Third-year progress report for thesis advisory committee

This manuscript (permalink) was automatically generated from mbhall88/TAC3_Report@0c2eae0 on October 6, 2020.

Authors

- **Michael B. Hall**
  0000-0003-3683-6208 · mbhall88 · mbhall88
  EMBL-EBI; University of Cambridge · Funded by EMBL International PhD Programme (EIPP)

Thesis Advisory Committee

- Zamin Iqbal (Supervisor) - EMBL-EBI
- John Marioni (Chair) - EMBL-EBI
- Georg Zeller - EMBL Heidelberg
- Estée Török - University of Cambridge

Starting Date: 12/10/2017
Qualifying Assessment Date: 06/07/2018
Second TAC Meeting: 15/10/2019
Third TAC Meeting: 13/10/2020
Current Contract Expiry: 11/04/2021

Executive summary

Progress

- The content of Chapter 1 is effectively finished and a paper covering the work (second-author) is in the process of submission. All that remains is to write the thesis chapter.
- The major publication from this thesis comprises the work in chapters 2 & 3. I am on track to have this paper complete and submitted by the end of quarter 1 2021.
- All results for chapter 2 are close to complete. I have written some of this chapter and hope to complete the methods and results by the end of 2020.
- The "start-up" work and method refinement for chapter 3 are mostly complete as they are intertwined with chapter 2. I expect to have chapter 3 complete in quarter 1 of 2021.

Timeline

Chapter 1: written by April 2021.
Chapters 2 & 3: written by March 2021.
Chapter 4: Begin work in February/March 2021 and will likely require 6 months of work to get it to a stage where I can write a chapter about it. Unlikely to reach publication stage within my PhD.

Taking the progress and timelines outlined above into account, I would like to ask for a 6-month extension from 3.5 to 4 years.

Part A: Progress Report

Examining bacterial variation with genome graphs

Motivation and Background

The idea of using a single, linear reference genome to represent a population of individuals is not ideal. Even more so in the context of a bacterium such as *Salmonella enterica*, where two individuals can differ to such a degree that they only share 16% of their genes [1,2]. This fluidity of genetic content leads to the concept of a pan-genome; defined as the set of all genes observed in a species. Some genes are more common than others within this pan-genome, leading to the distinction of the 'core' and 'accessory' genome with the core being those shared across (most) all individuals and accessory being everything else. The effect of using a single-reference genome to describe variation within samples from the same species is best illustrated in Figure 1. In this toy example, the maximum number of SNPs we can hope to discover is only 56%. Clearly, comparing many samples requires better methods than single-reference based ones.

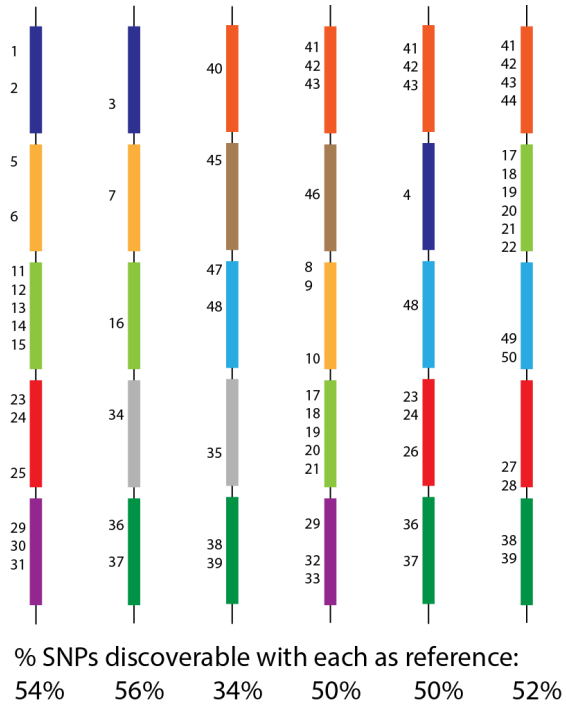

Figure 1: *An illustration of how reference bias can impact variant-calling. Each vertical column signifies an individual genome, with the coloured blocks representing genes. Numbers label 50 segregating SNPs. The percentages at the bottom express the proportion of SNPs which can be detected using each of the 6 'genomes' as a reference by mapping perfect reads from the remaining 5 to this reference.*

One of the first questions that arise from a computational point-of-view is how to apply current methods, which assume a single reference, to a pan-genome? An approach to answering this question that has been gaining considerable traction in recent years is that of genome graphs. This new paradigm uses a population reference graph (PRG) as its equivalent of a reference genome. A PRG is built to represent variation seen within a population, conceding the fact that no single genome can accurately represent an entire species. To construct such a PRG, one takes a multiple sequence alignment (MSA) and collapses shared sequence and creates "bubbles", or branch points, where they do not (see Figure 2 for an illustration of this process). We define a collection of PRGs as a pan-genome reference graph (PanRG), which we will interchangeably refer to as a pan-genome. Thus far, genome graph methods have focused mainly on eukaryotes. Given the rich diversity of genetic content in the prokaryote world, it would seem genome graphs are more suited to utilisation there.

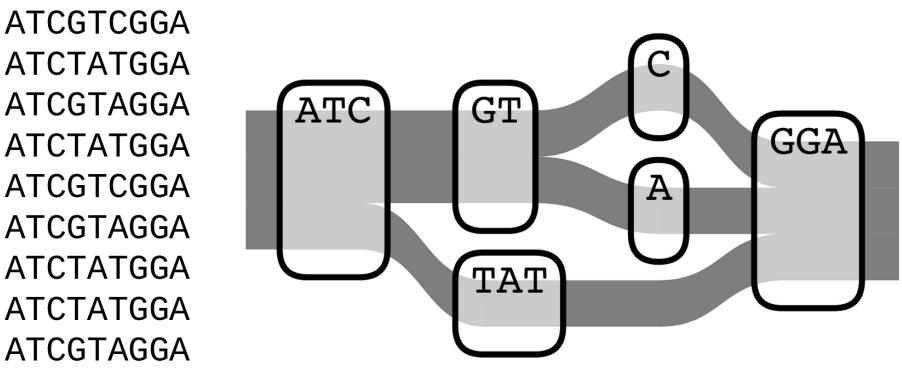

Figure 2: *An illustration of how a population reference graph (PRG) is constructed. Regions (columns) of shared sequence are collapsed into a single node. Those that differ are split into "bubbles", or branching nodes.*

In 2018, 1.4 million people died of tuberculosis (TB) globally, and over 10 million people fell ill to the disease, with 377,520 of those being multi-drug resistant (MDR) [3]. Standard of care requires phenotypic testing of the infecting organism against the four first-line drugs to ensure that appropriate treatment is prescribed. However, *Mycobacterium tuberculosis* (Mtb), the causative agent of TB, is a slow-growing organism and phenotypic testing takes around two months to complete. Whole-genome sequencing (WGS) offers a faster solution; recently it was shown that equivalent results are achievable by sequencing Mtb grown in liquid culture after two weeks of culture in contrast to the two-month traditional (Lowenstein-Jensen) culture method [4]. A number of genes are implicated in drug resistance and predicting resistance from sequencing data based on a catalogue of resistance SNPs and indels works with high specificity [5,6]. For the four first-line drugs, a study by the CRyPTIC consortium is the first of its kind to demonstrate that phenotyping is not required if genotype predicts susceptibility [7]. However, as the genetic basis for drug resistance is not entirely understood, there is still a sensitivity gap that differs drug-by-drug.

Nanopore sequencing yields ultra long reads with a mean/mode read identity in the range of 87-94%, while on a consensus level, it can achieve 99.94% identity with assembly polishing [8].
Variant calling with Nanopore sequencing data has seen a somewhat slow development. Currently, the main tools that have had reasonable testing done are `nanopolish` [9], Clair [10] and `medaka` [11]. A recent benchmark showed that Nanopore variant calling provides reliable diagnostic information for *Neisseria gonorrhoeae* [12]. However, to date, there has been no extensive Nanopore variant calling benchmark done for Mtb. Given the potential benefits of using genome graphs and long-read Nanopore sequencing for bacterial genomics, it makes sense to try and blend the two.

Pandora is a method being developed in the group to genotype across the *entire* pan-genome of a bacterial sample. It does this by working with a PRG, rather than a linear reference. The method is based on the following intuition: genomes evolve by recombination and mutation, and thus we ought to be able to approximate a $N+1$ genome as a mosaic of the first N genomes. `pandora` maps Nanopore reads to a graph encoding of a PRG, infers a mosaic, and provides genotypes at all variants in the PanRG. Mapping is done using minimising k-mers [13] in a similar vein to that done by `minimap` [14], and is therefore fast. By using Nanopore sequencing data, it is also possible to infer gene order as, in general, a single read will contain multiple genes, as opposed to Illumina sequencing where multiple reads are required to span a single gene. Note `pandora` does not (prior to this work) include any facility for discovering novel variation.

Summary

This PhD seeks to develop new methods to enable Nanopore-based diagnostics and epidemiology for Mtb and other bacteria. By using PRGs as a strong prior [15,16], we should be able to mitigate the Nanopore error biases and indel issues. In the process, we aim to construct a high-quality reference

pan-genome for Mtb that we hope will open previously inaccessible parts of its genome for investigation.

Chapter 1: Variant discovery in genome graphs

Variation in bacterial genomes can arise through a diverse range of processes. Mutations can occur during replication and are inherited vertically, genetic material can be transferred horizontally, and homologous recombination can lead to eukaryote-like gene conversion [17]. This breadth of ways in which bacteria can acquire new and varied genetic material results gives rise to the phenomenon of a pan-genome. Bacterial species with an "open" pan-genome may have individuals in their population who can share as little as 16% of their genes (*S. enterica*) [1]. Not all species' pan-genomes are this open, but it does raise the question: what do we use as a "reference" genome for such a species? One solution is to use the reference genome for the specific strain of interest. This works fine when dealing with a single sample or multiple samples of the same strain. However, when expanding to many samples from varying strains, the reference is no longer representative. An alternative solution is to focus solely on the core genome. The issue with this approach is the loss of information about variation in all of the non-core genes, which could be a large number if the pan-genome is open.

