_x \neq C_y$, let $v_{x,y}(T)$ be the mean of all the values $w_{i,j}(T)$ for $(i, j) \in P(T)$ where $\sigma(i) = C_x$ and $\sigma(j) = C_y$.

We can now define the tree vector of T to be

$$v(T) = (v_{1,2}, v_{1,3}, \ldots, v_{|\mathcal{C}|-1,|\mathcal{C}|})$$

Definition 6. *Let $\mathcal{T}$ be a set of comparable trees. For $T_x, T_y \in \mathcal{T}$ we define the distance between T_x and T_y to be:*

$$d(T_x, T_y) = ||v(T_x) - v(T_y)|| \ ,$$

where $|| \cdot ||$ is the Euclidean distance (l^2-norm).

We note some properties of this distance function. Suppose we have a collection of comparable trees $\mathcal{T} = T_1, \ldots, T_n$, $n \geq 2$, with individual labels $\mathcal{S}$ and associated $\mathcal{C}$ and $\sigma : \mathcal{S} \to \mathcal{C}$. When each collapsed tree $T_1^*, \ldots, T_n^*$ has exactly one tip label from each category $C_1, \ldots, C_m \in \mathcal{C}$ then d is equivalent to the topological tree vector d_0 from [25], where it is proven to be a metric (positive, symmetric, zero if and only if the trees are isomorphic, and satisfies the triangle inequality). Therefore, under these conditions d is a metric on the collapsed trees $T_1^*, \ldots, T_m^*$, and a pseudo-metric on the original trees

$T_1, \ldots, T_m$, which may still differ *within* the structure of their monophyletic category clades even when the distance d is zero.

Whenever there are duplicate category labels in any of the collapsed trees $T_1^*, \ldots, T_n^*$, it is possible to find 'false zeroes', i.e. a distance of zero between non-isomorphic trees. This is essentially because taking the mean of the depths results in a loss of information. For example, the distance between collapsed trees $T_1^* = ((A, B), C)$ and $T_2^* = ((A, A, B), C)$ (written in Newick format) would be zero. Thus d is in general a pseudo-metric on the collapsed trees.

Concordance measure

The tip collapsing operation also enables us to define a measure of *concordance* between a collection of comparable trees $\mathcal{T}$ and a reference tree R. The labels of the reference tree R must also be comparable: they must be the associated set of categories $\mathcal{C}$, with the additional condition that each category label must occur precisely once. The concordance value ranges from ≈ 0 if the trees strongly disagree in their evolutionary histories to a value of 1 when they are in 'complete agreement', as we explain below.

Definition 7. *Let $\mathcal{T}$ be a set of comparable trees and let R be a fixed reference tree whose labels are precisely the associated set of categories $\mathcal{C}$. Extending the earlier notation, let $w_{x,y}(R)$ be the depth of the (unique) MRCA of categories C_x and C_y in R. Then for any tree $T \in \mathcal{T}$ we define an indicator for a pair of tips $(i, j) \in P(T)$:*

$$\delta(i, j) = \begin{cases} 1 & \text{if } w_{i,j}(T) = w_{\sigma(i),\sigma(j)}(R) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

We can now define a measure of concordance between T and R. It counts the fraction of tip pairs $(i, j) \in P(T)$ (pairs of individuals in T from different categories) whose relative positions in T differ from the relative positions of those categories in R. Specifically, it counts the fraction of tip pairs for which the depth of the MRCA in tree T is the same as the depth of the MRCA of the corresponding categories $\sigma(i)$ and $\sigma(j)$ in R.

Definition 8. *The concordance of T with R is given by*

$$conc(T, R) = \frac{1}{|P(T)|} \sum_{(i,j) \in P(T)} \delta(i, j) \ .$$

When $conc(T, R) = 1$, T and R are *fully concordant*, meaning that they describe the same evolutionary relationships at the category level.

Theorem 1. *Let $\mathcal{S}$ be a set of individuals and $\mathcal{C}$ a set of categories, with an associated function $\sigma : \mathcal{S} \to \mathcal{C}$. Let T be a rooted tree with labels in $\mathcal{S}$ and R a rooted tree whose labels are exactly the set $\mathcal{C}$ (each occurring exactly once). If $conc(T, R) = 1$ then:*

1. *if $\mathcal{T}^*$ is a binary tree, then $\mathcal{T}^* = R$. That is, for each category C_j, all its individuals i such that $\sigma(i) = C_j$ form a monophyletic clade in T.*

2. *if $\mathcal{T}^*$ is not binary then for each category C_j, all its individuals i such that $\sigma(i) = C_j$ form one or more monophyletic clades in T, the MRCAs of which are all sisters (they share the same parent node). Thus, if a further collapsing operation were performed on $\mathcal{T}^*$ to create a tree $\mathcal{T}^{**}$, where any sister tips with the same category labels were collapsed into a single tip with that category label, then $\mathcal{T}^{**} = R$.*

Proof. Consider the binary case (a). For each category C_j, let the set of individuals i such that $\sigma(i) = C_j$ be denoted S_j. Suppose for a contradiction that $conc(T, R) = 1$ but there exists a set of individuals $S_j \subseteq \mathcal{S}$ such that S_j does not form a monophyletic clade in T. Let N be the MRCA of S_j in T. Then by definition the direct descendants of N are themselves tips from S_j or have descendants which are tips in S_j. Since S_j is not a monophyletic clade, N has at least one descendant tip that is not in S_j: call this tip k, with $\sigma(k) \neq C_j$. Note that tip k cannot be a direct descendant of N, otherwise the other direct descendant of N would be the MRCA of S_j.

Without loss of generality, let N_1 be the direct descendant of N that is an ancestor of k, and let N_2 be N's other direct descendant. Node N_1 has descendants in S_j; call these S_{j1}, and node N_2 is either a tip itself in S_j or has descendants in S_j; call these tip(s) S_{j2}. We have that $S_{j1} \bigcup S_{j2} = S_j$. The MRCA of any tip $a \in S_{j1}$ and k is N_1. In contrast, the MRCA of k and any tip $b \in S_{j2}$ is N. Since N_1 and N have different depths, we have $w_{k,a}(T) \neq w_{k,b}(T)$. Thus, at most one of $w_{k,a}(T), w_{k,b}(T)$ can be equal to $w_{\sigma(k),C_j}(R)$, hence either $\delta(k, a)$ or $\delta(k, b)$ is zero, so $conc(T, R) < 1$.

Now consider the non-binary case (b). Again, assume that $conc(T, R) = 1$ but that there is a set of individuals S_j from the same category C_j which is neither monophyletic in T, nor forms multiple sister monophyletic clades in T. Again let N be the MRCA of S_j. Then there exists a direct descendant of N, call it N_1, with at least one descendant tip x in S_j and at least one descendant tip k from a different category, i.e. with $\sigma(k) \neq C_j$. Because N is the MRCA of all tips in S_j, it has at least one other descendant y in S_j which is not a descendant of N_1. Then the MRCA of x and k has depth greater than or equal to the depth of N_1, whereas the MRCA of k and y is N. As in case (a) this contradicts $con(T, R) = 1$.

