Imperial College London

Evolutionary analysis of a large whole-genome dataset of *Aspergillus fumigatus*

Wenhua Zhou

CID: 00731413

August 27, 2020

Keywords: Tree sequences, Phylogenetics, Large datasets, Genomes, *Aspergillus fumigatus*

Wordcount: 5858

A thesis submitted for the partial fulfillment of the requirements for the degree of Master of Research at Imperial College London

Formatted in the journal style of Systematic Biology.
Submitted for the MRes in Computational Methods for Ecology and Evolution

Declaration

The primary method for inferring tree sequence is issued by Dr Jerome Kelleher and his group in Oxford Big Data Institute. The *Aspergillus fumigatus* datasets are provided by Prof. Matthew Fisher's group in the Department of Infectious Disease and Epidemiology in Imperial College. The data processing pipeline is based on Dr Johanna Rhodes' work from with adjustments for cluster servers and Kelleher et al.'s method. Both my supervisor Prof. Tim Barraclough and co-supervisor Dr James Rosindell provide consistent guidance on the methodology of evolutionary biology and empirical insights in the scientific literature. Dr Jerome Kelleher, Dr Yan Wong, Dr Johanna Rhodes, Dr Reuben Nowell and Wilder Wohns also provide methodology support for this project. I certify that all materials in this dissertation which is not my own work have been properly acknowledged.

Acknowledgement

I really appreciate the support from my supervisor Prof. Tim Barraclough and co-supervisor Dr. James Rosindell along the project. I would also like to thank the help from Dr. Jerome Kelleher's group with the method and Prof. Matthew Fisher's group with the data. I would also thank the support of all members from my supervisors' lab groups. In addition, thanks for the mental support from my girlfriend and her cat Dapao, and they are the motivation to keep myself continuing my academic career. 2020 is a tough year for everyone, the college shut down beacause of COVID-19, all courses and meetings are held online. I am happy make it through together with all the friends in Silwood Park.

Contents

1 Abstract

Fungal pathogens are one of the significant threats to human health and food biosafety. They are receiving more attention in current medical and biological research. Tracing evolutionary pattern using genomic models is an essential method to discover the factors causing diseases before developing controls. Current methods in phylogenetics are limited to a small number of sequence reads, and they are too computational inefficient to process for a whole-genome dataset with millions of DNA sequence reads. Besides, recombination events across the whole-genome sequence will lead to more than one phylogenetic tree when recovering the relationships within species. These limitations inspires the idea of inferring a 'sequence' of phylogenetic trees from the whole-genome DNA sequences, where each tree encodes information for a part of the DNA sequence. This method was initially introduced within a population of humans, which is a single homogeneous species(Kelleher et al., 2019). Here I apply this approach on whole-genome sequences of 219 different isolates of *Aspergillus fumigatus*, a fungus that causes disease in immuno-compromised humans and that previous work has shown to encompass two cryptic species. It turns out Kelleher et al.'s method encodes sufficient information for evolutionary analysis for multiple fungal species, and provide reasonable accuracy.

2 Introduction

There are approximately 2.2 to 3.8 million fungal species (Hawksworth and Lücking, 2017), and over 300 are known to be pathogenic to humans. *Aspergillus*, *Candida* and *Cryptococcus* are the most common pathogenic species, causing allergic disease and infection to humans with weakened immune systems. In pharmaceutical development, pathogenic fungi are usually drug-resistant (Anderson, 2005). Tracing the evolutionary patterns of pathogenic fungi can find the heritable traits among different fungal species across generations. These traits are related to test drug resistance. In epidemiology and ecology, evolutionary analysis testing drug resistance is essential to predicting future mutations and developing next-generation drugs.

The genus *Aspergillus* is a group of conidial fungi, comprises hundreds of different species with the shape of an *aspergillum*(holy water sprinkler). *Aspergillus* species share a standard asexual spore-forming structure, including one-third of them with a sexual stage (Geiser, 2009). Among those species, more than 60 species are considered as pathogens. For humans, *Aspergillus* species can cause a range of diseases such as infection to the external ear and ulcers. The research focus of fungal pathogens is mainly on *Aspergillus fumigatus* species since *Aspergillus fumigatus* causes the most significant number of infections among all *Aspergillus* species (Latgé, 1999).

In evolutionary biology, using evolutionary models to analyse DNA sequences is an increasing trend. From the last few decades, different modern technologies are introduced to measure biological and genomic data. These modern technologies lead to massive growth in the availability of whole-genome data, which brings the potential to infer more information about the traits among multiple species. For example, several improvements have been made for Illumina gene sequencing platforms to handle large genome sequences (Fadrosh et al., 2014; Quail et al., 2008; Van Dijk et al., 2014). However, current methods performing accurate evolutionary analysis are difficult to compute whole-genome sequences with multiple individuals or species. Phylogenetic trees are a common way to illustrate the evolutionary relationships among different fungal species, but they struggle to handle long DNA sequences with a larger number of different samples. The whole-genome sequence contains millions of DNA reads, taking too much time to infer phylogenetic trees for current computational resources. Also, during a sexual reproduction process, genes from different individual samples come in different combinations. The recombination events cause multiple ancestor histories across the genome. Various ancestor histories for samples are impossible to be expressed by a single tree but requires a network or multiple trees to encode instead.

Although whole-genome sequences include a substantial quantity of information, most of the information is not relevant to the evolutionary analysis. Only the variants among different

examples determine the analytical results. Therefore, taking variants information as the input instead of samples of whole-genome sequence can minimise the memory requirements for large scale datasets.

Recently, a 'tree-sequence method' (Kelleher et al., 2016) was introduced in genealogical analysis for large sample size. The current state-of-art phylogenetic approaches are mostly based on ancestral recombination graphs(ARG) (Griffiths and Marjoram, 1996), which model a phylogenetic network by encoding the recombination and common ancestor events in all samples (Arenas, 2013; Minichiello and Durbin, 2006). However, the problem of finding a phylogenetic network with the minimum recombination events is NP-hard (Bordewich and Semple, 2005; Wang et al., 2001). Therefore, the computational time of ARG based methods increases rapidly with the number of samples and recombination events, limiting ARG's use by the prohibitive high computational cost. Phylogenetic networks aim to handle more complex evolutionary scenarios rather than the whole-genome sequences and large sample size.

Instead of encoding all the genetic information, Kelleher et al.'s tree-sequence method presents the evolutionary patterns of DNA sequences by inferring a sequence of correlated trees. All of the individuals in Kelleher et al.'s tree sequences are within a sexual reproducing species. Each tree represents the evolutionary relationships between a part of DNA sequences of all individuals. Along a chromosome, ancestry changes at the points where recombination events occur, resulting in a new correlated tree. Only a small part of ancestries with the affected individuals change for the new tree, and only the changing part is encoded in this method. Using Kelleher et al.'s different encoding method increases the computational efficiency significantly and also encodes sufficient genetic information for further evolutionary analysis. This leads to great potential in dealing with large whole-genome sequences and multiple examples.

I adapt and apply Kelleher et al.'s tree-sequence method into *Aspergillus* whole-genome dataset from Professor M.C. Fisher and Dr Johanna Rhodes in Imperial's Department of Infectious Disease and Epidemiology. For Imperial's *Aspergillus* whole-genome datasets, there are a total of 219 different individual samples each with millions of DNA pair reads. The dataset is collected in the UK and Republic of Ireland from 2005 to 2017 with metadata available (Rhodes, 2019; Sewell et al., 2019a,b). Tree sequence method is ideal here, given that it provides a considerable advantage in computational efficiency for large datasets (Kelleher et al., 2019). How is the tree-sequence method going to perform within a more complex, structured species of *Aspergillus fumigatus*? Can tree-sequence method provide robust evolutionary relationship for multiple species? How fast is tree-sequence method comparing to other tree-plotting algorithms?

3 Methodology

3.1 Data preparation

The original *Aspergillus* datasets are from Professor M.C. Fisher's lab group in the Department of Infectious Disease and Epidemiology of Imperial College London. They were collected from different labs in England, Scotland, and the Republic of Ireland from 2005 to 2017. There are 219 samples of *Aspergillus* whole-genome sequences in total, and each sample contains a pair of DNA sequence sin compressed FASTQ format. All compressed FASTQ files from 219 examples take about a total of 250GB space to store.

Kelleher et al.'s tree-sequence method take observed genetic variation data as the input to infer tree sequences. Therefore, a data preparation step is essential to transfer whole-genome sequences in FASTQ file format into one variant call format(VCF) file, including variant information for all 219 samples of *Aspergillus fumigatus*. Dr. Johanna Rhodes in Professor M.C. Fisher's lab group developed a similar whole-genome sequencing pipeline to get the variant information on each *Aspergillus* sample. I adapted her pipeline to combine all the variant information into one VCF file and apply the VCF file to the tree-sequence method. Due to the large size of compressed FASTQ files, running the pipeline on a local computer is prohibitively expensive. I upload *Aspergillus* datasets to Imperial's Research Data Store and use Imperial's cluster server to run the pipeline for all samples in parallel.