Prior work: Mosaic approximations and genotyping

As mentioned in Motivation and Background, `pandora` is a method developed by a previous PhD student in the lab, Rachel Colquhoun. It works on the premise of approximating a genome as a hierarchical mosaic. At a high-level, it represents a mosaic of loci - usually genes and intergenic regions - while at the locus-level, it is a mosaic of previously-seen genomes.

`pandora` aims to infer a consensus sequence from a PRG for a single sample or a collection of samples. In the case of a collection of samples, the consensus sequence will be one that best fits the collection of samples. `pandora` can additionally perform genotyping of the sample(s) with respect to this inferred consensus and produces a Variant Call Format (VCF) file. If a gene is present in only 2 of 50 samples then genotyping information is provided for those 2 samples and null for the other 48.

While `pandora`, before the work in this chapter, allows comparison of genomes to a level of detail provided by no other tool, there is still a significant shortcoming: it cannot discover novel variation. If a sample contains a variant not present in the PRG, the best `pandora` can do is select the path that is closest to that variant. The work in this chapter outlines a method for removing this limitation and provides an analysis of the gain in recall and precision by incorporating *de novo* variant discovery.

Local *de novo* variant discovery in a genome graph

There are two significant difficulties in discovering *de novo* variants on a graph. The first is finding regions within the graph that look like the reads mapping to them contain variation we do not have in the PRG. Secondly, we need to generate new paths for these regions and add them back into the PRG for consideration when remapping.

Finding candidate regions

We define a candidate region, r, as an interval within a local graph where coverage on the maximum likelihood path is less than a given threshold, c, for more than l consecutive positions. Any r within a specified distance of each other are then merged. For a given read that has a mapping to r, we define s_r to be the subsequence of the read mapping to r. We define the pileup P_r as the set of all $s_r \in r$.

Enumerating paths through candidate regions

We construct a de Bruijn graph from each P_r. Using anchor k-mers either side of the candidate region, find all paths that exist in the de Bruijn graph between the anchors, using a depth-first search (DFS) tree - starting with the left anchor. If the path is longer than a certain threshold, we discard it. Or, if there are too many paths, we abandon variant discovery for that region altogether.

Pruning the path-space in a candidate region

As the path enumeration step has the potential for combinatorial explosion due to cycles in the graph, and erroneous reads causing "dead-ends", we prune the DFS tree. To do this, we produce a distance map D_r by running reversed breadth-first search (BFS) on the de Bruijn graph, beginning from the *right* anchor. Each entry in the distance map describes the shortest path from that node to the end anchor. So as we walk along the DFS tree, if a node is not reachable within our maximum distance threshold, we abandon path enumeration.

I programmed the above methods in C++ and added them into the code base for `pandora`. They constitute 1325 lines of source code and 3486 lines of test code. I had help with the implementation of multi-threading the *de novo* component and the BFS pruning from Leandro Ishi.

Evaluation

Simulated data

I have previously presented results from evaluating this new *de novo* variant discovery method on simulated data to ensure if does indeed do as it advertises. For the sake of keeping this report concise, I will refrain from repeating the results here, except to say that this method does indeed find a large number of variants not already in the PRG.

Empirical data - multi-sample comparison

This section will be the major focus for the evaluation of both *de novo and* `pandora` and will constitute (part of) the results section of the `pandora` paper (near completion). We show the benefits of using a PRG, along with `pandora`'s performance compared to Nanopore- and Illumina-based variant callers. We compare 20 samples, from across the *E. coli* phylogeny. As other variant callers do not use a PRG we carefully selected 24 single-reference genomes representing the diversity of the *E. coli* tree as best as possible. Each variant caller was run once for each reference genome to illustrate the variance in results depending on the reference genome used.

In the process of developing an evaluation framework, we produced a python package called `varifier` (https://github.com/iqbal-lab-org/varifier). Briefly, for precision, `varifier` creates 'probes' for each variant in the VCF, using the genome the variants were called with-respect-to. It then maps these probes to the truth genome for the sample and determines the distance between the variant component of the probe and the part of the truth genome it maps to. For recall evaluation, `varifier` collects all differences in the pairwise alignment between the truth and VCF-reference genomes. Probes are created for these differences (based on the truth genome) and they are mapped to an augmented version of the VCF-reference genome, which has had the variants applied to it. The mappings are then evaluated in the same way as for precision.
As `pandora` calls variants for each sample with respect to an inferred best approximation sequence, creating the set of truth variants is slightly different. We perform a pairwise alignment for all pairs of samples and collect all the differences from this alignment. We then deduplicate this panel to ensure

variants are not “double counted”, meaning core genome variants would have an unbalanced effect on the overall precision and recall. We then follow the same probe-mapping approach from `varifier` with these truth variants.

Figure 3 shows, for a subset of four samples, two important results regarding the effect of (unfiltered) *de novo* variant discovery in `pandora`. Firstly, is shows that the choice of the Nanopore basecalling model has a sizeable impact - at least for *E. coli*. When using a methylation-aware model (not default), there is a significant increase in recall and decrease in error rate. While this has been previously described for Enterobacteriaceae [8], it highlights a weakness in the *de novo* variant discovery process. If the default model is used (green line in Figure 3), `pandora` has a higher error rate if discovering variants is enabled, however, we do get an increase in recall. This means that our discovery process is indeed picking up known systematic biases in Nanopore reads and we will need to apply careful filters to negate this. Removal of a large amount of this methylation-related systematic bias - by using a methylation-aware model (orange line in Figure 3) - shows that enabling discovery of novel variants improves both error rate and recall for `pandora`.

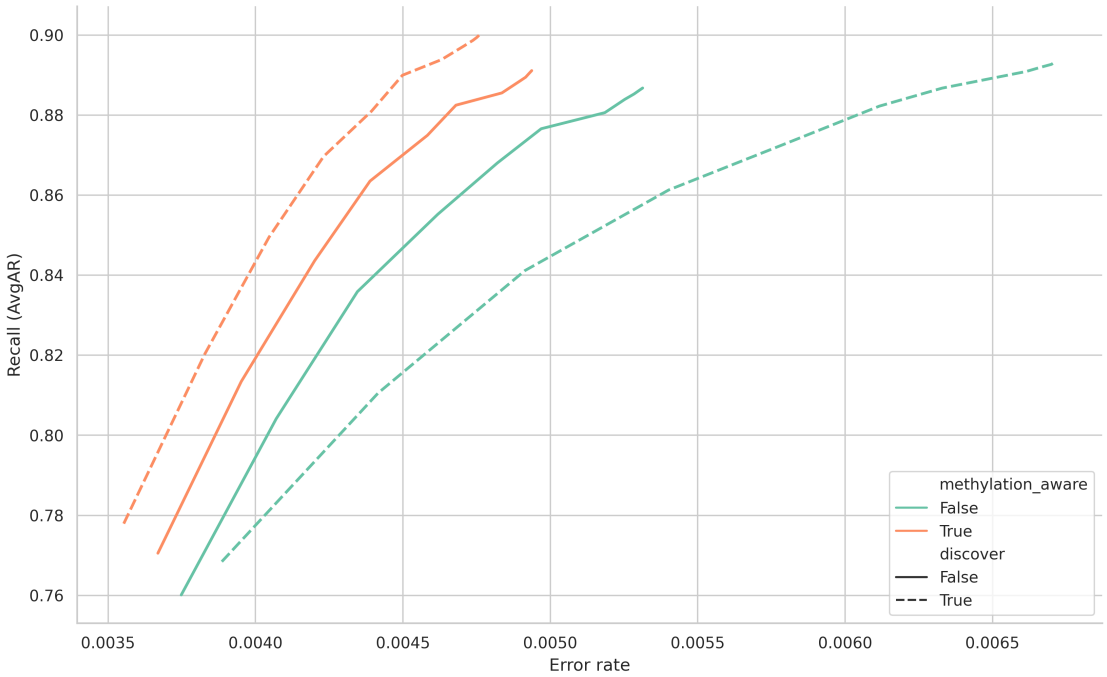

Figure 3: *Effect of Nanopore basecalling model on `pandora` variant calling error rate (1-precision; X-axis) and recall (Y-axis). The default model is in green, whilst the methylation-aware model is in orange. The dashed line represents using the de novo variant discovery method in `pandora`. The line shows the effect of increasing the genotype confidence score threshold - i.e. moves line towards bottom left.*

For comparison with other variant callers we filtered `pandora` (Nanopore/Illumina) variants based on the following criteria:

- Depth less than 10x/5x
- Less than 5%/5% of reads are on one strand
- 60%/80% or more of k-mers on the allele have zero coverage

As the main claim with the `pandora` method is that the use of genome graphs allows for more power to discover variants in the accessory genome, we assess the recall in that light. We group loci based on the number of samples they are found in and calculate the recall for each group. The more samples a locus is found in, the more likely it is a core genome locus. Figure 4 shows that in loci shared by less than 13 (65%) samples, `pandora` has recall at least inline with other variant callers for both Nanopore and Illumina reads. However, when the number of shared samples is less than 6 (30%), `pandora`’s recall is significantly greater for both sequencing technologies.

