Models of deletion for visualizing bacterial variation: an application to tuberculosis spoligotypes

Josephine F Reyes[1,2], Andrew R Francis[3] and Mark M Tanaka[1,2]
[1] School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney 2052 Australia
[2] Evolution & Ecology Research Centre, University of New South Wales, Sydney 2052 Australia
[3] School of Computing and Mathematics, University of Western Sydney, South Penrith DC, NSW 1797 Australia

[1] School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney 2052 Australia
[2] Evolution & Ecology Research Centre, University of New South Wales, Sydney 2052 Australia
[3] School of Computing and Mathematics, University of Western Sydney, South Penrith DC, NSW 1797 Australia

Email: Josephine F Reyes*- j.reyes@student.unsw.edu.au; Andrew R Francis- a.francis@uws.edu.au; Mark M Tanaka - m.tanaka@unsw.edu.au;

*Corresponding author

Abstract

Background: Molecular typing methods are commonly used to study genetic relationships among bacterial isolates. Many of these methods have become standardized and produce portable data. A popular approach for analyzing such data is to construct graphs, including phylogenies. Inferences from graph representations of data assist in understanding the patterns of transmission of bacterial pathogens, and basing these graph constructs on biological models of evolution of the molecular marker help makes these inferences. Spoligotyping is a widely used method for genotyping isolates of *Mycobacterium tuberculosis* that exploits polymorphism in the direct repeat region. Our goal was to examine a range of models describing the evolution of spoligotypes in order to develop a visualization method to represent likely relationships among *M. tuberculosis* isolates.

Results: We found that inferred mutations of spoligotypes frequently involve the loss of a single or very few adjacent spacers. Using a second-order variant of Akaike's Information Criterion, we selected the Zipf model as the basis for resolving ambiguities in the ancestry of spoligotypes. We developed a method to construct graphs of spoligotypes (which we call spoligoforests). To demonstrate this method, we applied it to a tuberculosis data set from Cuba and compared the method to some existing methods.

Conclusions: We propose a new approach in analyzing relationships of *M. tuberculosis* isolates using spoligotypes. The spoligoforest recovers a plausible history of transmission and mutation events based on the selected deletion model. The method may be suitable to study markers based on loci of similar structure from other bacteria. The groupings and relationships in the spoligoforest can be analyzed along with the clinical features of strains to provide an understanding of the evolution of spoligotypes.

Background

The visualization of relationships among genotypes of bacterial isolates is a useful approach to addressing both evolutionary and epidemiological questions. Inferences from graph representations of data assist in understanding the patterns of transmission of bacterial pathogens. Presently, there are two approaches to visualization. The first class of methods is sequence-based, and these methods often produce phylogenetic trees, or dendrograms. These trees are used extensively to represent relatedness of isolates that have been identified by almost any typing procedure. However, the models of sequence evolution upon which phylogenetic methods depend are not appropriate for many markers that are not sequence-based. Also, because many of these markers evolve rapidly enough to generate intra-specific variation, it is preferable to show direct relationships between genotypes. Bacterial isolates often cluster into the same genotype, and dendrograms are not suited to showing these clusters.

The second class of methods produce "network-like" graphs that show direct relationships between clusters of genotypes. Some examples of this second class of methods are found in the works of Zhu et al. [1], Tanaka and Francis [2] and Excoffier and Smouse [3]. The genoclouds in Zhu et al. [1] consist of related isolates of *Nesseria meningitidis* that are grouped according to criteria that minimize genetic, temporal and physical distances. The result is a parsimonious tree that depicts the relationships between the genoclouds. Similarly, Tanaka and Francis [2] proposed cluster-graphs where isolates of *Mycobacterium tuberculosis* sharing the same genotype are assigned into clusters, and all possible close relationships between these clusters are shown. In both these methods, clusters are associated with epidemiologically linked cases of infection. Guernier et al. [4] developed a technique of representation that is based on the cluster-graph, with two additional elements included: (1) concentric circles to show the number of possible mutation steps between spoligotypes, and (2) hypothetical intermediate nodes to visualize possible links between clades of spoligotypes known to be related. Excoffier and Smouse [3] used an analysis of molecular variance to construct minimum spanning trees and networks to represent genetic relatedness. See Posada et al. [5] for a more general discussion of graphical methods to represent relationships.

The eBurst package [6, 7] is designed for visualizing data from multi-locus sequence typing (MLST). Isolates that have similar sequence types are assigned to disjoint groups, where similarity depends on the number of shared alleles in the MLST profile. The radial layout of eBurst diagrams shows groups or subgroups of related genotypes, centered around the inferred founding genotypes. The complexity of an eBurst diagram suggests the age of the clonal complex; a clonal complex is considered young when its structure is simple and older when its structure is complex. There are many genotyping technologies enabling the study of genetic variation in bacteria. Here, we focus on spacer oligonucleotide typing (spoligotyping), a technique that exploits polymorphism in the direct repeat (DR) region of *M. tuberculosis* [8, 9]. This method has gained widespread use for differentiating isolates of *M. tuberculosis* over the last decade [10]. The DR region is composed of numerous identical 36-base-pair direct repeats, interspersed by nonrepetitive short sequences or direct variable repeats (DVRs) called spacers. Mechanisms known to cause variation in this locus are homologous recombination between adjacent or spatially distant DRs leading to deletion, and transposition and recombination of IS*6110* elements in the DR locus [9, 11]. The DR locus has been identified as a hotspot for the integration of insertion elements in the chromosome of *M. tuberculosis* complex strains [12]. Such insertion into a spacer sequence can lead to the apparent deletion of that spacer [13]. It is presumed that spacers cannot be recovered when lost, since there is little or no recombination observed between strains [14].

It is possible that there is a relationship between the presence of some number of spacers in specific positions and the transmission rate of a strain, as seen by the lack of a length of spacers in the W-Beijing strain, which is prevalent in many data sets. However, in our model, we assume that a deletion event has no relation with the transmission rate, and any such relationship is beyond the scope of our study. We refer the reader to papers that discuss the importance of studying the W-Beijing type and its transmission [15, 16].

In this paper we examine the problem of determining a plausible evolutionary history of a sample of tuberculosis spoligotypes using an explicit model of the evolution of the DR locus. We start with the cluster-graph construct of Tanaka and Francis [2] to represent all possible mutation events in a sample of spoligotypes. Nodes of a cluster-graph represent distinct spoligotypes in a sample, and edges drawn between nodes determine the possible mutation events. By mutation we mean a deletion of one or more spacers in a spoligotype. Even for moderately large samples, this can lead to a tangled network of relationships between spoligotypes, which can hinder further analysis. In particular, many spoligotypes appear to have originated from multiple parent spoligotypes. One solution to this problem is to randomly sample edges from a set of multiple inbound edges [17]. However, some edges (mutation events) may be more likely than others to explain the origin of a given spoligotype. We formulate a variety of models to describe the deletion processes that generate variation in the DR locus, and identify an appropriate model using Akaike's Information Criterion. The selected model can then be used to choose a single inbound edge into a specific spoligotype. Applying this procedure to each spoligotype with multiple inbound edges in a sample, we can refine the cluster-graph. We call the resulting graph a spoligoforest.

Methods

We present several candidate deletion models of spoligotype evolution, then compare them using a second-order form of the Akaike's Information Criterion (AIC_c) and data from selected published spoligotype samples (see Table 1). In this section we begin with the underlying assumptions about spoligotypes and their evolution. We then outline the procedure for model selection and finally describe the models.

Assumptions on the evolution of spoligotypes

A spoligotype consists of 43 binary characters. Each binary character denotes the presence or absence of a spacer in the DR locus of *M. tuberculosis*. The copy number of a spoligotype refers to the number of spacers present in its binary pattern. It is assumed that a mutation event involves the deletion of any number of adjacent spacers from the spoligotype; deleted spacers are not recovered, so that the spoligotype resulting from a mutation always contains fewer spacers than the parent. For our purposes we regard the different mechanisms that influence variation in the DR locus to be indistinguishable. Furthermore, deletion is the only source of variation. In our model we assume that a deletion event has no relation with the transmission rate. We assume that the mutation rate is low enough relative to the transmission rate that infected individuals carry only a single strain of *M. tuberculosis* with a specific spoligotype. When this infecting strain mutates, it is replaced by a strain with a different spoligotype that has not been observed elsewhere in the sample. Consequently, in any sample, a given spoligotype can have at most one possible parent spoligotype, but potentially many descendants.

Data sets and cluster-graphs

Given that spoligotypes mutate by deletions of adjacent blocks of spacers, we would like to know whether some lengths of deleted adjacent blocks are more probable than others. Specifically, we would like to find the frequency of each deletion size. There are many published tuberculosis data sets using spoligotyping as a marker containing the information required for our purposes. We selected fifteen published data sets that provide the spoligotyping pattern of each *M. tuberculosis* isolate in the sample and the number of isolates that cluster into each pattern (see Table 1). We consider that individuals within a sample are sufficiently close to each other for transmission to occur. These data sets come from various parts of the world, and vary in statistical features such as RTI_{n-1} (in the range (0.3279,0.8055)), number of singletons (7,105), average cluster size (1.8,6.7) and θ-estimate (2.73,66.25). Some of these quantities are discussed in [2] and [10].

We use cluster-graphs as described in Tanaka and Francis [2]. We group isolates that have the same spoligotype into clusters; each cluster is drawn as a node, and a possible single-event deletion that relates two clusters (spoligotypes) is represented by a directed edge. Possible deletion events are established by pairwise comparisons of spoligotype patterns. We define a *spoligoforest* to be a cluster-graph in which a single parent is chosen for each cluster having one or more parents. Some clusters in a cluster-graph already have a unique parent, and are likely to represent true deletions. This set of unambiguous deletion events forms the sample of deletion lengths for the model selection. Table 2 (column 4) shows the number of such edges from each data set. We assume that mutations occur independently of the state of the population, and hence edges, which represent mutations, are independent. The edges from the different data sets, representing independent deletion events, are pooled together in order to find analyze the frequency of deletion lengths. We obtain an empirical distribution of deletion lengths represented by the unambiguous edges from the fifteen data sets. The total pool of analyzed unambiguous edges consisted of 339 deletion events.

Candidate models for spacer deletion length in pattern mutations.

Our goal is to find a model that best describes the underlying process generating the distribution of lengths of spacer deletions in the inferred mutation events. We formulated several candidate models based on standard discrete distributions and various hypotheses on spacer deletion lengths. For each model we found the maximum likelihood estimators (MLEs) of the parameters, analytically when possible, and numerically otherwise. Let the observed number of deletions of length i be x_i, where i can take values from

1 to 43, let m be the total number of mutations analyzed (the unambiguous edges, in this case 339), and let $\bar{x}$ be the mean deletion length. Let the random variable K describe the deletion length associated with a mutation event. In each of the candidate models, let $P(K=k)$ (or $P(k)$) be the probability mass function. The corresponding likelihood function is $\mathcal{L}(\mathbf{p}|x) = \prod_{k=1}^{\infty} [P(k)]^{x_k}$, where $\mathbf{p}$ is the vector of parameters, x is the frequency of deletion lengths collected from the data sets, k is the deletion length, and x_k is the frequency of the class of deletions with length k.