Consider collapsing T to T^*. The tips of T^* are the MRCA(s) of the monophyletic clade(s) of each S_j, labelled by the corresponding category C_j. Now, performing the further collapse T^{**}, where any sister tips with the same category labels are collapsed into a single

tip with that category label, means that T^{**} and R have the same set of tip labels and each occurs precisely once. We then have $v(T^{**}) = v(R)$, which by the Kendall Colijn metric [25] means that T^{**} and R are isomorphic. $\qquad\square$

Finally, we note that whilst the concordance measure can equal one and can reach values close to zero, it will never in fact equal zero.

Lemma 1. *The concordance measure $conc(R, T)$ gives values in the interval $(0, 1]$.*

Proof. It is straightforward to construct examples of R and T such that $conc(R, T) = 1$. It follows from Theorem 1 that for any T, letting $T^* = R$ gives $conc(T, R) = 1$.

Suppose for a contradiction that we have comparable, rooted trees R and T with $conc(R, T) = 0$. Since R is rooted we can pick a pair of category labels in R whose MRCA is the root (depth 0): call these categories C_x and C_y. Since $conc(R, T) = 0$, the depth of the MRCA of each pair of individual tip labels (i, j) in T with $\sigma(i) = C_x$ and $\sigma(j) = C_y$ must be strictly greater than zero, i.e. they are always found on the same 'side' of the root in T.

Since the root has outdegree at least 2 (Definition 1), there must be at least one tip in T which descends from a different 'side' of the root. Call this tip k and say that it belongs to category C_z, where $C_x \neq C_z \neq C_y$. Now the depth of the MRCA of tip k with any individuals from $S_x = \{i : \sigma(i) = C_x\}$ or $S_y = \{i : \sigma(i) = C_y\}$ is zero, and so $conc(R, T) = 0$ tells us that the depths of the MRCAs of (C_x, C_z) and (C_y, C_z) in R are both strictly greater than zero. However, we picked C_x and C_y as categories whose MRCA is the root, and tip C_z cannot be on both sides of the root, so we have our contradiction. $\qquad\square$

We note that requiring every category to be present exactly once in R limits the applicability of the method. However, our rationale can be explained by the following simple example. Consider a tree T and let $R = T^*$. Then by Theorem 1, $conc(R, T) = 1$. Now consider a tree T' obtained from T by pruning away all the tips belonging to one of its categories c, along with the internal nodes whose tip descendants all belonged to c. This necessarily removes at least one internal node, with the consequence that at least one remaining internal node will have a shallower depth in T' than in T. Thus $conc(R, T') < 1$ despite the *remaining* tips sharing the same relative phylogenetic relationships in R and T' at the category level. Where this is not a concern, the condition could be relaxed.

Implementation

RAxML trees were inferred using [47]. All other analysis and the presentation of our results were performed using R R Core Team 2014. To calculate distances between

concatenated WGSs we used the "K80" [26] distance of `dist.dna` from *ape* [31]. The "maximum" and "Euclidean" distances between MIRU types were calculated using `dist` from the R *stats* package R Core Team 2014. The NJ trees were estimated using `nj` from *ape* [31]; UPGMA clustering and ML tree estimation were performed using `upgma` and pml respectively from *phangorn* [31, 46]. The MDS of tree distances (Figure 7b) was plotted using `plotGrovesD3` from *treespace* [24]. Other plotting functions used were from *ape* [31] and *ggplot2* [49].

We have implemented our tree distance and concordance functions in the R package *treespace* [24]. Reproducible code for the small examples in this paper is supplied as a vignette with the package.

Data

For our analysis of *M. tuberculosis* we chose a dataset representing a cluster of 129 clinical isolates belonging to Lineage 3 that had been characterized both by WGS and MIRU-VNTR typing [3]. A total of 1418 variable sites were concatenated into a single multifasta file by Ayabina et al. and used as input in our analyses.

Results

Illustrative examples

We present some small examples to help provide intuition and clarification about the measures we have proposed.

Tree distance measure

In Figure 2 we present six comparable trees with individuals set

$$\mathcal{S} = a1, \ldots, a9, b1, \ldots, b9, \ldots, h1, \ldots, h9$$

where individuals $a1, \ldots a9$ belong to category A and so on. In Figure 3 we give the matrix of distances between each pair of trees. These figures capture some simple principles behind the behaviour of the distance measure, as we note in the following observations.

Trees T_1 and T_2 are very similar, in fact $d(T_1, T_2) = 0$ because $T_1^* = T_2^*$. The phylogenetic relationships between the categories are identical despite T_2 having more tips than T_1.
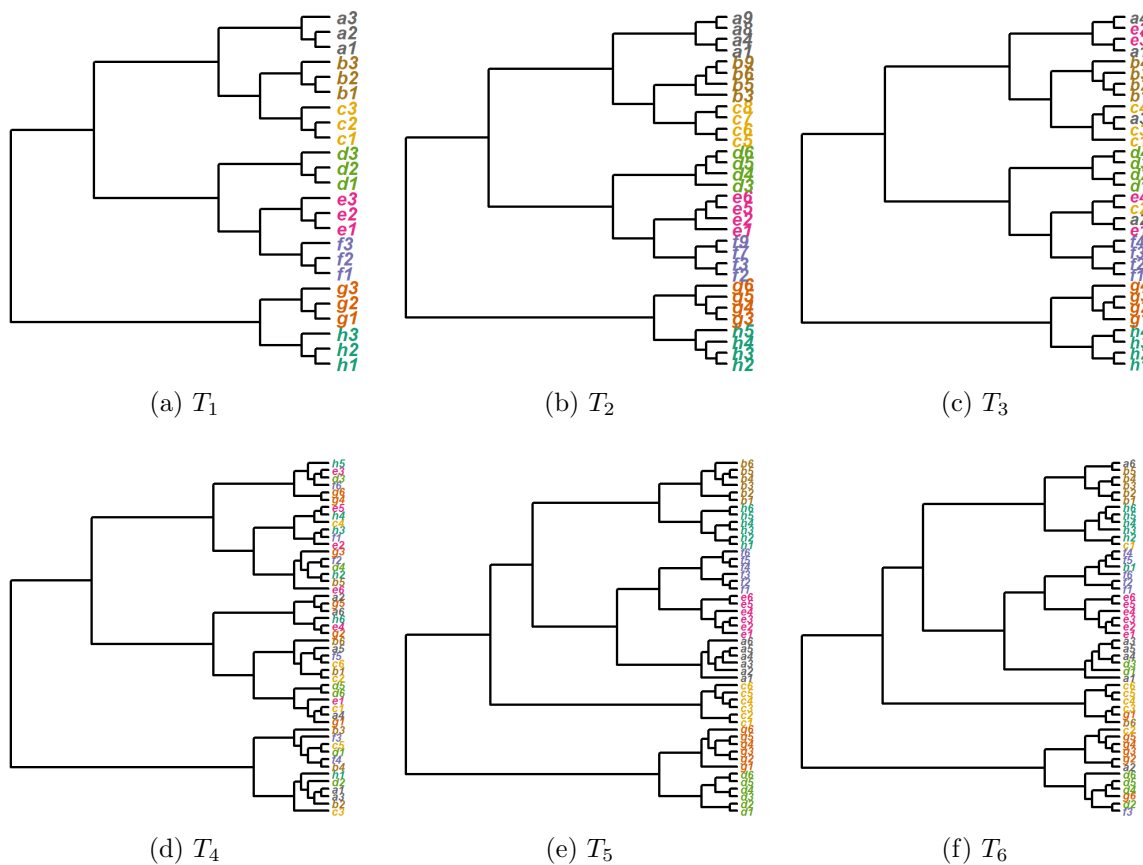
(a) T_1 (b) T_2 (c) T_3