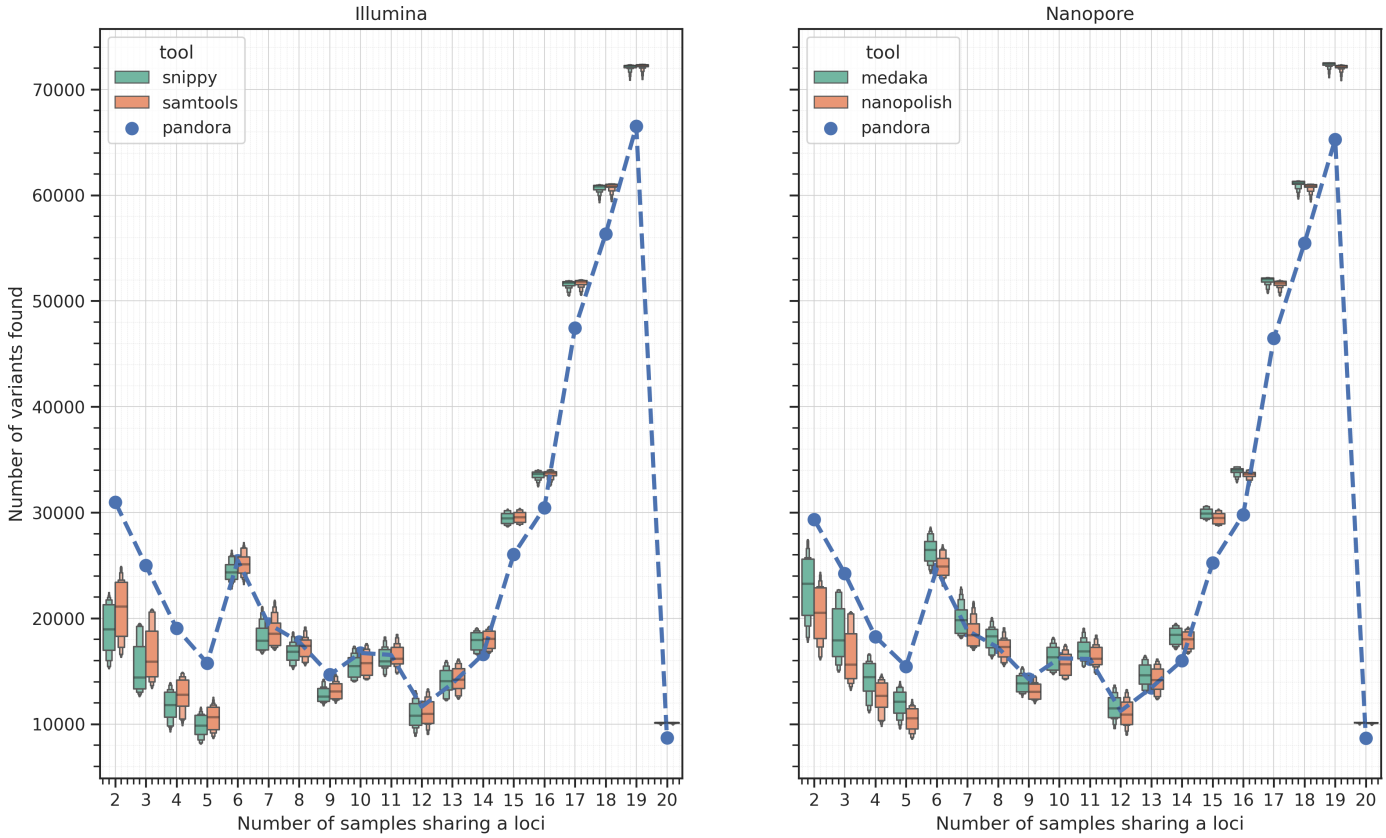

Figure 4: *Variant caller recall across the pan-genome. The Y-axis shows the number of variants found. The X-axis shows, for each locus, how many samples is it found in (i.e. left is more "accessory" and right is more "core"). The panels split the results by Illumina (left) and Nanopore (right) data and variant callers. The boxes represent using different reference genomes from across the phylogenetic tree to call variants (for non-* `pandor` *callers).*

We then show in Figure 5 that the variants called by `pandora` have significantly higher precision than `nanopolish` and `medaka` for Nanopore data and `samtools` for Illumina. `snippy` (Illumina) does have a superior precision to `pandora` though. One other important point is that, aside from `snippy` all other variant caller's precision varies quite a lot depending on the choice of reference genome. `pandora`, on the other hand, maintains consistent precision regardless of the phylogroup of a sample.

Both of these results highlight that the choice of reference genome can have a big effect on the results of variant calling. However, using a genome graph to represent a pan-genome, you gain access to variation from the accessory genome that is "invisible" to standard linear reference-based approaches.

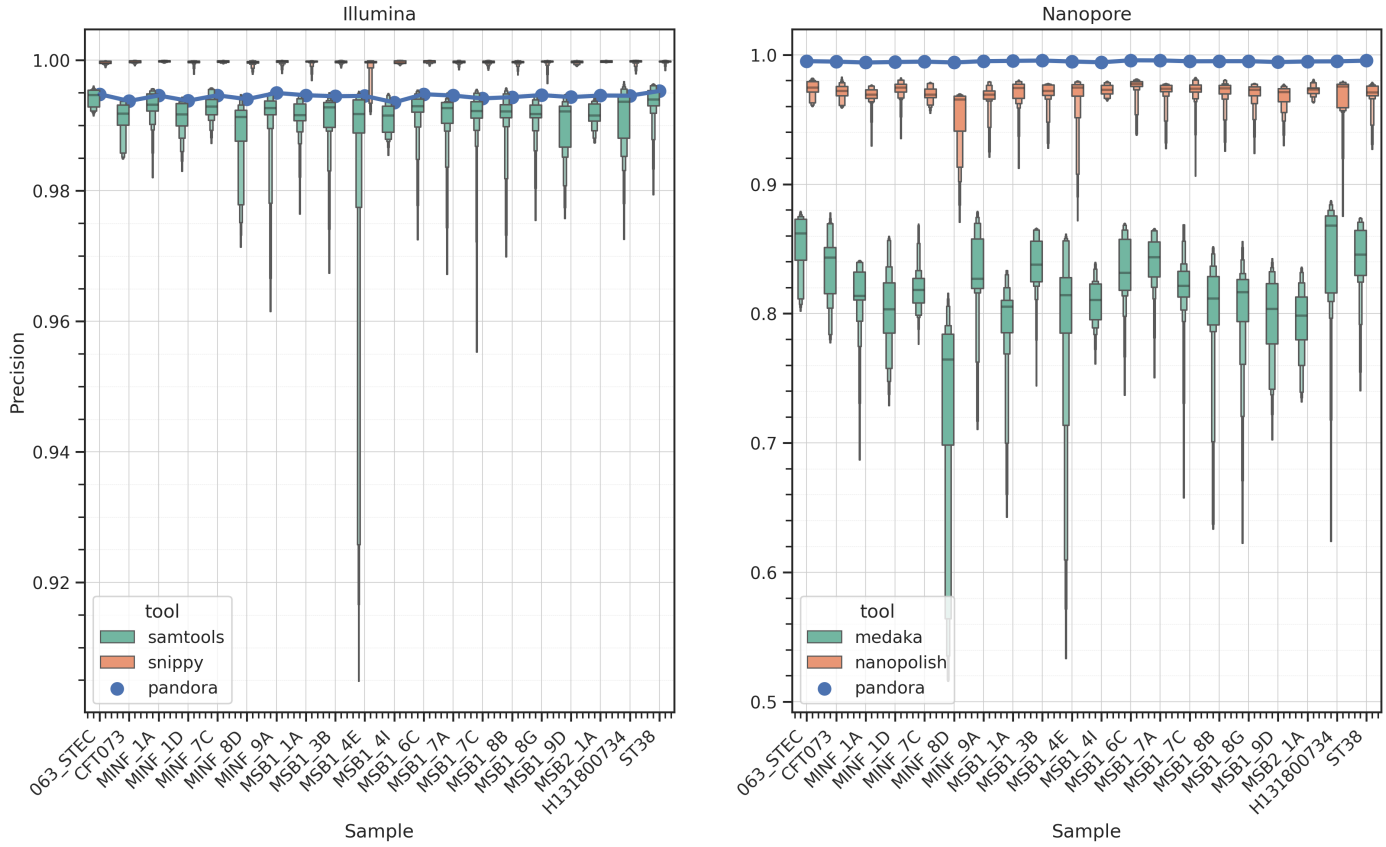

Figure 5: *Precision (Y-axis) per-sample for Illumina (left) and Nanopore (right) data. The boxes represent using different reference genomes from across the phylogenetic tree to call variants (for non-* `pandora` *callers) and are coloured by variant caller.*

Whilst I wrote nearly all of the code and associated tests for evaluating the recall for this analysis, a lot of it has since been refactored by Leandro Ishi and by Martin Hunt in the form of `varifier`. Due to the novelty of the method we are proposing, this has taken quite a lot of time and thought. I wrote 1250 lines of code to evaluate the methods, along with 3500 lines of test code to ensure there are no bugs in our evaluation. Additionally, I built the original `snakemake` [18] pipeline of approximately 3500 lines of codes to orchestrate the entire evaluation and simulations (However, much of this has been rewritten by Leandro Ishi).

Outstanding work

The major work still outstanding for this project is the direct integration of *de novo* candidates back into the PRG. The current procedure requires a fairly laborious, multi-step process for adding *de novo* candidates into the PRG, requiring the user to run a separate pipeline from `pandora`. Ultimately this will need to be handled all within the `pandora` program with no intervention from the user.
Lastly, there is an ongoing refinement of the *de novo* variant discovery process from the work in Chapters 2 and 3. The analysis in these chapters lean heavily on `pandora`, but for Mtb, which has a very different pan-genome to that of *E. coli* - which was used for most of the development of `pandora`'s methods.