For each of the models, we computed the value of the second-order variant of Akaike's Information Criterion (AIC_c) [18] to select a parsimonious model. The AIC_c is given by the formula

$$AIC_c = -2\ln\mathcal{L}(\hat{\mathbf{p}}|x) + \frac{2ms}{m-s-1} \tag{1}$$

where $\mathcal{L}$ is the likelihood, $\hat{\mathbf{p}}$ is the vector of parameters at the maximum likelihood, x is the frequency of deletion lengths collected from the data sets, m is the sample size (the number of edges) and s is the number of parameters in the model. Models with low relative AIC_c values are favoured. A summary of the probability mass functions and (MLEs) can be found in Table 6.

Geometric model. Consider a deletion mechanism that moves along the DR region, making independent attempts to remove a spacer. Let k be the number of spacers this mechanism is able to remove, and the constant probability of the removal of a spacer is p. The probability mass function is $P(k) = p^{k-1}(1-p)$, where $k \geq 1$ and $0 < p < 1$. The MLE for p is $\hat{p} = 1 - 1/\bar{x}$.

Negative binomial model. To generalize the geometric model, we define the negative binomial parameters p and r. Consider a deletion mechanism that involves r rounds of spacer deletion, so that in total k spacers are removed. Each of these r rounds removes a geometrically distributed number of spacers with parameter p. The probability mass function is

$$P(k) = \frac{(1-p)^r}{1-(1-p)^r}\binom{k+r-1}{r-1}p^k \tag{2}$$

where $k, r \geq 1$ and $0 < p < 1$. The MLEs were found by solving $\hat{p} = 1 - \hat{r}/\bar{x}$ for $\hat{r}$ and $\hat{p}$. This was done by considering the equations conditionally on $1 \leq \hat{r} \leq 10$ to solve for the values of $\hat{p}$ and inspecting the likelihood.

Conditional Poisson model. In this model, the deletion process results in the loss of k spacers distributed as a Poisson parameter $\lambda > 0$, conditional on $k \geq 1$. The probability mass function is

$$P(k) = e^{-\lambda}\frac{\lambda^k}{k!(1-e^{-\lambda})} \tag{3}$$

with MLE $\hat{\lambda}$ such that $\bar{x} = \hat{\lambda}/(1-e^{-\hat{\lambda}})$.

Logarithmic model. Suppose the deletion process causes the loss of k spacers, following a logarithmic distribution, given by the probability mass function

$$P(k) = -\frac{\theta^k}{k\log(1-\theta)}, \tag{4}$$

with $0 < \theta < 1$ and $k \geq 1$. The MLE is found by solving the equation $\bar{x} = \frac{-\hat{\theta}}{(1-\hat{\theta})\log(1-\hat{\theta})}$ for $\hat{\theta}$.

Zeta model. Suppose the deletion process results in the loss of k spacers distributed as a zeta parameter ρ. The probability mass function is

$$P(k) = \frac{k^{-\rho}}{\sum_{d=1}^{\infty} d^{-\rho}} \tag{5}$$

with $\rho > 1$ and $k \geq 1$. The MLE $\hat{\rho}$ can be calculated numerically.

Zipf model. If we restrict the loss of k spacers to the interval $1 \leq k \leq 43$, then the probability mass function for the zeta model can be written with a finite sum in its denominator, i.e.,

$$P(k) = \frac{k^{-p}}{\sum_{d=1}^{43} d^{-p}}. \tag{6}$$

The MLE $\hat{p}$ can be found numerically.

Uniform model. We consider a deletion process that cuts spacers in lengths distributed uniformly across deletion lengths k up to some endpoint a. The probability mass function is $P(k) = 1/a$ where k and a are integers such that $1 \leq k \leq a$ and $1 \leq a \leq 43$. We denote this model as Uniform (V), where the endpoint a is a parameter, and is allowed to vary. The MLE $\hat{a}$ was found numerically. We also included a variation of the uniform model where $P(k) = p$ for all values of $1 \leq k \leq 43$, with MLE $\hat{p}$. We refer to this model as Uniform (F), where the endpoint a is fixed at $a = 43$.

Empirical model. We include in the analysis a model that completely represents the empirical values of deletion frequencies in the fifteen data sets used as reference (see Table 1). The probability mass function is $P(k) = \frac{x_k}{m}$ where x_k is the frequency of k-deletions and m is the total number of deletions, with $1 \leq k \leq 43$. Setting the parameters to be p_k for each k, the values of the MLEs $\hat{p_k}$ are found analytically. This model has 42 free parameters.

Results

We begin with some general observations about the relative frequencies of different deletion lengths. We discuss the outcomes of the model selection procedure, and then apply the chosen model to a new visualization method for representing relationships of isolates of tuberculosis.

Inferred pattern mutations of spoligotypes frequently involve short spacer deletions.

The selected data sets independently show a high frequency of deletions of a single spacer. The pooled set of edges are shown in the gray bars of Figure 1. The 339 single-event deletions are distributed as shown in Table 3. Deletion lengths not appearing in Table 3 are not observed. The average number of spacers deleted is $\bar{x} = 2.46$ and the standard deviation is 3.376. The skewed distribution of deletion sizes indicate a high number of short deletions, and very few longer deletions. Note, it is conceivable that spoligotypes that exist in the population but not sampled are intermediate in state between two sampled spoligotypes. If such spoligotypes exist and are sampled then the distribution would shift further towards shorter deletion lengths.

Model selection.

We selected the model with the lowest value of AIC_c (see Equation 1). This model is the Zipf model (Equation (6)). AIC_c values and parameter estimates for some of the models are shown in Table 4.
Figure 1 shows a plot of the relative frequencies for deletion lengths estimated by each of the models as well as the actual empirical values (gray bars). We verified that the selected model fits the individual data sets well by repeating the analysis separately for the individual data sets. The Zipf model often has the lowest AIC_c, while the logarithmic series and geometric models are selected in some individual data sets (see Table 5).

Visualizing relationships among spoligotypes.

The selected model can now be used in a method to visualize relationships among *M. tuberculosis* isolates.
For a specific data set consisting of *M. tuberculosis* isolates typed using spoligotyping, we represent each spoligotype by a node with area proportional to the number of isolates with that spoligotype pattern.
Inferred possible mutation events are represented by directed edges between nodes, with the arrowheads pointing to descendant spoligotypes. This specifies the cluster-graph [2].
Multiple inbound edges into a node are reduced to a single inbound edge. We use a heuristic that chooses a single inbound edge that has maximum weight. We define the weight w of an edge e_{AB} in a cluster-graph from spoligotype A to its descendant B to be:

$$w(e_{AB}) = \frac{P(d(e_{AB})) \times n_A}{\sum_{i \in S} P(d(e_{iB})) \times n_i} \tag{7}$$

where $P(\cdot)$ is the model of deletion, $d(\cdot)$ is the deletion length represented by the edge, n_i is the cluster size of spoligotype i, and S is the set of nodes that are possible parents nodes of spoligotype B. Ties in the maximum weight are broken arbitrarily. The resulting graph is what we refer to as a spoligoforest. Code for automatically constructing cluster-graphs and spoligoforests from a sample of tuberculosis spoligotypes was implemented using the visualization software GraphViz [19] The method has been implemented on a web server and is publicly available at http://emi.unsw.edu.au/spolTools (see [20] for details).

Application of the method to tuberculosis spoligotypes

We applied our new method for constructing forests to data from several data sets. To illustrate the method, we first use published data from a study on the transmission of tuberculosis in Cuba [21]. Isolates collected over a year were typed using both spoligotyping and IS*6110* typing. One-hundred and fifty-seven isolates were classified into 47 spoligotype patterns. The clusters of isolates sharing the same spoligotype are nodes in the diagram, labelled using shared type (ST) numbers in SpolDB4 [22] wherever possible. When the spoligotype is not found in SpolDB4, we labelled it as 'Or' with a number (e.g. Or1). Orphan spoligotypes are unique alleles without an ST number [23]. Following the description in Tanaka and Francis [2], we constructed the cluster-graph for these data in a hierarchical layout as shown in Figure 2, with edges labelled with the weights computed using our selected model. The size of each node reflects the number of isolates in that node. The resulting graph is a complex network showing all possible relationships of spoligotypes under our assumptions about the spoligotype mutation processes. The Zipf model is used to calculate the weights of the edges, as given in Equation (7), of the cluster-graph. In this cluster-graph, there are 19 nodes with multiple inbound edges. The nodes are labelled according to the shared type (ST) identifiers used in SpolDB4 [22]. For example, ST 718 has 18 possible parents, while STs 47, 1, 62, 791, 2, 132 and 209, each has 3 possible parents. Of the 83 edges in the cluster-graph, 37 were chosen for the spoligoforest (see Figure 3). As with the cluster-graph, the nodes in the spoligoforest represent the number of isolates that share the same spoligotype pattern. If the weight of the edge is equal to 1, we draw a solid edge, if the weight is greater than or equal to 0.5 but less than 1, a dashed line, and if less than 0.5, a dotted line.

The spoligoforest consists of two trees (connected components) and eight disconnected nodes. The large tree has ST 53 at the root, suggesting that ST 53 is the oldest spoligotype in this tree. Seven spoligotypes are descended from ST 53, two of which have comparably large cluster sizes: ST 50 with 16 isolates and ST 42 with 14 isolates. These two spoligotypes form two distinct lineages diverging from ST 53. A comparison with the families in SpolDB4 [22] identified these two lineages to be the Haarlem and LAM (Latino-American and Mediterranean) families. ST 50 and its descendants ST 47 and ST 3 belong to the Haarlem family of strains, whereas ST 42 and its descendants STs 81, 20, 74, 33, and 17 are from the LAM family. ST 80, a disconnected node representing 6 isolates, is also of the Haarlem family. The separate smaller tree on the right includes ST 1 with 20 isolates. This is the spoligotype of the W-Beijing strain, known to be widely distributed around the world.

Comparative analyses

In this section, we compare the spoligoforest to two other methods of visualisation, namely phylogenies and cladograms. We illustrate that using models with AIC_c values close to that of the Zipf model has minimal effect on the edges of a spoligoforest.