(d) T_4 (e) T_5 (f) T_6

Figure 2: An example of six comparable trees

Tree T_3 was generated by taking $T_1^* = T_2^*$, grafting on four individual per category, and then swapping 20% of the tip labels. Accordingly, some of the categories (B,D,F,G,H) remain monophyletic but others are paraphyletic, so that $d(T_1, T_3) > 0$ (similarly for T_2).

Tree T_4 was generated in a similar way to T_3 but by shuffling 100% of the tip labels. None of the categories remain monophyletic, and accordingly $d(T_1, T4) > d(T_1, T_3)$ (see Figure 3).

Tree T_5 was generated by simulating a different, random category tree and appending individual tips to form monophyletic clades. In particular, the pairing of clades for categories G and H in T_1 and T_2 on one side of the root is replaced by a pairing of category clades G and D in T_5. We therefore find that there is a rather large distance between T_1 and T_5, since both consist of monophyletic category clades but with different evolutionary histories.

Tree T_6 was generated like T_5 but by then shuffling 30% of the tip labels. Accordingly, the distance between trees T_5 and T_6 is quite small (a little larger than $d(T_1, T_3)$, where 20% of tips were shuffled). Notice that the shuffling has created some

	Tree 2	Tree 3	Tree 4	Tree 5	Tree 6
Tree 5					3
Tree 4				6.6	4.6
Tree 3			5.2	6.9	5.8
Tree 2		2.1	5.3	7.1	5.9
Tree 1	0	2.1	5.3	7.1	5.9

Figure 3: Table of tree distances from Figure 2

similarities in category relationships between trees T_4 and T_6, so that $d(T_4, T_6) < d(T_5, T_6)$.

Concordance measure

In Figure 4 we demonstrate a simple reference tree R and three comparable trees T_1, T_2, T_3. We show the concordance of each with R in the captions. It can be seen that T_1 is fully concordant with R (Figure 4b) because the relationships between the categories are identical: $T^* = R$. In T_2 (Figure 4c) the individuals belonging to categories A and B are no longer each monophyletic, but the overall root separation of categories A and B from a monophyletic category C is preserved, resulting in fairly high concordance. In T_3 (Figure 4d) the root separates category A individuals from those in categories B and C, and individuals from B and C are paraphyletic, resulting in much lower concordance with R.

Finally, we demonstrate the effect of 'mixing' tip labels in trees so that they are increasingly less concordant with a reference tree. To do this we created a random reference tree R with ten tips corresponding to ten categories A to J. To create our comparable trees we then grafted on ten 'individual' tips per category. We permuted the labels of either 0%, 20%, 40%, 60%, 80% or 100% of the individual tips, then calculated the concordance of the resulting tree with R. We repeated this experiment 100 times for each proportion of labels to be permuted. Figure 5 is a box plot of the tree concordance results, grouped by the proportion of tip labels permuted. We see that, in general, the more tips that are permuted, the less concordant the tree to the original reference tree R. This reflects the fact that category-level relationships become weaker: as the number of permutations increase, the monophyly of each category gets broken. Some tree similarities remain (and hence the concordance does not decrease below approximately 0.15) due to the fact that each tree is simulated based on the same reference tree R.
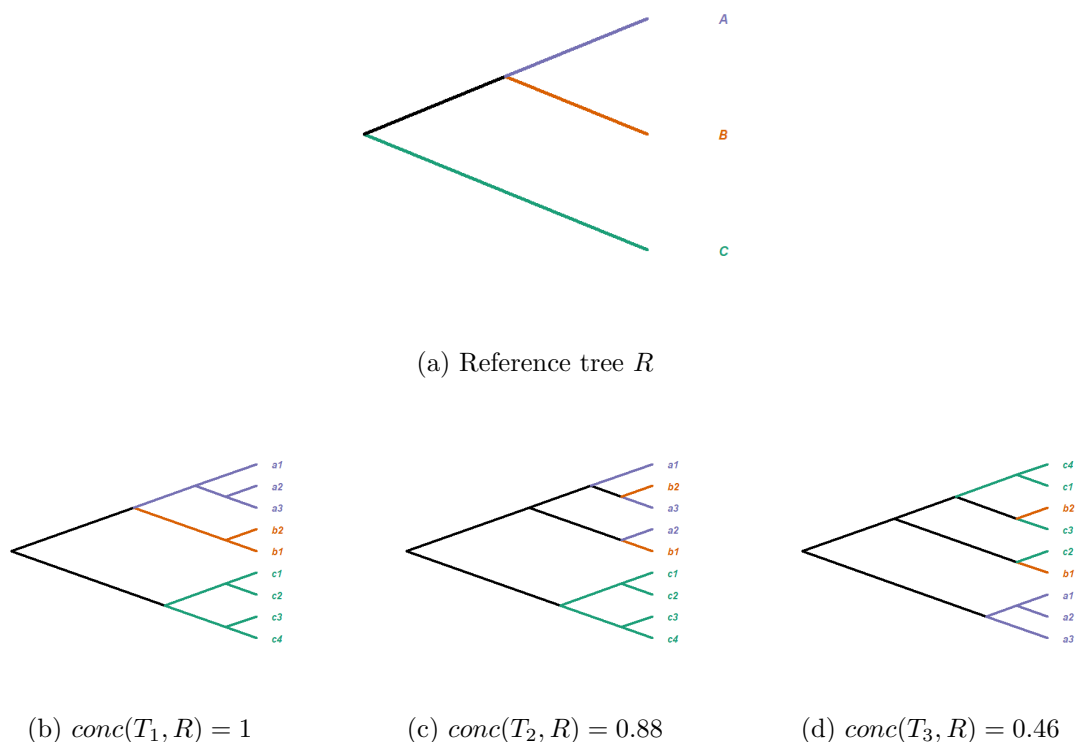
(a) Reference tree R

(b) $conc(T_1, R) = 1$ (c) $conc(T_2, R) = 0.88$ (d) $conc(T_3, R) = 0.46$

Figure 4: A toy example of a reference tree R and three comparable trees T_1, T_2, T_3 with their respective concordance ((c) and (d) rounded to two decimal places).

We note that, as explained by Lemma 1, the concordance does not quite decrease to zero; the minimum value found in this experiment was 0.183111. This tells us that 0.183111 of the individual tip pairs from different categories retained their relative MCRA depths from before the permutation. This number is still relatively high, partly because we did not change the structure of the underlying tree in our experiments. We found that picking a new random 'base' tree from which to create our individuals tree each time typically resulted in lower concordance values.

