

Efficient analysis of microbial
whole-genome sequence data using
de Bruijn graphs



Phelim Bradley

Trinity College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2017

To my family: Mary, Colin, Maeve, Niall; and my partner Katia.

This thesis, as well as many other good things in my life, would not have been possible without their constant support and encouragement.

Acknowledgements

My route to and throughout my DPhil has been supported and encouraged by so many people, it would be difficult for me to thank them all here. Many past teachers and great friends have inspired me to where I am, but they are too numerous to list.

Heartfelt thanks to my supervisors Zamin Iqbal and Gil McVean for being selfless with their time, and for their thoughtful guidance of my research. Thank you Zam for your boundless enthusiasm, optimism, support, and advice.

Thanks to the rest of the Iqbal and McVean groups, as well as other colleagues in the Wellcome Trust Centre for Human Genetics, who gave so much feedback on my work and made my time there so enjoyable. In particular, I'd like to thank the people who make the Iqbal group amazing: Sorina, Rachel, and Carlos; and the more recent additions: Martin, and Robyn, who were such entertaining company and supportive colleagues.

I am indebted to many of the staff at Oxford, whose logistic support made the programme possible; notably, the Research Computing team

who made the technical infrastructure, which was so vital to my work, run so smoothly.

My DPhil benefited from many people's advice, insight, and teaching. I am especially grateful to have had the opportunity to collaborate with the exceptionally talented people in the Modernising Medical Microbiology group in Oxford.

I thank the Wellcome Trust for their generous financial support.

Finally, a huge thanks to Katia, who depleted her red pens while proof-reading this thesis.

Abstract

Antimicrobial resistance (AMR) is a persistent and growing threat to global health. Whole genome sequencing (WGS) has the potential to dramatically improve our ability to detect, understand, and monitor AMR. However, microbial diversity and complexity means that the analysis and interpretation of their genomes is challenging. In this thesis, I explore applications of de Bruijn graphs (DBGs) to the analysis of these data.

First, I present a tool, **Mykrobe predictor**, that uses DBGs to rapidly identify species and AMR from WGS data. I show that it is accurate, flexible, and efficient.

Next, I explore an extension of **Mykrobe predictor** to long read sequencing of direct clinical samples of *M. tuberculosis*. In doing so, I show that one could reduce the turn-around time for susceptibility testing of an *M. tuberculosis* isolate from 2 weeks to 12 hours.

Finally, I explore the challenges of DNA search in very large collections (millions) of microbial data sets. In particular, I address the super-linear scaling of existing k -mer indexing tools and present a novel representation and implementation of a probabilistic coloured de Bruijn graph, “Coloured Bloom Graph” (CBG). I demonstrate its scalability by building a CBG of all publicly accessible microbial WGS data (almost half a million samples) and use it to run millisecond searches in these data.

Contents

1	Introduction	1
1.1	Antimicrobial resistance	1
1.2	Mechanisms of resistance	4
1.3	Clinical microbiology for patient management	6
1.4	Clinical microbiology for pathogen surveillance	9
1.5	Whole-genome sequencing in clinical microbiology	10
1.5.1	Whole-genome sequencing for species identification and drug susceptibility testing	12
1.5.2	Whole-genome sequencing for pathogen surveillance	18
1.6	Clinical microbiology bioinformatics	20
1.6.1	Reference mapping	21
1.6.2	De Novo assembly	22
1.6.3	<i>K</i> -mer methods	23
1.6.4	Graph methods	24
1.6.4.1	De Bruijn Graphs	25
1.7	Overview of this work	27
2	Rapid antibiotic resistance predictions from high throughput genome sequence data	29
2.0.1	Publication note and acknowledgements	30
2.1	Introduction	31
2.2	<i>Staphylococcus</i> and <i>Mycobacterium</i> data sets represent global diversity	33
2.2.1	Overview of <i>Staphylococcus aureus</i> data sets	33
2.2.1.1	<i>Staphylococcus aureus</i> phylogeny	34
2.2.2	Overview of <i>M. tuberculosis</i> data sets	34
2.2.3	<i>M. tuberculosis</i> phylogeny	35

2.3	Species identification from WGS data using coloured de Bruijn graphs	38
2.3.1	Species probe generation	38
2.3.2	<i>Staphylococcus</i> species identification	39
2.3.3	<i>M. tuberculosis</i> species identification	40
2.4	Using population genome graphs for genotyping	43
2.4.1	Genotyping at mutations	45
2.4.1.1	Constructing variant probe sets	45
2.4.2	Genotyping at genes	48
2.5	Resistance prediction	50
2.5.1	Resistance calling at mutations	52
2.5.2	Resistance calling at genes	52
2.5.3	Mykrobe predictor <i>Staphylococcus</i> predictions match consensus phenotype	53
2.5.3.1	Detecting virulence elements in <i>S. aureus</i>	57
2.5.4	Mykrobe predictor <i>M. tuberculosis</i> resistance predictions match commercial assays	58
2.6	Power to detect minor populations	63
2.6.1	Heteroresistant <i>S. aureus</i>	65
2.6.2	Minor alleles increase power to distinguish XDR from MDR TB	67
2.6.2.1	Slow growth <i>rpoB</i> SNPs and limitations of the gold standard	68
2.6.3	Very low frequency <i>rpoB</i> variation in <i>M. tuberculosis</i>	69
2.7	Nanopore sequencing of <i>Staphylococcus</i>	71
2.7.1	Resistance calling from ONT data	71
2.8	Software performance and usability	73
2.9	Discussion	75
2.10	Extended methods	80
2.10.1	K-mer size	80
2.10.2	<i>S. aureus</i> data sets	80
2.10.2.1	Building the <i>Staphylococcus aureus</i> phylogeny	83
2.10.2.2	Resistance catalogues	84
2.10.3	<i>Mycobacterium tuberculosis</i> data sets	84
2.10.3.1	Building the <i>M. tuberculosis</i> phylogeny	84
2.10.3.2	Resistance catalogues	85
2.10.4	NTM species	85
2.10.5	Resistance calling at genes	86
2.10.6	Nanopore sequencing	86

3	Same-day diagnostic and surveillance data for <i>M. tuberculosis</i> via whole-genome sequencing of direct respiratory samples	89
3.0.1	Publication note and acknowledgements	90
3.1	Introduction	91
3.2	Results	94
3.2.1	DNA extraction protocol and evaluation of Illumina sequencing output	94
3.2.1.1	Contamination in direct and MGIT samples	95
3.2.1.2	Recovery of <i>M. tuberculosis</i> genome	97
3.2.2	Concordance of results from direct and MGIT samples	98
3.2.2.1	No evidence of higher diversity in direct samples	101
3.2.2.2	Detection of <i>M. tuberculosis</i> in culture-positive/-negative samples	101
3.2.2.3	Antibiotic resistance prediction	101
3.2.3	Sub-24 hour turnaround time with Illumina MiniSeq and ONT MinION	105
3.2.4	Sub-24 hour turnaround time with Illumina MiniSeq	106
3.2.5	Nanopore MinION sequencing of spiked samples	107
3.2.5.1	Modified Mykrobe predictor genotyping model for ONT data	108
3.2.5.2	Analysis of modified method for ONT MinION	110
3.2.5.3	12 hour turnaround time with ONT MinION	111
3.2.6	MinION turnaround estimates using empirical <i>M. tuberculosis</i> read proportion data	115
3.3	Discussion	119
3.4	Extended Methods	123
3.4.1	Sample selection and processing	123
3.4.2	DNA extraction and Illumina MiSeq sequencing	124
3.4.3	DNA extraction for ONT MinION and Illumina MiniSeq sequencing	125
3.4.4	MiniSeq sequencing	125
3.4.5	MinION Sequencing	125
3.4.6	Bioinformatic analysis of MinION data	127
4	Enabling rapid DNA search of all sequenced bacteria and viruses	129
4.0.1	Publication note and acknowledgements	131
4.1	Introduction	131

4.1.1	Microbial diversity is a challenge for building searchable indexes	132
4.2	Background	136
4.2.1	Probabilistic coloured de Bruijn graphs	136
4.2.2	Bloom filters	137
4.2.2.1	Bloom filter construction and querying	137
4.2.2.2	Choosing bloom filter parameters	138
4.2.3	Probabilistic de Bruijn graphs	139
4.2.4	Coloured de Bruijn Graphs	140
4.2.5	Coloured probabilistic de Bruijn Graphs	141
4.2.6	Sequence bloom trees	141
4.3	Results: A space-efficient representation of a probabilistic coloured de Bruijn graph	143
4.3.1	Coloured Bloom Graph construction and querying	144
4.3.2	Coloured Bloom Graph sequence search algorithm	146
4.3.2.1	Inexact search algorithm	147
4.3.2.2	Exact search algorithm	147
4.3.2.3	Variant search and genotyping algorithm	147
4.3.2.4	Controlling the false positive-rate of variant queries	148
4.3.3	Choosing CBG parameters	148
4.4	Results: Computational performance and validation	150
4.4.1	Simulated scaling to 1 million samples	150
4.4.2	Empirical scaling benchmark	151
4.4.3	CBG pseudo alignment scores correlate strongly with megaBLAST scores	154
4.4.4	Assessing gene genotyping accuracy	157
4.4.5	Assessing variant genotyping accuracy	158
4.5	Results: A searchable index of the entire microbial ENA/SRA	158
4.5.1	Taxonomic analysis of microbial ENA data sets	159
4.6	Application 1: Fast gene search in the microbial ENA	160
4.7	Application 2: Surveillance of antimicrobial resistance genes and variants	161
4.7.1	Rapid variant search and genotyping variants enables monitoring of allele frequency in a population	161
4.7.2	A survey of antimicrobial resistance genes in the SRA/ENA	164
4.8	Application 3: Measuring the host range of conjugative elements	168
4.8.1	Measuring plasmid host range	168
4.8.1.1	Prevalence of conjugative systems	170

4.8.1.2	MOB types may define conjugative elements' host range	174
4.9	Application 4: Surveillance of multi-drug-resistant plasmids	177
4.10	Discussion	180
4.10.1	Comparison to web search	182
4.10.2	Future improvements	183
4.10.3	Surveillance and applications	185
4.11	Extended methods	187
4.11.1	Simulated scaling comparison	187
4.11.2	Empirical scaling comparison SBTs and CBGs	188
4.11.3	Assessing genotyping accuracy	189
4.11.4	ENA download and processing	190
4.11.4.1	Approximate counting of unique <i>k</i> -mers	192
4.11.4.2	Species assignment of ENA data sets	192
4.11.5	Plasmid search and exclusion of contaminated samples	193
4.11.6	<i>Yersinia pestis</i> plasmid search	194
5	Conclusion	195
5.1	WGS as a diagnostic	196
5.2	A cloud-based analysis and data collection platform using de Bruijn graphs	198
A	Chapter 2 Appendix	202
A.1	Accessions for Simulation 1	202
A.2	Figures and tables	202
B	Chapter 3 Appendix	211
	Bibliography	213

List of Figures

1.1	A typical workflow for processing bacterial pathogens.	7
1.2	Timelines for sequencing-based analysis and culture-based drug susceptibility testing <i>Staphylococcus</i>	13
1.3	Timelines for sequencing-based analysis and culture-based DST <i>M. tuberculosis</i>	14
1.4	A toy example of a de Bruijn Graph	24
2.1	Phylogeny of <i>Staphylococcus aureus</i> with drug resistance indicated in concentric rings around the phylogenetic tree	33
2.2	Counts of each clonal complex in <i>Staphylococcus</i> training set St_{A1} and validation set St_{B1}	35
2.3	Phylogeny of <i>M. tuberculosis</i> with drug resistance indicated in concentric rings around the phylogenetic tree	36
2.4	Mykrobe predictor species predictions for <i>S. aureus</i>	40
2.5	Mykrobe predictor species predictions for <i>M. tuberculosis</i>	41
2.6	A cartoon of the genetic diversity in a bacterial species and two options for building a reference variation structure.	44
2.7	Within-sample frequency of resistant alleles in the training set	51
2.8	Comparison of Mykrobe predictor, Disc, and Phoenix antimicrobial resistance predictions	53
2.9	Comparison of Mykrobe predictor and SeqSphere antimicrobial resistance predictions	57

2.10	Comparison of Mykrobe predictor and KvarQ <i>M. tuberculosis</i> antimicrobial resistance predictions	59
2.11	Comparison of Mykrobe predictor and Hain <i>M. tuberculosis</i> antimicrobial resistance predictions	60
2.12	Power to detect low frequency features	64
2.13	Photograph of BSAC disc test showing hetero-resistant phenotype.	66
2.14	Percentage of true positive resistant calls in <i>M. tuberculosis</i> validation set due to minor alleles.	67
2.15	A pile-up showing a low frequency mutation at <i>rpoB</i> -Leu-449	69
2.16	Overview of data sets used for training/validation of species identification and resistance prediction.	82
3.1	DNA extracted from MGIT cultures and direct clinical samples.	94
3.2	Proportion of reads assigned to various taxonomic categories in each sample sorted by increasing total count of MTBC reads.	96
3.3	Recovery of <i>M. tuberculosis</i> genome in direct samples and robustness to contamination.	97
3.4	Histogram of genetic (SNP) differences between direct and paired MGIT samples	98
3.5	Direct/MGIT pairs placed on a phylogenetic tree	99
3.6	Coverage distribution across <i>M. bovis</i> BCG strain reference genome	107
3.7	Distribution of genotype confidence across 68950 SNPs genotyped on R9.4 MinION reads	109
3.8	Cumulative yield when sequencing culture negative sputum spiked with 15% BCG	110
3.9	Identity distribution for 1D reads in pure BCG sequencing run with MinION	111
3.10	Extrapolating from MinION R9.4 results to estimate performance for a realistic distribution of proportion of reads from <i>M. tuberculosis</i>	116
3.11	Timelines and cost for drug susceptibility testing from WGS of <i>M. tuberculosis</i>	118

LIST OF FIGURES

4.1	Sharing of genes in the <i>Escherichia coli</i> pan-genome	133
4.2	A schematic of a bloom filter	138
4.3	A schematic of a Sequence Bloom Tree	143
4.4	A schematic showing Coloured Bloom Graph construction and querying vs naïve approach	145
4.5	Simulated scaling of CBG, SBT, and mccortex to 1 million data sets with high/low proportion of sharing of k -mers between samples.	150
4.6	Build and query time for CBG and SBT	153
4.7	CBG scores vs megaBLAST scores	156
4.8	Counts of the most frequent bacterial genera in the SRA/ENA	160
4.9	Proportion of <i>M. tuberculosis</i> classified by genotypes as resistant (R), pan-susceptible (S), multi-drug-resistant (MDR),extensively-drugresistant (XDR) in microbial-CBG by date of first public availability	162
4.10	Proportion of <i>M. tuberculosis</i> resistance alleles in the microbial-CBG by date of first public availability	165
4.11	Counts of accessions with AMR genes across 5 genera	166
4.12	Proportion 10 AMR genes from CARD in the microbial-CBG across 5 genera split by year of upload to the ENA	167
4.13	Plasmid sequences found at least 5 times in more than one genus in the SRA/ENA	172
4.14	Distribution of MOB types among phyla	175
4.15	Source and locations of samples containing <i>Yersinia pestis</i> plasmid showing proportion sources as pie charts	179
A.1	Screenshot of the Mykrobe predictor <i>S. aureus</i> desktop app showing drugs split by resistant or susceptible prediction.	208
A.2	Screenshot of the Mykrobe predictor TB desktop app showing drugs split by first and second line (TB) alongside resistant or susceptible prediction.	209
A.3	Screenshot of the Mykrobe predictor <i>S. aureus</i> desktop app with evidence for each of the resistance calls.	210

B.1 Number of confident (genotype confidence > 1) heterozygous SNPs called
in paired vs direct/MGIT samples 212

CHAPTER 1

Introduction

1.1 Antimicrobial resistance

Infectious diseases caused by microbial pathogens are a persistent threat to global health. The need to improve testing, surveillance, and understanding of these diseases is particularly pressing now, because of the growing threat of a phenomenon which may cast us back into the dark ages of medicine: antimicrobial resistance. Modern medicine has relied on effective antibiotics—that is, drugs that treat serious infections—since the mass production of penicillin began in 1945 (Aminov 2010). However, our reliance on these ‘miracle’ drugs is at risk. Once-potent drugs are becoming increasingly ineffective and there are few new drugs arriving to take their place (Bartlett et al. 2013; Smith et al. 2015). Now, once treatable infectious diseases include nearly untreatable strains (Brown et al. 2016). People are dying as a result, with approximately 700,000 deaths from drug-resistant infections in 2014 (O’Neill et al. 2014).

1. INTRODUCTION

Antimicrobial resistance is present globally, and multi-drug-resistant (MDR) isolates have been seen across a wide range of bacterial species, including: *Mycobacterium tuberculosis*, *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Neisseria gonorrhoeae*, *Acinetobacter baumannii*, and *Pseudomonas aeruginosa* (Nathan et al. 2014). Resistance to carbapenems (e.g., meropenem) and polymyxins (e.g., colistin), two of classes of our “drugs of last resort”, have been observed in gram-negative bacteria, such as *Escherichia coli* and *Klebsiella pneumoniae* (Liu et al. 2016; Yong et al. 2009). Worryingly, convergence of multi-drug-resistance and hyper-virulence—traits that cause both aggressive and difficult to treat infections—has recently been observed in a highly transmissible strain of *Klebsiella pneumoniae*. The five patients infected did not respond to any antibiotic treatment and died over the course of several weeks (Gu et al. 2017).

The economic and societal consequences of widespread AMR would be enormous. It is estimated that by 2050, unless action is taken, 10 million people may die annually as a result of antimicrobial resistance, more than from cancer (O’Neill et al. 2016). The resulting cumulative global cost to economic output could exceed 100 trillion USD (O’Neill et al. 2016).

Antimicrobial resistance has been noted as an imminent threat and a strategic priority by many countries and global institutions (McArthur et al. 2017; World Health Organization 2012). The causes of spreading antibiotic resistance are complex and the strategies to combat it are multi-faceted, including: raising awareness through improvements in public health policy and communication, in particular amongst prescribers; improved hygiene to prevent spread; reduction of the use of

antimicrobials in agriculture; improved surveillance and monitoring; development of new rapid diagnostics for improved treatment decisions; development and use of vaccines; improved training of AMR stewards; increased funding for basic research; accelerating the drug discovery process for novel antibiotics; and greater global collaboration to address the problem (Livermore et al. 2013; O'Neill et al. 2016; Spellberg et al. 2013; World Health Organization 2012). Most of these topics, although vital to addressing AMR, are outside the scope of this thesis. Understanding and utilising the underlying genetic mechanisms of resistance can aid the surveillance of antimicrobial resistance, by identifying and tracking the relevant genes and variants, and is a potential avenue toward rapid diagnostic tests through molecular testing or DNA sequencing, and this is what my thesis will focus on.

The process of testing bacteria for susceptibility has been largely unchanged in decades, often using processes which are over 100 years old. The vast majority of antimicrobial prescriptions are made without using a diagnostic tool and “empiric” therapies, which are prescribed based on intuition and professional judgement. This is not due to irrationality of clinicians—there simply are not sufficient accurate, rapid, and affordable tests. It is often cheaper, faster, and less risky to prescribe broad-spectrum antimicrobials ‘just in case’.

Bacteria must be cultured for 24 hours or more to confirm the type of infection and the drugs to which they are susceptible—far too long for many clinical decisions. Thus, many patients are given antimicrobials that they do not need, and others are given antimicrobials that are ineffective against the infecting pathogen, resulting in further delays and potentially additional ineffective empiric therapies. As a result,

the World Health Organisation (WHO), the UK Longitude Prize, and others have called for new rapid point-of-care diagnostics. Whole-genome sequencing, a new technology which brings unprecedented resolutions into the genetics of bacteria offers a potential route to this goal, as well as to improvements in surveillance of the genetic mechanisms of resistance, both of which I will discuss further below.

1.2 Mechanisms of resistance

It is important to note that, although exacerbated by human activity, antimicrobial resistance is not a human creation—it is a natural adaptation, evolved for defence against other organisms in the bacterias' niches (Davies et al. 2010). As a result, there are sources of antibiotic resistance mechanisms that have existed long before the prevalent use of antimicrobials. For instance, a bacterial penicillinase was identified before penicillin had been introduced as a therapeutic (Abraham et al. 1940), analysis of β -lactamase genes show that they have existed for millions of years (Barlow et al. 2002), and genes for β -lactam, tetracycline, and glycopeptide resistances have been isolated from 30,000 year old DNA (D'Costa et al. 2011).

Resistance to antibiotics is driven by a number of mechanisms: antibiotic permeability, where the antibiotics are excluded by an impermeable barrier; alteration of target molecules, where mutations in a protein inhibit the binding of an antibiotic to its target or enhance binding to off-target elements; enzymatic degradation of the antibiotics, where the bacteria's cell machinery dismantle the antibiotic molecule; and efflux of antimicrobials from the cell, where the antimicrobials are rejected from

the cell via a chemical pump (Wright 2003). Our extensive use of antimicrobials in medicine and agriculture has accelerated the appearance and spread of resistance (Davies et al. 2010).

In large part, this escalation is made possible due to some remarkable genetic capacities that bacteria have. Although bacteria can acquire resistance from de novo mutations, which are then passed on to offspring by descent, many mechanisms are acquired by horizontal gene transfer — the acquisition of DNA from another unrelated bacterium — via mobile genetic elements such as plasmids or transposons that can rapidly be exchanged between bacteria (Davies et al. 2010). Importantly, mobile genetic elements, which can carry genes that encode for antimicrobial resistance, can be exchanged between members of different species of bacteria. This phenomenon of horizontal gene transfer has profound implications for the surveillance of antimicrobial resistance, as it means that many antimicrobial resistance mechanisms are not limited to the clade (a group of lineal descendants from a common ancestor) in which they are first observed.

Mobile genetic elements exist in various forms and can be seen in the ‘core genome’ (the non-conjugative chromosome(s) that are conserved within a species) as transposons or integrative conjugative elements, or in the ‘accessory genome’ as plasmids or phages (Brown-Jaque et al. 2015; Stokes et al. 2011; Touchon et al. 2014). Some transposons can jump from the chromosome to plasmids, and back again (Stokes et al. 2011; Touchon et al. 2014). These complex genetic mechanisms pose significant challenges to analysis and surveillance.

1.3 Clinical microbiology for patient management

One of the key goals of clinical microbiology is rapid characterisation of isolates in order to direct the management of individual infected patients. To oversimplify, clinicians need to know (as quickly as possible):

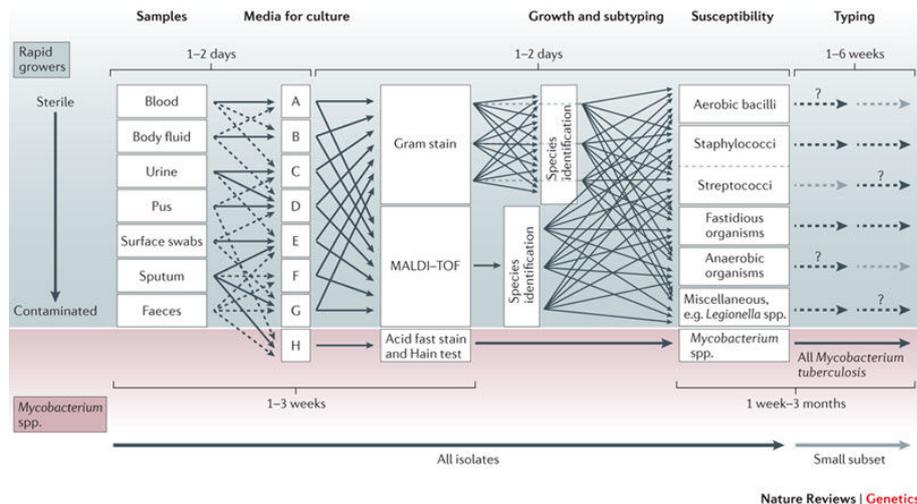
- What organism is causing my patient's infection?
- What drugs can be used to treat it and which are the most appropriate?
- How is the pathogen related to other similar infections?

As a result, species identification, typing, antimicrobial susceptibility testing, and subsequent recording and auditing of the results are central to both the treatment of serious bacterial infections and the ongoing monitoring of antimicrobial resistance. To answer these questions there are a diverse and fragmented set of assays (Didelot et al. 2012). The detailed methodologies required to characterise a pathogen are often species- and source-specific and require highly knowledgeable and skilled technicians to implement. These processes, many of which are up to a hundred years old, are often laborious, complex, and slow (Didelot et al. 2012; Köser et al. 2012b).

In bacteria, these tests principally involve: sample collection (e.g., blood/faeces), culturing in media, species identification, susceptibility testing, and epidemiological typing (Figure 1.1). The culture step can be complex and depends on many factors, including the sample origin, the likely pathogen species, and growth time.

1.3. Clinical microbiology for patient management

Figure 1.1: A typical workflow for processing bacterial pathogens. These tests generally involve species-specific processes for culture, species identification, drug susceptibility testing, and strain typing. Culture depends on sample type. For instance, samples that are likely to be contaminated with other flora may require selective media to encourage the growth of the suspected pathogen. The sample's species is then identified via gram-staining and other biochemical testing or via MALDI-TOF. Susceptibility testing, typically via growth inhibition, is then performed, often with species-specific considerations. Strain typing is not routinely performed in the majority of cases and is performed in an ad-hoc manner to follow up a subset of isolates for outbreak investigation. From Didelot et al. 2012. Reproduced with permission.



Species identification can be performed with a broad range of tests, which determine expressed phenotypes of the organism, including: colony morphology; growth (e.g., growth only in certain contexts); gram-staining, which differentiates groups of pathogens based on constituents of their cell-wall; rapid biochemical reactions, such as polymerase chain reaction (PCR); serology, where antibodies are used to detect the presence of particular proteins; and mass spectrometry (e.g. MALDI-TOF), which detects patterns of molecules uniquely present in the different species. These

1. INTRODUCTION

techniques are of variable cost, speed, resolution, and complexity, but all provide limited additional information about the sample, such as antibiotic susceptibility or virulence.

From the perspective of patient management, determining the antimicrobial resistance profile, or “antibiogram”, is of utmost importance. Falsely recording an organism as susceptible to an antibiotic, a false negative (“very major error” in the antimicrobial resistance susceptibility testing nomenclature (US-FDA et al. 2007)), presents the serious risk of the patient being given ineffective treatment. The reverse, a false positive resistance call (or “major error”), risks the patient being given a non-optimal treatment, which could be less effective or have greater side effects.

The vast majority of antimicrobial susceptibility tests are based on growth inhibition of the bacteria when exposed to the varying concentrations of the desired drug, determining the “minimum inhibitory concentration” (MIC). Typically, “breakpoints” (MIC thresholds) are then used to define a binary (resistant or susceptible) result (*EUCAST* 2017). Testing based on growth inhibition *in vitro* has advantages because it is, in theory, independent of mechanism and gives information on both resistance and susceptibility to drugs. However, even with ‘gold standard’ susceptibility testing, the *in vitro* MIC of an isolate is likely to differ from the clinical, *in vivo*, MIC. Molecular tests also exist, for example, detection of *mecA* to determine methicillin resistance in *S. aureus* (Bode et al. 2012) and Hain line probe assays to detect single-nucleotide polymorphisms (SNPs) predicting resistance in *M. tuberculosis* (Abreu Maschmann et al. 2013; Chryssanthou et al. 2012; Miotto et al. 2012; Rodwell et al. 2014). However, they are often species- and mechanism-

specific and cannot be updated quickly as new mechanisms become understood.

1.4 Clinical microbiology for pathogen surveillance

Tracing the source of disease outbreaks and their mode of transmission is vital to prevent further spread. Key questions include:

- What is the prevalence of antimicrobial resistance mechanisms in a population?
- Are the mechanisms changing over time?
- Which of the mechanisms are being transmitted, and to where?
- Have similar strains been seen before? If so, where and when? What are the imminent risks to public health?

Pathogen surveillance and outbreak investigations are typically supported by reference laboratory genotyping (Köser et al. [2012b](#)). Turnaround times of at least a week, along with additional delays due to batching, mean that timely evidence of the introduction and transmission of a new strain is difficult to obtain (Köser et al. [2012b](#)). The limited resolution of current typing tools (e.g., spa-typing and pulsed-field gel electrophoresis) mean many samples are likely misclassified and many outbreaks are missed (Didelot et al. [2012](#); Schürch et al. [2010](#)). In addition, only a

few laboratories globally perform routine typing, in most cases isolates are chosen extemporaneously, and the data are often siloed making national or global collaboration on surveillance challenging.

1.5 Whole-genome sequencing in clinical microbiology

Whilst current microbiology assays can answer some of the vital questions above, they are often fragmented, non-digital and slow. Individual, often species-dependent tests, are required to answer each biological question. However, in theory, the genome of a bacterium contains a tremendous amount of information, which could be used in order to manage patients and track outbreaks, although how to interpret these data is not always known. Once decoded, the genome of a pathogen, in principle, reveals a host of clinically relevant information, including: species, resistance to different drugs, relatedness to pathogens isolated from other patients or the environment, and ability to cause illness or virulence. Potential for antimicrobial resistance is mediated by key genes or mutations, and as such, is encoded in the DNA of the bacteria.

Whole-genome sequencing (WGS), a process that reads the entire genome of a sample, has recently become fast and affordable and has advantages over the 100-year-old phenotypic tests: it is versatile, accurate, and digital. WGS gives unprecedented resolution to the genomics of pathogens and the relationship between

them. There are now technologies which provide high-throughput, fast, affordable, and portable WGS (Bradley et al. 2015; Didelot et al. 2012; Köser et al. 2012b; Pankhurst et al. 2016; Quick et al. 2016). These include: sequencing by synthesis, single-molecule real-time sequencing, and nanopore sequencing, amongst others. The two technologies discussed in this thesis are: sequencing-by-synthesis and nanopore sequencing. Sequencing-by-synthesis (patented by Illumina) fragments the genome into short ‘reads’ (fragments of DNA sequence) 50-600 bp long which are then read with high accuracy, high throughput and low cost. Nanopore sequencing (Oxford Nanopore Technologies) reads longer DNA fragments (1kbp-1Mbp) by measuring voltage change as the ‘read’ is pulled through a biological ‘pore’, but with lower accuracy and throughput.

These technologies raise the potential of replacing or augmenting the plethora of current tests with a single rapid assay, normally following culture, which can then be cheaply stored, shared, compared, and interrogated in a multitude of different ways using software. Since WGS data can be stored and interrogated digitally, it is well suited to be used in a collaborative system for surveillance systems to track disease trends and enable large-scale bacterial genome-wide association studies.

As a result, WGS is now commonplace in pathogen genomic research. The combination of rapidly declining cost (now <£40/bacterial WGS data set) and speed (now <24 hours for a high-throughput run) means that WGS is likely to soon become part of routine clinical microbiology (Köser et al. 2012b). WGS has already found many applications, including: elucidating infectious disease outbreaks (Gardy et al. 2011; Grad et al. 2012), uncovering novel antimicrobial resistance

determinants (Billal et al. 2011; Walker et al. 2015), and investigating virulence (Harris et al. 2010), amongst others (Köser et al. 2012b; Olsen et al. 2012). A WGS-based workflow is currently being piloted by Public Health England’s Tuberculosis Section in parallel to its existing workflow. However, there are still many challenges to overcome before it becomes part of routine care or surveillance.

As a result of the rapidly declining cost of sequencing, the bottleneck for implementation is shifting toward software and analysis of these large data sets (Didelot et al. 2012; Green et al. 2011; Olsen et al. 2012; Pop et al. 2008). Bacterial genomes are comparatively small when compared with human genomes, but their genomes are highly complex and plastic, making interpretation demanding. Although analysis challenges are varied, in this thesis I will focus on those surrounding the computational analysis of WGS data, in particular addressing the speed, accessibility, and scalability of these analyses.

1.5.1 Whole-genome sequencing for species identification and drug susceptibility testing

Making better use of the limited antibiotics we have is of critical importance. As a result, a key step in the process of using WGS in routine clinical practice is making antimicrobial resistance predictions from these data easy, fast, accurate, and quickly updateable when new information becomes available. Until a clinician is aware of the species and antibiotic resistance profile of the pathogen infecting their patient, they must resort to an empiric therapy of broad spectrum antibiotics.

1.5. Whole-genome sequencing in clinical microbiology

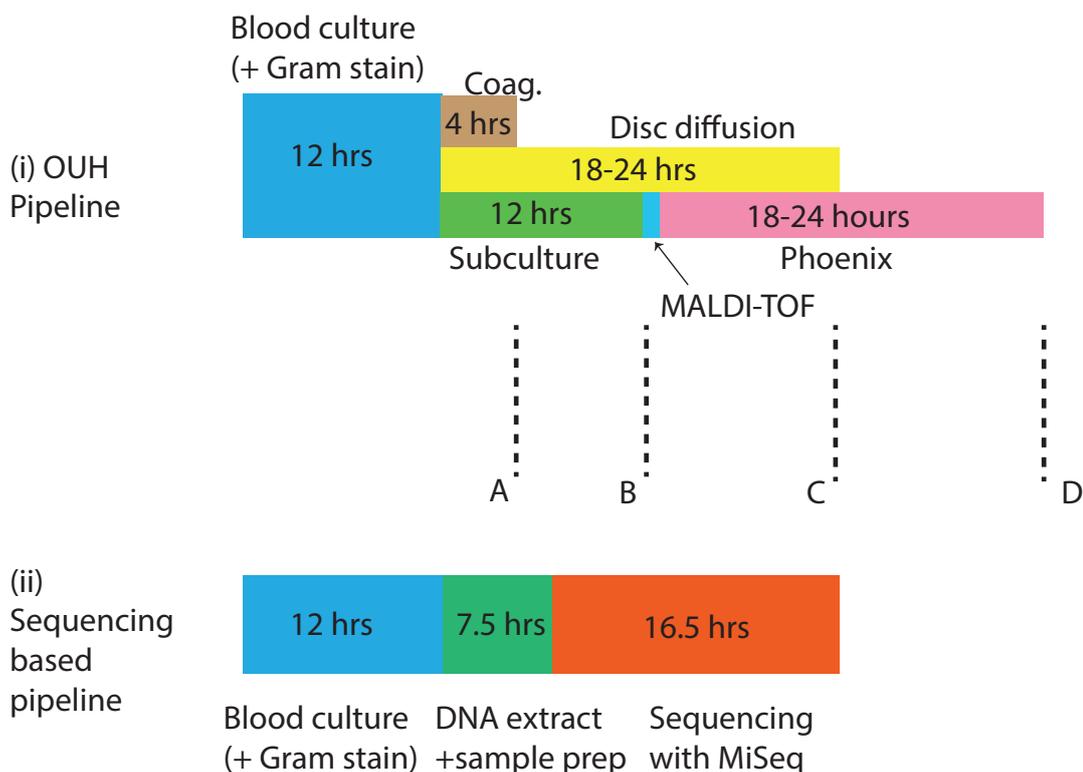


Figure 1.2: Timelines for sequencing-based analysis and culture-based drug susceptibility testing (DST) *Staphylococcus*: Both culture-based (i) and sequencing-based (ii) options involve 12h of blood culture. After this, the culture-based approach (at Oxford University Hospitals clinical laboratory) follows with a direct coagulase test (Coag.) that provides a presumptive species identification at 4h (marked ‘A’). Concurrently, blood culture is subcultured to blood agar, and MALDI-TOF confirms the species at 12h (‘B’). A disc diffusion test for five antimicrobials (including methicillin) is performed directly from a positive blood culture providing first-line susceptibility information 18–24h later (‘C’), assuming an acceptable inoculum. Finally, post-subculture samples undergo extended susceptibility testing by automated broth microdilution (brandname ‘Phoenix’), giving final results after another 18–24h (‘D’). For the sequencing-based workflow (ii), the DNA extraction plus sample preparation takes 7.5h because samples are from blood culture, not colony isolates. With the Illumina MiSeq v3 reagents, a 16.5h run is possible (giving paired 75 bp reads, adequate for this purpose), giving full susceptibility results at the same time as direct disc tests provide results for five drugs. From Bradley et al. 2015.

1. INTRODUCTION

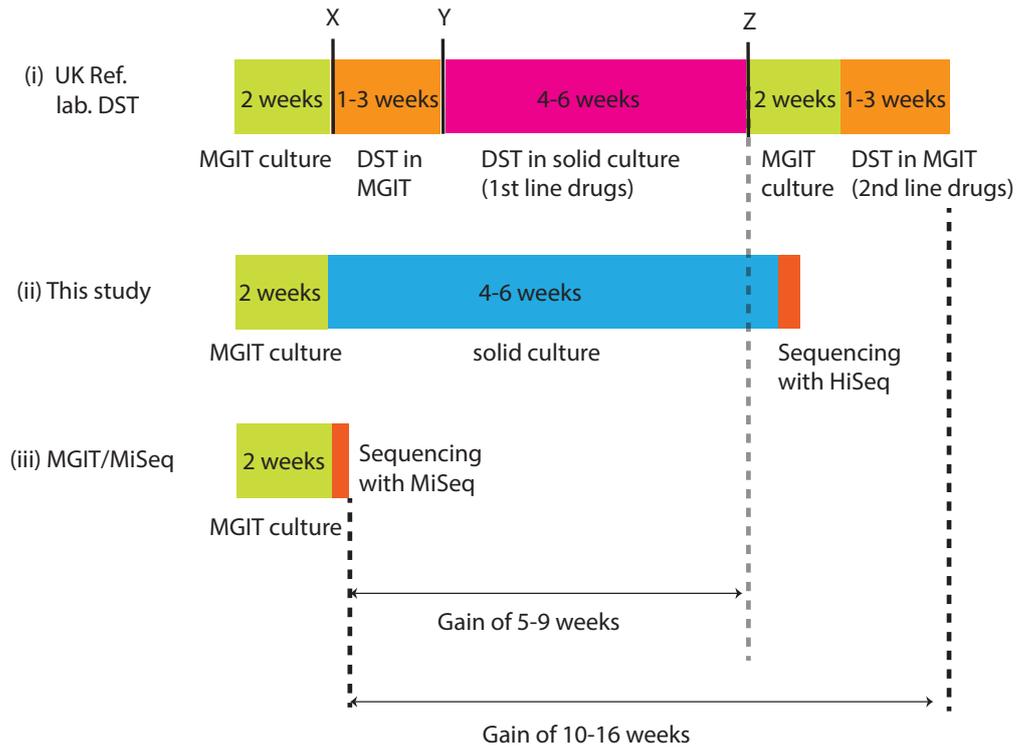


Figure 1.3: Timelines for sequencing-based analysis and culture-based DST *M. tuberculosis*: i) The culture-based process (in a typical UK reference laboratory) starts with two weeks of mycobacterial growth indicator tube (MGIT) culture, followed by a species identification test ('X'). If the species belongs to the MTBC, then DST is run in MGIT, and at decision point 'Y'. If, at this point, the sample tests susceptible to all first-line drugs, no further testing is done. MGIT DST is repeated for pyrazinamide if the first test revealed resistance to this drug. If there is resistance to any other drug, then solid culture DST is performed. If these tests show there is resistance to rifampicin then another round of MGIT culture followed by MGIT DST is done for second-line drugs. For sequencing-based approaches I show timelines for an HiSeq workflow (ii) and a MiSeq alternative (iii), which would reduce time-to-results to just over 2 weeks. "This study" refers to Bradley et al. 2015, which is discussed in more detail in chapter 2.

This risks that the empiric therapy is inadequate, providing insufficient treatment, or increasing exposure of the pathogen to antibiotics and the patient to unnecessary side effects.

There are two primary situations where a WGS-based drug susceptibility testing would be of immediate benefit. Firstly, when the microbe is difficult to grow, phenotypic drug susceptibility testing can be slow and demanding (e.g., *M. tuberculosis*). Secondly, WGS-based drug susceptibility testing is of benefit whenever the mechanism of resistance, or potential for resistance, informs clinical decisions. For example, *S. aureus* can be induced to be clindamycin resistant, if first treated with erythromycin and if its genome encodes *erm* genes (Lewis et al. 2005). Therefore, the presence of *erm* would lead to a decision to not treat with clindamycin, if the patient was already exposed to erythromycin, and perhaps not at all. Another example is when a resistance gene can exist on either a plasmid or on the chromosomes; plasmid-mediated resistance can affect the decision to implement infection control measures. There are also many other reasons why sequencing of clinical samples may become routine, such as typing for outbreaks and surveillance. If this is the case, and since the data would already be available, WGS for drug susceptibility testing may be used in other contexts faster than would otherwise be expected.

At present, phenotyping tests take at least 1–2 days to complete for rapidly growing bacteria, such as *S. aureus* (Gordon et al. 2014), and they can take weeks in slow-growing bacteria, such as *M. tuberculosis* (Lee et al. 2017). A WGS-based test could be done in a similar time frame as phenotyping tests for fast growing organisms, such as *S. aureus* (with sufficiently fast bioinformatic analysis) (see Figure 1.2),

but has the capacity to dramatically decrease the time taken for antimicrobial resistance prediction for slow growing organisms, like *M. tuberculosis* (Figure 1.3), as well as providing other useful information such as species and relatedness (Didelot et al. 2012). Sequencing-by-synthesis runs can take between 16–48 hours with Illumina technologies and a variable length of time with ‘random access’ technologies like Oxford Nanopore’s MinION (where the user has control over the sequencing time). Typically, this is done after a short culture step to improve inoculum size. However, there are “direct-from-sample” extraction methods in development, which would allow turnaround times of less than 24 hours (Brown et al. 2015). I discuss this in more detail in chapter 3, where I apply *k*-mer based drug susceptibility testing from WGS to *M. tuberculosis* isolates sequenced directly from sputum.

The key challenge in using WGS to predict antimicrobial resistance is understanding the genotype to phenotype correspondence. We need a model that can use genotypic information to accurately and rapidly predict phenotype, and which will need to be updated as novel resistance mechanisms arise. Perhaps surprisingly, for many species/drug combinations we already have a sufficiently comprehensive understanding of the genetic mechanisms of resistance in order to infer antibiograms with high accuracy. For example, Gordon et al. 2014 achieved a 97% (95%–97%)¹ sensitivity and 99% (99%–100%) specificity predicting resistance to a panel of 12 drugs in *S. aureus*, Walker et al. 2015 achieved 92.3% (90.7%–93.7%) sensitivity and 98.4% specificity (98.1%–98.7%) for *M. tuberculosis* resistance prediction, Stoesser et al. 2013 achieved 96% (0.94%–0.98%) sensitivity and 97% (0.95%–0.98%) speci-

¹95% confidence intervals

ficity in *Escherichia coli* and *Klebsiella pneumoniae*. There are also ongoing efforts to improve genotype to phenotype models using genome-wide association studies (Earle et al. 2016) and machine learning (Davis et al. 2016).

In addition to an accurate model, we also need reliable, fast, and easy-to-use bioinformatic tools to infer the presence/absence of the relevant genetic features. A lack of user-friendly and automated software is a significant barrier to the use of WGS in routine settings (Fricke et al. 2014; Köser et al. 2012b; Török et al. 2012). Several methods have been proposed for this purpose but broadly speaking they follow two general strategies: 1) reference-based read mapping and 2) de novo assembly. Most of these methods perform adequately.

However, assembly followed by BLAST for genes and variants (Gordon et al. 2014; Leopold et al. 2014) is computationally intensive and assumes the sequence is derived from a clonal sample—potentially ignoring low frequency strains which may have important clinical information. Mapping to a reference genome to detect point mutations of genes (Kohl et al. 2014; Köser et al. 2013) assumes that data comes from a single haploid genome (Pop 2009), and so it is ill-suited for mixed samples, and mapping to a single reference results in error rates that depend on genetic distance of the sample from the reference (Bertels et al. 2014).

In addition, the currently available tools that use the two strategies above require high levels of expertise to install and run, and typically assume a contamination-free isolate (Inouye et al. 2014; Leopold et al. 2014; Steiner et al. 2014). However, contamination-free data are difficult to obtain with bacterial WGS and will become increasingly challenging as we move to culture-free sequencing for diagnostics. As a

result, there is still work to be done to enable fast and accurate resistance prediction in an easy-to-use software, which is also flexible and takes into account the needs of clinicians and the complexities of the data for various species. I describe my work on improving antimicrobial resistance prediction software from WGS in chapter 2.

1.5.2 Whole-genome sequencing for pathogen surveillance

Traditional typing techniques have limited resolution as they typically only interrogate small regions of the genome. In contrast, WGS provides the ultimate resolution of the relationships between pathogens. For instance, strains that look identical under virulence gene content, serotyping, multilocus sequence typing, rep-PCR, pulsed-field gel electrophoresis, optical mapping, and antimicrobial susceptibility testing can be resolved using WGS, helping to elucidate the epidemiology of an outbreak (Grad et al. [2012](#)). There are many other examples where WGS has provided vital insight (Dettman et al. [2013](#); Gardy et al. [2011](#); Harris et al. [2010](#); Holden et al. [2013](#); Snitkin et al. [2012](#)).

Moving to WGS-based diagnostics opens unprecedented opportunities for data acquisition, surveillance, and discovery of antimicrobial resistance determinants. The rise of cloud computing, the declining cost of sequencing, and the increasing utility of WGS for diagnostics mean we have the capacity to collect, analyse, and learn from potentially millions of bacterial WGS data sets. This will feed back into the WGS-based drug susceptibility testing which, in order to match the sensitivity of existing phenotypic-based tests, will require comprehensive databases of genetic

data linked with clinically relevant metadata. However, there are many technical challenges to overcome in order to make these databases accessible, scalable, and rapidly updating.

Although databases of bacterial WGS data do exist, for example the European Nucleotide Archive with close to 500,000 bacterial data sets, these are almost entirely inaccessible to sequence search and have limited structured metadata. Consequently, it is challenging to use them to rapidly compare any isolate with the corpus of existing data, search for genes or mobile genetic elements of interest, or use them in applications such as surveillance or genome-wide association studies.

Surveillance databases such as the National Antimicrobial Resistance Monitoring System for Enteric Bacteria (NARMS) have more extensive metadata than the SRA/ENA archives, however it comprises a limited number of data sets and the data are siloed and difficult to link to the genomic archives. Being able to utilise and collaborate on large collections of data in order to gain insight into the genomics of phenotypic traits is a vital component to make best use of sequencing in the clinic, for public health, and for science.

So far, the applications of WGS to surveillance of pathogens, have primarily been retrospective analyses of outbreaks, usually consisting of less than a thousand isolates (Dettman et al. 2013; Gardy et al. 2011; Harris et al. 2010; Holden et al. 2013; Snitkin et al. 2012). It seems that cheap WGS may be bringing the Malthusian principle to genomics: our capacity to generate data may be outstripping our ability to interpret it at equivalent scales. There is considerable work required to build a genomic surveillance database and link WGS data with extensive clinical metadata.

However, the potential benefits of such a system could be tremendous. Enabling alerts when, for example, a plasmid is first seen in a new country or in a new species would have global significance. However, despite the large quantity of data we have already collected we are far from enabling this type of alert without considerable focused effort. In chapter 4, I describe a novel method to index and search millions of data sets with the goal of opening progressively larger data sets for analysis.

1.6 Clinical microbiology bioinformatics

A major challenge in making the applications described above a reality is to develop fast, scalable data structures and algorithms that can handle the exabytes of data expected to be generated in the coming years. Thanks to their comparably small genome, bacteria can be sequenced quickly and cheaply and, as a result, are relatively accessible to interrogation by WGS. However, their diversity and complexity means that analysis and interpretation are challenging. Contamination is a persistent problem with microbial WGS data, from cohabiting microbes, the host, or the technician performing the sequencing. It is also possible for samples to have mixed populations of the same species, involving two or more strains of the microbe simultaneously (Tarashi et al. 2017). This is a result of the fact that, unless performing single-cell sequencing (which I will not consider in this thesis), the sequencing data represent genomes of cell populations, which are not necessarily identical.

Perhaps due to our anthropocentric bias, the majority of bioinformatics tools for analysis of resequencing data (WGS from a species with a reference genome), which

were designed with human genomes in mind, have been based on the assumption that any given individual genome can be treated as a small perturbation away from a reference genome. This is mostly true for human genomes, where the diversity is low, but in bacteria levels of diversity can be much higher, genomic rearrangements are more common, and problems are caused by the presence/absence of variation of regions containing entire genes, which leads to the concept of a pan-genome—breaking the small perturbation assumption. In order to make our analysis consistent with the biology of bacteria this complexity must be considered.

1.6.1 Reference mapping

A standard approach to the analysis of bacterial WGS is to map and align sequencing reads to a reference genome—an exemplar of the population from which data can be compared (Köser et al. 2012a; Li et al. 2010a; Snitkin et al. 2012; Wyres et al. 2014). This approach is primarily suited to identifying positions where the sample contains simple variant sequences: discovery and genotyping of SNPs; and short insertions or deletions. The popularity of the mapping approach is likely a result of the primary current use case of WGS: the detection of pathogen transmission events and the investigation of outbreaks. Here, we are primarily interested in the identification of discriminatory simple variants in the core genome in order to rule patients into, or out of, an outbreak (Harris et al. 2013). Mapping is well established as an accurate method of doing so (Gardy et al. 2011; Grad et al. 2012; Harris et al. 2013; Li et al. 2010a; Lunter et al. 2011). However, the popularity of

reference mapping may also be driven by the fact that the relevant tools are popular and well-validated due to their application in other areas of genomics.

This approach has limitations though, in particular when applied to bacterial data sets, as horizontal transfer events mean that the sample may be highly diverged from the reference (Beiko et al. 2005; Ge et al. 2005). The breaking of the small perpetuation assumption and reference bias from mapping to a single representative genome is also a problem outside of microbial genomics (Dilthey et al. 2015; Novak et al. 2017).

It has led to the development of “population reference graphs” (Novak et al. 2017), which include multiple samples within the reference, represented as a graph. These have been shown to improve the ability to analyse variation in areas of high sequence diversity, such as the major histocompatibility complex region (Dilthey et al. 2015), and complex regions in malaria (Maciuca et al. 2016), amongst others.

1.6.2 De Novo assembly

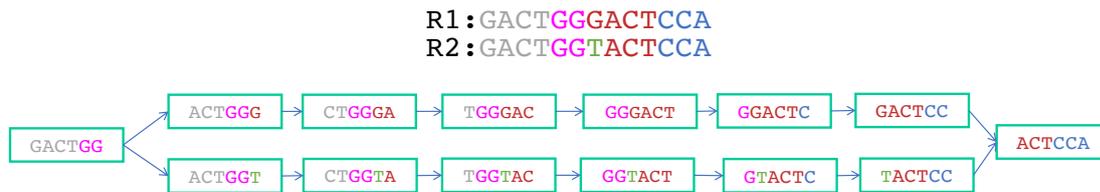
A second approach is *de novo* assembly. De novo assembly involves merging the sequencing reads in order to reconstruct the original sequence, which does not require a reference. However, this approach also has limitations. First, assembly tools almost exclusively focus on consensus assembly, treating the sequence as if it is clonal and removing any sequence which is at a low frequency. This is a reasonable assumption in human data, where normally it can be expected that all the cells have the same genome (with the exception of cancer genomes). However, in

bacteria, these low frequency genotypes can have important clinical consequences. For example, a study of fluoroquinolone resistance found that only 50% of resistant samples had a single fixed ($> 95\%$ frequency) mutation, while the others had evidence for multiple alleles (Eilertson et al. 2014). Deep sequencing to detect low-frequency alleles improved their power to detect resistance by 19% (Eilertson et al. 2014). Also, bacterial samples are intrinsically a mixture of multiple genomes. As a result, methods will need to take into account in-host diversity, minor populations, accessory elements, and metagenomics in order to give an accurate representation of a pathogen’s potential for resistance, virulence, or transmission—information which can be lost in consensus assembly.

1.6.3 *K*-mer methods

K-mer based methods involve fragmenting the sequencing reads into smaller sequences of fixed length (k) and performing analysis on these short sequences. If the k -mers overlap by $k - 1$ bases, then they form a de Bruijn graph, which I will discuss below. Although less popular than mapping or de novo assembly approaches, they can be useful for some applications, in particular where computational efficiency is paramount. For example, **kraken** (Wood et al. 2014), a popular metagenomic sequence classifier, shreds reads into constituent k -mers, which it can rapidly classify, and **KvarQ** (Steiner et al. 2014) uses k -mers to identify antimicrobial resistance variants in *M. tuberculosis*.

