Imperial College London

Examining recombination and intra-genomic conflict dynamics in the evolution of anti-microbial resistant bacteria

Department of Infectious Disease Epidemiology

School of Public Health

Imperial College London

Joshua D'Aeth

8th July 2022

Thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy of Imperial College London

and the

Diploma of Imperial College

Abstract

The spread of antimicrobial resistance (AMR) among pathogenic bacterial species threatens to undercut much of the progress made in treating infectious diseases. AMR genes can disseminate between and within populations via horizontal gene transfer (HGT). Selfish mobile genetic elements (MGEs) can encode resistance and spread between host cells. Homologous recombination can alter the core genes of pathogens with resistant donors via HGT too. MGEs may be cured from host genomes through transformation. Hence, MGEs may be able to avoid deletion by disrupting transformation. This work aims to understand how the dynamics of these processes affect the epidemiology of AMR pathogens.

To understand these dynamics, I co-developed a new version of the popular recombination detection tool Gubbins. Through simulation studies, I find this new version to be both accurate in reconstructing the relationships between isolates, and efficient in terms of its use of computational resources.

I then apply Gubbins to both AMR lineages and species-wide datasets of the pathogen *Streptococcus pneumoniae*. I find that recombination frequently occurs around core genes involved in both drug resistance and the host immune response. Additionally, an MGE was able to successfully spread within a population by disrupting the transformation machinery, preventing its loss from the host.

Finally, I investigate two recent examples of MGEs disrupting transformation in the gram-negative species *Acinetobacter baumannii* and *Legionella pneumophila*. I find that while these insertions may decrease the efficiency of transformations within cells, the observed recombination rates largely reflect the selection pressures on isolates. With MGEs only partially able to inhibit these observable transformation events.

These results show how selection pressures from clinical interventions shape pathogen genomes through diverse, often interspecies, recombination events. The spread of MGEs can also be favoured by both these selection pressures, and their ability to disrupt host cell machinery.

Declaration

I hereby certify that the material presented in this thesis, which I now submit for the award of Doctor of Philosophy of Imperial College London, is entirely my own work unless cited, acknowledged or declared otherwise within this thesis.

Joshua Charles D'Aeth

Copyright Statement

Acknowledgements

Firstly I'd like to thank my supervisor Nick Croucher. He has been an unfailingly wise and kind source of guidance through-out these last four and bit years. I am constantly amazed by the breadth and depth of his knowledge, as well his ability to steadfastly not wash up a mug! I'd also like to thank all past and present members of the bacterial evolutionary epidemiology group. You have always offered great advice and have helped to shape this work. In particular I'd like to thank John Lees, whose help with PopPUNK has been invaluable. I'm sure I will still find ways to not fully understand a fit. To the Global Pneumococcal Sequencing team at the Wellcome trust Sanger institute thank-you too, it has been amazing to get the opportunity to work on this great dataset. The Wellcome trust have also generously funded this work, and have been very understanding during the disruptions from COVID-19. The administration team within DIDE, particularly Julie, have been invaluable during this period as well.

Thanks too, to all the PhD students in my cohort, you have always made lunchtime something to look forward to! Hannah, Ben, Janetta and Daniela, I can't think of better people to have experienced the, at times exhausting, ride of being a PhD student with. EECID massive for life. My housemates over the years too, Chris, Jon and Allan, have been invaluable sources of support. To Allan, a thank-you, at this time of year, at this time of day, localized entirely within this section.

Finally I'd also like to thank my family, none of this would've been possible without them. To my sister Alex for her love and kindness and my brother Noah for indulging my interest in basketball (among other things), thank-you. To my parents Emma and Jack, thank-you for keeping me going, this is dedicated to you!

Contents

1 Introduction **20**

1.1 The history of antimicrobial resistance 21

1.1.1 The discovery of antimicrobials 21

1.1.1.1 The first antimicrobials 21

1.1.1.2 The Golden Age of antibiotic discovery 22

1.1.2 Mechanisms of antibiotic action 24

1.1.2.1 Inhibitors of nucleic acid synthesis 24

1.1.2.2 Inhibitors of protein synthesis 25

1.1.2.3 Inhibitors of cell wall synthesis 26

1.1.3 Mechanisms of AMR 27

1.1.3.1 Modification of antibiotic target 28

1.1.3.2 Minimization of antibiotic concentration 29

1.1.3.3 Direct inactivation of an antibiotic 30

1.1.4 The evolution of AMR 31

1.1.4.1 The resistome 31

1.1.4.2 Modelling the spread of resistance among bacterial populations 32

1.1.4.3 Modelling the spread of MDR 35

1.1.5 The impact of AMR on public health 37

1.1.5.1 The effect of AMR on patient outcomes 37

1.1.5.2 Predictions of the burden of AMR 39

1.2 Genomic data analysis 40

1.2.1 Sequencing bacterial genomes . . . 40
1.2.1.1 Sequencing methods . . . 40
1.2.2 Predicting AMR from genomes . . . 42
1.2.2.1 Direct association methods . . . 43
1.2.2.2 Predictive modelling methods . . . 44
1.2.3 Genomic epidemiology . . . 46
1.2.3.1 Inferring the relationship between isolates . . . 47
1.2.3.2 Phylodynamics . . . 50
1.2.3.3 Insights into bacterial population structure from sequencing 51
1.3 Bacterial horizontal gene transfer . . . 53
1.3.1 Phage and Transduction . . . 53
1.3.1.1 Mechanisms of Transduction . . . 53
1.3.1.2 DNA transferred by transduction . . . 55
1.3.1.3 Phage-host conflict . . . 56
1.3.2 Conjugation and conjugative elements . . . 57
1.3.2.1 Mechanism of Conjugation . . . 57
1.3.2.2 Plasmids . . . 58
1.3.2.3 Integrative conjugative elements . . . 59
1.3.3 Transformation and homologous recombination . . . 61
1.3.3.1 Competence regulation and DNA uptake . . . 61
1.3.3.2 Incorporation of DNA via homologous recombination . . . 63
1.3.3.3 The evolution of transformation . . . 63
1.4 Summary . . . 65

2 Extending methods to detect recombination in bacterial genomes 67

2.1 Introduction . . . 68
2.1.1 Methods of detecting recombination: . . . 69
2.1.1.1 Detecting exchanges . . . 69
2.1.1.2 Detecting imports . . . 70
2.1.1.3 Gubbins algorithm . . . 72
2.1.2 Extensions to the Gubbins algorithm . . . 76

2.2 Methods . . . 77
2.2.1 Simulating artificial sequences . . . 77
2.2.2 Assessing differences in phylogenies . . . 79
2.3 Results . . . 80
2.3.1 Assessing the phylogenies produced from Gubbins . . . 80
2.3.2 Assessing the choice of substitution model on phylogenies produced by Gubbins . . . 85
2.3.3 The accuracy of recombination statistic estimation . . . 88
2.3.3.1 Accuracy by phylogeny builder . . . 89
2.3.3.2 Accuracy by reconstruction . . . 91
2.3.3.3 Assessing the accuracy of SNP classification . . . 92
2.3.4 Time and memory usage of Gubbins models . . . 96
2.3.5 Conclusion . . . 99

3 The emergence of resistance among core genes of the pneumococcus 102

3.1 Introduction . . . 103
3.1.1 Core gene resistance and recombination . . . 103
3.1.2 PMEN3 population . . . 104
3.1.3 PMEN9 collection . . . 104
3.1.4 GPS collection . . . 105
3.2 Methods . . . 105
3.2.1 Isolate collection and sequencing . . . 105
3.2.2 Generation of annotations and alignments . . . 106
3.2.3 Phylogenetic and phylodynamic analyses . . . 106
3.2.4 Antibiotic resistance analyses . . . 108
3.2.4.1 Penicillin resistance . . . 108
3.2.4.2 Co-trimoxazole resistance . . . 108
3.2.5 Detecting interspecies recombination events . . . 109
3.2.5.1 Pipeline for detecting interspecies events at the *pbp* loci . 109
3.2.5.2 Detecting interspecies recombination at *murM* . . . 110
3.3 Results . . . 110

3.3.1 Genomic epidemiology of the PMEN3 and PMEN9 lineages 110
3.3.2 Variable recombination dynamics across the PMEN lineages 116
3.3.3 Hotspots of recombination 119
3.3.4 Investigating β-lactam resistance emergence in the pneumococcus 120
3.3.4.1 Determining β-lactam resistance levels 120
3.3.4.2 Emergence of β-lactam resistance in PMEN3 and PMEN9 123
3.3.4.3 Role of interspecies transformation in β-lactam resistance in PMEN3 and PMEN9 127
3.3.4.4 The levels and origin of β-lactam resistance in the GPS collection 127
3.3.5 Evolution of resistance through recombination at other core loci . . . 129
3.3.5.1 Penicillin resistance and *murM* 129
3.3.5.2 Co-trimoxazole resistance in PMEN3 and PMEN9 129
3.3.5.3 Co-trimoxazole resistance in the GPS collection 131
3.4 Conclusions 131

4 The spread of resistance through mobile genetic elements in pneumococcal populations 134

4.1 Introduction 135
4.1.1 MGEs, recombination and resistance 135
4.1.2 The Tn*916* element 136
4.1.3 The Tn*1207.1* element 136
4.2 Methods 139
4.2.1 MGE identification 139
4.2.2 Antibiotic consumption data 139
4.2.3 Phylodynamic analysis 140
4.2.4 MGE insertion site identification 140
4.3 Results 142
4.3.1 MGE distribution among PMEN3 and PMEN9 142
4.3.2 Expansion of macrolide resistance in German pneumococci 146

4.3.3 Multiple acquisitions of the Tn*1207.1* and Tn*916* elements across the GPS collection . . . 151
4.3.3.1 Defining unique hits . . . 151
4.3.3.2 MGE distribution across the GPS collection . . . 155
4.3.3.3 Frequent insertion of MGEs via recombination . . . 160
4.3.4 The interspecies origin of MGE importation events . . . 161
4.4 Conclusions . . . 166

5 Investigating host MGE conflict in gram-negative bacterial species 175
5.1 Introduction . . . 175
5.1.1 *Acinetobacter baumannii* and AbaRs . . . 176
5.1.2 *Legionella pneumophila* and pLPL . . . 178
5.2 Methods . . . 179
5.2.1 Isolate collections . . . 179
5.2.2 Population structure and quality control of assemblies . . . 180
5.2.3 Detecting disrupted competence machinery . . . 181
5.2.4 Detecting recombination dynamics . . . 182
5.3 Results . . . 182
5.3.1 Population structure of *Acinetobacter baumannii* . . . 182
5.3.2 Population struction of the *Legionella pnuemophila* collection . . . 186
5.3.3 The distribution of MGEs across the collections . . . 191
5.3.4 Recombination dynamics in the GC2 strain of *Acinetobacter baumannii* . . . 192
5.3.5 The effect of *comM* disruption on recombination dynamics in *Acinetobacter baumannii* . . . 195
5.3.6 Recombination dynamics in *L. pneumophila* strain 1. . . . 199
5.3.7 RocRp disrupts recombination dynamics in *L. pneumophila* . . . 202
5.4 Conclusions . . . 205