Comparing MIRU and WGS trees of Mycobacterium tuberculosis

We demonstrate an application of our method to typing and sequence data for the bacterium *M. tuberculosis*, the causative agent of tuberculosis. Globally, *M. tuberculosis* remains a leading infectious killer, and molecular epidemiological surveillance is of key importance for identifying and breaking transmission chains. Mycobacterial interspersed repetitive unit-variable number of tandem repeat (MIRU-VNTR) typing is the most commonly applied method for molecular epidemiological surveillance, but whole-genome
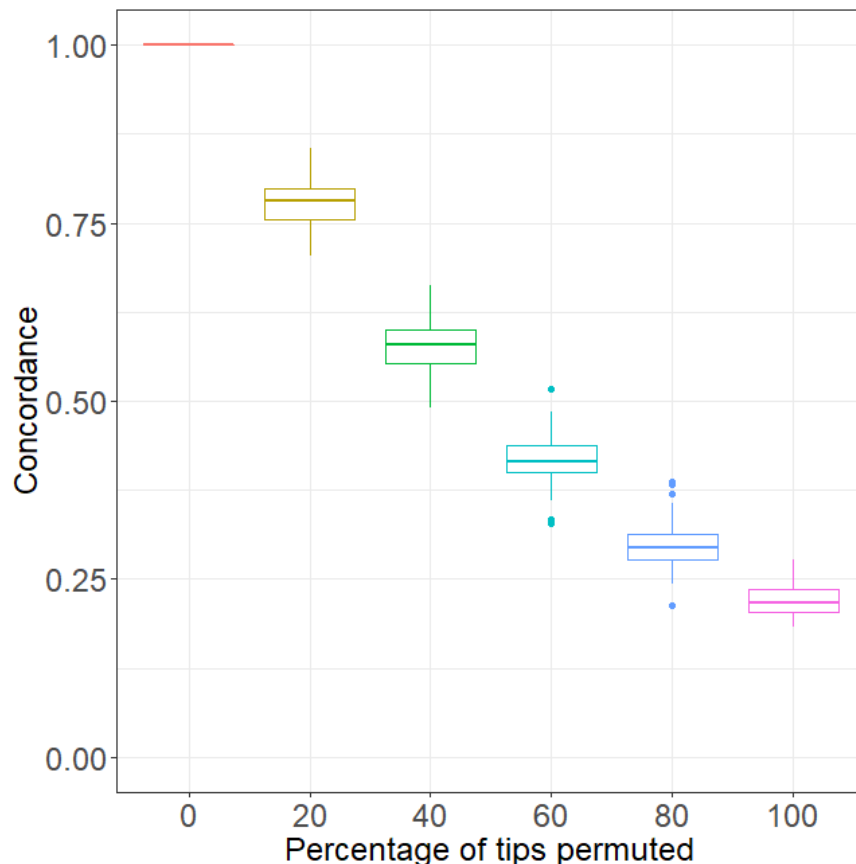
Figure 5: Concordance to a reference tree after a given percentage of random tip permutations

sequencing (WGS) is rapidly becoming the method of choice in resource-rich countries, due to its superior resolution and power. Since WGS and MIRU-VNTR are very different technologies and MIRU-VNTR samples repetitive regions of the genome that are difficult to capture with short read sequencing, it is natural to expect that trees produced by these input types will differ. This hypothesis is supported by the findings of various studies [5, 43, 14, 18]. Our methods enable us to quantify and visualise such tree differences.

There are a variety of ways to measure distances between whole-genome sequences (see for example the options available for `dist.dna` in the R package *ape* [31]), a variety of measures of distance between MIRU types (including those given in `dist` in R R Core Team 2014), and multiple methods for building phylogenetic trees from a set of distances. We tested as many combinations of these methods as feasible. For simplicity of the narrative we omit here many distances and methods which produced repetitive results. All of our findings and any further combinations can be easily reproduced following the

method given in the *treespace* [24] vignette. Full details of the functions and R packages used are given at the end of this section.

We can straightforwardly verify that the differences between isolates according to WGS distances versus MIRU type distances have very low correlation using simple scatterplots, as demonstrated in Figure 6. Here we found the distances between the concatenated whole genome alignment and the "maximum" and "Euclidean" distances between MIRU types.
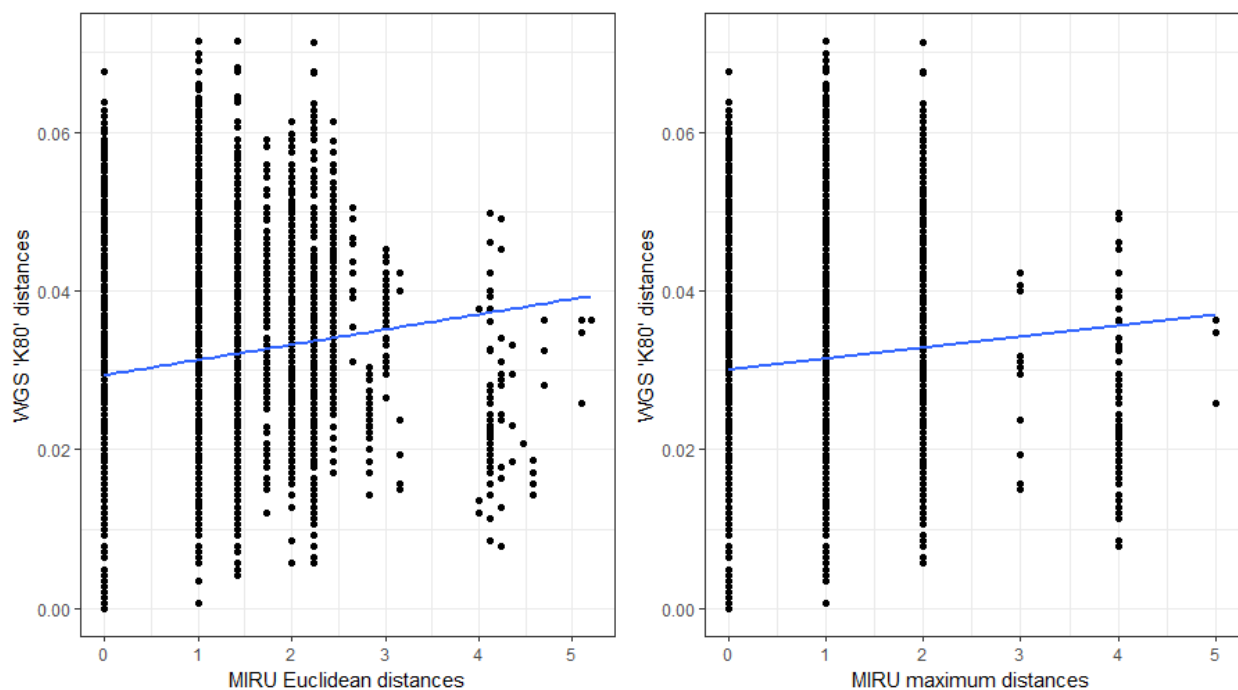