Figure 1 is the general flowchart of my pipeline. The pipeline is written in a shell script for the cluster server, converting input FASTQ files for 219 samples into one VCF file. The major tool I used for this genome sequencing pipeline is the Genome Analysis Toolkit(GATK) (McKenna et al., 2010), and the reference genome for variant calling is *Aspergillus fumigatus* Af293 from National Center for Biotechnology Information(NCBI), including Mitochondrial DNA and eight different reference sequences (Nierman et al., 2005). The GATK package can call variants from BAM files. A BAM file is the binary version of a SAM file, which is a text file of sequence alignment data. Before calling the variant, I map FASTQ files of each sample to the reference genome. This process is accomplished by using the Burrows-Wheeler Alignment(BWA) (Li and Durbin, 2009) tool package, and I get a SAM file as the output. SAMtools and Picard are the tools to manipulate alignment data in SAM/BAM format. I can import the SAM files to BAM files and sort the alignment information in order by SAMtools(Li et al., 2009) and add read groups names as sample names for the samples by Picard. The GATK package then calls variants including the Single Nucleotide Polymorphisms(SNPs) and insertions or deletions(INDELs) of bases in the genome from the processed BAM files. In most phylogenetic methods, the tree-sequence method uses SNPs to trace evolutionary patterns (Anderson, 2005).

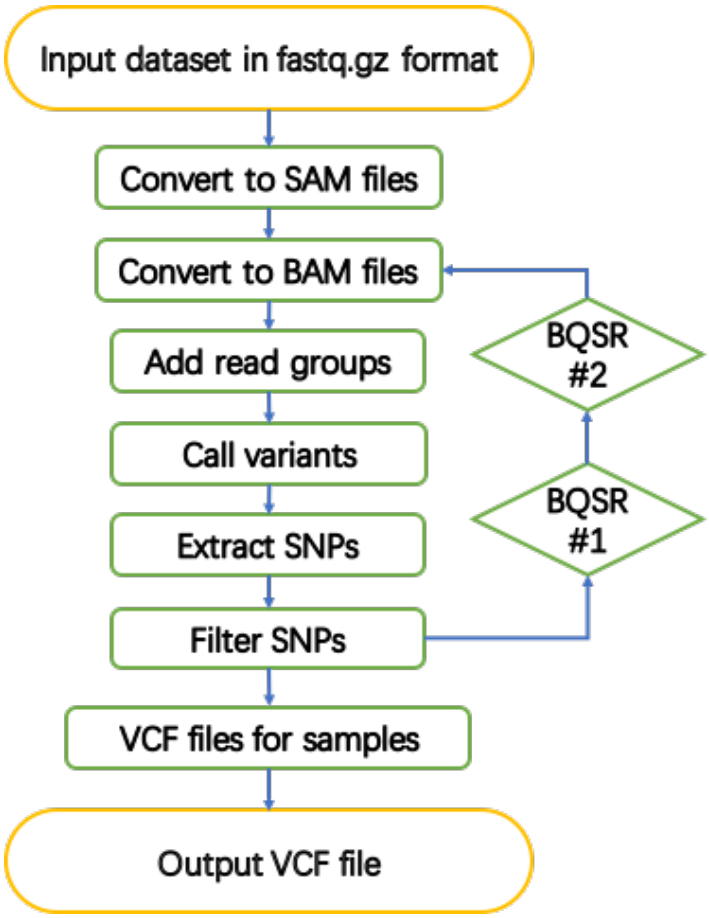


Figure 1: Flowchart of data preparation pipeline of *Aspergillus* whole-genome sequences. After filtering SNPs for the first time, the pipeline apply BQSR two times and backtrack the BAM file to improve base quality. When filter SNPs for the second time, the pipeline continues directly to the VCF file for each *Aspergillus* sample.

Therefore, I extract SNPs and INDELs and only record SNPs for further downstream analysis methods. Low-quality SNPs will affect the accuracy in future analyses, so quality control is always necessary for the pipeline. In my pipeline, I separate the quality control into two parts: filtering SNPs and base quality score recalibration(BQSR). After extracting SNPs from the variants, I filter the SNPs with the filtering value utilised by Dr Johanna Rhodes, followed by two BQSR consecutively. A BQSR process is a data pre-processing step that detects systematic errors made by the sequencing machine by estimating the accuracy of each base call. A BQSR process applies machine learning to model the error and adjust the quality scores, it recalls the original BAM file to filtered SNPs file and outputs an adjusted BAM file with base quality score control which improves the accuracy in variant calling. After applying the backtracking BQSR, I call the variants from the new BAM file again, extract and

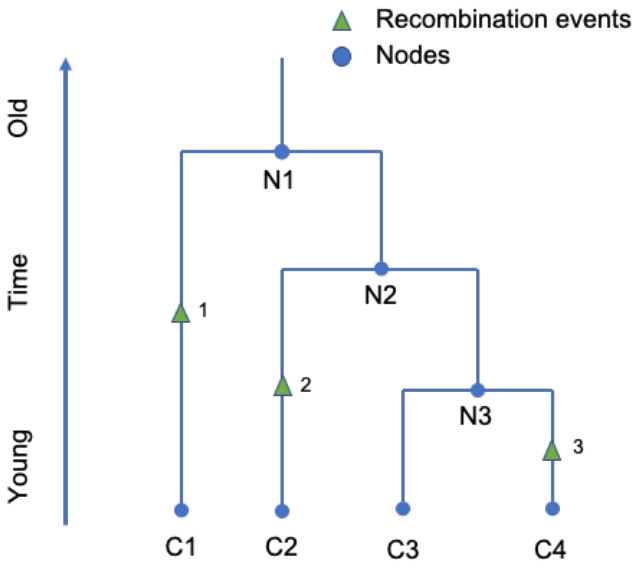
filter SNPs a second time to get the output VCF file for each sample. The VCF file contains the SNPs for each sample whole-genome *Aspergillus* sequence. Combining the 219 VCF files results in one VCF file. Note that VCF file only records the SNPs for each sample. This fact means that all other positions of the chromosome have the same alleles as the reference sequence. Therefore in the combination process, SNPs positions not recorded in some of the individual samples are counted as reference alleles.

3.2 Tree-sequence Method

The tree-sequence inference method by Kelleher et al. takes the variation data as an input, giving a sequence of marginal trees. The nodes of trees are the different individuals, and the branches represent the relationships between these individuals. Instead of encoding all the ancestral histories and recombination events as ARG based methods do, the tree-sequence method encodes information of a sequence of trees.

The sequence of trees is represented by two tables. The first table includes information for the nodes, recording the node individuals and their closest ancestor nodes, respectively. Along the whole-genome sequence, new trees are introduced by recombination events. Therefore, information of nodes also includes the sequence interval information. Only the changing relationships between nodes are added to the table, and the same relationships between trees are recorded in the previous existing row with enlarged interval. The second table records recombination events with the related descendant nodes. Along the chromosome, recombination events are the key in changing of ancestral histories. The tree-sequence method encodes all recombination events in the second table together with the related descendant nodes, which may be affected by the recombination events. The position information of a recombination event is also added to put all combination events into the right order.

Figure 2 shows a simple example of how a tree is encoded in two tables. The nodes C1, C2, C3, C4 in the example are the individual sample nodes, and N1, N2, N3 are their ancestral nodes. The recombination events are recorded with its sequence position in the second table with the position in the order. Recombination events may change the relationship between nodes, affecting the branch they are on. For instance, the position of the first recombination event is 4, and the interval of individual sample node C1 is (0,4). Suppose there is a different tree after the recombination event, such as C1 connects with a new node N4 after the recombination. This new relationship is recorded in a new row of the first table. It is expressed in a new tree with node C1, related node N4 and interval (4,76).



Nodes

Node	Related Node	Interval
C1	N1	(0,4)
C2	N2	(0,23)
C3	N3	(0,162)
C4	N3	(0,76)
N1	-	(0,242)
N2	N1	(0,189)
N3	N2	(0,93)

Recombination events

Site	Related Node	Position
1	C1	4
2	C2	23
3	C4	76

Figure 2: A genealogical encoding example of trees, with a simple version of tree and its related information recorded in two tables. The nodes C1,C2,C3,C4 are the individual samples, and N1,N2,N3 are the ancestral nodes.

The inference algorithm starts with an input variation dataset, which I already have from the data preparation step. The ancestral haplotypes of a DNA sequence are essential to inferring ancestral histories, but we do not always know them. The first step of the algorithm uses a simple trick to infer ancestors from variation data. The trick is to replace the relative age of the corresponding ancestor by the frequency between reference and derived allele. The trick is based on the fact that the derived allele started as a single mutation, higher frequency means it takes longer computational time since the mutation first arose. Hence use frequency instead of relative age is reasonable. Then, each ancestor inferred has a related site, I can scan from the related site to both sides and reconstruct an ancestral haplotype for the ancestor.