The branches in a phylogeny show indirect relationships between isolates via implicit common ancestors, whereas the edges in the spoligoforest describe direct relationships among clusters of spoligotypes. However, related spoligotypes in the spoligoforest are consistent with inferences on clustered isolates from a phylogeny. Figure 4 shows a phylogenetic tree based on IS*6110*-typing and Figure 5 is a spoligoforest using data from a prison in Azerbaijan [24]. The tree depicts genetic relatedness of isolates with each other based on similarities of IS*6110* banding. The leaves of this tree have been renamed using STs (shared types from SpolDB4) of the spoligotypes, so that isolates sharing the same spoligotype may appear in different leaves of the tree. An inspection of the branch lengths in the phylogenetic tree indicate that ST 42 is most related to ST 254 (2 isolates of differing IS*6110* bands.). The spoligoforest in Figure 5 is consistent with this observation: ST 42 is chosen as a parent for ST 254, with weight 0.7064 in the cluster-graph (not shown). Furthermore, the spoligoforest shows that ST 42 is likely to have evolved from ST 53, which cannot be inferred from Figure 4. Direct links between spoligotypes are also seen in the spoligoforest, for instance the edge from ST 35 to ST 1050. In the phylogenetic tree, this relationship can be seen in the leftmost group with STs 35, 62 and 1050. Also, ST 1051 shown to be distant from the other types. It may be worth investigating whether ST 1051 is more related to the ST 53 group, as shown in the spoligoforest.

A similar network-like technique of visualization to the spoligoforest is the cladogram in Figure 6 of [25]. The method of construction of the cladogram involves using information from nested clades and geographic location. The main difference between the cladogram and the spoligoforest is that the cladogram involves the introduction of intermediate steps between types, accounting for possibly unsampled spoligotypes. The spoligoforest for this data set is shown in Figure 6. The LAM3 and LAM9 groups identified in the cladogram are also evident as a subtree in the spoligoforest, with ST 42 at the top of this subtree (see highlighted region in Figure 6). The relationship of ST G4 with other spoligotypes is different in the two figures. In the spoligoforest, ST G4 is linked with a dotted line (computed weight of 0.2387 in cluster-graph, not shown) to ST 42. In the cladogram, however, ST G4 is related to ST 45 through a conjectured intermediate type. It may be interesting to assess whether ST G4 may be more related to the LAM3 and LAM9 groups than is shown by the cladogram.

In order to assess the choice of model among the best four, we applied the method to several data sets using a range of possible models. This procedure has revealed that model selection has minimal impact on the edges of spoligoforests. We constructed the spoligoforests for six data sets, using the Zeta, Logarithmic series, Geometric and Empirical models. Table 7 shows the number of differing edges in spoligoforests constructed from these alternative models, relative to that constructed using the selected model (Zipf). Clearly, the Zipf and Zeta models are similar, as the only difference between them is that the domain of the Zipf distribution is finite (see Table 6). The spoligoforest for the data set from Madagascar [26] using the selected Zipf model is shown in Figure 7. The spoligoforest using a Logarithmic series model (Figure 8) for the same data set differs from Figure 7 by 4 edges, the highest number of edge differences among the data sets and models we tested.

Discussion

This paper proposes a new method of visualizing the relationships among genotypes of tuberculosis by selecting a model of evolution of spoligotypes. The selected model is the Zipf model with parameter p for deletion length. We have made the spoligoforest application available in the spolTools website

`http://www.emi.unsw.edu.au/spolTools.`

In this website, users can search through the repository of spoligotype data sets in spolTools as well as manipulate their own data sets. These data sets can be processed to construct spoligoforests.

A spoligoforest recovers a plausible history of transmission and mutation events. The area of each node is proportional to the number of isolates (cluster size); edges between nodes reflect evolutionary relationships between spoligotypes with arrowheads pointing to descendants. A single edge is chosen from multiple inbound edges using the deletion model, resulting in a forest – that is, a collection of acyclic graphs, or trees.

Information about the age of a spoligotype is contained in three aspects of a spoligoforest. First, if its node is large, the strain with that spoligotype may have been transmitted extensively over a long time. Second, a large number of descendants (outbound edges) suggests the strain has had a long period over which to generate mutations. Third, the location of a node also offers clues as to age: the closer it is to a root node, the older it is. For example, ST 1 in Figure 3 is a root and potentially the oldest spoligotype in this forest. If a spoligotype node size is large yet located at a tip of the spoligoforest, this mixed signal may indicate that the strain with the spoligotype is transmitting faster than the other strains in the data set [17]. For example, ST 42 in Figure 3 has cluster size 14 and 6 outbound edges, whereas ST 81 with 10 isolates only has 1 outbound edge. ST 81 could therefore be an "emerging strain". Application of the analysis of Tanaka and Francis [17] did not, however, identify any rapidly spreading strains in this data set. In this case, therefore, there is no evidence for the presence of emerging strains. Note that the algorithm for choosing edges proposed in this study could be used to refine the method of Tanaka and Francis [17].

One way to improve the analysis of strain age and emergence would be to consider spoligotypes in conjunction with other molecular markers. For example, consider the same two spoligotypes discussed above (ST 42 and ST 81 from Figure 3). ST 42 has 12 different IS*6110* profiles in the data set we used, while ST 81 has only one. This suggests once again that ST 81 may be associated with a higher transmission rate than ST 42. Further quantitative analysis would be needed to verify this point.

We note the limitations of our method. First, in choosing a single edge from multiple edges, we assumed that homoplasy (i.e., a spoligotype arising from more than one parent) does not occur. Because the number of spacers is finite and the deletion process is discrete, homoplasy may occasionally occur, but it is likely to be infrequent. The occurrence of homoplasy may have only a minor effect on graph-construction, producing a small number of cycles if such events could be properly identified. Second, we always choose one edge (parent) among possible inbound edges into a given spoligotype. It is conceivable, however, that the given spoligotype did not descend from any of the candidate parents. An improvement to the method would incorporate a procedure for not choosing any edges when appropriate. Third, as in any statistical analysis involving samples of data, there could be a bias in sampling. An overrepresentation of a spoligotype in a sampled data set can lead to biased selection of a parent node. Fourth, our methodology cannot be applied to markers such as Variable Numbers of Tandem Repeats (VNTR), which is commonly used to type various bacteria. The mutation process for VNTRs is better modeled using a stepwise mutation model rather than a deletion model.

Our method may, however, be suitable for markers based on loci of similar structure in some other bacteria. The direct repeat region of *M. tuberculosis* is among a family of repetitive genome sequences that are called Clustered Regularly Interspersed Short Palindromic Repeats (CRISPRs) found in many different species of bacteria and archaea [27–29]. Recently, CRISPR systems have received increased attention due to evidence that links these loci with the acquisition of resistance in bacteria to infection by phages [30]. Examples of these structures have been studied in *M. tuberculosis* [12,13], *Haloferax mediterranei* [31], *Methanocaldococcus jannaschii* [27,32], and *Yersinia pestis* [33].

The mechanisms that are believed to be involved in the evolution of CRISPR systems involve a frequent deletion of spacer-repeat motifs (thought to be necessary to prevent over-inflation of the CRISPR locus [29]) as well as the insertion of new spacers next to the leader sequence due to uptake of phage DNA [33].

Typing methods similar to spoligotyping for other bacterial isolates with CRISPRs are being developed. One such typing method (also called spoligotyping) has been applied to *Corynebacterium diphtheriae* strains, in which the location and structure of two CRISPR loci have been identified [34]. These loci consist of 27 spacers (the DRA with 21 spacers and the DRB with 6 spacers) in two different regions of the genome. The spoligotyping method used in this particular study is similar to the method used for *M. tuberculosis*. At present, there is yet to be an analysis of the evolution of these DR loci in *C. diphtheriae*. It has also been speculated that in some CRISPRs, new repeat motifs can appear, like those in *Yersinia pestis* [33]. Investigations into how these loci evolve may allow the development of methods similar to that described here.

As with other visualization methods, the groupings and relationships that are seen in the spoligoforest can be analysed along with the known clinical features of strains. Such analyses are valuable when an understanding of the history of transmission and mutation of strains is required.

Conclusions

There is a lack of tools for visualizing relationships among tuberculosis isolates that employ a model describing evolution of a specific marker. Current understanding of the evolution of spoligotypes led us to a method for visualizing relationships of isolates within a sample. The methodology presented in this paper may be applied to loci that have the same structure as the DR region of *Mycobacterium tuberculosis*, and whose evolution involves the deletion of spacer-repeat motifs. The groupings and relationships that are seen in the spoligoforest can be analysed along with the clinical features of strains to understand the evolution of strains.

Authors' contributions

JR implemented all the methods, performed the analytical and computational work and wrote the initial draft of the manuscript. MT, AF and JR designed the study and edited the manuscript.

Acknowledgements

We thank Fabio Luciani for helpful discussions and comments on the manuscript. This work was supported by the Australian Research Council through the Discovery scheme (DP0556732).

References

1. Zhu P, van der Ende A, Falush D, Brieske N, Morelli G, Linz B, Popovic T, Schuurman I, Adegbola R, Zurth K, Gagneux S, Platonov A, Riou J, Caugant D, Nicolas P, Achtman M: **Fit genotypes and escape variants of subgroup III *Neisseria meningitidis* during three pandemics of epidemic meningitis**. *Proc Natl Acad Sci* 2001, **98**(9):5234–5239.
2. Tanaka MM, Francis AR: **Methods of quantifying and visualising outbreaks of tuberculosis using genotypic information**. *Infect Genet Evol* 2005, **5**:35–43.
3. Excoffier L, Smouse P: **Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: molecular variance parsimony**. *Genetics* 1994, **136**:343–359.
4. Guernier V, Sola C, Brudey K, Guegan JF, Rastogi N: **Use of cluster-graphs from spoligotyping data to study genotype similarities and a comparison of three indices to quantify recent tuberculosis transmission among culture positive cases in French Guiana during a eight year period.** *BMC Infect Dis* 2008, **8**:46.
5. Posada D, Crandall K: **Intraspecific gene genealogies: trees grafting into networks**. *Trends Ecol Evol* 2001, **16**:37–44.
6. Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG: **eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data**. *J Bacteriol* 2004, **185**(5):1518–1530.
7. Spratt BG, Hanage WP, Li B, Aanensen DM, Feil EJ: **Displaying the relatedness among isolates of bacterial species – the eBURST approach**. *FEMS Microbiol Lett* 2004, **241**(2):129–134.
8. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M, van Embden J: **Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology**. *J Clin Microbiol* 1997, **35**(4):907–914.
9. Groenen PM, Bunschoten AE, van Soolingen D, van Embden JD: **Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method**. *Mol Microbiol* 1993, **10**(5):1057–1065.
10. Luciani F, Francis AR, Tanaka MM: **Interpreting genotype cluster sizes of *Mycobacterium tuberculosis* isolates typed with IS*6110* and spoligotyping**. *Infect Genet Evol* 2008, **8**(2):182–190.
11. Fang Z, Morrison N, Watt B, Doig C, Forbes KJ: **IS*6110* transposition and evolutionary scenario of the direct repeat locus in a group of closely related *Mycobacterium tuberculosis* strains**. *J Bacteriol* 1998, **180**(12):2102–2109.
12. Hermans PW, van Soolingen D, Bik EM, de Haas PE, Dale JW, van Embden JD: **Insertion element IS*987* from *Mycobacterium bovis* BCG is located in a hot-spot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains**. *Infect Immun* 1991, **59**(8):2695–2705.
13. Warren RM, Streicher EM, Sampson SL, van der Spuy GD, Richardson M, Nguyen D, Behr MA, Victor TC, van Helden PD: **Microevolution of the direct repeat region of *Mycobacterium tuberculosis*: implications for interpretation of spoligotyping data.** *J Clin Microbiol* 2002, **40**(12):4457–4465.