Figure 1.4: A toy example of a de Bruijn Graph. Here, we show the resulting 6-mer de Bruijn graph from 2 reads (R1,R2). Single nucleotide differences will cause the graph to form “bubbles”.



1.6.4 Graph methods

In this thesis, I will examine a strategy based on genome graphs. Genome graphs offer both computational efficiency and the flexibility to represent diversity in collections of genomes. Genome graphs involve the representation of sequencing data as graphs, nodes of DNA strings, connected by edges. These nodes can be reads in the case of *overlap graphs* (Kececioglu et al. 1995), arbitrary sequence in the case of *sequence graphs* (Novak et al. 2017), or k -mers in the case of *de Bruijn graphs* (Chaisson et al. 2008; Zerbino et al. 2008), which is the data structure underlying the methods discussed in future chapters. Since I will focus on de Bruijn graphs, it could also be considered a k -mer centric approach. I will demonstrate that de Bruijn graphs can be used to efficiently and accurately achieve many of the goals of our analysis, including species identification, genotyping for AMR prediction, and indexing very large collections of sequence data for search.

1.6.4.1 De Bruijn Graphs

De Bruijn graphs are the fundamental data structure underlying much popular assembly software (Butler et al. 2008; Chaisson et al. 2008; Chikhi et al. 2013b; Simpson et al. 2009; Zerbino et al. 2008). Their use in bioinformatics was pioneered by Idury, Pevzner, and Waterman (Idury et al. 1995; Pevzner et al. 2001) using graph theory originally developed by de Bruijn in the context of combinatorial mathematics (Bruijn 1946).

De Bruijn graphs are straightforward to construct. DNA sequence reads are broken into chains of consecutive k -mers which overlap by $k-1$ bases. Edges link k -mers that are neighbours in the underlying read data, see Figure 1.4. Construction requires looking up exact matches between k -mers and, as a result, can be performed in linear time using a hash table which allows for constant time k -mer lookups. Due to their simplicity and performance, de Bruijn graphs have become a popular tool in bioinformatics—in particular for genome assembly (Butler et al. 2008; Chaisson et al. 2008; Chikhi et al. 2013b; Simpson et al. 2009; Zerbino et al. 2008).

Although computationally efficient, de Bruijn graphs can require a lot of memory. Since there is a node for every unique k -mer in the genome, de Bruijn graphs of large eukaryotic genomes can have billions of nodes, requiring hundreds of gigabytes of memory (Simpson et al. 2009). In addition, each sequencing error will create unique, erroneous k -mers—increasing the memory burden.

Coloured de Bruijn graphs extend classical de Bruijn graphs by ‘colouring’ nodes in the graph by the samples in which they are seen (Iqbal et al. 2012). It is inter-

esting to note that an implicit coloured de Bruijn graph, where edges are not stored explicitly, is equivalent to an inverted index—where k -mers are mapped to the data sets which contain them. Inverted indexes are commonly used in text search (Brin et al. 2012; Goodwin et al. 2017), a fact I will explore more in Chapter 4. Coloured de Bruijn graphs suffer from similar memory challenges as de Bruijn graphs, since the union of unique nodes in all colours needs to be stored. If the k -mers are shared amongst many of the colours in the graph, then the memory burden is similar to that of a single-colour de Bruijn graph. However, if the union of unique k -mers is much larger than the unique k -mers in any colour (e.g., in a bacterial pan-genome), then the memory required can be significantly greater.

Several approaches have been suggested to manage the high memory requirements, including: error cleaning, distributed computing, and compressed or probabilistic encodings. Error cleaning of the erroneous can significantly reduce the memory requirement of the graph (Iqbal et al. 2012; Li et al. 2010a). Distributed computing can spread the memory load across multiple machines (Simpson et al. 2009). And novel encodings such as succinct data structures (Belk et al. 2016), sparse bit vectors (Conway et al. 2011), bloom filters, or compression by segregation patterns, (Almodaresi et al. 2017) can decrease the memory requirement per k -mer (Chikhi et al. 2013b).

1.7 Overview of this work

In this thesis, I describe novel data structures and algorithms, and explore applications of graphs, specifically de Bruijn graphs and coloured de Bruijn graphs, to some of the computational challenges discussed in the previous sections. In chapter 2, I use de Bruijn graphs to rapidly determine species and antimicrobial resistance from short read WGS of cultured clinical isolates. I present a tool, “Mykrobe predictor”, that can be used to rapidly identify species and antimicrobial resistance from WGS data sets, for *M. tuberculosis* and *S. aureus*. It is easy to use, extremely efficient, and can be augmented with new antimicrobial resistance catalogues easily (Bradley et al. 2015).

In chapter 3, I explore the additional issues encountered applying this approach to sequencing of direct clinical samples of *M. tuberculosis* with the goal of reducing the time required for the end-to-end test (Votintseva et al. 2017). This includes applications to data from noisy, long nanopore reads.

In chapter 4, I explore the challenges in scaling coloured de Bruijn graphs, or “inverted k -mer indexes” to very large collections (millions) of data sets with applications to DNA sequence search. In particular, I address the superlinear scaling with number of samples of existing tools with and present a novel representation and implementation of a probabilistic coloured de Bruijn graph, a “Coloured Bloom Graph”, that can scale sequence search to millions of bacterial sequence data sets. I demonstrate the scalability of this tool by building a Coloured Bloom Graph of all publicly accessible microbial WGS data, almost half a million samples, and use

1. INTRODUCTION

it to run sub-millisecond k-mer searches in this data.

Finally, I conclude by discussing a potential framework for combining **Mykrobe predictor**, or similar tools, with Coloured Bloom Graphs and other probabilistic data structures to create a distributed, scalable, cloud-based bacterial surveillance platform and offer concluding remarks.

CHAPTER 2

*Rapid antibiotic resistance predictions
from high throughput genome sequence
data*

2.0.1 Publication note and acknowledgements

The majority of the work described in this chapter was previously published in Bradley et al. 2015. I briefly discuss some extensions and projects involving `Mykrobe` predictor, of which I am a co-author (Lipworth et al. 2017; Mason et al. submitted; Quan et al. 2017) in section 2.9. Figures 2.1 – 2.14 and some of the text in this chapter comes from this publication. Although the work was collaborative in parts, the work described here is the sole work of myself, with the guidance of my supervisors, unless otherwise specified below:

- The *S. aureus* and *M. tuberculosis* phenotyping described in Section 2.10.2 was performed by N. Claire Gordon and Laura Dunn.
- The nanopore sequencing described in Section 2.10.6 was performed by Maria Teresa de Cesare and Paolo Piazza.
- Luke Anson and Antonina A. Votintseva performed sample preparation for MinION and MiSeq.
- Sarah Earl and Tanya Golubchik ran the RaxML to generate the *S. aureus* and *M. tuberculosis* phylogenies.
- The Windows and Mac interface shown in Figures A.1 – A.3 was created by Simon Heys.

2.1 Introduction

In this chapter, I consider the problem of predicting the species and antibiotic resistance of a pathogen from whole-genome sequencing (WGS) in a clinical setting. I consider two exemplar species: *M. tuberculosis*, a slow growing microbe which evolves resistance via point mutations during the course of infection; and *Staphylococcus aureus*, a fast growing organism which has a diverse range of mechanisms, from point mutations in core and accessory genes to mobile genetic elements.

First, I describe how we can use coloured de Bruijn graphs to build clinically informed taxonomy ‘probe sets’ and use these to rapidly infer complex, species, and strain in a targeted manner. Metagenomic classification tools such as Kraken (Wood et al. 2014) aim for a balance of sensitivity and specificity across a broad taxonomic range with no prior on which species are most likely. In contrast, our goal with `Mykrobe predictor` was to tune sensitivity and specificity for clinical considerations where there is a prior knowledge of the likely pathogen and where some error modes have more impact than others. For example, misidentifying species or lineage within MTBC has limited impact on choice of treatment, but it would be more problematic to misidentifying a *M. tuberculosis* as a nontuberculous mycobacterium (NTM).

I then describe how, by using a catalogue of diverse resistance elements, I built a reference de Bruijn graph and used it to rapidly genotype our sample. I used these genotypes to infer antibiograms that are as accurate as phenotypic assays in *Staphylococcus aureus*, and line probe assays in *M. tuberculosis* (Abreu Maschmann et al. 2013; Chryssanthou et al. 2012). Using depth of coverage information contained

2. RAPID ANTIBIOTIC RESISTANCE PREDICTIONS FROM HIGH THROUGHPUT GENOME SEQUENCE DATA

in the de Bruijn graph, I show that `Mykrobe predictor` can detect contamination and minor populations, and that these could have clinical consequences. The software, `Mykrobe predictor`, was implemented with an easy-to-use “drag and drop” interface and I discuss the performance of this software as well as its impact since publication in 2015, including its inclusion in Public Health England’s routine *M. tuberculosis* pipeline.

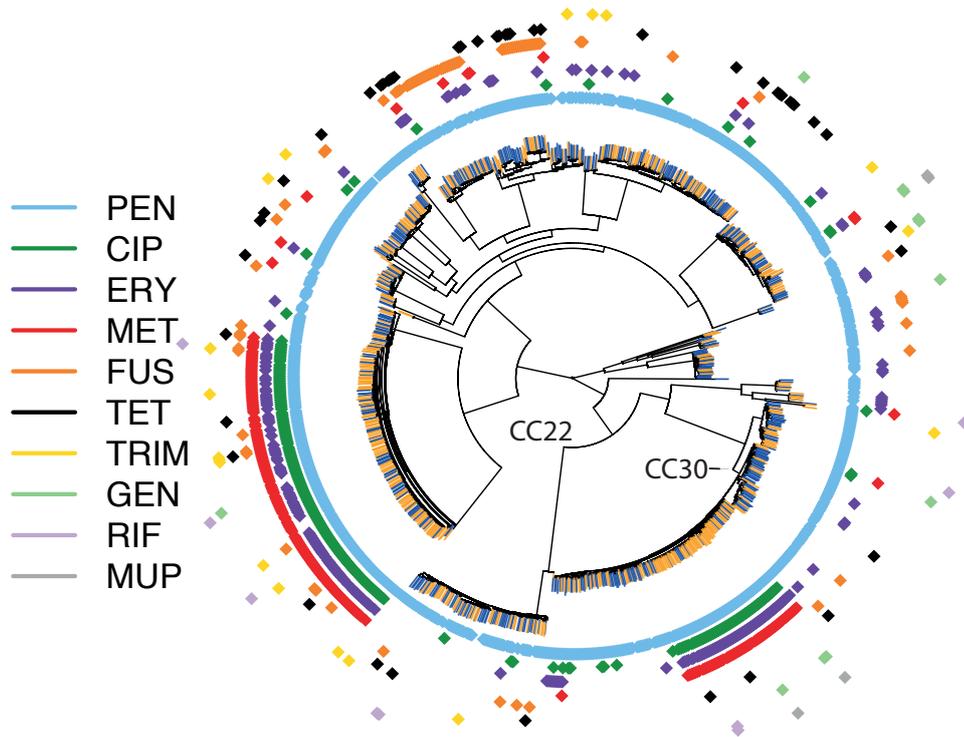


Figure 2.1: Phylogeny of *Staphylococcus* samples used in evaluating resistance prediction, with tips marked orange or blue to represent samples in training set (St_{A1} , $n=495$) or validation set (St_{B1} , $n=471$). Drug resistance is indicated in concentric rings around the phylogenetic tree; plasmid-mediated resistance (erythromycin in purple, tetracycline in black) is distributed across the whole tree. The two multi-drug-resistant clades are in UK hospital clonal complexes CC22 and CC30.

2.2 *Staphylococcus* and *Mycobacterium* data sets represent global diversity

2.2.1 Overview of *Staphylococcus aureus* data sets

In order to train and validate our species and drug susceptibility prediction software Mykrobe predictor for *S. aureus*, I used a collection of 1,423 *S. aureus*

isolates. I used a training set (St_{A1}) of 495 and a validation set (St_{B1}) of 471 *S. aureus* isolates that had been sequenced and phenotyped after being collected in Bristol, UK, and Oxfordshire, UK, for resistance prediction. These collections were supplemented with 457 coagulase-negative staphylococci (CoNS) for species identification analysis. See Section 2.10.2, Figure 2.16, and Table 2.4, for an overview of the data sets and associated metadata.

2.2.1.1 *Staphylococcus aureus* phylogeny

I show a tree constructed from St_{A1} and St_{B1} in Figure 2.1, which also displays membership of training or validation sets and phenotypic drug resistance. The training (orange tips) and validation (blue tips) samples are distributed across the phylogeny, which includes all major clonal complexes (Figure 2.2). The collection is enriched for the UK hospital-associated *S. aureus* clonal complexes (CC22/CC30), where multidrug resistance is prevalent.

2.2.2 Overview of *M. tuberculosis* data sets

In order to train and validate Mykrobe predictor for *M. tuberculosis* I used a collection of 4,056 *Mycobacterium* isolates. I use a ‘training’ data set $MTBC_A$ of 1,920 *M. tuberculosis* complex (MTBC) isolates with Illumina sequence data and associated drug susceptibility test (DST) data (Figure 2.16, Table 2.4) from Oxfordshire, Birmingham, Sierra Leone, and South Africa to train the resistance prediction algorithm. I used a separate data set ($MTBC_B$) of 1,609 further isolates

2.2. *Staphylococcus* and *Mycobacterium* data sets represent global diversity

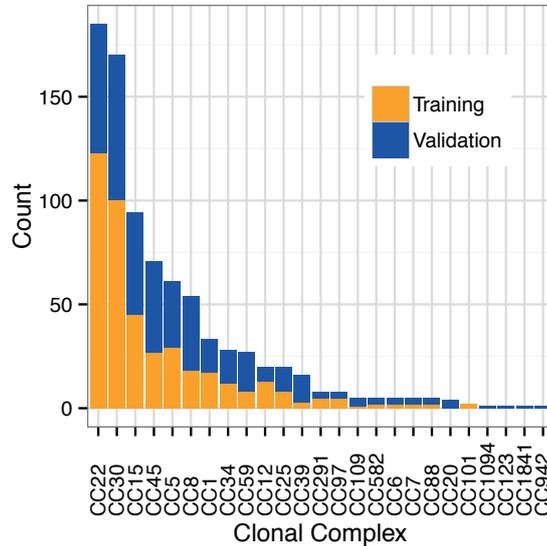


Figure 2.2: Counts of each clonal complex in *Staphylococcus* training set St_{A1} and validation set St_{B1} . All major clonal complexes were represented.

from Uzbekistan, Germany, South Africa, and the UK for the validation data set. All of these samples had previously been collected for an independent study on the discovery of mutations predictive of resistance (Walker et al. 2015). For species identification training and validation these were supplemented with an additional 527 non-tuberculosis mycobacterium (NTM) data sets $MYCO_{SRA}$ and $MYCO_{RETRO}$ (see Extended Methods).

2.2.3 *M. tuberculosis* phylogeny

Figure 2.3 shows a phylogeny of $MTBC_A$ and $MTBC_B$, with training and validation samples coloured at the branch tips in orange and blue, respectively (see Section 2.10.3.1 for details on construction). The validation set shows some clus-

2. RAPID ANTIBIOTIC RESISTANCE PREDICTIONS FROM HIGH THROUGHPUT GENOME SEQUENCE DATA

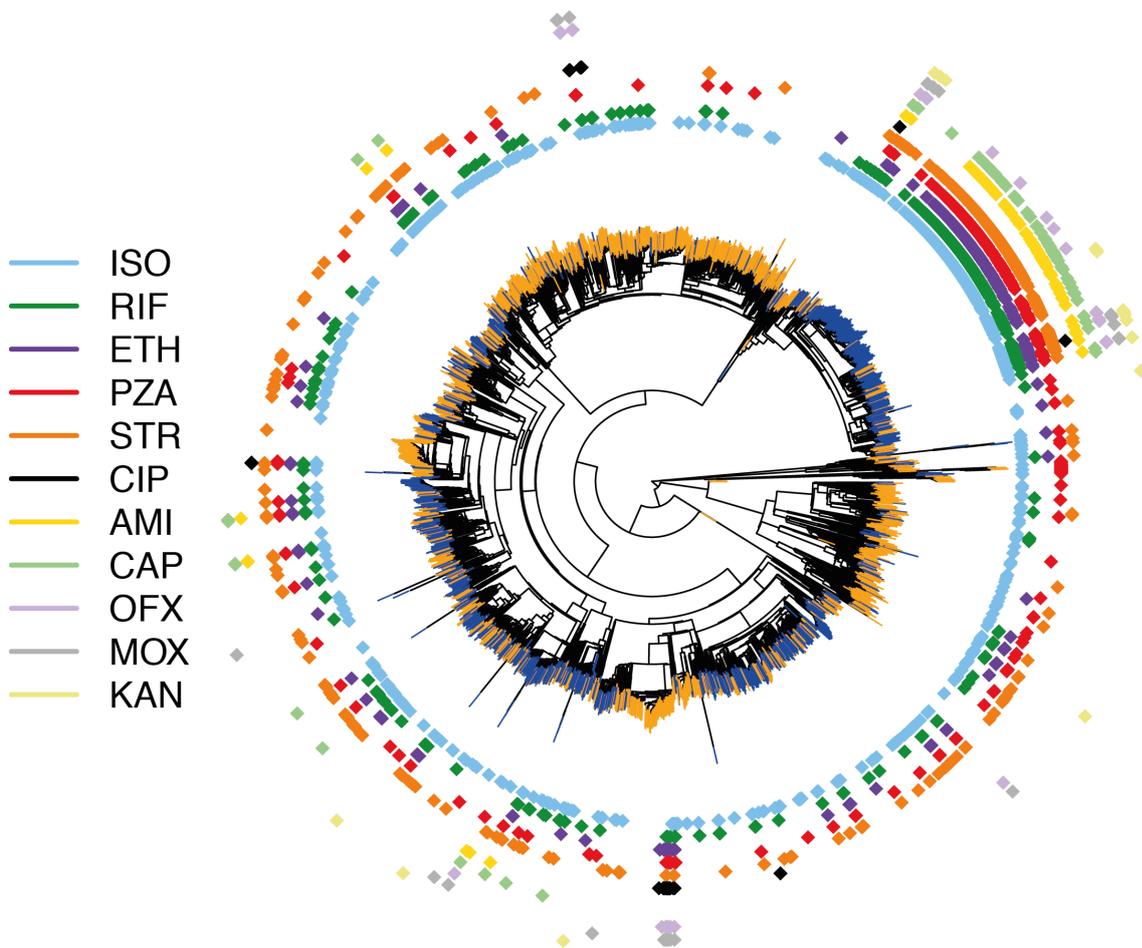


Figure 2.3: Phylogeny of MTBC samples with phenotype data, with tips marked orange or blue to indicate training set ($MTBC_A$, n=1,920) or validation set ($MTBC_B$, n=1,609). Drug resistance is shown in concentric rings around the phylogenetic tree. Resistance exists across the phylogeny, especially against isoniazid (light blue), with a clustering of multi-drug resistance in the Beijing lineage.

2.2. *Staphylococcus* and *Mycobacterium* data sets represent global diversity

tering within the phylogeny due to the large number of samples from Uzbekistan in the validation set, with a resulting high number of XDR TB in the validation set.

2.3 Species identification from WGS data using coloured de Bruijn graphs

Our goal with `Mykrobe predictor`'s species identification was to develop software that was sensitive to differences between highly similar species and lineages, but also lightweight enough to be run locally on computers with limited resources. In order to ensure sensitivity to low frequency contaminating species, we wanted to use more than one probe (as opposed to common methods based on single genes, e.g., *rpoB*, *gyrA*, etc.), which contained more sequence than a single gene.

2.3.1 Species probe generation

To achieve this, I developed a system for designing a hierarchy of markers (contigs), which separated phylogroups (e.g., *S. aureus* from coagulase-negative Staphylococci, or MTBC from NTM). I first built a de Bruijn graph by pooling several hundred samples from both phylogroups, and pulled out all unique and unambiguous contigs ("unitigs") with `cortex`.

For each group I calculated the frequency of each contig in each phylogroup (e.g., the frequency of each contig within MTBC and NTM). I chose the 2,000 most highly differentiated contigs to form marker panels to distinguish the groups. This process was run to find contigs informative at the complex (e.g., MTBC), sub-complex (e.g., *Mycobacterium avium* complex), species (e.g., *M. tuberculosis*), and lineage/sub-species level (e.g., European/American *M. tuberculosis*), using the tree

structure of the phylogeny.

2.3.2 *Staphylococcus* species identification

The above process was applied to 731 staphylococcal isolates in training set St_A , which combines data sets St_{A1} (532 clinical *S. aureus* isolates) and St_{A2} (199 coagulase-negative staphylococci (CoNS) isolates), using `cortex` (Iqbal et al. 2012) with kmer-size 15, to produce probes (contigs) for phylogroups (*S. aureus* versus coagulase negative staphylococci) and species (*S. aureus*, *S. epidermidis*, *S. haemolyticus*, other coagulase-negative species). I used presence of the catalase gene to confirm presence of staphylococci. I plotted the proportion of sequence in the probe panel found in each training sample (“recovery”), and ignored extreme outliers as possible errors in the SRA metadata. Detection thresholds were chosen based on recovery in the training set: 90% for *S. aureus*, 30% for *S. epidermidis* and *S. haemolyticus*, 10% for other staphylococci, and 20% for the catalase gene.

I then evaluated our predictions on a separate validation set (St_B), combining 471 *S. aureus* samples (St_{B1}) and 221 CoNS (St_{B2}), and show the results in Figure 2.4. This confirmed an appropriately low rate of missing a true *S. aureus* sample (0/492, upper 97.5% CI 0.7%). I studied the 3 non-*S. aureus* samples that appeared to be misclassified by `Mykrobe` predictor as *S. aureus*, and concluded that they were mislabelled in the NCBI Short Read Archive (SRA) (our “truth”), as both BLAST (Altschul et al. 1990) and `OneCodex` (Minot et al. 2015) agreed with `Mykrobe` predictor that these were *S. aureus*.

2. RAPID ANTIBIOTIC RESISTANCE PREDICTIONS FROM HIGH THROUGHPUT GENOME SEQUENCE DATA

	Mykrobe				
	S.aur	S.epi	S.hae	O.St.	Non-St.
S.aur	492	0	0	0	0
S.epi	0	54	0	0	2
S.hae	0	5	68	0	1
O.St.	3	2	5	60	0
Non-St.	0	0	0	0	0

Figure 2.4: A confusion matrix showing species predictions for *Staphylococcus* vs our “Truth” from SRA metadata: species classification results on species validation set St_B ($n=692$). Red shading of box indicates errors we wish to minimise. Abbreviations: S.aur: *S. aureus*, S.epi: *S. epidermidis*, S.hae: *S. haemolyticus*, O.st : other staphylococcus, Non-st: Non staphylococcal. “Truth (SRA)” is the species as annotated in the SRA metadata, which was used as truth for comparisons.

2.3.3 *M. tuberculosis* species identification

Species within the *M. tuberculosis* complex (MTBC) cause tuberculosis, but clinical samples may consist of other mycobacterial species. Co-infection with both MTBC and NTM, which is known to occur (Jun et al. 2009; Maiga et al. 2012), should be reported, if present. Consequently, we chose to identify 4 MTBC species (*M. tuberculosis*, *M. africanum*, *M. bovis*, *M. caprae*) and 40 NTM species (See Section 2.10.4).

In terms of desired error profile, the main aim was to minimize misclassifying a MTBC as a NTM, or vice versa. Misidentifying species within MTBC has limited impact on choice of treatment, except that *M. bovis* is known to be intrinsically re-

2.3. Species identification from WGS data using coloured de Bruijn graphs

		Mykrobe					
		M.tb.	M.bv.	M.af.	MTBC	NTM	Other
Truth	M.tb.	1192	0	1	2	0	0
	M.bv.	0	7	0	0	0	0
	M.af.	0	0	0	4	0	0
	MTBC	10	1	0	0	0	0
	NTM	0	0	0	0	84	0
	Other	0	0	0	0	0	3

Figure 2.5: A confusion matrix showing species predictions for *M. tuberculosis*: Species classification results on a validation set ($MTBC_{A2}$ + Myco_Retro, $n=1304$). Colours indicate misclassifications between NTM/MTBC (red), concordance with “truth” (dark green), or greater resolution from Mykrobe predictor than PCR (light green). Abbreviations: M.tb.: *M. tuberculosis*, M.af.: *M. africanum*, M.bv.: *M. bovis*.

sistant to pyrazinamide, and some substrains of the Bacille Calmette–Guérin (BCG) strain of *M. bovis* are known to be resistant to isoniazid.

Marker panels were generated for MTBC and NTM (“phylo groups”) as well as for individual species within those groups, using the same method as for staphylococci (Section 2.3.1) based on a training set consisting of data sets $MTBC_{A1}$ (338 MTBC clinical isolates) and $MYCO_{SRA}$ (380 *Mycobacterium* samples downloaded from the Short Read Archive (SRA)).

The required percentage of probe sequence to be found in each of the “phylo group” panels was set as 70% for MTBC, 25% for the NTM panel, and 30% for all species panels. Above this threshold the phylo group or species was predicted to be

2. RAPID ANTIBIOTIC RESISTANCE PREDICTIONS FROM HIGH THROUGHPUT GENOME SEQUENCE DATA

present. I also used the lineage-informative SNPs defined by Stucki et al. (Stucki et al. 2012) to assign *M. tuberculosis* lineages: Beijing/East Asia, East Africa/Indian ocean, Delhi/Central Asia, European/American, West Africa 1 & 2, or Ethiopian.

I then evaluated species prediction on the union of data sets $MTBC_{A2}$ (1157 MTBC clinical isolates) and $MYCO_{RETRO}$ (147 *Mycobacterium* isolates), where species had been identified by Hain assay, showing results in Figure 2.5. No samples were misclassified between MTBC and NTM, but there were four *M. africanum* and two *M. tuberculosis* samples that were only resolved to MTBC, and one *M. tuberculosis* sample misidentified as *M. africanum*. Finally, I tested our identification of the lineages as defined by Comas et al. 2013 by comparing with the lineage as identified by their tool, KvarQ (Steiner et al. 2014) and found 100% concordance between the two tools. These lineages' probesets have recently been extended and validated by Lipworth et al. 2017 using Mykrobe predictor.

2.4 Using population genome graphs for genotyping

Once a sample has been identified as *M. tuberculosis* or *S. aureus*, the next step is to genotype a set of variants or genes from which its antibiogram can be inferred. Various methods have been used for genotyping resistance features: mutations and genes have been detected by whole genome assembly (Gordon et al. 2014), genes by assembly and BLAST (Leopold et al. 2014), or SNPs and indels by mapping (Kohl et al. 2014; Köser et al. 2013). As I discussed in section 1.5, these approaches have risks of bias, so here I take a reference graph approach.

In Figure 2.6, I show a cartoon of the genetic diversity in a bacterial species and two options for building a reference variation structure. In option i) I show a standard approach, similar to that taken by Kohl et al. 2014 and Köser et al. 2013, where we select an arbitrary strain (Strain 1) to be the reference genome, along with one copy of each plasmid gene. This approach (option i), whereby sequence reads are mapped to the reference genome and genes, requires the mapping and inference to cope with the divergence between sample strains and the reference.

With `Mykrobe predictor`, we take a reference graph based approach, shown in option ii). We start with a curated knowledge base of resistant/susceptible alleles, and assemble a de Bruijn graph (Iqbal et al. 2012; Turner et al. 2017) of them on different genetic backgrounds, along with many examples of resistance genes. This forms our reference graph. In Figure 2.6b, I show the corresponding analyses of a

2. RAPID ANTIBIOTIC RESISTANCE PREDICTIONS FROM HIGH THROUGHPUT GENOME SEQUENCE DATA

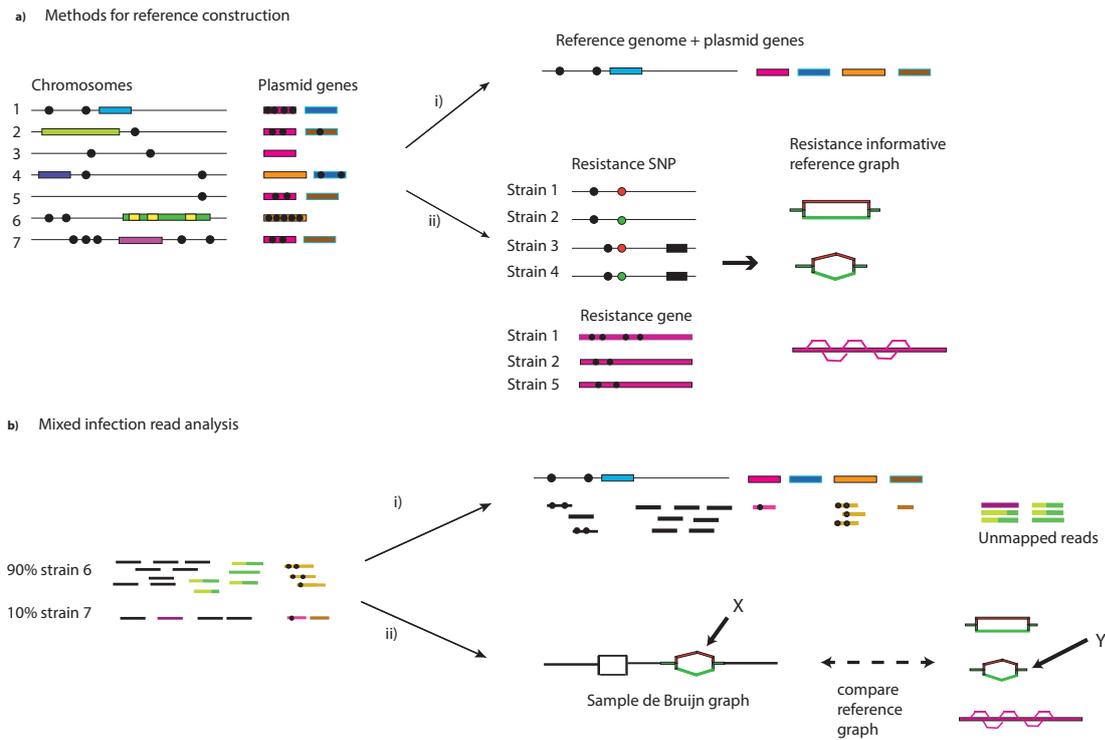


Figure 2.6: Representation and analysis of bacterial genetic variation. (a) Reference construction methods. Left: chromosomes with SNPs (black circles) and genes (coloured blocks) from strains of a bacterial species. Option (i) picks strain 1 to be reference, plus one example of each plasmid resistance gene. In option (ii), our method is to build the de Bruijn graph of all strains, restrict to loci of interest, and annotate resistance (red) and susceptible (green) alleles. For SNPs, local graph topology is determined by adjacent SNPs (black dots) and indels (black blocks). (b) Mixed infection read analysis. Left: sequence data from a clinical sample harbouring major (90%) and minor (10%) strains. Right: option (i) maps the reads to the reference genome to detect SNPs and genes. In option (ii), our approach, we construct the de Bruijn graph of the sample and compare with the reference graph. We see a specific SNP is present both in the sample and the reference graph (marked X,Y). Both the resistant (red) and susceptible (green) alleles are present in the sample, and within-sample frequency is estimated from sequencing depth on each allele.

mixed sample. Our approach (option ii) directly compares the de Bruijn graph of the sample with the reference graph. This results in statistical tests for the presence of resistance alleles that are unbiased by choice of reference or assumptions of clonality. Moreover, these tests will improve as the catalogue of diversity in the species grows.

In order to control memory use, the implementation of `Mykrobe predictor` first builds a target graph from the genes, alleles, and species identifying contigs, and then only loads sample data that intersects it, restricting RAM use to less than 100Mb, regardless of the size of the sample’s data. The inverse could also be done, where the sample’s de Bruijn graph is built and the reference graph of resistance elements is intersected with it. This would mean that memory use would be much higher, but it would have the advantage that the data structure could be also be used to perform variant discovery. Since our goal with `Mykrobe predictor` is only to genotype an existing catalogue of genomic elements, I report results using the former method only.

2.4.1 Genotyping at mutations

2.4.1.1 Constructing variant probe sets

In order to use a de Bruijn graph to genotype a catalogue of variants we first convert the list of variant sites into a set of sequence ‘probes’ of length $2k+1$. If the resistance mutations are defined in amino acid space, we first convert these into a set of possible codon changes in DNA space. In the simple case, we then represent these DNA SNPs into two short sequences (of length $2k+1$), one “reference” or

2. RAPID ANTIBIOTIC RESISTANCE PREDICTIONS FROM HIGH THROUGHPUT GENOME SEQUENCE DATA

“susceptible” allele, and the other “alternate” or “resistant” allele. However, if there is common variation within k bases of the variant site, any sample with this variation would have different susceptible or resistant allele sequence. In order to be more robust to this variation, we also build equivalent susceptible or resistant alleles for each variant within k bases of our SNP of interest, in any sample in the training set. These leads to at least two alleles for each SNP, but often several more.

From this set of probes, we construct our reference de Bruijn graph and calculate the percentage of k -mers with depth of coverage at least 1, and the median depth on each probe. For all mutations in the panel we iterate through all possible nucleotide changes that would generate the specified amino acid change (if appropriate), and find the resistant allele and susceptible allele with the highest coverage. We then compare three competing models: pure susceptible, minor resistant (frequency=10%), and major resistant (we used frequency=75%, but we expect that values from 60%-100% would result in identical model choice). In this and subsequent sections, a subscript MAJ, MIN, or S refers to the major resistant, minor resistant or susceptible models—also referred to as 1/1, 0/1, and 0/0, respectively. We use the following simple Poisson model for the likelihoods for all three models. Susceptible model: Likelihood specified by Poisson coverage on S allele, plus errors driving both coverage loss on S allele and coverage on R allele.

S:

$$Cov(S \text{ allele}) \sim Pois(D(1 - \varepsilon)^k),$$

$$Cov(R \text{ allele}) \sim Pois\left(\frac{D\varepsilon(1-\varepsilon)^{k-1}}{3}\right);$$

MAJ:

$$\begin{aligned} Cov(R \text{ allele}) &\sim Pois(D\varepsilon(1-\varepsilon)^k), \\ Cov(S \text{ allele}) &\sim Pois\left(\frac{D(1-\varepsilon)^{k-1}}{3}\right); \end{aligned}$$

and for minor resistant Poisson coverage on both alleles scaled by frequency:

MIN:

$$\begin{aligned} Cov(S \text{ allele}) &\sim Pois(D(1-f)(1-\varepsilon)^k), \\ Cov(R \text{ allele}) &\sim Pois(Df(1-\varepsilon)^k); \end{aligned}$$

where $Cov()$ is a function returning median depth of coverage on an allele, D is the expected depth of coverage, ε is the per-base error rate, k is the k -mer size, and the frequency f of the resistance allele is 0.1/0.75 for the minor/major resistant model. The likelihoods of the models are modified by the following indicator functions:

$$P_S = 1,$$

$$P_{MAJ} = I(\text{perc}(R) = 100\%),$$

$$P_{MIN} = I(\text{perc}(R) = 100\% \text{ AND } \text{perc}(S) = 100\%),$$

where $I()$ is an indicator function, and $\text{perc}(R)$ and $\text{perc}(S)$ are the percentage of the k -mers in the resistant/susceptible alleles that are seen in the sample. This ensures that to have a non-zero likelihood for the MAJ and MIN model the relevant

probes must have 100% coverage. In addition, because, with `Mykrobe predictor`, we also have knowledge of the presence of any contaminating species, we ignore the minor resistant model if any are observed. The maximum likelihood model is chosen.

2.4.2 Genotyping at genes

Because we often have multiple versions of the same gene in our resistance catalogue, the first step to determine the presence of genes is to choose the gene version with the most k -mers with non-zero depth. If there are more than one such gene version, we choose the version with the highest median depth. The expected proportion of k -mers in a gene which is observed is:

$$\gamma = 1 - P(\text{gap}) = 1 - \exp(-Df).$$

We use the following priors:

$$Pr_S = 1,$$

$$Pr_{MAJ,MIN} = I[\max_{G_i}(\text{perc}(G_i)) > \gamma K(G)],$$

where each gene G has multiple exemplars G_i representing diversity of that gene, I is an indicator function, and $\text{perc}()$ is a function returning the percentage of k -mers present in the sample. K is the minimum percentage of k -mers expected to be recovered for a gene, based on the empirical level of diversity observed in a training set.

2.4. Using population genome graphs for genotyping

The likelihoods for major and minor models depends on the probability of having the observed median coverage across the gene. The likelihood for the major resistant, minor resistant, and susceptible models are given by:

$$Cov(gene) \sim Pois(Df),$$

where $Cov()$ is a function returning median depth of coverage on the gene, D is the expected depth of coverage, and f is the expected frequency of the gene where $f(R) = 1, f(r) = 0.1, f(S) = \varepsilon$, and ε is the error rate. The maximum a posteriori model is chosen.

2.5 Resistance prediction

A description of the catalogues of genomic elements from which resistance was inferred can be found in Section 2.10.2.2 and Section 2.10.3.2. From these, I built variant and gene “probe-sets” which `Mykrobe predictor` uses to genotype via the methods described in the previous section. From the genotypes we can infer resistance or susceptibility to the associated drugs.

Performance of resistance prediction is typically assessed via their major error (ME) and very major error (VME) rates (US-FDA et al. 2007). A major error’s reference (“truth”) result is S and prediction result is R. A very major error’s reference (“truth”) result is R and prediction result is S. Below I use “major error” interchangeably with “false positive”, and “very major error” with “false negative”. The ME rate is defined as:

$$\text{Major Error Rate} = \frac{FP}{TN + FP} = 1 - \text{specificity} = 1 - \frac{TN}{TN + FP},$$

and VMR rate as:

$$\text{Very Major Error Rate} = \frac{FN}{TP + FN} = 1 - \text{sensitivity} = 1 - \frac{TP}{TP + FN}.$$

The U.S. Food and Drug Administration office requires rates that are < 3% for ME and < 1.5% for VME. They also require the 95% confidence interval for VME rates to be below 7.5% for the upper boundary and below 1.5% for the lower boundary (US-FDA et al. 2007).

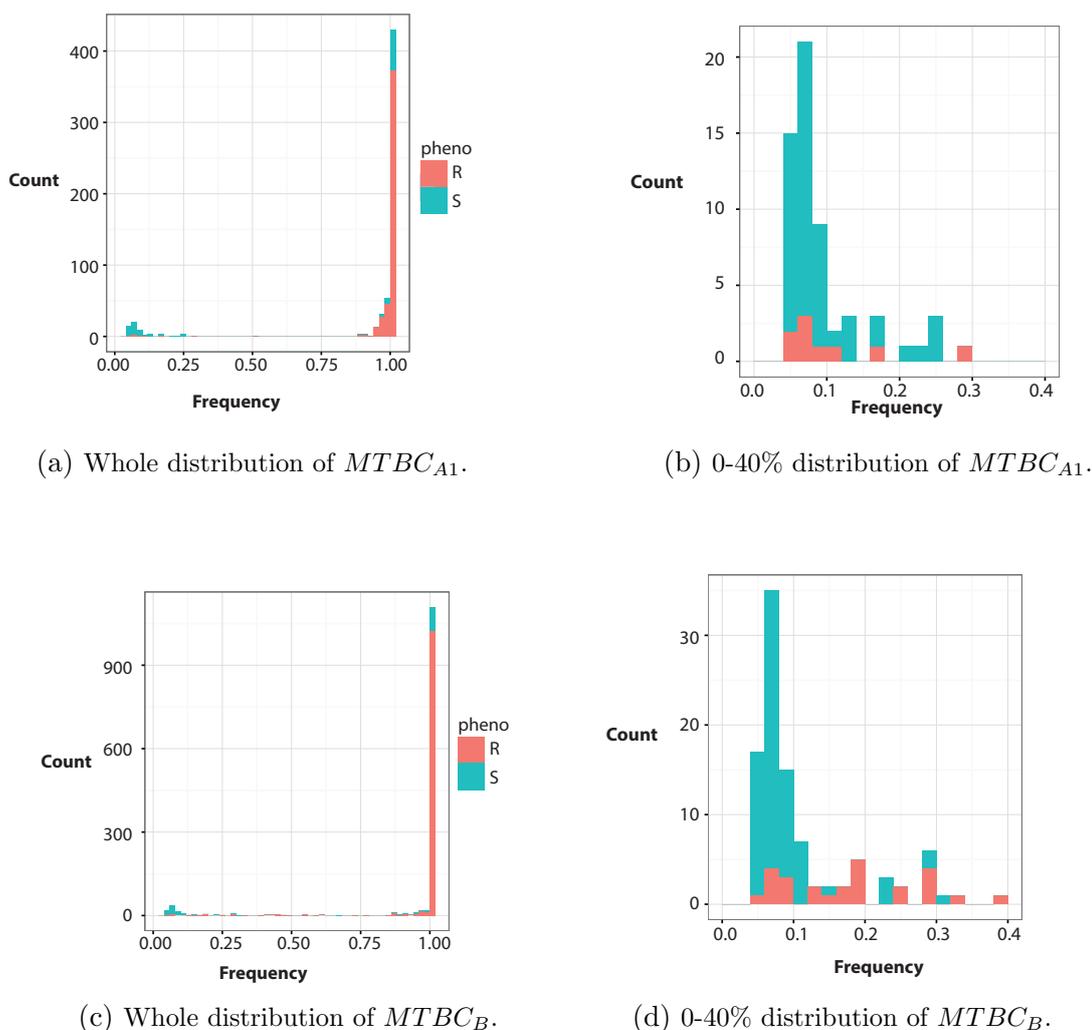


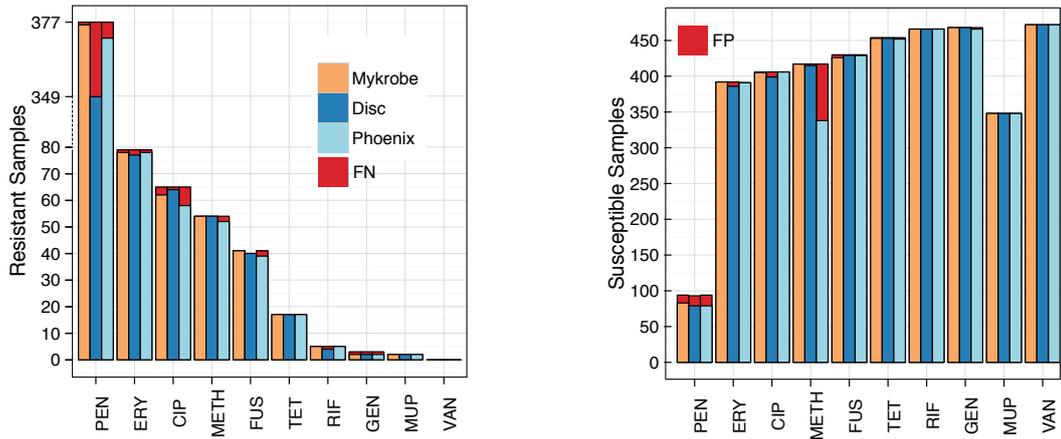
Figure 2.7: Within-sample frequency of resistant alleles in the training set $MTBC_{A1}$ and validation set $MTBC_B$, coloured by associated phenotype. Alleles at frequency $> 90\%$ with an associated susceptible phenotype are as follows (where X means any amino acid): Training set: embB M306X: 31, fabG1 C-15X: 5, fabG1 G-17X: 1, fabG1 T-8X: 2, gyrA A90X: 3, gyrA D94X: 9, gyrA S91X: 2, katG S315X : 1, rpoB D435X: 3, rpoB H445X: 3, rpoB L430X: 1, rpoB L452X : 3, rpoB S450X: 2, rrs G1484X : 1; Validation set: embB M306X: 54, fabG1 C-15X: 9, katG S315X: 8, rpoB L430X: 3, rpoB L452X: 3, rpoB Q429X : 1, rpoB Q432X: 1, rpoB S450X: 4, rpsL K43R: 6, rrs A1401X : 7, rrs C1402X : 1, rrs C517X: 1. The dominant mutations, embB M306V and M306I, are a known phenomenon, as the Minimum Inhibitory Concentration (MIC) of resistance caused by these mutations is very close to the critical concentration—causing stochastic “flip-flopping” of the test, depending on whether resistance is just above or below the threshold. This is an artifact of forcing a binary classification on a quantitative trait.

2.5.1 Resistance calling at mutations

If the frequency of the genotyped variant is below a threshold (T , determined below) Mykrobe predictor reports the mutation but predicts a susceptible phenotype. Otherwise it predicts a resistant phenotype, and mentions whether it classified this as a minor or major population. For *S. aureus*, the protocol undergone by samples in the standard clinical workflow removed almost all mixture, and so there were almost no minor alleles to either train or validate on—I set an arbitrary threshold of 10%. For *M. tuberculosis*, I did not have enough data to estimate per-drug thresholds, and we knew that the phenotyping data was imperfect. I examined the two frequency distributions of resistance alleles present in phenotypically resistant/susceptible samples in the training set (Figures 2.7a and 2.7b, and selected a single threshold of 10% for all drugs. The corresponding distributions for the validation set can be seen in Figures 2.7c and 2.7d.

2.5.2 Resistance calling at genes

Genes in *S. aureus* can either be in the chromosome, at copy number 1, or on a plasmid, with higher copy number. In the training set, a threshold frequency was chosen for each gene, such that above this frequency the sample was more likely to be resistant than sensitive (see Section 2.10.5). If the gene was genotyped as Minor Resistant but the gene’s estimated frequency (based on median coverage over overall depth of coverage) was below this threshold a susceptible phenotype was reported (see Section 2.10.5). Otherwise, a (major or minor) resistant phenotype



(a) Proportion of resistant *Staphylococcus* samples correctly identified as resistant. VME/FNs in red.

(b) Proportion of susceptible *Staphylococcus* samples correctly identified as susceptible. ME/FPs in red.

Figure 2.8: Comparison of Mykrobe predictor, Disc, and Phoenix antimicrobial resistance predictions in St_{B1} : Mykrobe predictor (orange), disc test (dark blue), and Phoenix (light blue) compared with consensus, with false negatives in red. Note the break in the y axis between 80 and over 300 to show penicillin on same plot. PEN, penicillin; ERY, erythromycin; CIP, ciprofloxacin; METH, methicillin; FUS, fusidic acid; CLIN, clindamycin; TET, tetracycline; RIF, rifampicin; GEN, gentamicin; MUP, mupirocin; TRIM, trimethoprim; VAN, vancomycin.

was reported.

2.5.3 Mykrobe predictor *Staphylococcus* predictions

match consensus phenotype

Considering each resistance mutation and gene in turn, our prediction algorithm first genotypes a sample into one of three categories: clonal-susceptible, minor-frequency resistant allele, or major-frequency resistant allele following the methods

2. RAPID ANTIBIOTIC RESISTANCE PREDICTIONS FROM HIGH THROUGHPUT GENOME SEQUENCE DATA

Table 2.1: Resistance prediction results for *Mykrobe predictor* on the *S. aureus* validation set (St_B1), treating the consensus phenotype as gold standard except for trimethoprim (which Phoenix does not test) where the disc test was used as truth. FN: False negative calls. R: total number of resistant samples. FP: false positives. S: total number of susceptible samples. VME: very major error rate (false negative rate), only shown where R > 10. ME: major error rate (false positive rate), only shown where S > 10. PPV: positive predictive value. NPV: negative predictive value. N/A: Not Applicable.

Error rates shown with 95% CI calculated by Clopper-Pearson.

Error rates' CIs meeting FDA requirements are in bold.

Abbreviations – PEN: penicillin, ERY: erythromycin, CIP: ciprofloxacin, METH: methicillin, FUS: fusidic acid, CLIN: clindamycin, TET: tetracycline, RIF: rifampicin, GEN: gentamicin, MUP: mupirocin, TRIM: trimethoprim, VAN: vancomycin

Drug	FN(R)	FP(S)	VME (%)	ME (%)	PPV (%)	NPV (%)
PEN	1 (377)	11 (94)	0.3 (0.0-1.5)	11.7 (6.0-20.0)	97.2 (95.0-98.6)	98.8 (93.5-100.0)
ERY	1 (79)	0 (392)	1.3 (0.0-6.9)	0.0 (0-0.9)	100.0 (95.4-100)	99.7 (98.6-100.0)
CIP	3 (65)	1 (406)	4.6 (1.0-12.9)	0.2 (0.0-1.4)	98.4 (91.5-100.0)	99.3 (97.9-99.8)
METH	0 (54)	0 (417)	0.0 (0-6.6)	0.0 (0-0.9)	100.0 (93.4-100)	100.0 (99.1-100)
FUS	0 (41)	4 (430)	0.0 (0-8.6)	0.9 (0.3-2.4)	91.1 (78.8-97.5)	100.0 (99.1-100)
CLIN	0 (25)	1 (97)	0.0 (0-13.7)	1.0 (0.0-5.6)	96.2 (80.4-99.9)	100.0 (96.2-100)
TET	0 (17)	1 (454)	0.0 (0-19.5)	0.2 (0.0-1.2)	94.4 (72.7-99.9)	100.0 (99.2-100)
RIF	0 (5)	0 (466)	0.0 (0-52.2)	0.0 (0-0.8)	100.0 (47.8-100)	100.0 (99.2-100)
GEN	1 (3)	0 (468)	33.3 (0.8-90.6)	0.0 (0-0.8)	100.0 (15.8-100)	99.8 (98.8-100.0)
MUP	0 (2)	0 (348)	0.0 (0-84.2)	0.0 (0-1.1)	100.0 (15.8-100)	100.0 (98.9-100)
TRIM	0 (1)	1 (188)	0.0 (0-97.5)	0.5 (0.0-2.9)	50.0 (1.3-98.7)	100.0 (98.0-100)
VAN	0 (0)	0 (472)	NaN	0.0 (0-0.8)	NaN	100.0 (99.2-100)

outlined above in Section 2.4. It then predicts a resistant phenotype for samples containing resistant alleles of sufficiently high frequency (See Section 2.10.5). Note that for antibiotics where resistance is mediated by genes on variable copy-number plasmids (e.g., erythromycin and tetracycline), a minor population with high copy-number of a resistance-carrying plasmid may sometimes be called as major resistant. After using the training data set to estimate parameters for our statistical model, I applied `Mykrobe predictor` to the validation set.

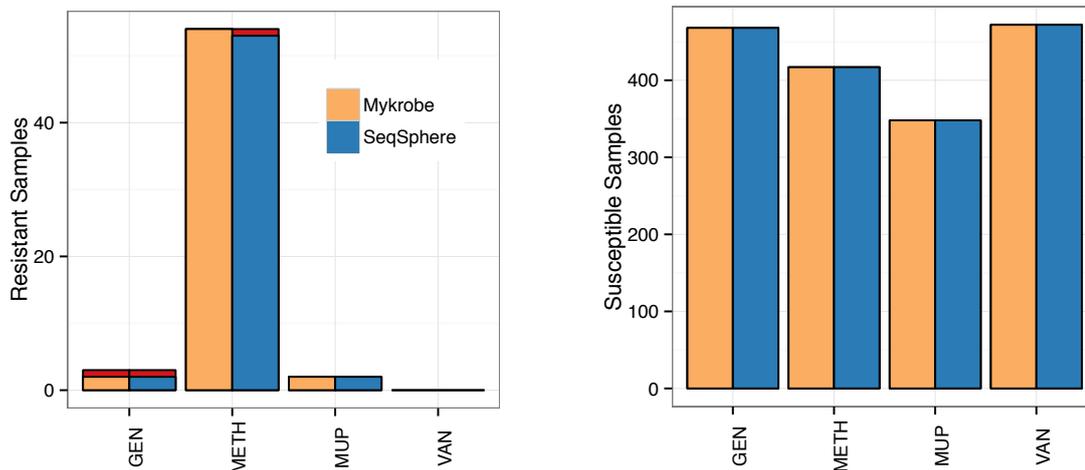
Figure 2.8 shows in red the false negative calls (Figure 2.8a) and false positive calls (Figure 2.8b) for `Mykrobe predictor` and the two laboratory methods: disc and Phoenix. If we consider Figure 2.8a first (showing very major errors), and focus on the 7 drugs with more than 10 resistant samples, then `Mykrobe predictor` misses fewer resistant calls than the other individual phenotypic methods for all drugs except ciprofloxacin. Ciprofloxacin had a false negative rate of 4.6%—I was unable to determine the reason for these missed resistant predictions, although we note that disc and Phoenix had similar problems, and that this drug has at least one uncharacterized mechanism for resistance (Pidcock et al. 2002). For the 3 drugs (methicillin, penicillin, erythromycin), for which our study has enough resistant samples to meet US Food and Drug Administration (FDA) criteria (VME rate $< 1.5\%$, upper 95% confidence interval $< 7.5\%$), `Mykrobe predictor` met these criteria (US-FDA et al. 2007). The data underlying this plot are presented in Table 2.1. The results for the Disc and Phoenix tests on the validation set are shown in Table A.2 and Table A.3.