Chapter 2: Applications to *M. tuberculosis* Nanopore variant calling

Public health applications for genome sequencing of Mtb generally focus on three use-cases: species identification, prediction of drug resistance, and clustering of samples for epidemiological purposes. In this chapter, we plan to focus on how the methods developed in `pandora` can be used to improve clustering of samples - generally referred to as "transmission clusters" - while, in the next chapter, we will address the drug resistance prediction component. The intention is to be able to use Nanopore data for public health. Therefore, this chapter will focus on a head-to-head comparison of Nanopore

and Illumina sequencing technologies for classifying transmission clusters for Mtb. What we aim to show is that, contrary to current dogma, Nanopore sequencing technology has advanced to the point where it can be applied to this use-case to a standard acceptable by public health authorities.

Genetic clustering of samples

Although there is scientific interest in the question of identifying transmission chains from genetic data, all the actionable public health information exists in the identification of transmission clusters [19,20].

The first step towards clustering a set of genomes is determining a distance matrix. For the majority of bacteria, there is a necessary step of identifying recombination tracts - which will contain a high density of SNPs - and removing them. Removal of these SNPs is necessary as they will have arrived at a different rate to the putative molecular clock and will artefactually extend branch lengths on the phylogenetic tree [19,20]. In the case of Mtb, however, there is virtually no recombination [21,22,23], so this step in not required.

For this chapter, we define genetic distance to be the sum of genetic discordances, where missing data and heterozygosity do not cause discordance and study the clustering this definition generates.

Data

Each sample was sequenced on both Nanopore and Illumina platforms from the same isolate and DNA extraction. In total, we received 118 samples from Madagascar, 83 from South Africa, and 46 from the National Tuberculosis Reference Lab in Birmingham; giving us a total of 247 samples. As these samples are not reference isolates, we need to be able to compare both Illumina and Nanopore to a truth. To establish how each platform compares to the truth, we have additionally sequenced 35 of the Malagasy isolates with PacBio and will use the high-quality assemblies for validation of variant calls.

Quality Control

The first step in quality control (QC) was to exclude samples where the Nanopore and Illumina data were not perfectly matched. In total, we excluded 40 samples at this stage.

The remaining 207 samples were processed through a QC pipeline. The first step in the pipeline is decontamination of sequencing reads. We used the decontamination database from `clockwork` (https://github.com/iqbal-lab-org/clockwork), which contains a wide range of organisms, including viral, human, Mtb, non-tuberculosis Mycobacterium (NTM), and nasopharyngeal-associated bacterial genomes. Reads were mapped to the database and was output to a final decontaminated fastq file if it had any mapping to a non-contaminant genome. All decontaminated fastq files were subsampled to a depth of 60x (Illumina) and 150x (Nanopore) using `rasusa` [24].

The last step in the QC pipeline is to assign lineages for each sample. A panel of lineage-defining SNPs [25,26,27] was used in conjunction with a sample's Illumina VCF from the Baseline variant analysis for the lineage assignment.

We exclude samples from further analysis if they had coverage below 20x (Illumina) or 30x (Nanopore), or if they could not be assigned a single lineage. This filtering criteria led to a further 57 samples being excluded; leaving us with a total of 150 samples to use for the remainder of this work. In addition to the QC of the Illumina/Nanopore data, we sadly had to exclude 26/35 PacBio sequencing datasets due to mismatched Illumina/Nanopore data or PacBio coverage lower than 20x.

Baseline variant analysis

The truth set of variants for the Illumina data in this chapter come from running the Public Health England pipeline, COMPASS [28]. This pipeline will act as a guide to inform us about whether the results from the Nanopore data are comparable with those being used in real public health settings. COMPASS effectively uses `samtools` [29] to call variants and then applies a series of complex variant filtering. As a baseline for the Nanopore data, we use `bcftools` [30], with some filtering of variants to remove low-quality calls and with a mask to avoid repetitive and structurally variable regions of the Mtb genome.

Nanopore SNP concordance with Illumina

Whilst the Illumina SNP calls from COMPASS are filtered as part of the pipeline, we had to settle on filters for the Nanopore SNP calls from `bcftools`. We used the methodology from the section Comparing Illumina and Nanopore SNPs to truth assemblies to refine the filters in an iterative process. In the end we filter all SNPs with quality (QUAL column in VCF) below 60, a read position bias less than 0.05, a segregation-based metric above -0.5, or a variant distance bias below 0.002.

To assess how well the Nanopore SNPs agree with Illumina we first look at SNP concordance. Two metrics of interest here are the call rate - what proportion of COMPASS alternate alleles does `bcftools` make a reference/alternate call - and the concordance - what proportion of COMPASS alternate alleles does `bcftools` genotype agree with. We found that concordance is very high between the two technologies, with nearly all samples having a concordance greater than 99.5%. Call rate is a little lower than this, with the majority of samples being above 97%.

As transmission clusters are ultimately defined based on a SNP distance matrix, it is important to understand how such matrices differ between Illumina and Nanopore variant calls. To investigate this, a consensus sequence was generated from the filtered VCFs, with repetitive regions masked [31]. We then generate a pairwise SNP distance matrix for each sequencing technology from their respective consensus genomes using `snp-dists` [32].
Figure 6 shows the relationship of the pairwise distance of samples based on the sequencing technology used. While the relationship across all samples is interesting, in the context of defining transmission clusters, it is slightly misleading. Transmission clusters are defined by grouping together samples that are within a certain number of SNPs. The threshold used for this grouping is generally in the order of tens-of-SNPs [19] so it makes more sense to look at the distance relationship for samples that are closer to each other. In Figure 7, we limit to samples within an Illumina SNP distance of 100. It shows that, at this scale, the relationship between Illumina- and Nanopore-defined SNP distance is much closer. The correlation between the two can be quantified by the linear equation $y = 0.93x + 0.84$, where y is the predicted Nanopore distance between two samples, given the Illumina distance x. We can use this equation as a way of translating transmission cluster SNP thresholds for Illumina data to Nanopore. For instance, if clusters are defined as samples within 12 SNPs of each other, we can use this as x and define our Nanopore transmission clusters as $y = 0.93 \times 12 + 0.84 = 12.0$. So at a threshold of 12 SNPs, the Nanopore threshold would be the same as Illumina.

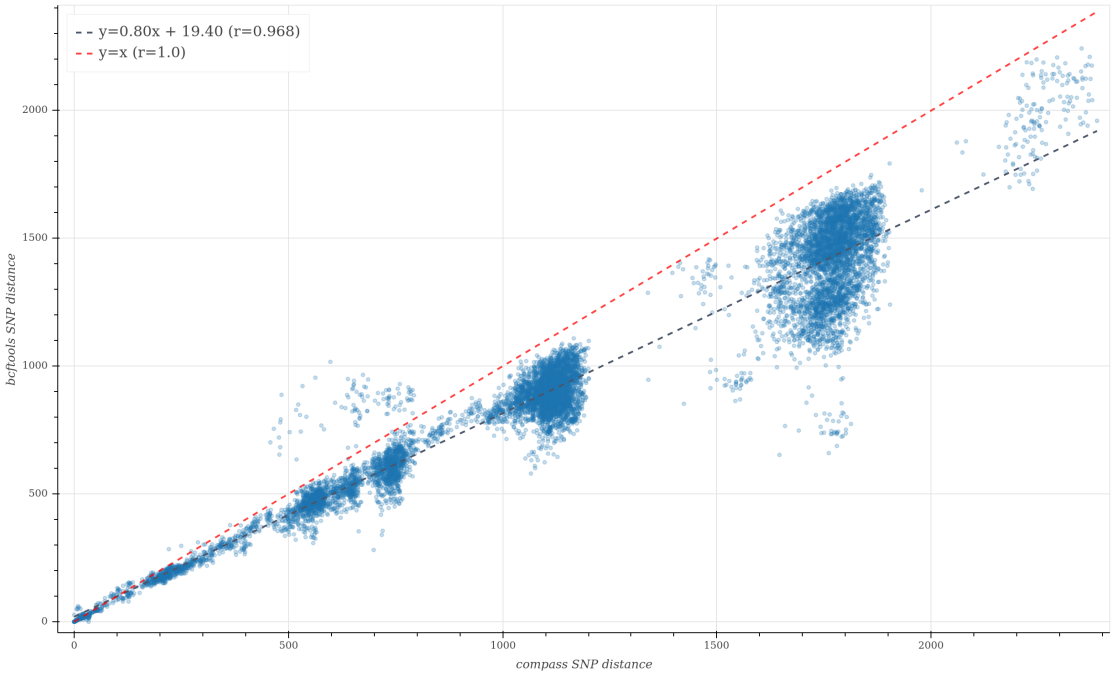

Figure 6: *Relationship between pairwise SNP distance for Illumina (COMPASS; X-axis) and Nanopore (`bcftools`; Y-axis). Each point represents a pair of samples. The red diagonal line is the identity line, which is where the points should lie if the distance between samples is the same for each technology. The black line shows the line of best fit for the data. The legend also shows the equations for these lines, along with their correlation coefficient (r).*