14. van Embden JD, van Gorkum T, Kremer K, Jansen R, van Der Zeijst BA, Schouls LM: **Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria**. *J Bacteriol* 2000, **182**(9):2393–2401.

15. Bifani PJ, Mathema B, Kurepina NE, Kreiswirth BN: **Global dissemination of the *Mycobacterium tuberculosis* W-Beijing family strains.** *Trends Microbiol* 2002, **10**:45–52.

16. Kremer K, Glynn JR, Lillebaek T, Niemann S, Kurepina NE, Kreiswirth BN, Bifani PJ, van Soolingen D: **Definition of the Beijing/W lineage of *Mycobacterium tuberculosis* on the basis of genetic markers.** *J Clin Microbiol* 2004, **42**(9):4040–4049.

17. Tanaka MM, Francis AR: **Detecting emerging strains of tuberculosis by using spoligotypes.** *Proc Natl Acad Sci* 2006, **103**(41):15266–15271.

18. Burnham K, Anderson D: *Model selection and inference*. New York: Springer-Verlag, 2nd. edition 1998.

19. Gansner ER, North SC: **An open graph visualization system and its applications to software engineering**. *Software Pract Exper* 1999, **1**:1–5.

20. Tang C, Reyes JF, Luciani F, Francis AR, Tanaka MM: **spolTools: online utilities for analyzing spoligotypes of the *Mycobacterium tuberculosis* complex.** *Bioinformatics* 2008, **doi: 10.1093/bioinformatics/btn434**.

21. Diaz R, Kremer K, de Haas PE, Gomez RI, Marrero A, Valdivia J, van Embden JD, van Soolingen D: **Molecular epidemiology of tuberculosis in Cuba outside of Havana, July 1994-June 1995: utility of spoligotyping versus IS*6110* restriction fragment length polymorphism.** *Int J Tuberc Lung Dis* 1998, **2**(9):743–750.

22. Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, Al-Hajoj SA, Allix C, Aristimuno L, Arora J, Baumanis V, Binder L, Cafrune P, Cataldi A, Cheong S, Diel R, Ellermeier C, Evans JT, Fauville-Dufaux M, Ferdinand S, de Viedma DG, Garzelli C, Gazzola L, Gomes HM, Guttierez MC, Hawkey PM, van Helden PD, Kadival GV, Kreiswirth BN, Kremer K, Kubin M, Kulkarni SP, Liens B, Lillebaek T, Ho ML, Martin C, Martin C, Mokrousov I, Narvskaia O, Ngeow YF, Naumann L, Niemann S, Parwati I, Rahim Z, Rasolofo-Razanamparany V, Rasolonavalona T, Rossetti ML, Rusch-Gerdes S, Sajduda A, Samper S, Shemyakin IG, Singh UB, Somoskovi A, Skuce RA, van Soolingen D, Streicher EM, Suffys PN, Tortoli E, Tracevska T, Vincent V, Victor TC, Warren RM, Yap SF, Zaman K, Portaels F, Rastogi N, Sola C: ***Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology.** *BMC Microbiol* 2006, **6**:23.

23. Sola C, Filliol I, Legrand E, Lesjean S, Locht C, Supply P, Rastogi N: **Genotyping of the Mycobacterium tuberculosis complex using MIRUs: association with VNTR and spoligotyping for molecular epidemiology and evolutionary genetics.** *Infect Genet Evol* 2003, **3**(2):125–133.

24. Pfyffer G, Strassle A, van Gorkum T, Portaels F, Rigouts L, Mathieu C, Mirzoyev F, Traore H, van Embden J: **Multi-drug resistant tuberculosis in prison inmates, Azerbaijan**. *Emerg Infect Dis* 2001, **7**(5):855–861.

25. Duchene V, Ferdinand S, Filliol I, Guegan J, Rastogi N, Sola C: **Phylogenetic reconstruction of *Mycobacterium tuberculosis* within four settings of the Carribean region: tree comparative analyse and first appraisal on their phylogeography**. *Infect Genet Evol* 2004, **4**:5–14.

26. Ferdinand S, C CS, Chanteau S, Ramarokoto H, Rasolonavalona T, Rasolofo-Razanamparany V, Rastogi N: **A study of spoligotyping-defined *Mycobacterium tuberculosis* clades in relation to the origin of peopling and the demographic history in Madagascar.** *Infect Genet Evol* 2005, **5**(4):340–348.

27. Jansen R, van Embden J, Gaastra W, Schouls L: **Identification of a Novel Family of Sequence Repeats among Prokaryotes**. *OMICS* 2002, **6**:23–33.

28. Kunin V, Sorek R, Hugenholtz P: **Evolutionary conservation of sequence and secondary structures in CRISPR repeats**. *Genome Biol* 2007, **8**(4):R61.

29. Grissa I, Vergnaud G, Pourcel C: **The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats**. *BMC Bioinformatics* 2007, **8**:172.

30. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P: **CRISPR provides acquired resistance against viruses in prokaryotes.** *Science* 2007, **315**(5819):1709–1712.

31. Mojica FJ, C F, Juez G, Rodriguez-Valera F: **Long stretches of short tandem repeats are present in the largest replicons of Archaea *Haloferax mediterranei* and *Haloferax volcanii* and could be involved in replicon partitioning**. *Mol Microbiol* 1995, **17**:85–93.

32. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, Fitzgerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb JF, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen NS, Venter JC: **Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii***. *Science* 1996, **273**(5278):1058–73.

33. Pourcel C, Salvignol G, Vergnaud G: **CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA and provide additional tools for evolutionary studies**. *Microbiology* 2005, **151**:653–663.

34. Mokrousov I, Limeschenko E, Narvskaya O: ***Corynebacterium diphtheriae* spoligotyping based on combined use of two CRISPR loci**. *Biotechnol J* 2007, **2**:901–906.

35. Soini H, Pan X, Amin A, Graviss E, Siddiqui A, Musser J: **Characterization of *Mycobacterium tuberculosis* isolates**. *J Clin Microbiol* 2005, **43**:95–100.

36. David S, Ribeiro DR, Antunes A, Portugal C, Sancho L, de Sousa JG: **Contribution of spoligotyping to the characterization of the population structure of *Mycobacterium tuberculosis* isolates in Portugal.** *Infect Genet Evol* 2007, **7**(5):609–617.

37. Jou R, Chiang C, Huang W: **Distribution of the Beijing Family Genotypes of *Mycobacterium tuberculosis* in Taiwan**. *J Clin Microbiol* 2005, **43**:95–100.

38. Sajduda A, Brzostek A, Poplwaska M, Rastogi N, Sola C, Augustynowicz-Kopec E, Zwolska Z, Dziadek J, Portaels F: **Molecular epidemiology of drug-resistant *Mycobacterium tuberculosis* strains isolated from patients with pulmonary tuberculosis in Poland: a 1-year study**. *Int J Tuberc Lung Dis* 2004, **8**(12):1448–1457.

39. Nikolayevskyy VV, Brown TJ, Bazhora YI, Asmolov AA, Balabanova YM, Drobniewski FA: **Molecular epidemiology and prevalence of mutations conferring rifampicin and isoniazid resistance in *Mycobacterium tuberculosis* strains from the southern Ukraine.** *Clin Microbiol Infect* 2007, **13**(2):129–138.

40. Easterbrook P, Gibson A, Murad S, Lamprecht D, Ives N, Ferguson A, Lowe O, Mason P, Ndudzo A, Taziwa A, Makombe R, Mbengerenwa L, Sola C, Rastogi N, Drobniewski F: **High rates of clustering of strains causing tuberculosis in Harare, Zimbabwe: a molecular epidemiological study**. *J Clin Microbiol* 2004, **42**(10):4536–4544.

41. Mokrousov I, Jiao WW, Sun GZ, Liu JW, Valcheva V, Li M, Narvskaya O, Shen AD: **Evolution of drug resistance in different sublineages of *Mycobacterium tuberculosis* Beijing genotype**. *Antimicrob Agents Chemother* 2006, **50**(8):2820–2823.

42. Godreuil S, Torrea G, Terru D, Chevenet F, Diagbouga S, Supply P, Van de Perre P, Carriere C, Banuls A: **First molecular epidemiology study of *Mycobacterium tuberculosis* in Burkina Faso**. *J Clin Microbiol* 2007, **45**(3):921–927.

43. Toungoussova O, Mariandyshev A, Bjune G, Sandven P, Caugant D: **Molecular epidemiology and drug-resistance of *Mycobacterium tuberculosis* isolates in the Archangel Prison in Russia: predominance of the W-Beijing clone family**. *Clin Infect Dis* 2003, **37**:665–672.

44. Sola C, Ferdinand S, Mammina C, Nastasi A, Rastogi N: **Genetic diversity of *Mycobacterium tuberculosis* in Sicily based on spoligotyping and variable number of tandem DNA repeats and comparison with a spoligotyping database for population-based analysis.** *J Clin Microbiol* 2001, **39**(4):1559–1565.

45. Millet J, Miyagi-Shiohira C, Yamane N, Sola C, Rastogi N: **Assessment of mycobacterial interspersed repetitive unit-QUB markers to further discriminate the Beijing genotype in a population-based study of the genetic diversity of *Mycobacterium tuberculosis* clinical isolates from Okinawa, Ryukyu Islands, Japan.** *J Clin Microbiol* 2007, **45**(11):3606–3615.

46. Sun YJ, Lee AS, Ng ST, Ravindran S, Kremer K, Bellamy R, Wong SY, van Soolingen D, Supply P, Paton NI: **Characterization of ancestral *Mycobacterium tuberculosis* by multiple genetic markers and proposal of genotyping strategy.** *J Clin Microbiol* 2004, **42**(11):5058–5064.

47. Banu S, Gordon S, Palmer S, Islam R, Ahmed S, Alam K, Cole S, Brosch R: **Genotypic analysis of *Mycobacterium tuberculosis* in Bangladesh and prevalence of the Beijing strain**. *J Clin Microbiol* 2005, **42**(2):674–682.

48. Douglas J, Qian L, Montoya J, Musser J, van Embden J, van Soolingen D, Kremer K: **Characterization of the Manila family of *Mycobacterium tuberculosis***. *J Clin Microbiol* 2003, **41**(6):2723–2726.