Figure 6: WGS distances and MIRU type distances using the K80 distance for sequences and two choices of MIRU distance.
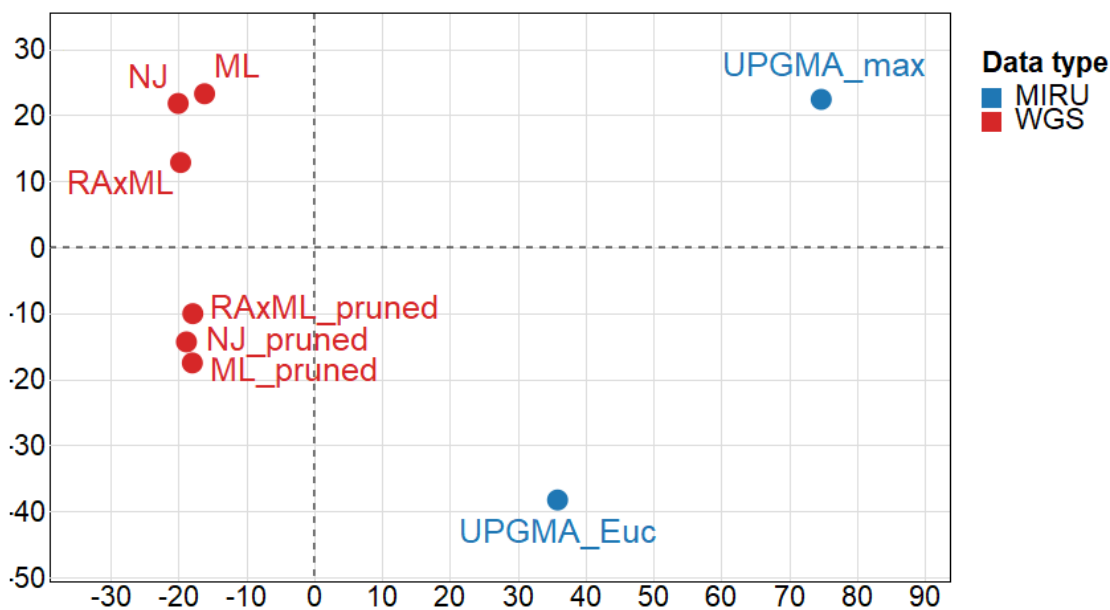
Tree distances

We consider the set of 24 unique MIRU types as 'categories', so that each of the 129 isolates can be assigned to a single MIRU category. We can now compare trees at the MIRU category level, even when they have multiple tips per MIRU type.

We made trees of all the isolates (unless stated otherwise) according to the following methods: UPGMA maximum distance on MIRU types; UPGMA Euclidean distance on MIRU types; neighbour-joining (NJ) on WGSs; maximum likelihood (ML) on WGSs in R and maximum likelihood using RAxML [47]. For the latter, only the 107 unique WGSs were used, as required by the software.

We rooted the trees using the ancestral MIRU type '1064-32' as an outgroup [3]. For the trees created from MIRU distances this was straightforward because all the tips
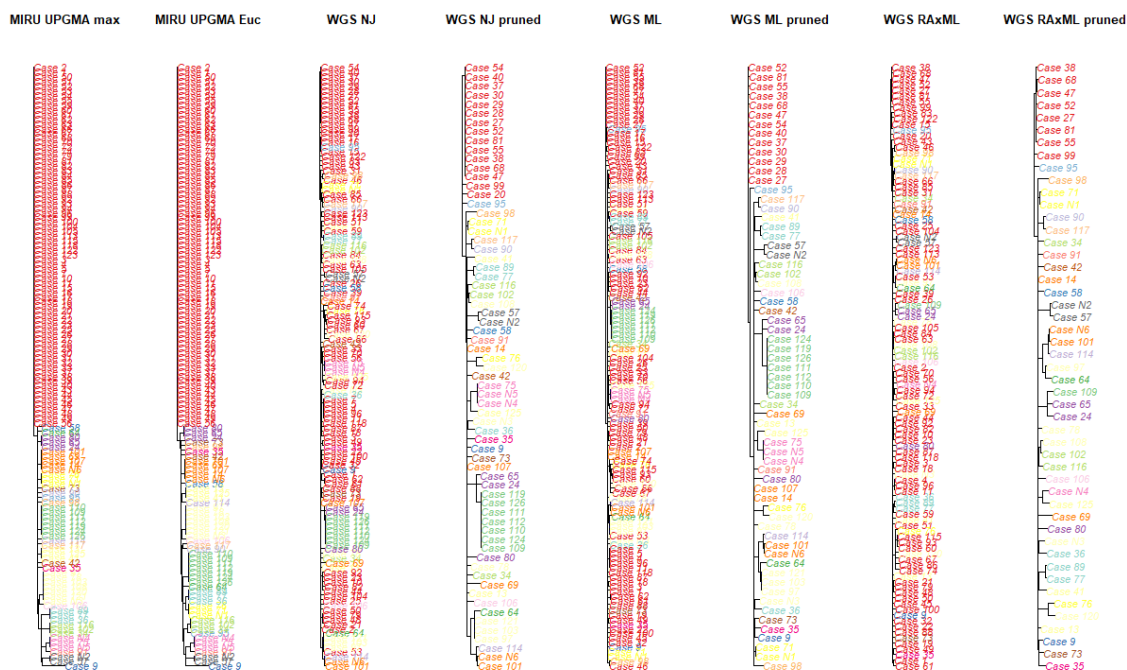
(a) Trees of individual isolates, with tip labels coloured according to MIRU type. The ancestral MIRU type '1064-32' is shown in red and can be seen to be highly paraphyletic in the WGS trees before its smaller clades were pruned.

(b) MDS plot of the relative distances between each of the trees built from WGS or MIRU data.

Figure 7: Comparing trees from MIRU and WGS data

belonging to this category were monophyletic. However, for the WGS trees this was never the case. Instead, we rooted each of these trees using the *largest* monophyletic clade of tips belonging to category '1064-32' as the outgroup. Additionally, we included in our analysis a copy of each of these trees with the remaining tips from category '1064-32' (from smaller clades) pruned out; we see later in the concordance measure that this seems to reduce the bias away from focusing so strongly on the paraphyletic nature of this category. The trees are shown in Figure 7a.