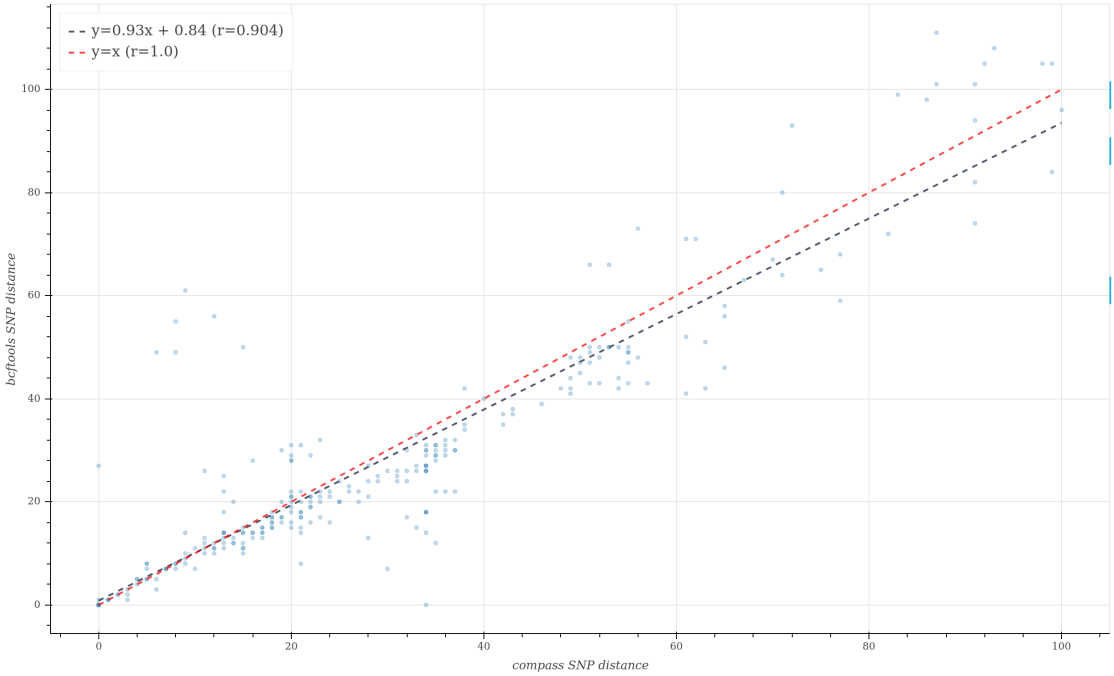

Figure 7: *Relationship between pairwise SNP distance for Illumina (COMPASS; X-axis) and Nanopore (`bcftools`; Y-axis) for samples within 100 SNPs of each other (based on Illumina distance). Each point represents a pair of samples. The red diagonal line is the identity line, which is where the points should lie if the distance between samples is the same for each technology. The black line shows the line of best fit for the data. The legend also shows the equations for these lines, along with their correlation coefficient (r).*

Comparing Illumina and Nanopore SNPs to truth assemblies

The analyses so far have treated the Illumina SNPs as a kind of "truth". In order to get a sense of how "correct" the SNP calls are for each technology we need to compare then to a "truth". For the nine samples with PacBio CCS data that passed QC, we generated assemblies (using only the CCS reads) with `flye` [33]. We masked any positions in the assembly where mapped Illumina reads did not have more than 90% agreement with the assembly, or had less than 10 reads. One sample was excluded due to the detection of other species contigs within the assembly. We then used `varifier`

to assess the precision and recall of the SNP calls for the eight samples with high quality assemblies. Figure 8 shows that the precision and recall for Nanopore is slightly lower than for Illumina, however, the difference is not large.

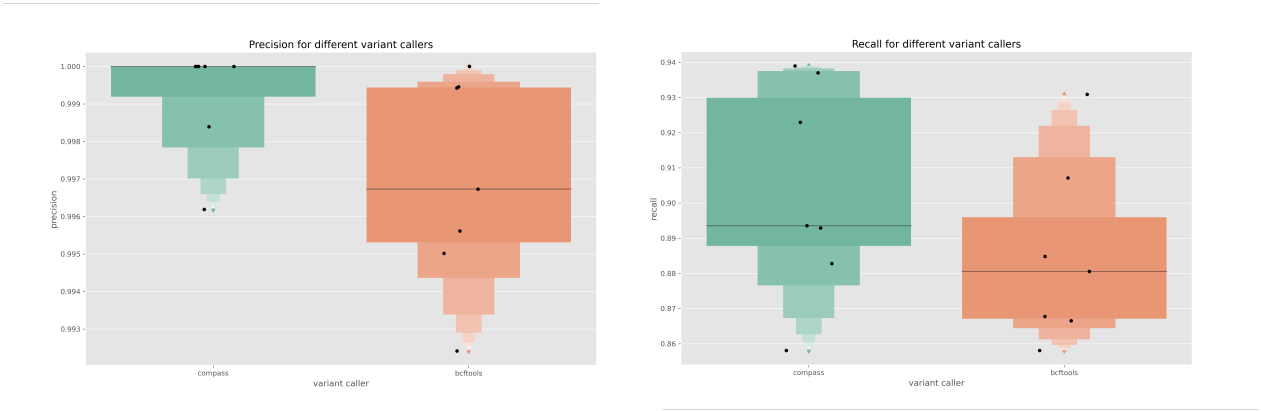

Figure 8: *Evaluation of the Illumina (COMPASS; green) and Nanopore (`bcftools`; orange) SNP calls to the PacBio CCS assemblies for eight samples using `varifier`. The left plot shows precision and the right is recall. Each point is one of the eight samples.*

Per-sample variant calls with `pandora`

The aim of this section will be to try varying degrees of PRG complexity for Mtb sample analysis. At this stage, we have two varieties in mind:

- Sparse PRG - H37Rv and all variants from a random selection of 100 samples from each lineage in the CRyPTIC (Comprehensive Resistance Prediction for Tuberculosis: an International Consortium) dataset [7].
- Dense PRG - The same as Sparse PRG, but 500 samples from each lineage.

In all of the above PRGs, we will apply the same mask from the baseline analysis and divide the genome into genes and intergenic regions, with a local PRG for each.

For each of the PRGs, we plan to perform the following analysis. Quantify the number of SNPs and indels `pandora` calls per-sample and see how this compares to the baseline and truth. We will report on the concordance rate, which is the proportion of shared sites between `pandora` and the truth that agree. Additionally, we will investigate how the complexity of the PRG effects the call rates and what the cost in computational performance is.

Multi-sample comparison

Multi-sample comparison suffers from two main challenges. First, large chunks of DNA may be present or absent across samples - this is the pan-genome effect. This effect causes significant issues with single-reference approaches, as outlined previously, but `pandora` was developed to address this. Second, when comparing a set of samples, the choice of reference affects how one describes variants. `pandora`, by design, chooses a reference for each loci PRG based on the current dataset, intending to maximise the succinctness of variant descriptions (see Figure 1). For example, we want to see SNPs as what they are - single-base variants - not as a nested region (as in Figure 1).

As mentioned, we handle these cases implicitly in `pandora`, however, the aim of this chapter and the next, in addition to comparing across technologies, will also be to compare methods for gaining drug resistance and epidemiological clustering information. To be able to compare with other methods though, we will need to be able to compare variants called with respect to a single-reference by other tools, with those from `pandora`.

With that in mind, this section will focus on producing a distance matrix for all samples using the result of the `pandora compare` routine and contrast this to those obtained from the single-reference methods.

Reproducing "truth" Illumina transmission clusters

In this section we will examine how well we can recreate the transmission clusters produced from COMPASS SNP calls with Nanopore data. Ultimately we will conclude with a recommendation of which method to call variants with: `bcftools`, `pandora map` (single-sample), or `pandora compare` (multi-sample). And what SNP threshold is required to ensure clusters are as similar as possible - if it *is* possible to get comparable clusters.

All of the analysis in this chapter was performed by myself - with the exception of the running of COMPASS - and all of the pipelines can be found at https://github.com/mbhall88/head_to_head_pipeline.

Outstanding work

I am currently producing the `pandora` variant calls. This has required an iterative refinement of the *de novo* variant discovery process - thus improving it's ability to detect clustered variants. Additionally, there is likely to be some software development time required to allow the `compare` routine to deal with samples of varying coverage. This will entail refining some prototyping that has been done already on genotype confidence percentile, which allows for normalising over coverage.
Once the `pandora` variants are in-hand, there just remains the clustering of samples. We don't anticipate any impediments to generating clusters and it *should* be a fairly quick process.

Chapter 3: Applications to improving *M. tuberculosis* drug resistance prediction

The genetic basis for drug resistance in Mtb is only partially understood. For the four first-line drugs, it is possible to detect the majority of resistant strains with high confidence [7], but for second-line, novel, and repurposed drugs, this is much harder.

Previous work from our group has shown that using WGS it is possible to create a panel of resistance markers and then successfully use this panel to predict drug resistance from Mtb sequence data [5,6]. This work involved the development of a software program called `mykrobe` to automate this inference and can use either Illumina or Nanopore data [34]. During the previous year, the predictive power of `mykrobe` has expanded and become even more accurate [35] (I played a small role in improving the likelihood calculations). This update of the `mykrobe` panel was mostly due to the recently published work from the CRyPTIC project [7]. CRyPTIC aims to perform drug susceptibility testing and WGS on 40,000 Mtb samples (many MDR) from all over the globe, and combine this with WGS data from another 60,000 samples. The goal of the project is to improve genotypic resistance prediction by expanding our catalogue of resistance mutations.

The proposed work for this chapter is based on the assumption that a large part of the work by this consortium (which our group is a critical part of) will be available. While there has already been a significant amount of data produced from CRyPTIC, there will be more coming during the remainder of my PhD. Given the collection of SNPs and indels identified as being necessary for resistance to the 14 major drugs tested, we want to show that we can detect them as well with Nanopore data as we can with Illumina.

Limitations of existing methods

Although there are many tools available for predicting drug resistance in Mtb from Illumina WGS, only two support the use of Nanopore data - `mykrobe` and `tb-profiler` [36]. In the recent update to `mykrobe` from our group [35], we have shown that the primary factor determining how well a tool performed was the catalogue of resistance mutations used. However, given the same panel, different tools do not perform identically, and therefore methodology is still important. While some studies have focused specifically on Nanopore data, all used a small sample size (Votintseva *et al.* 2017 n=5, Hunt *et al.* 2019 n=5, `tb-profiler` 2019 n=3). Additionally, `tb-profiler` and `mykrobe` both have limitations with their methods. Both `tb-profiler` and `mykrobe` only genotype with respect to known variants - i.e. they cannot detect novel variants. The CRyPTIC consortium recently introduced a new approach whereby if an unknown mutation is identified in a gene known to be involved in resistance, they refuse to make a call and instead send the sample for phenotyping [7]. On their 10,000 samples, this achieved a specificity and sensitivity for first-line drugs that was acceptable for clinical usage. This method is now in use at Public Health England for all Mtb samples in England. Hunt *et al.* 2019 [35] quantified the cost of the pure-genotyping approach of `mykrobe`, showing that 2.4-4.6% of resistant samples were missed. By introducing *de novo* discovery into `pandora`, I enable us to address this issue, and that is the focus for this chapter.