6 Discussion 208
6.1 Summary of results . . . 208
6.2 Implications of research . . . 210

6.2.1 Selection and recombination 210
6.2.2 Adaptive evolution to public health interventions 210
6.2.3 Comparisons of recombination properties across species 212
6.2.4 Conflict between hosts and MGEs 213
6.2.5 Impact of vaccines on resistance 215
6.3 Limitations and Future work 215
6.3.1 Detecting MGEs 215
6.3.2 Assembly consistency 216
6.3.3 Gubbins improvements 217
6.3.4 Genomic epidemiology 218
6.4 Conclusion 218

Bibliography **220**

List of Figures

1.1 **Timeline of the decades new classes of antibiotics reached the clinic.** 23

1.2 **Relationship between β-lactam consumption and PNSP invasive isolates across Europe in 2020.** 33

1.3 **Frequency distribution of genes within GC2 lineage of *A. baumannii*.** . 52

1.4 **Summary of the different HGT methods, split by DNA uptake and integration steps.** 54

2.1 **Kendall-Colijn distance with Gubbins models grouped by first iteration phylogeny builder** 83

2.2 **Kendall-Colijn distance with Gubbins models grouped by main iteration phylogeny builder** 84

2.3 **Kendall-Colijn distance with Gubbins models grouped by ancestral state reconstruction method** 85

2.4 **Kendall-Colijn distance with Gubbins models grouped by total model** 86

2.5 **Comparison of the Phylogenies produced from a GTR and JC substitution model and the true phylogeny.** 88

2.6 **Recombination statistics from simulated and reconstructed data, with Gubbins models grouped by first iteration phylogeny builder.** 90

2.7 **Recombination statistics from simulated and reconstructed data, with Gubbins models grouped by main iteration phylogeny builder.** 92

2.8 **Recombination statistics from simulated and reconstructed data, with Gubbins models grouped by ancestral state reconstruction method.** . 93

2.9 **Recombination statistics from simulated and reconstructed data, with Gubbins models grouped by full model used.** 94
2.10 **Correlation plots of the value of ρ against *r* for different models.** . . . 95
2.11 **Sensitivity and PPV scores for Gubbins reconstructions.** 95
2.12 **The difference between the excess number of *m* SNPs reconstructed by Gubbins models and the number of *r* SNPs present in recombinations with less than three SNPs.** 97
2.13 **The analysis times and memory usages of Gubbins models** 98

3.1 **Phylogenomic analysis of the PMEN3 lineage.** 112
3.2 **Root to tip analysis of PMEN3 lineage.** 113
3.3 **Serotype switching events across the PMEN3 lineage.** 114
3.4 **Phylogenomic analysis of PMEN9 lineage.** 115
3.5 **Root to tip analysis of the PMEN9 lineage.** 116
3.6 **Serotype switching events across the PMEN9 lineage.** 117
3.7 **Summary of recombination differences between the PMEN3 and PMEN9 lineages** 118
3.8 **Penicillin resistance prediction using different models.** 122
3.9 **PMEN3 resistant lineages through time.** 124
3.10 **Histograms of recorded MIC values for penicillin across the PMEN3 and PMEN9 lineages.** 126
3.11 **Origin of *pbp* genes for penicillin resistant isolates.** 128
3.12 **Analysis of the origin of the *murM* gene across the PMEN3 and PMEN9 Lineages.** 130

4.1 **Linear and circular representation of the Tn*916* element.** 137
4.2 **Linear representation of the Tn*1207.1* element.** 138
4.3 **Outline of insertion point pipeline.** 143
4.4 **Phylogenomic analysis of PMEN3 lineage.** 144
4.5 **Phylogenomic analysis of PMEN9 lineage.** 145
4.6 **Consumption of macrolide and β-lactam antibiotics across Europe.** . . 147
4.7 **Ratio of macrolide to β-lactam consumption in Europe.** 148

4.8 **Root to tip analysis of 162 German isolates within PMEN9.** 149
4.9 **Expansion of a macrolide resistant clade in Germany pre-vaccine.** . . 150
4.10 **Skygrowth analysis incorporating macrolide consumption data** 152
4.11 **Skygrowth analysis incorporating β-lactam consumption data** 153
4.12 **Insert of Tn*1207.1* within the PMEN9 reference genome.** 154
4.13 **Insertion points of classified Tn*1207.1* hits within *S. pneumoniae.*** . . . 157
4.14 **Insert of Tn*1207.1* as Mega within *tag*.** 158
4.15 **Insertion points of classified Tn*916* hits within *S. pneumoniae.*** 159
4.16 **Comparison of length and SNP density of recombination events.** . . . 162
4.17 **The likely origin of MGE insertions.** 164
4.18 **Flanking region origin for Tn*1207.1 tag* insertions.** 165
4.19 **Insert of Tn*916* downstream of *recJ*** 169
4.20 **Insert of Tn*916* downstream of *gmuF*** 170
4.21 **Insert of Tn*916* upstream of *gidB*** 171
4.22 **Insert of Tn*916* upstream of *rplL* (1)** 172
4.23 **Insert of Tn*916* upstream of *rplL* (2)** 173
4.24 **Insert of Tn*916* upstream of *rplL* (3)** 174

5.1 **Distance calculations for *Acinetobacter baumannii* and *Legionella pneumophila* populations.** 184
5.2 **Population structure of *Acinetobacter baumannii* collection.** 185
5.3 **Distribution of the within strain and between strain distances for isolates within the *A. baumannii* GC2 clade.** 186
5.4 **Core distance phylogeny of the *L. pneumophila* collection** 188
5.5 **Distribution of the between strain pairwise distances based on *L. pneumophila* subtype.** 189
5.6 **Population structure of *Legionella pneumophila* collection.** 190
5.7 **Within and between-strain distances for *L. pneumophila* strain 2 and strain 2-like strains** 191
5.8 **Recombination dynamics in the *A. baumannii* GC2 lineage.** 194

5.9 **Distributions of the lengths of recombination events within *A. baumannii*** . . . 197

5.10 **Distributions of the SNP density of recombination events within *A. baumannii*** . . . 198

5.11 **Distributions of the differences between *A. baumannii* WT*comM* isolates and disrupted *comM* isolates *r*/*m* values.** . . . 199

5.12 **Distributions of the differences between *A. baumannii* WT*comM* isolates and disrupted *comM* isolates ρ/*m* values.** . . . 200

5.13 **Recombination dynamics in strain 1 of *L. pneumophila*.** . . . 201

5.14 **Distribution of the lengths of recombination events, split by RocRp possession across the three largest *L. pneumophila* strains.** . . . 203

5.15 **Distributions of the SNP density of recombination events, split by RocRp presence across the three largest *L. pneumophila* strains.** . . . 204

5.16 **Distributions of the differences between *L. pneumophila* WT and RocRp present isolates' *r*/*m* values.** . . . 205

5.17 **Distributions of the differences between WT and RocRp present isolates' ρ/*m* values.** . . . 205

6.1 **Outline of the main steps in moving from assemblies to strain level phylogenies, for a collection of sequences.** . . . 219

List of Tables

1.1 **Grouping antibiotic classes by pathway targeted.** 25

2.1 **Species counts of the 52 *Streptococcus* genomes used as recombination donors** 78

2.2 **Models benchmarked for Gubbins v3.2.0.** 81

2.3 **Kendal-Colijn metric distances between the true tree and each tree formed from the model listed** 87

2.4 **Runtimes and memory usage for multiple cores** 99

3.1 **Isolate sources for PMEN3 and PMEN9** 106

4.1 **Closest species match to 500 bp region upstream of *tag* disrupting Tn*1207.1* insertion.** 166

5.1 **Assembly levels of the collections for *A. baumannii* and *L. pneumophila* from GenBank.** 180

5.2 **Numbers of *A. baumannii* isolates split by *comM* status** 196

5.3 **Distribution of the transformation inhibiting sRNA RocRp across the three largest *L. pneumophila* strains.** 202

List of Acronyms

3GREC Third generation cephalosporin-resistant *E. coli*

A. baumannii *Acinetobacter baumannii*

AMR Antimicrobial resistance

CDC Centers for Disease Control and Prevention

CFRs Case fatality ratios

CSP Competence stimulating peptide

DHPS dihydropteroate synthase

dsDNA double stranded DNA

E. coli *Escherichia coli*

EARS-Net European antimicrobial resistance surveillance network

ECDC European Centres for Disease Control and prevention

EF-G Elongation factor G

EM Expectation-Maximisation

ENA EMBL Nucleotide Sequence Database

ESBL Extended spectrum β-lactamase

EUCAST European committee on antimicrobial susceptibility testing

GBD Global burden of disease

GC1 Global Clone 1

GC2 Global Clone 2

GPS Global pneumococcal sequencing project

GPSC Global pneumococcal sequencing cluster

GTA Gene transfer agent

GTR General time-reversible

H. influenzae *Haemophilus influenzae*

HGT Horizontal gene transfer

HMM Hidden Markov model

ICE Integrative conjugative element

IHME Institute of Health Metrics and Evaluation

INSDC International Nucleotide Sequence Database Collaboration

IPD Invasive pneumococcal disease

IPP C_{55}-isoprenyl pyrophosphate

IQR Inter-quartile range

JC Jukes-Cantor

K. pneumoniae *Klebsiella pneumoniae*

L. pneumophila *Legionella pneumophila*

LD Linkage disequillibrium

LeD Legionnaire's disease

LRS Long-read sequencing

M. tuberculosis *Mycobacterium tuberculosis*

MCMC Markov chain Monte Carlo

MDR Multi-drug resistant

MGE Mobile genetic element

MIC Minimum inhibitory concentration

MLEE Multilocus Enzyme Electrophoresis

MLST Multi-locus sequence typing

MP Maximum parsimony

MPF Mating pair formation

MRSA Methicillin-resistant *Staphylococcus aureus*

MSSA Methicillin-sensitive *Staphylococcus aureus*

N. gonorrhoeae *Neisseria gonorrhoeae*

N. meningitidis *Neisseria meningitidis*

nBSIs non-Blood stream infections

NGS Next-generation sequencing

NPET Nascent peptide exit tunnel

ODE Ordinary differential equation

ONT Oxford Nanopore technology

ORF Open reading frame

oriT Origin of transfer

P. aeruginosa *Pseudomonas aeruginosa*

PABA ρ-aminobenzoic acid

PBP Penicillin binding protein

PDR Pan drug resistant

PFGE Pulsed-field gel electrophoresis

PI Pathogenicity island

PNSP Penicillin non-susceptible pneumococci

PPV Positive predictive value

PSA Proportion of shared ancestry

QC Quality control

R-M Restriction modification system

RBP Receptor binding proteins

RF Random forest

RND Resistant nodulation division

S. aureus *Staphylococcus aureus*

S. mitis *Streptococcus mitis*

S. oralis *Streptococcus oralis*

S. pneumoniae *Streptococcus pneumoniae*

S. pseudopneumoniae *Streptococcus pseudopneumoniae*

SBT Sequence based typing

SMRT Single-molecule real-time

SNP Single nucelotide polmorphism

sRNA Small RNA

ssDNA Single-stranded DNA

ST Sequence type

T4CP Type IV coupling protein

T4SS Type IV secretion system

TB Tuberculosis

TPD Trans-peptidase domains

V. cholerae *Vibrio cholerae*

WGS Whole genome sequence

WHO World Health Organization

XDR Extremely drug resistant

TDR Totally Drug resistant

Chapter 1

Introduction

There is a very neat story often told about the discovery of antibiotics. It goes, broadly, that one summer, during his childhood, Winston Churchill found himself in difficulty in the family swimming pool and was on the verge of drowning when a young Scottish pool boy dived in and saved him. The Churchill family, in a sign of their immense gratitude to the pool boy, offered to subsequently pay for his medical school education. Some years later, when Churchill was prime minister during the Second World War, this pool boy saved him yet again. This time, however, through his discovery of the antibiotic penicillin, used to treat Churchill's pneumonia.