Using the distance function (Definition 6) we found the pairwise distances between each of these trees and present them in Figure 7b using a multidimensional scaling (MDS) plot [10]. As in [2, 21, 22, 25], similar trees are plotted closer together, and we can clearly see the difference in signal between MIRU and WGS input data, and the close similarity in the WGS trees after the pruning of non-basal tips of MIRU type '1064-32'.
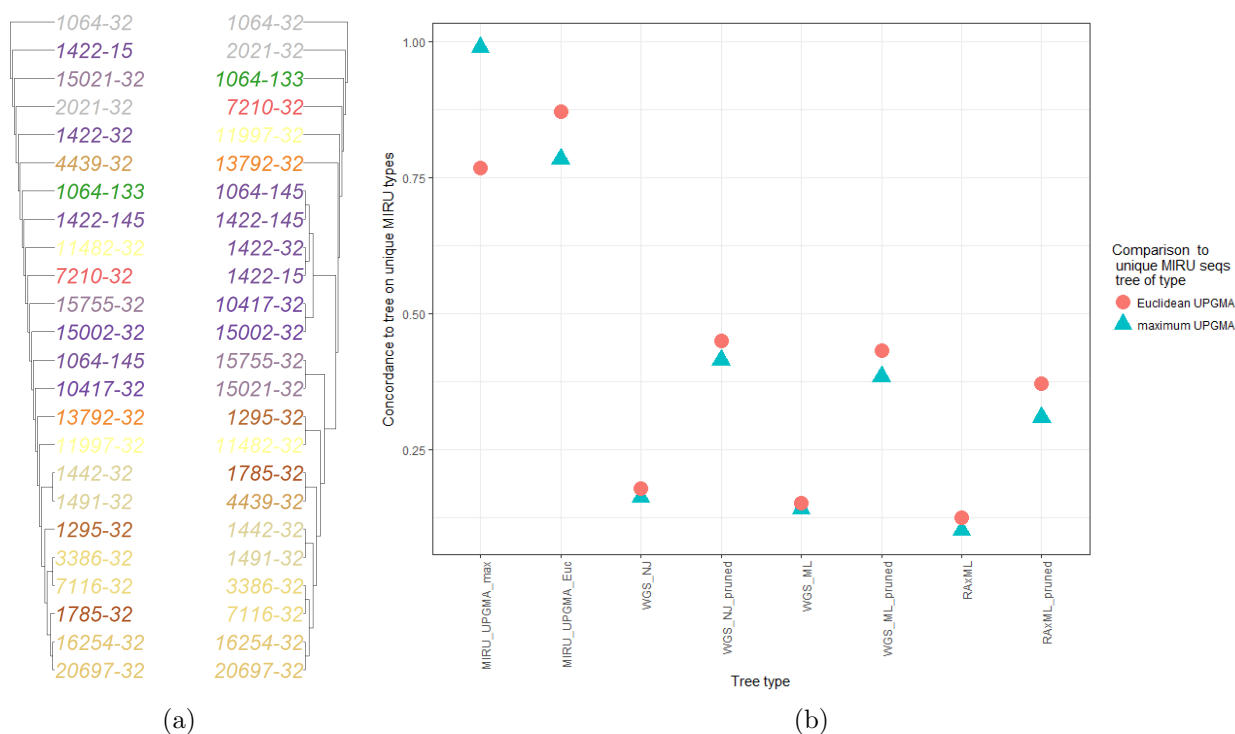
Concordance

(a)   (b)

Figure 8: Concordance to MIRU reference trees. The reference trees are given in Figure (a) ("maximum" left and "Euclidean" right), plotted using `plotTreeDiff` from *treespace* to highlight topological differences. Tips are coloured according to their number of ancestral differences between the two trees, so tips of the same colour have the same number of differences in their set of MRCA depths between the two trees. Figure (b) shows the concordance of individual isolate trees to the two reference trees.

Using one of each of the 24 unique MIRU types, we built two 'MIRU reference trees', one from the the "maximum" and one from the "Euclidean" distance measure and both with UPGMA linkage. These are shown in Figure 8a. We then used the concordance measure to find the concordance of the trees in Figure 7a to these reference trees. The results are plotted in Figure 8b.

The concordance in this setting is a measure of two characteristics of the large trees of individual isolates: first, the extent to which they display MIRU types as monophyletic, and second the extent to which the ancestral relationship between each pair of MIRU types matches the relationship of the corresponding pair in the reference tree. The highest concordances are found between the large trees created from MIRU distances. This was expected because the reference trees were also created from MIRU distances. These MIRU distance large trees did indeed have the property that each category is monophyletic. The highest concordance (0.98) was found between the two trees inferred from 'maximum' distance. This reflects the fact that these trees display very similar evolutionary histories per category. The trees from WGS data have noticeably lower concordance, although the concordance is dramatically improved by the pruning of non-basal tips from category '1064-32'. We see that, for each large tree, the value of the concordance to the "maximum" reference tree (blue triangles) is quite similar to the value of the concordance to the "Euclidean" reference tree (red circles).

DISCUSSION

We have presented methods for quantitatively comparing trees according to their heirarchical patterns of a set of categories. We demonstrated one application for *M. tuberculosis*, comparing the phylogenetic signals from MIRU and WGS data. MIRU typing has been noted to be discordant with WGS for tuberculosis previously [5, 43, 14, 18]; our method can assess this quantitatively. This could form a basis for assessing candidate typing schemes, allowing researchers to seek a scheme with the most reliable and high concordance to core genome phylogenies derived from WGS.

Further potential applications are varied and wide-ranging, including new methods for selecting consensus trees [20] and constructing supertrees [45]. In common with the method of Garba et al. [17], the methods also lend themselves to analyses where some or all of the trees have one or more 'missing' or 'additional' tips due to presence or absence of input data or after the pruning of 'rogue' taxa [1, 47]. The methods presented here will place little to no importance on such variations if the 'missing' tips (where present) are located close to other tips from their category, and similarly if 'additional' tips are grouped close to others from their category. To capture tip presence information, future developments to the methods could include weighting factors to magnify tree distances where there are significant differences between tip sets.

Most bacteria are divided into sub-types according to phenotypic or molecular typing schemes that pre-date next-generation sequencing technologies (examples include the MIRU-VNTR typing we have discussed, as well as MLST typing, capsular typing that defines serogroups of *S. pneumococcus*, sequence types of *S. aureus*, and so on). Consider comparing the evolution of these groups in two geographical areas. In this context it may be important to know how concordant the typing scheme is with a phylogeny (perhaps derived from whole-genome sequencing), how these groups evolved, and whether the loci defining the typing scheme are moving through the population differently than other loci. Datasets from different regions would be expected to mostly share the set of high-level categories, such as serogroups, but will not have the same individual tips. In such applications, therefore, variation in the deep structure of the tree may be of key interest but differences in clades of highly similar taxa from the same category may not be particularly important. Our approach quantifies the concordance at the higher level of categories rather than at the fine-grained level of individual taxa.

At the macroevolutionary level, phylogenetic trees are typically used to represent speciation, with taxa corresponding to species and internal nodes corresponding to inferred speciation events. However, each species contains many genes, and molecular phylogenies derived from gene sequences contain sequences from many orthologous groups of protein-coding genes for each species. Indeed, reconciling gene trees and species trees, and inferring robust species trees from molecular data, is a major challenge for phylogenetics in a macro-evolutionary setting [44, 11, 12, 30]. Nodes in gene family trees may correspond to speciation or to gene duplication, and species' names occur multiple times in the tips of these trees [9]. Quantitatively comparing gene family trees to each other (for different groups of genes) and to species trees therefore requires coping with trees whose species labels are repeated.