Drug susceptibility prediction for *M. tuberculosis* using `pandora`

The work in this chapter will aim to predict drug-resistance for Mtb using `pandora` and its new *de novo* component introduced in Chapter 1. The first step in this will be producing a gene-succinct PRG that includes variants from the above-mentioned CRyPTIC work that are known to cause resistance or susceptibility. This PRG will be easy to build as the alleles and probes for these variants-of-interest are already defined for `mykrobe`. I will write a software program (either an extension of `pandora` or a standalone tool) that takes the output of `pandora` used with this PRG and makes predictions about resistance, susceptibility, and whether phenotyping should be performed. As we know whether an allele causes resistance or susceptibility, this prediction will be straightforward to implement.

I plan to validate this approach with the data from Chapter 2, comparing concordance with `mykrobe` for Illumina and Nanopore. In particular, the analysis will focus on the (hopefully) lower coverage required by `pandora` to achieve the same, or better, results as `mykrobe`, and the increased detection power provided by *de novo* variant discovery.

Chapter 4: Construction of a *M. tuberculosis* reference pan-genome

Although Mtb has a closed pan-genome due to its lack of recombination and horizontal gene transfer [21,22,23], there are reasons why a pan-genome would be useful to the community. First of all, some genes exist within the pan-genome that are not present in the H37Rv reference genome [37]. Secondly, approximately 10% of the genome consists of so-called *pe/ppe* genes. These genes have a high GC-content, are very repetitive, and have been implicated in immune evasion and virulence [38,39]. The *pe/ppe* genes also harbour a disproportionately large amount of genetic diversity between isolates and a nucleotide diversity approximately 2-fold higher than the rest of the genome [39]. They are sufficiently similar that short reads fail to map, and are frequently masked out of the genome for variant calling. The ability to accurately map sequencing reads to these genes would likely improve our ability to perform variant calling in Mtb and therefore better determine how isolates relate to each other.

For this chapter, the aim is to build a high-quality pan-genome for Mtb, to allow variant discovery in *all* genes - ideally including the *pe/ppe* genes.

Assembly and multiple sequence alignment of high-quality *M. tuberculosis* genomes

To facilitate the desired outcomes of this chapter, we will first assemble the highest quality Mtb genomes from Chapter 2. These genomes, in the worst-case, have matched Illumina and Nanopore data and, in the best-case, PacBio too. The idea is that these assemblies will serve as the scaffold for the Mtb pan-genome, with the addition of other high-quality genomes outside of this thesis (2 PacBio assemblies from every lineage), and population variants discovered from the CRyPTIC consortium.

After assembly, we plan to perform large-scale multiple sequence alignment of these genomes to investigate how stable the Mtb genome is when ignoring the *pe/ppe* genes. The unknown quantity going into this chapter will be how easily *pe/ppe* genes can be assigned and matched across genomes. It may end up being necessary to assign genes from these genomes using synteny and parsimony with existing gene identifiers.

Ultimately, the overall gene ordering is not of great importance for the construction of a `pandora`-friendly pan-genome. The design of `pandora` is such that we divide the pan-genome into discrete pieces (loci) - i.e. genes and intergenic regions. Due to the low nucleotide diversity of Mtb, it may end up being necessary to split the Mtb genome into synteny blocks rather than the standard gene/intergenic approach we have used for other bacteria.

A genome graph map of *pe/ppe* genes

One question which will be of particular interest for this section will be whether reads covering one *pe/ppe* gene map to various others. If, as shown by others, *pe/ppe* genes arose through gene conversion [40], we would expect this to be the case. However, having high-quality assemblies built from a combination of long- and short-read technologies, we hope we can improve on the current nucleotide resolution and allow more accurate mapping to these genes. The main deliverable from this section will be a collection of high-quality *pe/ppe* PRGs with information about what read length will provide reliable mapping, and whether Illumina data can be reliably mapped to them. For some genes, this may be a no, but for others, we expect it will be possible to reliably map shorter reads than before.

Re-analysis of head-to-head data

Using this newly constructed, high-quality pan-genome, without the *pe/ppe* genes masked, we will re-analyse some of the data from the head-to-head analysis in Chapter 2. The aim is to see how many more variants we find, and whether we are better able to cluster samples as a result of having access to high-quality *pe/ppe* PRGs.

pe/ppe genetic variation in 10000 genomes

As mentioned previously, the CRyPTIC consortium are sequencing tens of thousands of Mtb genomes. In this section, we will look at the *pe/ppe* variation across 10,000 Mtb genomes using our newly constructed Mtb pan-genome and present the patterns we find. This work will be aided by a mycobacteriologist postdoc in our group who has extensive knowledge of Mtb biology.

Part B: Training and career development

Publication Strategy

- The paper covering the work in Chapter 1 is currently in preparation - titled *Nucleotide-resolution bacterial pan-genomics with reference graphs*. This paper covers the entirety of the `pandora` method. As the bulk of this method was produced by a previous PhD student in the lab, Rachel Colquhoun, I will be the second author. The work I will have contributed to this paper include the addition of the *de novo* variant discovery, and a large amount of the evaluation. We aim to submit the paper in October 2020.
- The work that will constitute Chapter 2 and Chapter 3 will also be contained in a paper, of which I will be the first author. We aim to have this work completed and a manuscript submitted by the end of November 2020 - or the first quarter of 2021 at the latest.
- The work in Chapter 4 is a bit harder to put an approximate date on, but we would hope to have it completed sometime in the third quarter of 2021, with myself being the first author.

List of publications, papers in press, preprints, manuscripts submitted/in preparation to date

In addition to the major "thesis publications" listed above, I am also involved in the following publications:

- The recent publication of `mykrobe` [35], in which I assisted with the improvement of the genotyping model for Nanopore data.
- A preprint [41] from a student-led project investigating the use of Nanopore sequencing for freshwater monitoring. I did a large part of the data analysis for this work.
- Supplementary to the work in Chapter 2 we also aim to submit a paper that will be a "crowd-sourcing" call for training species-specific Nanopore basecalling models. I used the data we have good PacBio assemblies for to try and train a Mtb-specific basecalling model, as it has been previously shown (in *K. pneumoniae*) [8] that this can improve read and consensus accuracy. This attempt did not yield a more accurate model than the default, but we think it is important to make this result available and challenge others to do the same for their species-of-interest.
- I am a co-author on a new manuscript for the workflow management system `snakemake` [18]. The manuscript is in the process of being submitted to *Nature Communications*. I contributed to this work by co-developing a tool called `snakefmt` (https://github.com/snakemake/snakefmt) that is used to provide automatic formatting of `snakemake` code - improving the readability of workflows. Additionally, I also developed and maintain the `snakemake` "profile" that configures job submission and status-checking for the LSF cluster system.
- A paper in review at *Emerging Infectious Diseases* by some of our collaborators in Madagascar about a lemur that contracted and died of TB. I performed the WGS analysis that was required to address previous major revisions they had been asked to make.
- I plan to submit an "Application Note" to *Bioinformatics* for `rasusa` [24]. This is a tool I wrote in the Rust programming language to randomly subsample sequencing reads to a specified coverage.

Work plan and timeline for thesis submission

I started working on my thesis in April 2020. For the last two months I have been spending approximately 2 hours a week writing.
The writing I have done includes:

- The methods section for *de novo* variant discovery in Chapter 1.

- The methods and results for sections in Chapter 2 covering the training of an Mtb-specific basecalling model, an assembly method benchmark for the samples we have PacBio data for, and the quality control of the data.

I plan to continue writing the results and methods for Chapter 2 and 3 in parallel with the daily work I am doing. As the work in these chapters will be published together (see Publication Strategy) I hope to have completed a well-polished draft for them both by the end of the year.

As a lot of the work in Chapters 2 and 3 leverage (and improve) the methods in Chapter 1, I will revisit this chapter once the work in these two chapters is complete. By this point, the `pandora` paper will be submitted, as will the paper for Chapter 2 and 3. This will allow me to write a "finalised" methods and results section for Chapter 1. As such, I plan to take some time at the start of next year (2021) to work solely on getting Chapter 1 to a near-finished state.

Chapter 4 will be approached in a similar manner to Chapters 2 and 3 - writing methods and results in parallel with the work I am doing. This means I will hope to have this chapter in a near-complete state, regardless of whether I "finish" the project before when I need to submit.

Lastly, I will work on the introductory chapter. I prefer to do this once I have a clear picture of all of the topics touched on in the body of the work.

I plan to submit (pending a six-month extension) my thesis in October 2021.

List of scientific courses and conferences attended to date and planned for next year

- RECOMB, Padova, Italy - April 2021
- Computational Pan-genomics workshop, Bielefeld, Germany - February 2021
- Applied Bioinformatics and Public Health Microbiology, Hinxton, UK - May 2021
- TBScience - October 2020 (virtual)
- Genome Informatics - September 2020 (virtual)
- CRyPTIC Consortium Annual Meeting, Hyderabad, India - October 2019
- Computational Pan-genomics workshop, Bielefeld, Germany - October 2019

List of additional training, teaching and other relevant activities to date

- Hackseq, October 2019 (virtual). I was involved in a project that created an automated bioinformatics bug discovery tool called `hypothesis-bio`.

Career development plan

Please describe your current long-term career aims (i.e. 3-5 years after PhD).