This story, painting Alexander Fleming as a guardian angel to Churchill, is a myth [1]. Fleming himself was seven years younger than Churchill, while Churchill was actually treated with sulphonamides during the war. Nevertheless, this myth is not far off conveying the immense impact the discovery of antibiotics would have. From their discovery in the 20th century up to today, they have saved countless lives and enabled a great raft of innovation in the treatment of patients. Their use though, has spawned a looming public health crisis with the rise of antimicrobial resistance (AMR). In this introduction I will detail the history of AMR, look at how bacteria can disseminate resistance genes among their populations, and look into how genomic sequencing has enabled us to track its spread. I will start with a look at the history of the earliest antimicrobials.

1.1 The history of antimicrobial resistance

1.1.1 The discovery of antimicrobials

1.1.1.1 The first antimicrobials

An antimicrobial is a medical agent used to kill or stop the growth of microorganisms, including bacteria, viruses and fungi. Antibiotics are themselves a subclass of antimicrobials that specifically target bacterial infections within humans. The first effective antimicrobial used in clinical practice in the UK were the sulphonamides, introduced by Leonard Colebrook in 1937, after work by Gerhard Domagk demonstrated their selective antibacterial nature [2, 3]. While their mode of action was not fully understood at the time of their introduction [4], they are now known to act via inhibition of the dihydropteroate synthase (DHPS) enzyme [5]. This enzyme is essential to the folate biosynthesis pathway of many prokaryotes [6]. Sulphonamides competitively inhibit DHPS via their structural analogy to the substrate ρ-aminobenzoic acid (PABA) [5,7]. Crucially mammalian cells lack the DHPS folate biosynthesis pathway, instead taking up preformed folate from dietary sources, giving sulphonamides a broad spectrum of activity against bacterial pathogens [8].

Sulphonamides proved to be inexpensive, easily produced in bulk and effective treatments, playing a key role in reducing deaths from wounds in the Second World War [9]. They were used as prophylactics to prevent upper respiratory infections [10], and, as mentioned above, in the treatment of Winston Churchill for bacterial pneumonia in 1943 [9].

Resistance to sulphonamides, though, was quick to develop after their widespread use. Streptococci were among the first bacteria to exhibit resistance. For instance, resistant isolates were detected within six months of a program of widespread chemoprophylaxis in naval training centres in New York in 1944 [11]. Resistant isolates of *Neisseria gonorrhoeae* and *Neisseria meningitidis* were also reported in the 1940s following increased clinical treatment and prophylaxis of gonorrhoea and meningitis infections with sulphonamides [12, 13]. There was early debate surrounding whether these resistant isolates were arising *de novo* or pre-existing resistant isolates were being selected for [10]. Strains of *N. gonorrhoeae* and *N. meningitidis* isolated before the introduction of

sulphonamides had proven to be resistant for instance [10, 12].

A mechanism of sulphonamide resistance was first described in *Streptococcus pneumoniae* isolates. Here extracts of resistant isolates were seen to increase their production of a sulphonamide inhibitor [10], later confirmed to be PABA [14]. This increase in PABA production enables the substrate to outcompete the sulphonamide inhibitor of DHPS and allows the folate biosynthesis pathway of the bacterium to continue. In resistant *Escherichia coli* isolates however another route to resistance was identified. Here, the DHPS enzyme itself was found to be modified, this reduced the enzyme's affinity to sulphonamides to a large extent, allowing PABA to outcompete the inhibitor [15, 16]. This alteration also, however, made DHPS less heat stable and less efficient than the wild type enzyme, illustrating how initial resistance mutations are often accompanied by a loss in fitness. Today the mechanism of sulphonamide resistance is mainly that of DHPS alteration, as opposed to PABA overproduction [17, 18].

Sulphonamides were quickly replaced by the new "Magic Bullet" penicillin [2] in the first line treatment of many bacterial infections from the 1940s onward. Penicillin was the first antibiotic capable of killing gram-positive bacteria, including the causative agents of gonorrhoea and syphilis [19]. Additionally, penicillin was found to be free of tissue toxicity, had no known antagonists and could be five times more potent than certain sulphonamides for the same causative agent [2]. However, resistance to penicillin was observed very early on in its development as an effective clinical drug. Abraham & Chain discovered a penicillinase produced by *E. coli* in 1940 [20, 21], before it was clinically introduced in 1941 [19]. While in 1942, hospitalized patients were found to be infected with penicillin-resistant *Staphylococcus aureus* [21, 22]. Bacteria were quick to respond to these clinical interventions.

1.1.1.2 The Golden Age of antibiotic discovery

The discovery and adaptation for clinical use of penicillin is thought to have initiated the "Golden Age" of antibiotic discovery [23], which spanned from the 1940s to late 1960s (Figure 1.1). Fleming, in the immediate aftermath of his discovery of penicillin, pioneered techniques to widely sample soil biomes, searching for biocidal compounds [19]. Similarly, Selman Waksman, and his lab [24], began to systematically study soil microbes in order

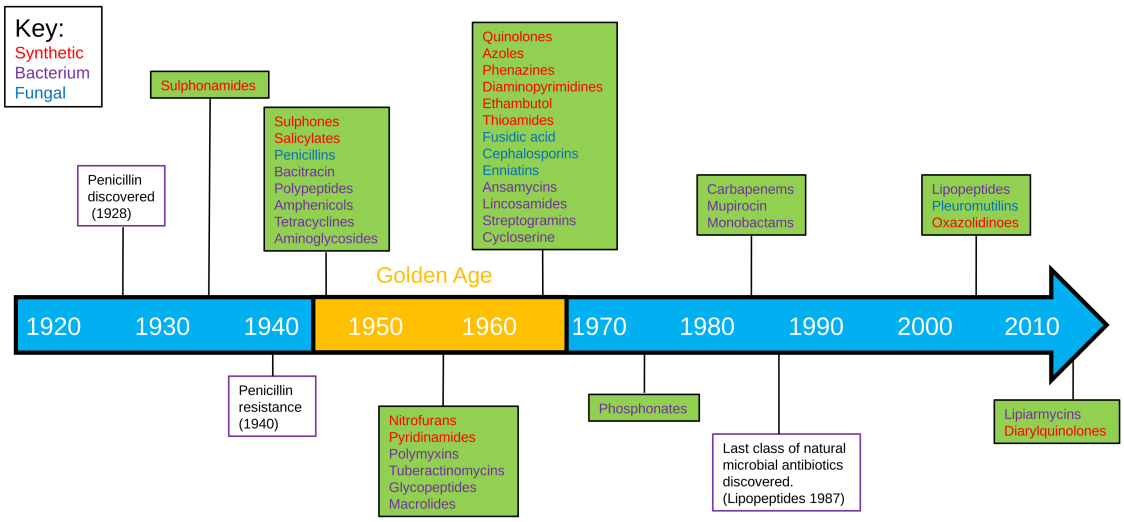

Figure 1.1: Timeline of the decades new classes of antibiotics reached the clinic. Individual dates of important drug discovery and resistance are highlighted. Discovered classes are in green boxes, with the text colour highlighting the source of the novel class. Figure is adapted from Hutchings, Truman & Wilkinson 2019 [25]

to identify possible antibiotics [25]. Within the incredibly diverse bacterial actinomycetes order [26], the Waksman group identified the genus *Streptomyces* as producers of a range of secondary metabolites with antibiotic activity [25]. Streptomycin, an aminoglycoside, was isolated from *Streptomyces griseus* in 1944 and became the first antibiotic cure for tuberculosis (TB) [23, 27, 28]. Beside the aminoglycosides, *Streptomyces* species have also yielded chloramphenicol, from *Streptomyces venezuelae*, tetracycline, from *Streptomyces rimosus*, and the first macrolide, erythromycin, from *Streptomyces erythraeus*, although this species has now been reclassified as *Saccharopolyspora erythraea* [29, 30]. The tetracycline terramycin was the first antibiotic to be described as broad-spectrum, effective against a range of gram-positive and gram-negative pathogens [31]. This epithet was not initially a medical term, but instead was coined by Arthur Sackler, patriarch of the Sackler family, in his role as advertiser for the drug [31].

In total, of all the antibiotics discovered during the Golden Age, two out of three came from the actinomycetes [29], with *Streptomyces* alone accounting for 55% of all known antibiotics [25, 32]. With this rate of discovery the mood at the time in the medical community was largely triumphal, with Frank Macfarland Burnet, a Nobel prize winner in 1960 for his work on immunological tolerance, declaring that we had seen the ”virtual elimination of the infectious diseases as a significant factor in social life” in 1962 [33]. However, it has

been suggested that the rapid discovery of new antibiotics during this period, led to less than prudent usage which has contributed to the rise of antibiotic resistance [25]. Indeed, resistance to streptomycin was first observed soon after its introduction as a treatment for TB, with George Orwell for instance infected with a resistant strain of TB while writing 1984, in 1948 [18]. Penicillin-resistant *S. aureus* had also become a pandemic by the late 1950s and early 1960s, with the clone Φ80/81 driving this spread [34]. While the introduction of methicillin in 1961 led to this penicillin resistant clone disappearing from the population [35], almost immediately methicillin resistant clones emerged [34, 36, 37].