The reconstructing process is shown in figure 3, with a simple example. I use only 5 individual samples with a shorter sequence. In the example, the focal site has position 5, only C1, C3, C5 in position 5 have derived alleles. Starting from the left side for these three haplotypes, position 4 has a lower frequency of derived alleles, so I can just keep moving to position 3. Position 3 has a higher frequency than the focus site. Only samples with a derived allele at this position can keep on scanning left. Therefore, sample C5 with a reference allele on position 3 ends its scan here. Similarly, the rest sample haplotypes C1, C3 can both scan through position 2, but C3 will stop at position 1. At this position, only C1 is left. The proportion of sample haplotypes is 1/3, which is not more than 1/2. Therefore, the scanning

process on the left side ends in position 1. On the other side, C3 and C1 stop at position 7 and 8 respectively. From the scanning process, the length of this ancestral haplotype is 8, which is from position 1 to 8. When the frequency of derived alleles is lower or equal to the focal site, the inference result of related position is the ancestral allele. Otherwise, when the frequency is higher, the inference result of related position is the more frequent allele among the scanned sample haplotypes. If both frequencies are 50%, the result takes the ancestral allele. Although this trick is an unusual approach, it is shown to be robust with reasonable accuracy (Kelleher et al., 2019).

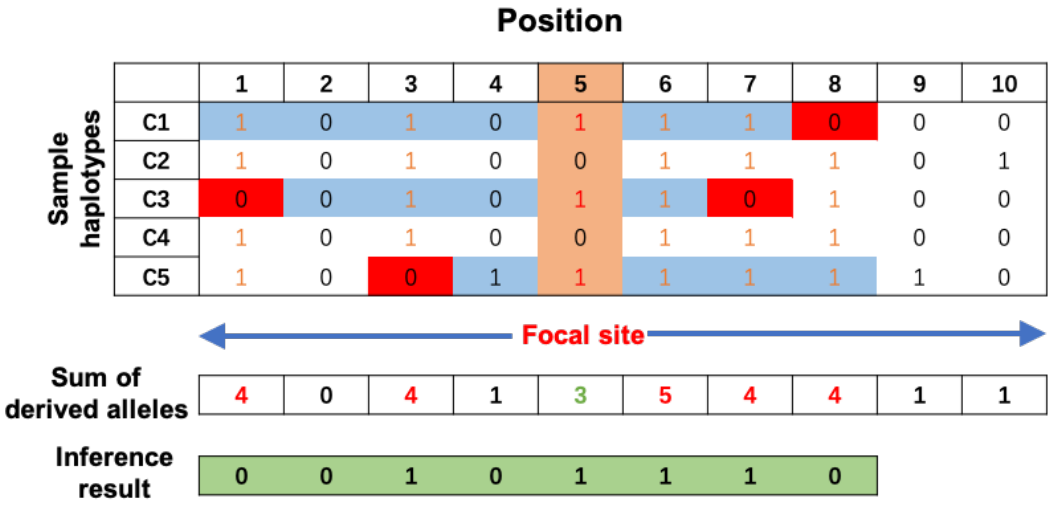


Figure 3: A simple example process of reconstructing an ancestral haplotype from a focal site with a smaller number of sample haplotypes and shorter position sequence. In this sample position 5 is the focal site, C1,C3,C5 are the sample haplotypes with the focus site. Scanning from focal site to both sides by the precise rules, an ancestral haplotype is inferred with accurate acestors.

For whole-genome datasets, the inferred ancestral haplotypes usually do not span the whole genome due to a large number of recombination events. Inferred ancestors with high relative age also tend to get shorter sequence length. To get the tree sequence out of those ancestral haplotypes, Kelleher et al. also introduce an ancestral copying process. Before the copying process, a reference panel aligning all inferred ancestral haplotypes from young to old age along the whole-genome sequence is required. The panel is based on a statistical model of linkage disequilibrium among SNPs with multiple samples (Li and Stephens, 2003). By using the reference panel, it is clear to observe the relationships from young to old ancestral haplotypes along the whole-genome sequence. The relationships show how each haplotype is copied from its ancestors at a certain interval of sequence. By getting the copying history

of haplotypes with all older haplotypes, we can consider each copying path as an edge for a related tree, and all the edges generate a sequence of succinct trees directly. It is guaranteed to generate a tree sequence for every position of the sequence by finding all copying paths for all the ancestors.

3.3 Application of *Aspergillus fumigatus* datasets

The tree-sequence inferring method has a package called `tsinfer` (Kelleher et al., 2019), and it has a related toolkit called `tskit` (Kelleher et al., 2016). From the data preparation step, I transfer the *Aspergillus* samples into a VCF file containing all the variant information. However, the variant information input is required to be in a *.samples* file including three types of information of all SNPs. The first information is the position number of SNPs in a chromosome. The reference *Aspergillus fumigatus* genome includes 8 different reference chromosomes and mitochondrial DNA. To ensure succinct position numbers, I apply `tsinfer` to reference sequence separately and focus on mitochondrial DNA and the most extended sequence for further analysis. After separating SNPs in reference sequences, position numbers are easy to obtain from the VCF file. The second information is the name of ancestral and derived allele, which is a list of two letters. However, the *Aspergillus fumigatus* genome does not contain information of ancestral alleles. In this case, I consider the ancestral alleles to be more frequent alleles for all samples. There is another way to infer ancestral alleles, which is to simply set the ancestral alleles to be the reference alleles. Both of the methods are acceptable since in most sites the reference alleles are exactly the more frequent alleles. The last information is the genotypes information, which is a list of numbers for each sample. '0' in the list stands for the ancestral allele, and '1' in the list stands for the derived allele. When combining samples of *Aspergillus fumigatus* into one VCF, samples having the reference alleles for specific sites record '.' instead of '0'. Change all '.'s to '0's and then genotypes information is obtainable. The inference algorithm for tree-sequence method is only available for biallelic SNPs when inferring the ancestral haplotypes. Therefore, it is also important to filter out all SNPs with more than one type of derived alleles at this stage.

3.4 Comparison with conventional phylogenetic methods

The tree sequence inference method is an approximation algorithm that aims to achieve reasonable accuracy with a low computational cost. Therefore, I compare the tree sequence method to a standard tree inference method for its accuracy. `Fasttree` (Price et al., 2010) is a tool package to infer phylogenies for large alignments with up to hundreds of thousands of sequences. It uses an approximation of maximum-likelihood methods to handle a large sequence of alignments. The tool package itself takes multiple sequence alignments in fasta

format as input and get a phylogenetic tree in the Newick format. Since I have my existing pipeline for 219 *Aspergillus fumigatus* isolates, I backtrack the sequence of alignments from the output VCF file of my pipeline containing all the SNPs information only by SAMtools. From the SNPs alignment simply apply `fasttree`, and it results in a phylogenetic tree for 219 individual samples.

However, the output of Kelleher et al.'s tree sequence method is a tree sequence, which is not comparable to one phylogenetic tree from `fasttree`. There is not a common method to analyse a sequence of trees, since Kelleher et al.'s tree sequence method is a novel approach to handle large whole-genome datasets. Therefore, rather than trying to analyse a sequence of trees, I combine the tree sequence into one tree and compare it with the tree using `fasttree`. This is accomplished by drawing the consensus tree from the tree sequence. A consensus tree provides an estimated relationship between sample nodes from a sequence of trees. In a tree sequence, if a particular combination occurs over a certain rate among all trees, this combination is recorded in the consensus tree. The percentage is usually from 50% to 100%, where the consensus tree is defined as majority rule and strict rule respectively. The number of trees in the tree sequence is relatively large, so I only apply the consensus tree to tree sequence by majority rule.

3.5 Genealogical nearest neighbours

Consensus tree can not possess all the relationships simultaneously. Therefore only the closest relationship in the tree sequence can be recorded by drawing a consensus tree. Alternatively, we can characterize the ancestral relationship by evaluating the genealogical nearest neighbours(GNN). GNN is a statistic based on the topological properties to identify nearest neighbours between multiple samples from a sequence of reference sets (Kelleher et al., 2019). The average GNN along a tree sequence $\mathbb{T}$ is defined as

$$G_{u,k} = \frac{1}{L} \sum_{t \in \mathbb{T}} L^t G^t_{u,k} \tag{1}$$

where t is the tree in tree sequence $\mathbb{T}$ with focal node u and a set of its reference nodes k. Each tree covers L^t reads of genome sequence and $L = \sum_{t \in \mathbb{T}} L^t$. $G^t_{u,k}$ stands for the GNN statistic for given starting node u and tree t. This can be defined as

$$G^t_{u,k} = \frac{C^t_{v,k} - [u \in R_k]}{\sum_{k=1}^{K} C^t_{v,k} - [u \in \bigcup_{j=1}^{K} R_j]} \tag{2}$$

where R is a list of K sets of reference nodes, $C^t_{v,k}$ is the number of nodes descending from target node v in tree t from list R_k. $[u \in R_k]$ stands for the Iversonian condition, where

$[u \in R_k] = 0$ if false and $[u \in R_k] = 1$ if true. Note that the genealogical nearest neighbour statistic is a number from 0 to 1 according to the definition. GNN close to 1 represents the relationship are extremely close, and GNN close to 0 stands for a relatively distant relationship. If the metadata of the dataset are provided, GNN can be applied to represent the relationships between groups by changing the reference nodes k to be a list of samples in each group. *Aspergillus fumigatus* samples have different sets of metadata, and I applied three different types of metadata: the Azole mutation resistance (AMR) status, the year each isolate measured, and the clustering groups by discriminant analysis of principal components (DAPC).