The equivalent plot for false positive calls is shown in Figure 2.8b. For all

2. RAPID ANTIBIOTIC RESISTANCE PREDICTIONS FROM HIGH THROUGHPUT GENOME SEQUENCE DATA

drugs except penicillin and methicillin, all methods have low ME rates, below the FDA threshold of 3%. All methods had high error rates for penicillin (11.7% for `Mykrobe predictor`, 15.1% for disc, 16.0% for Phoenix). However, it is known that for penicillin, phenotyping methods may under-detect resistance (El Feghaly et al. 2012; Haveri et al. 2005; Kaase et al. 2008), and so the apparent high false positive rate is likely to be artefactual—i.e., under-detecting resistance by disc, Phoenix and nitrocefin might lead to an (incorrect) consensus susceptible call. Indeed it has been previously shown (Gordon et al. 2014) that for some samples with weak beta-lactamase activity (exhibited by a very slowly developing nitrocefin test), resistance was not detected by either disc or Phoenix. `Mykrobe predictor` had an acceptable false positive rate for methicillin of 0.0%, compared with disc (0.5%), and Phoenix (18.9%). This high false positive rate for Phoenix was unexpected; either both disc and Etest under-detect resistance (they are both diffusion methods and might have correlated errors), or these were indeed false calls from Phoenix.

The commercial software SeqSphere recently demonstrated resistance gene detection (Leopold et al. 2014) and susceptibility testing when for six drugs resistance was gene-based. The template (the SeqSphere equivalent of a gene panel) used was not publicly released but the authors were kind enough to provide us with it. I ran SeqSphere on the sequence data from the validation set St_{B1} (471 *S. aureus* isolates) in pipeline mode. I followed the methods in the original paper (Leopold et al. 2014) with the exception that the Velvet assembly was run by SeqSphere with the default parameters. Although used in that paper, erythromycin, and therefore clindamycin, were excluded from the comparison since all samples were called as re-



(a) Proportion of resistant *Staphylococcus* samples correctly identified as resistant. VME/FNs in red.

(b) Proportion of susceptible *Staphylococcus* samples correctly identified as susceptible. ME/FPs in red.

Figure 2.9: Comparison of Mykrobe predictor and SeqSphere antimicrobial resistance predictions in St_{B1} : Mykrobe predictor (orange), SeqSphere (dark blue) compared with consensus, with false negatives in red. SeqSphere reported resistance to 6 drugs but erythromycin and clindamycin were excluded from the comparison since all samples were called as Resistant by SeqSphere. METH, methicillin; GEN, gentamicin; MUP, mupirocin; VAN, vancomycin.

sistant. Since SeqSphere predicted all 471 samples to be resistant to erythromycin and clindamycin, I excluded these drugs. As can be seen in Figure 2.9, other results were broadly comparable to Mykrobe predictor and the phenotyping methods.

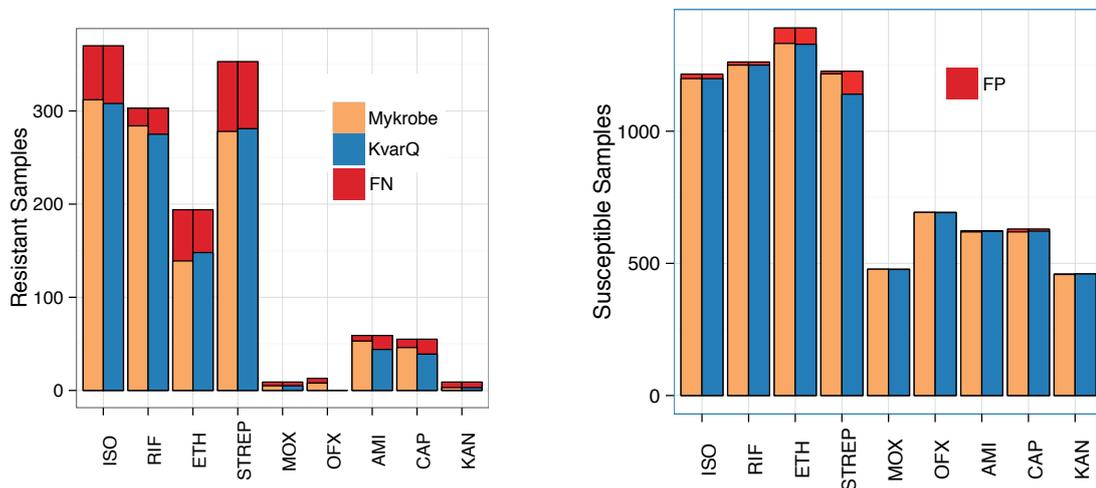
2.5.3.1 Detecting virulence elements in *S. aureus*

Antimicrobial resistance is not the only medically relevant phenotype that might be revealed by sequencing. *S. aureus* has a large number of virulence elements, which

might prove valuable to genotype. As an example, we considered Panton-Valentine Leukocidin (PVL), a cytotoxin that kills leukocytes and is associated with tissue necrosis (Lina et al. 1999). I incorporated tests for presence of the PVL genes lukPV-S and lukPV-K into `Mykrobe predictor` and applied them to sequence data from 67 *S. aureus* clinical isolates from an outbreak. The results were 100% concordant with PCR tests for the presence of these genes (23 negative and 44 positive). Mason et al. [submitted](#) have also evaluated `Mykrobe predictor` for the detection of a large panel of virulence elements showing high concordance with comparable tools.

2.5.4 `Mykrobe predictor` *M. tuberculosis* resistance predictions match commercial assays

Our understanding of the genetic basis for resistance in MTBC is incomplete. Common resistance mutations are on commercial line probe assays, and explain approximately 85-95% of observed resistance to the two primary first line drugs (isoniazid, rifampicin) (Abreu Maschmann et al. 2013; Chryssanthou et al. 2012; Rodwell et al. 2014). These assays have lower sensitivity for the third first-line drug (ethambutol) and second-line drugs (Miotto et al. 2012), and do not attempt to predict resistance for the fourth first-line drug (pyrazinamide), which is poorly understood. I built a panel of resistance mutations based on the Hain and AID line probe assays, with a small number of additional mutations from the literature (see Section 2.10.3.1 for details). For comparison with a method using a similar panel, but without minor calls, I also ran the `KvarQ` (version 0.12.3a1) (Steiner et al. 2014)



(a) Proportion of resistant *M. tuberculosis* samples correctly identified as resistant. VME/FNs in red.

(b) Proportion of susceptible *M. tuberculosis* samples correctly identified as susceptible. ME/FPs in red.

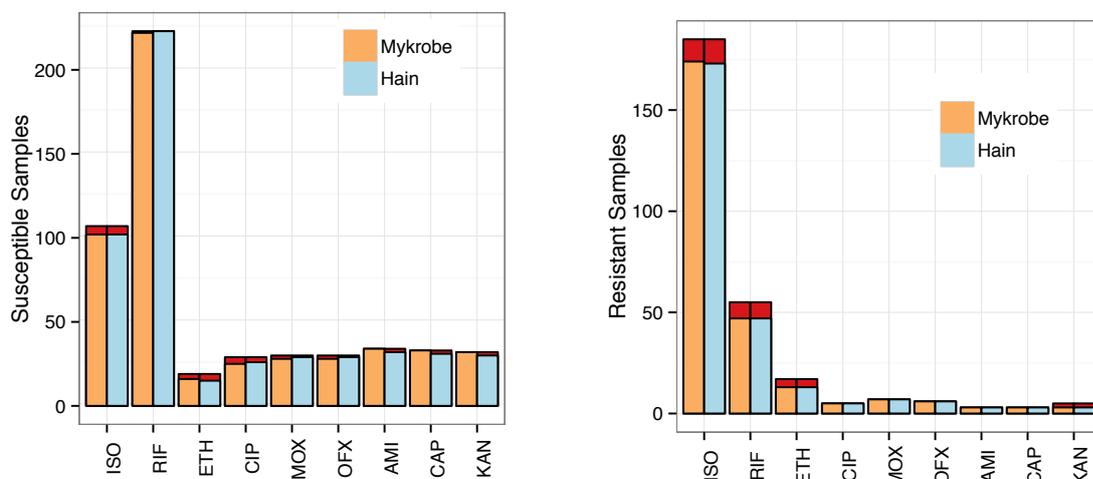
Figure 2.10: Comparison of Mykrobe predictor and KvarQ antimicrobial resistance predictions in $MTBC_A$: Mykrobe predictor (yellow) and KvarQ (light blue) compared with DST phenotype. ISO, isoniazid; RIF, rifampicin; ETH, ethambutol; STREP, streptomycin; MOX, moxifloxacin; OFX, ofloxacin; AMI, amikacin; CAP, capreomycin; KAN, kanamycin.

on the training and validation MTB fastq files using the provided ‘MTBC’ testsuite.

I used our training data set $MTBC_A$ to fit the frequency threshold, above which a resistance allele is modelled as causing phenotypic resistance (see Section 2.5.1). I chose an underlying frequency of 10% for the minor resistant model as this gave appropriately low false positive rates when comparing with phenotypes in the training set (Appendix Table A.5, Figure 2.7a).

Figure 2.10 shows the proportion of resistant and susceptible samples that were called correctly for each drug. As expected, for first-line drugs rifampicin, isoniazid,

2. RAPID ANTIBIOTIC RESISTANCE PREDICTIONS FROM HIGH THROUGHPUT GENOME SEQUENCE DATA



(a) Proportion of phenotypically resistant samples correctly identified as resistant by Mykrobe predictor (yellow) and Hain assay (blue) on data set $MTBC_{A1}$. Height of bars gives total resistant samples; and false positive calls are shaded red.

(b) Proportion of phenotypically susceptible samples correctly identified as susceptible by Mykrobe predictor (yellow) and Hain assay (blue) on data set $MTBC_{A1}$. Height of bars gives total susceptible samples; and false positive calls are shaded red.

Figure 2.11: Comparison of Mykrobe predictor and Hain antimicrobial resistance predictions in $MTBC_A$: Mykrobe predictor (yellow) and KvarQ (light blue) compared with DST phenotype. Since Mykrobe predictor was run using a nearly identical catalogue, similar results are expected. ISO, isoniazid; RIF, rifampicin; ETH, ethambutol; STREP, streptomycin; MOX, moxifloxacin; OFX, ofloxacin; AMI, amikacin; CAP, capreomycin; KAN, kanamycin.

Table 2.2: Results for Mykrobe predictor on the *M. tuberculosis* validation set (MTBC_B). Resistance prediction results for Mykrobe predictor on the *M. tuberculosis* validation set MTBC_B compared against the consensus phenotype.

FN: False negative calls. R: total number of resistant samples. FP: false positives. S: total number of susceptible samples. VME: very major error rate (false negative rate). ME: major error rate (false positive rate). PPV: positive predictive value. NPV: negative predictive value. N/A: Not Applicable.

Error rates shown with 95% CI calculated by Clopper-Pearson; FN/FP rate only shown where number of resistant/susceptible samples > 10.

RIF, rifampicin; ETH, ethambutol; STREP, streptomycin; MOX, moxifloxacin; OFX, ofloxacin; AMI, amikacin; CAP, capreomycin; KAN, kanamycin.

Drug	FN(R)	FP(S)	VME (%)	ME (%)	PPV (%)	NPV (%)
ISO	58 (370)	17 (1216)	15.7 (12.1-19.8)	1.4 (0.8-2.2)	94.8 (91.9-97.0)	95.4 (94.1-96.5)
STREP	75 (353)	9 (1227)	21.2 (17.1-25.9)	0.7 (0.3-1.4)	96.9 (94.1-98.6)	94.2 (92.8-95.4)
RIF	19 (303)	12 (1262)	6.3 (3.8-9.6)	1.0 (0.5-1.7)	95.9 (93.0-97.9)	98.5 (97.7-99.1)
ETH	55 (194)	59 (1391)	28.4 (22.1-35.2)	4.2 (3.2-5.4)	70.2 (63.3-76.5)	96.0 (94.9-97.0)
AMI	6 (59)	6 (623)	10.2 (3.8-20.8)	1.0 (0.4-2.1)	89.8 (79.2-96.2)	99.0 (97.9-99.6)
CAP	9 (55)	13 (630)	16.4 (7.8-28.8)	2.1 (1.1-3.5)	78.0 (65.3-87.7)	98.6 (97.3-99.3)
OFX	5 (13)	0 (693)	38.5 (13.9-68.4)	0.0 (0-0.5)	100.0 (63.1-100)	99.3 (98.3-99.8)
MOX	4 (9)	0 (478)	44.4 (13.7-78.8)	0.0 (0-0.8)	100.0 (47.8-100)	99.2 (97.9-99.8)
KAN	6 (9)	2 (460)	66.7 (29.9-92.5)	0.4 (0.1-1.6)	60.0 (14.7-94.7)	98.7 (97.2-99.5)

and ethambutol, the two methods (Mykrobe predictor and KvarQ) have similar power to detect resistance (93.7%, 84.3%, 71.6% versus 90.8%, 83.2%, 76.3%), and similar false positive rates (1.0%, 1.4%, 4.2% versus 1.0%, 1.4%, 4.5%)—in line with expected performance of the Hain assay. Fewer samples were phenotyped for second-line drugs, but Mykrobe predictor had noticeably higher sensitivity for amikacin and capreomycin (89.8% and 83.6% respectively) than KvarQ (74.6% and 70.9%). This is possibly due to Mykrobe predictor’s higher sensitivity to low frequency alleles as I will discuss in Section 2.6. There were very few false calls for

2. RAPID ANTIBIOTIC RESISTANCE PREDICTIONS FROM HIGH THROUGHPUT GENOME SEQUENCE DATA

second-line drugs for either method, except for a high (7%) error rate for KvarQ for streptomycin. See Table 2.2 and Table A.5) for full results.

Since the underlying panel is almost identical, we expect **Mykrobe predictor** to perform equivalently to the HAIN test. Comparing on $MTBC_A$, **Mykrobe predictor** in and HAIN have similar power to detect resistance (85.5%, 94.1%, 76.5% versus 85.5%, 93.5%, 76.5%) for first-line drugs rifampicin, isoniazid and ethambutol respectively—see Figure 2.11.

The sensitivity and specificity of **Mykrobe predictor** is driven primarily by the knowledge of the catalogue of resistance variants, which can easily be augmented. **Mykrobe predictor** has also been run with an extended catalogue from Walker et al. 2015, leading to an increase of sensitivity of 12.3% and a decrease in specificity of 0.9% across all drugs when run on $MTBC_A$ and $MTBC_B$ when compared with the “Hain” panel shown above. These values are likely to improve further as we gain a more comprehensive understanding of the mechanisms of resistance in *M. tuberculosis*.

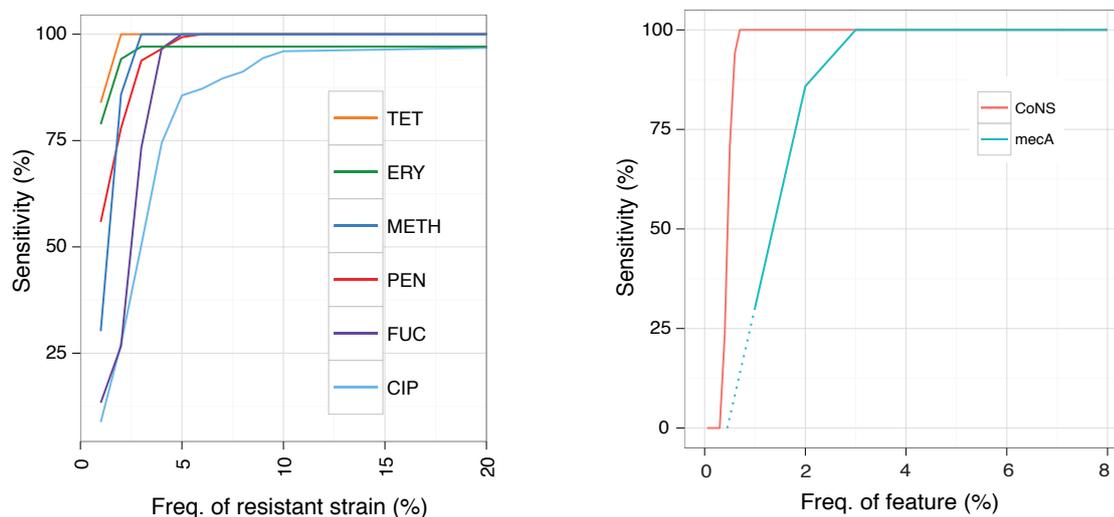
2.6 Power to detect minor populations

It is possible for patients to have mixed infection involving two or more strains of the pathogen simultaneously (Tarashi et al. 2017). These strains may have different susceptibility patterns, and in *M. tuberculosis* this has been identified as a major cause of treatment failure (Cohen et al. 2012). Although the true rate of mixed infection is difficult to estimate (as culturing of samples often removes this diversity), and their impact on clinical outcomes is unclear, one of our goals with `Mykrobe predictor` was to be able to detect low frequency genotypes which may have come from a minor strain.

In order to determine the power of `Mykrobe predictor` to detect minor resistant populations in *Staphylococcus*, I took 450 samples from the *S. aureus* data set St_{B1} from the subset of samples which had sequencing depth of at least $100\times$ coverage across the genome, and for each sample subsampled the reads to precisely $100\times$. I then took 1000 random pairs of samples from this set, and for each pair, I combined subsets of their reads so as to create 27 different mixtures with ratios ranging from 1:99 to 99:1. I ran `Mykrobe predictor` on these mixtures to determine the frequency at which rare SNPs/genes were detected, and to confirm that this sensitivity did not cause false positive predictions of phenotypic resistance (Simulation 1).

I show in Figure 2.12a `Mykrobe predictor`'s power to detect alleles at low frequency. Our method has greatest power to detect resistance genes which lie on multi-copy plasmids, with detection power reaching 94% and 100% by the time the population frequency is 2% for tetracycline and erythromycin, respectively. For

2. RAPID ANTIBIOTIC RESISTANCE PREDICTIONS FROM HIGH THROUGHPUT GENOME SEQUENCE DATA



(a) Power to detect minor resistant alleles (b) Power to detect low-frequency CoNS

Figure 2.12: Power to detect low frequency features. a) Power to detect minor resistant alleles in 27,000 in silico mixtures created by taking 1,000 pairs of *Staphylococcus* samples and mixing each pair in 27 different ratios. Power is greatest for the drugs where resistance genes reside on multi-copy plasmids, namely erythromycin and tetracycline. b) Power to detect low-frequency coagulase-negative species (red, simulation 1, N=540, described above) is consistently higher than power to detect *mecA* (blue, simulation 2, N=27,000, frequencies down to 1% only due to large sample numbers; dotted lines extrapolate linearly from points at 1 and 2%), which causes methicillin resistance in *Staphylococcus*. Thus, the risk of detecting *mecA*, but not detecting the coagulase-negative species it comes from, is limited. Tet=tetracycline; Ery=erythromycin; Meth=methicillin; Pen=penicillin; Fuc=fusidic acid; Cip=ciprofloxacin.

other drugs, detection power exceeds 90% once the subpopulation exceeds 8% frequency. This ability to genotype low frequency alleles did not come at the cost of false positive phenotypic resistance predictions. Apart from the few false positive calls that Mykrobe predictor made in pure samples (see Table 2.1), there were no additional false positives out of 189,000 calls (27,000 mixtures x 7 drugs).

Since minor frequency alleles can come from a contaminating species, one goal for `Mykrobe predictor` was to avoid the combination of detecting *mecA* from *mecA*-containing-CoNS while failing to detect the CoNS species itself—thus causing a miscall of MRSA. To test for this, I took 18 CoNS samples from the St_{B2} data set (9 *S. epidermidis* and 9 *S. haemolyticus*), and 9 *S. aureus* samples from St_{B1} , and associated random pairs of samples from these sets (always one *S. aureus* and one CoNS, mixture pairs below). I then made in silico mixtures of these pairs with the CoNS samples at 30 frequencies ranging from 0.5% to 20% (Simulation 2). I ran `Mykrobe predictor` on all 540 mixtures to determine sensitivity to detect CoNS at each frequency. Accessions can be found in Appendix Section [A.1](#).

I used the simulated mixtures of *S. aureus* and CoNS, to estimate the discovery power for low frequency CoNS species, and compared that with low frequency *mecA* in Simulation 1—results are shown in Figure [2.12b](#). I was able to confirm that in these mixtures such miscalls were indeed unlikely. At 1% frequency, the estimated power to detect the presence of a CoNS species was 100% (red curve), but power to detect *mecA* was 33% (blue curve). Above 3% frequency, power to detect each was 100%.

2.6.1 Heteroresistant *S. aureus*

Our strong prior expectation was that there would be limited within-sample diversity in our *S. aureus* St_{A1} and St_{B1} samples, due to blood culture followed by storage processes, and removal of contaminated samples. `Mykrobe predictor` con-

2. RAPID ANTIBIOTIC RESISTANCE PREDICTIONS FROM HIGH THROUGHPUT GENOME SEQUENCE DATA

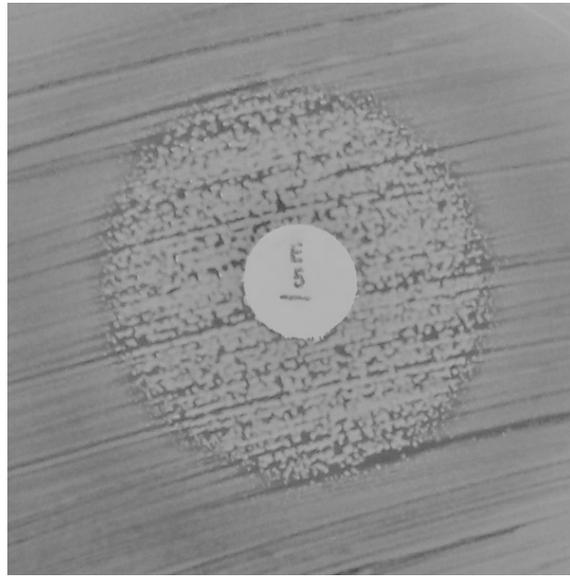


Figure 2.13: Photograph of BSAC disc test showing hetero-resistant phenotype: Seen on re-running Erythromycin disc test on a sample (accession: ERS398183) where Mykrobe predictor had called a false positive (resistant) that neither disc nor Phoenix had called.

firmed this expectation and made only 6 minor calls out of 11592 (12 drugs times 966 training and validation samples). However, we noted with interest that for the four samples where Mykrobe predictor made false positive (major) resistant calls that were not made by Disc or Phoenix, re-running the disc test resulted in contradictory results. Two changed to resistant (ciprofloxacin, erythromycin) and two produced heteroresistant phenotypes (erythromycin, tetracycline, see Figure 2.13). This behaviour is consistent with a disc test presented with mixed strains or with variable plasmid loss (which would explain the 3 erythromycin/tetracycline results). All 4 samples had low levels of chromosomal diversity (between 12 and 25 “heterozygous” SNPs), ruling out contamination, unless by a closely related strain.

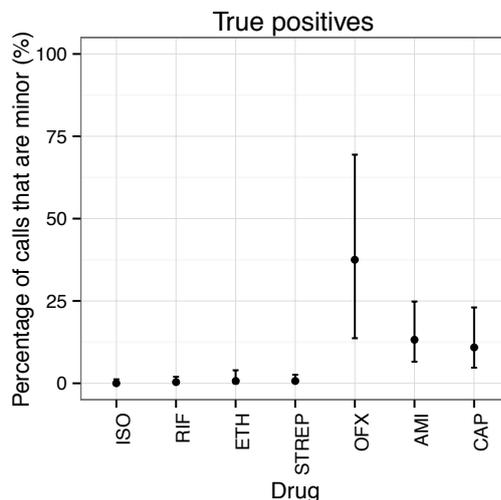


Figure 2.14: Percentage of true positive resistant calls in *M. tuberculosis* validation set due to minor alleles: Confidence intervals are calculated using the Clopper–Pearson interval. Drugs with < 10 resistant samples were excluded to avoid overly large confidence intervals. For aminoglycosides and quinolones, minor populations explain between 11–38% of true positive resistance predictions. ISO=isoniazid, RIF=rifampicin, ETH=ethambutol, STREP=streptomycin, OFX=ofloxacin, AMI=amikacin, and CAP=capreomycin.

2.6.2 Minor alleles increase power to distinguish XDR from MDR TB

Mykrobe predictor predicted that 56 samples were phenotypically resistant due to minor alleles, across the 9 drugs and 1609 samples in the $MTBC_B$ validation set (Table 2.2). Whole genome analysis of these samples found a median of 16 heterozygous sites per sample, consistent with mixed infections (local transmission or in-host evolution) (Walker et al. 2015), although we cannot exclude the possibility of contamination with a closely related strain. I show in Figure 2.14 the proportion

2. RAPID ANTIBIOTIC RESISTANCE PREDICTIONS FROM HIGH THROUGHPUT GENOME SEQUENCE DATA

of true positives due to minor resistant calls in the validation set, showing a clear demarcation between first- and second-line drugs. Power to predict phenotypic resistance to second-line drugs amikacin and capreomycin was significantly increased by the detection of minor populations: from 74.5% to 83.6% for capreomycin and 78.0% to 89.8% for amikacin. In addition, although here the numbers were small ($N=13$), power increased from 38.5% to 61.5% for ofloxacin. This increase in sensitivity did not come at the price of a loss of specificity. However, the effect was only seen in our validation set that had the majority of XDR (Extensively Drug-Resistant) samples in our data—these minor alleles were mostly (27/39) found in validation samples from Uzbekistan.

2.6.2.1 Slow growth *rpoB* SNPs and limitations of the gold standard

There were 12 false positive rifampicin resistance calls which are, all major calls in the gene *rpoB*. Three were L452P mutations, known to cause only low level resistance (Ohno et al. 1996; Sun et al. 2012). However, on examination, we found the remaining 9 calls may reflect limitations of gold standard culture-based phenotyping. These were either mutations known to slow growth (S450L ($n=3$), S450W ($n=1$), Q432 ($n=1$)) (Gagneux et al. 2006; Mariam et al. 2004; Song et al. 2014), or overlap a 10bp deletion affecting growth (Q429H ($n=1$), L430P ($n=3$)) (Malshetty et al. 2010)). Since the proportion method used in *M. tuberculosis* susceptibility tests fundamentally measures growth rate as a proxy for resistance (Canetti et al. 1969), slow growth can lead to false susceptible DST results for samples with these mutations (Van Deun et al. 2009). Thus, 75% of the false positive rifampicin calls

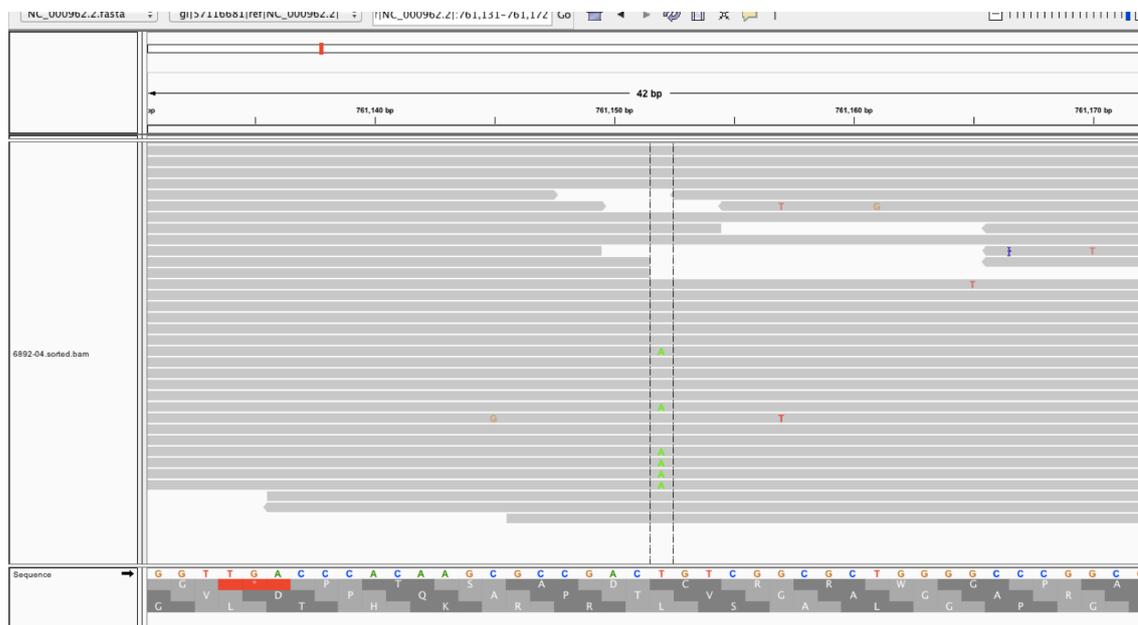


Figure 2.15: A pile-up showing a low frequency mutation at Leu-449. Low frequency (< 10%) variation was found in *rpoB*-Leu-449 in (11%) 177/1609 samples from $MTBC_B$. This mutation was not seen in phenotypically resistant samples, or at major frequency in any samples in $MTBC_A$ or $MTBC_B$.

from Mykrobe predictor may actually be resistant in vivo, but called susceptible by DST due to the nature of the test. Indeed, there is evidence that such *rpoB* mutations may be associated with poor outcomes (Williamson et al. 2012).

2.6.3 Very low frequency *rpoB* variation in *M. tuberculosis*

In the *M. tuberculosis* validation set, I found there were 177 (11%) samples rifampicin resistance alleles between 5-10% frequency (below our frequency threshold of 10% set by the training set). All of them looked like true low frequency muta-

2. RAPID ANTIBIOTIC RESISTANCE PREDICTIONS FROM HIGH THROUGHPUT GENOME SEQUENCE DATA

tions (see Figure 2.15), but all corresponded to phenotypic susceptibility. Of these, 146 (79%) were mutations at amino acid Leu-449, one amino acid away from the most common resistance mutation Ser-450, which by contrast was found in 250 phenotypically resistant samples. Leu-449 was never seen as a major allele, or in a phenotypically resistant sample. These samples were from Uzbekistan and South Africa. It would be hard to explain contamination of 146 samples with either one strain or multiple Leu-449 strains. Furthermore, a similar phenomenon can be seen in the training set, which was sequenced in Oxford (15 samples with Leu-449 out of 44 samples with low frequency alleles). Collaborators (personal correspondence, S. Niemann) found Leu-449 low frequency mutations in 830 MDR samples, but did not find it select for amongst 60 mono-rifampicin-resistant mutants.

2.7 Nanopore sequencing of *Staphylococcus*

2.7.1 Resistance calling from ONT data

Having evaluated performance on two species using Illumina data, we tested `Mykrobe predictor` on data from the Oxford Nanopore Technologies (ONT) MinION single-molecule sequencing machine. Since the per-base error rate was high (between 10-30% per base at the time of sequencing, depending on whether the molecule has been sequenced in one or two directions, termed “1d reads” and “2d reads” respectively), I modified `Mykrobe predictor` to expect an error rate of 10% and to ignore the quality score for ONT data. We took a multi-drug-resistant *S. aureus* isolate from a clinical sample taken in 2014, and sequenced its genome with both the Illumina MiSeq and a ONT MinION (see Section 2.10.6 for details), and then ran `Mykrobe predictor`. The MiSeq run took 24 hours and produced, after cutting reads at bases with quality below 10, 368x of 122bp reads. The MinION run took 24 hours and generated 39x of 2d reads, with min/mean/max length 113bp/4.7kb/48kb. In both cases, `Mykrobe predictor` correctly predicted that the sample was resistant to penicillin, methicillin, gentamicin, trimethoprim, erythromycin, ciprofloxacin, and clindamycin, and that it was susceptible to fusidic acid, rifampicin, tetracycline, vancomycin, and mupirocin. All of the resistance calls were due to detection of genes, except for ciprofloxacin where a S→L mutation at position 84 in the gene *gyr* was detected. No false positive resistance SNPs were called. Furthermore, by subsampling the MinION output by time from the start of

2. RAPID ANTIBIOTIC RESISTANCE PREDICTIONS FROM HIGH THROUGHPUT GENOME SEQUENCE DATA

sequencing, we showed that these results could have been obtained with just seven hours of sequencing.

Table 2.3: Performance and feature comparison of **Mykrobe predictor**, **SeqSphere**, and **KvarQ** software. We show elapsed time for one sample on a laptop (Macbook Air with 8GB RAM) and a server (Dell PowerEdge R820 with 32 cores, 1Tb RAM), and then for the entire *S. aureus* and *M. tuberculosis* validation sets. We ran **KvarQ** on one thread for ease of parallelization and comparison, as recommended by the authors. However we ran **SeqSphere** on four threads because to use one would have taken a prohibitively long time. Sa=*Staphylococcus aureus*, Mtb=*Mycobacterium tuberculosis*. MTBC=*Mycobacterium tuberculosis* complex.

	M predictor (Sa)	SeqSphere (Sa)	M predictor (Mtb)	KvarQ (Mtb)
RAM	100Mb	8Gb	100Mb	30Mb
Time/sample on laptop	1.5 mins	-	2.75 mins	40 mins
Time/sample on server	44 sec	19 mins	47s	23 mins
CPU time (Mtb validation set)	-	-	1 day	30 days
CPU time (Sa validation set)	5.8 hrs	6.2 days	-	-
Speciation of clinical samples	Yes	No	Yes	MTBC only

2.8 Software performance and usability

The **Mykrobe predictor** software is freely available (open-source) at www.github.com/iqbal-lab/Mykrobe-predictor. **Mykrobe predictor** is implemented in a Linux command-line version and desktop “drag-and-drop” applications for 64-bit Windows and Mac OS X (see screenshots in Appendix Figures A.1 – A.3).

Mykrobe predictor memory use depends only on the size of the variant catalogue, and not on the input data. For our *S. aureus* and *M. tuberculosis* catalogues discussed above this is comparable with that typically used by a web browser, such

2. RAPID ANTIBIOTIC RESISTANCE PREDICTIONS FROM HIGH THROUGHPUT GENOME SEQUENCE DATA

as Chrome with multiple tabs open ($\approx 100\text{MB}$), and CPU requirements are low ($\approx 1\text{min}$ for a $100\times$ data set). `Mykrobe predictor` has been run on a Google Nexus 10 tablet, a Samsung Core Duos phone, and a Raspberry Pi Model B. I give in Table 2.3 some performance statistics for *S. aureus* (abbreviated Sa) and *M. tuberculosis* (abbreviated Mtb), with comparison data from alternative tools `SeqSphere` (Leopold et al. 2014), which uses whole genome assembly, and `KvarQ` (Steiner et al. 2014), which uses k -mer detection.

`Mykrobe predictor` has also been extensively validated against two other closed-source *S. aureus* DST tools: `Typewriter` and `Genefinder` (Mason et al. submitted). The three tools have high concordance with each other and no significant difference in sensitivity or specificity. However, `Mykrobe predictor` had significantly better computational performance (at least $2\times$ faster) and also performed species and contamination identification, which `Typewriter` and `Genefinder` did not.

2.9 Discussion

Rapid determination of antimicrobial resistance is of critical importance to patient care for many serious bacterial infections and has wider implications for determination of treatment protocols and national surveillance. We have developed a software application, extensible to many bacterial species, called **Mykrobe predictor**, which can identify species, resistance profile, and other genomic features, such as virulence elements and phylogenetic lineage defining SNPs, within 3 minutes on a standard laptop. Above I described the application of the tool, to *S. aureus* and *M. tuberculosis* DST, and validated them extensively against clinical gold standards. Sensitivity and specificity on *S. aureus* are above 99% and are comparable with, or better than, phenotyping methods (BSAC disc test, Phoenix). For *M. tuberculosis*, specificity is high (98.5%) and sensitivity of 82.6% matches the line probe assays from which our resistance panel was constructed, but it still is below that of the gold standard of DST based on solid (Lowenstein-Jensen) culture.

We designed **Mykrobe predictor** to be easy to use, highly performant, easy to extend, and accurate. Consequently, it has been included in Oxford University Hospitals' routine WGS analysis pipeline for *M. tuberculosis* and *S. aureus*. It has also been included in Public Health England's National Mycobacterial Reference Service where, between 20th April 2015 and 31st March 2016, it analysed 1965 *M. tuberculosis* samples, which were processed both by conventional testing and WGS (Quan et al. 2017). Here, **Mykrobe predictor** was used exclusively for species identification where it had a 99.5% concordance with the organism identified from

2. RAPID ANTIBIOTIC RESISTANCE PREDICTIONS FROM HIGH THROUGHPUT GENOME SEQUENCE DATA

the routine lab for *M. tuberculosis* samples.

There are some limitations to the current implementation of **Mykrobe predictor**. Our sensitivity for *M. tuberculosis* is low (82.6% across all drugs) compared with traditional DST, and completely excludes the first-line drug pyrazinamide because known mutations are poorly predictive. This issue is shared by all molecular assays. This can be resolved by large-scale sequencing and phenotyping studies, such as Walker et al's study of 10,225 *M. tuberculosis* isolates, which demonstrates sensitivity greater than the WHO-endorsed in-silico assays MTB/RIF Xpert, MTBDRplus, and MTBDRsl assays (Walker et al. [in review](#)). Genome-wide association studies (Earle et al. [2016](#)), directed evolution experiments, or analysis of protein structure (Orencia et al. [2001](#)), also show promise in improving our understanding of how phenotype can be predicted from genotype.

There is now a plethora of AMR prediction tools including, ABRicate (Seeman [2017](#)), ARIBA (Hunt et al. [2017](#)), Antibiotic Resistance Gene-ANNOTation (Gupta et al. [2014](#)), CASTB (Iwai et al. [2015](#)), KvarQ (Steiner et al. [2014](#)), ARGs-OAP (Yang et al. [2016](#)), RAST (Davis et al. [2016](#)), PhyResSe (Feuerriegel et al. [2015](#)), ResFinder (Zankari et al. [2012](#)), RGI(McArthur et al. [2013](#)), SSTAR (Man et al. [2016](#)), SRST2 (Inouye et al. [2014](#)), and TB Profiler (Coll et al. [2015](#)). All of these tools have their own source of gene and variant catalogue, which they use to infer resistance, resulting in a disperse ecosystem of tools, which provide very similar, but not identical results. Current tools (including **Mykrobe predictor**) focus on detecting published and characterised antimicrobial resistance gene sequences. As new resistance-causing mutations and genes are determined, **Mykrobe predictor**

and other software-based prediction tools can be easily updated and tested with automatic workflows, and unlike a line-probe assay or the automated Xpert-Mtb/Rif (Cepheid) assay, they are unaffected by number of resistance-causing mutations in the panel. However, they do not currently aid in the detection of novel threats, nor automatically update when new threats emerge, such as MCR-1 (Liu et al. 2016). The next major steps to improve this ecosystem of tools are to create are to i) create tools that are community-driven and decentralised so they can be quickly updated when new threats, such as MCR-1 (Liu et al. 2016), emerge; and ii) further clinical testing, and then obtaining regulatory approval. These two steps are possibly mutually exclusive, as it is not yet clear if obtaining regulatory approval will possibly require an unchanging catalogue on which to test extensively, or if each variant or gene can be approved individually.

Despite the explosion in available tools, `Mykrobe predictor` is still the only tool that explicitly tests for minor alleles. In our study, `Mykrobe predictor` classified 3.4%/4.9% of our *M. tuberculosis* training/validation samples as having minor resistant populations, with median frequency 6.8%/9.2%, respectively. However, in order to match the results of the in vitro gold standard, we only predicted a resistant phenotype for alleles above 10% frequency. It remains an open question as to whether the lower frequency alleles have a bearing on patient outcome, despite generally failing to cause in vitro resistance. In our validation set, which contained the bulk of our XDR samples, we found that minor alleles improved power to predict resistance by over 12% for second line drugs (amikacin, capreomycin, and ofloxacin, see Figure 2.14), though this should be replicated with a larger data set of XDR

2. RAPID ANTIBIOTIC RESISTANCE PREDICTIONS FROM HIGH THROUGHPUT GENOME SEQUENCE DATA

isolates. We also found low frequency variation ($< 10\%$) in *rpoB* in 177 validation set samples, 79% of which were at the same location *rpoB*-449. These low frequency mutations were not correlated with phenotypic resistance. Longitudinal patient data would be required to determine if low frequency alleles (below the 10% frequency threshold we set) rise in frequency and cause resistance later in the course of an infection, but are below the detection threshold of the phenotypic-based tests at the earlier time point.

We detail in Figure 1.2 how the timelines for a proposed sequencing workflow would compare with the clinical protocols implemented at Oxford University Hospitals (OUH) Clinical Laboratory. Using an Illumina MiSeq 16.5 hour run, our proposed sequencing workflow would provide a full set of predictions for all drugs at 36 hours, ahead of OUH by 12 hours. At other institutions, one might use MALDI-TOF or alternate rapid methods to identify the species or even methicillin resistance directly on blood culture (Bhowmick et al. 2013), but these cannot give the full susceptibility profile (nor any information on ancestry, epidemiology, or virulence).

The acceleration provided by sequencing is even greater for *M. tuberculosis*, where the standard process is to run first-line drug tests and, if necessary, the second-line tests afterwards, which can take months (see Figure 1.3). By contrast, for the sequencing workflow, most clinical isolates become MGIT positive within two weeks. Thus, if sample preparation and sequencing is completed within two days of a positive MGIT result (Votintseva et al. 2015), one can get results in two weeks. This is a gain of somewhere between 5 and 17 weeks compared with gold standard DST, depending on whether the sample is resistant to first-line drugs. However,

further improvements can be made. One possible improvement is to sequence directly from sputum, removing the the 2 week MGIT culture step. Another is to use nanopore sequencing, which has the potential to reduce the ≈ 24 sequencing step. Direct sequencing is expected to lead to lower yield, more diverse and more contaminated samples, but could dramatically reduce analysis time. Nanopore sequencing is expected to lead to lower yield and more error-prone data, but allows for “random access” sequencing, where there is no requirement to batch samples and sequencing can be stopped when sufficient data have been collected. Both of these approaches introduce challenges to bioinformatic analysis, which I will discuss in more detail in the following chapter.

2.10 Extended methods

2.10.1 K-mer size

A k -mer size of 15 and 21 for *Staphylococcus aureus* and *M. tuberculosis* respectively was chosen by using KmerGenie (Chikhi et al. 2013a) to create abundance histograms for many values of k from exemplar data sets from the training set. A suitable k -mer size was chosen by choosing the smallest k which had few repeats within the exemplar *Staphylococcus aureus* and *M. tuberculosis* data sets.

2.10.2 *S. aureus* data sets

In order to train and validate Mykrobe predictor for *S. aureus* I used a collection of 1,423 *S. aureus* isolates (see Figure 2.16 and Table 2.4). I removed six samples from the St_{A1} and nine samples from St_{B1} , as BLAST of the assembly contigs confirmed presence of non-staphylococcal contamination. I also removed eleven samples from St_{B1} , which had greater than 40 confident heterozygous SNPs as called by the Cortex variation assembler (independent workflow, $k=31$, ploidy=2, automatic error cleaning, using “bubble caller” calling algorithm). I used this threshold of 40 heterozygous sites to determine whether a sample was contaminated by an unrelated strain—less than that we considered a conceivable level of in-host diversity. In total, I removed six samples from the training data set and twenty samples from the validation set that were not confidently identified as isolates.

Initial phenotyping of the training and validation sets was described in detail in

Gordon et al. (Gordon et al. 2014). The training set was phenotyped using either the Vitek automated system (bioMerieux) or the Stokes method disc diffusion (Gosden et al. 1998), whereas all validation samples were phenotyped using two methods: a British Society for Antimicrobial Chemotherapy (BSAC) disc test (Andrews 2001) and the Phoenix Automated Microbiology System (BD Biosciences, Sparks, MD, USA). For trimethoprim only disc testing was performed (Table 2.4). E-tests were performed when disc and Phoenix were discrepant for all drugs except for penicillin, where a nitrocefin test was performed (Gordon et al. 2014).

In addition, the following additional analyses were run that were not included in the Gordon et al. 2014 publication: all samples where there was discordance between Phoenix and disc for any drug in Gordon et al. 2014 were rerun on Phoenix (for all drugs), and previous results from Gordon et al. (Gordon et al. 2014) were discarded.

We defined a “consensus” phenotype to be the consensus of disc/Phoenix where they agreed, or the result of an Etest and/or nitrocefin (for penicillin) where disc and Phoenix were discrepant. This allowed us to estimate error rates for disc and Phoenix, as well as for *Mykrobe predictor*. Samples were sequenced on the Illumina HiSeq 2000 platform, with mean read length (after cutting reads at bases with quality score below 10) of 87 bp and mean depth of 87 [sic], as described in Gordon et al. 2014.

2. RAPID ANTIBIOTIC RESISTANCE PREDICTIONS FROM HIGH THROUGHPUT GENOME SEQUENCE DATA

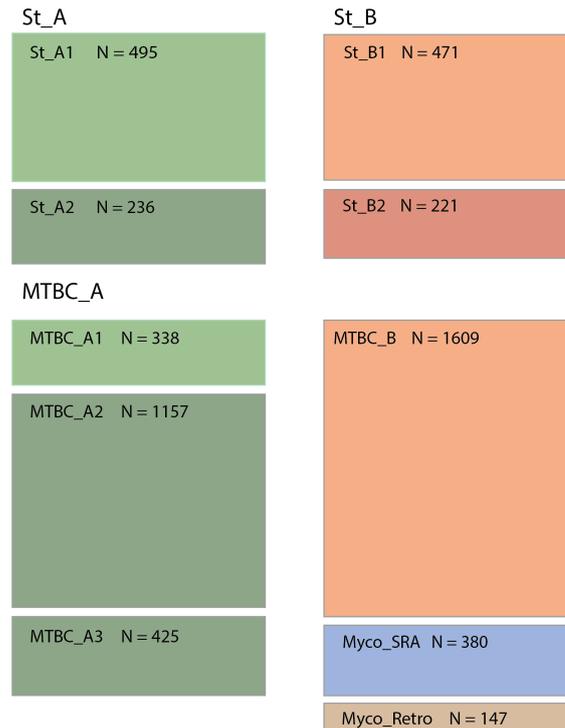


Figure 2.16: Overview of data sets used for training/validation of species identification and resistance prediction. Left hand column shows data sets used for training, and right hand column those used for validation. Abbreviations used are St=Staphylococcus, Myco=Mycobacteria, and MTBC=*M. tuberculosis complex*. See Table 2.4 for details on what phenotype and metadata these data sets each have.

S. aureus species training: $St_A = St_{A1} + St_{A2}$,

S. aureus species validation: $St_B = St_{B1} + St_{B2}$,

S. aureus resistance training: St_{A1} ,

S. aureus resistance validation: St_{B1} .

M. tuberculosis species training: $MTBC_{A1} + MYCO_{SRA}$,

M. tuberculosis species validation: $MTBC_{A2} + MYCO_{RETRO}$,

M. tuberculosis resistance training: $MTBC_{A1} + MTBC_{A2} + MTBC_{A3}$,

M. tuberculosis resistance validation: $MTBC_B$.

Table 2.4: Summary of metadata for each data set used. See Figure 2.16 for how these data sets were combined for different analyses.

Set	Species truth	Phenotype information
St_{A1}	Mapping to reference	Vitek or Disc
St_{A2}	Mapping to reference or SRA metadata	None
St_{B1}	Mapping to reference	Disc+Phoenix
St_{B2}	Mapping to reference or SRA metadata	None
St_{BVL}	NaN	PCR for PVL
$MTBC_{A1}$	Hain	Traditional DST
$MTBC_{A2}$	Hain	Traditional DST
$MTBC_{A3}$	None	Traditional DST
$MTBC_B$	None	Traditional DST
	SRA metadata	None
$MYCO_{RETRO}$	PCR	None

2.10.2.1 Building the *Staphylococcus aureus* phylogeny

A conservative set of SNPs were called for each sample in the training set (St_{A1}) and validation set (St_{B1}) by mapping reads to the MRSA252 reference genome with `stampy` (version 1.0.17) (Lunter et al. 2011), and then calling variants with `samtools` (Li et al. 2009) (version 0.1.18). SNPs with less than five reads' support, or without at least one read on each strand, were filtered, as were multiallelic SNPs, SNPs with at least 5 reads on both alleles, and SNPs in repetitive regions. A phylogeny was then constructed using `RAxML` (Paradis et al. 2004) with the following command-line:

```
RAxML-7.7.6/raxmlHPC-PTHREADS-SSE3 -s phylipFile
-n outputPrefix -m GTRCAT -p 12345 -c 1 -T 2
```

-D ON -f c -F ON -V.

2.10.2.2 Resistance catalogues

The set of genetic elements we considered is broadly described in Gordon et al. 2014. From the variant catalogue, I made the following changes: B434N in *fusA* was changed to D434N (a typo in the original publication). Q456K in *rpoB* was removed as Q was not the amino acid at position 456 of the referenced *rpoB* gene. I added *rpoB* N474K, described by Villar et al. (Villar et al. 2011). I also excluded any variants which were reported in the literature to modify the minimum inhibitory concentration (MIC), but not cause resistance in isolation as we did not have MIC phenotypes with which to compare.

For resistance genes I took all versions/alleles of the gene from NCBI that were not explicitly annotated as existing in a susceptible strain and did not have stop codons.

2.10.3 *Mycobacterium tuberculosis* data sets

2.10.3.1 Building the *M. tuberculosis* phylogeny

We used the underlying phylogeny of samples in sets $MTBC_A$ and $MTBC_B$, which was constructed using RAxML (version 8.0.5) using a (General Time Reversible-CAT) GTRCAT model (Paradis et al. 2004). For this study, we combined this tree with data set membership and phenotypic resistance metadata using the Analyses of Phylogenetics and Evolution (APE) package (Paradis et al. 2004)

to produce Figure 2.3

2.10.3.2 Resistance catalogues

We used a catalogue of MTBC resistance variants from the HAIN (Chryssanthou et al. 2012), Cepheid (Marlowe et al. 2011), and AID (Ritter et al. 2014) assays supplemented by others from the literature (Feuerriegel et al. 2012; Plinke et al. 2010; Zaunbrecher et al. 2009).

2.10.4 NTM species

The NTM species included in our panel were: *M. abscessus*, *M. africanum*, *M. aromaticivorans*, *M. avium*, *M. bovis*, *M. branderi*, *M. caprae*, *M. chelonae*, *M. chlorophenolicum*, *M. chubuense*, *M. colombiense*, *M. crocinum*, *M. flavescens*, *M. fluoranthenvivorans*, *M. fortuitum*, *M. gilvum*, *M. gordonae*, *M. hodleri*, *M. interjectum*, *M. intracellulare*, *M. kansasii*, *M. lentiflavum*, *M. leprae*, *M. malmoense*, *M. marinum*, *M. mucogenicum*, *M. pallens*, *M. peregrinum*, *M. phage*, *M. pyrenivorans*, *M. rufum*, *M. rutilum*, *M. scrofulaceum*, *M. senegalense*, *M. smegmatis*, *M. sphagni*, *M. szulgai*, *M. triplex*, *M. tuberculosis*, *M. tusciae*, *M. ulcerans*, *M. vaccae*, and *M. xenopi*.

2.10.5 Resistance calling at genes

The values for K_G used in Section 2.4.2 based on empirical diversity in our training set were:

$$K_G = 0.3 \quad \text{for } blaZ$$

$$K_G = 0.6 \quad \text{for } fusB, fusC$$

$$K_G = 0.8 \quad \text{otherwise.}$$

The minimum frequency thresholds for each drug were: erythromycin: 0.19, fusidic acid: 0.03, gentamicin: 0.04, methicillin: 0.06, mupirocin: 0.21, penicillin: 0.04 , tetracycline: 0.13, otherwise: 0.03.