49. Caws M, Thwaites G, Stepniewska K, Nguyen TN, Nguyen TH, Nguyen TP, Mai NT, Phan MD, Tran HL, Tran T, van Soolingen D, Kremer K, Nguyen VV, Nguyen TC, Farrar J: **Beijing genotype of *Mycobacterium tuberculosis* is significantly associated with human immunodeficiency virus infection and multidrug resistance in cases of tuberculous meningitis.** *J Clin Microbiol* 2006, **44**(11):3934–3939.

50. Storla DG, Rahim Z, Islam MA, Plettner S, Begum V, Mannsaaker T, Myrvang B, Bjune G, Dahle UR: **Heterogeneity of *Mycobacterium tuberculosis* isolates in Sunamganj District, Bangladesh.** *Scand J Infect Dis* 2006, **38**(8):593–596.

Figures

Figure 1 - Models

Relative frequencies of the lengths of spacer deletions (shown here for deletion lengths 1 to 15) estimated by the models. The gray bars represent the empirical values. Uniform (V) refers to the uniform model with parameter $\hat{a} = 10$, while Uniform (F) is the uniform model with $\hat{p} = 1/43$.

Figure 2 - Cluster-graph of Cuban data in Diaz et al. [21] with weighted edges

Nodes are labelled with the ST identifier as indicated in SpolDB4 [22], with the cluster size enclosed in parentheses. Where the spoligotype does not appear in SpolDB4, it is called an orphan strain, hence labelled here 'Or' with a number. Sizes of nodes reflect the number of isolates sharing the spoligotype pattern associated with that node. Edges are labelled with corresponding weights that are computed as explained in the text. For example, ST 1 is inferred to have arisen either from ST 1484 (with weight 0.1666), ST 702 (with weight 0.2752) or Or1 (with weight 0.2809). The lengths of edges do not represent evolutionary distance.

Figure 3 - The spoligoforest generated from the Cuban data in Diaz et al. [21]

Edges with weights less than 0.5 are drawn as dotted lines, those with weights greater than 0.5 but less than 1 are dashed, and those where no decision was required to be made are solid. For example, ST 1 is resolved to have mutated from Or1, and is drawn as a dashed line because it has a weight equal to 0.5582. As in the cluster-graph, the lengths of edges do not represent evolutionary distance.

Figure 4 - IS6110-based phylogenetic tree from the Azerbaijan data in Pfyffer, et. al. [24]

IS*6110*-based phylogenetic tree for data set in [24]. Each tip or leaf of the tree represents an isolate typed with both IS*6110* and spoligotyping. The leaves are labelled with shared types (STs) from SpolDB4. Those spoligotype patterns not appearing in SpolDB4 (orphans) are labelled as Or1 and Or2. Forty-six isolates consisting of 25 different IS*6110* profiles are all represented by ST 1.

Figure 5 - Spoligoforest of the Azerbaijan data in Pfyffer, et. al. [24]

Spoligoforest for data set in [24]. The labels of the nodes are the same as in Figure 4. ST 1, the W-Beijing spoligotype, is represented by the largest node. There are at least two clades revealed in the spoligoforest: the clade with ST 35 and ST 1050, and the clade with ST 53, 42, 254, 118 and 1051.

Figure 6 - Spoligoforest of Caribbean data in Duchene et al. [25]

Spoligoforest for the Caribbean data set (from Cuba, Haiti and French Antilles) in [25]. The clustering of spoligotypes from the LAM group appears in the subtree with ST 42 at the top. This group is highlighted in the spoligoforest, and corresponds to clade 2-1 discussed in [25], which includes the LAM3 and LAM9 families.

Figure 7 - Spoligoforest of Madagascar data in Ferdinand et al. [26] using Zipf model(Equation (6))

Spoligoforest for the Madagascar data set in [26], generated using the selected Zipf model. The edges that are different from the spoligoforest in Figure 8 are F109→F1, F86→F1202, F47→F46 and F53→F237.

Figure 8 - Spoligoforest of Madagascar data in Ferdinand et al. [26] using alternate Logarithmic series model. (Equation (4))

Spoligoforest for the Madagascar data set in [26], generated using the alternative Logarithmic series model. This graph differs from the spoligoforest in Figure 7 by the following edges:
F1521→F1,F156→F1202,F62→F46 and F50→F237.

Tables

Table 1 - Spoligotype data sets used in this analysis

Publication	Isolates[a]	Spoligotypes	Location
Soini et al. [35]	1283	227	USA
David et al. [36]	665	159	Portugal
Jou et al. [37]	420	113	Taiwan
Sajduda et al. [38]	251	91	Poland
Nikolayevskyy et al. [39]	225	73	Ukraine
Easterbrook et al. [40]	224	79	Zimbabwe
Mokrousov et al. [41]	123	14	China
Godreuil et al. [42]	120	39	Burkina Faso
Toungoussova et al. [43]	114	17	Russia
Sola et al. [44]	104	56	Italy
Millet et al. [45]	100	21	Japan
Sun et al. [46]	68	41	Singapore
Pfyffer et al. [24]	65	13	Azerbaijan
Banu et al. [47]	48	18	Bangladesh
Douglas et al. [48]	11	8	Philippines

Table 2 - Data sets and their graph features

Published data set (First Author)	Cluster-graph edges[a]	Spoligoforest edges [b]	Unambiguous edges[c]
Soini	445	126	56
David	403	129	60
Jou	366	84	45
Sajduda	342	63	28
Nikolayevskyy	212	59	29
Sola	137	45	22
Easterbrook	90	53	32
Sun	42	20	15
Godreuil	28	22	18
Mokrousov	27	10	6
Millet	22	13	7
Toungoussova	22	11	8
Banu	13	8	6
Douglas	11	6	3
Pfyffer	8	5	4
Pooled edges			339

[a] Number of single-event deletions in the cluster-graph.

[b] Number of single-event deletions in any spoligoforest derived from the data.

[c] Number of nodes in the cluster-graph having a single parent node.

Table 3 - Frequency of lengths of spacer deletions

Length	1	2	3	4	5	6	9	10	11	12	13	14	15	16	17	26	32
Frequency	199	61	21	25	4	9	3	2	5	1	1	2	2	1	1	1	1

Table 4 - AIC_c values of models with respect to pooled edges

Model name	No. of parameters	Parameter estimate	AIC_c
Geometric[a]	1	$\hat{p} = 0.5935$	1128.82
Negative binomial[a]	2	$\hat{r} = 1;\hat{p} = 0.5935$	1130.82
Conditional Poisson[a]	1	$\hat{\lambda} = 3.3697$	1597.9
Logarithmic series[a]	1	$\hat{\theta} = 0.7967$	1044.49
Zeta[a]	1	$\hat{\rho} = 2.0696$	1033.46
Zipf $k \in [1, 43]$	1	$\hat{\rho} = 1.9962$	1025.23[b]
Uniform (V) $k \in [1, 10]$	1	$\hat{a} = 10$	1770.39
Uniform (F) $k \in [1, 43]$	1	$\hat{p} = 1/43$	2351.77
Empirical-based $k \in [1, 43]$	42	$\hat{p_k} = \frac{x_k}{m}$	1066.94

[a]Model has infinite support for k

[b]Selected model

Table 5 - AIC_c values of models in individual data sets for the best model and the Zipf model

Data set	Best model	Lowest AIC_c	AIC_c of Zipf
Soini	Log series	175.119	179.154
David	Zipf	155.579	155.579
Jou	Zipf	126.111	126.111
Sajduda	Zipf	93.6016	93.6016
Nikolayevskyy	Zipf	107.79	107.79
Sola	Zipf	55.8004	55.8004
Easterbrook	Log series	82.5029	84.0715
Sun	Zipf	21.6428	21.6428
Godreuil	Log series	60.6045	62.8431
Mokrousov	Zipf	20.1675	20.1675
Millet	Zipf	21.7799	21.7799
Toungoussova	Log series	34.4701	35.2256
Banu	Geometric	31.0965	33.2797
Douglas	Log series	19.8291	20.0476
Pfyffer	Geometric	22.8621	23.697

Table 6 - Probability mass functions and maximum likelihood estimators of the models considered

Model name	support	Probability mass function	Maximum likelihood estimator
Geometric	$k \in [1, \infty)$	$P(K = k) = P(k) = p^{k-1}(1-p)$	$\hat{p} = 1 - \frac{1}{\bar{x}}$
Negative binomial	$k \in [1, \infty)$	$P(k) = \frac{(1-p)^r}{1-(1-p)^r} \binom{k+r-1}{r-1} p^k$	
		where $k, r \geq 1$	$\hat{p} = 1 - \frac{\hat{r}}{\bar{x}}$
Conditional Poisson	$k \in [1, \infty)$	$P(k) = e^{-\lambda} \frac{\lambda^k}{k!(1-e^{-\lambda})}$	
		where $k \geq 1, \lambda > 0$	Solution to $\bar{x} = \hat{\lambda}/(1 - e^{-\hat{\lambda}})$
Logarithmic series	$k \in [1, \infty)$	$P(k) = -\frac{\theta^k}{k \log(1-\theta)}$	Solution to $\bar{x} = \frac{-\hat{\theta}}{(1-\hat{\theta}) \log(1-\hat{\theta})}$
Zeta	$k \in [1, \infty)$	$P(k) = \frac{k^{-\rho}}{\sum_{d=1}^{\infty} d^{-\rho}}$	
		where $\rho > 1$	Estimated numerically
Zipf	$k \in [1, 43]$	$P(k) = \frac{k^{-\rho}}{\sum_{d=1}^{43} d^{-\rho}}$	
		where $\rho > 1$	Estimated numerically
Uniform	$k \in [1, 43]$	$P(k) = \frac{1}{43}$	$\prod_{k=1}^{43} (\frac{1}{a})^{x_k}$
Uniform	$k \in [1, a]$	$P(k) = \frac{1}{a}$	
		where $1 \leq a \leq 43$	Estimated numerically
Empirical	$k \in [1, 43]$	$P(k) = \frac{x_k}{m}$	$\prod_{k=1}^{43} P(k)^{x_k}$

Table 7 - The number of edge differences between spoligoforests using alternate models as compared with selected model (Zipf)

Data set	Zeta	Logarithmic	Geometric	Empirical	No. of edges
Diaz et al. [21]	1	0	2	2	37
Duchene et al. [25]	0	3	3	2	42
Ferdinand et al. [26]	0	4*	0	0	50
Caws et al. [49]	0	2	2	1	44
Storla et al. [50]	0	0	2	2	42
Godreuil et al. [42]	0	0	1	1	22

*Spoligoforest shown in Figure 8.