The growing burden of resistance is now more worrisome given the faltering supply line of new antibiotics from the 1970s onwards (Figure 1.1). Natural produce antibiotics are those derived from secondary metabolites. The most recent class of these, the lipopeptides, was discovered 35 years ago in 1987 [25]. The Golden Age of discovery is thought to have harvested all the low-hanging fruit of antibiotics, with pharmaceutical companies now also diverting funding away from discovery departments [38–40]. However, sampling species in more diverse environments [41, 42], as well as recent efforts to activate cryptic secondary metabolite production not normally produced *in vitro* [43], offers some hope for the discovery of novel classes of antibiotics [25].

1.1.2 Mechanisms of antibiotic action

Building on from the Golden Age of antibiotic discovery, there are 38 different classes of antibiotic in clinical use today [25]. While these differ in their precise mode of action and molecular targets, the majority of these can be broadly split into disrupting three main essential prokaryotic cellular pathways: nucleic acid synthesis, protein synthesis and cell well synthesis [44] (Table 1.1).

1.1.2.1 Inhibitors of nucleic acid synthesis

In total 10 different classes of antibiotics target pathways involved in nucleic acid synthesis (Table 1.1). The sulphonamides described above, along with the salicylates, sulphones and diaminopyrimidines (which include trimethoprim), all target the folate biosynthesis pathway [45]. This pathway is essential to production of nucleic acids [46]. All these folate targeting compounds are synthetically derived. Other synthetic antibiotics target

Pathway targeted	Antibiotic class	Number of Classes
Nucleic acid synthesis	Ansamycins, Azoles, Diaminopyrimidines, Fluoroquinolones, Lipiarmycins, Nitrofurans, Phenazines, Salicylates, Sulphonamides, Sulphones	10
Protein Synthesis	Aminoglycosides, Amphenicols, Fusidic acid, Lincosamides, Macrolides, Mupirocin, Oxazolidinones, Pleuromutilins, Streptogramins, Tuberactinomycins, Tetracyclines	11
Cell Wall synthesis	Bacitracin, Carbapenems, Cephalosporins, Cycloserines, Enniatins, Ethambutol, Glycopeptides, Lipopeptides, Monobactams, Penicillins, Phosphonates, Polymyxins, Polypeptides, Pyridinamides, Thioamides	15
ATP synthesis	Diarylquinolines	1
Not known	Arsphenamines	1

Table 1.1: Grouping antibiotic classes by pathway targeted. Data are adapted from Hutchings, Truman & Wilkinson [25]

DNA directly. The azole metronidazole causes DNA damage when treating infections from the *Giardia* genus [47], while the Nitrofurans also cause DNA damage [48]. The synthetic fluoroquinolones target the prokaryotic DNA gyrase enzyme, which is essential for maintaining DNA supercoiling within the bacterial cell [49]. Finally the phenazines, such as clofazimine, target and bind to guanine bases preventing further DNA replication [50].

Other antibiotics derived from bacterial products target the synthesis of RNA. The ansamycins, which include rifamycin, target the RNA polymerase enzyme, binding and preventing further RNA synthesis [51]. This gives rifamycin biocidal activity against a range of gram-positive and gram-negative pathogens [51]. Similarly, lipiarmycins target the RNA polymerase enzyme, although these bind to a different region, the switch region, giving no cross-resistance with ansamycins [52].

1.1.2.2 Inhibitors of protein synthesis

Antibiotic classes targeting protein synthesis within pathogens are derived from bacterial, fungal, and synthetic products (Table 1.1). Broadly these antibiotics either target the 30S ribosomal subunit or the 50S ribosomal subunit involved in protein translation [44]. The 30S subunit is involved in the binding of tRNAs to matching mRNA codons, and

the translocation of bound mRNA and tRNAs [53]. It is composed of 16S rRNA and 19 proteins [54], and is targeted by aminoglycosides, such as streptomycin, tetracyclines and tuberactinomycins [25]. Typically these compounds bind to the 16S rRNA component, preventing the binding of tRNAs and arresting protein synthesis [55].

The 50S ribosomal subunit acts as a peptidyl transferase to catalyze the binding of amino acids together [56]. The subunit is made up of 5S rRNA, 23S rRNA and 31 proteins [57]. The amphenicol class (which includes chloramphenicol), the lincosamides, the streptogramins, the pleuromutillins and the synthetic oxazolidinones all target overlapping regions of the 23S rRNA of the 50S subunit, inhibiting its peptidyl transferase activity [58–61]. The macrolide class of antibiotics target the nascent peptide exit tunnel (NPET) of the 50S subunit, arresting further translation of mRNA [62].

Outside of directly targeting the ribosome, fusidic acid binds to elongation factor G (EF-G) [63]. EF-G is a GTPase that catalyzes the complete translocation of mRNA-tRNA when peptide chains are elongated [63]. Similarly, mupirocin, also known as pseudomonic acid, binds to the isoleucyl-tRNA synthetase enzyme, which catalyzes the formation of isoleucyl-tRNA [64]. This prevents further tRNA synthesis, arresting the production of proteins [65].

1.1.2.3 Inhibitors of cell wall synthesis

The cell wall is the component of bacterial cells most frequently targeted by antibiotics, with 15 different classes disrupting it (Table 1.1). Some classes of antibiotics act to disrupt the cell wall directly, such as the lipopeptides, polypeptides, polymyxins and enniatins. These antibiotics tend to function by binding to the outer membrane and either disrupting the permeability of the outerlayer or directly forming pores within to expose the periplasm [66–69]. Given the differences in cell wall construction between gram-positive and gram-negative bacteria, these antibiotics are specific to either architecture, with lipopeptides, polypeptides and enniatins effective against gram-positive bacteria, while polymyxins are effective against gram-negatives [66–69].

Other antibiotic classes target steps in the synthesis of the cell wall, as opposed to the cell wall directly. The β-lactam grouping of antibiotics represents the most utilised

antibiotics, accounting for 65% of all prescriptions of injectable antibiotics in the US from 2004-2014 [70]. This grouping includes the penicillins, carbapenems, monobactams and the cephalosporins, which bind to an array of different penicillin binding proteins (PBPs) [71]. These PBPs are enzymes involved in the terminal steps of peptidoglycan cross-linking, playing a key role in both gram-positive and gram-negative cell wall formation [70]. Other enzyme targets include: alanine racemase and D-alanine-D-alanine ligase enzymes involved in peptidogylcan biosynthesis, which are inhibited by cycloserines [72]; the MurA enzyme, which catalyzes the first committed step in peptidoglycan synthesis, and is targeted by phosphonates such as fosfomycin [73]; the NADH-dependent enoyl-ACP reductase, encoded by *inhA*, that is a part of the mycolic acid biosynthesis pathway in *Mycobacterium* species, and is targeted by thioamides and pyridinamides [74]; and Arabinosyl transferase, encoded by *embB*, involved in the biosynthesis of arabinogalactan in the cell wall of *Mycobacterium tuberculosis*, which is targeted by Ethambutol [75].

Apart from enzymes, substrates in the cell wall biosynthesis pathway are also targeted. Glycopeptides, such as vancomycin, bind to the D-alanine-D-alanine terminus of the lipid II monomer, preventing cross-linking of the cell wall [76]. Bacitracin binds to C_{55}-isoprenyl pyrophosphate (IPP), preventing its use as a carrier during the synthesis of peptidoglycan repeat subunits [77].

Across these three main synthesis pathways, antibiotics have been discovered that disrupt an array of molecular targets. In response to increased use of these antimicrobials in clinical, and other settings, bacteria have evolved a host of countermeasures.

1.1.3 Mechanisms of AMR

There are a variety of routes through which bacteria can be resistant to the effects of antibiotics. In some cases this can be due to inherent structural or functional properties that prevent biocidal activity by the antibiotic [78]. This is known as intrinsic resistance. For example, as mentioned above, lipopeptide antibiotics, such as daptomycin, are not effective on gram-negative species [68]. This is due to gram-negative bacteria having a lower proportion of negatively charged anionic phospholipids in their membrane, which reduces the efficiency of the Ca^{2+} mediated insertion of daptomycin into the membrane [78, 79]. It has also been argued that naturally occurring gene amplification, increasing the pro-

duction of target molecules to overcome the effect of an antibiotic, can be considered an intrinsic resistance mechanism [18]. One example of this occurs in Group B Streptococcus (*Streptococcus agalactiae*), where natural amplifications of the folate biosynthesis genes have arisen, conferring resistance to sulphonamides and diaminopyrimidines [80].

Apart from intrinsic routes, resistance can evolve through three main mechanistic groups: modification of the molecular target of the antibiotic; minimizing the intracellular concentration of antibiotic; and direct inactivation of the antibiotic [78].

1.1.3.1 Modification of antibiotic target

Alterations to target structure can affect the affinity of antibiotic binding, reducing the drug's ability to disrupt key cellular processes. These alterations can occur either as a result of mutational change in the gene sequence encoding a target, or through post-translational change, such as methylation, of a target [78]. For instance, resistance to mupirocin in *S. aureus* can arise through mutations in the *ileS* gene encoding the isoleucyl-tRNA synthetase enzyme [64, 81]. Stepwise mutations in the *ileS* gene are also indicated in the evolution of very high levels of resistance to mupirocin [82]. Any mutation, especially in the vital molecular pathways targeted by antibiotics, will tend to be deleterious [83]. Hence, there is often assumed to be a fitness cost associated with gains of resistance [84]. Experimental work in *Salmonella enterica* serovar Typhimurium identified substantial losses in fitness associated with *ileS* resistant mutations, with compensatory mutations required to offset this loss [85, 86]. However, similar experiments in *S. aureus* indicate no significant fitness cost was associated with even very high levels of mupirocin resistance [82].

M. tuberculosis resistance is thought to evolve solely through point mutations in the chromosome, with mutations in *embB*, most commonly at codon 306 in clinical isolates, driving ethambutol resistance for instance [87]. The missense mutations at this codon can drive very high minimum inhibitory concentrations (MICs) for ethambutol, up to 16 μg/ml, but these are associated with fitness costs [88]. Indeed many of the *M. tuberculosis* resistance mutations are associated with fitness costs [89, 90]. The spread of extremely drug resistant (XDR) and totally drug resistant (TDR) strains therefore highlights the immense selection pressure antibiotic consumption places on *M. tuberculosis* populations, but also

the ability of evolution to select for compensatory mutations that can alleviate this fitness cost [91, 92].

As well as mutational changes to a target's structure, resistance can also arise through modification after transcription or translation [78]. This is commonly seen with the methylation of rRNA components of the prokaryotic ribosome. The erythromycin ribosome methylase (erm) for instance, dimethylates the A2058 nucleotide of the 23S rRNA within the 50S ribosomal subunit, preventing the binding of lincosamides, macrolides and streptogramins [93]. Modification to outer membrane antibiotic targets can also occur, for instance with glycopeptides modification of the lipid II monomer can lead to high-levels of resistance. In this case the terminal D-alanine residue of the lipid II monomer is replaced with an isosteric D-lactate which lowers the affinity glycopeptides, such as vancomycin, for the peptide stem of the monomer by 10^3 fold [94].