2.10.6 Nanopore sequencing

The DNA library was prepared using the Genomic DNA Sequencing Kit SQK-MAP005, according to the manufacturer's protocol (Version MN005_1115_revC_26Nov2014), with small modifications. Without undergoing any shearing process, 2 µg of DNA were treated with PreCR Repair Mix (New England BioLabs, NEB), to repair possible damage to the DNA that could interfere with the sequencing process. Following Oxford Nanopore's recommendation for the optional PreCR treatment, each 1 µg of DNA was first diluted in nuclease-free water to a volume of 85 µl, to which 10 µl ThermoPol Reaction Buffer, 1 µl NAD+, 1 µl 10 µM dNTPs, and 2 µl PreCR Repair Mix were added, and the two reactions were incubated at 37 °C for 30 min. The repaired DNA was purified with 1 volume (100 µl)

Agencourt AMPure XP beads (Beckman Coulter, UK), according to manufacturer's instructions. The purified DNA from each of the two reactions was eluted from the magnetic beads in 40 μ l EB buffer and pooled together.

The 80 μ l of DNA were end repaired using the NEBNext End Repair (NEB) module from the NEBNext DNA Library Prep Master Mix Set (New England BioLabs, UK) for Illumina by adding 10 μ l buffer and 5 μ l of enzyme mix, and nuclease-free water to a final volume of 100 μ l, and incubating the reaction at 20 °C for 30 min. The end-repaired DNA was purified with 1 volume (100 μ l) Agencourt AMPure XP beads and the purified product eluted in 25 μ l EB buffer. dA-tailing on the purified DNA was performed in a final volume of 30 μ l using the dA-tailing module of the NEBNext DNA Library Prep Master Mix Set for Illumina: 3 μ l buffer and 2 μ l Klenow Fragment (3' 5' exo-) were added and the reaction was incubated at 37 °C for 30 min. The dA-tailed DNA was then transferred to Eppendorf LoBind tubes. The ligation of the Oxford Nanopore adapter was performed by adding 10 μ l Adapter Mix, 2 μ l HP ("hairpin" adapter), 50 μ l Blunt/TA Ligase Master Mix (NEB), and water to a final volume of 100 μ l, with a 10 min incubation at room temperature.

The fragments with a hairpin were selectively pulled down using Dynabeads His-Tag Isolation and Pulldown. A 10 μ l aliquot of beads was washed twice using 200 μ l of a 1:2 dilution of Oxford Nanopore Bead Binding Buffer (BBB) and resuspended in 100 μ l undiluted BBB before adding to the sample. After a 5 min incubation at room temperature, the tube was placed on a magnetic rack and the supernatant removed. The beads were then washed twice with 200 μ l diluted BBB, and any excess of buffer was removed by aspiration with a pipette. The beads were resuspended

2. RAPID ANTIBIOTIC RESISTANCE PREDICTIONS FROM HIGH THROUGHPUT GENOME SEQUENCE DATA

in 25 μ l Oxford Nanopore Elution Buffer and left for 10 min at room temperature. The tube was placed on a magnetic rack and the library was transferred to a new tube. In parallel with the library preparation, the MinION was made ready for sequencing. A new flow cell (R7.3 chemistry) was loaded on the MinION and the Platform QC protocol in the MinKNOW software was run to assess the number of available pores for sequencing. At the end of the QC, the flow cell was primed by loading twice 150 μ l of a mix of 75 μ l Oxford Nanopore Running Buffer (2X), 72 μ l nuclease-free water, 3 μ l Fuel mix, and leaving at least a 10 min interval between subsequent loadings of buffer or library. Once the flow cell was ready, a mix of 6 μ l library, 3 μ l Fuel Mix, 75 μ l Oxford Nanopore Running Buffer (2X), nuclease-free water to a final volume of 150 μ l was loaded on the flow cell and the 48h sequencing protocol was started. Additional aliquots of library were loaded after 4h and 24h to increase yield.

Once the sequencing run had begun, the Metrichor program (<https://metrichor.com>) was started and the raw data were automatically uploaded for base calling (workflow 2D Basecalling rev 1.14).

CHAPTER 3

*Same-day diagnostic and surveillance
data for M. tuberculosis via
whole-genome sequencing of direct
respiratory samples*

3.0.1 Publication note and acknowledgements

The majority of the work described in this chapter was previously published in Votintseva et al. 2017. Related work, direct sequencing of *S. aureus*, has been accepted for publication in Anson et al. 2018 and is discussed briefly in Section 3.3.

Figures 3.1 – 3.10 and some of the text in this chapter comes from this publication. Although the work was collaborative, the work described here is the sole work of myself, with the guidance of my supervisors, unless otherwise specified below.

- The sample preparation, extraction, and sequencing of Illumina data reported in Section 3.4.1 and Section 3.4.2 was performed by Antonina A. Votintseva.
- The sample preparation, extraction, and sequencing of Oxford Nanopore Minion data and MiniSeq data reported in Section 3.4.4 was performed by Louise Pankhurst.

3.1 Introduction

Phenotypic drug susceptibility testing (DST) is currently the “gold standard” for determining resistance to antibiotics in *M. tuberculosis*. However, it is time-consuming and laborious and can place significant delays on the “turnaround time” of the sample, that is, the time between the sample being isolated and its species and antibiogram being determined. In addition, it is recognised that phenotyping is liable to error, especially for certain drugs or in the context of certain genomic variants (Miotto et al. [2014](#); Schön et al. [2016](#)).

Molecular assays such as the GenoType MTBDRplus, MTBDRsl assays (Hain Lifescience GmbH, Germany), and Xpert MTB/RIF (Cepheid, USA), partially address this problem. They can detect *M. tuberculosis* directly from clinical samples, and genotype the most common drug resistance conferring mutations. However, these have limited scope, detecting only a subset of all resistance mutations and are slow to update as our knowledge of antimicrobial resistance improves.

As a result of these limitations in existing *M. tuberculosis* diagnostics, the World Health Organization (WHO) has called for development of universal drug susceptibility testing (Uplekar et al. [2015](#); World Health Organization [2015](#)). Whole-genome sequencing (WGS) (or amplicon sequencing), has the potential to provide the best of both worlds: a test which can be applied directly to clinical samples, as is done with Xpert MTB/RIF, but which can also be used to detect a wide range of species and genetic elements. To achieve this potential improvement in turnaround time, there are several challenges to overcome. One particular challenge is the requirement

3. SAME-DAY DIAGNOSTIC AND SURVEILLANCE DATA FOR *M. TUBERCULOSIS* VIA WHOLE-GENOME SEQUENCING OF DIRECT RESPIRATORY SAMPLES

to isolate and culture suspected bacteria in advance of sequencing, which imposes delays of two weeks or more in *M. tuberculosis*.

Direct sequencing without the need for culture is challenging for both sequencing preparation and analysis. Samples are likely to contain large amounts of human and other bacteria cells. As a result, samples require enriching for *M. tuberculosis* (or your pathogen of interest) during sample preparation and the analysis following this should also be robust to contamination. Recently, there have been examples of successful sequencing of *M. tuberculosis* without the need for culture (Brown et al. 2015; Doughty et al. 2014). Doughty et al. 2014 were able to achieve only 0.002-0.7X depth of coverage for *M. tuberculosis* with 20.3-99.3% of sequences mapping to the human genome. Brown et al. 2015 applied a SureSelect target enrichment method (Agilent, USA) to capture *M. tuberculosis* DNA prior to WGS. This was effective such that 20/24 smear-positive samples achieved 90% genome coverage with $\geq 20\times$ depth; a result that was sufficient for prediction of species and antibiotic susceptibility. However, the protocol was slow (2-3 days) and may be prohibitively expensive in low-income settings, costing over £200 per sample.

In this chapter, I describe my work analysing the WGS data from 40 smear-positive, primary respiratory samples from *M. tuberculosis*-infected patients via a simple low-cost DNA extraction developed by Votintseva et al. 2017. I evaluate the protocol in terms of DNA obtained, species assignment of the sequenced reads, and our ability to obtain key clinical data (detection of *M. tuberculosis* and antibiotic susceptibility prediction), along with epidemiological information (placement on phylogenetic tree). I show that these data would enable a single test to deliver the

core information for both patient and public health in < 48 hours using Illumina-based WGS. I also describe extensions of `Mykrobe predictor` to Oxford Nanopore Technologies' (ONT) MinION on potentially low quality data with a goal to apply it to directly sequenced samples. Finally, I detail a simulated workflow with empirical data, which could provide clinical results within 12 hours—a dramatic reduction in turnaround time—using direct sequencing with an ONT MinION and analysed by `Mykrobe predictor`.

3. SAME-DAY DIAGNOSTIC AND SURVEILLANCE DATA FOR *M. TUBERCULOSIS* VIA WHOLE-GENOME SEQUENCING OF DIRECT RESPIRATORY SAMPLES

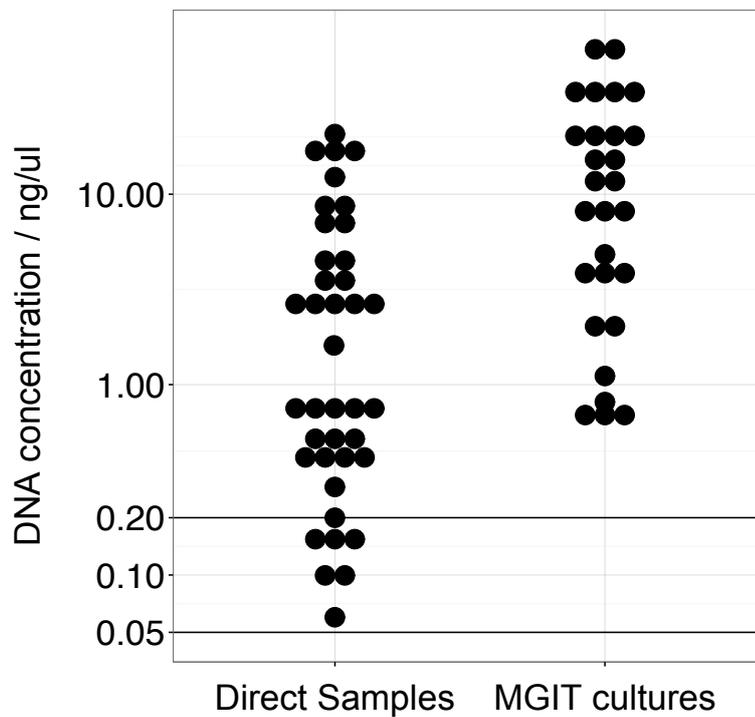


Figure 3.1: DNA extracted from MGIT cultures and direct clinical samples. Each dot represents a single extraction. Horizontal line at $0.2 \text{ ng } \mu\text{l}^{-1}$ represents the DNA concentration theoretically required for MiSeq library preparation. Horizontal line at $0.05 \text{ ng } \mu\text{l}^{-1}$ represents minimum DNA concentration used for MiSeq library preparation from direct samples in this study. One sample not shown as DNA was below detection limits.

3.2 Results

3.2.1 DNA extraction protocol and evaluation of Illumina sequencing output

DNA was extracted from 40 (Ziehl-Neelsen) ZN-positive direct respiratory samples, of which 38 were culture-confirmed *M. tuberculosis* (“culture-positive”) and

2 were culture-negative. DNA was also extracted from 28 available corresponding MGIT cultures. All direct samples were the remainder of specimens available after processing by the routine laboratory, and therefore had variable volume (median 1.5 ml, IQR 0.5-3.1, range 0.25-15) and age (median 30 days from collection to processing, IQR 15-45, range 0-67). Most direct samples (78%; 31/40) were suboptimal on the basis of either low volume (≤ 1 ml) or long storage time (≥ 30 days) or both. After DNA extraction, 33/40 (83%) direct samples and all 28 MGIT cultures yielded ≥ 0.2 ng μl^{-1} DNA, the amount recommended for MiSeq Illumina library preparation (Figure 3.1).

In total 39/40 direct samples with detectable DNA (37 culture-positive, 2 culture-negative) and 27/28 MGIT cultures were sequenced on an Illumina MiSeq. One MGIT culture was not sequenced because the corresponding direct sample failed to yield measurable DNA. All sequenced direct samples produced ≥ 1.5 million reads (median 3.6 million, IQR 2.9-5.0, range 1.5-12), as did all MGIT cultures (median 3.1 million, IQR 2.8-3.3, range 2.0-4.1).

3.2.1.1 Contamination in direct and MGIT samples

To determine levels of contamination and *M. tuberculosis* in samples, reads were mapped with `bwa-mem` (Li et al. 2010a) to the human reference genome GRCh37 (hg19), and human reads were counted and discarded. Remaining stored reads were then mapped to the *M. tuberculosis* H37Rv reference strain (GenBank NC_018143.2), and any unmapped reads were then mapped to nasal, oral, and mouth flora available in the NIH Human Microbiome Project database

3. SAME-DAY DIAGNOSTIC AND SURVEILLANCE DATA FOR *M. TUBERCULOSIS* VIA WHOLE-GENOME SEQUENCING OF DIRECT RESPIRATORY SAMPLES

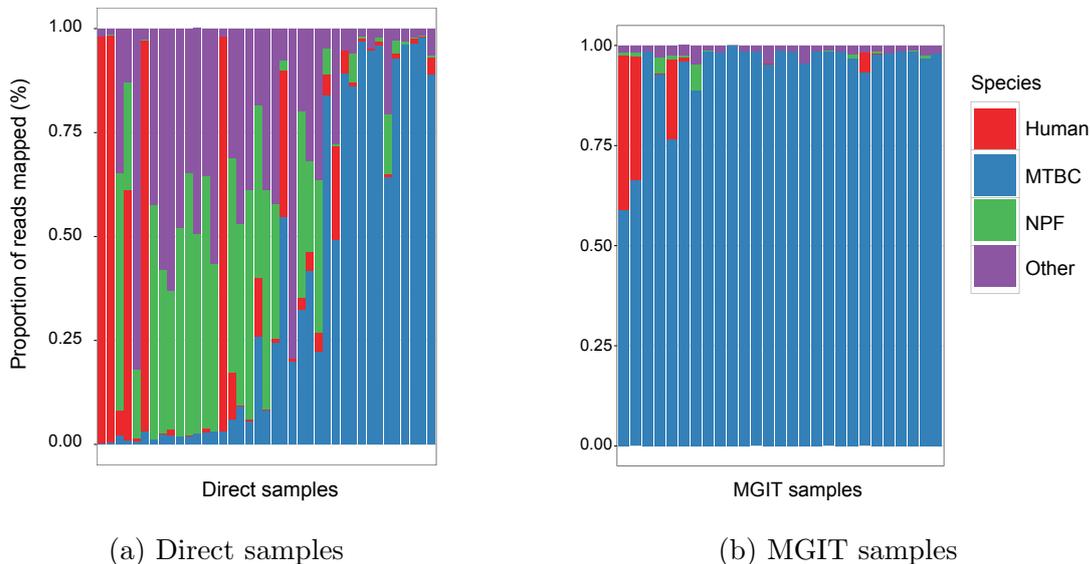


Figure 3.2: The proportion of reads assigned to various taxonomic categories in each sample sorted by increasing total count of MTBC reads. a) Direct samples show removal of human DNA (red) has been broadly successful, but removal of NPF (green) and other bacteria (purple) had more variable success. b) MGIT samples show much more uniform dominance of *M. tuberculosis* reads, as expected after 2 weeks of culture designed to favour mycobacterial growth.

(<http://www.hmpdacc.org/>).

Based on the mapping results, I assigned reads to the following categories: *M. tuberculosis*, human, naso-pharyngeal flora (NPF), and “other”. 77% (30/39) of direct samples contained < 10% human reads. However, only 46% (18/39) contained < 10% NPF and other bacterial reads, and 26% (10/39) contained > 40% of reads from non-mycobacterial, non-NPF bacteria (Figure 3.2a). By comparison, MGIT culture samples showed much less contamination, as expected after two weeks of culture designed to favour mycobacterial growth (Figure 3.2b).

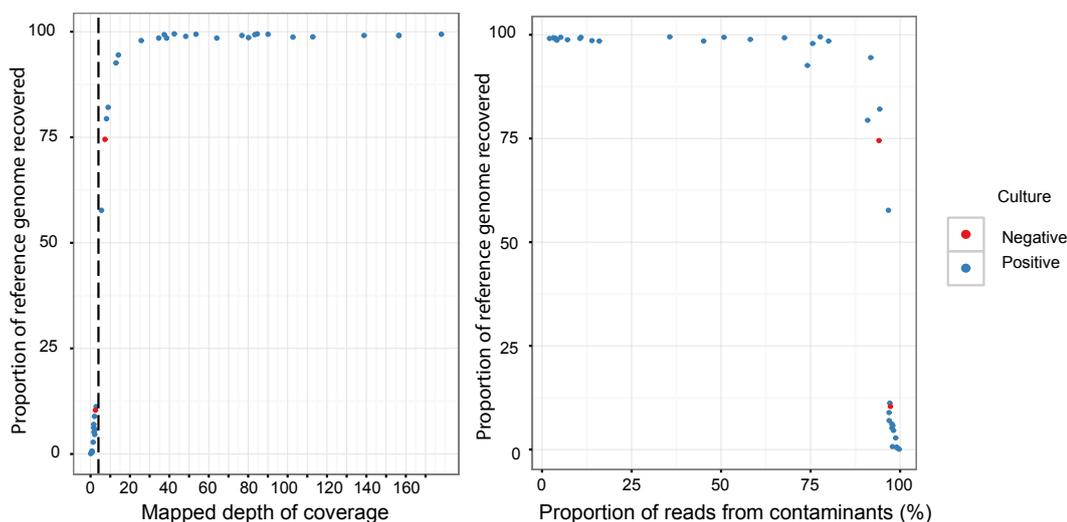


Figure 3.3: Recovery of *M. tuberculosis* genome in direct samples and robustness to contamination. Depth versus proportion of the *M. tuberculosis* reference recovered (at $> 5\times$ depth). Vertical dotted line at $3\times$ depth is threshold used for resistance prediction in this study. Proportion of contamination (reads not mapping to *M. tuberculosis* reference) versus proportion of genome recovered is shown. Samples with $< 95\%$ of the *M. tuberculosis* genome recovered all have $> 75\%$ contaminated reads.

3.2.1.2 Recovery of *M. tuberculosis* genome

Figure 3.3a shows the distribution of the *M. tuberculosis* reference genome depth of coverage across direct samples. Samples either have more than $10\times$ depth and recover more than 90% of the genome, or have $< 3\times$ depth and recover less than 12% of the genome. The vertical dotted line delineates a threshold of $3\times$ coverage, below which resistance predictions were not made. Figure 3.3b shows the amount of contamination (reads not mapping to *M. tuberculosis*) per sample. Ten samples had $< 15\%$ contaminant reads, although contamination levels increased as high

3. SAME-DAY DIAGNOSTIC AND SURVEILLANCE DATA FOR *M. TUBERCULOSIS* VIA WHOLE-GENOME SEQUENCING OF DIRECT RESPIRATORY SAMPLES

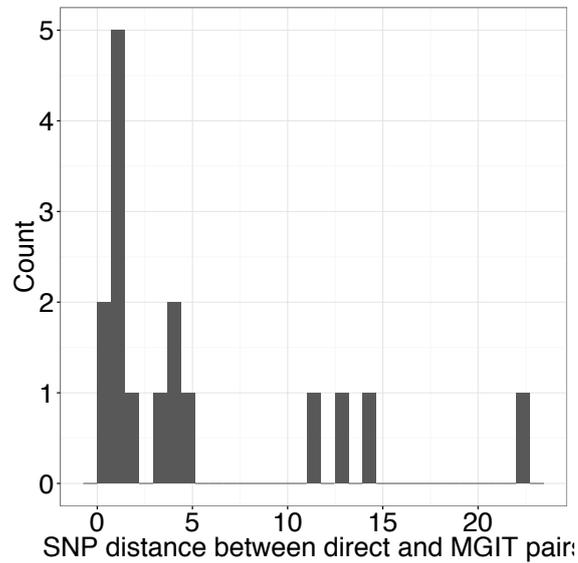


Figure 3.4: Genotypic concordance between direct and paired MGIT samples. Histogram of genetic (SNP) differences, excluding the one pair which differs by 1106 SNPs; median (and modal) difference is 1—implying direct sequencing is identifying the same strain of *M. tuberculosis* as culture-based sequencing would.

as 75% before the proportion of the *M. tuberculosis* genome recovered started to drop. Low numbers of *M. tuberculosis* reads could also reflect poor DNA quality from samples stored for long periods, as most of the samples with $< 80\%$ reference genome coverage (12/17, 71%) were more than 3 weeks old before extraction.

3.2.2 Concordance of results from direct and MGIT samples

In order to compare concordance of genotypes between direct and MGIT pairs, I used `Mykrobe predictor` to genotype all samples on a set of SNPs. To generate

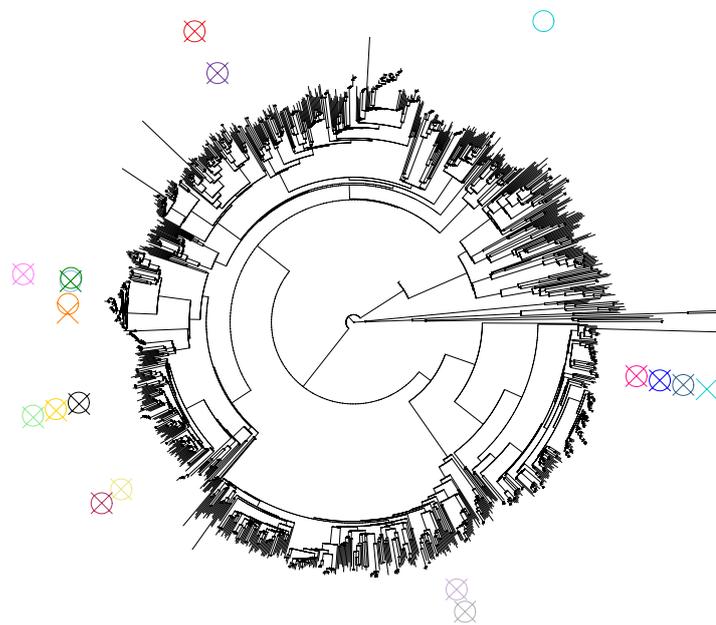


Figure 3.5: Genotypic concordance between direct and paired MGIT samples. Placing direct/MGIT pairs on a phylogenetic tree of 3480 samples shows distribution of samples across world diversity. Circles indicate the sequence from the direct sample and crosses the sequence from the corresponding MGIT sample—overlapping circles and crosses indicate pairs which were placed close to each other on the tree. The 1 pair (of 17) which does not overlap, had 1106 SNP differences (turquoise, 1/4 o'clock on the tree) between direct and MGIT. The MGIT sample of this pair places very close to other samples (0 SNP differences, see 4 o'clock on the tree), and so is possibly due to a labelling error.

3. SAME-DAY DIAGNOSTIC AND SURVEILLANCE DATA FOR *M. TUBERCULOSIS* VIA WHOLE-GENOME SEQUENCING OF DIRECT RESPIRATORY SAMPLES

this set of SNPs, conservative SNP calls were made using Cortex (Iqbal et al. 2012) (independent workflow, $k=31$) on 3480 samples from Walker et al. 2015. Singleton variants were discarded, and a de-duplicated list of 68695 SNPs was constructed. All samples (from our study and from Walker et al. 2015) were genotyped at these sites using the `Mykrobe predictor` genotyping model (Section 2.4.1) (Bradley et al. 2015). I excluded pairs where the direct sample had $< 5\times$ coverage to ensure like-with-like analysis, leaving 17 pairs which I compared. I measured the number of SNP differences between the paired direct and MGIT samples, counting only sites where both genotypes had confidence in our Illumina model greater than 1, and neither site was called as heterozygous. This allowed us to calculate a genetic distance between the 17 paired MGIT and direct samples. The median (and modal) SNP difference was 1 (Figure 3.4), with one outlier pair of samples that differed by 1106 SNPs (discussed below), and all other differences ≤ 22 SNPs.

I placed the 17 paired direct and MGIT samples on the phylogeny from section 2.2.3 (Figure 2.3) by identifying the leaf with the fewest SNP differences, across the 68695 sites. Placement therefore returns a closest leaf, and a SNP distance to that leaf. The samples were distributed across global diversity (Figure 3.5; tree thinned to aid visibility). For the pair with 1106 SNP differences, the MGIT sample was placed very closely to 3 other pairs (0 SNP difference to one sample, and 5 SNP differences to the others). Although this outlier might result from different strains being present within the host, a within-laboratory labelling error or cross-contamination is also possible and seems more likely given the close relationship to other samples in this study.

3.2.2.1 No evidence of higher diversity in direct samples

Comparing direct/MGIT pairs where both samples had at least $20\times$ mean depth of coverage on the *M. tuberculosis* reference, the median number of high confidence heterozygous sites was 25 in both direct and MGIT samples. There was no clear trend of greater genome-wide diversity in direct samples (Figure B.1).

3.2.2.2 Detection of *M. tuberculosis* in culture-positive/-negative samples

I ran `Mykrobe predictor` (v0.3.5) on all sequenced culture-positive (37/39) direct *M. tuberculosis* samples and 27 MGIT samples. All direct samples were successfully identified to complex level (37/37) and 95% to species level (35/37), including 13/37 (35%) where the mean depth of coverage was $< 3\times$. All MGIT cultures were identified as *M. tuberculosis*. `Mykrobe predictor` was also able to identify *M. tuberculosis* in 2/2 direct samples with low acid-fast bacilli (AFB) scores (+1) and no growth in MGIT culture; these may represent dead bacilli from a patient undergoing treatment.

3.2.2.3 Antibiotic resistance prediction

Mycobacterial species and antibiotic resistance to isoniazid, rifampicin, ethambutol, pyrazinamide, streptomycin, aminoglycosides (including capreomycin, amikacin and kanamycin), and fluoroquinolones (including moxifloxacin, ofloxacin, and ciprofloxacin) were predicted using `Mykrobe predictor` v0.3.5 using a catalogue

3. SAME-DAY DIAGNOSTIC AND SURVEILLANCE DATA FOR M. TUBERCULOSIS VIA WHOLE-GENOME SEQUENCING OF DIRECT RESPIRATORY SAMPLES

Table 3.1: Antibiotic susceptibility predictions by reference laboratory from MGIT cultures and by WGS directly from clinical samples (n=37), MGIT cultures. I=Isoniazid,R=Rifampicin,E=Ethambutol,S=Streptomycin,A=Aminoglycosides,F=Fluoroquinolones. S=sensitive, R=resistant, r= mixed result (presence of S and R variants), F=failed for the reference laboratory DST /falls below Mykrobe predictor confidence threshold for WGS, .=not tested.

Sample	Reference laboratory DST							WGS direct samples							WGS MGIT culture						
	I	R	E	P	S	A	F	I	R	E	P	S	A	F	I	R	E	P	S	A	F
F13757	S	S	S	S	.	.	.	F	F	F	F	F	F	F	S	S	S	S	S	S	S
M5979	S	S	S	S	.	.	.	F	F	F	F	F	F	F	S	S	S	S	S	S	S
M5992	S	S	S	S	.	.	.	S	S	S	S	S	S	S	S	S	S	S	S	S	S
L48766	S	S	S	S	.	.	.	F	F	F	F	F	F	F	S	S	S	S	S	S	S
H32593	S	S	S	S	.	.	.	F	F	F	F	F	F	F	S	S	S	S	S	S	S
L11705	S	S	S	S	.	.	.	S	S	S	S	S	S	S	S	S	S	S	S	S	S
F33033	R	S	S	S	.	.	.	F	F	F	F	F	F	F	S	S	S	S	S	S	S
600223	S	S	S	S	.	.	.	S	S	S	S	S	S	S	S	S	S	S	S	S	S
600225	S	S	S	S	.	.	.	S	S	S	S	S	S	S	S	S	S	S	S	S	S
600494	S	S	S	S	.	.	.	S	S	S	S	S	S	S	S	S	S	S	S	S	S
600529	S	S	S	S	.	.	.	S	S	S	S	S	S	S	S	S	S	S	S	S	S
602112	R	S	R	S	.	.	.	R	R	R	r	S	r	R							
602188	S	S	S	S	.	.	.	S	S	S	S	S	S	S							
602378	S	S	S	S	.	.	.	S	S	S	S	S	S	S							
602497	S	S	S	S	.	.	.	S	S	S	S	S	r	S	S	S	S	S	S	S	S
602532	R	S	S	S	.	.	.	R	S	S	S	S	S	S	R	S	S	S	S	S	S
602861	S	S	S	S	.	.	.	F	F	F	F	F	F	F	S	S	S	S	S	S	S
602905	R	R	R	S	.	.	.	R	R	R	r	S	r	R							
602907	R	R	S	S	.	.	.	F	F	F	F	F	F	F	R	R	S	S	S	S	S
602908	R	R	S	S	.	.	.	F	F	F	F	F	F	F	R	R	S	S	S	S	S
603183	R	S	S	S	.	.	.	R	S	S	S	S	S	S	R	S	S	S	S	S	S
603184	R	S	S	S	.	.	.	R	S	S	S	S	S	S	R	S	S	S	S	S	S
614114	R	S	S	S	.	.	.	R	S	S	S	S	r	S	R	S	S	S	S	S	S
614115	R	S	S	S	.	.	.	R	S	S	S	S	r	S	R	S	S	S	S	S	S
614341	S	S	S	S	.	.	.	S	S	S	S	S	S	S	S	S	S	S	S	S	S
614406	S	S	S	S	.	.	.	S	S	S	S	S	S	S	S	S	S	S	S	S	S
614509	R	R	R	R	.	.	.	R	R	R	r	S	r	R							
617126	S	S	S	S	.	.	.	F	F	F	F	F	F	F	S	S	S	S	S	S	S
617332	S	S	S	S	.	.	.	S	S	S	S	S	S	S	S	S	S	S	S	S	S
L26916d	S	S	S	S	.	.	.	S	S	S	S	S	S	S	S	S	S	S	S	S	S
H54502	R	S	S	S	.	.	.	F	F	F	F	F	F	F	R	S	S	S	S	S	S
L26183	S	S	S	S	.	.	.	S	S	S	S	S	S	S	S	S	S	S	S	S	S
L26556	S	S	S	S	.	.	.	S	S	S	S	S	S	S	S	S	S	S	S	S	S
L26916	S	S	S	S	.	.	.	S	S	S	S	S	S	S	S	S	S	S	S	S	S
L61938	S	S	S	S	.	.	.	F	F	F	F	F	F	F	S	S	S	S	S	S	S
L66375	R	R	R	R	.	.	.	F	F	F	F	F	F	F	R	R	R	R	S	S	S
M26794	S	S	S	S	.	.	.	F	F	F	F	F	F	F	S	S	S	S	S	S	S

Table 3.2: Antibiotic resistance conferring mutations identified by WGS in direct samples and if available, corresponding MGIT cultures. “m” after sample ID indicates MGIT culture; Resistance-conferring mutation in the relevant gene, R:S:C represents the number of reads on the R allele; the number of reads on the S allele; and C represents Mykrobe Illumina-model confidence score (this is, the difference between log likelihoods of most and next-most likely models, so any value greater than 1 is confident).

Patient	Sample ID	Mutation_R:S:C2						
		Isoniazid	Rifampicin	Ethambutol	Pyrazinamide	Aminoglycosides	Fluoroquinolones	
1	614114	katG_S315T_21:0:99						rrs_C1402A_19:22:43
	614114m	katG_S315T_78:0:99						
	614115	katG_S315T_48:2:69						rrs_C1402A_55:59:124
1	614115m	katG_S315T_85:0:99						
	614509	fabG1_C15T_35:0:99	rpoB_I491F_27:0:99	embB_M306L30:1:44	pncA_V7L_17:13:23	rrs_A140IG_19:39:39	gyrA_D94G_31:0:99	rrs_A140IG_36:42:81
2	602112	fabG1_C- 15T_35:4:40	rpoB_I491F_50:1:76	embB_M306L_34:0:99	pncA_V7L_27:9:8	rrs_A140IG_36:42:81	gyrA_D94G_41:0:99	
		katG_S315T_38:44:85			pncA_T135P_11:37:20			
2	602905	fabG1_C- 15T_130:0:99	rpoB_I491F_168:0 :99	embB_M306L181:0:99	pncA_V7L_119:57:32	rrs_A140IG_73:91:164	gyrA_D94G_156:0:99	
	602532	fabG1_C- 15T_174:0:99						
3	602532m	fabG1_C- 15T_98:1:152						
	603183	fabG1_C- 15T_155:0:99						
4	603183m	fabG1_C- 15T_33:0:99						
	603184	fabG1_C- 15T_129:0:99						
4	603184m	fabG1_C- 15T_84:2:126						

3. SAME-DAY DIAGNOSTIC AND SURVEILLANCE DATA FOR *M. TUBERCULOSIS* VIA WHOLE-GENOME SEQUENCING OF DIRECT RESPIRATORY SAMPLES

of AMR variants from Walker et al. 2015. For samples where the estimated depth of kmer-coverage of *M. tuberculosis* reported by `Mykrobe predictor` fell below $3\times$, no resistance predictions were made (see Figure 3.3).

In total, 168 predictions for first-line (n=96) and second-line (n=72) antibiotic susceptibility were made for the 24/37 (65%) direct samples which had at least $3\times$ depth (Tables 3.1,3.2). For the 13/37 (35%) samples that had $< 3\times$ depth, no resistance predictions were made. This included 1/2 culture-negative samples.

First-line antibiotics were concordant with the reference laboratory DST in 92/96 (96%) predictions. The four mismatches (three pyrazinamide mixed genotypes with both R and S alleles present, and one rifampicin-resistant genotype with a sensitive phenotype) were found across three samples, all from the same patient (patient 2 in Table 3.2), who had a variable phenotype for rifampicin and pyrazinamide. The resistant genotype for rifampicin was consistent across all three samples from this patient (*rpoB*-I491F). There is evidence that this mutation causes resistance, but that the phenotype is often reported as sensitive (Walker et al. 2015). The mixed genotype for pyrazinamide was again consistent with presence of both R and S alleles on *pncA*-V7L across all three samples, whereas the phenotype varied. This mutation is also known to confer resistance in samples reported as phenotypically sensitive (Walker et al. 2015). Further, 1/3 samples from this patient (sample 602112, Table 3.2) contained two additional mutations conferring resistance to isoniazid and pyrazinamide, *katG*-S315T and *pncA*-T135P respectively, which were not detected in the previous or following sample. This variation between genotypes from same-patient samples taken over time may represent ongoing evolution, changing population size,

and within-patient diversity of *M. tuberculosis*. This has been previously observed by Eldholm et al. 2014.

In addition to the above, WGS provided 72 predictions for second-line antibiotics where phenotypic DST was not attempted. The 13/37 samples that yielded insufficient WGS data for resistance prediction had a higher proportion of other bacterial DNA than the samples with more than $3\times$ depth (Figure 3.2; median 96%, IQR 38-70%, vs median 12%, IQR 0-67%, in those where resistance prediction was possible, rank-sum $p=0.01$).

3.2.3 Sub-24 hour turnaround time with Illumina MiniSeq and ONT MinION

The results above demonstrated that it is possible to accurately predict AMR resistance from direct samples within 48 hours (Figure 3.11). Admittedly, the failure rate is high—35% (13/37) of direct samples did not generate sufficient data for resistance prediction. However, with improvements to the enrichment step and with higher quality DNA it is likely this failure rate will improve. The most time consuming step of the Illumina MiSeq workflow is the 16-hour sequencing run (Figure 3.11). We set out to simulate similar workflows using Illumina MiniSeq and Oxford Nanopore’s Minion, both of which have shorter turnaround times.

3. SAME-DAY DIAGNOSTIC AND SURVEILLANCE DATA FOR *M. TUBERCULOSIS* VIA WHOLE-GENOME SEQUENCING OF DIRECT RESPIRATORY SAMPLES

Table 3.3: Yield from pure BCG, and from negative sputum spiked with BCG. Sequenced on Illumina MiniSeq

	Fmol	Yield (Mb)	Read length (bp)	BCG coverage (%)
100% N716	800	381	101	84
50 % N718	800	244	101	31
50% N719	800	257	101	33

3.2.4 Sub-24 hour turnaround time with Illumina MiniSeq

Illumina MiniSeq sequencing for three samples (single run; 1 pure BCG, 2 negative sputum human DNA spiked with BCG DNA) was completed in 6 hours 40 minutes (see Methods Section 3.4.4). BCG reference genome coverage was 31 \times and 33 \times in the two spiked samples, and 84 \times in pure BCG (Table 3.3). In all cases the species/strain was correctly identified as *M. bovis* strain BCG, and pyrazinamide resistance was correctly identified due to mutation H57D in *pncA*.

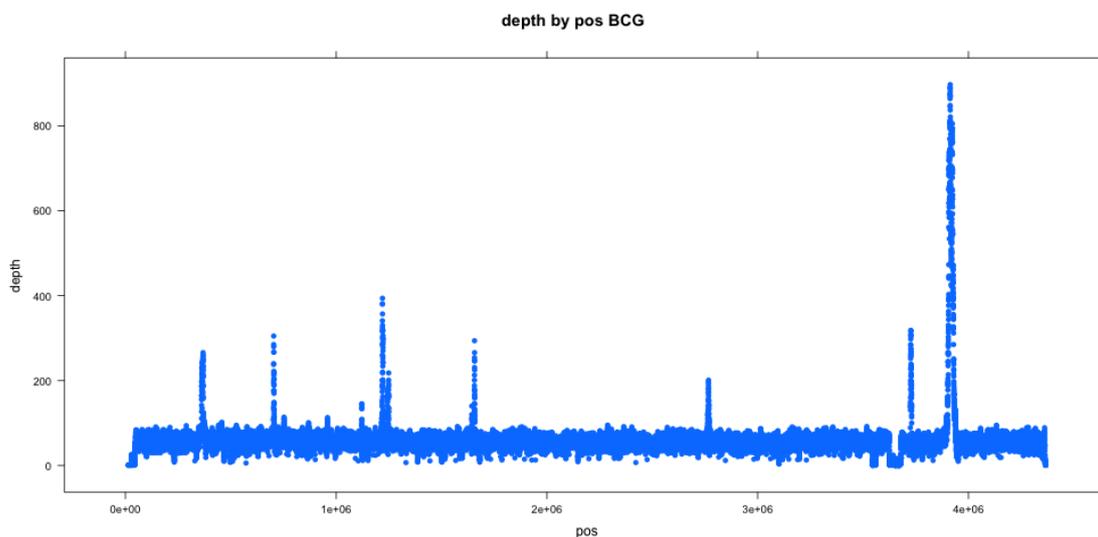


Figure 3.6: Coverage distribution across *M. bovis* BCG strain reference genome of MinION reads from pure *M. bovis* BCG strain culture, showing even coverage across the genome apart from 7 loci with peaks of coverage.

3.2.5 Nanopore MinION sequencing of spiked samples

In order to test the feasibility of using Oxford Nanopores’s (ONT) MinION to sequence direct samples, several human sputum samples spiked with *bovis* at various proportions (See Methods Section 3.4.5, Table 3.4) were sequenced. A new PCR-based rapid 1D MinION protocol was used by extracting BCG DNA, ZN-negative sputum DNA spiked with BCG DNA, and R9 flowcells (See Methods Section 3.4.5). Analysis of genome-wide coverage distribution confirmed that use of PCR had not led to significant coverage bias (Figure 3.6), and that > 95% of the reference genome attained coverage $5\times$ for all samples.

3.2.5.1 Modified Mykrobe predictor genotyping model for ONT data

Since the expected error rate for ONT data was higher than for Illumina data, and the expected yield was lower, I made modifications to the predictor genotyping model.

I modelled this as three competing k -mer producing Poisson processes:

0/0:

$$KmerCount(ref) \sim Pois(\mathcal{D}(1 - \varepsilon)),$$

$$KmerCount(alt) \sim Pois\left(\frac{\mathcal{D}\varepsilon}{3}\right);$$

1/1:

$$KmerCount(ref) \sim Pois\left(\frac{\mathcal{D}\varepsilon}{3}\right),$$

$$KmerCount(alt) \sim Pois(\mathcal{D}(1 - \varepsilon));$$

0/1:

$$KmerCount(ref) \sim Pois\left(\frac{\mathcal{D}}{2} + \frac{\mathcal{D}\varepsilon}{2 \cdot 3}\right),$$

$$KmerCount(alt) \sim Pois\left(\frac{\mathcal{D}}{2} + \frac{\mathcal{D}\varepsilon}{2 \cdot 3}\right);$$

where $KmerCount()$ is a function returning the number of k -mers observed from a given allele (i.e., the sum of the k -mer coverage), \mathcal{D} is the expected number of k -mers given an expected depth of coverage D ($\mathcal{D} = kD$), and ε is the expected

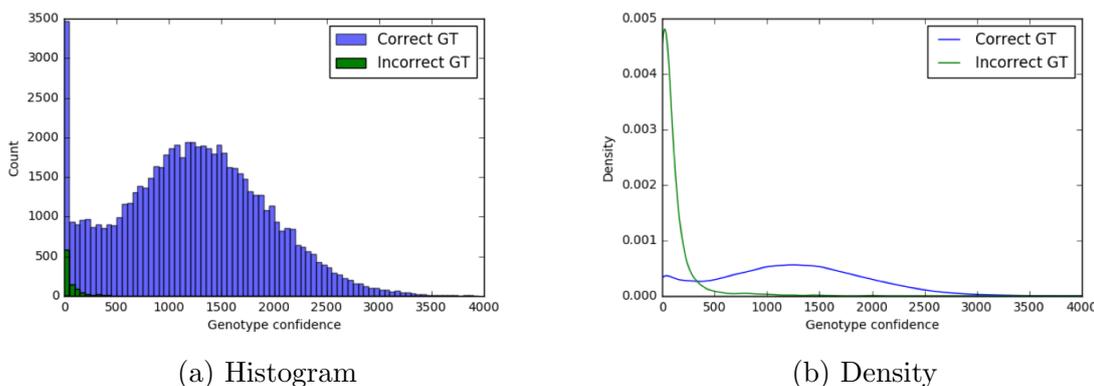


Figure 3.7: Distribution of genotype confidence across 68950 SNPs genotyped on R9.4 MinION reads. To clarify, if under our ONT-model, the log likelihood of a particular A/G SNP having allele A is (say) -0.01 , and the log likelihood of the allele being G is -1000 , then the SNP is genotyped as having allele A with confidence $-0.01 - (-1000) = 999.99$. Larger values of genotype confidence signify more certain results. Here, bars are split by whether the genotype was correct (blue) or incorrect (green) (using MiniSeq genotypes as truth). `Mykrobe predictor` uses a default genotype confidence threshold of 100 for ONT genotyping.

error rate. k is the k -mer size of the `Mykrobe predictor` de Bruijn graph. The log-likelihood of each allele is summed, and the maximum likelihood model is chosen. The confidence is given by the difference between the log-likelihoods of the two most likely models.

I compared the genotype calls from ONT MinION run using the model above to genotypes already generated from a MiniSeq run on the same sample. Figure 3.7 shows the genotype confidence distribution split by whether the genotype is correct or not, using the MiniSeq genotype as truth. For the analysis of ONT runs below, I used a genotype confidence threshold of 100, below which `Mykrobe predictor` did not make a call.

3. SAME-DAY DIAGNOSTIC AND SURVEILLANCE DATA FOR *M. TUBERCULOSIS* VIA WHOLE-GENOME SEQUENCING OF DIRECT RESPIRATORY SAMPLES

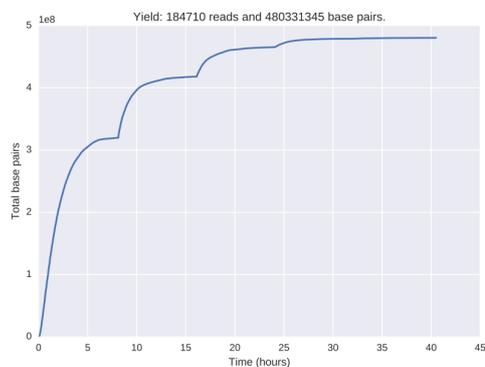


Figure 3.8: Cumulative yield (in megabases) from the MinION (version R9 flow cell) when sequencing culture negative sputum spiked with 15% BCG, using the Phusion PCR enzyme. Approximately 65% of the final yield is obtained in 8 hours and 80% in 10 hours.

3.2.5.2 Analysis of modified method for ONT MinION

In all cases `Mykrobe predictor` correctly identified the species/strain as *M. bovis* strain BCG (Table 3.3). Amplification with Phusion High-Fidelity master mix resulted in the highest yield (760Mb, with 68x coverage of BCG). All MinION experiments resulted in correct identification of the H57D mutation in *pncA* that confers pyrazinamide resistance in BCG, but in the 5% spike this call was filtered out as it only had kmer coverage of 1× on the resistant allele, and did not achieve the required confidence threshold. In all cases, no false resistance calls were made, but deep coverage was needed to be able to genotype all 175 mutations in the catalogue. Although the pure BCG/R9 and 15% BCG/Phusion/R9 runs missed only 3/175 and 1/175 mutations, respectively (Table 3.3), only the R9.4 sequencing run (below) allowed all mutations to be typed.

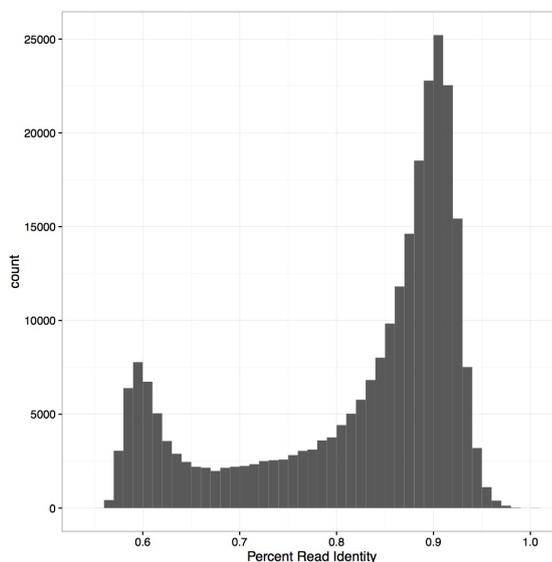


Figure 3.9: Identity distribution for 1D reads in pure BCG sequencing run with MinION (version R9 flow cell). Each read is mapped to the BCG reference genome and percentage identity is measured, and then the distribution of these scores is shown.

In all 5 samples sequenced on R9 flowcells, data yield was highest at the start of the run, with consistent yield profiles. For the Phusion/15% run, over 65% of the data was obtained in 8 hours, and 80% in 10 hours (Figure 3.8). Despite the high sequencing error rate (Figure 3.9), high accuracy genotyping of known SNPs/indels was achievable as described above.

3.2.5.3 12 hour turnaround time with ONT MinION

A single sample (15% BCG spiked ZN-negative sputum) was sequenced on the latest R9.4 MinION flowcell (see Methods Section 3.4.5). Yield was 1.3Gb in 48 hours. Mykrobe predictor was able to detect the *M. tuberculosis* complex, iden-

3. SAME-DAY DIAGNOSTIC AND SURVEILLANCE DATA FOR *M. TUBERCULOSIS* VIA WHOLE-GENOME SEQUENCING OF DIRECT RESPIRATORY SAMPLES

Table 3.4: Yield from pure BCG, and from negative sputum spiked with BCG—both sequenced with MinION 1D protocol.

Model	Sample	Fmol loaded	Read count	Yield (Mb)	Avg read length (kb)	BCG covg depth	H57D R-allele kmer covg+	% mutations typed (# not typed)
R9	Pure cultured BCG	++	297,239	360	1.2	80	17	99 (1)
R9	5% BCG LongAmp	82	182,670	559	2.0	19	1*	47 (93)
R9	10% BCG LongAmp	76	180,507	467	1.8	10	3	56 (77)
R9	15% BCG LongAmp	51	203,285	627	2.0	35	3	90 (18)
R9	15% BCG Phusion	27	184,895	758	2.4	68	10	98 (3)
R9.4	15% BCG Phusion	43	754,338	1306	1.7	147	16	100 (0)

tify the strain as BCG, detect the correct pyrazinamide resistance mutation, and correctly place the sample on the phylogenetic tree after 1 hour of sequencing. After 3 hours, 170/175 mutations were genotyped confidently; after 4 hours, definitive results for all drugs except streptomycin; and after 6 hours, definitive results for streptomycin were obtained and sequencing could be stopped. One pyrazinamide mutation remained un-genotyped, but since `Mykrobe predictor` already had made a confident resistance call for pyrazinamide, there would be no need to continue. Sufficient coverage on the final mutation was obtained after a further 3 hours (9 hours total sequencing; Table 3.5). Incorporating 6.5 hours for decontamination, DNA ex-

Table 3.5: Susceptibility prediction at timestamps during R9.4 run.

Hour	%AMR	# mutations un-typed (/175)	Un-genotyped mutations	Drugs awaiting results
1	57.1	75	*	All except PYZ
2	88.5	20	katG (S700, L141, V633, W191, D142, L704) gid (L26, V41, G34, R47, G117, A205, R118, Q125) rpoB (H445) embB (D328, G406) rpsL (K43) pncA (T47, K48)**	ISO, STREP, RIF, ETH
3	97.1	5	embB (D328) gid (G34, A205) katG (W191) pncA (T47)	ISO, STREP, ETH
4	98.2	3	gid (G34, A205) pncA (T47)	STREP
5	98.8	2	gid (G34) pncA (T47)	STREP
6	99.4	1	pncA (T47)	-
9	100	0	-	-

* We omit list of un-genotyped mutations here; 75 is too many to list.

** further un-genotyped pncA mutations could be ignored, as H57D had already been detected at 1 hour—sample was already predicted to be pyrazinamide resistant. Thus, pyrazinamide not listed in column 5.

3. SAME-DAY DIAGNOSTIC AND SURVEILLANCE DATA FOR *M. TUBERCULOSIS* VIA WHOLE-GENOME SEQUENCING OF DIRECT RESPIRATORY SAMPLES

Table 3.6: Prediction of time taken to identify *M. tuberculosis* (and provide initial susceptibility predictions) and of time taken to give final susceptibility predictions with MinION R9.4. for a range of different proportions of *M. tuberculosis*. Estimates are based on the proportion of reads that mapped to the *M. tuberculosis* reference genome (column 1), from each of the 39 samples sequenced with Illumina MiSeq. Each sample is assumed to generate 1.3 Gigabases of reads, arriving according to the yield curve from our R9.4 run (Figure 3.8). From these, and the depth of coverage at the hourly timestamps at which our R9.4 run detected Mtb/produced final predictions, one can predict the equivalent time for all 39 samples. If the result is not achieved within 48 hours, we mark it as N/A. We highlight two pairs of statistics. First, for samples with $\geq 20\%$ reads from *M. tuberculosis*, identification of *M. tuberculosis* would take 20 minutes, and complete results would take 150 minutes. Second, for samples with $\geq 84\%$ reads from *M. tuberculosis*, identification of *M. tuberculosis* would also take 20 minutes, but full results would take 93 minutes. ID=Identify Mtb+initial susceptibility predictions (hours).

Reads mapping to <i>M. tub</i> (%)	Identify (hours)	Complete DST (hours)
0.28	NaN	NaN
0.64	NaN	NaN
0.83	32.28	NaN
0.95	25.08	NaN
1.23	12.90	NaN
1.83	8.16	NaN
2.06	6.89	NaN
2.14	6.51	NaN
2.30	5.95	NaN
2.32	5.89	NaN
2.69	4.90	NaN
2.88	4.53	NaN
3.07	4.20	NaN
3.09	4.18	NaN
3.24	3.98	NaN
5.67	2.31	NaN
5.91	2.21	NaN
8.23	1.65	28.72
9.07	1.52	24.64
19.99	0.80	6.63
22.30	0.72	5.75
24.42	0.67	5.14
25.91	0.64	4.80
32.30	0.54	3.75
41.86	0.46	2.88
49.19	0.41	2.48
54.89	0.38	2.25
64.35	0.35	1.95
84.01	0.30	1.55
86.04	0.30	1.52
89.13	0.30	1.48
89.40	0.30	1.47
92.90	0.29	1.42
94.78	0.29	1.39
95.86	0.29	1.37
96.18	0.29	1.37
96.50	0.29	1.36
96.79	0.29	1.36
97.89	0.28	1.34

traction, and sample preparation (Figure 3.11), this would give a turnaround time of 7.5 hours for identifying species, phylogenetic placement, and initial susceptibility predictions, and 12.5 hours for complete results.