Due to continuing improvements in sequencing technology, it is possible to characterise pathogen diversity within individual hosts. Particularly for HIV, deep sequencing is increasingly used to understand transmission and determine rates of adaptation and immune escape [51, 40]. In deep sequencing studies, different regions of the genome are amplified to produce many reads for each patient, but by the nature of short-read sequencing, reads from one locus are not linked to reads at another. In consequence, while the patient IDs correspond from one locus to another, the individual reads for each patient do not. Applications of deep sequencing are numerous and range from studying the dynamics of immune escape [16, 40], estimating the date of infection [35], comparing diversity in individuals to global diversity, examining reversions to ancestral forms and recombination within individuals [53], and using deep sequencing to infer transmission events [15, 51]. While these tools are most established for HIV, their use in other pathogens, including bacteria, is rapidly becoming established [32, 51, 52].

We note that a single distance between a pair of trees or a concordance value

between 0 and 1 can be challenging to interpret. For example, a low concordance value for a tree could be caused by paraphyly of one or more categories, or by significant differences in the relationships between categories in comparison to the reference, or (commonly) a mixture of both. Without plotting the trees it is not straightforward to assess the magnitude of each of these (non-independent) contributions. Another measure or visualisation tool like `plotTreeDiff` from [24] would help to make such results more immediately informative. Similarly, it is not straightforward to define the expected value of either of these measures. Although it may be possible to derive a theoretical expected value (essentially a mean across all possible trees), we would expect trees constrained by data to be much more similar to each other than a pair of random trees. Thus, an *objective* distance x between two trees is not especially informative (unless it is zero) but *relative* distances $x_1, x_2, \ldots, x_{\binom{n}{2}}$ between n trees can be more readily visualised and interpreted e.g. for identifying clustering or outliers.

The current implementation in R is rather slow because it relies on checking for the presence of all possible individual tips supplied in a data frame and repeated vector look-ups. Speeding up the implementation would be desirable and seems likely to be practical.

Finally, we note that our tree distance as we have defined it uses the mean of MRCA depths where there is more than one tip belonging to a given category; other scalar summaries such as the median could also be used. Further, it may be possible to achieve deeper comparisons by retaining the full sets of MRCA positions between tips of one type and those of another, rather than using scalar summaries. Comparing two such sets could be done using the symmetric set difference of the two sets of MRCA depths (in the case that these are discrete) or applying a smoothing operation to continuous depths and comparing sets of depths in the two trees using something like the Kullback-Leibler distance [27]. Additionally, the individual tip-pair contributions could have weighted contributions to the mean (or other summary statistic) according to, for example, the read counts from deep sequencing data, or a measure of importance or credibility on each tip.

$*$

References

[1] Aberer, A. J., D. Krompass, and A. Stamatakis. 2013. Pruning Rogue Taxa Improves Phylogenetic Accuracy: An Efficient Algorithm and Webservice. Systematic Biology 62:162–166.

[2] Amenta, N. and J. Klingner. 2002. Case study: visualizing sets of evolutionary trees. Pages 71–74 *in* IEEE Symposium on Information Visualization, 2002. (InfoVis'02).

[3] Ayabina, D., J. O. Ronning, K. Alfsnes, N. Debech, O. B. Brynildsrud, T. Arnesen, G. Norheim, A.-T. Mengshoel, R. Rykkvin, U. R. Dahle, C. Colijn, and V. Eldholm. 2017. Genome-based transmission modeling separates imported tuberculosis from recent transmission within an immigrant population. bioRxiv preprint.

[4] Billera, L. J., S. P. Holmes, and K. Vogtmann. 2001. Geometry of the Space of Phylogenetic Trees. Advances in Applied Mathematics 27:733–767.

[5] Bryant, J. M., S. R. Harris, J. Parkhill, R. Dawson, A. H. Diacon, P. van Helden, A. Pym, A. A. Mahayiddin, C. Chuchottaworn, I. M. Sanne, C. Louw, M. J. Boeree, M. Hoelscher, T. D. McHugh, A. L. C. Bateson, R. D. Hunt, S. Mwaigwisya, L. Wright, S. H. Gillespie, and S. D. Bentley. 2013. Whole-genome sequencing to establish relapse or re-infection with Mycobacterium tuberculosis: a retrospective observational study. The Lancet Respiratory Medicine 1:786–792.

[6] Cardona, G., A. Mir, F. Rossello Llompart, L. Rotger, and D. Sanchez. 2013. Cophenetic metrics for phylogenetic trees, after Sokal and Rohlf. BMC Bioinformatics 14:3.

[7] Chakerian, J. and S. Holmes. 2012. Computational tools for evaluating phylogenetic and hierarchical clustering trees. Journal of Computational and Graphical Statistics 21:581–599.

[8] Colijn, C. and G. Plazzotta. 2018. A Metric on Phylogenetic Tree Shapes. Systematic Biology 67:113–126.

[9] Cotton, J. A. and R. D. Page. 2006. The shape of human gene family phylogenies. BMC Evolutionary Biology 6:1–8.

[10] Cox, T. F. and M. A. A. Cox. 2000. Multidimensional scaling. CRC Press.

[11] Degnan, J. H. and N. A. Rosenberg. 2006. Discordance of species trees with their most likely gene trees. PLOS Genetics 2:e68.

[12] Degnan, J. H. and N. A. Rosenberg. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology and Evolution 24:332–340.

[13] Drummond, A. J. and A. Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology 7:214.

[14] Eldholm, V., J. H.-O. Pettersson, O. B. Brynildsrud, A. Kitchen, E. M. Rasmussen, T. Lillebaek, J. O. Rønning, V. Crudu, A. T. Mengshoel, N. Debech, K. Alfsnes,

J. Bohlin, C. S. Pepperell, and F. Balloux. 2016. Armed conflict and population displacement as drivers of the evolution and dispersal of Mycobacterium tuberculosis. Proceedings of the National Academy of Sciences 113:13881–13886.

[15] English, S., A. Katzourakis, D. Bonsall, P. Flanagan, A. Duda, S. Fidler, J. Weber, M. McClure, R. Phillips, and J. Frater. 2011. Phylogenetic analysis consistent with a clinical history of sexual transmission of HIV-1 from a single donor reveals transmission of highly distinct variants. Retrovirology 8:54.

[16] Fischer, W., V. V. Ganusov, E. E. Giorgi, P. T. Hraber, B. F. Keele, T. Leitner, C. S. Han, C. D. Gleasner, L. Green, C.-C. Lo, A. Nag, T. C. Wallstrom, S. Wang, A. J. McMichael, B. F. Haynes, B. H. Hahn, A. S. Perelson, P. Borrow, G. M. Shaw, T. Bhattacharya, and B. T. Korber. 2010. Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. PLOS One 5:e12303.

[17] Garba, M. K., T. M. W. Nye, and R. J. Boys. 2017. Probabilistic distances between trees. Systematic Biology Page syx080.

[18] Grandjean, L., R. H. Gilman, T. Iwamoto, C. U. Köser, J. Coronel, M. Zimic, M. E. Török, D. Ayabina, M. Kendall, C. Fraser, S. Harris, J. Parkhill, S. J. Peacock, D. A. J. Moore, and C. Colijn. 2017. Convergent evolution and topologically disruptive polymorphisms among multidrug-resistant tuberculosis in Peru. PLOS ONE 12:e0189838.