Antibiotics target a limited number of sites, hence modifications at a single target may also lead to resistance to multiple antibiotics developing. Within *Haemophilus influenzae* for example, alterations to the PBP proteins can lead to resistance against a broad range of β-lactams [95]. Additionally, mutations altering the structure of the 23S rRNA of the 50S ribosomal subunit can confer resistance to macrolides and streptogrammins in *Streptococcus pneumoniae* (the pneumococcus) [96].

1.1.3.2 Minimization of antibiotic concentration

The minimization of the intracellular concentration of antibiotics can occur through either decreased membrane permeability to the antibiotic or effective efflux pumping of the antibiotic from the cytoplasm [97]. Gram-positive bacteria tend to be intrinsically more permeable than gram-negative to antibiotics, as such antibiotics have to enter gram-negative cells through porin channels in the membrane [98]. Gram-negative species therefore can acquire resistance to antibiotics through either downregulating porin production or switching to more selective porin channels [78]. For instance in Enterobacteriace, high levels of carbapenem resistance have been observed in isolates that lost or altered porins, in the absence of a carbapenemase that might otherwise have explained this resistance [99]. While in *H. influenzae* isolates extracted from cystic fibrosis patients, alterations in the porin amino acid sequences led to decreased susceptibility to streptomycin [100].

Efflux pumps are one of the most ubiquitous types of resistance elements, present in gram-negatives and gram-positives [101, 102]. These pumps can be specific in nature, such as the Tet and Mef exporters which target tetracyclines and macrolides respectively [103]. They can also be broader in nature, known as multi-drug resistance (MDR) efflux pumps. Within gram-positives, the most well studied MDR efflux pump is the NorA pump first identified in *S. aureus* in 1986 [101, 104]. This pump can export fluoroquinolones and other antiseptics such as ethidium bromide and quaternary ammonium compounds from the cell [105]. Resistance is often caused by overexpression of this NorA pump, which can be in response to exposure to biocidal compounds [106], or to iron limitation, either of which indicates that *S. aureus* is within a host environment [78, 107].

Within gram-negatives, the resistant nodulation division (RND) MDR pumps are the most well characterised [78]. Members of this RND pump family have been shown to export fluoroquinolones, tetracyclines, chloramphenicol and some β-lactams too [108]. Similar to the NorA pump, resistance is often conferred by overexpression of RND pumps [109]. In *Escherichia coli*, indole presence, which is a stationary-phase extracellular signal, causes the overexpression of the AcrAB RND pump conferring MDR to these cells [110].

1.1.3.3 Direct inactivation of an antibiotic

The final mechanism of resistance is whereby bacteria act directly on antibiotics, either to modify or destroy them. The first identified enzyme to catalyze the hydrolysis of an antibiotic was the penicillinase discovered by Abraham & Chain in 1940 as penicillin was being manufactured [20]. Following the introduction of other β-lactams, such as the cephalosporins, extended-spectrum β-lactamases (ESBLs) were discovered in the 1960s on plasmids in Enterobacteriaceae [111, 112]. This earliest ESBL was designated TEM-1 , and could hydrolyse the β-lactam ring of penicillins and cephalosporins [113]. Another ESBL class which was initially common when discovered, were SHV-type enzymes first found in *Klebsiella pneumoniae* and *E. coli* [111]. Today though, both these classes are becoming less and less common, with TEM-type ESBLs detected in less than 1% of *E. coli* and *K. pneumoniae* in Europe [114]. Instead the most commonly found ESBLs are in the CTX-M family, with CTX-M14 and CTX-M15 the most widely isolated of this

family [78, 114]. These ESBLs confer resistance to 3rd generation cepahlosporins and have been associated with a number of successful pathogenic clones, such as ST131 *E. coli* [115–117].

As well as hydrolysing an antibiotic, enzymes can also act to modify their structure preventing binding to molecular targets. Aminoglycosides are particular targets of these enzymes, with three main classes targeting them: acetlytransferases, phosphotransferases and nucleotidyltransferases [118]. These enzymes are very diverse in nature and found across a range of host taxa [118, 119].

We can see then, that in response to the challenge of antibiotics, bacteria have evolved a whole host of mechanisms to survive in their presence. The spread of these mechanisms to important human and animal pathogens has led us to the looming public-health challenge of AMR.

1.1.4 The evolution of AMR

1.1.4.1 The resistome

Resistance genes are everywhere [120]. From the clinically relevant sites of hospital wards and operating theatres [121–123], to the less relevant, but still important, reaches of outer space [124] and isolated cave systems [125]. This reflects the millions, or potentially billions, of years over which bacteria and fungi have been competing with each other [126, 127]. Indeed, molecular clock studies looking at the emergence of penicillin have dated the machinery used for its biosynthesis to at least one billion years ago [126]. Resistance genes are ancient [128]. What is not ancient, however, is the spread of resistance genes within pathogenic bacteria, which tend to be much less diverse in terms of biosynthetic gene clusters and more adept at immune evasion [126]. Historical collections of pathogenic bacteria, such as the Murray collection of *Enterobacteriaceae* [129], indicate that before the widespread introduction of antibiotics these pathogenic species were largely antibiotic susceptible [130]. Much of the spread of resistance in pathogenic species has been from the mobilisation of the ”resistome” of commensal or environmental bacteria [131]. The CTX-M ESBL genes described above for instance, originated in the environmental genus *Kluyvera* and were mobilised on the insertion sequence IS*Ecp*1 [132–134].

The important human pathogen the pneumococcus, which is a leading cause of death in children under the age of five globally [135, 136], is another species which has gained resistant genes from commensal species. Resistant mosaic forms of the *pbp1a*, *pbp2b* & *pbp2x* genes, which are targeted by penicillins, all arose initially through recombination with related streptococcal species, such as *S. mitis* and *S. oralis* [137–139]. Penicillin-non-susceptible pneumococci (PNSP) were first detected in Massachusetts in the 1960s, with isolates with very high MICs, up to 12 μg/ml, found in South Africa by the 1970s [140]. Since then penicillin-resistant MDR clones of the pneumococci have gone on to spread globally, with the PMEN1 lineage responsible for 40% of penicillin resistant invasive pneumococcal disease (IPD) in the US between 1996 and 1997 [141, 142].

1.1.4.2 Modelling the spread of resistance among bacterial populations

In general, empirical evidence suggests a linear relationship between antibiotic usage and antibiotic resistance [143]. Indeed, when I plot the latest 2020 European Centres for Disease Control and prevention (ECDC) data on β-lactam consumption and the proportion of invasive pneumococci which are PNSP, there is a significant relationship (Adjusted R^2 = 0.2309, F(1,24) = 8.506, p = 0.007562) (Figure 1.2). Countries with the highest rates of antibiotic prescription tend to have the highest rates of resistance [144]. However, these are often just snapshots of resistance and consumption. Longitudinal surveillance data instead demonstrates the coexistence of susceptible and resistant lineages [143, 145]. Indeed for many pathogens, such as the pneumococcus, *E. coli* and enterococci, the prevalence of resistant phenotypes has remained stable in Europe over the last 15 years [146–150]. This dynamic is not reproducible through simple person-person models which result in competitive exclusion of susceptible strains by resistant strains [145].

As such there has been a broad exploration of different mechanisms that could explain this coexistence. Colijn *et al* 2010 [146] built on previous work exploring how more structurally neutral models, which do not implicitly build in coexistence, could elucidate the factors which underlie this phenomenon [146, 151]. Of the five models they tested, the framework which best produced coexistence allowed for simultaneous co-infection with susceptible and resistant strains, although this coexistence only occurred in 20% of their simulations [146]. Recent work has also explored the (non-exclusive) impact of pop-

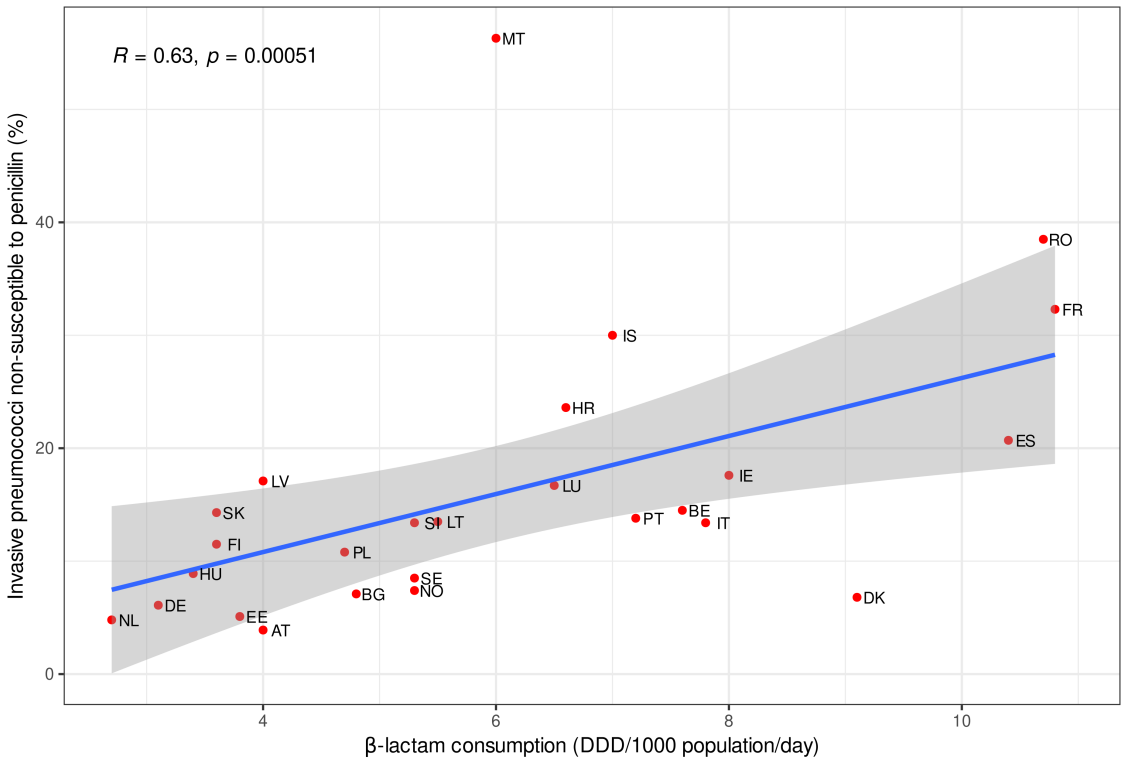

Figure 1.2: Relationship between β-lactam consumption and PNSP invasive isolates across Europe in 2020. The percent of invasive pneumococci recorded as PNSP across 24 European countries comes from the ECDC Antimicrobial resistance in Europe report 2022. The consumption of β-lactams for 2020 also comes from the ECDC, combining countries reporting on both hospital and community level prescriptions. Countries values are labelled by their eurostat two letter country code. PNSP used the European committee on antimicrobial susceptibility testing (EUCAST) guidelines, where intermediate or resistant isolates, those with an MIC > 0.06 μg/ml, were deemed non-susceptible. The Spearman's rank correlation coefficient and p value for this correlation are displayed in the top left hand corner of the plot. The blue line represents the fitting of a linear model.

ulation structure on the maintenance of both susceptible and resistant strains [152, 153]. Cobey *et al* 2017 [152] explored how age assortative mixing and age specific treatment can maintain strain coexistence among a simulated pneumococcus population. While they did observe resistant and susceptible coexistence, the resistant fractions, and the costs of resistance at which these fractions were maintained, were generally unrealistic. Blanquart *et al* 2018 [153] built a more generalizable ordinary differential equation (ODE) model incorporating population structure. While this framework could also produce coexistence between resistant and susceptible strains, the population class structure was too rigid and their rates of inter-class transmission that favour coexistence were unrealistically low.