Within 3-5 years of completing my PhD I hope to be in a postdoctoral position. Geographically, I will be based in Queensland, Australia with my (soon-to-be) wife. I would like to continue working on Mtb genomics. Some Mtb-related topics I would like to pursue would be real-time analysis from a Nanopore device and also method development for improving our ability to analyse mixed infections. I also have a strong desire to do research within the reproducible bioinformatics field.

Please comment on the types of position you would like to apply to for after the PhD and your expected application timeline?

I have already been in discussions with my Master's supervisor about a job post-PhD. He recently received funding for a project aimed at implementation, in Papua New Guinea, of rapid direct-from-sample DNA sequencing approaches to characterise drug resistant tuberculosis, and evaluation of the accuracy of this approach compared to standard culture-based approaches. The bioinformatics component of this project is not yet filled. We are in discussions about the possibility of me joining the project in either a postdoc or research officer capacity.

There are also a couple of Australian-based postdoctoral fellowships I would like to apply for within the next year or two:

- Australian Research Council (ARC) Discovery Early Career Researcher Awards (DECRA). This grant provides focused research support for early career researchers in both teaching and research, and research-only positions. The applications for funding beginning 2023 closes in October 2021.
- NHMRC Investigator Grant: These grants provide the investigator with flexibility to pursue important new research directions as they arise and to form collaborations as needed, rather than being restricted to the scope of a specific research project. This grant has been postponed, but applications are likely to open in February/March 2021. I will need to begin discussions with a university partner in order to be able to apply for this funding.

What do you see as your strengths (2-3 skills)?

- Nanopore bioinformatics method development and analysis
- Bacterial variant-calling evaluation
- Software engineering

What do you see as your areas for improvement (2-3 areas)?

- Structural organisation of the Mtb genome
- Phylogenetics
- Writing and reviewing academic publications

What are your career development priorities until the end of your contract?

- Improve on my understanding of Mtb genomics at the nucleotide level, but also macro elements such as the pan-genomic landscape. This skill set will benefit both my PhD and my future postdoctoral aspirations to work on Mtb further.
- Reviewing and writing papers will be a valuable set of skills for applying to postdoctoral fellowships and is something that can fit into the completion of my PhD nicely.

What actions will you take to develop these skills?

- Chapters 2, 3, and 4 of my thesis will all dramatically improve upon my understanding of Mtb genomics.
- The TBScience conference I am (virtually) attending in October 2020 is another action I am taking to increase my Mtb knowledge. It is part of the Union World Conference On Lung Health and so will cover a wide range of Mtb topics.
- I am currently reviewing a paper together with Zam to learn how this process works. I plan to be involved in other reviewing opportunities before the end of my PhD.
- Aside from writing my thesis, I have at least three publications that I would like to first-author before the end of my PhD. The practise of writing and submitting these will no doubt help me

when it comes time to write a grant proposal for a postdoctoral position.
- I have registered for the EMBL “Applying to postdoc positions” workshop (late-October) and am also waiting for the next “Grant Writing” workshop to become available.

Part C: Impact of COVID-19 pandemic

Our research lab began working from home full-time on 16/03/2020. My partner, who lives in Australia, was due to arrive in Cambridge in early April for seven weeks. This trip had to be cancelled due to Australian border restrictions. The cancellation meant that I would not be seeing her for an unknown amount of time - having not seen each other since the beginning of 2020. As such, I organised to travel back to Australia. Trying to organise international travel at this time ended up being quite disastrous and after 3 flight cancellations, I was due to fly back to Australia on 27/04/2020. In the meantime, I found it very difficult to stay focused working at home. As I live in college accommodation, my office and bedroom are one-and-the-same. Adding the stress of trying to get back to my partner, I would say this was the most unproductive time of my PhD to-date. The day before my flight to Australia, on 27/04, I started displaying COVID-19 symptoms - fever, lethargy, lack of appetite - and had to isolate for a week. Needless to say, I did not work for this week and I flew to Australia the following week on 03/05. I was in hotel quarantine for 16 days in Australia, but this was surprisingly productive from my PhD perspective. However, my first 6 weeks after arriving in Australia were not as productive as I would have liked as my living arrangement with my partner was difficult to adjust to accommodate both of our working requirements. In the end, I started paying (from my predoc budget) to work from a co-working space three days a week and have additionally bought new home office equipment that has improved my ability to work from home. From July onwards I would say that my productivity has returned to pre-pandemic levels. I began meticulously tracking my work hours at the beginning of the pandemic and I am now doing roughly twice as many hours of productive work than I was prior to July.

I will remain in Australia until the beginning of January 2021 and continue to work three days a week from the co-working space - provided my predoc budget does not run out. Zam and I meet twice a week and I still attend group meetings every week - our lab will remain remote until January 2021.

Overall, I estimate that COVID-19 has caused me to lose about 6-8 weeks worth of time spent on my PhD.

References

1. **Why prokaryotes have pangenomes**
James O. McInerney, Alan McNally, Mary J. O'Connell
Nature Microbiology (2017-03-28) https://doi.org/gfw8gq
DOI: 10.1038/nmicrobiol.2017.40 · PMID: 28350002

2. **panX: pan-genome analysis and exploration**
Wei Ding, Franz Baumdicker, Richard A Neher
Nucleic Acids Research (2018-01-09) https://doi.org/gczkbr
DOI: 10.1093/nar/gkx977 · PMID: 29077859 · PMCID: PMC5758898

3. **Global tuberculosis report 2019**
Organisation mondiale de la santé
(2019)
ISBN: 9789241565714

4. **Mycobacterial DNA Extraction for Whole-Genome Sequencing from Early Positive Liquid (MGIT) Cultures**
Antonina A. Votintseva, Louise J. Pankhurst, Luke W. Anson, Marcus R. Morgan, Deborah Gascoyne-Binzi, Timothy M. Walker, T. Phuong Quan, David H. Wyllie, Carlos Del Ojo Elias, Mark Wilcox, … Derrick W. Crook
Journal of Clinical Microbiology (2015-04) https://doi.org/f65dtt
DOI: 10.1128/jcm.03073-14 · PMID: 25631807 · PMCID: PMC4365189

5. **Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis**
Phelim Bradley, N. Claire Gordon, Timothy M. Walker, Laura Dunn, Simon Heys, Bill Huang, Sarah Earle, Louise J. Pankhurst, Luke Anson, Mariateresa de Cesare, … Zamin Iqbal
Nature Communications (2015-12-21) https://doi.org/f755tg
DOI: 10.1038/ncomms10063 · PMID: 26686880 · PMCID: PMC4703848

6. **Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study**
Timothy M Walker, Thomas A Kohl, Shaheed V Omar, Jessica Hedge, Carlos Del Ojo Elias, Phelim Bradley, Zamin Iqbal, Silke Feuerriegel, Katherine E Niehaus, Daniel J Wilson, … Tim EA Peto
The Lancet Infectious Diseases (2015-10) https://doi.org/f3jjtq
DOI: 10.1016/s1473-3099(15)00062-6 · PMID: 26116186 · PMCID: PMC4579482

7. **Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing**
The CRyPTIC Consortium and the 100,000 Genomes Project
New England Journal of Medicine (2018-10-11) https://doi.org/d9kj
DOI: 10.1056/nejmoa1800474 · PMID: 30280646 · PMCID: PMC6121966

8. **Performance of neural network basecalling tools for Oxford Nanopore sequencing**
Ryan R. Wick, Louise M. Judd, Kathryn E. Holt
Genome Biology (2019-06-24) https://doi.org/gf4jwm
DOI: 10.1186/s13059-019-1727-y · PMID: 31234903 · PMCID: PMC6591954

9. **Real-time, portable genome sequencing for Ebola surveillance**
Joshua Quick, Nicholas J. Loman, Sophie Duraffour, Jared T. Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, Raymond Koundouno, Gytis Dudas, Amy Mikhail, … Miles W. Carroll
Nature (2016-02-03) https://doi.org/f88652
DOI: 10.1038/nature16996 · PMID: 26840485 · PMCID: PMC4817224

10. **Exploring the limit of using a deep neural network on pileup data for germline variant calling**
Ruibang Luo, Chak-Lim Wong, Yat-Sing Wong, Chi-Ian Tang, Chi-Man Liu, Chi-Ming Leung, Tak-Wah Lam
Nature Machine Intelligence (2020-04-06) https://doi.org/d9kq
DOI: 10.1038/s42256-020-0167-4

11. **nanoporetech/medaka**
Oxford Nanopore Technologies
(2020-10-04) https://github.com/nanoporetech/medaka

12. **High precision *Neisseria gonorrhoeae* variant and antimicrobial resistance calling from metagenomic Nanopore sequencing**
Nicholas D. Sanderson, Jeremy Swann, Leanne Barker, James Kavanagh, Sarah Hoosdally, Derrick Crook, Teresa L. Street, David W. Eyre, The GonFast Investigators Group
Genome Research (2020-09) https://doi.org/ghcbjr
DOI: 10.1101/gr.262865.120 · PMID: 32873606

13. **Reducing storage requirements for biological sequence comparison**
M. Roberts, W. Hayes, B. R. Hunt, S. M. Mount, J. A. Yorke
Bioinformatics (2004-07-15) https://doi.org/dkhs8w
DOI: 10.1093/bioinformatics/bth408 · PMID: 15256412

14. **Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences**
Heng Li
Bioinformatics (2016-07-15) https://doi.org/f8zxc3
DOI: 10.1093/bioinformatics/btw152 · PMID: 27153593 · PMCID: PMC4937194

15. **A Natural Encoding of Genetic Variation in a Burrows-Wheeler Transform to Enable Mapping and Genome Inference**
Sorina Maciuca, Carlos del Ojo Elias, Gil McVean, Zamin Iqbal
Lecture Notes in Computer Science (2016) https://doi.org/d9ks
DOI: 10.1007/978-3-319-43681-4_18