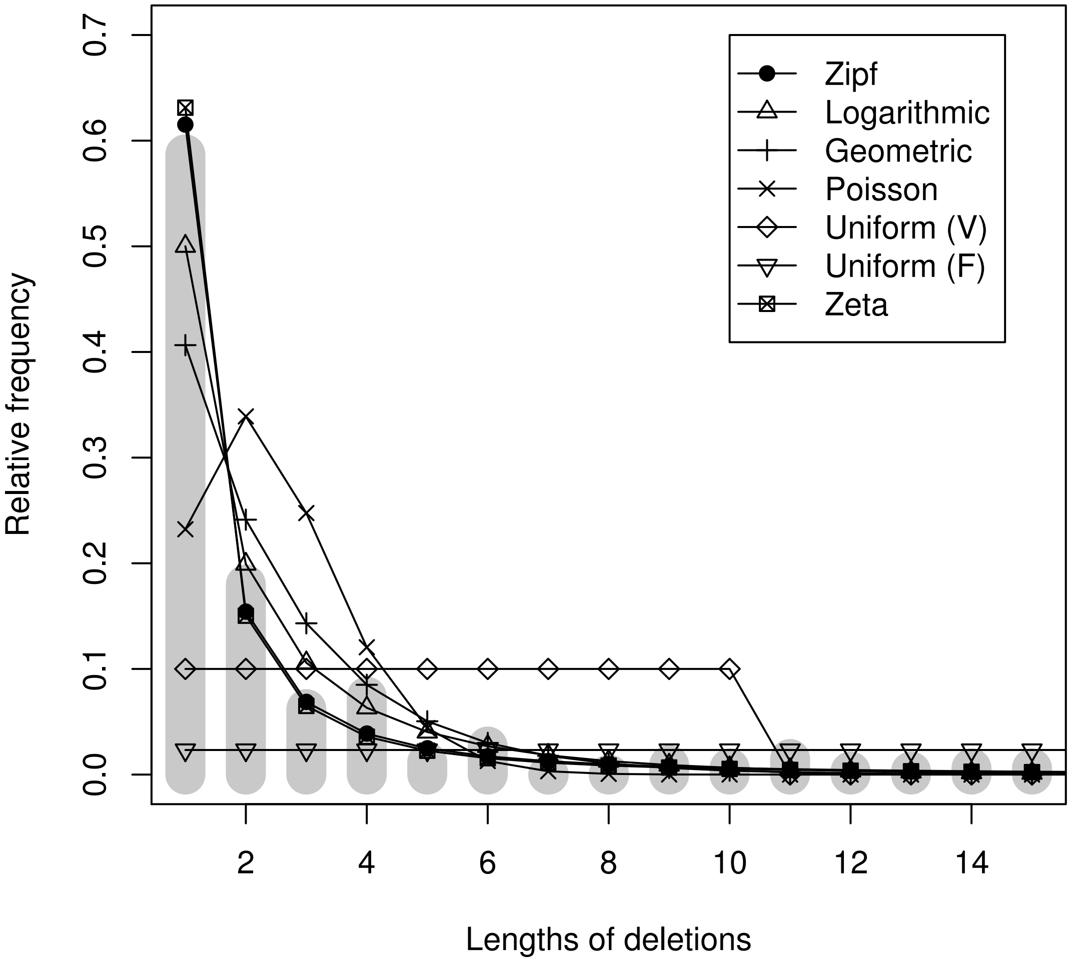

Relative frequency
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
2
4
6
8
10
12
14
Lengths of deletions
● Zipf
△ Logarithmic
+ Geometric
× Poisson
◇ Uniform (V)
▽ Uniform (F)
☒ Zeta

Figure 1

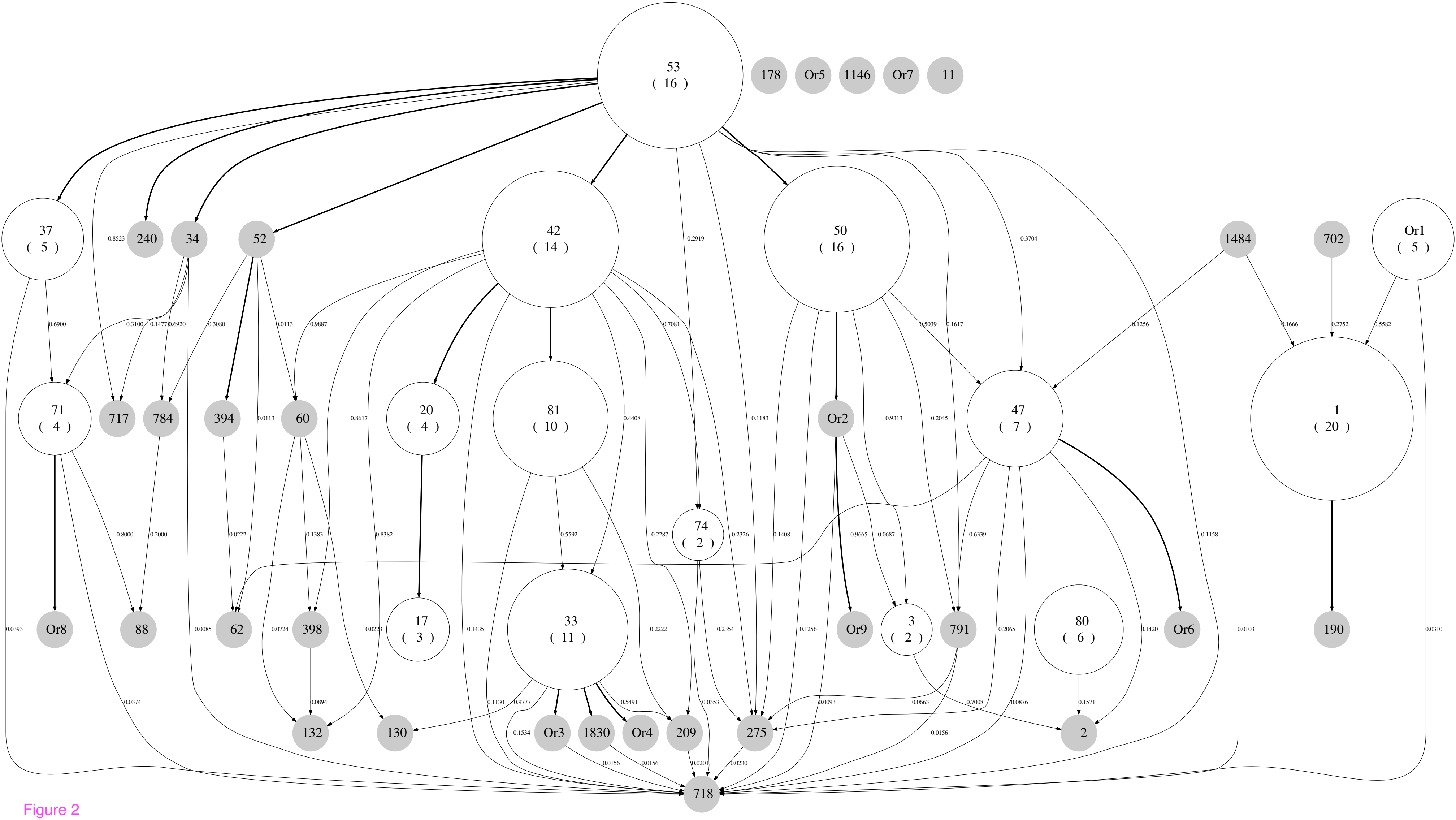

53 (16)
178
Or5
1146
Or7
11
37 (5)
240
34
52
42 (14)
50 (16)
1484
702
Or1 (5)
0.8523
0.2919
0.3704
0.6900
0.3100
0.1477
0.6920
0.3080
0.0113
0.9887
0.7081
0.5039
0.1617
0.1256
0.1666
0.2752
0.5582
71 (4)
717
784
394
0.0113
60
0.8617
20 (4)
81 (10)
0.4408
0.1183
Or2
0.9313
0.2045
47 (7)
1 (20)
74 (2)
0.8000
0.2000
0.0222
0.1383
0.8382
0.5592
0.2287
0.2326
0.1408
0.9665
0.0687
0.6339
0.1158
Or8
88
62
398
17 (3)
33 (11)
Or9
3 (2)
791
80 (6)
Or6
190
0.0393
0.0085
0.0724
0.0223
0.1435
0.2222
0.2354
0.1256
0.2065
0.1420
0.0103
0.0310
0.0374
0.0894
0.1130
0.9777
0.5491
0.0353
0.0093
0.0663
0.7008
0.0876
0.1571
132
130
Or3
1830
Or4
209
275
2
0.1534
0.0156
0.0156
0.0156
0.0201
0.0230
718