4 Results

4.1 Inferred tree sequence

The ARG based methods encode the information of common ancestor and recombination events by a conventional data storage matrix. This matrix has size $m * n$, which m is the number of variant sites and n is the number of individual samples. For the application of *Aspergillus fumigatus* datasets, each row represents each SNP along the whole-genome chromosome, and each column is one of the 219 individual samples of *Aspergillus fumigatus* isolates. The matrix is consists of 0s and 1s, where 0 stands for the ancestral allele and 1 stands for the derived allele.

The Kelleher et al.'s tree sequence method results in two tables with the form similar to Figure 2. The second table of recombination events has m rows, where each row represents the same site of each recombination event as the conventional matrix. The first table records both individual sample nodes and ancestral nodes. In addition, the table also encodes new relationships between nodes along the tree sequence. Therefore, the first table has k rows, where k represents the number of total succinct edges in the tree sequence and $k > n$.

In comparison, the data for ARG based conventional matrices and tree sequence inference tables take $O(m \times n)$ and $O(k + m)$ space to store respectively. Symbol O stands for the order of computational complexity to compute the matrices and tables. The longest reference chromosome for *Aspergillus fumigatus* has 63542 (m) variable sites with 219 (n) individual samples. The tree sequence methods have 105701 (k) succinct edges for the nodes table. The size of the conventional matrix for this chromosome is about 100MB, where tree sequence tables take only 5MB storage space. Due to the different orders of computational complexity, tables from the tree sequence method have the potential to save more storage space than the conventional matrix for even more samples and sites.

Information from two encoding tables shows all the relationships between nodes with its

related intervals. Hence I can obtain a tree sequence by the two tables easily. Figure 4 shows an example of a phylogenetic tree in the tree sequence of the mitochondrial DNA with only 31892 bases for the sequence in a circular cladogram. By the information of tree legend, this tree is the first tree out of the total 6 trees, which covers a sequence interval from the beginning to the 7321st base pair in the alignment of the mitochondrial DNA. The circular cladogram is an ideal layout to observe a large number of samples clearly (Pavlopoulos et al., 2010).

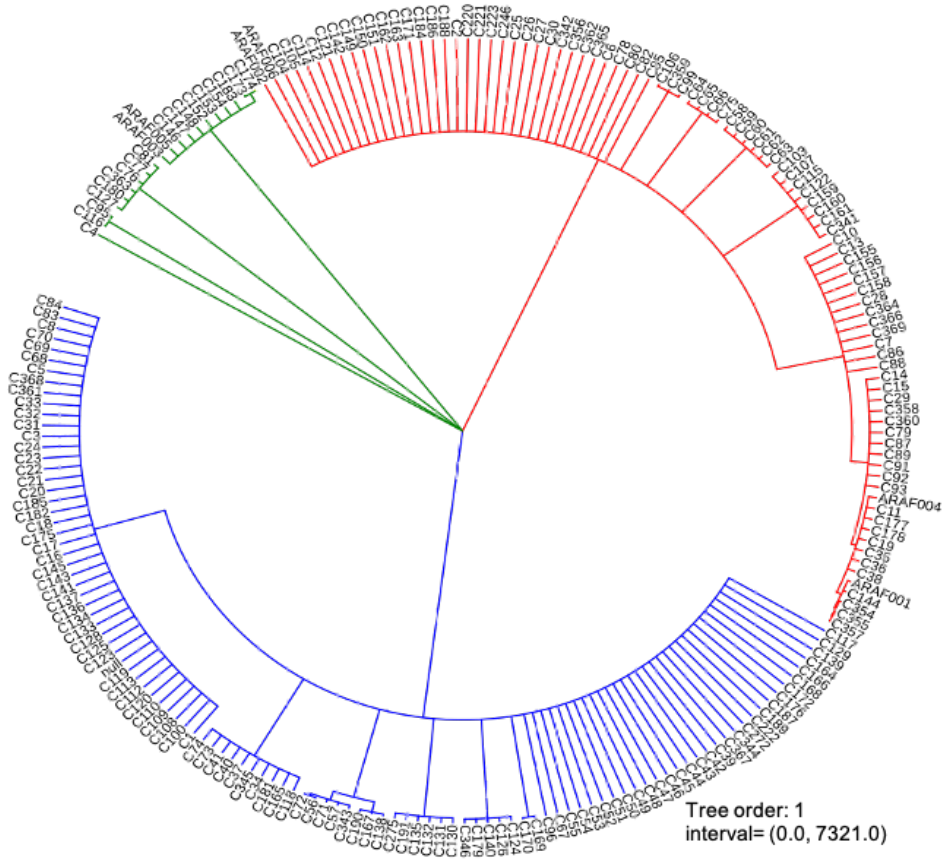


Figure 4: An example of a circular phylogenetic tree in the tree sequence, showing the relationships between 219 different isolates of *Aspergillus fumigatus* for part of their whole-genome sequences. Two main clades of the circular cladogram are shown in red and blue color, and the remaining nodes are shown in green.

4.2 Compute consensus tree

The result of the consensus tree of mitochondrial DNA by majority rule (figure 5) illustrates relationships for more than half of the samples. Compared to the general phylogenetic

tree by `fasttree` (figure 6), the clades including related nodes in the consensus tree are directly mapped to the phylogenetic tree, and the mapped clades have the same nodes. It is confident to say the related individual samples in the consensus tree are also shown to be closely related to each other in the phylogenetic tree. This means the Kelleher et al.'s tree sequence method shows the correct relationships between samples and has the potential to perform phylogenetic analysis. However, there are many unresolved samples in the consensus tree, and the consensus tree can not show relationships for all individual samples. The trees are even less resolve by strict rule (see supplementary figure 11). The tree sequence inferred for the whole chromosome has more than 10,000 trees, and it is harder to find a typical relationship among a larger number of trees (see supplementary figure 12).

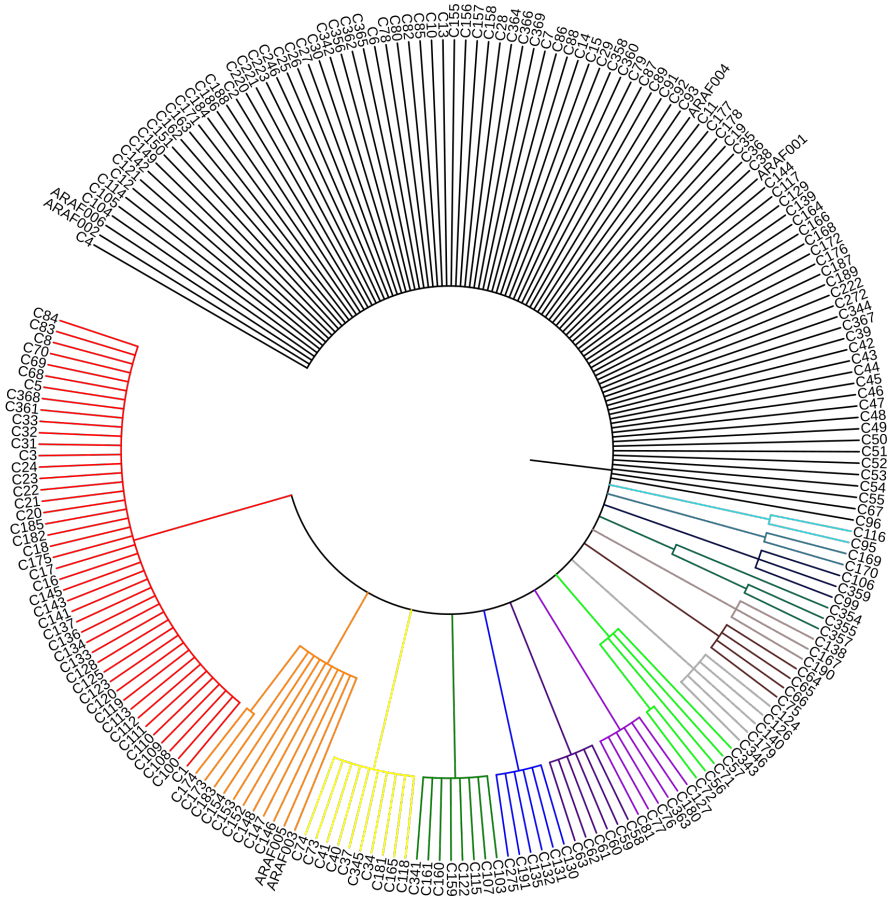


Figure 5: A majority rule consensus tree for the tree sequence of mitochondrial DNA with 219 *Aspergillus fumigatus* samples. Each clade in consensus tree are coloured differently, in order to map the nodes in the clade to figure 6. The unresolved samples stay uncoloured, almost half of the samples are uncoloured in this consensus tree.