I took the phylogenetic placement of the MiniSeq BCG data as truth, which was 4 SNPs distant from a known BCG sample on the predefined tree from Walker et al. 2015. After 1 hour of sequencing with R9.4, I was able to confidently genotype 22694 of the 68695 SNPs, placing the sample at the correct leaf of the tree, at an estimated distance of 3 SNPs. Thus, genotyping on 1D nanopore reads had at most 7 errors (=3+4) out of 22694 SNPs—an error rate below 0.03%.

3.2.6 MinION turnaround estimates using empirical *M. tuberculosis* read proportion data

The proportion of *M. tuberculosis* reads found in our clinical samples varied over a considerable range (Figure 3.2, blue), with between 0.3% and 97.9% of sequenced DNA coming from *M. tuberculosis*. To model how this distribution might translate into MinION performance, I used hourly timestamps on the R9.4 MinION total DNA yield curve (Figure 3.8), and coverage needed to detect *M. tuberculosis*, pyrazinamide resistance and full susceptibility results, to estimate the turnaround times for all Illumina-sequenced samples. This assumed they all were to yield 1.3Gb of MinION reads with the same proportion of reads from *M. tuberculosis*, as seen in Figure 3.2.

Finally, based on the performance of the 1.3Gb R9.4 sequencing run, I esti-

3. SAME-DAY DIAGNOSTIC AND SURVEILLANCE DATA FOR *M. TUBERCULOSIS* VIA WHOLE-GENOME SEQUENCING OF DIRECT RESPIRATORY SAMPLES

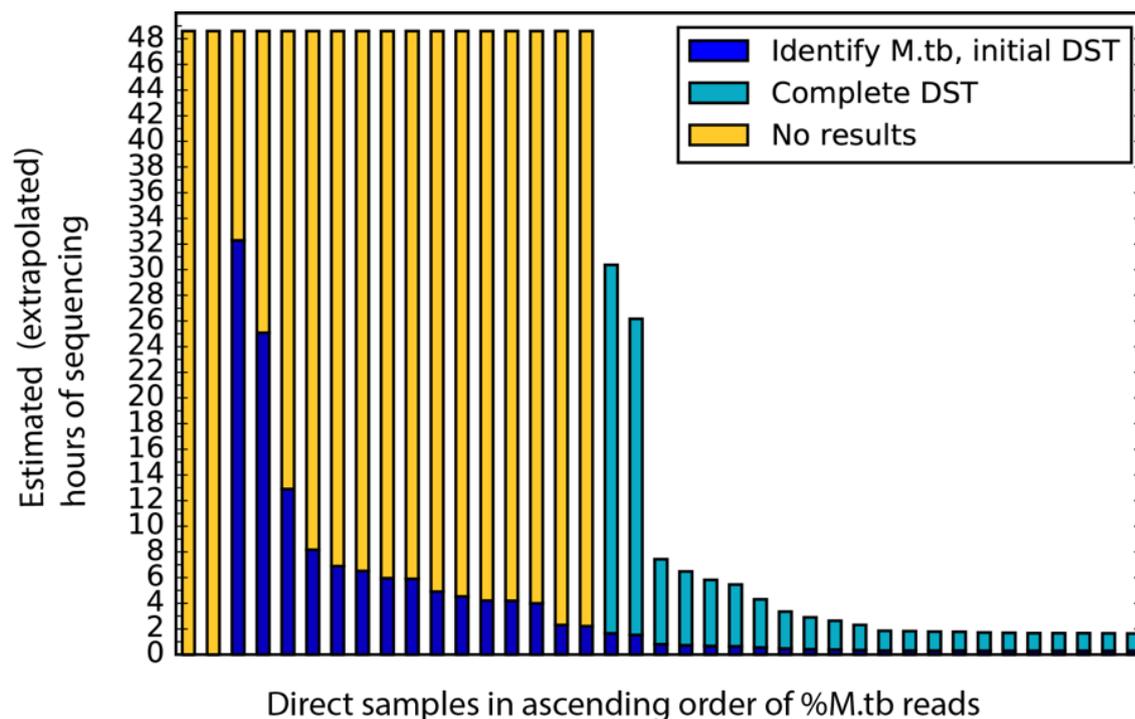


Figure 3.10: Extrapolating from MinION R9.4 results to estimate performance for a realistic distribution of proportion of reads from *M. tuberculosis*. Each bar corresponds to one of the 39 samples sequenced with Illumina MiSeq sorted by increasing proportion of *M. tuberculosis*. The simulation assumes that if sequenced with a R9.4 MinION, each of the 39 samples would yield as much as our R9.4 run (1.3Gb), and infer the yield of *M. tuberculosis* reads for each sample from the proportion (blue), shown in Figure 2a. This gives us estimates of sequencing time needed to detect *M. tuberculosis* (dark blue), and provide susceptibility predictions (light blue). Time spent sequencing without attaining a result is marked in yellow. For a fixed yield, this plot gives an idea of how sequencing time and success vary with the proportion of reads coming from *M. tuberculosis*.

mated (see Methods Section 3.4.6) that full susceptibility prediction would fail to be generated for 17/39 of the sputum samples sequenced here (MiSeq) with 8% *M. tuberculosis*. However, for the 11/39 samples with > 84% *M. tuberculosis*, species identification and initial susceptibility predictions would be obtained within 20 minutes of sequencing, and full results within 93 minutes (Figure 3.10 and Table 3.6).

3. SAME-DAY DIAGNOSTIC AND SURVEILLANCE DATA FOR *M. TUBERCULOSIS* VIA WHOLE-GENOME SEQUENCING OF DIRECT RESPIRATORY SAMPLES

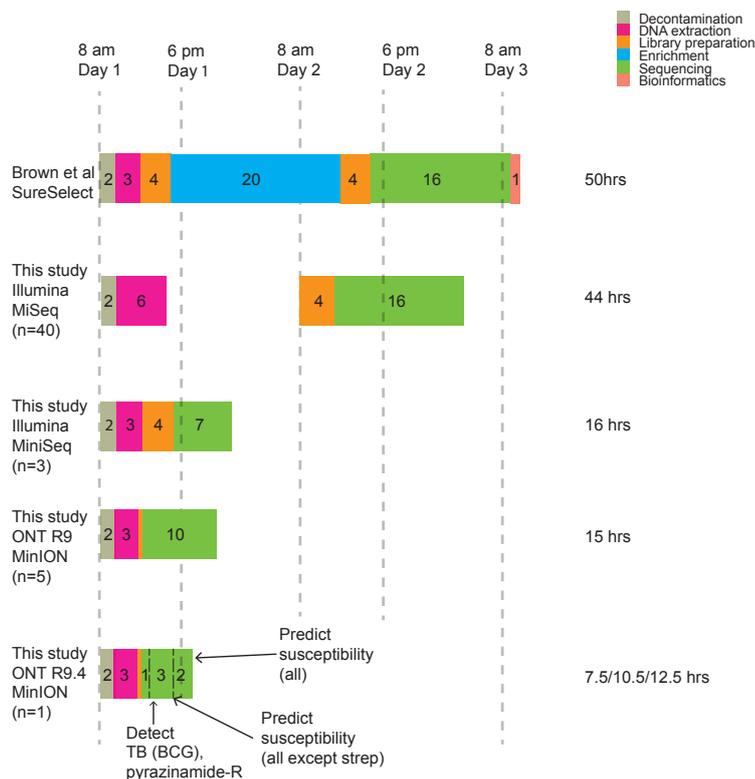


Figure 3.11: Timelines and cost of the results are shown here, using the Illumina MiSeq and MiniSeq, and the ONT MinION, compared with the method of Brown et al. 2015. It is assumed that no step of the process can be initiated after 6pm or before 8am. The method of Brown et al. 2015 has a rapid extraction step, but also a 20-hour enrichment step, resulting in a 50-hour turnaround time, at an estimated cost of £203/sample. By contrast, our extraction method with MiSeq sequencing provides results at £94/sample (Votintseva et al. 2017). The DNA extraction process was updated for the MiniSeq and MinION experiments, removing the ethanol precipitation step. In normal use this would take 3 hours. The thin orange rectangle on the MinION timelines is the 10 minute sample preparation step. This figure is intended to show comparable real-use timelines, and so the MiniSeq/MinION timelines are shown with 3 hour extraction steps. The MiniSeq enables a 16-hour turnaround time, by sequencing for only 7 hours. The R9 MinION also delivers sub-24 hour results, but requires one flow-cell per sample. The R9.4 MinION gives an 8-hour turnaround time (3 hours of sequencing with real-time (i.e., simultaneous) basecalling when used on a single sample).

3.3 Discussion

In high-incidence settings, diagnosis, DST, and treatment of *M. tuberculosis* are of utmost importance. There is a role for WGS here if the accuracy, speed, and cost can become competitive with existing techniques. In particular, WGS could improve DST where the existing approaches are either slow or have limited predictive power. Phenotypic DST is the current “gold standard” for determining resistance to antibiotics, but can take weeks to months to get results. Rapid molecular tests are used to complement culture-based DST. They provide early identification of *M. tuberculosis*, but have limited DST sensitivity on a subset of drugs. As we already saw in the previous chapter, WGS is as accurate as the current molecular assays and, as such, has the potential to overcome their limitations if the turnaround time is fast enough. Since the genomic sites are analysed in software, WGS-based DST can change the size of its catalog without impacting speed, which allows us to rapidly take advantage of improving genotype to phenotype models.

We have seen that using a novel method extracting and purifying mycobacterial DNA from primary clinical samples (developed by a collaborator, Antonina A. Votintseva) can provide sufficient data for analysis in the majority of cases. This was despite the fact that the majority of samples had poor DNA quality due to storage for long periods before processing. With higher quality samples, it seems likely that fewer samples would have had insufficient yield. The genotypes, inferred from these direct samples, were concordant with the equivalent cultured sample (when yield was sufficient to cover the whole genome). In addition, I showed that Mykrobe

3. SAME-DAY DIAGNOSTIC AND SURVEILLANCE DATA FOR *M. TUBERCULOSIS* VIA WHOLE-GENOME SEQUENCING OF DIRECT RESPIRATORY SAMPLES

`predictor` can accurately predict AMR from these data, with a turnaround time of 48 hours, two weeks faster than the approach described in the previous chapter. This approach was estimated to cost < \$100 per sample for consumables (Votintseva et al. 2017). Including consumables, staff time and overheads this is comparable to that of traditional phenotyping (and MIRU-VNTR)—£495 vs £518 (Pankhurst et al. 2016; Votintseva et al. 2017). Using an Illumina MiniSeq, the turnaround time could be improved further, albeit at an increased cost.

There are several aspects of the Oxford Nanopore’s MinION sequencer which are advantages in the context of diagnostics: 1) it is “random-access” so batching is not required; 2) it is “real-time”, so data can be analysed before the run is stopped; and, 3) it can analyse reads going through its pore—allowing for enrichment at the point of sequencing. With Illumina technology, the depth of sequencing is determined in advance. As a result, in samples with a small amount of *M. tuberculosis*, insufficient coverage could be generated resulting in failures, as we saw in 13/37 direct samples above. MinION sequencing allows sequencing to continue until sufficient coverage has been obtained, potentially avoiding this type of failure when the load is low and allowing for faster results when the load is high. In the 15% BCG-spiked sputum sample sequenced with the R9.4 MinION flowcell we saw a potential turnaround time of 12.5 hours, using a modified `Mykrobe predictor` genotyping model for ONT data.

Using simulations based on empirical R9.4 yields and empirical *M. tuberculosis* proportions (Figure 3.2), I showed that species and initial DST could be generated after only 20 minutes of sequencing in the samples with high proportions of *M.*

tuberculosis reads (Figure 3.10). In addition, this simulation also showed only 2/37 samples failing to generate results within 48 hours of sequencing, an improvement over the 13/37 failures we saw with the Illumina data, where sequencing depth could not be controlled. Although these predictions are based on a single sequencing run, the data show clear potential for a point of care test.

Both the nanopore sequencing and direct sequencing technologies are young and have scope for improvement. One possible avenue for improvement is using real-time filtering of contamination. Loose et al. 2016 demonstrated that nanopore sequencing can be used to select for particular reads at sequencing time. They did this reversing the polarity on the pore in order to reject unwanted reads. This method could be used to further enrich for mycobacterial reads, deplete non-mycobacterial reads, or even select for mycobacterial in a particular region of interest (e.g., AMR genes). The new VolTRAX device (*VolTRAX* 2017), allows automatic library preparation for the MinION. This could help potential adoption as a diagnostic. Also, improvements in the library preparation could be made to make the mycobacterial enrichment and high-sequencing yield seen in several of the samples in this study (Figure 3.2) more reliable.

In conclusion, diagnostic and surveillance information could now be obtained in 16-48 hours on Illumina MiniSeq/MiSeq platforms directly from patient specimens, a considerable step forward. This technology is still a proof of concept and more realistic trials need to be performed. In addition, the ONT's MinION sequencer may offer the same information in as little as 12 hours with several avenues of potential improvement. Currently, costs may still be prohibitive in low-resource settings,

3. SAME-DAY DIAGNOSTIC AND SURVEILLANCE DATA FOR *M. TUBERCULOSIS* VIA WHOLE-GENOME SEQUENCING OF DIRECT RESPIRATORY SAMPLES

in particular for the faster MiniSeq and MinION workflows. However, these data provide initial optimism that, at least, high-income countries can realise some of the promise of WGS-based diagnostics, and that this progress can accelerate implementation in lower-resource, high-incidence settings where there is greater potential benefit.

There is also promise that direct sequencing approaches can be used in other species, such as *S. aureus* and *Escherichia coli*. Anson et al. 2018 have recently developed a method for the preparation of bacterial DNA for WGS-based diagnostics directly from liquid blood culture. Their data were also analysed by Mykrobe predictor; 95% of AMR predictions in *S. aureus* made by Mykrobe predictor were concordant with phenotypes in 12 samples sequenced on Illumina MiSeq, and 97% concordant on 3 samples sequenced by ONT MinION.

3.4 Extended Methods

3.4.1 Sample selection and processing

Direct respiratory Ziehl-Neelsen (ZN)-positive samples with acid-fast bacilli (AFB) scorings from +1 to +3 used in this study had been originally collected from patients with subsequently confirmed *M. tuberculosis* infections at the John Radcliffe Hospital, Oxford Universities NHS Foundation Trust, Oxford, UK (n=18), and Birmingham Heartlands Hospital NHS Foundation Trust, Birmingham, UK (n=22). 2/18 Oxford samples were culture negative specimens taken 2.5 months apart from the same patient undergoing treatment for *M. tuberculosis*. If available, corresponding Mycobacterial Growth Indicator Tube (MGIT) cultures were collected for each direct sample (Oxford n=11, Birmingham n=17). Two ZN and culture-negative direct respiratory samples were also collected from the John Radcliffe Hospital.

The discarded direct samples were collected only after sufficient material had been obtained for the routine diagnostic workflow, including the requirement to ensure that enough sample volume remained if re-culture was requested. Consequently, study samples were of lower volume and quality than would be the case if the method were used routinely. While waiting for the routine laboratory results, samples were stored at +4C and later processed in batches of 5-12. All ZN-positive samples were digested and decontaminated with NAC-PAC RED kit (AlphaTec, USA). Direct samples and corresponding MGIT culture aliquots (1 mL) were heat inactivated in a thermal block after sonication (20 min, 35 kHz) for 30 min and

2h at 95C, respectively. MGITs were inactivated for 2 hours owing to their high bacterial load. Before DNA extraction samples were stored at +4C.

A lower than recommended DNA concentration threshold for MiSeq library preparation ($0.05 \text{ ng } \mu\text{l}^{-1}$ was used rather than $0.2 \text{ ng } \mu\text{l}^{-1}$) on the basis of previous experience of sequencing mycobacterial cultures with suboptimal amounts of DNA. 6/40 (15%) samples yielded DNA below the $0.2 \text{ ng } \mu\text{l}^{-1}$ threshold.

3.4.2 DNA extraction and Illumina MiSeq sequencing

Mycobacterial DNA from MGIT cultures was extracted using a previously validated ethanol precipitation method (19). DNA from ZN-positive direct samples was extracted using a modified version of this protocol. These modifications included a saline wash followed by MolYsis Basic5 kit (Molzym, Germany) treatment for the removal of human DNA, and addition of GlycoBlue co-precipitant (LifeTechnologies, USA) to the ethanol precipitation step.

Libraries were prepared for the MiSeq Illumina sequencing using a modified Illumina Nextera XT protocol (19). Samples were sequenced using the MiSeq Reagent Kit v2, 2 x 150bp in batches of 9-12 per flow-cell. Median library size (Tapestation, Agilent, USA) was 627 bp (IQR 495 – 681). Median reads available per sample were 3.2 million (IQR 2.8 – 4.1 million); this would yield a median depth of approximately 213, given pure *M. tuberculosis*, although it was anticipated that non-mycobacterial reads would be present.

3.4.3 DNA extraction for ONT MinION and Illumina MiniSeq sequencing

ZN/culture-negative sputum and BCG (Pasteur strain; cultivated at 37 °C in MGIT tubes) DNA was extracted using a modified version of that in Votintseva et al. [2015](#). Following a saline wash, samples were re-suspended in 100 µl of molecular grade water and subjected to three rounds of bead-beating at 6 m/s for 40 seconds. The beads were pelleted by centrifugation at 16,100 xg for 10 minutes and 50 µl supernatant cleaned using 1.8x volume AMPure beads (Beckman Coulter, UK). Samples were eluted in 25 µl molecular grade water, and quantified using the Qubit fluorimeter (Thermo Fisher Scientific, USA).

3.4.4 MiniSeq sequencing

Extracted ZN-negative sputum DNA and pure BCG DNA were combined in a 50:50 ratio (0.5 ng each) and libraries prepared alongside pure BCG DNA (1 ng) using a modified Illumina Nextera XT protocol (19). BCG and two BCG+sputum DNA samples were sequenced at Illumina Cambridge Ltd. UK, using a Mid Output kit (FC-420-1004) reading 15 tiles and with 101 cycles.

3.4.5 MinION Sequencing

All MinION sequencing utilised the best available sample preparation kit for our samples and flow cells (R9/R9.4 flowcells and PCR-based sample preparation,

3. SAME-DAY DIAGNOSTIC AND SURVEILLANCE DATA FOR *M. TUBERCULOSIS* VIA WHOLE-GENOME SEQUENCING OF DIRECT RESPIRATORY SAMPLES

as described below). A single ZN-negative sputum extract was divided into three equal concentration aliquots (187 ng), and BCG DNA added at 5%, 10%, and 15% of the total sputum DNA concentration. These 5-15% spikes represent the lower end of the spectrum seen in the MiSeq samples above (see Figure 2a). These samples, along with pure BCG DNA, were prepared following ONTs PCR-based protocol for low-input libraries (DP006_revB_14Aug2015), using modified primers supplied by ONT, a 20 ng DNA input into the PCR reaction, and LongAmp Taq 2X Master Mix (New England Biolabs, USA). PCR conditions were as follows: initial denaturation at 95 °C for 3 minutes, followed by 18 cycles of 95 °C for 15s, 62 °C for 15s, and 65 °C for 2.5 minutes, and a final extension at 65 °C for 5 minutes. Samples were cleaned in 0.4x volume AMPure beads and the PCR product assessed using the Qubit fluorimeter and TapeStation (Agilent, UK). The final elution was into 10 µl 50 µM NaCl, 10 µM Tris.HCl pH8.0. Finally, 1 µl of PCR-Rapid Adapter (PCR-RAD; supplied by ONT) was added and samples incubated for 5 minutes at room temperature to generate pre-sequencing mix. The pre-sequencing mix was prepared for loading onto flow cells following standard ONT protocols, with a loading concentration of 50–100 fmol.

Using the 15% BCG spiked sputum DNA prepared above, amplification was repeated using Phusion High-Fidelity PCR Master Mix with DMSO (New England BioLabs, USA). Gradient PCR was performed to identify the optimal annealing temperature for recovery of BCG DNA (data not shown). Final PCR conditions were as follows: initial denaturation at 98 °C for 30s, followed by 18 cycles of 98 °C for 10s, 59 °C for 15s, and 72 °C for 1.5 minutes, and a final extension of 72 °C for

10 minutes. Following PCR, the sample was prepared for sequencing, as described above. The final loading concentration was approximately 27 fmol.

The above samples were sequenced using R9 spot-on generation flow cells and the 48-hour protocol for FLO-MIN105 (ONT, UK). Base-calling was performed via the Metrichor EPI2ME service (ONT, UK) using the 1D RNN for SQK-RAD001 v1.107 workflow.

Subsequently, a new 15% BCG spiked sputum was prepared, as described above, using Phusion Master Mix with DMSO. Sequencing was performed using R9.4 spot-on generation flow cells and the 48-hour FLO-MIN106 protocol (ONT, UK). Final loading concentration was 43 fmol. Base-calling was performed after sequencing was complete using Albacore (ONT, UK), as base-calling via Metrichor failed. Subsequent tests on other samples (data not shown) showed that base-calling could have been performed in real-time, i.e., during the run.

3.4.6 Bioinformatic analysis of MinION data

`Mykrobe predictor` (version v0.5.0-6-g6b19d83) was used to predict resistance from the MinION base-called reads using the modified genotyping model above (Section 3.2.5.1). Yield and timing were analysed using `Poretools` (Loman et al. 2014). For the R9.4 sample, `Mykrobe predictor` was applied to the cumulative read output at each hour. Yield of BCG was measured by mapping to a BCG reference (accession BX248333.1). Phylogenetic placement of the 15% spike BCG sample sequenced on MinION R9.4 was achieved as for the Illumina data—by genotyping

3. SAME-DAY DIAGNOSTIC AND SURVEILLANCE DATA FOR *M. TUBERCULOSIS* VIA WHOLE-GENOME SEQUENCING OF DIRECT RESPIRATORY SAMPLES

68695 SNPs, and choosing the leaf with the fewest SNP differences across those sites.

CHAPTER 4

Enabling rapid DNA search of all sequenced bacteria and viruses

In this chapter I consider the problem of sequence search in very large collections of whole-genome sequencing (WGS) data sets. As DNA sequencing becomes ubiquitous due to falling costs, bacterial sequence archives—a huge trove of information on infection, disease, biology, and evolution—are growing exponentially, doubling every two years (*EBI Statistics 2017*). However, although technically public, most of these data are almost entirely inaccessible to search.

Basic local alignment search (BLAST) and similar tools provide search within reference databases (Altschul et al. 1990). These contain curated, mostly non-redundant data—exemplars of species that often do not represent the full diversity of the species’ pan-genome. However, reference databases constitute only 20% of the data sets in the microbial archives (110,898 assemblies versus 554,680 raw read data

4. ENABLING RAPID DNA SEARCH OF ALL SEQUENCED BACTERIA AND VIRUSES

sets in the European Nucleotide Archive (ENA) as of October 2017), and cannot be indexed by alignment search tools like BLAST and, as a result, are currently not indexed for search.

In this chapter, I use probabilistic k -mer indexing to enable search in millions of diverse sequence data sets with a novel implementation of a coloured probabilistic de Bruijn graph, “Coloured Bloom Graph” (CBG). I use this data structure to index all 447,833 bacterial and viral whole-genome sequenced data that had been generated and deposited in the European Nucleotide Archive as of December 2016 (which I will refer to as the `microbial-CBG`); an achievement orders of magnitude beyond the ability of previous methods. Coloured Bloom Graphs enable fast search in these data. I validate this tool by demonstrating accurate variant typing and multi-locus sequencing typing compared with validated mapping tools.

I demonstrate its use with several applications: i) rapid sequence search in the ENA, ii) antimicrobial resistance surveillance of both genes and SNPs, iii) investigating the host range of mobile genetic elements, and iv) showing how to set up a threat-alert for cross-host transmission of multi-drug-resistant plasmids from *Salmonella* in cattle to *Y. pestis*. In doing so, I hope to demonstrate that these archives and future collections of data can be made accessible to arbitrary sequence and variant search, opening up new possibilities for retrospective analysis and future surveillance.

¹The ‘Coloured Bloom Graph’ (CBG) is referred to as ‘Bitsliced Genomic Signature Index’ (BIGSI) in this publication. There is no difference between the tools.

4.0.1 Publication note and acknowledgements

A preprint of this work has been published in Bradley et al. 2017 and has been submitted for publication¹. Although the work was collaborative in parts, the work described here is the sole work of myself, with the guidance of my supervisors. However, I would like to acknowledge: Henk den Bakker for his assistance with accessing the NARMS metadata discussed in section 4.9, Eduardo Rocha for sharing T4SS and relaxase genes used in section 4.8.1.1, as well as both their consultation on this work; Jerome Kelleher for his advice on improving the performance of CBG with SSDs; and Robert Esnouf and Guy Cochrane for their enormous assistance with data download from the EBI.

The microbial-CBG discussed in section 4.5 has already been used by Wang et al. 2017 in their work on characterising the global distribution and spread of the colistin resistance gene *mcr-1*. My small contribution to this work is described in section 4.7.2.

4.1 Introduction

Sequence search is a fundamental operation of genomic research: ascertaining which isolates in a collection have a specific sequence, determining reference sequence with high sequence similarity to a query sequence, estimating population frequencies of alleles, and metagenomic read assignment can all be described in terms of sequence search. BLAST (Altschul et al. 1990), and its variants PSI-BLAST (Altschul et al.

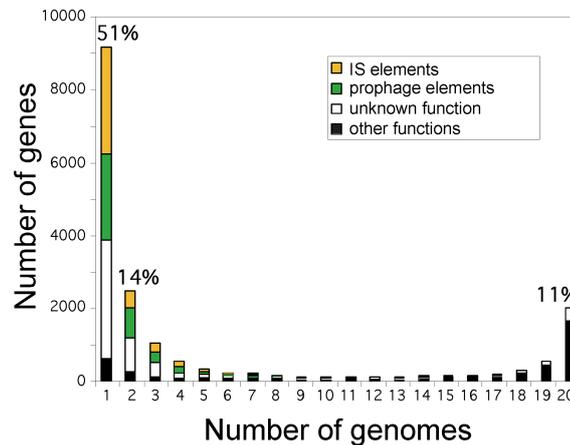
1997), megaBLAST (Zhang et al. 2000), and impala (A. Schäffer et al. 1999), are among the most popular tools that provide this function and are currently the default method for searching the two major assembled sequence databases: the National Center for Biotechnology Information’s (NCBI) RefSeq (O’Leary et al. 2015) and the European Nucleotide Archive’s genome assembly database (ENA).

However, BLAST cannot search the two major raw read databases NCBI’s “Short Read Archive” (SRA) and ENA’s “Read” database. As a result, it is currently not possible to search our raw read archives in reasonable computational time. SRA-BLAST provides search in the SRA, but is limited to a subset of manually chosen experiments and requires the read-length of the experiment to be greater than that of the query sequence (Bethesda 2011). It is also too slow for many applications; Solomon et al. 2016 estimated that SRA-BLAST would take over two days to search for a single transcript in 2,652 RNA-seq experiments. Thus, it has limited practical use since many interesting queries exceed read length and/or require fast search. Also, knowing the subset of samples to search in advance is not always possible.

4.1.1 Microbial diversity is a challenge for building searchable indexes

In order to index the microbial ENA/SRA, we first need to understand the dimension of the search space: the microbial pan-genome, i.e., the collection of all unique sequence observed in microbes. In bacteria, the diversity in the pan-genome is dominated by a large pool of low frequency genes and mobile genetic elements—

Figure 4.1: At one extreme of the x-axis are the genes present in a single genome, which are regarded as strain-specific genes (9054 genes: 51% of the pan-genome), while at the opposite end of the scale are situated the genes found in all 20 genomes, which represent the *Escherichia coli* core genome (1976 genes: 11% of the pan-genome). Coloured rectangles represent the proportion of insertion sequence (IS)-like elements (yellow), prophage-like elements (green), and genes of unknown/unclassified function (white). Black rectangles represent genes for which a function can be assigned. Strain-specific genes correspond to 2885 IS-like elements (32%), 2352 prophage-like elements (26%), and 3220 genes of unknown/unclassified function (35%). From (Touchon et al. 2009) (CC BY).



the accessory genome (Lapierre et al. 2009). The accessory genome contains far more unique sequence than the highly shared core (Lapierre et al. 2009). For example, Touchon et al. 2009 demonstrated that in *Escherichia coli* most genes of the pan-genome exist in very few genomes, and as a result the *Escherichia coli* pan-genome is dominated by rare sequence (see Figure 4.1). This is in contrast to species without an accessory genome, like *homo sapiens*, whose pan-genome is dominated by sequence shared between a high proportion of individuals (Li et al. 2010b). Later, I will refer to these regimes as “high k -mer sharing” and “low k -mer sharing” collections.

Attempts to estimate the size of the microbial pan-genome show that it is very large, likely consisting of at least a billion genes (Lapierre et al. 2009; Medini et al. 2005). Admittedly, these claims are built on a limited amount of data and depend heavily on the model parameters (Vernikos et al. 2015). However, if we consider that there are an estimated 10^{31} bacteriophages (Hendrix 2003), which infect 10^{24} bacteria per second, there is a huge incentive for evolution and gene flow of novel sequence in microbes. If, in addition, we add noise due to sequencing and analysis errors, it is not unreasonable to assume that the cardinality of unique k -mers contained within all of life's genomes may only be bounded by the theoretical maximum of 4^k (e.g. 10^{18} for $k = 31$) as the number of sequence data sets grow.

Although a bacterial species' pan-genome may be large, each individual genome within it contains a similar number of k -mers. For example, *Escherichia coli* genomes range in size only from 4.56 Mbp to 5.70 Mbp, but its pan-genome is orders of magnitude larger (Lukjancenko et al. 2010; Touchon et al. 2009). The largest known bacterial genome is 14.8Mbp (Han et al. 2013), and within a species or genus the maximum expected size (excluding contamination) of any data set can be estimated by looking at the distribution within a subset of samples. In contrast, the number of k -mers in the pan-genome is often not known and difficult to estimate.

This diversity of sequence and resulting large numbers of unique k -mers (“terms”, “n-grams”, or “shingles” in text-search terminology) poses a significant problem for traditional large-scale text indexing, such as web search, which primarily uses “inverted indexes” (Brin et al. 2012). Inverted indexes map all the unique terms in the corpus to the documents (also called “records” or “data sets”) in which they

are contained. Using inverted indexes to search the web relies on the fact that the number of unique terms grows slowly with the addition of new documents (after all the common terms have been indexed). Naively, index size would be bounded by the number words in natural languages ($\approx 10^6$ – 10^7). In practice, even with rare terms, the number of unique terms indexed by commercial search engines is on the same order as the number of indexed documents (Brin et al. 2012; Goodwin et al. 2017; Zobel et al. 1998) ($\approx 10^{10}$ – 10^{12}). Google reportedly indexes 10^{11} documents and requires 100PB of storage to do so (*Google–How Search Works–Crawling and Indexing* 2017) and likely indexes $\approx 10^{10} - 10^{12}$ unique terms if trends from Brin et al. 2012 have continued. This is far fewer terms than the number of possible k -mers in the same number of WGS data sets.

Recently, progress has been made toward enabling sequence search on large raw read archives. Solomon and Kingsford introduced the sequence bloom tree (SBT) and demonstrated its use in indexing and searching 2,652 raw read human RNA-seq data sets (Solomon et al. 2016) using a hierarchy of probabilistic data structures (bloom filters) and on-disk indexing. Unfortunately, as we will discuss in more detail in Section 4.2.6, SBTs either scale super-linearly with new data sets or suffer detrimental impact on performance as a result of bloom filter saturation. However, these limitations are unique to Solomon and Kingsford’s implementation and can be resolved, as I will discuss in section 4.3.

In section 4.3, I describe an index which takes advantage of the fact that microbial genomes have bounded size ($\approx 10^6$ k -mers), but a pan-genome orders of magnitude larger (possibly $\geq 10^{18}$ k -mers) and makes trade-offs that align closely

with their biology. I do this by developing an indexing structure, a novel implementation of a coloured probabilistic de Bruijn graph, whose size does not depend on the total number of k -mers (“terms”) in the collection, but which is restricted by a maximum per data set k -mer cardinality (which can be estimated and set in advance). In doing so, I remove the scaling dependency on the total number of k -mers, which causes scaling problems for other k -mer indexing techniques (including SBT).

4.2 Background

4.2.1 Probabilistic coloured de Bruijn graphs

De Bruijn graphs represent sequence data as a set of k -mer nodes (i.e., overlapping sub-sequences of a fixed length) and edges between k -mers that overlap by $k-1$ bps, as described in section 1.6.4.1. Since k -mers in a de Bruijn graph have an edge between them if they are overlapping, the edges of a de Bruijn graph do not need to be stored explicitly. By querying the set of k -mers in the de Bruijn graph for the presence of the 4 possible neighbours of each k -mer, the edges can be inferred. The set of all k -mers, without edges, represents an *implicit* de Bruijn graph.

In fact, as noticed by Pell et al. 2012, one does not even need to store k -mers themselves. If you can query for presence, or *set membership*, of a k -mer, then you can represent an implicit de Bruijn graph even if you do not store an exact representation of the set. Bloom filters provide exactly this operation.

4.2.2 Bloom filters

Bloom filters are space-efficient probabilistic data structures, which allow set membership queries (Bloom 1970). The data structure does not store the elements of the set explicitly. Instead, it stores a low-memory representation of the set, which supports set probabilistic membership queries. Elements can be added to the bloom filter, but they cannot be removed. There is a probability of a false positive from a set-membership query, which increases as more elements are added to the set. However, this error is one-sided and false-negative results cannot occur. As a result, bloom filters can tell you that an element is *probably* in a set, or that an element is not in the set. For a given cardinality (n), the false-positive rate can be modified by changing the bloom filter parameters (see Section 4.2.2.2). Despite its trade-offs, the small space requirements make the bloom filter desirable for many applications and have been used previously in the context of document retrieval and text indexing as a type of document ‘signature’ (Shepherd et al. 1989; Wong et al. 1985; Zobel et al. 1998), as well as in sequence graphs (Chikhi et al. 2013b; Pell et al. 2012), amongst other applications (Broder et al. 2004).

4.2.2.1 Bloom filter construction and querying

Each bloom filter is a bit vector of length m , initialised with 0 at all positions. To add an element to a bloom filter we apply η ² unique hash functions to the

²The parameter determining the number of hash functions in a bloom filter is commonly referred to as k . I have used η here to avoid any confusion with k -mer-size which I have denoted as k elsewhere.

4. ENABLING RAPID DNA SEARCH OF ALL SEQUENCED BACTERIA AND VIRUSES

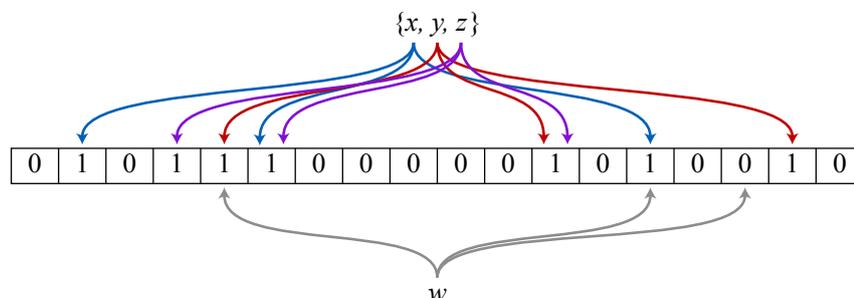


Figure 4.2: A schematic of a bloom filter: Elements $\{x, y, z\}$ are hashed $\eta = 3$ times and the bits at these positions are set to 1. w is queried by hashing 3 times by the same hash functions. Since not all bits at these positions are 1 w is not in the set. False-positive results can occur if the queried positions were all set to 1 by other elements. The probability of a false positive can be tuned by the bit vector's length and the number of hash functions applied to each element. By David Eppstein [Public domain], via [Wikimedia Commons](#).

element to get an array of η positions. The bits at the resulting positions are set to one. See Figure 4.2.

To query if an element (w) is a member of the set, we apply the same η hash functions to w to get an array of η positions. The result of the query is found by taking the bit-wise AND of the bits at the resulting positions: 0 if the element is not in the set and 1 otherwise. False positives occur when the bits at all η positions are set by insertion of other elements.

4.2.2.2 Choosing bloom filter parameters

There are two parameters that need to be set when constructing bloom filters. These are the bloom filter length (m) and the number of hash functions used (η). These determine the approximate false-positive rate (p) for a given set cardinality (n) as:

$$p \approx (1 - e^{-\frac{\eta m}{m}})^\eta,$$

as shown in (Bloom 1970).

As shown by Broder et al. 2004, this can be reversed, so given an expected set cardinality (n) and a desired false-positive rate (p), the optimal values of m and η can be calculated from:

$$m = \frac{-n \ln p}{\ln 2^2} \quad \eta = -\frac{\ln p}{\ln 2}. \quad (4.1)$$

4.2.3 Probabilistic de Bruijn graphs

Since an implicit de Bruijn graph can be represented by knowing which k -mers are present in a set of k -mers, a bloom filter can be used to encode an implicit de Bruijn graph with the possibility of false-positive nodes: a *probabilistic* (and implicit) de Bruijn graph. Pell et al. 2012 showed that a bloom filter can be used to represent an implicit de Bruijn graph in as little as 4bits per k -mer. However, they noted that the false positives resulting from the bloom filter occasionally result in false branching with the graph. Chikhi et al. 2013b showed that by extending this data structure (i.e., by storing the false branching nodes explicitly), they were able to represent an exact de Bruijn graph. This required a small amount of additional space to store the *critical false positive*, i.e., false-positive k -mers that caused false branching within the graph. Using this exact de Bruijn graph based on a bloom

filter, they were able to assemble a human genome using only 5.7GB of memory in less than 24 hours.

4.2.4 Coloured de Bruijn Graphs

A coloured de Bruijn graph generalises the de Bruijn graph to multiple samples, or ‘colours’, by associating each k -mer with its presence or absence in a set of samples. This allows storing of shared sequence in a single data structure where each unique k -mer is only stored once, along with its presence/absence in each colour. Coloured de Bruijn graphs have found application in variant discovery and genotyping with implementations such as `cortex` (Iqbal et al. 2012), `mccortex` (Turner et al. 2017), `BFT` (Holley et al. 2016), and `vari` (Belk et al. 2016). Implicit coloured de Bruijn graphs can all be thought of as inverted k -mer indexes, mapping k -mers to the data sets in which they are contained.

Coloured de Bruijn Graphs compress when there are few k -mers in many colours (since storing colour presence requires less space than a node). Human data, where one expects high sharing of sequence between samples (Li et al. 2010b), allow these tools to scale near-linearly with the number of colours/samples. However, they do not compress well when the union of k -mers is large, but each k -mer intersects colours infrequently. Microbes with more diverse pan-genomes, and a resulting large number of possible k -mers, archives with multiple species, and metagenomic data all present difficulties to coloured de Bruijn graphs and approaches based on inverted indexes. As such, most implementations of coloured de Bruijn graphs scale well

in the high k -mer sharing regime, but poorly otherwise. This is discussed in more detail in section 4.4.1, and a simulation of scaling in high and low k -mer sharing regimes is visualised in Figure 4.5.

4.2.5 Coloured probabilistic de Bruijn Graphs

In order to represent a coloured *probabilistic* de Bruijn Graph we need to index a collection of *probabilistic* de Bruijn Graphs so that colour membership can be queried from a single data structure, without explicitly storing the k -mers in the graph. Since an implicit coloured *probabilistic* de Bruijn Graphs maps k -mers to the data sets which contain them, they can be used to provide the same function as inverted indexes, and also be thought of as *probabilistic* inverted k -mer indexes. There is such a data structure which supports indexing of a set of probabilistic de Bruijn graphs, and thus implements a coloured probabilistic de Bruijn graph: the sequence bloom tree.

4.2.6 Sequence bloom trees

Recently, Solomon and Kingsford proposed the ‘sequence bloom tree’ (SBT) as a method for indexing a collection of sequence data sets (Solomon et al. 2016). The SBT indexes a collection of bloom filters constructed from all k -mers within each data set. The SBT is a binary tree, which indexes these bloom filters as leaves in the tree. The internal nodes are bloom filters that contain the union of their two children.

4. ENABLING RAPID DNA SEARCH OF ALL SEQUENCED BACTERIA AND VIRUSES

K -mer set membership queries in a SBT can rapidly traverse the tree to find the bloom filters they are contained within. Since a SBT indexes a set of k -mer bloom filters, which encode probabilistic de Bruijn graphs, it can be considered as an implementation of a coloured probabilistic de Bruijn graph, which can also be thought of as probabilistic inverted k -mer indexes, or “signature index”. Solomon and Kingsford used the SBT to search for transcripts in raw human RNA-seq data sets.

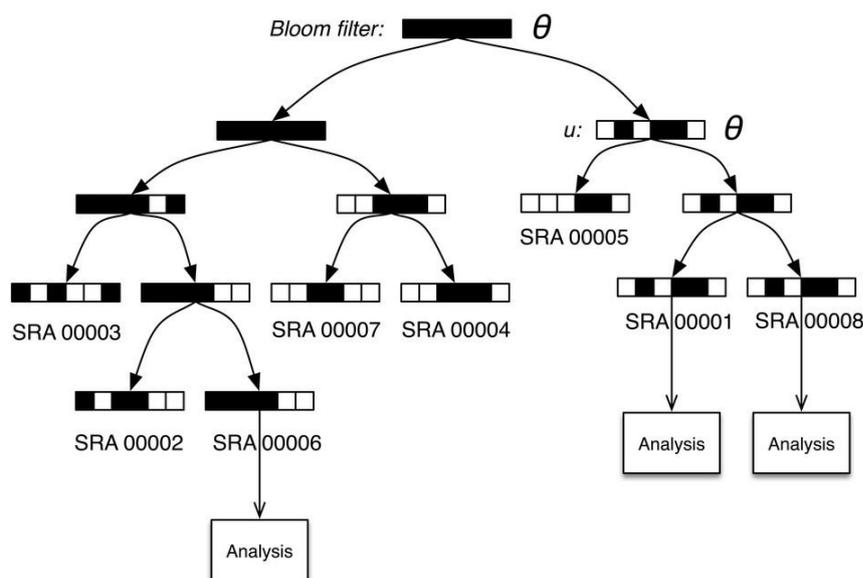
The disadvantage of SBTs is that in order to maintain the logarithmic time tree traversal, the bloom filters of the internal nodes must not saturate. Bloom filter saturation occurs when too many bits are set in the bloom filter’s bit vector and the probability of a false-positive set membership query approaches one, see Figure 4.3, where the root node exhibits saturation.

SBT saturation is avoided by choosing bloom filter parameters such that the bloom filter at the root of the tree can contain all of the k -mers in all descendant bloom filters without saturation. In the case where you have a collection of diverse sequences, such as bacteria with open pan-genomes or several species from a range of taxa, the bloom filters can become prohibitively large in order to prevent saturation. Also, in many cases the total number of k -mers in the entire collections of sequences is difficult to estimate in advance of construction. This requires rebuilding of the tree with different parameters if the internal nodes saturate; otherwise the query-time performance will decrease as more sequences are added.

As a result, the SBT has limitations, which prevent it from being a tool that can practically be used to index a set of samples as diverse as those contained within

4.3. Results: A space-efficient representation of a probabilistic coloured de Bruijn graph

Figure 4.3: A schematic of a Sequence Bloom Tree: each internal node holds the union of the two children. Saturation of the nodes at the top of the tree can be seen, requiring traversal down both children regardless of query. From Solomon et al. 2016 and reproduced with permission.



the microbial SRA/ENA archives. This is the constraint which I try to address with ‘Coloured Bloom Graphs’ described in the next section.

4.3 Results: A space-efficient representation of a probabilistic coloured de Bruijn graph

Below, I outline a novel representation of a coloured probabilistic de Bruijn graph, the Coloured Bloom Graph (CBG), along with algorithms for construction and querying (Figure 4.4). A CBG has a fixed storage requirement per colour and grows independently of the total number of k -mers contained within the collections

of sequences—a key distinction from the data structures discussed in section 4.2.4 and section 4.2.5. Despite also using bloom filters, CBGs do not suffer from saturation as SBTs do, and additional sequences can be inserted into a CBG in an incremental manner without re-construction, regardless of the number of new k -mers they contain.

This means a CBG's size does not depend on the number of k -mers in the pan-genome, an advantage when there are many more k -mers in the pan-genome than in any individual of the collection. This also has the secondary advantage that the maximum number of k -mers in each sample is easier to estimate than the cardinality of k -mers in the whole collection (with perhaps the exception of metagenomic data sets, which I discuss later in section 4.10).

CBG bloom filter parameters are determined by the maximum expected k -mers in each colour—not the expected k -mers in the pan-genome. This is achieved via 'bit-sliced' indexing (Wong et al. 1985) of a set of bloom filters, which I describe in more detail in the next section.

4.3.1 Coloured Bloom Graph construction and querying

The CBG indexes a set of N bloom filters, by position in the bloom filter. Conceptually, a CBG is a bit matrix (see Figure 4.4c), where each column is a bloom filter, indexed by row (or bloom filter position), such that row lookups are constant time.

Each bloom filter must be constructed with the same parameters (m, η) and can

4.3. Results: A space-efficient representation of a probabilistic coloured de Bruijn graph

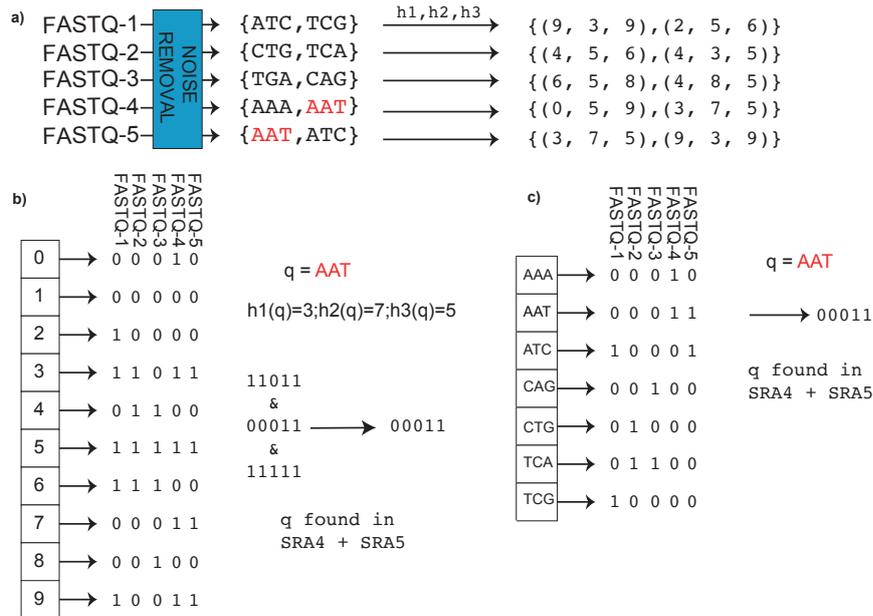


Figure 4.4: CBG encoding compared with naïve approach. a) CBG step 1: each input data set (could be raw sequence data (FASTQ format) or assembly) is converted to a non-redundant list of k -mers (with an optional denoising step to remove sequencing errors). A fixed set of η hash functions (h_1, h_2, \dots) is applied to each k -mer ($\eta = 3$ in this figure), giving a tuple of positions which are all be set to 1 in a bit-vector (a Bloom Filter). b) CBG step 2. Each data set is stored as a fixed length bloom filter, as a column in a rectangular matrix. To query the CBG for k -mer AAT, the η hash functions are applied to the query k -mer, returning η rows to be checked (namely 3,7,5 here). All columns (datasets) that have 1 in all of those η rows contain the query k -mer: these rows that are checked are called “bitslices”. Adding a new data set requires just adding a new column. c) Naïve encoding for contrast. A complete list of all k -mers in all datasets form the rows of a large matrix, and columns are datasets. For any given k -mer, entries are set to one for data sets containing that k -mer. When a new data set is added, the matrix grows vertically (new k -mers added) and horizontally (new column for new data set).

be considered a single-colour probabilistic de Bruijn Graph. Given a set of N bloom filters, to construct the bloom filter matrix, we column-wise concatenate the bloom filters into a matrix. The rows of this bit matrix can be indexed as a set of key-value pairs, where each key is the bloom filter position and value is a bit vector consisting of the 1/0 value from all of the N bloom filters at that position. The rows of this matrix are inserted into a hash table, or key-value store allowing $O(1)$ lookup at a position in the bloom filter of all colours. This set of key-value pairs can be stored on disk, in memory, or distributed across several machines. To insert a colour we simply append the bloom filter as a new column in the existing bit matrix.

To query the CBG for a k -mer we hash the k -mer η times, lookup the resulting keys in the key-value store, and take the bit wise AND of the resulting bit vectors. This results in a bit vector showing k -mer presence/absence in each colour in the CBG (see Figure 4.4c).

An open source implementation of CBG is available at <https://github.com/phelimb/CBG>.

4.3.2 Coloured Bloom Graph sequence search algorithm

To search for an arbitrary sequence query q of length L_q base-pairs in a CBG, we break q into $\mathcal{L} = L_q - k + 1$ k -mers. Then, we query the CBG for each of these \mathcal{L} k -mers independently following the procedure outlined in section 4.3.1, resulting in \mathcal{L} bit vectors. This step is parallelisable. A threshold (T) determines the proportion of k -mers required to be found in the query in order to record the query as present

in a given colour.

4.3.2.1 Inexact search algorithm

If $T < 1$ we take the sum of bit vectors resulting from the k -mer queries (converting to an integer array) and divide the resulting array by \mathcal{L} to get the proportion of k -mers present in each colour (f_c). If $f_c > T$ the query is found. K -mers could be weighted such that queries are returned if $f_c w_c > T$, but in all analyses in this thesis I used unweighted search.

4.3.2.2 Exact search algorithm

$T = 1$ is a special case of the query algorithm. If $T = 1$, we require every k -mer to be present in each colour and as a result we can take the bit-wise AND of the bit vectors resulting from each k -mer query (without converting to an integer array). This results in a single bit vector of 1s where the query q is present in the colour at that position and 0 otherwise. Exact search can be computed much faster than the inexact equivalent, as it simply requires $\eta \times \mathcal{L}$ bit-wise AND operations. Although bitwise AND operations are $O(N)$ operations, they are extremely efficient.

4.3.2.3 Variant search and genotyping algorithm

CBG variant search and genotyping uses a modification of the probe creation and genotyping algorithm described in section 2.4.1. The variant genotyping problem is converted into a sequence search problem by generating sequence ‘probe-sets’

for each variant. This *in-silico* version of genotyping arrays is described in section 2.4.1.1. Each allele of the probe-set is searched via the exact search algorithm described above, resulting in Boolean presence/absence of each allele. If only a reference allele is present, then the genotype is returned as 0/0, if only an alternate allele then as 1/1, if both 0/1, and if neither -/-.

4.3.2.4 Controlling the false positive-rate of variant queries

To control the false discovery rate for an allele from a probe set, we enforce all k -mers from an allele of length ($L=2k-1$, $\mathcal{L} = k$) resulting in a false-positive rate per colour of p^k . For example, with parameters $m=2.5 \times 10^7$; $\eta = 3$; $k = 31$; $K_{max} = 10^7$, the false-positive rate per k -mer is $p \approx 0.3412\dots$ and the expected false discovery rate of an allele is $0.3412^{31} \approx 10^{-15}$ per colour, which is well below the expected error rate from the underlying data.

4.3.3 Choosing CBG parameters

The choice of CBG parameters depends on: the maximum number of k -mers expected in any colour (K_{max}), the number of samples/colours (N) expected, the shortest length of the query sequence to be supported (L_{min}), your k -mer size (k), and your desired maximum number of false discoveries per query (v).

The parameters that can be chosen when first constructing a CBG are the length of the bloom filter (m), which is also the number of rows in the bit matrix, and the number of hash functions used (η).

