Discussion

This thesis has focused on novel methodologies (genome graphs) and technology (Nanopore sequencing) as applied to bacterial genomics and particularly *M. tuberculosis*. In this concluding section, we will take a step back and outline where we think these can have an impact over the coming years.

Pan-genome reference graphs, as used by Pandora, will eventually facilitate the simultaneous analysis of thousands to tens of thousands of bacterial samples. Such population-scale analyses will open the door for many novel insights into the bacterial pan-genome. One future application is genome-wide association studies (GWAS) in recombining bacteria. An example use case for this would be to understand the genetic basis for a phenotype of interest. Doing this via a simple GWAS design, with cases and controls, one would look for associations between core genome SNPs and the phenotype. However, the larger the cohort, the smaller the core genome becomes (Figure 1.2c), and thus the less effective the GWAS will be.

Reference-free alternatives to this approach have been proposed, whereby associations between *k*-mers and the phenotype are used instead of SNPs [223, 224]. However, there are *vast* numbers of *k*-mers in a pan-genome, which creates statistical issues with multiple testing. In addition, *k*-mers do not represent a biological entity. That is, if one finds a *k*-mer to be significant, its location must be found by mapping to the pan-genome to determine its function.

As shown throughout this thesis, genome graphs provide a means to genotype SNPs across a cohort. Pandora provides a multi-sample VCF that describes not just the core genome but (nearly) the *whole* pan-genome, and the GWAS tests can then be performed on this VCF. This capability will facilitate new insights into genotype-phenotype association in bacteria via analyses of prior datasets and future population-scale ones.

The pan-genomic model we have described in this thesis lets go of gene ordering and even the definition of a locus. Many bacterial species have loci in common and can even share these loci. Indeed, some metagenomic applications disregard species definitions entirely and focus purely on loci. Furthermore, the unit of selection (loci) could be one of many different classes - e.g., gene, strain, plasmid, transposon. Pandora provides the flexibility to handle all of these definitions. As such, meta-genome graphs are something we believe will provide many fruitful insights. One application we foresee meta-genome graphs enabling is longitudinal studies of the microbiome and its response to drugs in patients with chronic infections - e.g., chronic *Pseudomonas aeruginosa* infections in the lungs of cystic fibrosis patients and its impact on the lung microbiome [225].

This thesis took a narrow perspective in showing Nanopore's "non-inferiority" to Illumina for *M. tuberculosis* DST and clustering. However, Nanopore provides unique technological differences that will allow for exploring other very interesting, clinically-focused, biological questions. For one, polyclonal (mixed) infections commonly occur in tuberculosis [226] and in most pathogens [227]. Detecting such mixed infections from genomic data can be challenging, similar to phasing haplotypes in human genomics, but with unknown or complex ploidy. As long Nanopore reads span multiple informative variants, they help with such phasing problems [228]. In addition, we believe genome graphs will allow for determining such mixed infections in partnership with long reads based on support for different paths in a PRG. This would allow one to look at the longitudinal dynamics of different strains throughout infection and see how mutations compete within the host during treatment.

Another exciting possibility off the back of the methods and technologies in this thesis is the high-resolution investigation of *pe/ppe* genes. These genes have been notoriously difficult to probe due to the difficulty in mapping short reads to them. These 160 genes pose many difficulties for short reads due to their high GC content, homology, and repetitive nature. As a result, they are generally excluded from genomic analyses. Nanopore reads promise to improve interrogation resolution as they can span entire (and multiple) *pe/ppe* genes - leaving much less ambiguity about where a read aligns. Indeed there has been some initial investigation of exactly this [88]. While this work dramatically improved coverage of these genes, gaps remain. This is where genome graphs are likely to complement Nanopore; by increasing *a priori* knowledge and reducing ambiguity. Added to the fact that these genes have a higher mutation rate than the rest of the genome [88], genome graphs are ideally suited to handle the varying complexities of these genes. Increased resolution in the *pe/ppe* genes promises to improve our knowledge of *M. tuberculosis* host-pathogen interactions, virulence, AMR, and other yet-to-be-determined functions.

The prospect of real-time clinical diagnostics is a major attraction of Nanopore sequencing. While this functionality has been shown in theory, it remains to be implemented for *M. tuberculosis*. A critical component of this application will be direct-from-sputum Nanopore sequencing, which is still in its infancy. However, significant progress has been made [96, 215], and it is only a matter of time until this is routine. In aid of improving direct-from-sputum sequencing, we see *Read Until* [99, 100] as playing an important role. Read Until is an API to the flowcell that allows for artificial depletion/enrichment of a sample (or target regions) by permitting ejection of reads from a pore based on some criteria. Potential applications include rejection of non-*M. tuberculosis* reads or enriching AMR-associated or *pe/ppe* genes.

One final application that will likely advance *M. tuberculosis* research and clinical care is epigenetics. Nanopore sequencing inherently captures information about epigenetic modifications, and certain modifications can now be identified in other species [229]. Epigenetic modifications have been associated with drug resistance, virulence, and regulation of gene expression profiles in *M. tuberculosis* [230–232], but as yet, no exploration of this with Nanopore sequencing has been attempted. Combining this with direct RNA sequencing - another unique capability of Nanopore - the next decade should see a much better understanding of how gene expression and RNA modification influence *M. tuberculosis* functions.