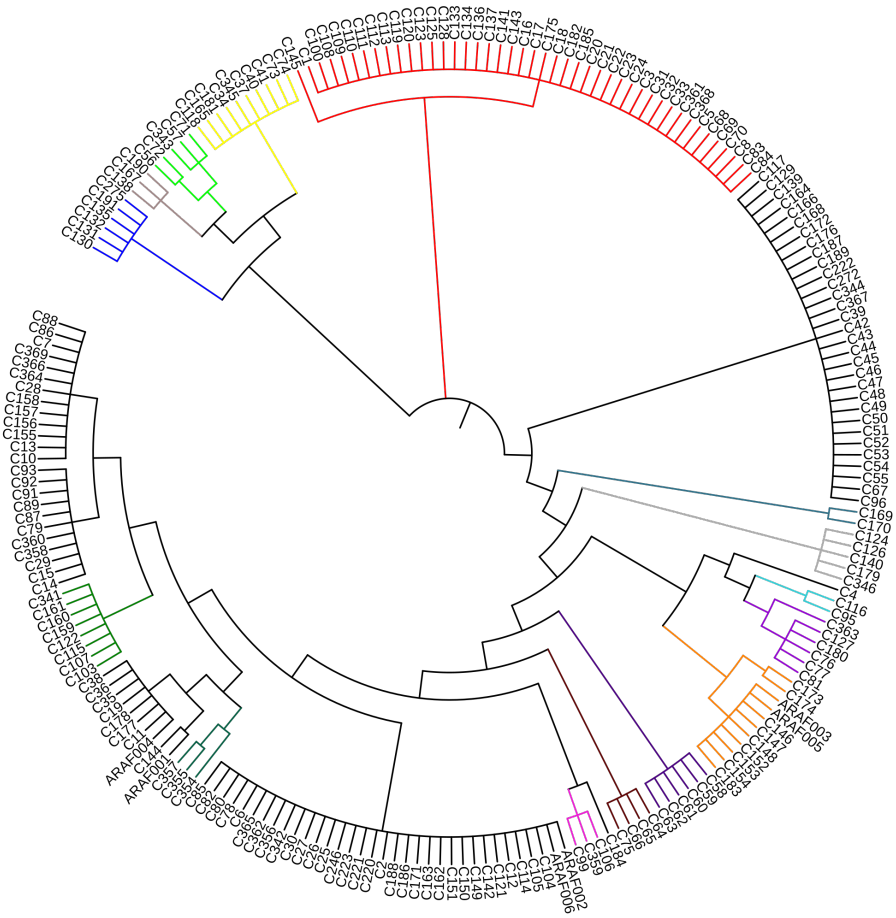


Figure 6: The phylogenetic tree plotted by `fasttree` from the SNP alignments of the mitochondrial DNA. Each coloured clade in Figure 5 are kept in this figure to show the relationship between consensus tree and this phylogenetic tree.

4.3 GNN results

Apart from drawing a consensus tree, computing genealogical nearest neighbours (GNN) is an alternative method to find the relationship between nodes in a sequence of trees. GNN is a statistic to evaluate the relationship between all samples nodes among a sequence of trees (details in Methodology section 3.5).

The full GNN results for *Aspergillus fumigatus* samples is a $219 * 219$ matrix, each index containing the relationship between nodes by numbers from 0 to 1. The sum of each row is 1, and a higher number stands for a closer relationship between sample nodes. Figure 7 shows GNN values among the selected number of nodes that each node at least has one GNN value larger than 0.2 (full heatmap in Supplementary figure 13). By reordering the selected samples, the heatmap of GNN values illustrates that most of the samples are not related, and

a few samples are closely related. In particular, node ARAF001 and C144, C95 and C116, C169 and C170, C173 and C174, C355 and C357, C56 and C72 are the six pairs of samples which are related closely. All the samples and relationships shown in the heatmap can also be found in a clade of the phylogenetic tree.

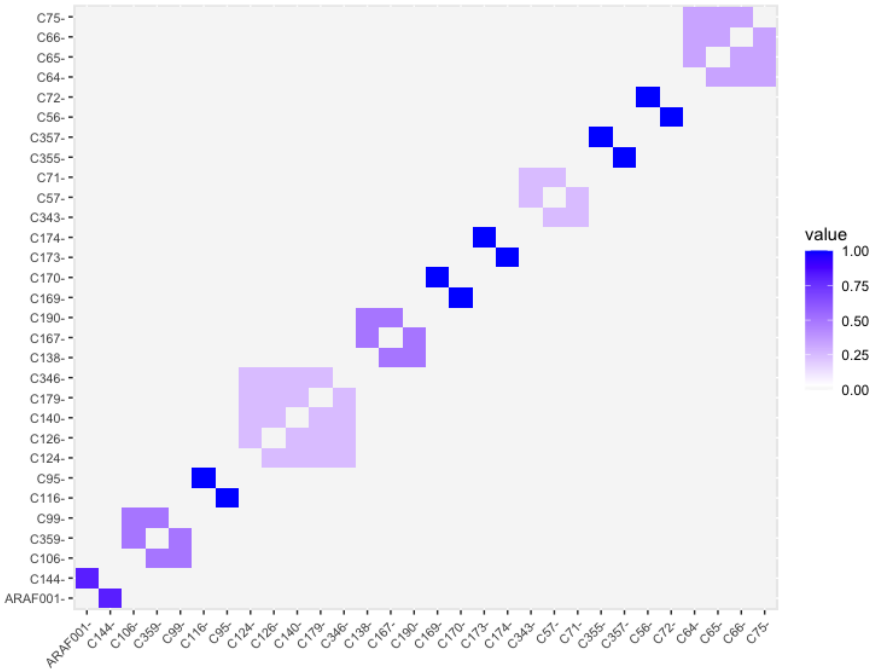


Figure 7: A heatmap representing part of GNN values for *Aspergillus fumigatus* mitochondrial DNA of 30 individual samples with the highest GNN values. The heatmap is reordered to combine related samples together to obtain a clearer observation.

The GNN statistic has another critical advantage, and it can be applied to find the relationship between groups of sample nodes by their metadata. Any common evolutionary pattern in *Aspergillus fumigatus* samples is not simple to obtain when I consider 219 samples separately. *Aspergillus fumigatus* datasets are measured carefully together with multiple different sets of metadata. In this project, I also map different metadata groups into the phylogenetic tree inferred by `fasttree` and compare it with GNN values using the same metadata to evaluate the performance of GNN results.

AMR	TR34	TR46	G54	Wildtype	Else
TR34	0.918691	0.011719	0.013013	0.055638	0.000939
TR46	0.198544	0.690294	0.011204	0.097584	0.002375
G54	0.085794	0.007111	0.571710	0.333709	0.001675
Wildtype	0.070610	0.008527	0.047222	0.853808	0.019832
Else	0.163588	0.016018	0.026122	0.789906	0.004366

Table 1: Whole-chromosome GNN statistics between different groups of Antimicrobial resistance (AMR) from metadata of *Aspergillus fumigatus* isolates. TR34, TR46 and G54 are all alternative mutations conferring Azole resistance. Wildtype stands for no mutations.

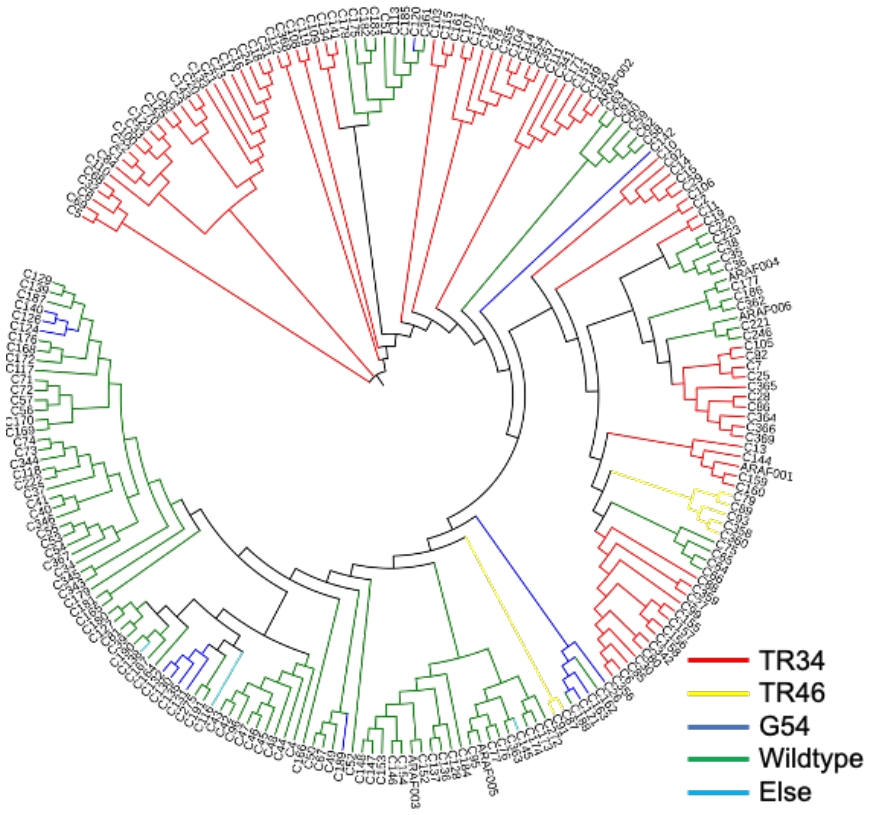


Figure 8: Phylogenetic tree by `fasttree` using *Aspergillus fumigatus* whole chromosome, coloured by different metadata Antimicrobial resistance (AMR) groups.