16. **Improved genome inference in the MHC using a population reference graph**
Alexander Dilthey, Charles Cox, Zamin Iqbal, Matthew R Nelson, Gil McVean
Nature Genetics (2015-04-27) https://doi.org/f7dt83
DOI: 10.1038/ng.3257 · PMID: 25915597 · PMCID: PMC4449272

17. **Genetic variation: molecular mechanisms and impact on microbial evolution**
Werner Arber
FEMS Microbiology Reviews (2000-01) https://doi.org/b4x7zf
DOI: 10.1111/j.1574-6976.2000.tb00529.x · PMID: 10640595

18. **Snakemake–a scalable bioinformatics workflow engine**
J. Koster, S. Rahmann
Bioinformatics (2012-08-20) https://doi.org/gd2xzq
DOI: 10.1093/bioinformatics/bts480 · PMID: 22908215

19. **Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions**
James Stimson, Jennifer Gardy, Barun Mathema, Valeriu Crudu, Ted Cohen, Caroline Colijn
Molecular Biology and Evolution (2019-03) https://doi.org/d9r7
DOI: 10.1093/molbev/msy242 · PMID: 30690464 · PMCID: PMC6389316

20. **Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review**
Hollie-Ann Hatherell, Caroline Colijn, Helen R. Stagg, Charlotte Jackson, Joanne R. Winter, Ibrahim Abubakar
BMC Medicine (2016-03-23) https://doi.org/f8gsk2
DOI: 10.1186/s12916-016-0566-x · PMID: 27005433 · PMCID: PMC4804562

21. **Bottlenecks and broomsticks: the molecular evolution of Mycobacterium bovis**
Noel H. Smith, Stephen V. Gordon, Ricardo de la Rua-Domenech, Richard S. Clifton-Hadley, R. Glyn Hewinson
Nature Reviews Microbiology (2006-09) https://doi.org/fhqqkv
DOI: 10.1038/nrmicro1472 · PMID: 16912712

22. **Evidence for Recombination in Mycobacterium tuberculosis**
Xiaoming Liu, Michaela M. Gutacker, James M. Musser, Yun-Xin Fu
Journal of Bacteriology (2006-12-01) https://doi.org/ftp6r2
DOI: 10.1128/jb.01062-06 · PMID: 16997954 · PMCID: PMC1698211

23. **Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of Mycobacterium tuberculosis**
Philip Supply, Michael Marceau, Sophie Mangenot, David Roche, Carine Rouanet, Varun Khanna, Laleh Majlessi, Alexis Criscuolo, Julien Tap, Alexandre Pawlik, … Roland Brosch
Nature Genetics (2013-01-06) https://doi.org/f4mrqv
DOI: 10.1038/ng.2517 · PMID: 23291586 · PMCID: PMC3856870

24. **Rasusa: Randomly subsample sequencing reads to a specified coverage**
Michael Hall
Zenodo (2020-03-27) https://doi.org/d9rz
DOI: 10.5281/zenodo.3731394

25. **Evolutionary pathway analysis and unified classification of East Asian lineage of Mycobacterium tuberculosis**
Egor Shitikov, Sergey Kolchenko, Igor Mokrousov, Julia Bespyatykh, Dmitry Ischenko, Elena Ilina, Vadim Govorun
Scientific Reports (2017-08-23) https://doi.org/gbvbxh
DOI: 10.1038/s41598-017-10018-5 · PMID: 28835627 · PMCID: PMC5569047

26. **Multiple Introductions of Mycobacterium tuberculosis Lineage 2–Beijing Into Africa Over Centuries**
Liliana K. Rutaihwa, Fabrizio Menardo, David Stucki, Sebastian M. Gygli, Serej D. Ley, Bijaya Malla, Julia Feldmann, Sonia Borrell, Christian Beisel, Kerren Middelkoop, … Sebastien Gagneux
Frontiers in Ecology and Evolution (2019-04-16) https://doi.org/d9r2
DOI: 10.3389/fevo.2019.00112

27. **Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages**
David Stucki, Daniela Brites, Leïla Jeljeli, Mireia Coscolla, Qingyun Liu, Andrej Trauner, Lukas Fenner, Liliana Rutaihwa, Sonia Borrell, Tao Luo, … Sebastien Gagneux
Nature Genetics (2016-10-31) https://doi.org/f9dg9j
DOI: 10.1038/ng.3704 · PMID: 27798628 · PMCID: PMC5238942

28. **Towards standardisation: comparison of five whole genome sequencing (WGS) analysis pipelines for detection of epidemiologically linked tuberculosis cases**
Rana Jajou, Thomas A Kohl, Timothy Walker, Anders Norman, Daniela Maria Cirillo, Elisa Tagliani, Stefan Niemann, Albert de Neeling, Troels Lillebaek, Richard M Anthony, Dick van Soolingen
Eurosurveillance (2019-12-12) https://doi.org/d9r9
DOI: 10.2807/1560-7917.es.2019.24.50.1900130 · PMID: 31847944 · PMCID: PMC6918587

29. **The Sequence Alignment/Map format and SAMtools**
H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup
Bioinformatics (2009-06-08) https://doi.org/ff6426
DOI: 10.1093/bioinformatics/btp352 · PMID: 19505943 · PMCID: PMC2723002

30. **Improving SNP discovery by base alignment quality**
H. Li

Bioinformatics (2011-02-13) https://doi.org/fw7k5k
DOI: 10.1093/bioinformatics/btr076 · PMID: 21320865 · PMCID: PMC3072548

31. **Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007–12, with whole pathogen genome sequences: an observational study**
Timothy M Walker, Maeve K Lalor, Agnieszka Broda, Luisa Saldana Ortega, Marcus Morgan, Lynne Parker, Sheila Churchill, Karen Bennett, Tanya Golubchik, Adam P Giess, … Christopher P Conlon
The Lancet Respiratory Medicine (2014-04) https://doi.org/f3hxn7
DOI: 10.1016/s2213-2600(14)70027-x · PMID: 24717625 · PMCID: PMC4571080

32. **Source code for snp-dists software**
Torsten Seemann
Zenodo (2018-09-09) https://doi.org/d9zj
DOI: 10.5281/zenodo.1411986

33. **Assembly of long, error-prone reads using repeat graphs**
Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin, Pavel A. Pevzner
Nature Biotechnology (2019-04-01) https://doi.org/gfzbrd
DOI: 10.1038/s41587-019-0072-8 · PMID: 30936562

34. **Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples**
Antonina A. Votintseva, Phelim Bradley, Louise Pankhurst, Carlos del Ojo Elias, Matthew Loose, Kayzad Nilgiriwala, Anirvan Chatterjee, E. Grace Smith, Nicolas Sanderson, Timothy M. Walker, … Zamin Iqbal
Journal of Clinical Microbiology (2017-05) https://doi.org/f94vt4
DOI: 10.1128/jcm.02483-16 · PMID: 28275074 · PMCID: PMC5405248

35. **Antibiotic resistance prediction for Mycobacterium tuberculosis from genome sequence data with Mykrobe**
Martin Hunt, Phelim Bradley, Simon Grandjean Lapierre, Simon Heys, Mark Thomsit, Michael B. Hall, Kerri M. Malone, Penelope Wintringer, Timothy M. Walker, Daniela M. Cirillo, … Zamin Iqbal
Wellcome Open Research (2019-12-02) https://doi.org/ggd835
DOI: 10.12688/wellcomeopenres.15603.1 · PMID: 32055708 · PMCID: PMC7004237

36. **Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs**
Jody E. Phelan, Denise M. O'Sullivan, Diana Machado, Jorge Ramos, Yaa E. A. Oppong, Susana Campino, Justin O'Grady, Ruth McNerney, Martin L. Hibberd, Miguel Viveiros, … Taane G. Clark
Genome Medicine (2019-06-24) https://doi.org/d949
DOI: 10.1186/s13073-019-0650-x · PMID: 31234910 · PMCID: PMC6591855

37. **Limitations of the Mycobacterium tuberculosis reference genome H37Rv in the detection of virulence-related loci**
Ronan F. O'Toole, Sanjay S. Gautam
Genomics (2017-10) https://doi.org/gchzjz
DOI: 10.1016/j.ygeno.2017.07.004 · PMID: 28743540

38. **The Enigmatic PE/PPE Multigene Family of Mycobacteria and Tuberculosis Vaccination**
Michael J. Brennan
Infection and Immunity (2017-03-27) https://doi.org/gbpvsd
DOI: 10.1128/iai.00969-16 · PMID: 28348055 · PMCID: PMC5442627

39. **Recombination in pe/ppe genes contributes to genetic variation in Mycobacterium tuberculosis lineages**
Jody E. Phelan, Francesc Coll, Indra Bergval, Richard M. Anthony, Rob Warren, Samantha L. Sampson, Nicolaas C. Gey van Pittius, Judith R. Glynn, Amelia C. Crampin, Adriana Alves, … Taane G. Clark
BMC Genomics (2016-02-29) https://doi.org/f8sf3h
DOI: 10.1186/s12864-016-2467-y · PMID: 26923687 · PMCID: PMC4770551

40. **Evolution and expansion of the Mycobacterium tuberculosis PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions**
Nicolaas C Gey van Pittius, Samantha L Sampson, Hyeyoung Lee, Yeun Kim, Paul D van Helden, Robin M Warren
BMC evolutionary biology (2006-11-15) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1660551/
DOI: 10.1186/1471-2148-6-95 · PMID: 17105670 · PMCID: PMC1660551

41. **Freshwater monitoring by nanopore sequencing**
Lara Urban, Andre Holzer, J Jotautas Baronas, Michael Hall, Philipp Braeuninger-Weimer, Michael J Scherm, Daniel J Kunz, Surangi N Perera, Daniel E Martin-Herranz, Edward T Tipper, … Maximilian R Stammnitz
bioRxiv (2020-02-07) https://doi.org/ghdcjv
DOI: 10.1101/2020.02.06.936302