Figure 2

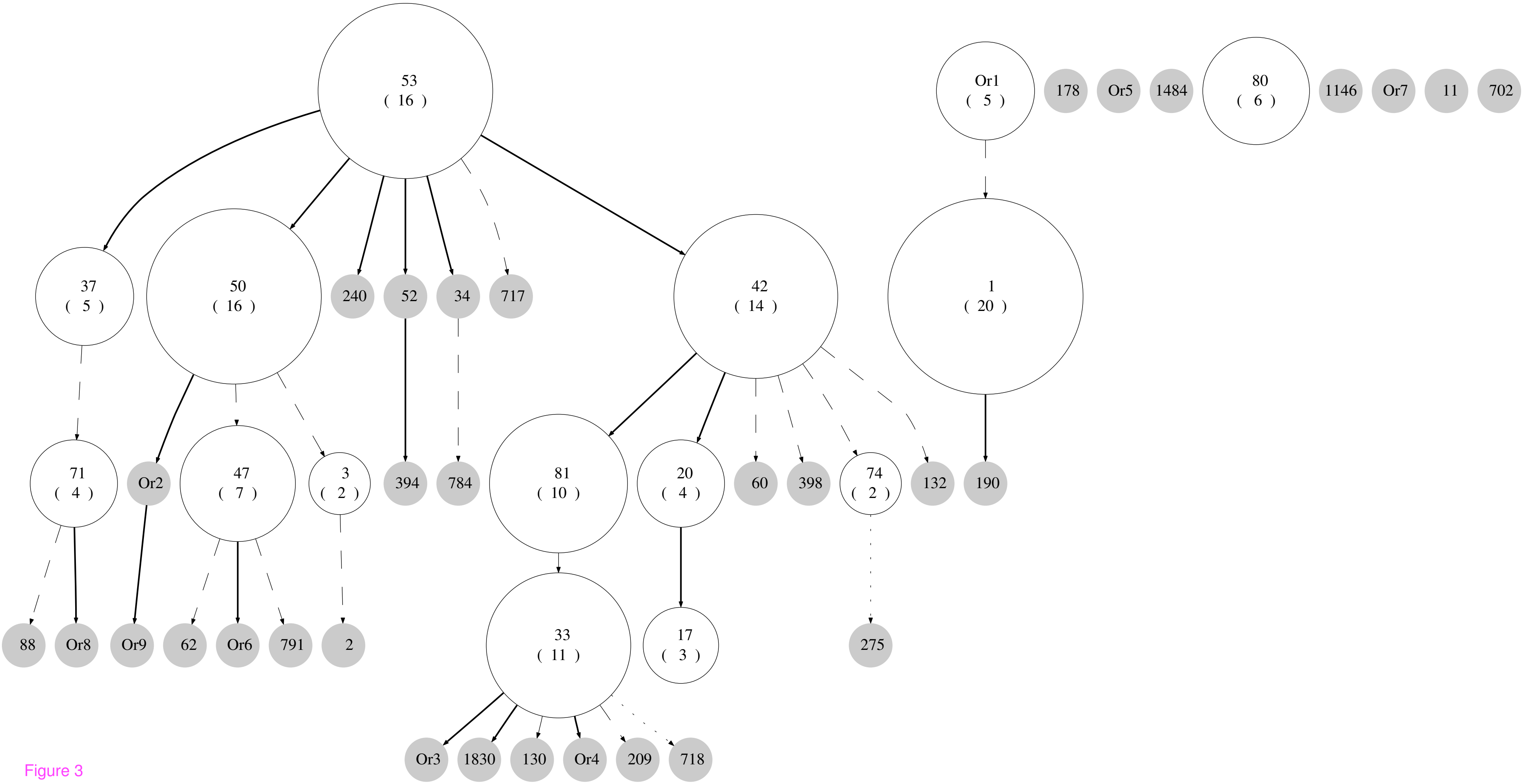

53
(16)
Or1
(5)
178
Or5
1484
80
(6)
1146
Or7
11
702
37
(5)
50
(16)
240
52
34
717
42
(14)
1
(20)
71
(4)
Or2
47
(7)
3
(2)
394
784
81
(10)
20
(4)
60
398
74
(2)
132
190
88
Or8
Or9
62
Or6
791
2
33
(11)
17
(3)
275
Or3
1830
130
Or4
209
718

Figure 3

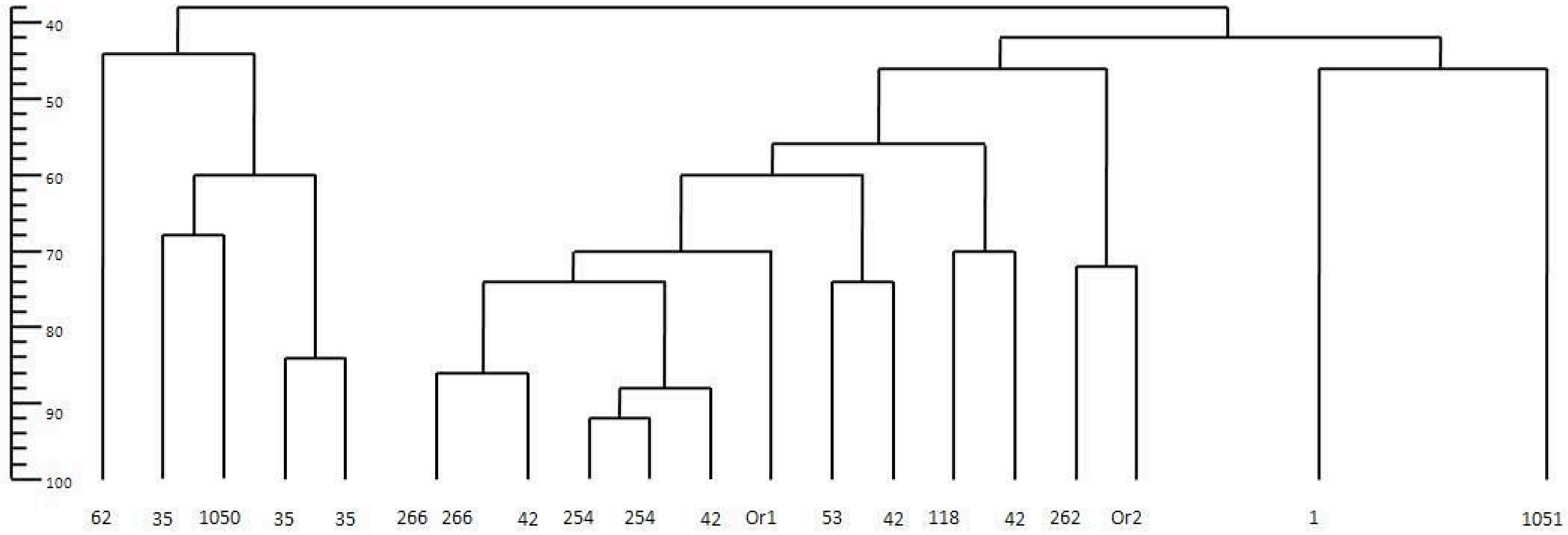

40
50
60
70
80
90
100
62
35
1050
35
35
266
266
42
254
254
42
Or1
53
42
118
42
262
Or2
1
1051

Figure 4

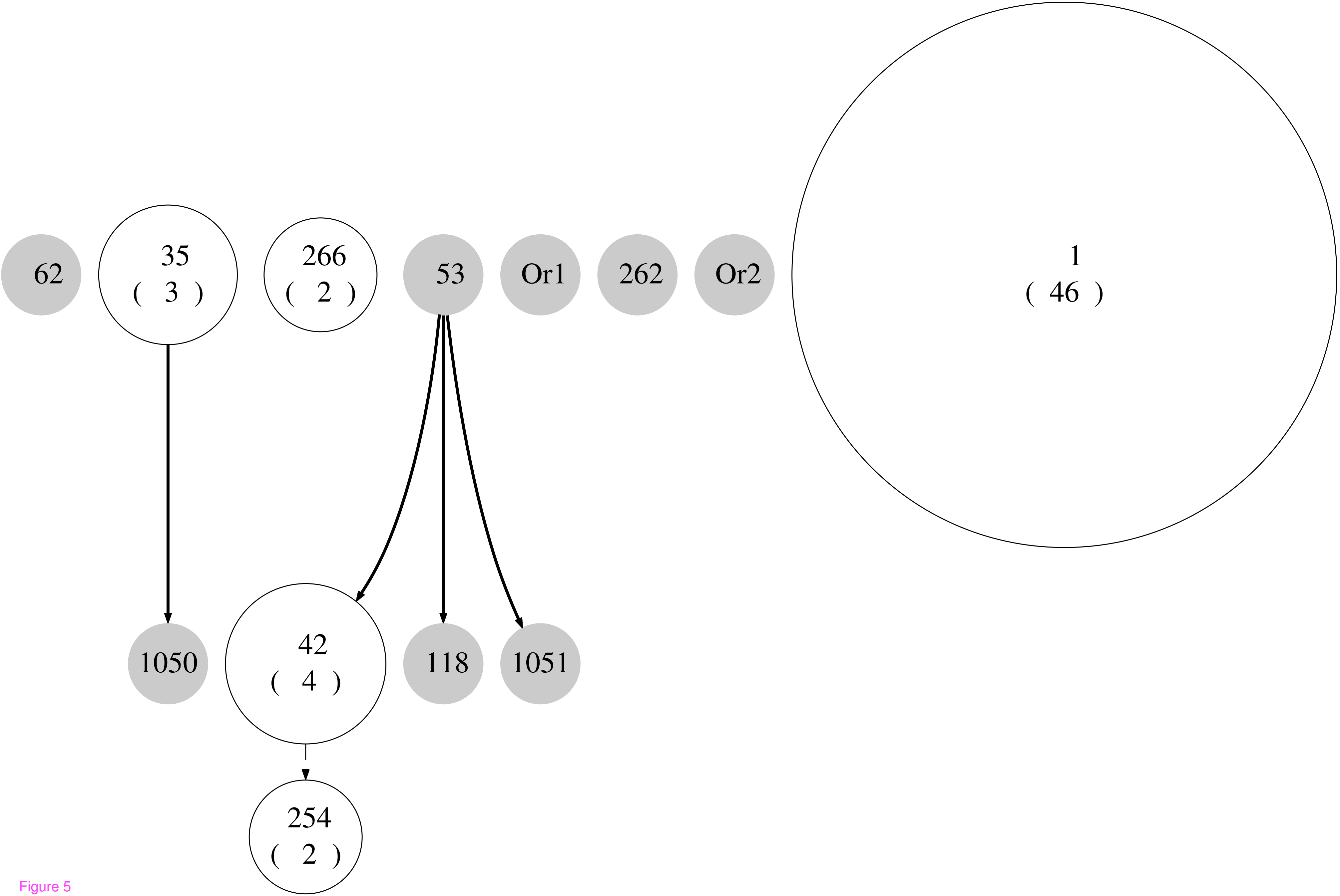
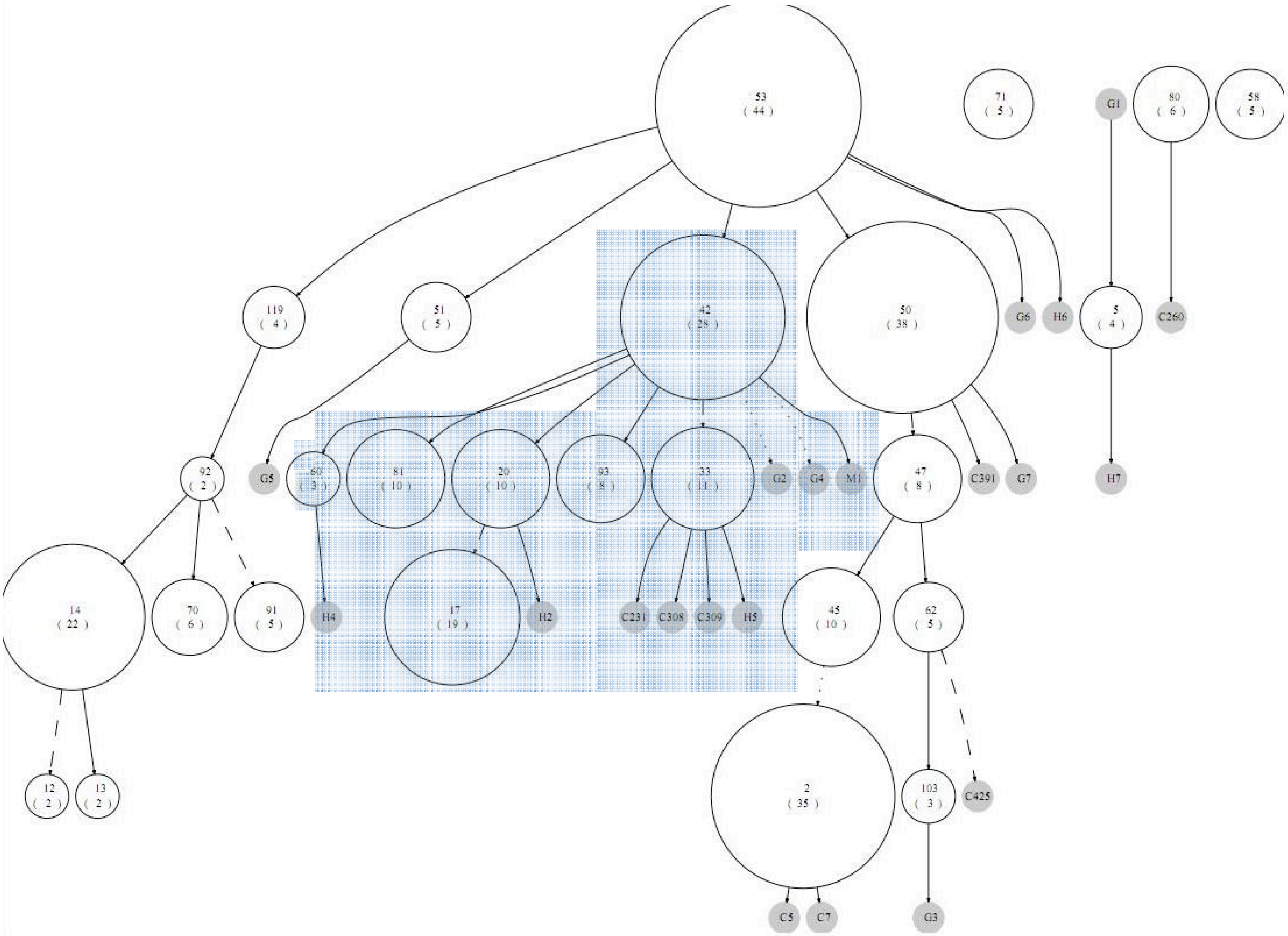