As well as population structure and co-infection dynamics, work has also focused on how genetic differentiation between susceptible and resistant strains can promote coexistence [147, 154]. Lehtinen *et al* 2017 [147] explore how genetic linkage between resistance genes and loci which influence carriage length could promote coexistence. They hypothesize that balancing selection on duration of carriage could maintain the linked resistance determining loci at intermediate frequencies in a population. Indeed this model is able to produce coexistence of the resistant and susceptible strains, however it is unable to explain resistance profile heterogeneity among isolates of the same serotype. In this framework, resistance could also be determined by horizontal gene transfer (HGT) rate, which in turn would be driven by carriage duration [154, 155]. To test this, Lehtinen *et al* 2020 [156] investigated a large population of pneumococci, looking at carriage episodes. From their measures of HGT they found only weak evidence for a link between higher rates of HGT and resistance in lineages, suggesting carriage duration was the main driver.

Recent work has also sought to include within-host dynamics of competition between strains to explain the population level coexistence of resistant and susceptible isolates [145]. Davies *et al* 2019 [145], building on previous work investigating malaria drug resistance [157], move on from a simple knock-out model of strain competition, and instead model individual hosts as having a carrying capacity while tracking strain frequencies within a host. This more closely mimics the colonization process of bacteria and it is able to capture the observed coexistence trends for *E. coli* and the pneumococcus. However, this model does make some broad assumptions about the fitness cost of resistance and host immunity [158].

Finally, in Davies *et al* 2021 [150] the authors compare four different models, determined to be biologically plausible, in their ability to explain how resistant and susceptible strains of the pneumococcus can coexist. The four models are: (i) treatment diversity, where a population is split into subpopulations varying in antibiotic treatment rate, similar to Cobey *et al* 2017 [152] & Blanquart *et al* 2018 [153]; (ii) pathogen diversity, where carriage duration and subsequent heterogeneity in the fitness effect of resistance is a factor in explaining resistance emergence, similar to the framework initially proposed in Lehtinen

et al 2017 [147]; (iii) treatment competition, where hosts can be colonized by resistant and sensitive strains and treatment determines the strain transmitted from the host, similar to *Colijn* et al 2010 [146]; and (iv) growth competition, an enhancement of iii where in the absence of treatment there is a fitness cost, in terms of growth rate, on resistance, similar to Davies *et al* 2019 [145]. They find that all four models, in their implementation, are able to recapitulate the trend in resistant pneumococci across European countries for 2007 equally well. However, they do not measure how well these models can match trends over a longer time period, or how competition with other nasopharyngeal species may affect these dynamics [150]. Taken together these studies suggest that multiple factors are needed to explain how resistance spreads through a bacterial population, while also serving to highlight how we do not fully understand the process [145, 159].

1.1.4.3 Modelling the spread of MDR

The evolution of MDR pathogens represents a particular threat to public health, as infections caused by these strains are more often challenging and expensive to treat [18, 160]. In 2017 the WHO published a list of six pathogens where the spread of MDR was especially concerning: *Enterococcus faecium*, *S. aureus*, *K. pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter* species [161]. These six ESKAPE pathogens are responsible for the majority of nosocomial infections globally [162, 163]. As well as being a public health challenge, the evolution of MDR also represents a scientific puzzle [160]. This is as resistance genes appear to aggregate in the same strains of bacteria, rather than being randomly distributed, leading to the idea of MDR over-representation [160, 164].

Numerous theories have been put forward to explain the evolution and dissemination of MDR pathogens, and these have been summarised previously in Chang *et al* 2015 [160] and Lehtinen *et al* 2019 [164]. One theory posits that MDR can be caused by a single resistance mechanism conferring resistance to multiple drugs, such as MDR efflux pumps [165] and ESBLs [166]. Genetic linkage between resistance genes is also thought to drive resistance. This is observed through the spread of mobile genetic elements (MGEs) containing multiple resistance genes, such as the large MDR plasmids seen in the Enterobacteriaceae [167]. Another theory is that highly mutagenic or recom-

binogenic lineages acquire more resistance genes. This is observed in the Beijing lineage of *M. tuberculosis* which is highly mutagenic and appears to accumulate resistance mutations faster than other lineages [168]. Treatment regimen failure is another theory, this can occur through combination therapy on populations with singly resistant lineages spontaneously developing further resistance genes, as is sometimes seen in *M. tuberculosis* treatment [169]. This can also occur when the resistance status of an infection informs the drug chosen, such that resistance to the chosen drug may also be acquired by the already resistant infection. Finally, cost epistasis is another mechanism put forward to explain MDR. This whereby there is a lower than expected fitness cost to the presence of multiple resistance determinants, this may occur with MDR plasmids where the fitness cost of the presence of the plasmid is the main cost, not the cost of each of the individual resistance genes [160, 164].

These explanations, however, are often applicable to only a subset of antimicrobials or pathogens, whereas MDR is seen across different drugs and species [164]. As such, recent modelling work by Lehtinen *et al* 2019 [164] has sought to explain the evolution of MDR, building on the framework they had previously developed for modelling the coexistence between resistant and susceptible strains [147]. Here they allow for different strata within their model based on host population structure and pathogen strain structure, with the fitness costs of resistance varying between these different strata. They develop concepts of resistance proneness for a strata, taking into account the antibiotic consumption in each strata and the clearance rate of a strain, and the resistance threshold for an antibiotic, taking into account the fitness cost of resistance and the proportion of consumption a specific antibiotic accounts for in the population. These concepts can produce nested patterns of resistance, which follow trends in pathogenic species, and also predict strains with a longer carriage duration are more resistance prone, again broadly matching observed trends. As with their previous model however, the framework can not explain resistance profile heterogeneity within a strain and assumes no cost epistasis between resistances.

A study by Jacopin *et al* 2020 [170] looks at incorporating cost epistasis into modelling the evolution of MDR. They model a large unstructured population initially, also allowing

for within-host strain competition and recombination, unlike Lehtinen *et al* 2019. They model only two different drugs, finding that dominance of a single doubly-resistant strain is the most frequent model outcome, with strain coexistence rare, which is unlike real-world data. When modelling a a structured host population, they find that negative cost epistasis greatly favours the evolution of MDR strain, however they still were not able to recapitulate the coexistence of sensitive and resistant strains.

Another recent study, McLeod & Gandon 2021 [171], focuses on how linkage disequilibrium (LD) dynamics can explain the population structure of MDR pathogens. They treat MDR and LD as synonymous and also use a metapopulation model to assess the evolution of MDR. They find that carriage duration is not necessary or sufficient for MDR to develop, however in testing this they do not include epistatic interactions between resistance loci.

As with resistant and susceptible strain coexistence, there is still more work to be done to fully understand the process of MDR evolution. A better understanding of this could lead to better treatment regimens aimed at preventing the spread of MDR pathogens.

1.1.5 The impact of AMR on public health

1.1.5.1 The effect of AMR on patient outcomes

Assessing how AMR affects clinical outcomes in patients is key in understanding the potential impacts of the spread of resistance. Studies looking at how resistance impacts morbidity and mortality can often be confounded by the lack of an adequate control population, with cases needing to be matched by aetiological agent and severity of infection [172]. This has led to a relative paucity of information on the effect of resistance both on morbidity and mortality, and on the relative costs associated with treatment [173]. Often the popular press will instead present an apocalyptic image of resistant ”superbugs” [174, 175]. A recent surveillance study, however, identified that only one patient out of a total of 85 infected with an extremely drug resistant (XDR) or pan drug resistant (PDR) infection died within the Marseilles hospital system between 2009 and 2015 [176]. This patient also had significant comorbidities. While another study looking into PDR infections in a Crete hospital over an eight year period, also found a lower than expected

mortality rate from PDR infections [177].

Meta-analyses focusing on specific species and antibiotic resistances have shown an increase in mortality in resistant infections though [178–181]. Cosgrove *et al* 2003 [178] found that, while 24 of the 31 studies they investigated saw no significant difference in the mortality for methicillin-resistant *S. aureus* (MRSA) and methicillin-sensitive *S. aureus* (MSSA), when pooled together there were significantly higher mortality rates for MRSA bacteraemia compared to MSSA bacteraemia. Similarly: DiazGranados *et al* 2005 [179] found higher mortality rates for bacteraemia caused by vancomycin-resistant enterococci compared to bacteraemia caused by vancomycin-sensitive enterococci bacteremia; Rottier *et al* 2012 [180] found bacteraemia caused by ESBL producing Enterobacteriaceae had a higher mortality rate than bacteraemia caused by non-ESBL producing Enterobacteriaceae; and Lemos *et al* 2014 [181] found carbapenem-resistant *A. baumannii* infections had a higher mortality that carbapenem-susceptible *A. baumannii* infections. Although for Lemos *et al* 2014, the authors were unable to control for all confounding factors, and they investigated a broader range of pathologies than the previous three studies.

Additionally, cohort studies appear to show resistant infections of *E. coli* [182], *P. aeruginosa* [183, 184], and *S. aureus* [185] are all associated with higher mortality rates. The increase in mortality associated with resistant strains is most likely a result of the delay in effective antibiotic treatment caused by resistance [186]. Within Cosgrove *et al* 2003 for instance, when the analysis was limited to outbreak settings where MRSA was known to be present, the difference in mortality between MSSA and MRSA was not significant [178]. Second-line antibiotics used to treat MRSA infections, such as vancomycin, are also less effective against MSSA infections [187]. However, factors not relating to treatment could also play a role, such as resistant strains being associated with a greater virulence. Plasmids are often seen to carry both resistance and virulence genes for example [188, 189]. Controlling for these factors would require a more holistic genomics based approach, with sequencing of invasive isolates followed by the identification of genetic factors other than resistance that may influence infection mortality.