342 Table 1 illustrates a close relationship within groups using Antimicrobial resistance (AMR)

metadata of *Aspergillus fumigatus* whole chromosome. The total number of samples for TR34, TR46, G54, Wildtype and the rest AMR groups are 90, 7, 106, 13 and 3 respectively. The groups in AMR metadata represents the status of azole resistance mutation where azoles are a class of anti-fungal compounds. If there are no mutations, it is said to be the wildtype. TR34 and TR46 stand for a resistance mechanism consisting of 34-bp and 46-bp tandem repeat respectively, and G54 is a Glycine to Tryptophan at codon 54 (can either be a Tryptophan switch or Tryptophan repeat). GNN statistics are relatively high within groups, and GNN values are higher for group TR34 and Wildtype, which have more samples. Furthermore, all of the GNN values are low between different groups. However, for the group which has only 3 samples, GNN values are difficult to show the relationship within the group.

By mapping the metadata groups into the phylogenetic tree by `fasttree` (figure 8), most related samples in a clade belong to the same AMR group. Therefore, the different AMR groups may affect the evolutionary patterns of *Aspergillus fumigatus* samples, proving that evolutionary relationships and patterns are related to drug resistance mutation.

Year	2005-2011	2012-2014	2015	2016-2017	N/A
2005-2011	0.695704	0.143418	0.104459	0.035777	0.020642
2012-2014	0.152226	0.373595	0.333638	0.079877	0.060663
2015	0.060460	0.135215	0.650127	0.104986	0.049211
2016-2017	0.063940	0.102087	0.272825	0.496689	0.064459
N/A	0.080594	0.182977	0.340670	0.169436	0.226324

Table 2: Whole chromosome GNN statistics between different groups of measuring years from metadata of *Aspergillus fumigatus* isolates. N/A represents the 14 individuals not having the metadata of accurate time measured.

Metadata of year measured for *Aspergillus fumigatus* samples are also recorded. However, some of the samples do not have accurate time data. The data are measured from 2005 to 2017, where most of the samples are measured in 2015. Therefore, I merge the year metadata measured into 4 different year groups, and each group is a period of time with a reasonable number of samples. The samples without year information are put in the fifth group.

The GNN values within the same year groups are large for the whole chromosome, but they are not strongly related compared to the AMR metadata (table 2). This difference can be clearly shown if I map the year groups to the same phylogenetic tree of the whole chromosome (figure 9). Some of the clades still contain samples within the same year group, but the others

have sample nodes from all different year groups.

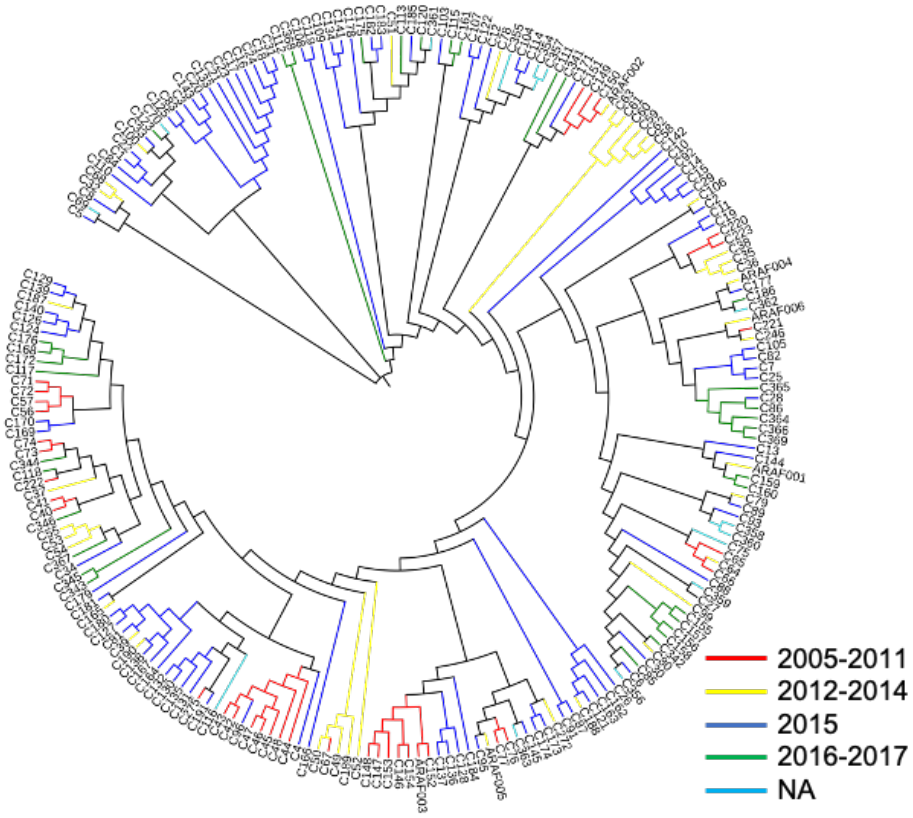


Figure 9: Phylogenetic tree by `fasttree` using *Aspergillus fumigatus* whole chromosome, coloured by different metadata year groups.

The *Aspergillus fumigatus* samples also provide metadata of clustering results for the whole chromosome. The clustering method is the discriminant analysis of principal components (DAPC) (Jombart et al., 2010). It is a hypothesis-free approach to see how many distinct genetic groups there were in the dataset without using phylogenetics. In this case, GNN values can be applied to evaluate the relationship between DAPC and `fasttree`. DAPC separates the samples into three groups, and all of the groups have an extremely high GNN value within groups (table 3). This can be mapped to the phylogenetic tree, almost divide the tree by the colours of different cluster groups. Note that there is a sample missing the clustering information in this case. From figure 10, it should belong to the second cluster. And the GNN values give the same results, the N/A group has the highest GNN value for cluster 2.

DAPC cluster	1	2	3	N/A
1	0.940305	0.038344	0.021049	0.000301
2	0.038813	0.909853	0.050464	0.000869
3	0.032672	0.098105	0.869079	0.000143
N/A	0.406615	0.460920	0.132465	0.000000

Table 3: Whole-chromosome GNN statistics between different groups of DAPC clustering result from metadata of *Aspergillus fumigatus* isolates, with one of the clustering metadata missing.

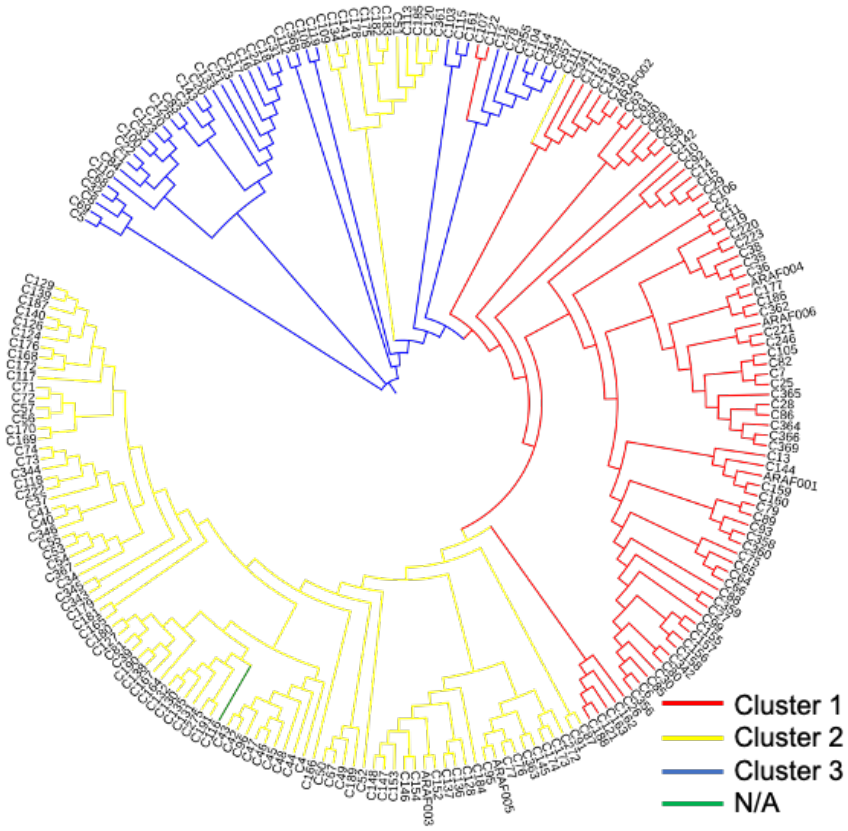


Figure 10: Phylogenetic tree by `fasttree` using *Aspergillus fumigatus* whole chromosome, coloured by different discriminant analysis of principal components (DAPC) clustering metadata groups.