4.3. Results: A space-efficient representation of a probabilistic coloured de Bruijn graph

Since each query L will consist of $\mathcal{L} = L - k + 1$ k -mers, and assuming the false-positive probability of each k -mer in each colour (p) is independent, the number of false discoveries (\mathcal{V}) for any query can be calculated as:

$$\mathcal{V} = E[V] = Np^{\mathcal{L}}.$$

The number of expected false discoveries for a query (\mathcal{V}) increases with shorter query size (L), so to keep \mathcal{V} below a chosen threshold v ($v < \mathcal{V}$) for a given N and k , p must be chosen to satisfy v for L_{min} ($\mathcal{L}_{min} = L_{min} - k + 1$):

$$p^{\mathcal{L}_{min}} = (v/N). \tag{4.2}$$

The maximum false-positive rate per colour, p depends on your bloom filter size (m), number of hashes (η), and the total number of k -mers per sample (K_{max}). You can optimise your choice of m and η based on Equation 4.2.2.2

So, for example, given $N = 10^6$, $K_{max} = 10^7$, $L_{min} = 50bp$, and $k = 31$, $v = 10^{-6}$ the resulting expected number of false positives per k -mer per bloom filter (p) would be:

$$p = (v/N)^{\frac{1}{\mathcal{L}_{min}}} = \frac{1}{10^{\frac{3}{5}}} = 0.2511\dots \tag{4.3}$$

using Equation 4.2.2.2, this determines optimal CBG parameters of:

$$m = 28,755,176; \eta = 2$$

Queries where $L > L_{min}$ will have $\mathcal{V} < v$ and $L < L_{min}$ queries will have $\mathcal{V} > v$.

4. ENABLING RAPID DNA SEARCH OF ALL SEQUENCED BACTERIA AND VIRUSES

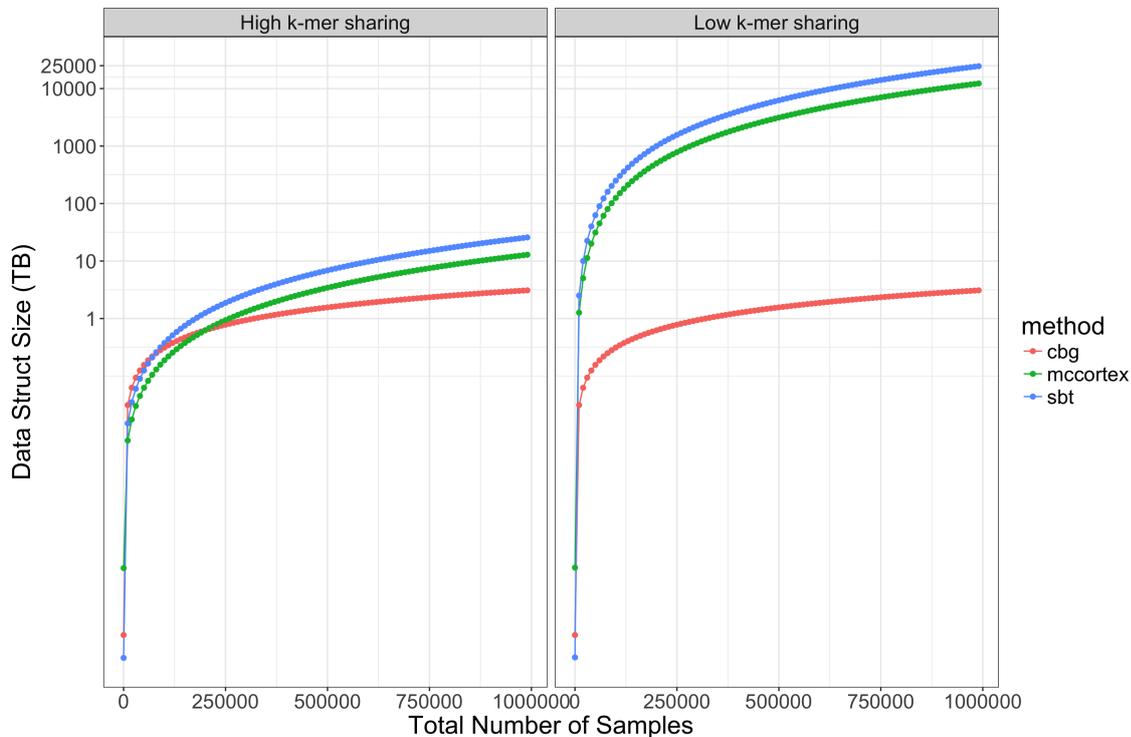


Figure 4.5: Simulated scaling of CBG, SBT, and `mccortex` to 1 million data sets with high/low proportion of sharing of k -mers between samples. In the high k -mer sharing regime, 10,000 new k -mers are introduced per sample, whereas the low k -mer sharing regime only introduces 100 new k -mers per sample. Since CBG scales linearly with N and independently of the total number of k -mers (N_k), it uses the same storage per sample in each regime. However, SBT and `mccortex` scale super-linearly with N as a result of their dependency on N_k .

4.4 Results: Computational performance and validation

4.4.1 Simulated scaling to 1 million samples

In order to estimate the storage required to index 1 million data sets, I simulated the scaling of storage requirements required to construct a `mccortex` graph, a SBT,

150

and a CBG, for data sizes up to 1 million genomes in two regimes: first, for genomes with high proportions of k -mer-sharing (e.g., *M. tuberculosis* or human), and second to species with lower proportions of kmer-sharing (e.g., most bacteria—see Methods Section 4.11.1 for details). The estimated peak storage required by each tool can be seen in Figure 4.5. CBG scales linearly with the number of samples, performing identically in both cases. In the low-kmer sharing regime (which is our focus), SBT would require 4 orders of magnitude more storage than CBG to construct (tens of Pb rather than 3Tb). In the high k -mer sharing regime the methods require similar orders of magnitude of storage, ranging from 3–25TB for 1 million samples. However, in the low k -mer sharing regime `mccortex` and SBT require 4 orders of magnitude more storage than CBG (10s of PB rather than 3TB) as a result of their dependence on N_k .

4.4.2 Empirical scaling benchmark

Because the underlying representation of a CBG can be viewed as a set of key-value pairs (keys=row index/bloom filter position, values=bit vector row), the implementation can use a variety of different back ends. CBG v0.1.2 can support disk-based (via Berkeley-DB (Olson et al. 1999)), in memory, or distributed in memory key-value stores (via `redis` (Sanfilippo 2017)). All the results discussed below use the Berkeley-DB, disk-based implementation.

To demonstrate empirical scaling of storage and speed I compared peak storage requirements (Table 4.1), build time (Figure 4.6a), and query time (Figure 4.6b)

4. ENABLING RAPID DNA SEARCH OF ALL SEQUENCED BACTERIA AND VIRUSES

N	# k -mers	SBT (GB)	CBG (GB)	SBT/CBG
10	55,788,985	0.1	0.03	4
100	499,164,156	12.5	0.3	40
500	2,465,376,242	308.2	1.6	197
1000	4,922,278,348	1230.6	3.1	394
2000	9,803,033,865	4901.5	6.2	784

Table 4.1: Peak storage requirements to construct a SBT and a CBG and total unique k -mer counts for various numbers of bacterial WGS data sets from the SRA/ENA. Because SBTs have a dependency on both total k -mers and number of samples in the collection, whereas CBGs depends only on the number of samples, the ratio of SBT size to CBG size increases when samples with new unique k -mers are added.

of a SBT and a CBG constructed from 2,000 WGS sequence data sets of *Enterobacteriaceae* bacterial samples randomly chosen from the SRA/ENA (see Methods, Section 4.11.2). For the collection of 2,000 samples SBT required 4.9 TB of intermediate storage to construct, and took 3 days to build. In contrast, CBG required 6.2GB space and took 5 hours to build—a 784× smaller and a 14× faster build time (see Table 4.1, Figure 4.6a).

At each increment I queried the SBT and CBG for 705 *Enterobacteriaceae* antimicrobial resistance genes with an average length of 847bps and total query length of 597,753bps. I then calculated the mean time take for the query to be returned. Query times for SBT, CBG exact search (T=100% of k -mers present), and inexact search (T=80% of k -mers present) can be seen in Figure 4.6. CBG had faster exact search times than SBT. SBT had slightly faster search times for inexact searches. Both methods' inexact query time appear to scale similarly with the number of

4.4. Results: Computational performance and validation

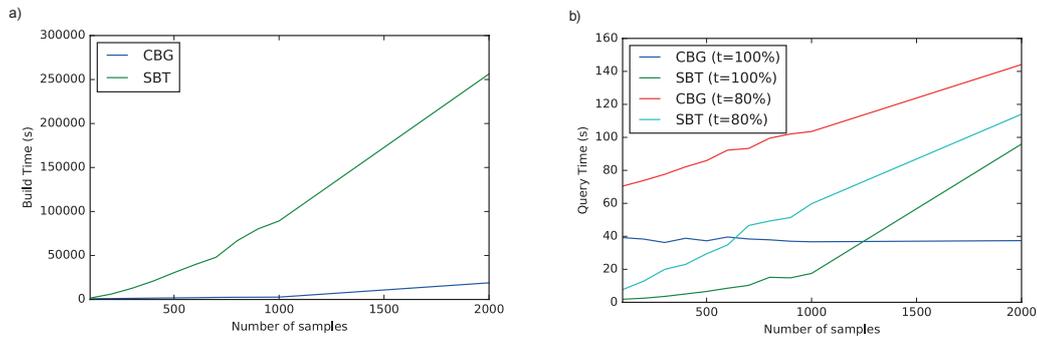


Figure 4.6: Build and query time for CBG and SBT: a) Build time, using a single process, for SBT and CBG data structures with increasing numbers of samples. Building a CBG requires creation and row-indexing of a bit matrix. SBT construction requires traversal of the tree and $\log(N)$ hamming distance operations for each insert. b) Query times for 705 antimicrobial resistance genes when searching in a SBT or CBG for exact (100% of k -mers) or inexact (80% of k -mers) matches. CBG exact searches ($T=100\%$) appear flat as a result of the bit vector optimisation possible for exact search (See Section 4.3.2.2).

colours/samples.

SBT and CBG returned identical hits from the exact match search, finding the antimicrobial resistance genes in the same samples. CBG returned two hits at $T=80\%$, one with 80% of k -mers and the other with 82% of k -mers, that were not returned by SBT. This was likely due to the different false positive rates of the underlying bloom filters as both were found by SBT when the threshold was lowered to 70%.

Constructing SBTs with larger numbers of diverse samples quickly becomes prohibitive in terms of both storage and construction time required. However, the CBG can store > 1.5 million bacterial sequences in 4.9TB, the same storage required to construct the SBT of the 2,000 *Enterobacteriaceae* data sets with similar or improved

4. ENABLING RAPID DNA SEARCH OF ALL SEQUENCED BACTERIA AND VIRUSES

AAT	ATA	TAG	AGC	GCT	CTA	TAA	AAC	ACC	CCC	CCG	CGC	GCT
1	0	0	0	1	1	1	1	1	0	0	0	1

$$\{0 : [3, 3], 1 : [1, 5, 1]\}$$

$$\text{Estimated number of mismatches} = 6/3 = 2$$

$$\text{Estimate number of matches} = 13$$

$$\text{Bit Score} = 13 - 4 = 9$$

Table 4.2: Toy example of CBG scoring

search time.

4.4.3 CBG pseudo alignment scores correlate strongly with megaBLAST scores

One of the advantages of BLAST and similar tools is that they generate alignments, which can then be scored. These scores can be assessed for significance and assigned p -values (Altschul et al. 1990; Zhang et al. 2000). Although CBG queries do not return a full alignment, it is possible to estimate BLAST-like scores for highly similar alignments.

To do this, we take the presence/absence vector for a query of N k -mers (as ones and zeros), remove any ones with zeroes on either sides and count the length of contiguous runs. From this, we estimate the approximate number of mismatches of the query with the sequence. See Table 4.2 for a toy example.

From these estimations of mismatches and matches we can estimate a score using -2 for mismatched and +1 for matched positions for an ungapped alignment. From

this, E-scores and p -values can be calculated using the same scheme as BLAST :

$$E = KNL e^{-\lambda s}, \quad (4.4)$$

and

$$P = 1 - e^{-E}, \quad (4.5)$$

from (Karlin et al. 1990), where $K = 0.621$, $\lambda = 1.330$, N is the number of colours in the graph, L is the length of the query, and s the score. K and λ are determined by the scoring scheme used, and here we use constants based on the NCBI BLAST's default values for ungapped alignment.

In order to compare CBG scores with megaBLAST scores I built a CBG of 67,146 bacterial assemblies from NCBI refseq (Oct 2016) (since it is not possible to run BLAST or megaBLAST on the microbial-CBG data set). The total uncompressed size of the CBG was 288GB. I queried this CBG and the megaBLAST database constructed from the same references for 100 random AMR genes from the CARD v1.1.7 database (McArthur et al. 2013). CBG exact/inexact queries took on average 1.25s and 4.92s, respectively. megaBLAST queries took 80.9s on average. Thus, CBG queries were 16× faster than megaBLAST for inexact search, and 64× faster for exact match search.

MegaBLAST returned many identical alignments so I de-duplicated the results and compared the resulting non-redundant set of scores. For each of the 100 de-duplicated gene query hits I compared the CBG score and the megaBLAST score on the same gene/reference combination. CBG and megaBLAST scores were highly correlated ($r=0.998$, Pearson) (see Figure 4.7).

4. ENABLING RAPID DNA SEARCH OF ALL SEQUENCED BACTERIA AND VIRUSES

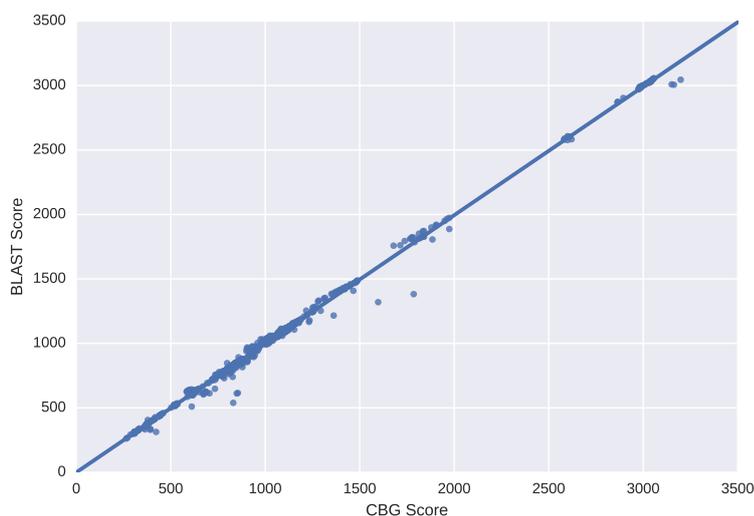


Figure 4.7: CBG vs `megaBLAST` scores: all unique `megaBLAST` scores in `megaBLAST` results for searches of 100 antimicrobial resistance genes vs. the equivalent CBG scores. Pearson correlation of the scores was $r=0.998$.

Currently, CBG scoring does not consider gapped alignments, but it may be possible to extend this using local assembly to find the path in the graph that best matches the query, and then running a true alignment on the query and assembled contig. This is discussed more in Section 4.10. It is also important to note that because each mismatch affects k k -mers, the ability to estimate the number of mismatches rapidly deteriorates as the query gets more dissimilar. Thus, CBG scores are only applicable to highly similar matches (the regime where `megaBLAST` is also used) and are not appropriate when the number of mismatches between the query and the hit is high.

4.4.4 Assessing gene genotyping accuracy

Multi-locus sequence typing (MLST) is a widely used method for categorising bacteria. It relies on accurate typing of gene alleles in order to categorise samples into sequence types. Many MLST typers use **BLAST** as the underlying algorithm, but they require assembly in advance (Jolley et al. 2012; Larsen et al. 2012; Page et al. 2016). Short Read Sequence Typing 2 (**SRST2**) takes raw reads and on a per sample basis it uses a mapping-based approach to align reads to the alleles in the MLST ‘scheme’ (Inouye et al. 2014).

In order to measure the accuracy of allele detection with **CBG**, I first searched the **microbial-CBG** (described in the following section) (with T=70% match) for a catalogue of *Escherichia coli* Multi Locus Sequence Type (MLST) alleles (Wirth et al. 2006) and chose the best scored allele for each gene. I then compared calls on a set of 954 samples with the MLST allele calls from a high-quality caller: **SRST2**. Where both methods made a call (6483/6678 alleles), there was 99.9% (6475/6483) agreement; otherwise **SRST2** failed (n=167), or **CBG** failed to find an allele version above T=70% (n=28).

The high concordance between **SRST2** and **CBG** MLST demonstrates the utility of **CBGs** in determining the distribution of sequence types in collections of WGS data sets without having to determine sequence types of each sample separately via mapping or assembly. It also demonstrates the accuracy of **CBG** allele typing more generally which is essential for antimicrobial resistance prediction, virulence determination, and many other analyses.

4.4.5 Assessing variant genotyping accuracy

We are often interested in searching samples for genetic variants as well as for sequence. In *M. tuberculosis* for example, most antimicrobial resistance determinants are amino acid substitutions in key proteins (Walker et al. 2015). We can express the variant search problem as a sequence search problem by searching for variant reference and alternate alleles (a *variant probe set*) on many genetic backgrounds (see section 2.4.1.1), and use the CBG to search and genotype these variant “probe sets” in a large collection of sequences, as discussed in section 4.3.2.3 above.

To assess genotyping accuracy, I took probes for 68,269 SNPs, built a CBG, and genotyped them using the algorithm described in section 4.3.2.3. I compared these genotypes with those from a `stampy` (Lunter et al. 2011) + `samtools` (Li et al. 2009) pipeline from (Walker et al. 2015), excluding filtered positions. See methods (section 4.11.3) for details. The mean concordance between methods was 99.997%, with a total of only 286/682,690 discrepancies across all samples. The majority of these discrepancies (203/286) were heterozygous calls from CBG, whereas `samtools` was run with a haploid model and did not make any heterozygous calls.

4.5 Results: A searchable index of the entire microbial ENA/SRA

To empirically illustrate the utility of CBG, we constructed a CBG from the entire bacterial and viral content of the ENA as of December 2016 (N=455,632), which

I will refer to as the `microbial-CBG` data set. Because the SRA and ENA mirror each other, this also contained all microbial SRA data.

Some samples were removed during pre-processing (see Extended Methods), and the resulting `microbial-CBG` of 447,833 error-cleaned samples required 1.5TB of storage, <1% of the original data size (170TB), while indexing more than 60 billion unique k -mers (see section 4.11.4 for details on construction and section 4.11.4.1 for details on approximate k -mer counting). Using the scaling equations described in section 4.4.1, I estimated that the intermediate storage required to build a SBT of the same data would have been >6.7PB of storage, more than 4400 \times that required by CBG. An exact de Bruijn graph or inverted index such as `mccortex` would have required ≈ 3.5 PB of storage, which is $\approx 2,300\times$ more.

4.5.1 Taxonomic analysis of microbial ENA data sets

The proportion of different species in each data set was determined by `bracken` (Lu et al. 2017a; Wood et al. 2014) (see Methods section 4.11.4.2). The distribution of genera within the SRA/ENA is unbalanced and can be seen in Figure 4.8. Over 90% of the isolates were from only 4% of the unique genera (20/496). The counts of the 20 most prevalent bacterial genera can be seen in Figure 4.8. Results in the following sections on the distribution of genetic elements must be viewed with this taxonomic bias in mind.

4. ENABLING RAPID DNA SEARCH OF ALL SEQUENCED BACTERIA AND VIRUSES

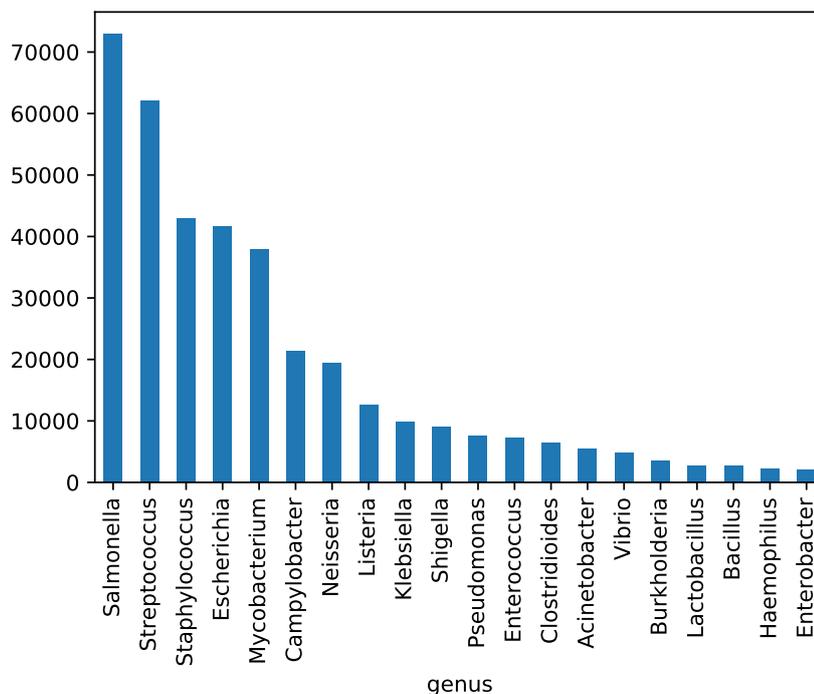


Figure 4.8: Counts of the most frequent bacterial genera in the microbial-CBG: Over 90% of the data sets were isolates of these 20 genera, and 65% were from the top 5 most prevalent genera, according to assignment by Bracken analysis.

4.6 Application 1: Fast gene search in the microbial ENA

Retrospective searching of collections of WGS data sets is often desired, but currently difficult to perform. For example, recent papers describing the emergence of plasmid-mediated colistin resistance genes MCR-1, MCR-2, and MCR-3 (Liu et al. 2016; Lu et al. 2017b; Xavier et al. 2016; Yin et al. 2017) sparked interest in retrospective analysis of databases to investigate dissemination of the genes (Hu

4.7. Application 2: Surveillance of antimicrobial resistance genes and variants

et al. 2016; Matamoros et al. 2017; Park et al. 2017; Suzuki et al. 2016). However, due to current limitations, these studies analysed less than 10% of the samples than in the SRA (between 410-43,099 genomes), raising concerns that the worldwide distribution of MCR may be underestimated (Hu et al. 2016).

Using the `microbial-CBG`, I ran an exact-match search of the AMR gene MCR-1 across the SRA for all three of the genes in 447,833 data sets. This took 1.73s seconds in total, scanning 10× more genomes than previous publications. CBG returned 166 hits for MCR-1 across 4 species: 139 *Escherichia coli*, 24 *Salmonella enterica*, 2 *Kluyvera ascorbata*, and 1 *Enterobacter aerogenes*, as well as finding MCR-3 in 38 *Escherichia coli*, 25 *Salmonella enterica*, and 3 *Klebsiella pneumoniae*. However, no results for MCR-2 were found, even when the threshold was reduced to 40%. The samples found were used in Wang et al. 2017 to help determine the global distribution of MCR-1.

4.7 Application 2: Surveillance of antimicrobial resistance genes and variants

4.7.1 Rapid variant search and genotyping variants enables monitoring of allele frequency in a population

Antimicrobial resistance in *M. tuberculosis* is driven primarily by amino acid mutations in key genes (Bradley et al. 2015; Walker et al. 2015). When combined

4. ENABLING RAPID DNA SEARCH OF ALL SEQUENCED BACTERIA AND VIRUSES

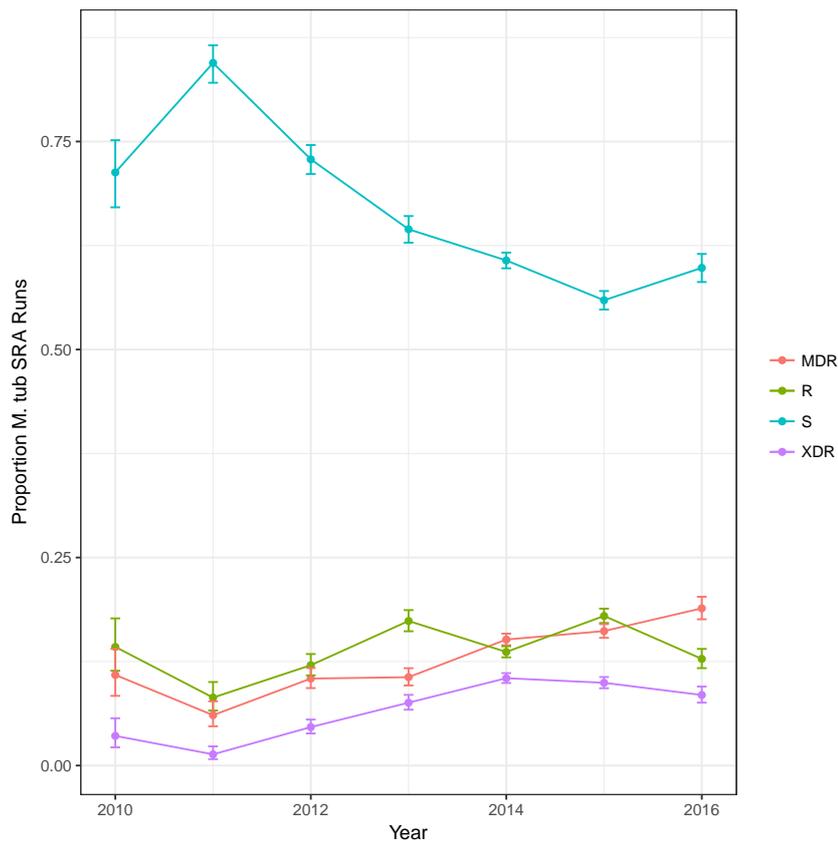


Figure 4.9: Proportion of *M. tuberculosis* classified by genotypes as resistant (R), pan-susceptible (S), multi-drug-resistant (MDR), and extensively drug-resistant (XDR) in `microbial-CBG` by date of first public availability: the `microbial-CBG` was queried for variants from the resistance catalogue from Walker et al. 2015. Resistant samples (R) are samples with at least one resistant allele but not MDR or XDR; susceptible samples (s) have no evidence for resistant alleles; MDR have rifampicin and isoniazid resistance alleles; and XDR have MDR mutations as well as resistance to one of the following: fluoroquinolones, amikacin, capreomycin, or kanamycin.

4.7. Application 2: Surveillance of antimicrobial resistance genes and variants

with date-time metadata, we can analyse trends in the frequency of these alleles in large collections of sequence data. For instance, by searching a collection of samples for a catalogue of antimicrobial resistance variants, we can rapidly screen for samples with resistance and susceptible amino acids as well as for the frequency of these alleles.

I searched the `microbial-CBG` for the variants from the catalogue described in Walker et al. 2015. The variant probes were generated using the procedure outlined in section 2.4.1.1 and genotyped using the procedure described in section 4.4.5. The genotyping took 103 minutes on a single core, which is $\approx 10000\times$ faster than running `Mykrobe predictor` on each sample individually.

I classified each of the data sets as resistant or susceptible to 12 antibiotics based on their genotypes. Based on their inferred antibiogram I further classified them as multi-drug-resistant (MDR) if resistant to rifampicin and isoniazid; extensively-drugresistant (XDR) if MDR and resistant to one of any fluoroquinolones, amikacin, capreomycin, or kanamycin; resistant (R) if resistant to any antibiotic, but not MDR or XDR, and susceptible otherwise. The results, split by year of upload to the SRA/ENA, can be seen in Figure 4.9. A trend of increasing resistance, MDR, and XDR can be observed in the `microbial-CBG` data set.

In Figure 4.10, I show equivalent trends in rifampicin and isoniazid resistance (left) and the most prevalent mutation inducing this resistance (right). From these figures we can deduce that rifampicin and isoniazid resistance is driven primarily by a single amino acid mutation, with favoured alleles, presumably via convergent selection as these mutations have arisen independently in multiple *M. tuberculosis*

lineages. The dominance of *katG*-315 and *rpoB*-450 in driving isoniazid and rifampin resistance has been observed previously (Walker et al. 2015).

Unfortunately, there are significant caveats to this analysis. First, the SRA/ENA metadata is very limited, so it was not possible to extract the date on which these data sets were sampled or sequenced. Thus, I used the date upon which they were first made available in the SRA/ENA as a proxy. This was the only ‘datetime’ metadata field consistently available across all samples. Second, the SRA/ENA is not an unbiased collection of samples, and trends in allele frequencies may simply reflect increased research interest in the related phenotype, rather than increasing frequency in the population. As a result, these analyses are intended to illustrate a potential use case of CBG: amalgamating sequence search with metadata to investigate trends. Interpretation from an epidemiological perspective is limited without improved metadata and/or rigorous sampling criteria. For instance, the latest WHO estimates for 2016 (World Health Organization 2017) put MDR prevalence at 6.6% (compare: 18.9% in 2016 in Figure 4.9).

4.7.2 A survey of antimicrobial resistance genes in the SRA/ENA

2,157 DNA sequences associated with antimicrobial resistance were downloaded from the comprehensive antibiotic resistance database [CARD] v1.1.7 database (McArthur et al. 2013), with a mean length of 937 base-pairs. I searched for these sequences in the microbial-CBG with thresholds of 100% and 80%. An exact match

4.7. Application 2: Surveillance of antimicrobial resistance genes and variants

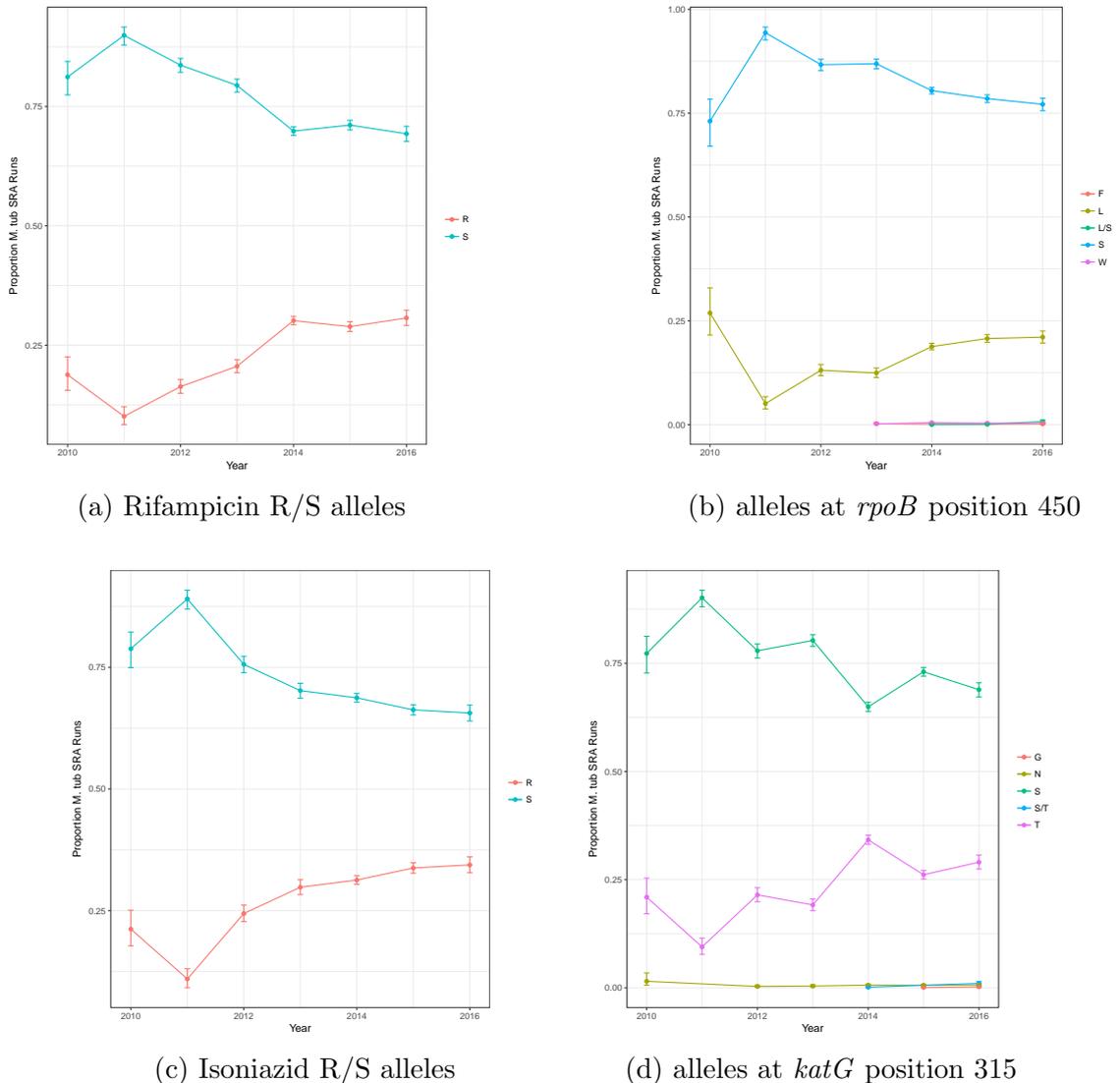


Figure 4.10: Proportion of *M. tuberculosis* resistance alleles in the microbial-CBG by date of first public availability: The microbial-CBG was queried for variants from the resistance catalogue from Walker et al. 2015. Proportions of resistance/susceptible alleles and the amino acid at the most frequently mutated site are shown. An increase in resistant alleles can be observed for rifampicin and isoniazid (left), primarily driven by a dominant mutation (right). The SRA/ENA is not an unbiased source and this may reflect increased interest and oversampling of resistant strains, rather than an increasing prevalence. X/Y indicates samples with evidence of more than one amino acid at that position, possibly from mixed samples.

4. ENABLING RAPID DNA SEARCH OF ALL SEQUENCED BACTERIA AND VIRUSES

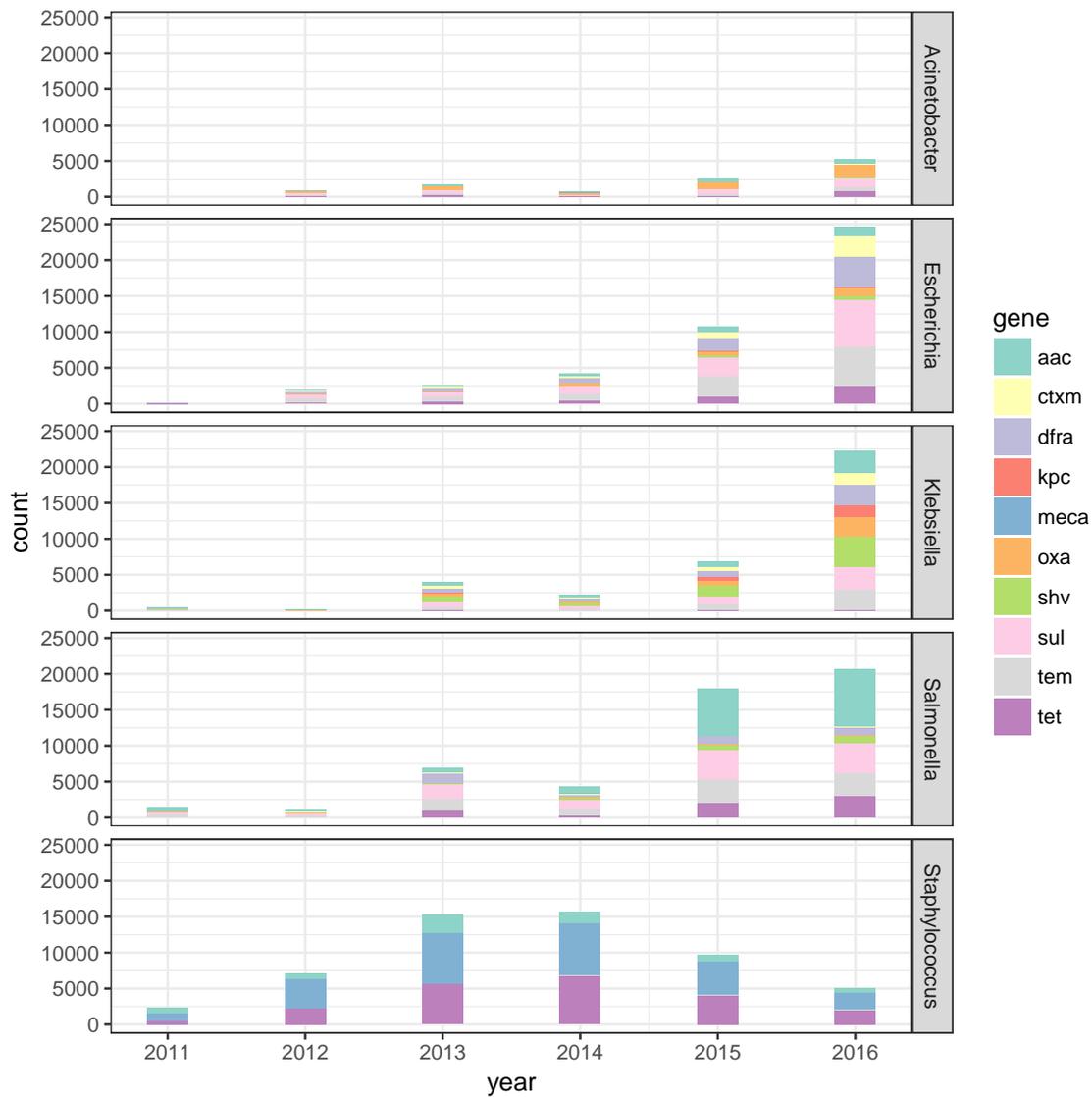


Figure 4.11: Counts of accessions with 10 AMR genes from CARD across 5 genera split by year of upload to the ENA.

4.7. Application 2: Surveillance of antimicrobial resistance genes and variants

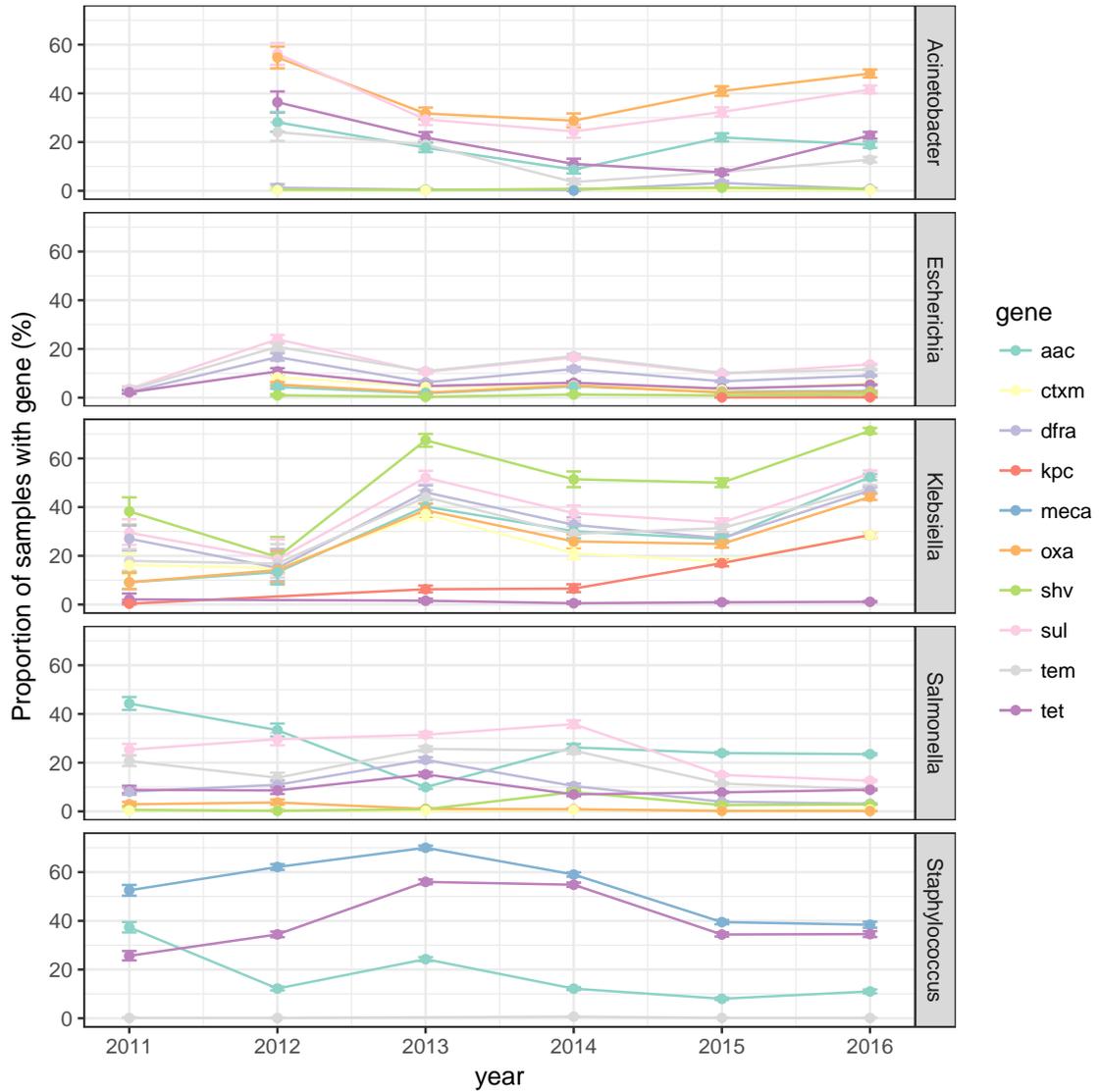


Figure 4.12: Proportion 10 AMR genes from CARD in the microbial-CBG across 5 genera split by year of upload to the ENA. Error bars show the 95% confidence interval calculated by the Wilson binomial confidence test (Wilson 1927).

search for a gene, requiring 100% of k -mers to be present, took 1.13 seconds and returned 438 hits on average. In total, this resulted in 944,862 hits in 193,582 unique accessions across 250 genera. An inexact search, requiring 80% of k -mers to be found, took 34.4 seconds and returned 5,320 hits on average.

Figures 4.11 and 4.12 show the absolute count and the frequency of 10 AMR genes across 5 genera split by the year the data set was made public in the ENA. Notable trends include the decreasing prevalence of *mecA* in *Staphylococcus* from 70.0% (7129/10183) in 2013, to 38.4% in 2016 (2316/6032), and the increasing prevalence of *kpc* in *Klebsiella* from 0.4% (1/285) in 2011, to 28.5% (1702/5968) in 2016. However, as already noted above, because the ENA is not an unbiased collection of data, it is possible these trends are related to increased or decreased interest in the gene, phenotype, or genera, rather than trends that are reflected in the population.

4.8 Application 3: Measuring the host range of conjugative elements

4.8.1 Measuring plasmid host range

Plasmids and other mobile genetic elements (MGEs), transferable between different bacterial species, are important components of microbial genomes. They have the ability to confer adaptive phenotypes such as antibiotic resistance or virulence and be drivers of bacterial outbreaks (Bennett 2008; Mathers et al. 2015). MGEs

are challenging for bioinformatic analysis. Their variable size, variable copy number, and repetitive sequence sometimes shared with the core genome mean that assembly algorithms can struggle to accurately assemble them (Antipov et al. 2016). This implies that they may be underrepresented in assembled sequence databases, hindering the ability to retrospectively search for them. Since **CBG** does not require assembly, and can search for arbitrary length query, it is an appropriate tool for running large scale retrospective surveys of plasmids and other mobile genetic elements. Searching for whole plasmids (queries of > 100kb) is possible, or, since plasmids can rearrange, for conserved backbones, or the genes that allow for conjugation (see section 4.8.1.1).

I took 2827 plasmids from the ENA and ran an inexact (T=90%) search for these in the **microbial-CBG**. The search took 2,120 CPU hours, 11 days real time on a single machine using 8 cores, using <1.5GB memory per process. The search returned 665,619 hits with 121,758 unique accessions across 258 genera. These results were filtered to exclude samples with contamination leaving 415,181 hits with little or no detectable cross-genus contamination (see Methods Section 4.11.5).

Figure 4.13 shows the frequency of 37 plasmids in 23 genera, which were found at least 5 times in more than one genus in the **microbial-CBG**. The majority of these plasmids, which are shared across multiple genera, are found in *Enterobacteriaceae*. However, this may either reflect the taxonomic bias within the **microbial-CBG** collection: 36.5% of all samples in the **microbial-CBG** were classified as *Enterobacteriaceae*, or bias in the plasmids reported on the EBI plasmid page.

Five plasmids were found across multiple families and orders (shown in Table 4.3). Plasmids GQ900420, U32369, and U40259 were only observed in *Bacilli*,

U09422 in *Bacilli*, *Clostridia*, and *Erysipelotrichia*. AF012911, otherwise known as pETHIS-1, is categorised by EBI as a known broad host range plasmids and was observed in 5 phyla and 10 taxonomic classes. It contains a *bla* gene encoding ampicillin resistance. However, it is used as a cloning vector, so the observations across multiple phyla may be a result of its use as an experimental construct (Schaller et al. 1999; Subramaniam et al. 2000). Therefore, its observed host range may not be representative of its “natural” range. U09422, otherwise known as Tn916, is in fact a known conjugative transposon (or integrative conjugative element), rather than a plasmid. First found in *Enterococcus faecium*, it is known to have a broad host range. Tn916 encodes for tetracycline resistance and has been previously observed in a diverse range of bacteria (Ciric et al. 2013).

4.8.1.1 Prevalence of conjugative systems

Integrative conjugative elements (ICEs), otherwise known as conjugative transposons, are a diverse group of mobile genetic elements found in both Gram-positive and Gram-negative bacteria (Johnson et al. 2015). Unlike conjugative plasmids, ICEs are integrated into the host’s core genome. They encode a type IV secretion system (T4SS) and a relaxase (MOB) in order to facilitate excision from the chromosome and conjugation. As a result, they are self-transmissible. Like plasmids and other mobile genetic elements, ICEs enable cross-clade transfer of phenotypically important genes, such as antimicrobial resistance genes.

In order to quantify the distribution and diversity of conjugative elements (both plasmids and ICEs), Guglielmini et al. 2011 searched for conjugative elements across

4.8. Application 3: Measuring the host range of conjugative elements

Table 4.3: The count of observations within various genera of 5 plasmids found at > 10% frequency in more than one taxonomic family. Observations are from a search in the microbial-CBG for these plasmids at T=90%.

	AF012911	GQ900420	U09422	U32369	U40259
Streptococcaceae	0	1497	4197	3	0
Staphylococcaceae	0	168	1743	76	5
Enterococcaceae	0	0	87	34	4
Peptostreptococcaceae	0	0	110	1	0
Bacillaceae	48	0	2	0	0
Lachnospiraceae	30	0	0	0	0
Enterobacteriaceae	20	0	0	0	0
Listeriaceae	0	0	19	0	0
Erysipelotrichaceae	0	0	12	0	0
Burkholderiaceae	9	0	0	0	0
Clostridiaceae	9	0	0	0	0
Bifidobacteriaceae	7	0	0	0	0
Kosmotogaceae	6	0	0	0	0
Ruminococcaceae	6	0	0	0	0
Phyllobacteriaceae	5	0	0	0	0
Flavobacteriaceae	4	0	0	0	0
Rhizobiaceae	3	0	0	0	0
Methylophilaceae	2	0	0	0	0
Methylococcaceae	2	0	0	0	0
Lactobacillaceae	0	0	2	0	0
Pseudomonadaceae	2	0	0	0	0
Geobacteraceae	2	0	0	0	0
Ectothiorhodospiraceae	2	0	0	0	0
Bradyrhizobiaceae	2	0	0	0	0

4. ENABLING RAPID DNA SEARCH OF ALL SEQUENCED BACTERIA AND VIRUSES

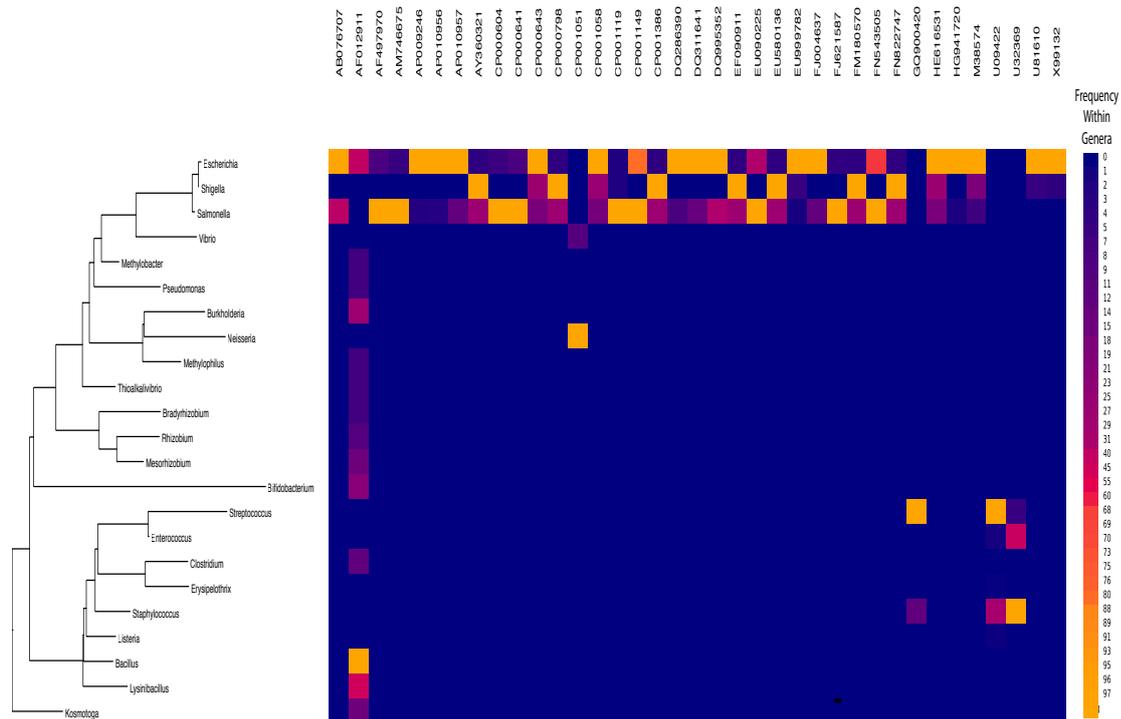


Figure 4.13: 37 plasmid sequences found at least 5 times in more than one genus in the microbial-CBG. The heatmap shows the frequency of each plasmid within each genus. The plasmids were hierarchically clustered using the UPGMA algorithm using euclidean distance metric (Day et al. 1984). The genera are from a RNA tree of bacteria from www.arb-silva.de. The plasmid on the left (AF012911) with an extremely wide phylogenetic distribution is a known cloning vector. The large amount of sharing between *Escherichia*, *Salmonella*, and *Shigella* is consistent with known promiscuity within *Enterobacteriaceae*.

4.8. Application 3: Measuring the host range of conjugative elements

1,124 prokaryotic genomes. They found a high frequency of conjugative elements in their data: 18% of genomes contained a putative ICE, and 12% contained a putative conjugative element. ICEs were found to be widely distributed across the bacterial phylogeny with examples in the major branches of proteobacteria, the bacteroidetes, and the firmicutes. They quantified MOB family distribution amongst clades, showing preference for certain MOB types in some clades. Their observations of a close interplay between ICEs and plasmids suggest that plasmids and ICEs might be a very similar type of element, and that plasmids often become ICEs and vice-versa. Limited by sample size, Guglielmini et al. 2011 described ICE distribution phylum level resolution. I ran a similar survey of conjugative systems within the microbial-CBG and compare insights on a larger collection of data sets, aiming for genus level resolution of the distribution of ICEs.

I searched the microbial-CBG for MOB type relaxases and T4SS sequences from Guglielmini et al. 2011, requiring an exact match (T=100%). I applied the same contamination filter described in section 4.8.1. 19.5% (36,030/184,652) of all data sets passing this filter had exact matches to at least one relaxase and T4SS. This indicated at least one putative conjugative system in these data sets, slightly more than the 18% reported by Guglielmini et al. 2011. However, this varied from $\approx 0.5\%$ in *Spirochaetes* to $\approx 31.7\%$ in *Firmicutes* (see Table 4.4).