[19] Harding, E. F. 1971. The Probabilities of Rooted Tree-Shapes Generated by Random Bifurcation. Advances in Applied Probability 3:44–77.

[20] Heled, J. and R. R. Bouckaert. 2013. Looking for trees in the forest: summary tree from posterior samples. BMC evolutionary biology 13:221.

[21] Hillis, D. M., T. A. Heath, and K. St John. 2005. Analysis and visualization of tree space. Systematic biology 54:471–482.

[22] Holmes, S. 2006. Visualising data. Pages 197–208 in Statistical Problems in Particle Physics, Astrophysics and Cosmology, Proceedings of PHYSTAT05 (L. Lyons and M. K. Ünel, eds.) Imperial College Press.

[23] Huber, K. T., A. Spillner, R. Suchecki, and V. Moulton. 2011. Metrics on multilabeled trees: interrelationships and diameter bounds. IEEE/ACM Transactions on Computational Biology and Bioinformatics 8:1029–1040.

[24] Jombart, T., M. Kendall, J. Almagro-Garcia, and C. Colijn. 2018. treespace: Statistical Exploration of Landscapes of Phylogenetic Trees. R package version 1.1.0.

[25] Kendall, M. and C. Colijn. 2016. Mapping phylogenetic trees to reveal distinct patterns of evolution. Molecular Biology and Evolution 33:2735–2743.

[26] Kimura, M. 1980. Journal of Molecular Evolution A Simple Method for Estimating Evolutionary Rates of Base Substitutions Through Comparative Studies of Nucleotide Sequences. Journal of Molecular Evolution 16:111–120.

[27] Kullback, S. and R. A. Leibler. 1951. On Information and Sufficiency. The Annals of Mathematical Statistics 22:79–86.

[28] Lanfear, R., X. Hua, and D. L. Warren. 2016. Estimating the Effective Sample Size of Tree Topologies from Bayesian Phylogenetic Analyses. Genome Biology and Evolution 8:2319–2332.

[29] Lewitus, E. and H. Morlon. 2016. Characterizing and comparing phylogenies from their laplacian spectrum. Systematic Biology 65:495–507.

[30] Nichols, R. 2001. Gene trees and species trees are not the same. Trends in Ecology and Evolution 16:358–364.

[31] Paradis, E., J. Claude, and K. Strimmer. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.

[32] Paterson, G. K., E. M. Harrison, G. G. R. Murray, J. J. Welch, J. H. Warland, M. T. G. Holden, F. J. E. Morgan, X. Ba, G. Koop, S. R. Harris, D. J. Maskell, S. J. Peacock, M. E. Herrtage, J. Parkhill, and M. A. Holmes. 2015. Capturing the cloud of diversity reveals complexity and heterogeneity of MRSA carriage, infection and transmission. Nature Communications 6.

[33] Pavoine, S., S. Ollier, D. Pontier, and D. Chessel. 2008. Testing for phylogenetic signal in phenotypic traits: new matrices of phylogenetic proximities. Theoretical population biology 73:79–91.

[34] Poon, A. F. Y. 2015. Phylodynamic inference with kernel ABC and its application to HIV epidemiology. Molecular Biology and Evolution 32:2483–2495.

[35] Poon, A. F. Y., R. A. McGovern, T. Mo, D. J. H. F. Knapp, B. Brenner, J.-P. Routy, M. A. Wainberg, and P. R. Harrigan. 2011. Dates of HIV infection can be estimated for seroprevalent patients by coalescent analysis of serial next-generation sequencing data. AIDS 25:2019–2026.

[36] Poon, A. F. Y., L. W. Walker, H. Murray, R. M. McCloskey, P. R. Harrigan, and R. H. Liang. 2013. Mapping the shapes of phylogenetic trees from human and zoonotic RNA viruses. PLOS One 8:e78122.

[37] Price, M. N., P. S. Dehal, and A. P. Arkin. 2009. Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. Molecular Biology and Evolution 26:1641–1650.

[38] R Core Team. 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria.

[39] Ren, Y., S. Zha, J. Bi, J. A. Sanchez, C. Monical, M. Delcourt, R. Guzman, and R. Davidson. 2017. A combinatorial method for connecting BHV spaces representing different numbers of taxa. arXiv preprint 1708.02626.

[40] Roberts, H. E., J. Hurst, N. Robinson, H. Brown, P. Flanagan, L. Vass, S. Fidler, J. Weber, A. Babiker, R. E. Phillips, A. R. McLean, J. Frater, and SPARTAC trial investigators. 2015. Structured observations reveal slow HIV-1 CTL escape. PLOS Genetics 11:e1004914.

[41] Robinson, D. F. and L. R. Foulds. 1979. Comparison of weighted labelled trees. Lecture Notes in Mathematics 748:119–126.

[42] Robinson, D. F. and L. R. Foulds. 1981. Comparison of phylogenetic trees. Mathematical Biosciences 53:131–147.

[43] Roetzer, A., R. Diel, T. A. Kohl, C. Rückert, U. Nübel, J. Blom, T. Wirth, S. Jaenicke, S. Schuback, S. Rüsch-Gerdes, P. Supply, J. Kalinowski, and S. Niemann. 2013. Whole Genome Sequencing versus Traditional Genotyping for Investigation of a Mycobacterium tuberculosis Outbreak: A Longitudinal Molecular Epidemiological Study. PLoS Medicine 10:e1001387.

[44] Rosenberg, N. A. 2013. Discordance of species trees with their most likely gene trees: a unifying principle. Molecular Biology and Evolution 30:2709–2713.

[45] Sanderson, M. J., A. Purvis, and C. Henze. 1998. Phylogenetic supertrees: Assembling the trees of life. Trends in Ecology & Evolution 13:105–109.

[46] Schliep, K. 2011. phangorn: phylogenetic analysis in R. Bioinformatics 27:592–593.

[47] Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

[48] Steel, M. A. and D. Penny. 1993. Distributions of tree comparison metrics-some new results. Systematic Biology 42:126–141.

[49] Wickham, H. 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

[50] Williams, W. T. and H. T. Clifford. 1971. On the comparison of two classifications of the same set of elements. Taxon 20:519–522.

[51] Worby, C. J., P. D. O'Neill, T. Kypraios, J. V. Robotham, D. De Angelis, E. J. P. Cartwright, S. J. Peacock, and B. S. Cooper. 2016. Reconstructing transmission trees for communicable diseases using densely sampled genetic data. Annals of Applied Statistics 10:395–417.

[52] Worby, C. J. and T. D. Read. 2015. SEEDY (simulation of evolutionary and epidemiological dynamics): An R package to follow accumulation of Within-Host mutation in pathogens. PLOS One 10:e0129745.

[53] Zanini, F., J. Brodin, L. Thebo, C. Lanz, G. Bratt, J. Albert, and R. A. Neher. 2015. Population genomics of intrapatient HIV-1 evolution. eLife 4.