1.1.5.2 Predictions of the burden of AMR

One of the first major attempts to assess the global threat of AMR was the O'Neill report on antimicrobial resistance [190]. Overseen by the economist Jim O'Neill, who tasked the analysis to two business consultancies, the report estimates that in 2015 700,000 people died globally from AMR and predicts that by 2050 some 10 million people globally will die from AMR infections, which would outstrip the death toll of cancer [190, 191]. These predictions are based on current rates of resistance evolution and assume resistance mutations will reach fixation in a population. This is contentious, as it treats the evolution of resistance as a simple process of mutant emergence and subsequent fixation [159, 192]. As seen above for pathogenic species however, there often appears to be long-term coexistence of resistance and susceptible lineages. Additionally the report often glosses over a lack of adequate data, assuming for instance that antibiotic consumption in humans matches that in agriculture, while its methods in general are opaque and not peer-reviewed.

In 2019 Cassini *et al* attempted to more rigorously estimate the number of deaths attributable to AMR infections in Europe [193]. Using: the European antimicrobial resistance surveillance network (EARS-Net), a literature review of publications relating to mortality attributable to resistance, and ECDC data on non-blood stream infections (nBSIs), among other sources, they estimated 33,110 deaths were attributable to AMR infections in 2015 across the EU. With results broken down by species, they estimated that third-generation cephalosporin-resistant *E. coli* (3GREC) caused the most deaths, at a median of 9,066, closely followed by MRSA at a median of 7,049 attributable deaths. While this is more rigorous than the O'Neill report in its estimation for 2015 attributable deaths, there are still limitations in the study, with limited data on a range of pathogens' attributable mortality scores and a broad conversion used to estimate the number of nBSIs.

Most recently, the Lancet global burden of disease (GBD) study, led by the Institute of Health Metrics and Evaluation (IHME), attempted to estimate the deaths attributable to AMR infections globally in 2019 [194]. To produce this figure they used: estimates of deaths from infectious and non-infectious syndromes from the GBD 2019 study [195], hospital discharge data to model case fatality ratios (CFRs) for individual pathogens, sales

and surveillance data to assess AMR spread, literature searches (taken from Cassini *et al* 2019 [193]) and clinical lab data to assess the mortality attributable to resistance. In total they estimated 1.27 million deaths were attributable to AMR in 2019. Unlike Cassini *et al* 2019 they found MRSA caused the most deaths, with 121,000 compared to 3GREC at 59,900 in 2019.

As with the previous studies though, there are flaws, particularly around the data used to create these estimates. The hospital discharge data used for instance, which informs CFR calculations and the distribution of pathogens, only comes from seven countries. None of these are in Africa, estimated to suffer the highest burden of AMR deaths. Due to data sparsity too, the authors were only able to produce a global estimate for the relative risk of death from an AMR infection, not a location specific estimate. Given the disparities in healthcare systems around the world, this may not be accurate [196, 197]. Finally, there have also been longstanding criticisms of the overall GBD modelling estimates, used to estimate incidence here, with these seen as opaque in terms of data sources used and the estimation methodology [198, 199].

Taken together these studies serve to highlight the general problem of AMR spread and the potential public health crisis it engenders. However, the absolute estimates of lives lost to AMR infections should be viewed sceptically.

1.2 Genomic data analysis

1.2.1 Sequencing bacterial genomes

1.2.1.1 Sequencing methods

The first bacterial genome sequenced and assembled was an *H. influenzae* isolate in 1995 [200, 201]. This was sequenced using the method first developed by Sanger *et al* 1977 [202], combined with a shotgun approach of breaking the whole genome up randomly into smaller lengths to sequence [201, 203]. Briefly, the Sanger method relies on using differently labelled dideoxy terminators, each label corresponding to a specific base, with random incorporation of these terminators into a strand to halt synthesis. These strands are then separated by size through electrophoresis and the sequence of bases can be determined by the patterns in the bands formed [204]. The shotgun approach can then

produce very accurate reads of up to 800 bp long that are subsequently assembled. The Sanger sequencing method is still considered the gold standard of sequencing, although it has low throughput, meaning it cannot be used in real-time clinically, and is expensive to perform [201].

These drawbacks prompted the development of high-throughput sequencing technologies, also known as next-generation sequencing (NGS), with Sanger sequencing termed the first generation [201, 205]. Initially NGS methods encompassed a range of different technologies: cyclic reversible termination (CRT), sequencing by ligation, and pyrosequecing for example [203, 205]. The generation of bacterial genomic data though, has become dominated by Illumina technology, which uses the CRT approach [203]. In this method, briefly, fragments of DNA and primers are attached onto a flow cell and then amplified to create many individual template strands [203]. A DNA polymerase then incorporates a fluorescently labelled and terminator modified nucleotide to a template strand fixed in place on a flow cell. The incorporated nucleotide's terminator group allows for fluorescence imaging to detect only the newly incorporated base. The fluorescence label and terminator group are then cleaved off, allowing for the extension of the complementary strand by further fluorescent terminator bases [205]. While this method did drastically reduce the cost per run and allow for a very high sequencing yield, the short reads produced, up to only 250 bp, are not as accurate as Sanger sequenced reads and the length ensures assembly is harder in regions of high repeats [203].

The most recent advances in sequencing technologies have concerned approaches that seek to combine the high-throughput nature of NGS and the longer reads of Sanger sequencing approaches. These are long-read sequencing (LRS) methods [206,207]. Two approaches in particular, Pacific Biosciences' (PacBio) single-molecule real-time (SMRT) sequencing and Oxford Nanopore Technologies' (ONT) nanopore sequencing, are predominant in the LRS market [207]. In PacBio sequencing, hairpin adapters are added to long dsDNA molecules to create a circular ssDNA molecule called a SMRTbell. These SMRTbells are then fed into a nanoscale observation chamber with a polymerase fixed to the bottom. Fluorescent nucleotides are subsequently incorporated and an illumination signal is recorded in real time to assess the base added [206]. In nanopore sequencing,

DNA sequences are sheared and end repaired with an adaptor complex, which consists of a tightly bound polymerase or helicase [206, 208]. These adaptors are then bound to nanopores within a membrane, known as a flow cell, with the adaptor enzyme ensuring ensuring stepwise movement of ssDNA through the pore. The movement of the base through the pore creates a change in the ionic current that is dependent on the base, which can be interpreted and, with further processing, used for base calling of the sequence [206].

Both these approaches can produce very long reads, with a recent benchmarking of *E. coli* sequencing producing a maximum read length of 297,344 bp for nanopore sequencing and 151,906 for PacBio sequencing [209]. However, both methods do tend to have very high error rates, with single pass methods in PacBio producing error rates as high as 15%, although this does drop to 0.1% with nine passes through the nanoscale observation chamber [210]. PacBio sequencing is seen to have a higher throughput than nanopore, however the machine costs are higher and the turnaround time for sequencing is a lot longer [211]. Indeed, the quick turnaround times, lower machine costs, and highly portable nanopore sequencing machines have allowed for real-time sequencing in outbreaks and *in situ* sequencing even in resource poor settings [212–216].

1.2.2 Predicting AMR from genomes

Traditionally, resistance phenotypes of bacteria have been assessed through antimicrobial susceptibility testing (AST) [217]. Broth dilution is currently the gold standard of AST, where bacteria are grown in the presence of different concentrations of an antibiotic to assess its effect on the growth of these isolates [218]. These results define the MIC values of an isolate, and reference bodies, such as the European Committee on Antimicrobial Susceptibility Testing (EUCAST), have developed stanardized protocols to determine the clinical resistance categories of isolates [219]. In a clinical setting AST methods are slow, typically requiring the culturing of isolates, which can take days, followed by these growth based assays which can also take days [220, 221]. More rapid automated AST methods have been developed, however these still often require the culturing, or incubation, step, which acts as a bottleneck on the process [220, 222]. This culturing also prevents the resistance phenotypes of unculturable bacteria from being explored. These bacteria

represent the majority of known bacterial diversity [223].

Improvements in sequencing methodologies, in terms of costs and runtimes, mean bacterial sequence data is much more readily available [224]. This has opened up the possibility of using whole genome sequence (WGS) data to assess the resistance profile of isolates [225]. Analysing WGS data has the potential to provide rapid results, not only on resistance but also for key virulence loci, in a clinical setting, while also allowing for the investigation of a broader range of bacterial species in surveillance studies [221,225]. Current methods to detect resistance genes from WGS data fall broadly into two categories: direct association, where the presence or absence of genetic variants known to cause resistance are detected; and the application of predictive models derived from machine learning approaches [226].

1.2.2.1 Direct association methods

Direct association methods typically involve mapping sequences onto curated reference databases of AMR genes, or known resistance mechanisms [227,228]. These databases can be species specific, such as MUBII-TB-DB [229] for *M. tuberculosis* and u-CARE for *E. coli* [230], antibiotic specific, such as TEMLacED for TEM ESBLs [231], or more broad in nature [225,228]. Two of the most widely used broad databases are the Comprehensive Antibiotic Resistance Database (CARD) and Resfinder databases [232–234]. The ResFinder database is part of the wider ResFinder tool, this database is updated regularly and currently contains over 3000 AMR genes curated from literature reviews. It does not contain efflux pump genes, but it has been merged recently with the PointFinder database, which contains data on point mutations and genetic variants linked to resistance [234]. The CARD database is ontology based and has over 6,500 terms, encompassing resistance genes as well as single nucleotide polymorphisms (SNPs) and genetic variants. It is updated monthly through expert feedback and literature reviews [232]. A recent systematic evaluation of the CARD and ResFinder databases found that CARD predictions had a higher major error rate (predicting resistance where an isolate is susceptible) but lower very major error rates (predicting susceptible when an isolate is resistant) compared with ResFinder [235]. Both databases though were not able to match current clinical standards for AST [235].

The various software tools which take in sequence data to compare to these databases can act on read data or assemblies [225,228]. Read data tools, such as SRST2 [236] and GeneFinder [237], can map sequences onto reference databases using Bowtie2 [238]. Other read based approaches use k-mer based mapping to reference databases, such as the ResFinder tool [234, 239]. Using mapping tools directly on linear representations of reference genes or variants can cause reference bias, masking important variation between subtypes which themselves can cause resistance [225, 240]. As such recent read-based approaches have also sought to build non-linear data structures to capture all the genetic variation in a reference database [241, 242]. In Mykrobe [241] a De Bruijn graph is constructed from query reads and compared to a De Bruijn graph created from a reference database to detect resistance, while Graphing Resistance Out of Metagenomes (GROOT) [242] query reads are mapped to a variation graph created from the reference database.