Consistent with observations by Guglielmini et al. 2011, the majority of data sets (79.8% - 28,758/36,030) with a putative conjugative system had more than one T4SS type indicating co-occurrence of conjugative systems. Table 4.4 shows that conjugative elements were more frequently found in Firmicutes, and Proteobacteria

4. ENABLING RAPID DNA SEARCH OF ALL SEQUENCED BACTERIA AND VIRUSES

Table 4.4: This table shows the number of observations of putative conjugative elements across phyla in the **microbial-CBG**. Dividing by the counts of members of this phyla gives the approximation of the proportion of samples of each phyla containing conjugative systems. Confidence intervals are calculated via the Wilson interval (Wilson 1927).

Phylum	Count	Total	Proportion
Firmicutes	25075	79022	31.7% (31.4%-32.1%)
Proteobacteria	10749	66386	16.2% (15.9%-16.5%)
Bacteroidetes	45	1375	3.3% (2.5%-4.4%)
Cyanobacteria	18	505	3.6% (2.3%-5.6%)
Fusobacteria	6	427	1.4% (0.6%-3.0%)
Acidobacteria	5	459	1.1% (0.5%-2.5%)
Spirochaetes	3	600	0.5% (0.2%-1.5%)

(even after accounting for bias within the SRA/ENA), but they were also found in Cyanobacteria, Acidobacteria, Fusobacteria, and others, at a lower frequency as previously reported in Guglielmini et al. 2011.

Figure 4.14 shows the distribution of MOB types found in each phylum. The distribution observed is broadly consistent with that observed by Guglielmini et al. MOB_B was found to be specific to *bacteroides* and MOB_V found mostly in *cyanobacteria*. MOB_T was the most prevalent relaxase in *firmicutes* and not found in other phyla while MOB_H was found in *proteobacteria*, but not in other phyla.

4.8.1.2 MOB types may define conjugative elements' host range

The data above can give valuable information about the possible spread of conjugative elements. For example, MOB_T is found in *streptococcus* and *staphylococcus*,

4.8. Application 3: Measuring the host range of conjugative elements

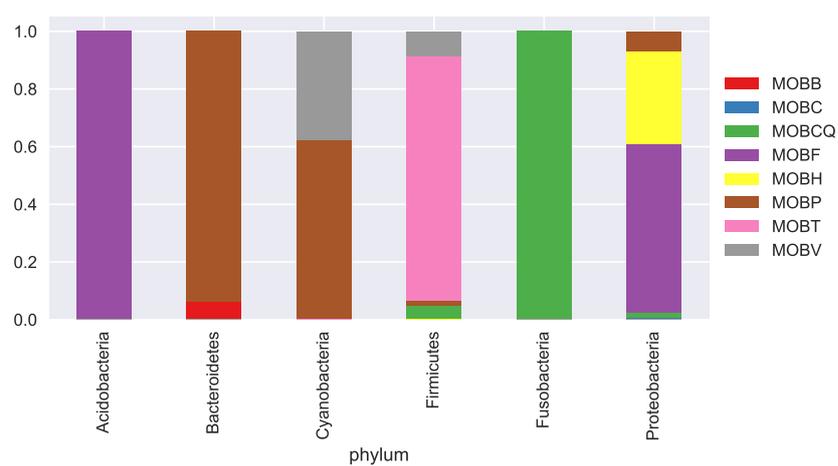


Figure 4.14: Distribution of MOB types among phyla based on a search of all known MOB types from Guglielmini et al. 2011 of the SRA/ENA.

4. ENABLING RAPID DNA SEARCH OF ALL SEQUENCED BACTERIA AND VIRUSES

Table 4.5: The number of observations of MOB genes across genera in the microbial-CBG. Genera with less than 6 observations were excluded. Note that the MOB types have well-defined host ranges. For example, MOB_T is found in *streptococcus* and *staphylococcus*, but not *salmonella*.

Gene	MOB_B	MOB_C	MOB_{CQ}	MOB_F	MOB_H	MOB_P	MOB_T	MOB_V
Acinetobacter	0	0	0	187	0	0	0	0
Alicyclophilus	0	0	0	0	0	8	0	0
Aliivibrio	0	0	0	0	0	14	0	0
Arcanobacterium	0	0	0	0	0	12	0	0
Bacillus	0	0	0	0	0	560	74	7
Bacteroides	0	0	0	0	0	14	0	0
Burkholderia	0	0	0	0	1	10	0	0
Campylobacter	0	0	0	0	0	59	0	0
Clostridioides	0	0	0	0	0	3	3165	0
Clostridium	0	0	0	0	0	16	0	0
Echinicola	0	0	0	0	0	16	0	0
Enterococcus	0	18	0	0	0	0	436	0
Erysipelothrix	0	0	0	0	0	0	72	0
Escherichia	0	106	63	72	36	32	0	0
Flavobacterium	0	0	0	0	0	12	0	0
Granulicella	0	0	0	8	0	0	0	0
Haemophilus	0	0	0	0	12	0	0	0
Helicobacter	0	0	0	0	0	6	0	0
Ilyobacter	0	0	6	0	0	0	0	0
Klebsiella	0	6	0	0	18	0	0	0
Lactobacillus	0	0	3	0	0	1	14	0
Legionella	0	0	0	320	8	18	0	0
Listeria	0	0	0	0	0	132	349	0
Mannheimia	0	0	0	0	56	0	0	0
Mycobacterium	0	0	0	0	0	116	0	0
Neisseria	0	0	0	0	5	1	0	0
Oscillatoria	0	0	0	0	0	6	0	0
Pasteurella	0	0	0	0	26	0	0	0
Pseudomonas	0	0	0	0	462	5	0	0
Pseudopedobacter	0	0	0	0	0	8	0	0
Ruminococcus	0	0	0	0	0	8	0	0
Salmonella	0	0	220	10975	3147	1020	0	0
Serratia	0	0	0	0	12	0	0	0
Shewanella	0	0	0	36	6	0	0	0
Staphylococcus	0	0	0	0	0	8	19273	3832
Stenotrophomonas	0	0	0	0	0	20	0	0
Streptococcus	0	0	1783	0	0	152	17628	20
Terriglobus	0	0	0	9	0	0	0	0
Thermoanaerobacter	0	198	0	0	0	2	0	0
Vibrio	0	0	0	0	1783	0	0	0
Yersinia	0	0	0	0	8	0	0	0

but not *salmonella*, indicating that an antimicrobial-resistant-carrying conjugative element encoding MOB_T has a higher risk of spreading to *streptococcus* and *staphylococcus* than to *salmonella*. In contrast, MOB_Q is observed in *salmonella* and *streptococcus*, but not in *staphylococcus*, indicating a different probability of cross-genus transfer. A table of MOB types across genera can be found in Table 4.5 which shows well-defined host range of many MOB types across genera. MOB_C , MOB_Q , MOB_{CQ} , and MOB_V are all observed in 5 or fewer genera.

4.9 Application 4: Surveillance of multi-drug-resistant plasmids

As discussed in section 4.8, mobile genetic elements are important drivers in the spread of antimicrobial resistance, virulence, and other phenotypes. Plasmid acquisition and increasing prevalence of specific mobile genetic elements can pose a substantial threat to human health, imparting antimicrobial resistance and virulence phenotypes to pathogen outbreaks (Gu et al. 2017; Liu et al. 2016; Mathers et al. 2015).

For instance, the need for surveillance to detect MDR plague has been raised previously (Galimand et al. 2006), as has the need to monitor these plasmids in *Vibrio cholerae* (Pan et al. 2008). This is particularly pressing given the ongoing (drug susceptible) plague outbreak in Madagascar which began in October 2017 (Roberts 2017). Here I show how CBG combined with metadata from The National

4. ENABLING RAPID DNA SEARCH OF ALL SEQUENCED BACTERIA AND VIRUSES

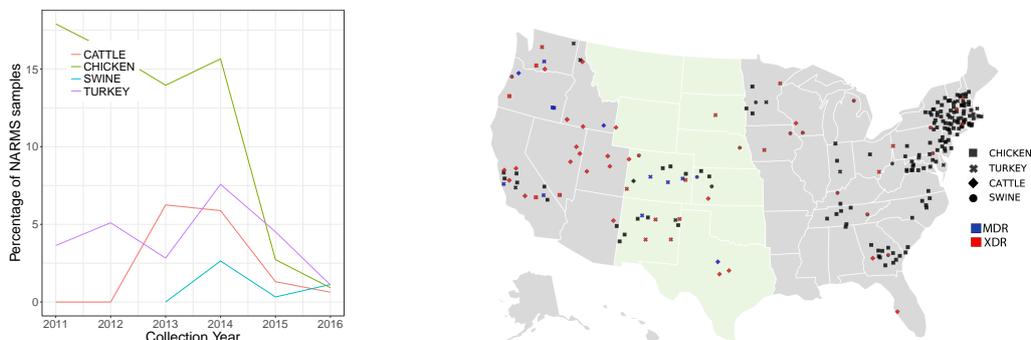
Antimicrobial Resistance Monitoring System (NARMS) can be used to retrospectively monitor the spread of a multi-drug-resistant plasmid from *Yersinia pestis*.

Welch et al. 2007 investigated a multi-drug-resistant plasmid found in numerous MDR enterobacterial pathogens and isolated from retail meat samples collected between 2002 and 2005 in the United States (Galimand et al. 1997; Welch et al. 2007). The plasmid shared a 113,320bp backbone with a *Yersinia pestis* plasmid—pIP1202—indicating acquisition of this plasmid from a shared ancestor and encoded resistance to 8 antimicrobials: sulfonamides (*sul1*, *sul2*); phenicols (*cat*, *floR*); tetracyclines (*tetRA*); aminoglycosides/aminocyclitols (*aacC*, *aadA*, *aphA*, *strAB*); quaternary ammonium compounds (*qacEdelta1*, *sugE1*, *sugE2*); β -lactams (*blaCMY-2-1*, *blaCMY-2-2*, *blaSHV-1*); trimethoprim (*dhfrI*); and mercury ions (*merRT-PABDE*, *merRTPCADE*). This plasmid was found in 70/125 *Salmonella* surveyed from retail meats, including chicken, turkey, pork, and ground beef, and in all ten states that participated in the NARMS at the time. I set out to investigate the occurrence of this plasmid in the microbial-CBG.

I searched for the whole 182,913 bp plasmid (T=50%), and found 2,341 hits in the microbial-CBG, dominated by *Salmonella enterica*, but also found hundreds of samples of *Escherichia coli*, *V. cholerae*, and *Klebsiella pneumoniae*. Of these hits, 10% (244/2,341) were also in NARMS, of which 243 were *Salmonella*.

Using metadata from NARMS, I was able to label by location, source (cattle, chicken, swine and turkey), and sampled date. In Figure 4.15a, I show the samples split by source. We first see the plasmid in cattle in 2012 and in swine in 2014, and there is an appreciable drop (18% (29/162) to 3% (27/990)) in the proportion

4.9. Application 4: Surveillance of multi-drug-resistant plasmids



(a) Histogram of the source of samples with *Yersinia pestis* plasmid pIP1202 by year. (b) Locations of samples containing *Yersinia pestis* plasmid pIP1202.

Figure 4.15: (a) The proportion of NARMS samples with a search hit for pIP1202 at $T=50\%$ in microbial-CBG split by source and year. We first see the plasmid in cattle in 2012 and in swine in 2014, and there is an appreciable drop ($\approx 17\%$ to $\approx 5\%$) in the proportion of chicken containing the plasmid in 2016. (b) Locations of 244 samples with search hits for plasmid pIP1202 in microbial-CBG. Location metadata was taken from NARMS; precise locations not shown, and instead each sample is placed with jitter within the correct state. Sample source, one of CHICKEN, TURKEY, CATTLE, and SWINE, is shown by marker shape. MDR or XDR samples are shown in blue and red, respectively. The samples were classified as MDR if the sample had resistance genes of 2 of the following: tetracycline, sulfonamides, and streptomycin; and XDR if MDR and resistance genes of at least 4 drugs (or 3 drugs and a multi-resistance gene) were present. States overlapping prairie dog range are highlighted.

of chicken containing the plasmid from 2011 to 2015. We can also observe the distribution of cases in the U.S., noting a high density of cases in the North East (see Figure 4.15b).

Combined with a real-time, or near real-time, database of sequence and metadata, CBG could be used to assist in surveillance and alerts for key information, such as horizontal gene transfer of antimicrobial resistance elements or significant trends. The states highlighted in green have populations of prairie dogs (Wimsatt

et al. 2009), which are commonly infected with *Yersinia pestis*. It is known that *Yersinia pestis* plasmids can transfer back and forth between *Escherichia coli* and *Yersinia pestis* in the gut of a flea (Hinnebusch et al. 2002) (though the conjugation experiment has not been done for this plasmid).

If free-ranging cattle bear this plasmid, with drug resistance genes, and in prairie-dog territory, then they are a potential risk for creating an MDR plague outbreak. In Figure 4.15, all coloured crosses represent samples from cattle containing genes conferring resistance to at least 2 of chloramphenicol, tetracycline, and streptomycin—the standard treatments for plague. The search does not guarantee the genes are on the plasmid, but it allows rapid warning that specific samples need further analysis. I used this example to show the power of global sequence search (e.g., via continuously updated CBG) combined with appropriate sampling and metadata storage. GenomeTrakr (*GenomeTrakr* 2017) already provides this metadata in a queryable format, but without sequence search, it can only enable very limited queries.

4.10 Discussion

Searching for sequences in a collection of assemblies, contigs, or raw read data sets, is a fundamental requirement to make optimal use of the huge quantities of sequence data that will be generated in the near future. Until recently, sequence search has been limited to local alignment based search in curated databases of assembled data using BLAST-like tools. However, as our public archives and private collections of sequence data grow, there is an increasing need to enable rapid search

on both raw read and assembled data. In order to address this issue, I developed a data structure, the Coloured Bloom Graph, and associated algorithms that enable search on millions of raw read or assembled data sets. I used a CBG to index and search hundreds of thousands of publicly available short read data sets, achieving a query time for a 1kb query of ≈ 1 s on a single machine. I also describe simulations scaling to millions of data sets.

The key difference between CBG and existing methods is that the storage requirements for a CBG scale linearly with the number of data sets or ‘colours’, regardless of the total number of k -mers in the union of the data sets. A major advantage of this approach is that determining maximum storage requirements in advance is trivial. Bloom filter size for the CBG, i.e., the number of rows in the bit matrix, is determined by the number of k -mers in each sample and the desired false-positive rate. This conveniently takes advantage of what we observe in bacteria: isolates have bounded size, between 1-15 million k -mers, but the union of these isolates—the pan-genome—contains an unbounded number of k -mers. In contrast, other k -mer indexing tools, such as the SBT (Solomon et al. 2016), BFT (Holley et al. 2016), cortex (Iqbal et al. 2012) and, vari (Belk et al. 2016), scale with both the total combined unique k -mers in all samples, and the number of samples, which can be prohibitively large when the pan-genome contains many more k -mers than are in each data set.

Although CBG have many beneficial properties, many technical improvements can be made to improve performance, both in terms of storage requirements, and speed. For instance, since k -mers are generated from sequencing reads each k -mer’s

neighbours (treating edge k -mers separately) can be used to infer the k -mer itself is part of the set. Thus, by querying for a k -mer, along with its possible neighbours (storing edge k -mers separately), we can reduce the false positive rate without requiring additional space. This extension, ‘ k -mer bloom filters’, is discussed in Pellow et al. 2016 and can be applied to CBGs.

4.10.1 Comparison to web search

To index for *natural language* queries, where terms can be syntactically separated and follow some statistical laws, inverted indexes have been demonstrated to be more space-efficient and performant (Brin et al. 2012; Zobel et al. 1998). Nonetheless, signature-based search, of which CBG could be considered an example, has previously been applied to text indexing and web search (Goodwin et al. 2017; Shepherd et al. 1989; Wong et al. 1985; Zobel et al. 1998). In particular, signature-based search using bloom filters and bit-sliced signatures have recently been explored by Goodwin et al. 2017, the team behind Microsoft’s Bing Search to improve their web search capacities .

Their tool “BitFunnel”, based on bloom filters and bit sliced signature, improved Bing’s search capacity by 10x. Although the dimensions of the search problem are different, BitFunnel addresses some of the challenges involved in scaling a bloom filter-based index to trillions of documents. Goodwin et al. 2017 introduced concepts of *higher rank rows* to reduce query execution time, and *frequency-conscious bloom filters* to reduce the memory footprint, improving performance 4-5x over their naive

data structure, which closely resembles the bitmatrix underlying **CBGs**. Both of these ideas can potentially be applied to **CBGs**, possibly improving performance on the scale of data addressed in the previous sections, and also allowing scaling to data sets orders of magnitudes larger.

Since the **CBG** size scales linearly with samples and can easily be distributed across multiple machines, it is feasible for this tool to scale to millions of bacterial genomes and potentially orders of magnitude more with additional extensions. Because one needs to set a maximum number of k -mers per data set in a **CBG**, additional considerations need to be made in order to apply **CBG** to a broader range of data sets, including eukaryote or metagenomic data sets, where the number of k -mers per data set can vary significantly. One solution to this problem is to build a nested structure with different bloomfilter parameters in advance (e.g., max k -mers = 10^5 , 10^6 , 10^7) and insert your sequence into the appropriate level by counting the k -mers in that data set. This approach, “sharding by document size”, has also been suggested by the authors of **BitFunnel** (Goodwin et al. 2017). I hope that this will enable large-scale surveys of the millions of sequence data sets being generated, beyond the exclusively bacterial and viral data indexed here.

4.10.2 Future improvements

The coloured probabilistic de Bruijn graph data structure I presented here has better storage performance than competing methods in the low- k -mer sharing regime. However, further compression is also possible, in particular in a high

k -mer sharing regime. For a collection of genomes with a high level of k -mer sharing between samples, the rows could be compressed with run-length encoding (Teuhola 1978), roaring bit vectors (Chambi et al. 2016), succinct bit vectors (Navarro et al. 2012), or other bit vector compression techniques because the positions of the 1's and 0's in the bloom filters would be expected to correlate strongly. I estimated that in species that have a high degree of k -mer sharing, compression ratios of $> 5\times$ could be achieved; allowing for pan genome indexing of hundreds of thousands of samples in $O(10\text{GB})$ of space. For example, using very naive compression by storing the set-difference between two bloom filters (params: $m = 2.5 \times 10^7; \eta = 3$), positions of $> 99\%$ similarity would require $< 0.6\text{MB}$ of storage (using 4byte integers), $5\times$ less than storing the full bloom filter.

Succinct data structures may also improve the intersection performance for CBG queries (in addition to space compression)—e.g., intersecting roaring bit vectors were found to be up to $900\times$ faster than other compressed bit vectors (Chambi et al. 2016). Furthermore, even if the samples in the collection were from different species, jaccard similarity (Ondov et al. 2016), or a similar distance metric, could cluster data sets by k -mer similarity. This would cluster similar columns together in the bit matrix, potentially enabling further compression of the rows, even if the global k -mer sharing was low.

There are also many improvements which could be made to the CBG scoring algorithm from k -mer alignments, including extensions for insertions and deletions. Ideally, we would be able to include a full alignment step in order to remove our dependency on the k -mer pseudo-alignment. Since each indexed data set in a CBG

is a probabilistic coloured de Bruijn graph, we could also use it to traverse de Bruijn graphs and generate a local assembly of the region best matched by the query, which could then be aligned to the query sequence. This would allow for an iterative approach to search where, for inexact matches, one can traverse the graph to generate contigs and assemble the region of interest in order to pull out all versions of a sequence close to the input query. This would enable both known-sequence search and discovery via assembly to be performed using the same data structure as well as more accurate scoring, via a full alignment, of query results. Chikhi et al. 2013b have proposed an extension to the probabilistic de Bruijn graph that takes advantage of the structure of de Bruijn graphs. Although the set of all possible false positives is large, the set of false positives that create false branching in the de Bruijn graph is bounded and comparably small. As a result, it is possible to enumerate and store these “critical false positives” in order to extend the bloom filter to encode a de Bruijn graph, which can be traversed without false positives. The storing of critical false positives could be used to extend **CBGs** and improve this assembly step by removing false graph traversals.

4.10.3 Surveillance and applications

Since the space required to create a **CBG** is so small, a search of thousands of experiments is possible on a laptop—enabling rapid hypothesis generation and testing that was previously not possible without expensive hardware. The small size of the graph would also allow sharing of the searchable index without requiring the shar-

4. ENABLING RAPID DNA SEARCH OF ALL SEQUENCED BACTERIA AND VIRUSES

ing of the underlying data. For example, combined with a distributed data-sharing system, such as the Beacon Project by the Global Alliance for Genomics Health (GA4GH) (*ga4gh-Beacon* 2017), which provides queries for allele presence (in human data), CBGs could provide presence/absence responses to arbitrary sequence queries aiding data-sharing without the risks of sharing the underlying data. This could also potentially enable archiving of the raw data, while maintaining search in $100\times$ less space.

Future applications of CBGs could range from bacterial surveillance to metagenomic read classification. K -mer based metagenomic read classifiers, such as **Kraken** (Wood et al. 2014), or **CLARK** (Ounit et al. 2015), or those based on pseudo-alignment, such as **kallisto** (Bray et al. 2016), require storing a mapping from k -mers to taxon. These databases often require large amounts of memory to hold; for example, the standard **kraken** database requires 122GB of RAM (Wood et al. 2014). Storing these k -mer-taxon classifications as colours in a CBG could allow for storing the index in a much smaller space.

Sharing of genomic elements can link samples in outbreaks, and surveillance of antimicrobial resistance genes and variants can help detect important trends, such as the emergence and spread of multi-drug-resistance (Jassal et al. 2009; Mathers et al. 2015; Souli et al. 2008). Despite its importance in understanding how the underlying genetic elements spread, information on the distribution and prevalence of these genomic mechanisms is often unavailable or difficult to interrogate (World Health Organization 2014). The CBG software also provides an application programming interface (API), which can be used to host a very large graph over the web on

a cluster of cloud-based servers. Combining this sufficient metadata could aid in investigations by enabling search for variants, antimicrobial resistance genes, and mobile genetic elements in a collection of very large collections of data sets, which can be incrementally updated and cross-referenced with relevant metadata.

4.11 Extended methods

4.11.1 Simulated scaling comparison

In order to estimate the peak storage requirements for `mccortex`, SBT and CBG we simulated scaling in two regimes: a high k -mer sharing, where each sample only adds 100 novel k -mers to the pan-genome, and a low k -mer sharing, where each sample adds 10,000 k -mers to the pan-genome (empirically fewer than what we observe in the `microbial-CBG` in Section 4.5). The storage requirements for each of the tools were calculated as follows.

As discussed in section 4.3, a CBG requires a new column of length m per sample so the number of bits required scales with the number of samples (N), independent of the number of k -mers. Here we use $m = 25 \times 10^6$, assuming $\approx 5 \times 10^6$ k -mers per sample (see Section 4.3.3):

$$CBG_{size}[bytes] = mN/8. \tag{4.6}$$

`mccortex` stores k -mer presence with N_k bit vectors representing the Boolean presence of each k -mer (see Figure 4.4(b)). `mccortex` also requires 9 bytes to store each

k -mer and associated edges. Thus, it scales with (N) and the number of k -mers (N_k) as:

$$McCortex_{size}[bytes] = (N_k(9 + (N/8))). \quad (4.7)$$

For SBT, I modelled the peak storage requirement of the data structure (without compression). Assuming a perfect binary tree SBT requires $2N$ bit vectors of length $(\approx N_k)$, as recommended by the authors (Solomon et al. 2016), it follows:

$$SBT_{size}[bytes] = 2NN_k/8. \quad (4.8)$$

4.11.2 Empirical scaling comparison SBTs and CBGs

To demonstrate empirical scaling of storage and speed I took 2,000 randomly chosen *Enterobacteriaceae* data sets SRA/ENA, and I then further randomly sub-sampled these into collections of sizes 10, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, and 2000 samples. For each set I built a SBT and a CBG from bloom filters using k -mers from error cleaned mccortex graphs. The mccortex graphs were cleaned using the mccortex auto-tuned thresholds for low coverage k -mers and graph tips.

Bloom filter construction was not included in the build time as this step was parallelised across many cores. As a result, construction time refers to the time taken to build each data structure from the original bloom filters, which was done using a single CPU process. The search time comparison was run with ‘bt query’

and ‘cbg search’, using k -mer thresholds 100% and 80%.

The CBG bloom filters were built with parameters $m = 25 \times 10^6$ and $\eta = 3$. SBT bloom filters were built with $\eta = 1$ and bloom filter size (m) equal to the count of the total number of k -mers in the collections of graphs, as is recommended in the SBT publication (Solomon et al. 2016). Jellyfish v2.2.5 was used to count the unique k -mers in the set of WGS data sets (Marçais et al. 2011), which was required to set the SBT bloom filter size. Construction time and query time analyses were run on a Dell PowerEdge R820 with 32 cores and 1 Tb RAM. CBG was run in server mode (‘hug cbg’) for the search profiling in order to exclude boot up time overhead.

4.11.3 Assessing genotyping accuracy

Single-colour de Bruijn graphs for 3,528 *M. tuberculosis* WGS data sets from Walker et al. 2015 were built and error cleaned with k -mer size 21 using `mccortex` v0.0.3-539-g22e27b7 (Turner et al. 2017). A CBG was built from the resulting 3,528 single-colour DBGs using parameters $k = 21$, $m=2.5 \times 10^7$, and $\eta = 3$. SNP calls were made using `cortex` (Iqbal et al. 2012) (independent workflow, $k=31$) on the 3,528 samples. We searched for the *M. tuberculosis* variant probe sets generated in section 2.4.1.1 in a CBG index built from these 3,528 samples. Singleton variants were excluded, and a de-duplicated list of 68,269 SNPs was constructed. *Variant probe sets* were made at each of these sites where each probe set had multiple possible references and alternate alleles for the SNP on multiple variant backgrounds (taken from the 3,528 Cortex VCFs). We then genotyped these variants using the method

described in section 4.3.2.3. The expected number of false discoveries for an allele from a probe set of length 43 with bloom filter parameters $m=2.5 \times 10^7, \eta = 3$ is 10^{-21} per colour, which is well below the expected error rate from the underlying data. Searching the 3,528 samples for 68,269 took just under 90minutes on a single CPU core—an effective genotyping rate of greater than 46,000 genotypes per second.

One hundred random *M. tuberculosis* samples were selected from the 3,528 genotyped samples. Their genotypes from the pipeline (described in Walker et al. 2015) at these variant positions were extracted from VCFs provided by the authors. The VCFs were generated with a pipeline using `samtools` v0.1.18 (Li et al. 2009) after mapping with `stampy` (Lunter et al. 2011) with a haploid genotyping model, and filtered for confident calls. Positions filtered by `samtools` and `CBG` in each sample were excluded and the remaining genotype calls were used to determine the genotype concordance between the two methods. We compared the genotypes of 100 random samples with those generated with the `stampy` + `samtools` pipeline from Walker et al. 2015, excluding filtered positions.

4.11.4 ENA download and processing

All bacterial and viral (N=455,632, December 2016) WGS data sets (170TB) were downloaded from the ENA’s Globus FTP archive via the ‘Pathogen’ endpoint. We refer to this data set as `microbial-CBG`. Single colour de Bruijn graphs for each data set were built and error cleaned with k -mer size 31 using `mccortex` v0.0.3-539-g22e27b7 (Turner et al. 2017). 31 was chosen to have a large number of distinct k-

mers across all microbes. However, it was not practical to count k -mer abundances across such a large collection of data sets so this choice was a heuristic. Error cleaning was done using `mccortex`, which removes short tips and low coverage unitigs that likely resulted from sequencing errors. Error cleaning was done primarily to reduce errors at query time; the index could also be built from uncleaned k -mer sets or de Bruijn graphs with the same space requirements. Since CBGs do not contain depth of coverage information, k -mer errors inserted into the graph could not be resolved at query time via consensus.

Bloom filters were built using the k -mers from the cleaned graphs with parameters $m = 2.5 \times 10^7$ and $\eta = 3$. Parameters m and η were chosen based on a maximum false-positive rate per- k -mer per-colour of $p = 0.3$ for maximum of $K_{max} = 10^7$ k -mers. Samples with more k -mers were excluded to keep an upper bound on the false-discovery rate (FDR) of $v = 5 \times 10^{-6}$ for a 40base-pair (10- k -mer) query ($L_{min} = 40$). 7,799 were masked in the CBG and excluded from further analyses as they exceeded 10^7 k -mers in their cleaned graph. This resulted in 447,833 bloom filters that were accessible to search in the final CBG.

The CBG was constructed from these bloom filters in a parallel manner, dividing the samples into 2,500 subsets, and merging the resulting graphs. These sub-graphs were merged by appending the rows for each index and inserting these into a combined Berkeley-DB, although it would have been possible to distribute queries across the sub-graphs without combining, if resource constraints had made it necessary.

³LogLog refers to the fact that the space required is $O(\log(\log(N_{max})))$ bits

4.11.4.1 Approximate counting of unique k -mers

Since the number of k -mers in the microbial-CBG was expected to be very large, exact counting was prohibitive. HyperLogLog³ is a probabilistic data structure used for estimating the cardinality of a set in fixed space (Flajolet et al. 2007). We inserted the k -mers from each of the cleaned de Bruijn graphs into a HyperLogLog data structure using `redis v3.2.6 (PFADD)` (Flajolet et al. 2007; Sanfilippo 2017). Merging the resulting HyperLogLogs allows approximate counting of the combined set of k -mers with a standard error of 0.81% (PFCOUNT). We found $6.05 \times 10^{10} \pm 5 \times 10^7$ unique k -mers in the microbial-CBG, resulting in an average unique k -mer contribution from each sample of $\approx 1.35 \times 10^5$ k -mers.

4.11.4.2 Species assignment of ENA data sets

The proportion of different species in each data set was determined by `kraken` and `bracken` (Lu et al. 2017a; Wood et al. 2014). Classification was done on the k -mers from each cleaned de Bruijn graph using the `minikraken_20141208` database. The resulting taxonomy labels assigned by `kraken` were analysed by `bracken` to estimate the number of k -mers originating from each species present in a sample. Taxonomic ranks were assigned using the EBI's taxonomy service API.

4.11.5 Plasmid search and exclusion of contaminated samples

2,826 plasmid sequences were taken from the EBI plasmid pages (www.ebi.ac.uk/genomes/plasmid.html; December 2016) and downloaded from the ENA. I then queried the `microbial-CBG` for these sequences with a proportion of k-mers threshold of 40% ($T=40\%$) and filtered for hits with $T \geq 90\%$ for downstream analysis. All of the queries were run on an Amazon Web Services i3.2xlarge virtual machine with 8vCPUs, 61GiB mem, and a 1.9TB non-volatile memory express solid state drive. GNU parallel was used to parallelise queries across the 8 available processors (Tange et al. 2011). Queries were run with 1 GB Berkeley-DB cachesize.

In order to determine the distribution of plasmids across taxa, we filtered these hits for accessions which were bacteria and had no secondary genus above 0.1% frequency. This criterion was chosen to avoid multi-copy plasmids from contaminating species, establishing false-positive hits within a non-host genus. 41% (184652/447,833) of accessions and 62% (415,181/668,720) of search hits passed this filter. We then filtered for all plasmids which had been seen at least 5 times each in more than one genus and had less than 99% of their observations in the most frequent genus. We found that 37 plasmids across 13 genera matched these criteria. By simulating mixtures of *Salmonella enterica* and *Escherichia coli* at relative abundances of 0.0001, 0.001, 0.01, 0.02, 0.05, 0.1, 0.15, 0.2, 0.25, and 0.3 we found we could observe the minority species above 2% frequency (the limit of detec-

tion was not lower because we had applied Kraken after error-cleaning of de Bruijn graphs). All 37 plasmids reported had at least one observation at a copy number of 5 (which, since we could detect contaminants at 2% frequency, would correspond to a copy number of above 250 if it came from a contaminant) and 16/37 had an observation at $2000\times$ copy number.

4.11.6 *Yersinia pestis* plasmid search

We searched for the 182,913bp *Yersinia pestis* biovar Orientalis str. IP275 plasmid pIP1202 (CP000603.1) at T=50% in the `microbial-CBG` resulting in 2341 hits. 244 of these accessions were also available in the NARMS database with extended metadata. These data were intersected with the results from “All AMR against all ENA” to find AMR genes in accessions, which also contain the plasmid. Resistance to the following drugs was inferred when any of the following genes were present: tetracycline: `tet`; chloramphenicol: `floR`; streptomycin: `aadA`, `AAC`, `APH`; beta lactamases: `CMY`, `TEM`; sulfonamides: `sul`; quinolone: `QnrB`; polymixin: `arnA`; macrolides: `msrE`; and multi-resistance: `emr`, `mdt`, `marA`. The samples were classified as MDR if the sample has resistance genes of 2 of tetracycline, sulfonamides and streptomycin, and XDR if MDR and resistance to at least 4 drugs (or 3 drugs and a multi-resistance gene).

CHAPTER 5

Conclusion

In this thesis I have explored applications of genome graphs, specifically de Bruijn graphs and probabilistic de Bruijn graphs, to large clinical microbiology data sets. With `Mykrobe predictor`, I explored their use in developing a fast and easy-to-use tool, which could be used to predict drug resistance from WGS data. With Coloured Bloom Graphs, I extended probabilistic de Bruijn graphs to multiple colours to create a probabilistic k -mer index, which can index very large collections of sequencing data for search. The volume of sequenced microbial genomes is growing exponentially, and is set to accelerate as sequencing-based diagnostics become more common. Sequencing data, often generated for one purpose (e.g., DST), has multiple uses and, if aggregated with appropriate metadata, can be used for more analysis than its original use case. However, currently, much of this data is siloed. As a result, it is difficult to search and analyse it simultaneously, limiting its use.

To finish this thesis, I will discuss how these tools may be integrated, along with other recent advances in bioinformatics (such as MASH (Ondov et al. 2016)), to enable analysis and collaboration on very large data sets, the likes of which may start to be generated if WGS becomes part of routine diagnostic workflows. Rapid in-situ diagnosis is a goal in and of itself, but it also creates huge opportunities for data collection and analysis.

5.1 WGS as a diagnostic

In high-incidence/endemic regions, or where analysis of samples is centralised (e.g., UK *M. tuberculosis* reference lab), it seems likely that batching samples to run

on high throughput, low cost per sample sequencers such as Illumina HiSeq/MiSeq, could prove a cost effective solution for diagnostics and DST in the near future, in particular for pathogens which are difficult or slow to culture. For example, Public Health England's National Mycobacterial Reference Service now sequences every *M. tuberculosis* case it receives and used **Mykrobe predictor** and other software to species ID and test for drug susceptibility. It is the first such centre to routinely use WGS in parallel to its existing workflow.

How a workflow like this will be formally accredited is still an open question and poses challenges to regulatory bodies. As the knowledge base improves, rapid versioning will be essential to make best use of the rapidly improving genotype to phenotype inferences. However, it is not yet clear if it will be the catalogue, each variant/gene of interest, the software, or the integrated whole which will require validation and accreditation. Since the inference is primarily dependent on software, this is could be automated and run much like the "continuous-integration"-testing common in software development and machine learning. However, this would be a significant step change from how validation has been performed on molecular assays, such as MTB/RIF Xpert.

A single assay providing diagnostic information, and data for surveillance and outbreak detection, is an attractive prospect. In high-incidence settings, where the upfront capital costs make financial sense, high throughput Illumina sequencing platforms may be applicable (e.g., in Mumbai there are upwards of 65,000 *M. tuberculosis* cases per year) . However, point-of-care testing via WGS could have also have a huge impact on the management of *M. tuberculosis*, where well-equipped

laboratories are not available. Rapid diagnostics and surveillance through real-time sequencing has recently been demonstrated by Quick et al. in *Ebola* and *Salmonella* using ONT's MinION (Quick et al. 2015, 2016). These real-world demonstrations of nanopore sequencing shows its use in settings, where conventional sequencing technologies are difficult to deploy. The work I presented in Chapter 3 shows this could also be applicable to *M. tuberculosis*, but widespread use would require additional improvements and validation studies. Diagnostics and outbreak investigation will become increasingly real-time and digital. Data can be generated and analysed in a matter of hours, and there is a tremendous opportunity to use these data to better inform clinical decisions in the context of a single sample, or a small outbreak. However, there is also an opportunity to generate prodigious amounts of useful data, which can be shared and integrated globally.

5.2 A cloud-based analysis and data collection platform using de Bruijn graphs

One major advantage of WGS-based diagnostic is that the data can be coupled to data-sharing frameworks, which can then be learned from, making the test more accurate, and creating a virtuous cycle. This principle can not only be applied to pathogens, but also to environmental samples, agricultural samples, or even, ignoring privacy issues, human data. However, the scale and complexity of the data mean that creating such a platform has significant challenges.

5.2. A cloud-based analysis and data collection platform using de Bruijn graphs

Tools like `Mykrobe predictor` and `CBG` might be used to practically set up the compute/network infrastructure to allow people to submit sample data over the internet, and in return receive an analysis of that sample as well as including it in a platform where it can be queried and compared with other samples. With `Mykrobe predictor`, we have an exemplar of a sample analysis (although, there could be many others).

`Mykrobe predictor` is currently implemented as a local application, rather than as a web service. The primary reason behind this decision was that in resource-limited areas, where `Mykrobe predictor` may be run, internet bandwidth is a significant concern. Uploading Gigabytes of sequencing data for every run is not a viable option. However, because the cleaned de Bruijn graph of a bacterial sample is $\approx 50\text{MB}$ in size, approximately $20\times$ smaller than the original data, graphs could be built locally and the cleaned de Bruijn graph uploaded to the cloud for analysis. Using compressed, succinct, or probabilistic de Bruijn graphs may result in a further $10\times$ compression.

A cloud-based analysis system would have many advantages. It would allow for centralised management of software versions and updates, enabling the AMR catalogues to be updated to the most improved version rapidly. It would also enable data to be uploaded to a centralised repository. With `CBG` we have a method for indexing all the graphs for search, allowing the de Bruijn graphs to be inserted into a global index of samples. Periodic polling on this index can allow alerts to be triggered when new samples are uploaded with certain properties (e.g., containing a MGE, or AMR gene of interest, or are part of an outbreak), possibly allowing for

5. CONCLUSION

a faster response to threats.

If the data are hosted in the cloud, then it may be possible to bring the analysis to the data, avoiding the high bandwidth costs of uploading and downloading the large data files associated with sequencing projects. Practically, this might be done with containers (Ali et al. 2016), where software can be bundled in contained environments, uploaded, and run in the cloud. This is the workflow already used by Illumina with their data storage and analysis platform basespace (*Illumina Basespace* 2017).

A desired feature of such a web-service would be to query for sample similarity (e.g., are there any uploaded samples similar to mine?), cluster samples by similarity, or build a similarity tree of a group of samples. There are several ways to approach this problem using de Bruijn graphs: 1) de novo variant calling can be performed on de Bruijn graphs using tools like Cortex, allowing SNP distances to be calculated; 2) genotyping on a pre-calculated set of variants can be performed by *Mykrobe predictor* on DBGs, or this set of variants can be genotyped by *CBG* to generate SNP distances on a fixed set of variants; or 3) k -mer based similarity can be rapidly evaluated with low resources MinHash approximations of Jaccard Similarity with tools like MASH (Ondov et al. 2016) (with k -mers extracted from the de Bruijn graph). MinHashes have also recently been shown to be useful for metagenomic containment, e.g., detecting genomes contained within a metagenome, also using k -mer sets (Koslicki et al. 2017; *Mash Screen* 2017) and for building similarity trees (Katz 2017; Ondov et al. 2016).

Combining these tools would allow for a web service using de Bruijn graphs

5.2. A cloud-based analysis and data collection platform using de Bruijn graphs

to provide: 1) AMR prediction or other gene/variant genotyping or k -mer based analysis, 2) similarity comparison, clustering, SNP distance comparison and, 3) arbitrary gene/variant search. This could form the core functions to answer many questions on these data efficiently, requiring reasonable space and time. If these could also be combined with extensive, structured-metadata, currently missing from our raw read archives, it could be a powerful tool in the fight against antimicrobial resistance.

APPENDIX A

Chapter 2 Appendix

A.1 Accessions for Simulation 1

Accession identifiers of the pairs: (SRR1182410, ERR410084), (SRR1182413, ERR410093), (SRR1182415, ERR410136), (SRR1609104, ERS398139), (SRR221652, ERS398155), (SRR398319, ERS398179), (SRR496759, ERS398307), (SRR496761, ERS398353), (SRR496889, ERS398370), (ERR085178, ERR410084), (ERR085180, ERR410093), (ERR085182, ERR410136), (ERR085188, ERS398139), (ERR085190, ERS398155), (ERR085192, ERS398179), (ERR085258, ERS398307), (ERR085260, ERS398353), (ERR085262, ERS398370)

A.2 Figures and tables

Table A.1: Results for **Mykrobe predictor** on the *S. aureus* training set (St_A1). Resistance prediction results for **Mykrobe predictor** on the *S. aureus* training set St_A1 treating Stokes Disc test as truth. FN: False negative calls. R: total number of resistant samples. FP: false positives. S: total number of susceptible samples. VME: very major error rate (false negative rate. ME: major error rate (false positive rate). PPV: positive predictive value. NPV: negative predictive value. N/A: Not Applicable. Error rates shown with 95% CI calculated by Clopper-Pearson; FN/FP rate only shown where number of resistant/susceptible samples > 10.

Drug	FN(R)	FP(S)	VME (%)	ME (%)	PPV (%)	NPV (%)
PEN	3 (437)	2 (58)	0.7 (0.1-2.0)	3.4 (0.4-11.9)	99.5 (98.4-99.9)	94.9 (85.9-98.9)
CIP	6 (170)	8 (325)	3.5 (1.3-7.5)	2.5 (1.1-4.8)	95.3 (91.0-98.0)	98.1 (96.0-99.3)
METH	0 (158)	2 (337)	0.0 (0-2.3)	0.6 (0.1-2.1)	98.8 (95.6-99.8)	100.0 (98.9-100)
ERY	0 (133)	1 (362)	0.0 (0-2.7)	0.3 (0.0-1.5)	99.3 (95.9-100.0)	100.0 (99.0-100)
CLIN	0 (88)	1 (89)	0.0 (0-4.1)	1.1 (0.0-6.1)	98.9 (93.9-100.0)	100.0 (95.9-100)
FUS	0 (39)	4 (456)	0.0 (0-9.0)	0.9 (0.2-2.2)	90.7 (77.9-97.4)	100.0 (99.2-100)
TET	0 (27)	0 (468)	0.0 (0-12.8)	0.0 (0-0.8)	100.0 (87.2-100)	100.0 (99.2-100)
TRIM	3 (13)	0 (304)	23.1 (5.0-53.8)	0.0 (0-1.2)	100.0 (69.2-100)	99.0 (97.2-99.8)
GEN	0 (7)	0 (488)	0.0 (0-41.0)	0.0 (0-0.8)	100.0 (59.0-100)	100.0 (99.2-100)
RIF	0 (2)	2 (493)	0.0 (0-84.2)	0.4 (0.0-1.5)	50.0 (6.8-93.2)	100.0 (99.3-100)
MUP	0 (2)	2 (176)	0.0 (0-84.2)	1.1 (0.1-4.0)	50.0 (6.8-93.2)	100.0 (97.9-100)
VAN	0 (0)	0 (495)	NaN	0.0 (0-0.7)	NaN	100.0 (99.3-100)

Table A.2: Results for Disc on the *S. aureus* validation set (St_B1). Resistance prediction results for Disc on the *S. aureus* validation set St_B1 compared against the consensus phenotype. FN: False negative calls. R: total number of resistant samples. FP: false positives. S: total number of susceptible samples. VME: very major error rate (false negative rate. ME: major error rate (false positive rate). PPV: positive predictive value. NPV: negative predictive value. N/A: Not Applicable. Error rates shown with 95% CI calculated by Clopper-Pearson; FN/FP rate only shown where number of resistant/susceptible samples > 10.

Drug	FN(R)	FP(S)	VME (%)	ME (%)	PPV (%)	NPV (%)
PEN	28 (377)	14 (93)	7.4 (5.0-10.6)	15.1 (8.5-24.0)	96.1 (93.6-97.9)	73.8 (64.4-81.9)
ERY	2 (79)	6 (392)	2.5 (0.3-8.8)	1.5 (0.6-3.3)	92.8 (84.9-97.3)	99.5 (98.2-99.9)
CIP	1 (65)	7 (406)	1.5 (0.0-8.3)	1.7 (0.7-3.5)	90.1 (80.7-95.9)	99.8 (98.6-100.0)
METH	0 (54)	2 (417)	0.0 (0-6.6)	0.5 (0.1-1.7)	96.4 (87.7-99.6)	100.0 (99.1-100)
FUS	0 (40)	1 (430)	0.0 (0-8.8)	0.2 (0.0-1.3)	97.6 (87.1-99.9)	100.0 (99.1-100)
CLIN	2 (21)	0 (96)	9.5 (1.2-30.4)	0.0 (0-3.8)	100.0 (82.4-100)	98.0 (92.8-99.8)
TET	0 (17)	1 (454)	0.0 (0-19.5)	0.2 (0.0-1.2)	94.4 (72.7-99.9)	100.0 (99.2-100)
RIF	1 (5)	0 (466)	20.0 (0.5-71.6)	0.0 (0-0.8)	100.0 (39.8-100)	99.8 (98.8-100.0)
GEN	1 (3)	0 (468)	33.3 (0.8-90.6)	0.0 (0-0.8)	100.0 (15.8-100)	99.8 (98.8-100.0)
MUP	0 (2)	0 (348)	0.0 (0-84.2)	0.0 (0-1.1)	100.0 (15.8-100)	100.0 (98.9-100)
VAN	0 (0)	0 (472)	NaN	0.0 (0-0.8)	NaN	100.0 (99.2-100)
TRIM	0 (0)	0 (0)	NaN	NaN	NaN	NaN

Table A.3: Results for Phoenix on the *S. aureus* validation set (St_B1). Resistance prediction results for Phoenix on the *S. aureus* validation set St_B1 compared against the consensus phenotype. FN: False negative calls. R: total number of resistant samples. FP: false positives. S: total number of susceptible samples. VME: very major error rate (false negative rate. ME: major error rate (false positive rate). PPV: positive predictive value. NPV: negative predictive value. N/A: Not Applicable. Error rates shown with 95% CI calculated by Clopper-Pearson; FN/FP rate only shown where number of resistant/susceptible samples > 10.

Drug	FN(R)	FP(S)	VME (%)	ME (%)	PPV (%)	NPV (%)
PEN	6 (377)	15 (94)	1.6 (0.6-3.4)	16.0 (9.2-25.0)	96.1 (93.7-97.8)	92.9 (85.3-97.4)
ERY	1 (79)	0 (391)	1.3 (0.0-6.9)	0.0 (0-0.9)	100.0 (95.4-100)	99.7 (98.6-100.0)
CIP	7 (65)	0 (406)	10.8 (4.4-20.9)	0.0 (0-0.9)	100.0 (93.8-100)	98.3 (96.5-99.3)
METH	2 (54)	79 (417)	3.7 (0.5-12.7)	18.9 (15.3-23.0)	39.7 (31.3-48.6)	99.4 (97.9-99.9)
FUS	2 (41)	1 (430)	4.9 (0.6-16.5)	0.2 (0.0-1.3)	97.5 (86.8-99.9)	99.5 (98.3-99.9)
CLIN	7 (25)	1 (97)	28.0 (12.1-49.4)	1.0 (0.0-5.6)	94.7 (74.0-99.9)	93.2 (86.5-97.2)
TET	0 (17)	2 (454)	0.0 (0-19.5)	0.4 (0.1-1.6)	89.5 (66.9-98.7)	100.0 (99.2-100)
RIF	0 (5)	0 (466)	0.0 (0-52.2)	0.0 (0-0.8)	100.0 (47.8-100)	100.0 (99.2-100)
GEN	1 (3)	2 (468)	33.3 (0.8-90.6)	0.4 (0.1-1.5)	50.0 (6.8-93.2)	99.8 (98.8-100.0)
MUP	0 (2)	0 (348)	0.0 (0-84.2)	0.0 (0-1.1)	100.0 (15.8-100)	100.0 (98.9-100)
VAN	0 (0)	0 (472)	NaN	0.0 (0-0.8)	NaN	100.0 (99.2-100)
TRIM	0 (0)	0 (0)	NaN	NaN	NaN	NaN

Table A.4: Validation (St_B1) *S. aureus* SeqSphere results. Resistance prediction results for SeqSphere on the *S. aureus* validation set St_B1 compared against the consensus phenotype. FN: False negative calls. R: total number of resistant samples. FP: false positives. S: total number of susceptible samples. VME: very major error rate (false negative rate. ME: major error rate (false positive rate). PPV: positive predictive value. NPV: negative predictive value. N/A: Not Applicable. Error rates shown with 95% CI calculated by Clopper-Pearson.

Drug	FN(R)	FP(S)	VME (%)	ME (%)	PPV (%)	NPV (%)
ERY	1 (79)	392 (392)	1.3 (0.0-6.9)	100.0 (99.1-100)	16.6 (13.3-20.3)	0.0 (0-97.5)
METH	1 (54)	0 (417)	1.9 (0.0-9.9)	0.0 (0-0.9)	100.0 (93.3-100)	99.8 (98.7-100.0)
CLIN	0 (25)	97 (97)	0.0 (0-13.7)	100.0 (96.3-100)	20.5 (13.7-28.7)	NaN
GEN	1 (3)	0 (468)	33.3 (0.8-90.6)	0.0 (0-0.8)	100.0 (15.8-100)	99.8 (98.8-100.0)
MUP	0 (2)	0 (348)	0.0 (0-84.2)	0.0 (0-1.1)	100.0 (15.8-100)	100.0 (98.9-100)
RIF	0 (0)	0 (0)	NaN	NaN	NaN	NaN
CIP	0 (0)	0 (0)	NaN	NaN	NaN	NaN
VAN	0 (0)	0 (472)	NaN	0.0 (0-0.8)	NaN	100.0 (99.2-100)
TRIM	0 (0)	0 (0)	NaN	NaN	NaN	NaN
PEN	0 (0)	0 (0)	NaN	NaN	NaN	NaN
TET	0 (0)	0 (0)	NaN	NaN	NaN	NaN
FUS	0 (0)	0 (0)	NaN	NaN	NaN	NaN

Table A.5: Results for Mykrobe predictor on the *M. tuberculosis* training set (MTBC_A1). Resistance prediction results for Mykrobe predictor on the *M. tuberculosis* training set MTBC_A1 compared against the consensus phenotype. FN: False negative calls. R: total number of resistant samples. FP: false positives. S: total number of susceptible samples. VME: very major error rate (false negative rate. ME: major error rate (false positive rate). PPV: positive predictive value. NPV: negative predictive value. N/A: Not Applicable. Error rates shown with 95% CI calculated by Clopper-Pearson; FN/FP rate only shown where number of resistant/susceptible samples > 10

Drug	FN(R)	FP(S)	VME (%)	ME(%)	PPV(%)	NPV(%)
ISO	33 (275)	11 (1636)	12.0 (8.4-16.4)	0.7 (0.3-1.2)	95.7 (92.4-97.8)	98.0 (97.2-98.6)
RIF	11 (102)	15 (1768)	10.8 (5.5-18.5)	0.8 (0.5-1.4)	85.8 (77.7-91.9)	99.4 (98.9-99.7)
STREP	32 (73)	0 (404)	43.8 (32.2-55.9)	0.0 (0-0.9)	100.0 (91.4-100)	92.7 (89.8-94.9)
ETH	19 (57)	35 (1832)	33.3 (21.4-47.1)	1.9 (1.3-2.6)	52.1 (40.0-63.9)	99.0 (98.4-99.4)
CIP	4 (22)	6 (252)	18.2 (5.2-40.3)	2.4 (0.9-5.1)	75.0 (53.3-90.2)	98.4 (96.0-99.6)
OFX	1 (16)	4 (114)	6.3 (0.2-30.2)	3.5 (1.0-8.7)	78.9 (54.4-93.9)	99.1 (95.1-100.0)
MOX	0 (15)	5 (116)	0.0 (0-21.8)	4.3 (1.4-9.8)	75.0 (50.9-91.3)	100.0 (96.7-100)
KAN	4 (9)	1 (95)	44.4 (13.7-78.8)	1.1 (0.0-5.7)	83.3 (35.9-99.6)	95.9 (89.9-98.9)
CAP	2 (7)	2 (99)	28.6 (3.7-71.0)	2.0 (0.2-7.1)	71.4 (29.0-96.3)	98.0 (92.9-99.8)
AMI	0 (6)	1 (106)	0.0 (0-45.9)	0.9 (0.0-5.1)	85.7 (42.1-99.6)	100.0 (96.5-100)

Figure A.1: Screenshot of the Mykrobe predictor *S. aureus* desktop app showing drugs split by resistant or susceptible prediction.