Figure 5

Figure 6



53
(44)
71
(5)
G1
80
(6)
58
(5)
119
(4)
51
(5)
42
(28)
50
(38)
G6
H6
5
(4)
C260
92
(2)
G5
60
(3)
81
(10)
20
(10)
93
(8)
33
(11)
G2
G4
M1
47
(8)
C391
G7
H7
14
(22)
70
(6)
91
(5)
H4
17
(19)
H2
C231
C308
C309
H5
45
(10)
62
(5)
12
(2)
13
(2)
2
(35)
103
(3)
C425
C5
C7
G3

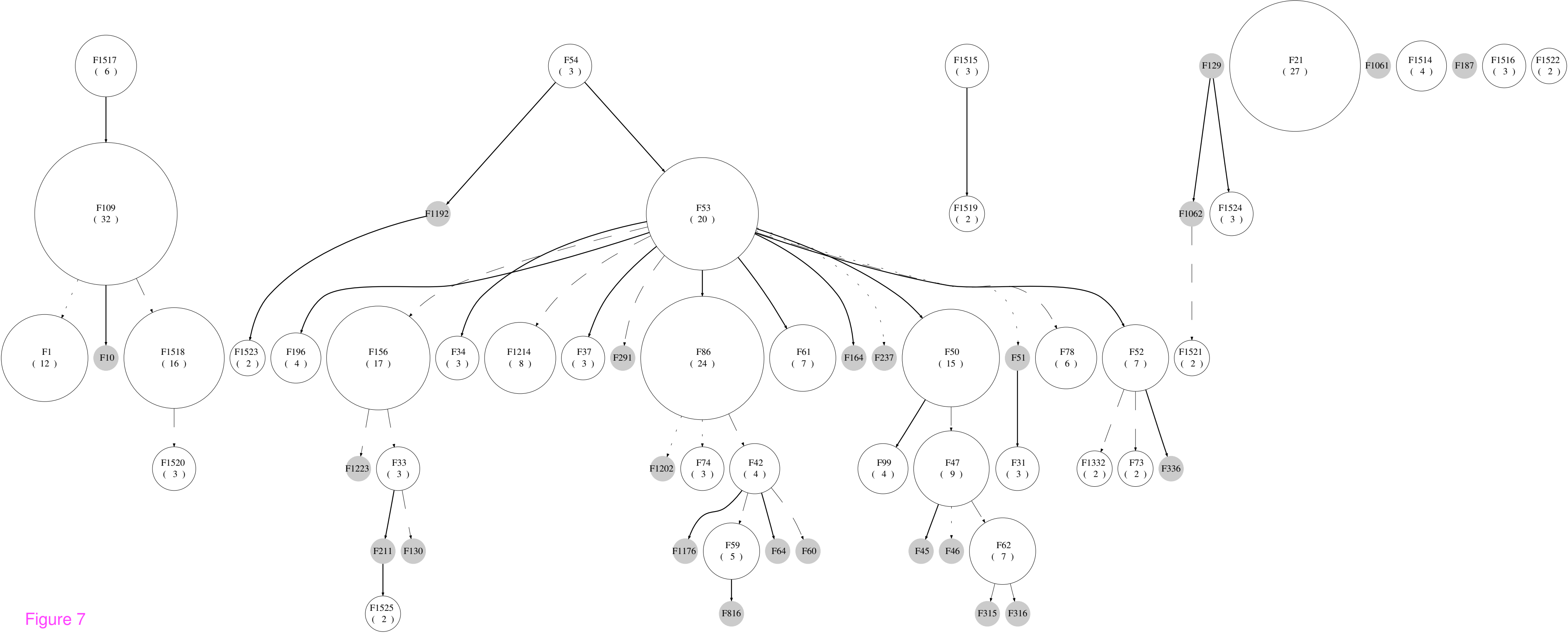

F1517 (6)
F54 (3)
F1515 (3)
F129
F21 (27)
F1061
F1514 (4)
F187
F1516 (3)
F1522 (2)
F109 (32)
F1192
F53 (20)
F1519 (2)
F1062
F1524 (3)
F1 (12)
F10
F1518 (16)
F1523 (2)
F196 (4)
F156 (17)
F34 (3)
F1214 (8)
F37 (3)
F291
F86 (24)
F61 (7)
F164
F237
F50 (15)
F51
F78 (6)
F52 (7)
F1521 (2)
F1520 (3)
F1223
F33 (3)
F1202
F74 (3)
F42 (4)
F99 (4)
F47 (9)
F31 (3)
F1332 (2)
F73 (2)
F336
F211
F130
F1176
F59 (5)
F64
F60
F45
F46
F62 (7)
F1525 (2)
F816
F315
F316

Figure 7

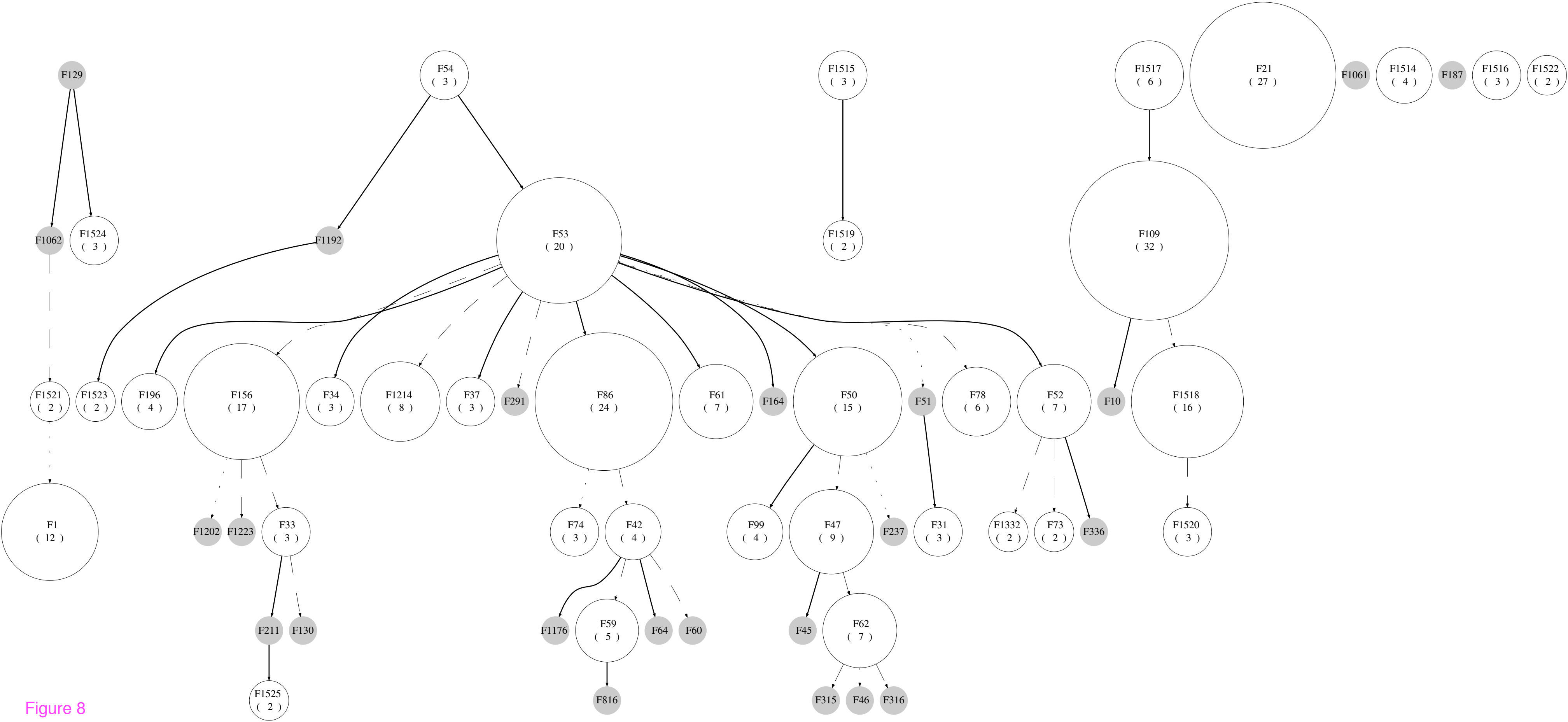

F129
F54 (3)
F1515 (3)
F1517 (6)
F21 (27)
F1061
F1514 (4)
F187
F1516 (3)
F1522 (2)
F1062
F1524 (3)
F1192
F53 (20)
F1519 (2)
F109 (32)
F1521 (2)
F1523 (2)
F196 (4)
F156 (17)
F34 (3)
F1214 (8)
F37 (3)
F291
F86 (24)
F61 (7)
F164
F50 (15)
F51
F78 (6)
F52 (7)
F10
F1518 (16)
F1 (12)
F1202
F1223
F33 (3)
F74 (3)
F42 (4)
F99 (4)
F47 (9)
F237
F31 (3)
F1332 (2)
F73 (2)
F336
F1520 (3)
F211
F130
F1176
F59 (5)
F64
F60
F45
F62 (7)
F1525 (2)
F816
F315
F46
F316

Figure 8