5 Discussion

The technologies for observing the genome sequences have been improved rapidly for the last 15 years, making many whole-genome sequences available including fungi genomes (Catlett et al., 2003). Current studies in phylogenetics are attempting to infer phylogenetic networks to obtain evolutionary relationships (Solís-Lemus et al., 2017). However, phylogenetic networks have relatively high computational cost, which is not suitable for analysis of whole-genome sequences. Kelleher et al. (2019)'s tree sequence method infers a sequence of trees by introducing an approximation to infer ancestral haplotypes. This method makes the further evolutionary analysis of whole-genome sequences feasible.

In this project, I applied Kelleher et al.'s method to 219 different isolates of *Aspergillus fumigatus* whole-genome sequences. Using the longest chromosome as a reference genome (Nierman et al., 2005), the SNPs of the isolates can be inferred into a sequence of 16212 different trees. The total size of trees in tables for all chromosomes only takes about 25MB space to store, while the filtered SNPs take more than 1GB. Therefore, the sequence of trees has the potential to save storage space for large datasets of whole-genome sequences, and tree sequence will perform even better in more massive datasets. The tree inference algorithm is based on an approximation, but this produces reasonable accuracy (Kelleher et al., 2019). Two methods are introduced to perform further evolutionary analysis for the tree sequence in my project. Tree sequence can be merged into a consensus tree, and related nodes in the consensus tree have the same evolutionary relationships compared to phylogenetic tree generated by other tree inference methods. Using different sets of metadata of *Aspergillus fumigatus* samples, a statistic called genealogical nearest neighbours (GNN) can be obtained to evaluate the relationships between different groups of metadata. GNN results of *Aspergillus* clearly show a relationship between evolutionary patterns and different metadata groups.

Tree sequence is not a regular output in phylogenetics studies, and there are not many different methods and models for tree sequence analysis at the moment. Both the consensus tree and GNN values have their limitations. The consensus tree can find the related samples correctly, it is usually drawn by the majority (p=0.5) or strict (p=1) rules. Nevertheless, for a large number of trees, it is hard for all samples to find their related samples. The longest whole chromosome infers 16212 trees, a relationship that can pass majority rule requires appearances in more than 8000 trees. In fact, only 1/4 of *Aspergillus fumigatus* samples are related to other samples. GNN values using metadata of year measured are not performing as good as GNN values of other metadata groups. There are multiple reasons for that. First of all, a few samples do not have accurate time information in the metadata. Current GNN values do not have the option to ignore the samples with missing metadata, since the samples in tree sequence all fully connected to each other. Hence, all 14 samples missing metadata are

put in a new group. Another reason is the unbalanced weight for different groups. Almost 100 *Aspergillus fumigatus* samples are measured in the year 2015. This leads the GNN values from other year groups to the year 2015 relatively high. The last reason is that the data are measured within 12 years, which is a short period to investigate the evolutionary patterns. Therefore, from GNN results, there is an evidently evolutionary pattern from different years measured for *Aspergillus fumigatus* samples, but the pattern is not significant.

The *Aspergillus fumigatus* datasets in this project are measured in the UK and Republic of Ireland. Many researchers from different countries are also studying the azole resistance of *Aspergillus fumigatus* isolates (Chen et al., 2015; Riat et al., 2018; Wiederhold et al., 2016). In future, if there is a larger *Aspergillus fumigatus* sample dataset combining isolates from different countries measured in a longer period of time, GNN can provide robust results of clear evolutionary patterns for *Aspergillus fumigatus* samples across time. In addition, more researches in this field may lead to a statistical model, which produces more analyses and better results than GNN statistics.

Evolutionary analysis of whole-genome sequences is still under development. Kelleher et al.'s tree sequence method is an excellent inspiration for new methods rather than inferring phylogenetic networks. In particular, *Aspergillus fumigatus* samples have the metadata of drug resistance mutation groups (AMR), and GNN results clearly prove that there is an evolutionary pattern between different AMR groups. Recently, many lab groups are focusing on pathogenic potency and drug-resistance of fungi, especially on *Aspergillus* species (Paulussen et al., 2017; Sanglard, 2016). Resistance to triazoles was reported in *Aspergillus fumigatus* isolates cultured from patients with invasive aspergillosis (Snelders et al., 2008). Current studies on *Aspergillus fumigatus* investigate azole resistance and try to provide treatments (Resendiz Sharpe et al., 2018; Sewell et al., 2019a). This method has the potential to predict the evolutionary relationship for future mutated isolates of *Aspergillus fumigatus* from current whole-genome isolates, which may be useful in producing next-generation drugs.

Kelleher et al.'s tree sequence method also has great potential to store evolutionary data for future analysis for whole-genome datasets with lots of examples. For current 219 examples of whole-genome sequences, the tables storing tree sequence data outperforms the original VCF (Danecek et al., 2011) file 40 times, and are almost as good as PBWT format (Durbin, 2014).

For the whole *Aspergillus fumigatus* chromosome, Kelleher et al.'s tree sequence method takes about 200 seconds to infer a tree sequence, where the `fasttree` takes about 360 seconds to infer the phylogenetic tree. `fasttree` is also known as one of the quickest way to generate phylogenetic trees (Price et al., 2010), therefore tree sequence method is computational efficient comparing to most of the tree inference methods.

For some genome samples having missing or undefined metadata, GNN can also be applied to backtrack the metadata. By considering each missing sample as a group, GNN results

illustrate the genealogical proportions of the missing sample. The group with the highest GNN value for the sample has the closest evolutionary relationship, and are likely to be the correct metadata group (example see in GNN results of DAPC clustering). Similar approaches of backtracking missing data are only introduced in complex network structure (Hric et al., 2016), but not applied to genome datasets yet.

6 Data and Code

The main code are uploaded in the project Github page

Bibliography

J. B. Anderson. Evolution of antifungal-drug resistance: mechanisms and pathogen fitness. *Nature Reviews Microbiology*, 3(7):547–556, 2005.

M. Arenas. The importance and application of the ancestral recombination graph. *Frontiers in genetics*, 4:206, 2013.

M. Bordewich and C. Semple. On the computational complexity of the rooted subtree prune and regraft distance. *Annals of combinatorics*, 8(4):409–423, 2005.

N. L. Catlett, O. C. Yoder, and B. G. Turgeon. Whole-genome analysis of two-component signal transduction genes in fungal pathogens. *Eukaryotic cell*, 2(6):1151–1161, 2003.

Y. Chen, H. Wang, Z. Lu, P. Li, Q. Zhang, T. Jia, J. Zhao, S. Tian, X. Han, F. Chen, et al. Emergence of tr46/y121f/t289a in an aspergillus fumigatus isolate from a chinese patient. *Antimicrobial agents and chemotherapy*, 59(11):7148–7150, 2015.

P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.

R. Durbin. Efficient haplotype matching and storage using the positional burrows–wheeler transform (pbwt). *Bioinformatics*, 30(9):1266–1272, 2014.

D. W. Fadrosh, B. Ma, P. Gajer, N. Sengamalay, S. Ott, R. M. Brotman, and J. Ravel. An improved dual-indexing approach for multiplexed 16s rrna gene sequencing on the illumina miseq platform. *Microbiome*, 2(1):6, 2014.

D. M. Geiser. Sexual structures in aspergillus: morphology, importance and genomics. *Medical mycology*, 47(sup1):S21–S26, 2009.

R. C. Griffiths and P. Marjoram. Ancestral inference from samples of dna sequences with recombination. *Journal of Computational Biology*, 3(4):479–502, 1996.

D. Hawksworth and R. Lücking. Fungal diversity revisited: 2.2 to 3.8 million species. *Microbiology Spectrum*, 5, 07 2017. doi: 10.1128/microbiolspec.FUNK-0052-2016.

D. Hric, T. P. Peixoto, and S. Fortunato. Network structure, metadata, and the prediction of missing nodes and annotations. *Physical Review X*, 6(3):031038, 2016.

T. Jombart, S. Devillard, and F. Balloux. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC genetics*, 11(1): 94, 2010.

J. Kelleher, A. M. Etheridge, and G. McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, 12(5):e1004842, 2016.

J. Kelleher, Y. Wong, A. W. Wohns, C. Fadil, P. K. Albers, and G. McVean. Inferring whole-genome histories in large population datasets. *Nature genetics*, 51(9):1330–1338, 2019.

J.-P. Latgé. Aspergillus fumigatus and aspergillosis. *Clinical microbiology reviews*, 12(2): 310–350, 1999.

H. Li and R. Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16): 2078–2079, 2009.

N. Li and M. Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, 2003.

A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

M. J. Minichiello and R. Durbin. Mapping trait loci by use of inferred ancestral recombination graphs. *The American Journal of Human Genetics*, 79(5):910–922, 2006.