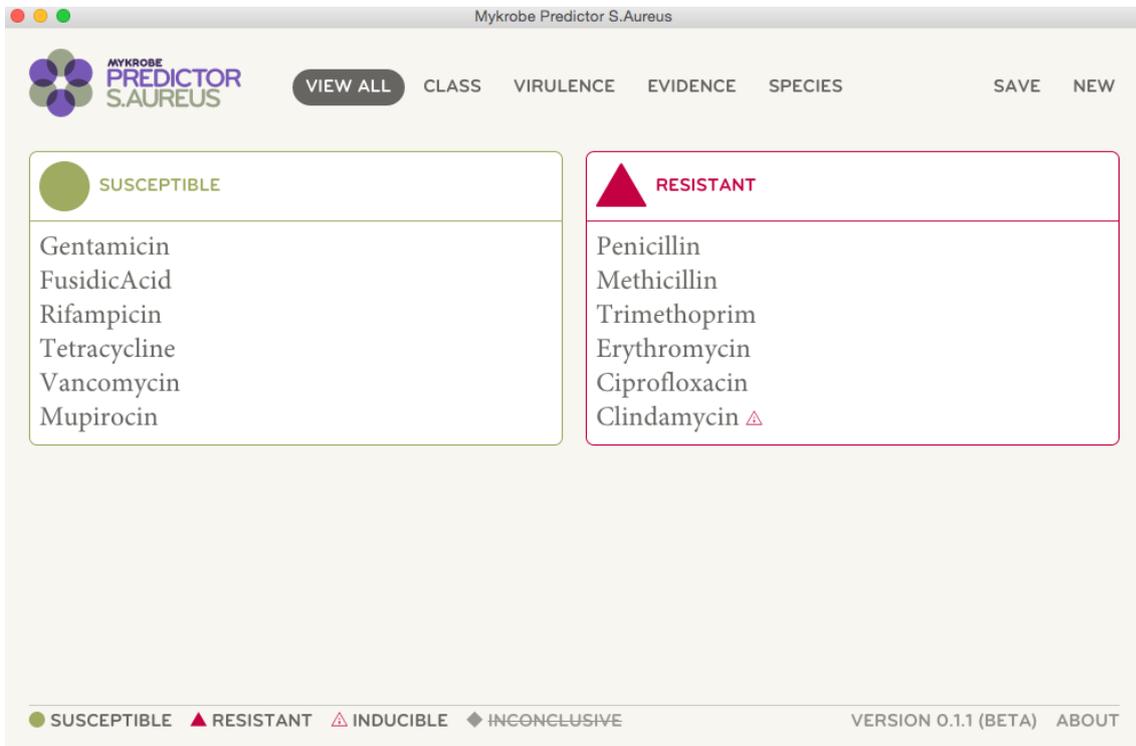
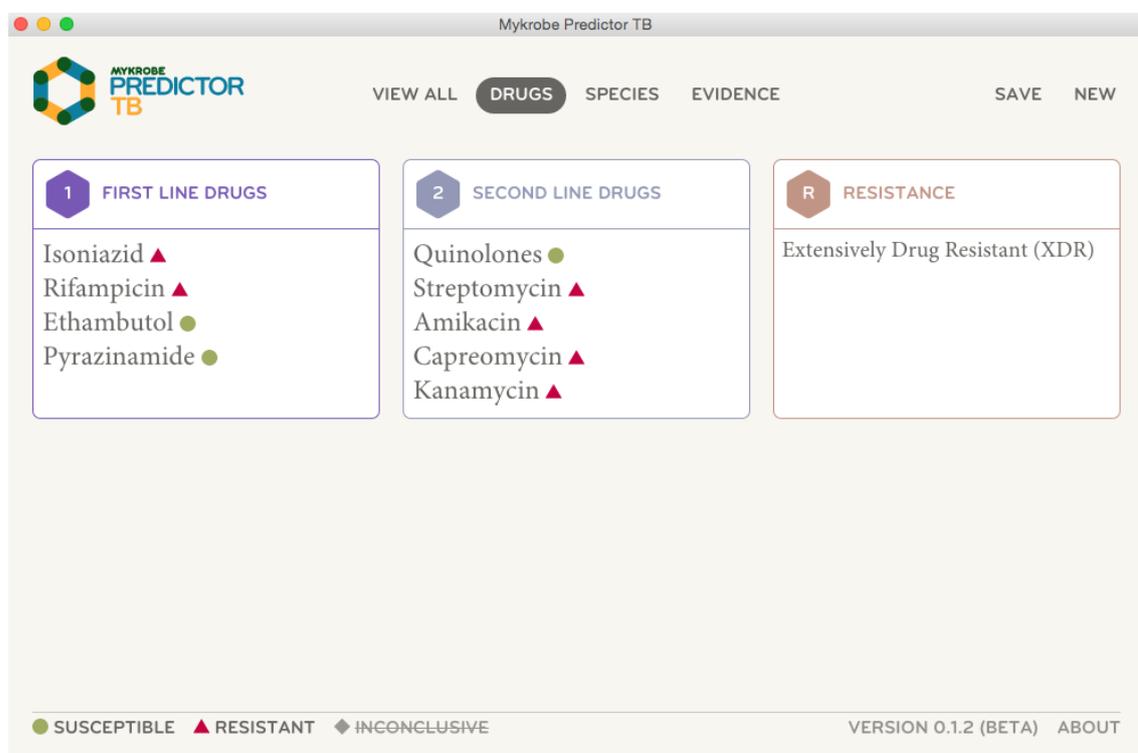


Figure A.2: Screenshot of the Mykrobe predictor TB desktop app showing drugs split by first and second line (TB) alongside resistant or susceptible prediction.



A. CHAPTER 2 APPENDIX

Figure A.3: Screenshot of the Mykrobe predictor S. aureus desktop app with evidence for each of the resistance calls.

The screenshot displays the Mykrobe Predictor S. aureus desktop application interface. The title bar reads "Mykrobe Predictor S.Aureus". The main header includes the logo, navigation tabs for "VIEW ALL", "CLASS", "VIRULENCE", "EVIDENCE" (which is selected), and "SPECIES", along with "SAVE" and "NEW" buttons. The evidence section is divided into five panels, each representing a different antibiotic resistance call with its associated gene and coverage statistics.

Antibiotic	Gene	Percent recovered	Median coverage
CIPROFLOXACIN	Resistance mutation found: S80F in gene <i>grlA</i> Resistant allele seen 33 times Susceptible allele seen 0 times	-	-
ERYTHROMYCIN	<i>ermA</i> gene found	100%	84
METHICILLIN	<i>mecA</i> gene found	99%	49
PENICILLIN	<i>blaZ</i> gene found	100%	48
TRIMETHOPRIM	Resistance mutation found: F99Y in gene <i>dfrB</i> Resistant allele seen 32 times Susceptible allele seen 0 times	-	-

At the bottom right of the application window, the text "VERSION 0.1.1 (BETA) ABOUT" is visible.

APPENDIX B

Chapter 3 Appendix

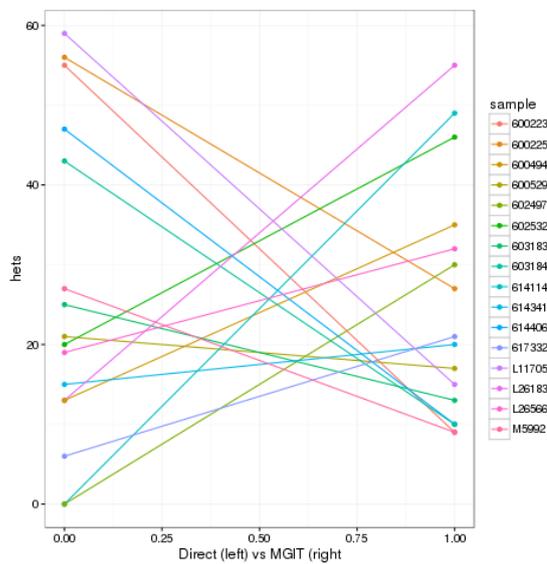


Figure B.1: Number of confident (genotype confidence > 1) heterozygous SNPs called in paired vs direct/MGIT samples where both samples have at least 5× depth of coverage. There was no consistent pattern of more heterozygotes in the direct samples.

Bibliography

- A. Schäffer, Alejandro et al. (1999). “IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices”. In: *Bioinformatics* 15.12, pp. 1000–1011.
- Abraham, Edward P and Ernst Chain (1940). “An enzyme from bacteria able to destroy penicillin”. In: *Nature* 146, p. 837.
- Abreu Maschmann, Raquel de et al. (2013). “Performance of the GenoType MTB-DRplus assay directly on sputum specimens from Brazilian patients with tuberculosis treatment failure or relapse”. In: *Journal of clinical microbiology* 51.5, pp. 1606–1608.
- Ali, Ahmed Abdullah et al. (2016). “The case for docker in multicloud enabled bioinformatics applications”. In: *International Conference on Bioinformatics and Biomedical Engineering*. Springer, pp. 587–601.
- Almodaresi, Fatemeh et al. (2017). “A space and time-efficient index for the compacted colored de Bruijn graph”. In: *bioRxiv*, p. 191874.

- Altschul, SF et al. (1990). “Basic local alignment search tool.” In: *J. Mol. Biol.* 215.3, pp. 403–10. ISSN: 0022-2836. DOI: [10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
- Altschul, S et al. (1997). “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs”. In: *Nucleic acids research* 25.17, pp. 3389–3402.
- Aminov, Rustam I (2010). “A brief history of the antibiotic era: lessons learned and challenges for the future”. In: *Frontiers in microbiology* 1.
- Andrews, Jennifer M (2001). “BSAC standardized disc susceptibility testing method”. In: *Journal of Antimicrobial Chemotherapy* 48.suppl.1, pp. 43–57.
- Anson, Luke et al. (2018). “DNA extraction from primary liquid blood cultures for blood stream infection diagnosis using whole genome sequencing”. In: *submitted*.
- Antipov, Dmitry et al. (2016). “plasmidSPAdes: assembling plasmids from whole genome sequencing data”. In: *Bioinformatics* 32.22, pp. 3380–3387.
- Barlow, Miriam and Barry G Hall (2002). “Phylogenetic analysis shows that the OXA β -lactamase genes have been on plasmids for millions of years”. In: *Journal of Molecular Evolution* 55.3, pp. 314–321.
- Bartlett, John G et al. (2013). “Seven ways to preserve the miracle of antibiotics”. In: *Clinical Infectious Diseases* 56.10, pp. 1445–1450.
- Beiko, Robert G et al. (2005). “Highways of gene sharing in prokaryotes”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.40, pp. 14332–14337.
- Belk, Keith et al. (2016). “Succinct Colored de Bruijn Graphs”. In: *Biorxiv*, p. 040071. DOI: [10.1101/040071](https://doi.org/10.1101/040071).

- Bennett, PM (2008). “Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria”. In: *British journal of pharmacology* 153.S1.
- Bertels, Frederic et al. (2014). “Automated reconstruction of whole-genome phylogenies from short-sequence reads”. In: *Molecular biology and evolution* 31.5, pp. 1077–1088.
- Bethesda, MD (2011). “Sequence Read Archive Submissions Staff. Searching using BLAST”. In: *SRA Knowledge Base [Internet]*. URL: <https://www.ncbi.nlm.nih.gov/books/NBK56920/?report=classic>.
- Bhowmick, T et al. (2013). “Controlled multicenter evaluation of a bacteriophage-based method for rapid detection of *Staphylococcus aureus* in positive blood cultures”. In: *Journal of clinical microbiology* 51.4, pp. 1226–1230.
- Billal, Dewan S et al. (2011). “Whole genome analysis of linezolid resistance in *Streptococcus pneumoniae* reveals resistance and compensatory mutations”. In: *BMC genomics* 12.1, p. 512.
- Bloom, Burton H (1970). “Space/time trade-offs in hash coding with allowable errors”. In: *Communications of the ACM* 13.7, pp. 422–426.
- Bode, Lonneke GM et al. (2012). “Rapid detection of methicillin-resistant *Staphylococcus aureus* in screening samples by relative quantification between the *mecA* gene and the SA442 gene”. In: *Journal of microbiological methods* 89.2, pp. 129–132.

- Bradley, Phelim et al. (2015). “Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*”. In: *Nature communications* 6, p. 10063.
- Bradley, Phelim et al. (2017). “Real-time search of all bacterial and viral genomic data”. In: *bioRxiv*. DOI: [10.1101/234955](https://doi.org/10.1101/234955). eprint: <https://www.biorxiv.org/content/early/2017/12/15/234955.full.pdf>. URL: <https://www.biorxiv.org/content/early/2017/12/15/234955>.
- Bray, Nicolas L et al. (2016). “Near-optimal probabilistic RNA-seq quantification”. In: *Nat Biotechnol* 34.5, pp. 525–527. ISSN: 1087-0156. DOI: [10.1038/nbt.3519](https://doi.org/10.1038/nbt.3519).
- Brin, Sergey and Lawrence Page (2012). “Reprint of: The anatomy of a large-scale hypertextual web search engine”. In: *Computer networks* 56.18, pp. 3825–3833.
- Broder, Andrei and Michael Mitzenmacher (2004). “Network applications of bloom filters: A survey”. In: *Internet mathematics* 1.4, pp. 485–509.
- Brown, Amanda C et al. (2015). “Rapid whole-genome sequencing of *Mycobacterium tuberculosis* isolates directly from clinical samples”. In: *Journal of clinical microbiology* 53.7, pp. 2230–2237.
- Brown, Eric D and Gerard D Wright (2016). “Antibacterial drug discovery in the resistance era”. In: *Nature* 529.7586, pp. 336–343.
- Brown-Jaque, Maryury et al. (2015). “Transfer of antibiotic-resistance genes via phage-related mobile elements”. In: *Plasmid* 79, pp. 1–7.
- Bruijn, FA de (1946). “A combinatorial problem”. In: *Proceedings of the Section of Sciences of the Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam* 49, pp. 758–764.

- Butler, Jonathan et al. (2008). “ALLPATHS: De novo assembly of whole-genome shotgun microreads”. In: *Biotechfor* 18.5, pp. 810–820. ISSN: 1088-9051. DOI: [10.1101/gr.7337908](https://doi.org/10.1101/gr.7337908).
- Canetti, G et al. (1969). “Advances in techniques of testing mycobacterial drug sensitivity, and the use of sensitivity tests in tuberculosis control programmes”. In: *Bulletin of the World Health Organization* 41.1, p. 21.
- Chaisson, Mark J and Pavel A Pevzner (2008). “Short read fragment assembly of bacterial genomes”. In: *Genome research* 18.2, pp. 324–330. ISSN: 1088-9051. DOI: [10.1101/gr.7088808](https://doi.org/10.1101/gr.7088808).
- Chambi, Samy et al. (2016). “Better bitmap performance with Roaring bitmaps”. In: *Software: practice and experience* 46.5, pp. 709–719.
- Chikhi, Rayan and Paul Medvedev (2013a). “Informed and automated k-mer size selection for genome assembly”. In: *Bioinformatics* 30.1, pp. 31–37.
- Chikhi, Rayan and Guillaume Rizk (2013b). “Space-efficient and exact de Bruijn graph representation based on a Bloom filter”. In: *Algorithms for Molecular Biology* 8.1, p. 22.
- Chryssanthou, Erja and KRISTIAN ÄNGEBY (2012). “The GenoType® MTB-DRplus assay for detection of drug resistance in Mycobacterium tuberculosis in Sweden”. In: *Apmis* 120.5, pp. 405–409.
- Ciric, Lena et al. (2013). “The Tn916/Tn1545 family of conjugative transposons”. In: *Madame Curie Bioscience Database*.

- Cohen, Ted et al. (2012). “Mixed-strain Mycobacterium tuberculosis infections and the implications for tuberculosis treatment and control”. In: *Clinical microbiology reviews* 25.4, pp. 708–719.
- Coll, Francesc et al. (2015). “Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences”. In: *Genome medicine* 7.1, p. 51.
- Comas, Iñaki et al. (2013). “Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans”. In: *Nature genetics* 45.10, pp. 1176–1182.
- Conway, Thomas C and Andrew J Bromage (2011). “Succinct data structures for assembling large genomes”. In: *Bioinformatics* 27.4, pp. 479–486.
- Davies, Julian and Dorothy Davies (2010). “Origins and evolution of antibiotic resistance”. In: *Microbiology and molecular biology reviews* 74.3, pp. 417–433.
- Davis, James J et al. (2016). “Antimicrobial resistance prediction in PATRIC and RAST”. In: *Scientific reports* 6, p. 27930.
- Day, William HE and Herbert Edelsbrunner (1984). “Efficient algorithms for agglomerative hierarchical clustering methods”. In: *Journal of classification* 1.1, pp. 7–24.
- D’Costa, Vanessa M et al. (2011). “Antibiotic resistance is ancient”. In: *Nature* 477.7365, pp. 457–461.
- Dettman, Jeremy R et al. (2013). “Evolutionary genomics of epidemic and nonepidemic strains of *Pseudomonas aeruginosa*”. In: *Proceedings of the National Academy of Sciences* 110.52, pp. 21065–21070.

- Didelot, Xavier et al. (2012). “Transforming clinical microbiology with bacterial genome sequencing”. In: *Nature reviews. Genetics* 13.9, p. 601.
- Dilthey, Alexander et al. (2015). “Improved genome inference in the MHC using a population reference graph”. In: *Nature genetics* 47.6, pp. 682–688.
- Doughty, Emma L et al. (2014). “Culture-independent detection and characterisation of *Mycobacterium tuberculosis* and *M. africanum* in sputum samples using shotgun metagenomics on a benchtop sequencer”. In: *PeerJ* 2, e585.
- Earle, Sarah G et al. (2016). “Identifying lineage effects when controlling for population structure improves power in bacterial association studies”. In: *Nature microbiology* 1, p. 16041.
- EBI Statistics* (2017). <https://www.ebi.ac.uk/ena/about/statistics>. Accessed: 2017-11-25.
- Eilertson, Brandon et al. (2014). “High proportion of heteroresistance in *gyrA* and *gyrB* in fluoroquinolone-resistant *Mycobacterium tuberculosis* clinical isolates”. In: *Antimicrobial agents and chemotherapy* 58.6, pp. 3270–3275.
- El Feghaly, Rana E et al. (2012). “Presence of the *blaZ* beta-lactamase gene in isolates of *Staphylococcus aureus* that appear penicillin susceptible by conventional phenotypic methods”. In: *Diagnostic microbiology and infectious disease* 74.4, pp. 388–393.
- Eldholm, Vegard et al. (2014). “Evolution of extensively drug-resistant *Mycobacterium tuberculosis* from a susceptible ancestor in a single patient”. In: *Genome biology* 15.11.
- EUCAST* (2017). <http://www.eucast.org>. Accessed: 2017-11-30.

- US-FDA et al. (2007). “Class II special controls guidance document: antimicrobial susceptibility test (AST) systems; guidance for industry and FDA”. In: *Rockville: US Food and Drug Administration*.
- Feuerriegel, Silke et al. (2012). “Sequence analysis for detection of first-line drug resistance in Mycobacterium tuberculosis strains from a high-incidence setting”. In: *BMC microbiology* 12.1, p. 90.
- Feuerriegel, Silke et al. (2015). “PhyResSE: a web tool delineating Mycobacterium tuberculosis antibiotic resistance and lineage from whole-genome sequencing data”. In: *Journal of clinical microbiology* 53.6, pp. 1908–1914.
- Flajolet, Philippe et al. (2007). “Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm”. In: *Analysis of Algorithms 2007 (AofA07)*, pp. 127–146.
- Fricke, W Florian and David A Rasko (2014). “Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions”. In: *Nature reviews. Genetics* 15.1, p. 49.
- ga4gh-Beacon* (2017). <https://github.com/ga4gh/beacon-team>. Accessed: 2017-11-25.
- Gagneux, Sebastien et al. (2006). “The competitive cost of antibiotic resistance in Mycobacterium tuberculosis”. In: *Science* 312.5782, pp. 1944–1946.
- Galimand, Marc et al. (1997). “Multidrug resistance in Yersinia pestis mediated by a transferable plasmid”. In: *New England Journal of Medicine* 337.10, pp. 677–681.

- Galimand, Marc et al. (2006). “Resistance of *Yersinia pestis* to antimicrobial agents”. In: *Antimicrobial agents and chemotherapy* 50.10, pp. 3233–3236.
- Gardy, Jennifer L et al. (2011). “Whole-genome sequencing and social-network analysis of a tuberculosis outbreak”. In: *New England Journal of Medicine* 364.8, pp. 730–739.
- Ge, Fan et al. (2005). “The cobweb of life revealed by genome-scale estimates of horizontal gene transfer”. In: *PLoS biology* 3.10, e316.
- GenomeTrakr (2017). <https://www.fda.gov/Food/FoodScienceResearch/WholeGenomeSequencingProgramWGS/ucm363134.htm>. Accessed: 2017-11-25.
- Goodwin, Bob et al. (2017). “BitFunnel: Revisiting Signatures for Search”. In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 605–614.
- Google–How Search Works–Crawling and Indexing (2017). <https://www.google.com/search/howsearchworks/crawling-indexing/>. Accessed: 2017-11-25.
- Gordon, NC et al. (2014). “Prediction of *Staphylococcus aureus* antimicrobial resistance by whole-genome sequencing”. In: *Journal of clinical microbiology* 52.4, pp. 1182–1191.
- Gosden, PE et al. (1998). “Comparison of the modified Stokes’ method of susceptibility testing with results obtained using MIC methods and British Society of Antimicrobial Chemotherapy breakpoints.” In: *The Journal of antimicrobial chemotherapy* 42.2, pp. 161–169.

- Grad, Yonatan H et al. (2012). “Genomic epidemiology of the Escherichia coli O104:H4 outbreaks in Europe, 2011”. In: *Proceedings of the national academy of sciences* 109.8, pp. 3065–3070.
- Green, Eric D and Mark S Guyer (2011). “Charting a course for genomic medicine from base pairs to bedside”. In: *Nature* 470.7333, p. 204.
- Gu, Danxia et al. (2017). “A fatal outbreak of ST11 carbapenem-resistant hypervirulent *Klebsiella pneumoniae* in a Chinese hospital: a molecular epidemiological study”. In: *The Lancet Infectious Diseases*.
- Guglielmini, Julien et al. (2011). “The Repertoire of ICE in Prokaryotes Underscores the Unity, Diversity, and Ubiquity of Conjugation”. In: *Plos Genet* 7.8, e1002222. ISSN: 1553-7390. DOI: [10.1371/journal.pgen.1002222](https://doi.org/10.1371/journal.pgen.1002222).
- Gupta, Sushim Kumar et al. (2014). “ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes”. In: *Antimicrobial agents and chemotherapy* 58.1, pp. 212–220.
- Han, Kui et al. (2013). “Extraordinary expansion of a *Sorangium cellulosum* genome from an alkaline milieu”. In: *Sci Reports* 3.1, p. 2101. DOI: [10.1038/srep02101](https://doi.org/10.1038/srep02101).
- Harris, Simon R et al. (2010). “Evolution of MRSA during hospital transmission and intercontinental spread”. In: *Science* 327.5964, pp. 469–474.
- Harris, Simon R et al. (2013). “Read and assembly metrics inconsequential for clinical utility of whole-genome sequencing in mapping outbreaks”. In: *Nature biotechnology* 31.7, pp. 592–594.

- Haveri, M et al. (2005). “Comparison of phenotypic and genotypic detection of penicillin G resistance of *Staphylococcus aureus* isolated from bovine intramammary infection”. In: *Veterinary microbiology* 106.1, pp. 97–102.
- Hendrix, Roger W (2003). “Bacteriophage genomics”. In: *Current opinion in microbiology* 6.5, pp. 506–511.
- Hinnebusch, B Joseph et al. (2002). “High-frequency conjugative transfer of antibiotic resistance genes to *Yersinia pestis* in the flea midgut”. In: *Molecular microbiology* 46.2, pp. 349–354.
- Holden, Matthew TG et al. (2013). “A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic”. In: *Genome research* 23.4, pp. 653–664.
- Holley, Guillaume et al. (2016). “Bloom Filter Trie: an alignment-free and reference-free data structure for pan-genome storage”. In: *Algorithms for Molecular Biology* 11.1, p. 3.
- Hu, Yongfei et al. (2016). “Dissemination of the *mcr-1* colistin resistance gene”. In: *The Lancet infectious diseases* 16.2, pp. 146–147.
- Hunt, Martin et al. (2017). “ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads”. In: *bioRxiv*, p. 118000.
- Idury, Ramana M and Michael S Waterman (1995). “A new algorithm for DNA sequence assembly”. In: *Journal of computational biology* 2.2, pp. 291–306.
- Illumina Basespace* (2017). <https://basespace.illumina.com/home/index>. Accessed: 2017-11-25.

- Inouye, Michael et al. (2014). “SRST2: Rapid genomic surveillance for public health and hospital microbiology labs”. In: *Genome Medicine* 6.11, p. 90. ISSN: 1756-994X. DOI: [10.1186/s13073-014-0090-6](https://doi.org/10.1186/s13073-014-0090-6). URL: <https://doi.org/10.1186/s13073-014-0090-6>.
- Iqbal, Zamin et al. (2012). “De novo assembly and genotyping of variants using colored de Bruijn graphs”. In: *Nat Genet* 44.2, pp. 226–232. ISSN: 1061-4036. DOI: [10.1038/ng.1028](https://doi.org/10.1038/ng.1028).
- Iwai, Hiroki et al. (2015). “CASTB (the comprehensive analysis server for the Mycobacterium tuberculosis complex): a publicly accessible web server for epidemiological analyses, drug-resistance prediction and phylogenetic comparison of clinical isolates”. In: *Tuberculosis* 95.6, pp. 843–844.
- Jassal, Mandeep and William R Bishai (2009). “Extensively drug-resistant tuberculosis”. In: *The Lancet infectious diseases* 9.1, pp. 19–30.
- Johnson, Christopher M and Alan D Grossman (2015). “Integrative and conjugative elements (ICEs): what they do and how they work”. In: *Annual review of genetics* 49, pp. 577–601.
- Jolley, Keith A et al. (2012). “Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain”. In: *Microbiology* 158.4, pp. 1005–1015.
- Jun, Hee-Jung et al. (2009). “Nontuberculous mycobacteria isolated during the treatment of pulmonary tuberculosis”. In: *Respiratory medicine* 103.12, pp. 1936–1940.

- Kaase, M et al. (2008). “Comparison of phenotypic methods for penicillinase detection in *Staphylococcus aureus*”. In: *Clinical microbiology and infection* 14.6, pp. 614–616.
- Karlin, Samuel and Stephen F Altschul (1990). “Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes”. In: *Proceedings of the National Academy of Sciences* 87.6, pp. 2264–2268.
- Katz, Lee (2017). *mashtree*. <https://github.com/lskatz/mashtree>.
- Kececioglu, John D and Eugene W Myers (1995). “Combinatorial algorithms for DNA sequence assembly”. In: *Algorithmica* 13.1-2, p. 7.
- Kohl, Thomas A et al. (2014). “Whole-genome-based *Mycobacterium tuberculosis* surveillance: a standardized, portable, and expandable approach”. In: *Journal of clinical microbiology* 52.7, pp. 2479–2486.
- Köser, Claudio U et al. (2012a). “Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak”. In: *New England Journal of Medicine* 366.24, pp. 2267–2275.
- Köser, Claudio U et al. (2012b). “Routine use of microbial whole genome sequencing in diagnostic and public health microbiology”. In: *PLoS pathogens* 8.8, e1002824.
- Köser, Claudio U et al. (2013). “Whole-genome sequencing for rapid susceptibility testing of *M. tuberculosis*”. In: *New England journal of medicine* 369.3, pp. 290–292.
- Koslicki, David and Hooman Zabeti (2017). “Improving Min Hash via the Containment Index with applications to Metagenomic Analysis”. In: *bioRxiv*. DOI: [10.1101/184150](https://doi.org/10.1101/184150). eprint: <https://www.biorxiv.org/content/early/2017/>

BIBLIOGRAPHY

09/04/184150.full.pdf. URL: <https://www.biorxiv.org/content/early/2017/09/04/184150>.

Lapierre, Pascal and Peter J Gogarten (2009). “Estimating the size of the bacterial pan-genome”. In: *Trends Genet* 25.3, pp. 107–110. ISSN: 0168-9525. DOI: [10.1016/j.tig.2008.12.004](https://doi.org/10.1016/j.tig.2008.12.004).

Larsen, Mette V et al. (2012). “Multilocus sequence typing of total genome sequenced bacteria”. In: *Journal of clinical microbiology*, JCM-06094.

Lee, Robyn S and Madhukar Pai (2017). “Real-time sequencing of Mycobacterium tuberculosis: are we there yet?” In: *Journal of Clinical Microbiology* 55.5, pp. 1249–1254.

Leopold, Shana R et al. (2014). “Bacterial whole-genome sequencing revisited: portable, scalable, and standardized analysis for typing and detection of virulence and antibiotic resistance genes”. In: *Journal of clinical microbiology* 52.7, pp. 2365–2370.

Lewis, James S and James H Jorgensen (2005). “Inducible clindamycin resistance in staphylococci: should clinicians and microbiologists be concerned?” In: *Clinical Infectious Diseases* 40.2, pp. 280–285.

Li, Heng and Richard Durbin (2010a). “Fast and accurate long-read alignment with Burrows–Wheeler transform”. In: *Bioinformatics* 26.5, pp. 589–595.

Li, Heng et al. (2009). “The sequence alignment/map format and SAMtools”. In: *Bioinformatics* 25.16, pp. 2078–2079.

Li, Ruiqiang et al. (2010b). “Building the sequence map of the human pan-genome”. In: *Nat Biotech* 28.1, pp. 57–63. URL: <http://dx.doi.org/10.1038/nbt.1596>.

- Lina, Gerard et al. (1999). “Involvement of Panton-Valentine leukocidin—producing *Staphylococcus aureus* in primary skin infections and pneumonia”. In: *Clinical Infectious Diseases* 29.5, pp. 1128–1132.
- Lipworth, Samuel I. W. et al. (2017). “A novel multi SNP based method for the identification of subspecies and associated lineages and sub-lineages of the *Mycobacterium tuberculosis* complex by whole genome sequencing”. In: *bioRxiv*. DOI: [10.1101/213850](https://doi.org/10.1101/213850). URL: <https://www.biorxiv.org/content/early/2017/11/06/213850>.
- Liu, Yi-Yun et al. (2016). “Emergence of plasmid-mediated colistin resistance mechanism *MCR-1* in animals and human beings in China: a microbiological and molecular biological study”. In: *The Lancet infectious diseases* 16.2, pp. 161–168.
- Livermore, David M and John Wain (2013). “Revolutionising bacteriology to improve treatment outcomes and antibiotic stewardship”. In: *Infection & chemotherapy* 45.1, pp. 1–10.
- Loman, Nicholas J and Aaron R Quinlan (2014). “Poretools: a toolkit for analyzing nanopore sequence data”. In: *Bioinformatics* 30.23, pp. 3399–3401.
- Loose, Matthew et al. (2016). “Real-time selective sequencing using nanopore technology”. In: *Nature methods* 13.9, pp. 751–754.
- Lu, Jennifer et al. (2017a). “Bracken: estimating species abundance in metagenomics data”. In: *PeerJ Computer Science* 3, e104.
- Lu, Xin et al. (2017b). “MCR-1.6, a New MCR Variant Carried by an IncP Plasmid in a Colistin-Resistant *Salmonella enterica* Serovar Typhimurium Isolate from a

- Healthy Individual”. In: *Antimicrobial Agents and Chemotherapy* 61.5, e02632–16.
- Lukjancenko, Oksana et al. (2010). “Comparison of 61 sequenced *Escherichia coli* genomes”. In: *Microbial ecology* 60.4, pp. 708–720.
- Lunter, Gerton and Martin Goodson (2011). “Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads”. In: *Genome research* 21.6, pp. 936–939.
- Maciucă, Sorina et al. (2016). “A natural encoding of genetic variation in a Burrows-Wheeler Transform to enable mapping and genome inference”. In: *International Workshop on Algorithms in Bioinformatics*. Springer, pp. 222–233.
- Maiga, Mamoudou et al. (2012). “Failure to recognize nontuberculous mycobacteria leads to misdiagnosis of chronic pulmonary tuberculosis”. In: *PloS one* 7.5, e36902.
- Malshetty, Vidyasagar et al. (2010). “Novel insertion and deletion mutants of RpoB that render *Mycobacterium smegmatis* RNA polymerase resistant to rifampicin-mediated inhibition of transcription”. In: *Microbiology* 156.5, pp. 1565–1573.
- Man, Tom JB de and Brandi M Limbago (2016). “SSTAR, a stand-alone easy-to-use antimicrobial resistance gene predictor”. In: *Msphere* 1.1, e00050–15.
- Marçais, Guillaume and Carl Kingsford (2011). “A fast, lock-free approach for efficient parallel counting of occurrences of k-mers”. In: *Bioinformatics* 27.6, p. 764. DOI: [10.1093/bioinformatics/btr011](https://doi.org/10.1093/bioinformatics/btr011). URL: <http://dx.doi.org/10.1093/bioinformatics/btr011>.

- Mariam, Deneke H et al. (2004). “Effect of rpoB mutations conferring rifampin resistance on fitness of Mycobacterium tuberculosis”. In: *Antimicrobial agents and chemotherapy* 48.4, pp. 1289–1294.
- Marlowe, Elizabeth M et al. (2011). “Evaluation of the Cepheid Xpert MTB/RIF assay for direct detection of Mycobacterium tuberculosis complex in respiratory specimens”. In: *Journal of clinical microbiology* 49.4, pp. 1621–1623.
- Mash Screen* (2017). <https://genomeinformatics.github.io/mash-screen/>. Accessed: 2017-11-25.
- Mason, Amy et al. (submitted). “Accuracy of different bioinformatics methods in detecting antibiotic resistance and virulence factors from Staphylococcus aureus whole genome sequences”. In: *submitted*.
- Matamoros, Sebastien et al. (2017). “Global Phylogenetic Analysis Of Escherichia coli And Plasmids Carrying The mcr-1 Gene Indicates Bacterial Diversity But Plasmid Restriction”. In: *bioRxiv*.
- Mathers, Amy J et al. (2015). “Klebsiella pneumoniae carbapenemase (KPC)-producing K. pneumoniae at a single institution: insights into endemicity from whole-genome sequencing”. In: *Antimicrobial agents and chemotherapy* 59.3, pp. 1656–1663.
- McArthur, Andrew G and Kara K Tsang (2017). “Antimicrobial resistance surveillance in the genomic age”. In: *Annals of the New York Academy of Sciences* 1388.1, pp. 78–91.
- McArthur, Andrew G. et al. (2013). “The Comprehensive Antibiotic Resistance Database”. In: *Antimicrobial Agents and Chemotherapy* 57.7, pp. 3348–3357.

BIBLIOGRAPHY

ISSN: 0066-4804. DOI: [10.1128/aac.00419-13](https://doi.org/10.1128/aac.00419-13). URL: <http://dx.doi.org/10.1128/aac.00419-13>.

- Medini, Duccio et al. (2005). “The microbial pan-genome”. In: *Current opinion in genetics & development* 15.6, pp. 589–594.
- Minot, Samuel S et al. (2015). “One codex: a sensitive and accurate data platform for genomic microbial identification”. In: *bioRxiv*, p. 027607.
- Miotto, Paolo et al. (2012). “GenoType MTBDRsl performance on clinical samples with diverse genetic background”. In: *European respiratory journal* 40.3, pp. 690–698.
- Miotto, Paolo et al. (2014). “Mycobacterium tuberculosis pyrazinamide resistance determinants: a multicenter study”. In: *MBio* 5.5, e01819–14.
- Nathan, Carl and Otto Cars (2014). “Antibiotic resistance—problems, progress, and prospects”. In: *New England Journal of Medicine* 371.19, pp. 1761–1763.
- Navarro, Gonzalo and Eliana Providel (2012). “Fast, Small, Simple Rank/Select on Bitmaps.” In: *SEA* 7276, pp. 295–306.
- Novak, Adam M et al. (2017). “Genome Graphs”. In: *bioRxiv*, p. 101378.
- Ohno, Hideaki et al. (1996). “Relationship between rifampin MICs for and rpoB mutations of Mycobacterium tuberculosis strains isolated in Japan.” In: *Antimicrobial Agents and Chemotherapy* 40.4, pp. 1053–1056.
- O’Leary, Nuala A et al. (2015). “Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation”. In: *Nucleic acids research* 44.D1, pp. D733–D745.

- Olsen, Randall J et al. (2012). “Bacterial genomics in infectious disease and the clinical pathology laboratory”. In: *Archives of pathology & laboratory medicine* 136.11, pp. 1414–1422.
- Olson, Michael A. et al. (1999). “Berkeley DB”. In: *Proceedings of the Annual Conference on USENIX Annual Technical Conference*. ATEC '99. Monterey, California: USENIX Association, pp. 43–43. URL: <http://dl.acm.org/citation.cfm?id=1268708.1268751>.
- Ondov, Brian D. et al. (2016). “Mash: fast genome and metagenome distance estimation using MinHash”. In: *Genome Biology* 17.1, p. 132. ISSN: 1474-760X. DOI: [10.1186/s13059-016-0997-x](https://doi.org/10.1186/s13059-016-0997-x). URL: <http://dx.doi.org/10.1186/s13059-016-0997-x>.
- O’Neill, Jim et al. (2014). “Review on antimicrobial resistance”. In: *Antimicrobial resistance: tackling a crisis for the health and wealth of nations*.
- O’Neill, Jim and Review on Antimicrobial Resistance (2016). *Tackling drug-resistant infections globally: final report and recommendations*. Review on Antimicrobial Resistance.
- Orencia, M Cecilia et al. (2001). “Predicting the emergence of antibiotic resistance by directed evolution and structural analysis”. In: *Nature Structural & Molecular Biology* 8.3, pp. 238–242.
- Ounit, Rachid et al. (2015). “CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers”. In: *BMC genomics* 16.1, p. 236.

- Page, Andrew J et al. (2016). “Multilocus sequence typing by blast from de novo assemblies against PubMLST”. In: *The Journal of Open Source Software* 1.8.
- Pan, Jing-Cao et al. (2008). “Vibrio cholerae O139 multiple-drug resistance mediated by Yersinia pestis pIP1202-like conjugative plasmids”. In: *Antimicrobial agents and chemotherapy* 52.11, pp. 3829–3836.
- Pankhurst, Louise J et al. (2016). “Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study”. In: *The Lancet Respiratory Medicine* 4.1, pp. 49–58.
- Paradis, Emmanuel et al. (2004). “APE: analyses of phylogenetics and evolution in R language”. In: *Bioinformatics* 20.2, pp. 289–290.
- Park, Kwang Seung et al. (2017). “How many mcr-1-harbouring bacteria were spreading geographically?” In: *Biomedical Research*.
- Pell, Jason et al. (2012). “Scaling metagenome sequence assembly with probabilistic de Bruijn graphs”. In: *Proceedings of the National Academy of Sciences* 109.33, pp. 13272–13277.
- Pellow, David et al. (2016). “Improving Bloom filter performance on sequence data using k-mer Bloom filters”. In: *International Conference on Research in Computational Molecular Biology*. Springer, pp. 137–151.
- Pevzner, Pavel A et al. (2001). “An Eulerian path approach to DNA fragment assembly”. In: *Proceedings of the National Academy of Sciences* 98.17, pp. 9748–9753.

- Piddock, Laura JV et al. (2002). “Novel ciprofloxacin-resistant, nalidixic acid-susceptible mutant of *Staphylococcus aureus*”. In: *Antimicrobial agents and chemotherapy* 46.7, pp. 2276–2278.
- Plinke, Claudia et al. (2010). “embCAB sequence variation among ethambutol-resistant *Mycobacterium tuberculosis* isolates without embB 306 mutation”. In: *Journal of antimicrobial chemotherapy* 65.7, pp. 1359–1367.
- Pop, Mihai (2009). “Genome assembly reborn: recent computational challenges”. In: *Briefings in bioinformatics* 10.4, pp. 354–366.
- Pop, Mihai and Steven L Salzberg (2008). “Bioinformatics challenges of new sequencing technology”. In: *Trends in Genetics* 24.3, pp. 142–149.
- Quan, T Phuong et al. (2017). “Evaluation of whole genome sequencing for *Mycobacterium tuberculosis* species identification and drug susceptibility testing in a clinical setting: a large-scale prospective assessment of performance against line-probe assays and phenotyping”. In: *Journal of clinical microbiology*, JCM-01480.
- Quick, Joshua et al. (2015). “Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*”. In: *Genome Biology* 16.1, p. 114.
- Quick, Joshua et al. (2016). “Real-time, portable genome sequencing for Ebola surveillance”. In: *Nature* 530.7589, p. 228.
- Ritter, C et al. (2014). “Evaluation of the AID TB resistance line probe assay for rapid detection of genetic alterations associated with drug resistance in *Mycobacterium tuberculosis* strains”. In: *Journal of clinical microbiology* 52.3, pp. 940–946.

BIBLIOGRAPHY

- Roberts, Leslie (2017). “Echoes of Ebola as plague hits Madagascar”. In: *Science* 358.6362, pp. 430–431. ISSN: 0036-8075. DOI: [10.1126/science.358.6362.430](https://doi.org/10.1126/science.358.6362.430). eprint: <http://science.sciencemag.org/content/358/6362/430.full.pdf>. URL: <http://science.sciencemag.org/content/358/6362/430>.
- Rodwell, Timothy C et al. (2014). “Predicting extensively drug-resistant Mycobacterium tuberculosis phenotypes with genetic mutations”. In: *Journal of clinical microbiology* 52.3, pp. 781–789.
- Sanfilippo, Salvatore (2017). *redis*. <https://github.com/antirez/redis>.
- Schaller, Alain et al. (1999). “Characterization of apxIVA, a new RTX determinant of *Actinobacillus pleuropneumoniae*”. In: *Microbiology* 145.8, pp. 2105–2116.
- Schön, Thomas et al. (2016). “Mycobacterium tuberculosis drug-resistance testing: challenges, recent developments and perspectives”. In: *Clinical Microbiology and Infection*.
- Schürch, Anita C and Roland J Siezen (2010). “Genomic tracing of epidemics and disease outbreaks”. In: *Microbial biotechnology* 3.6, pp. 628–633.
- Seeman, Torsten (2017). *ABRicate*. <https://github.com/tseemann/abricate>.
- Shepherd, Michael A. et al. (1989). “A fixed-size bloom filter for searching textual documents”. In: *The Computer Journal* 32.3, pp. 212–219.
- Simpson, Jared T et al. (2009). “ABySS: a parallel assembler for short read sequence data”. In: *Genome research* 19.6, pp. 1117–1123.
- Smith, Rachel A et al. (2015). “Antibiotic resistance: a primer and call to action”. In: *Health communication* 30.3, pp. 309–314.

- Snitkin, Evan S et al. (2012). “Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing”. In: *Science translational medicine* 4.148, 148ra116–148ra116.
- Solomon, Brad and Carl Kingsford (2016). “Fast search of thousands of short-read sequencing experiments”. In: *Nature biotechnology*.
- Song, Taeksun et al. (2014). “Fitness costs of rifampicin resistance in *Mycobacterium tuberculosis* are amplified under conditions of nutrient starvation and compensated by mutation in the $\beta 2032$ subunit of RNA polymerase”. In: *Molecular microbiology* 91.6, pp. 1106–1119.
- Souli, Maria et al. (2008). “Emergence of extensively drug-resistant and pandrug-resistant Gram-negative bacilli in Europe.” In: *European communicable disease bulletin* 13.47, pp. 5437–5453.
- Spellberg, Brad et al. (2013). “The future of antibiotics and resistance”. In: *New England Journal of Medicine* 368.4, pp. 299–302.
- Steiner, Andreas et al. (2014). “KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes”. In: *BMC genomics* 15.1, p. 881.
- Stoesser, N et al. (2013). “Predicting antimicrobial susceptibilities for *Escherichia coli* and *Klebsiella pneumoniae* isolates using whole genomic sequence data”. In: *Journal of Antimicrobial Chemotherapy* 68.10, pp. 2234–2244.
- Stokes, Hatch W and Michael R Gillings (2011). “Gene flow, mobile genetic elements and the recruitment of antibiotic resistance genes into Gram-negative pathogens”. In: *FEMS microbiology reviews* 35.5, pp. 790–819.

- Stucki, David et al. (2012). “Two new rapid SNP-typing methods for classifying *Mycobacterium tuberculosis* complex into the main phylogenetic lineages”. In: *PLoS One* 7.7, e41253.
- Subramaniam, Sumathi et al. (2000). “Characterization of a predominant immunogenic outer membrane protein of *Riemerella anatipestifer*”. In: *Clinical and diagnostic laboratory immunology* 7.2, pp. 168–174.
- Sun, Gang et al. (2012). “Dynamic population changes in *Mycobacterium tuberculosis* during acquisition and fixation of drug resistance in patients”. In: *The Journal of infectious diseases* 206.11, pp. 1724–1733.
- Suzuki, Satowa et al. (2016). “Investigation of a plasmid genome database for colistin-resistance gene *mcr-1*”. In: *The Lancet infectious diseases* 16.3, pp. 284–285.
- Tange, Ole et al. (2011). “Gnu parallel—the command-line power tool”. In: *The USENIX Magazine* 36.1, pp. 42–47.
- Tarashi, Samira et al. (2017). “Mixed infections in tuberculosis: The missing part in a puzzle”. In: *Tuberculosis*.
- Teuhola, Jukka (1978). “A compression method for clustered bit-vectors”. In: *Information processing letters* 7.6, pp. 308–311.
- Török, ME and SJ Peacock (2012). “Rapid whole-genome sequencing of bacterial pathogens in the clinical microbiology laboratory—pipe dream or reality?” In: *Journal of antimicrobial chemotherapy* 67.10, pp. 2307–2308.
- Touchon, Marie et al. (2009). “Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths”. In: *PLoS genetics* 5.1, e1000344.

- Touchon, Marie et al. (2014). “The chromosomal accommodation and domestication of mobile genetic elements”. In: *Current opinion in microbiology* 22, pp. 22–29.
- Turner, Isaac et al. (2017). “Integrating long-range connectivity information into de Bruijn graphs”. In: *bioRxiv*, p. 147777.
- Uplekar, Mukund et al. (2015). “WHO’s new End TB Strategy”. In: *The Lancet* 385.9979, pp. 1799–1801.
- Van Deun, A et al. (2009). “Mycobacterium tuberculosis strains with highly discordant rifampin susceptibility test results”. In: *Journal of clinical microbiology* 47.11, pp. 3501–3506.
- Vernikos, George et al. (2015). “Ten years of pan-genome analyses”. In: *Current opinion in microbiology* 23, pp. 148–154.
- Villar, María et al. (2011). “Epidemiological and molecular aspects of rifampicin-resistant *Staphylococcus aureus* isolated from wounds, blood and respiratory samples”. In: *Journal of antimicrobial chemotherapy* 66.5, pp. 997–1000.
- VolTRAX (2017). <https://nanoporetech.com/products/voltrax>. Accessed: 2017-12-03.
- Votintseva, Antonina A et al. (2015). “Mycobacterial DNA extraction for whole-genome sequencing from early positive liquid (MGIT) cultures”. In: *Journal of clinical microbiology* 53.4, pp. 1137–1143.
- Votintseva, Antonina A et al. (2017). “Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples”. In: *Journal of clinical microbiology* 55.5, pp. 1285–1298.

BIBLIOGRAPHY

- Walker, Timothy M and the CRyTPIC consortium (in review). “Whole-genome sequencing predicts M. tuberculosis drug susceptibility”. In: *in review*.
- Walker, Timothy M et al. (2015). “Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study”. In: *Lancet Infect Dis* 15.10, pp. 1193–1202. ISSN: 1473-3099. DOI: [10.1016/S1473-3099\(15\)00062-6](https://doi.org/10.1016/S1473-3099(15)00062-6).
- Wang, Ruobing et al. (2017). “The global distribution and spread of the mobilized colistin resistance gene *mcr-1*”. In: *bioRxiv*, p. 220921.
- Welch, Timothy J et al. (2007). “Multiple antimicrobial resistance in plague: an emerging public health risk”. In: *PloS one* 2.3, e309.
- Williamson, DA et al. (2012). “Clinical failures associated with *rpoB* mutations in phenotypically occult multidrug-resistant Mycobacterium tuberculosis”. In: *The International Journal of Tuberculosis and Lung Disease* 16.2, pp. 216–220.
- Wilson, Edwin B (1927). “Probable inference, the law of succession, and statistical inference”. In: *Journal of the American Statistical Association* 22.158, pp. 209–212.
- Wimsatt, Jeffrey and Dean E Biggins (2009). “A review of plague persistence with special emphasis on fleas”. In: *Journal of vector borne diseases* 46.2, p. 85.
- Wirth, Thierry et al. (2006). “Sex and virulence in Escherichia coli: an evolutionary perspective”. In: *Molecular microbiology* 60.5, pp. 1136–1151.
- Wong, Harry KT et al. (1985). “Bit Transposed Files.” In: *VLDB*. Vol. 85, pp. 448–457.

- Wood, Derrick E and Steven L Salzberg (2014). “Kraken: ultrafast metagenomic sequence classification using exact alignments”. In: *Genome biology* 15.3, R46.
- World Health Organization (2012). *The evolving threat of antimicrobial resistance: options for action*. Geneva: World Health Organization.
- (2014). *Antimicrobial resistance: global report on surveillance*. World Health Organization.
- (2015). *Gear up to end TB: introducing the end TB strategy*. World Health Organization.
- (2017). *Global tuberculosis report 2017*. World Health Organization.
- Wright, Gerard D (2003). “Mechanisms of resistance to antibiotics”. In: *Current opinion in chemical biology* 7.5, pp. 563–569.
- Wyres, Kelly L et al. (2014). “WGS analysis and interpretation in clinical and public health microbiology laboratories: what are the requirements and how do existing tools compare?” In: *Pathogens* 3.2, pp. 437–458.
- Xavier, Basil Britto et al. (2016). “Identification of a novel plasmid-mediated colistin-resistance gene, *mcr-2*, in *Escherichia coli*, Belgium, June 2016”. In: *Eurosurveillance* 21.27.
- Yang, Ying et al. (2016). “ARGs-OAP: online analysis pipeline for antibiotic resistance genes detection from metagenomic data using an integrated structured ARG-database”. In: *Bioinformatics* 32.15, pp. 2346–2351.
- Yin, Wenjuan et al. (2017). “Novel plasmid-mediated colistin resistance gene *mcr-3* in *Escherichia coli*”. In: *MBio* 8.3, e00543–17.

- Yong, Dongeun et al. (2009). “Characterization of a new metallo- β -lactamase gene, blaNDM-1, and a novel erythromycin esterase gene carried on a unique genetic structure in *Klebsiella pneumoniae* sequence type 14 from India”. In: *Antimicrobial agents and chemotherapy* 53.12, pp. 5046–5054.
- Zankari, Ea et al. (2012). “Identification of acquired antimicrobial resistance genes”. In: *Journal of antimicrobial chemotherapy* 67.11, pp. 2640–2644.
- Zaunbrecher, M Analise et al. (2009). “Overexpression of the chromosomally encoded aminoglycoside acetyltransferase eis confers kanamycin resistance in *Mycobacterium tuberculosis*”. In: *Proceedings of the National Academy of Sciences* 106.47, pp. 20004–20009.
- Zerbino, Daniel R and Ewan Birney (2008). “Velvet: algorithms for de novo short read assembly using de Bruijn graphs”. In: *Genome research* 18.5, pp. 821–829.
- Zhang, Zheng et al. (2000). “A greedy algorithm for aligning DNA sequences”. In: *Journal of Computational biology* 7.1-2, pp. 203–214.
- Zobel, Justin et al. (1998). “Inverted files versus signature files for text indexing”. In: *ACM Transactions on Database Systems (TODS)* 23.4, pp. 453–490.