W. C. Nierman, A. Pain, M. J. Anderson, J. R. Wortman, H. S. Kim, J. Arroyo, M. Berriman, K. Abe, D. B. Archer, C. Bermejo, et al. Genomic sequence of the pathogenic and allergenic filamentous fungus aspergillus fumigatus. *Nature*, 438(7071):1151–1156, 2005.

C. Paulussen, J. E. Hallsworth, S. Álvarez-Pérez, W. C. Nierman, P. G. Hamill, D. Blain, H. Rediers, and B. Lievens. Ecology of aspergillosis: insights into the pathogenic potency of aspergillus fumigatus and some other aspergillus species. *Microbial biotechnology*, 10(2): 296–322, 2017.

G. A. Pavlopoulos, T. G. Soldatos, A. Barbosa-Silva, and R. Schneider. A reference guide for tree analysis and visualization. *BioData mining*, 3(1):1, 2010.

M. N. Price, P. S. Dehal, and A. P. Arkin. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3):e9490, 2010.

M. A. Quail, I. Kozarewa, F. Smith, A. Scally, P. J. Stephens, R. Durbin, H. Swerdlow, and D. J. Turner. A large genome center's improvements to the illumina sequencing system. *Nature methods*, 5(12):1005–1010, 2008.

A. Resendiz Sharpe, K. Lagrou, J. F. Meis, A. Chowdhary, S. R. Lockhart, P. E. Verweij, and I. A. R. S. working group. Triazole resistance surveillance in aspergillus fumigatus. *Medical Mycology*, 56(suppl_1):S83–S92, 2018.

J. Rhodes. Aspergillusdata, 2019. URL `https://microreact.org/project/BritishIslesAsp`.

A. Riat, J. Plojoux, K. Gindro, J. Schrenzel, and D. Sanglard. Azole resistance of environmental and clinical aspergillus fumigatus isolates from switzerland. *Antimicrobial agents and chemotherapy*, 62(4), 2018.

D. Sanglard. Emerging threats in antifungal-resistant fungal pathogens. *Frontiers in medicine*, 3:11, 2016.

T. R. Sewell, Y. Zhang, A. P. Brackin, J. M. Shelton, J. Rhodes, and M. C. Fisher. Elevated prevalence of azole-resistant aspergillus fumigatus in urban versus rural environments in the united kingdom. *Antimicrobial Agents and Chemotherapy*, 63(9):e00548–19, 2019a.

T. R. Sewell, J. Zhu, J. Rhodes, F. Hagen, J. F. Meis, M. C. Fisher, and T. Jombart. Nonrandom distribution of azole resistance across the global population of aspergillus fumigatus. *MBio*, 10(3), 2019b.

E. Snelders, H. A. Van Der Lee, J. Kuijpers, A. J. M. Rijs, J. Varga, R. A. Samson, E. Mellado, A. R. T. Donders, W. J. Melchers, and P. E. Verweij. Emergence of azole resistance in aspergillus fumigatus and spread of a single resistance mechanism. *PLoS Med*, 5(11):e219, 2008.

C. Solís-Lemus, P. Bastide, and C. Ané. Phylonetworks: a package for phylogenetic networks. *Molecular biology and evolution*, 34(12):3292–3298, 2017.

E. L. Van Dijk, H. Auger, Y. Jaszczyszyn, and C. Thermes. Ten years of next-generation sequencing technology. *Trends in genetics*, 30(9):418–426, 2014.

L. Wang, K. Zhang, and L. Zhang. Perfect phylogenetic networks with recombination. *Journal of Computational Biology*, 8(1):69–78, 2001.

N. P. Wiederhold, V. G. Gil, F. Gutierrez, J. R. Lindner, M. T. Albataineh, D. I. McCarthy, C. Sanders, H. Fan, A. W. Fothergill, and D. A. Sutton. First detection of tr34 l98h and

tr46 y121f t289a cyp51 mutations in aspergillus fumigatus isolates in the united states. *Journal of clinical microbiology*, 54(1):168–171, 2016.

Supplementary information

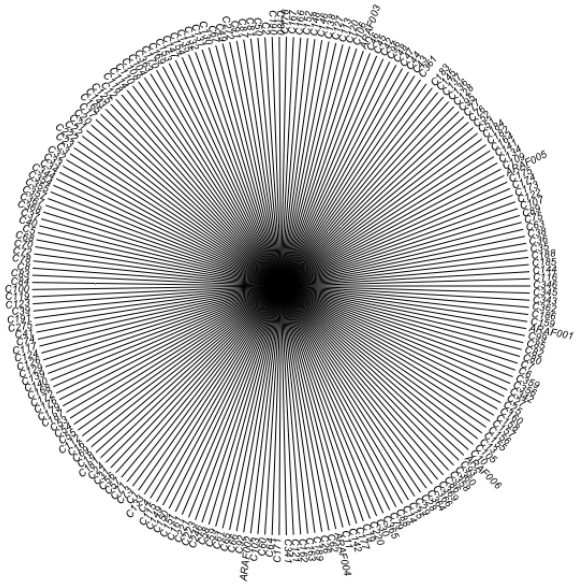


Figure 11: A strict rule consensus tree for the tree sequence of mitochondrial DNA of 219 *Aspergillus fumigatus* samples. None of the samples are shown with relationships

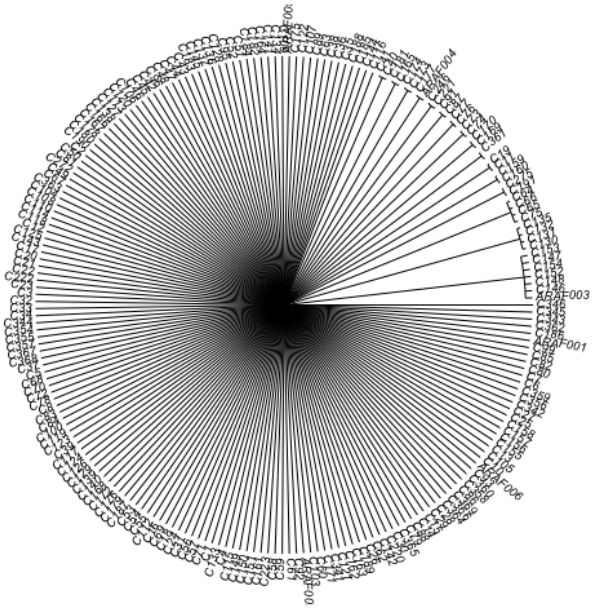


Figure 12: A majority rule consensus tree from tree sequence. The tree sequence is consists of 16212 trees inferring from whole chromosome of 219 *Aspergillus fumigatus* samples. Only less than 1/4 of the samples are related to each other

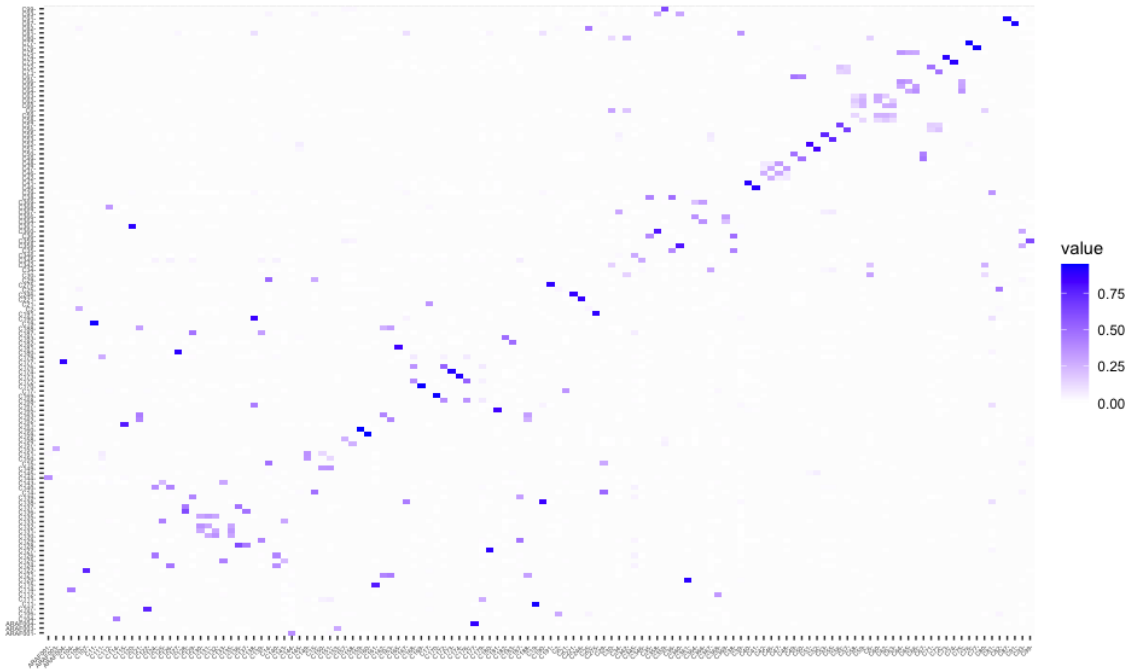


Figure 13: A heatmap representing GNN values for *Aspergillus fumigatus* whole chromosome with all 219 individual samples.