In contrast to read-based approaches, approaches that take in assembled sequences include the popular AMRFinderPlus [243] from the National Centre for Biotechnology information (NCBI) and CARD-RGI, which integrates directly from the CARD database [232]. These assembly based methods tend to use Basic Local Alignment Search tool (BLAST) for searching for resistance genes among assemblies [228], while some databases also contain a suite of Hidden Markov Models (HMMs) to find sequences with a similar function but low sequence identity [225,244]. Using assemblies to search for resistance genes can introduce potential biases from the *de novo* assembly steps, with duplicates of genes and repeat regions potentially confounding assembly methods [245–247]. Assemblies though, do allow for positional information to be retained, so upstream and downstream regulatory regions, and any potential epistatic effects can also be investigated [228]. Read-based methods however are typically faster, as the time consuming assembly step is skipped. Additionally read-based methods can be used with rarer species and in metagenoic studies where asssembly methods are limited [225, 248].

1.2.2.2 Predictive modelling methods

Direct association methods have been shown to be accurate for species, such as *M. tuberculosis* and *S. aureus* [237, 241, 249–251], and drugs, such as the β-lactams and

flouroquinolones [226, 252]. However, these methods do rely on the previous characterisation of resistance mechanisms, meaning species which are not as well studied may be erroneously predicted as susceptible [253]. Additionally, these methods are often not fully able to capture any epistatic interactions between loci, which may produce non-linear effects in determining resistance [254, 255]. Instead, approaches which build predictive models of resistance from sequence data, often using machine learning strategies, can model more diverse resistance mechanisms without the need for *a priori* knowledge of resistance loci [226].

These modelling methods typically involve training a classifier algorithm on a set of well-characterised sequence data, which could involve SNPs, genes, k-mers, amino acids or other features, with known resistance phenotypes [256]. Then using this trained model to predict unseen data, assessing the accuracy of the model based on this data [228]. There are a range of different classifier approaches available, including: logistic regression based approaches [257, 258], random forest (RF) approaches [259–261], support vector (SVM) machines [262, 263], rules-based approaches such as set covering machines (SCM) [264, 265], and deep learning approaches such as neural nets [266–268]. I will briefly describe two examples of applying these machine learning based approaches to predicting resistance phenotypes.

As mentioned above, much of the β-lactam resistance in pneumococci is driven by variations in the three PBP genes, *pbp1a*, *pbp2b* and *pbp2x* [140]. The variation in levels of resistance within clinical strains was further found to be primarily driven by alterations in the transpeptidase domains (TPDS) of these three genes [269]. Li *et al* 2016 used a large dataset of 2,528 pneumococci, with lab derived AST data, to extract their TPD domains and train three different classifiers on this data [269]. Their rules-based approach, whereby resistance was predicted as the modal MIC of isolates with the same TPD profile as the query, and an RF approach outperformed the elastic net, a derivative of logistic regression. Building on from this, in Li *et al* 2017 they expanded their data to 4,309 isolates and further tested the rules-based approach and the RF approach [270]. They found the RF approach to be more accurate than the rules-based approach, as it was able to more accurately predict resistance levels for novel TPD profiles [270]. Overall the major error

rate and very major error rate for the RF model were low, at 1.2% and 1.4% respectively across the six different antibiotics tested against.

In Hicks *et al* 2019 the authors set out to evaluate the accuracy of RF and SCM classifiers on three separate species, *N. gonorrhoeae* (the gonococcus), *K. pneumoniae* and *A. baumannii*, predicting resistance against the fluoroquinolone ciprofloxacin and the macrolide azithromycin [226]. Within gonococci, both classifiers did well predicting ciprofloxacin resistance, which is determined largely by mutations in the *gyrA* and *parC* genes [226, 271]. However, for azithromycin resistance, which has arisen multiple times through many different pathways in Gonococci, both classifiers performed poorly, with an average balanced accuracy (a combination of both sensitivity and specificity) significantly lower than that for ciprofloxacin [226]. When the the models were trained on the much more diverse *A. baumannii* and *K. pneumoniae* populations the models also performed significantly worse than on gonococci datasets for ciprofloxacin. The SCM model was roughly equivalent in terms of balanced accuracy for ciprofloxacin in *A. baumannii*, but the RF model was significantly worse, while in *K. pneumoniae* both models performed significantly worse than when predicting gonococci resistance [226]. In this case, the authors suggest more comprehensive sampling of these diverse populations, along with methods that can adjust for populations structure and other confounders, will hopefully improve these prediction methods to a level where they can be used in a clinical setting.

1.2.3 Genomic epidemiology

A bacterial WGS contains a wealth of information. From this sequence we can predict the AMR profile of the isolate in question, detect important pathogencity related genes, and assess key immunogenic loci, such as the capsular serotype, among others [243, 272, 273]. All of these phenotypic details are important in clinical settings. The continual advancement in sequencing technologies though, allows for the sequencing of a large number of genomes in a short time period, data that can be vital to the investigation of a local outbreak or the dissemination of pathogens globally [274]. Once this data has been assembled, one of the first and most important steps in providing clinically relevant data is assessing the relationships between isolates [275].

1.2.3.1 Inferring the relationship between isolates

Sequencing has been used to identify genetic variants linked to transmission in viral pathogens for decades now [276–278]. Their much smaller genome size made these approaches feasible in the early genomics era [279]. In bacteria, instead, other methods that utilised only a fraction of the WGS, or protein sequences, were initially used to delimit isolates in an outbreak [280]. One of the first methods for assessing the population genetic structure of bacteria was Multilocus Enzyme Electrophoresis (MLEE) [281]. MLEE assesses the differences between strains through electrophoretic variation of the amino acid sequences of key enzymes [281]. However, MLEE studies have very low resolution in terms of assessing genomic changes, with roughly only one in twenty of all possible mutations detected [282]. Pulsed-field gel electrophoresis (PFGE), which involves enzyme restriction of bacterial DNA followed by separation into bands through electrophoresis, was also a popular initial tool, and is still used by smaller labs today [283]. Although, PFGE struggles to discriminate between pandemic clones, such as the USA300 MRSA strain, and more clonal species, such as *M. tuberculosis* [274, 280]. Additionally, results for both PFGE and MLEE are difficult to standardize, due to the often only small differences observable between strains [274, 284].

Multi-locus sequence type (MLST) methods were one of the first sequencing-based methods for separating bacterial isolates [274]. MLST relies on determining the allele presence at core housekeeping loci [285]. The allelic profiles of these genes have been standardised with online databases of MLST profiles, which enables a common strain nomenclature. MLST approaches also enabled the building of phylogenies directly from sequences to highlight the relationships between strains [286].

Phylogenies are composed of branches representing the persistence of a lineage through time, and nodes which represent the birth of new lineages in the population [287]. Broadly, phylogenies can be built using two main classes of methods, distance-based and character-based [287, 288]. Distance-based methods first apply a model of nucleotide substitution to a set of sequences to calculate the genetic distance between these sequences. These substitution models can be relatively simple in nature, such as the Jukes-Cantor (JC) which assumes all substitutions occur at an equal rate [289], or more com-

plex, such as the general time reversible (GTR), which allows for different rates based on the mutation [290]. Once these distances have been calculated an algorithm, such as Neighbour-Joining (NJ), will cluster together sequences to form a tree [291]. These distance-based methods are fast, especially compared to character-based methods on larger datasets, although they struggle on highly diverse datasets [287].

Character-based methods search for the most-probable tree for a set of taxa given the sequence at each position [292]. Individual trees are scored based on a model of evolution, such as the nucleotide substitution models described above [291]. A common early character-based approach was maximum parsimony (MP) methods. This seeks to select a phylogeny that explains the sequence alignment in terms of the minimum number of substitutions possible [288]. MP approaches are thus computationally efficient [287]. However they struggle with divergent sequence and can not correct for multiple substitutions at the same site [293].

Maximum likelihood (ML) methods allow for more complex evolutionary models to be used to score a tree. These approaches attempt to find the parameters, the tree topology, branch lengths and substitution model parameters, that maximise the likelihood of obtaining the sequence data observed under the evolutionary model applied [294]. Most of the models applied in ML methods assume an independent evolution of sites, so the likelihood can be calculated as the product of the probabilities of all sites [287]. These ML methods are useful in that they can explore a range of evolutionary models, most of which will have explicit assumptions that can account for different evolutionary processes [287]. However they are computationally demanding and are reliant on the correct choice of evolutionary model [288].

Finally Bayesian methods are another approach to character-based phylogenetic reconstruction. In the Bayesian approach the parameters (the tree topology, branch lengths and evolutionary model parameters) are considered to be random variables that can be described by statistical distributions, rather than the fixed constants that ML methods assume [287]. Before analysis these parameters can be assigned prior distributions, which can allow for the incorporation of domain knowledge from users [295]. These priors are then combined with the data to generate a posterior distribution, representing the proba-

bility of a phylogeny given the data observed. This posterior distribution can not be calculated directly, instead these methods rely on Markov chain Monte Carlo (MCMC) sampling to approximate the posterior [287]. This algorithm allows for more extensive searching of the probability distribution, potentially finding the global optimum tree. Whereas ML methods perform a less extensive search, and can return only a local optimum tree. As with the ML methods, Bayesian methods also allow for the testing of a range of complex evolutionary models, while their probability distribution output is more easily interpreted than the likelihoods produced from ML methods. However, these methods are very computationally intensive with MCMC sampling struggling to converge on large alignments.

Early phylogenies built from MLST data were informative about the relationships between the strains of pathogenic species, while also helping to uncover the extent of recombination among bacterial isolates [286, 296, 297]. However, the MLST method has a fixed resolution and can struggle with both highly diverse and more clonal populations [298]. With the arrival of NGS and LRS, isolates can now be separated by individual SNP differences from WGS data. The phylogenies built from WGS are incredibly informative tools for transmission chain analysis [275]. For instance, Harris *et al* 2013 [299], used whole genome sequencing within a hospital-based outbreak of MRSA to show how, despite stringent infections control methods, the outbreak persisted through carriage in a staff member.

Sequencing of isolates from an outbreak will often be focused on very closely related cells, with sometimes only a few SNP differences between them [274]. In this case, phylogenies formed directly from sequence data are likely to be representative of the true relationship between strains. However, in studies focusing on community spread, or even the global spread of a pathogen, a phylogeny formed from sequence data is unlikely to be informative of the relationship between isolates. This is due to the often high rates of recombination across bacterial genomes [300]. Instead, categorising these isolates into closely related strains allow for further detailed analysis. MLST typing is again a popular approach to separating these larger sequencing datasets into strains [285]. However, its limited resolution may reduce its power to detect strain structure in certain populations. Extensions of this approach include core-genome MLST (cgMLST) and whole-genome

MLST (wgMLST), which use a wider range of sequence data to inform their clustering and appear to offer higher resolution in the determination of strains [301, 302]. Other approaches which utilise the WGS of isolates are also growing in popularity, such as hierBAPs and PopPUNK [298, 303]. These approaches promise to be more scalable to the increasing amount of WGS data from global surveillance programs.

1.2.3.2 Phylodynamics

Bacterial pathogens tend to be rapidly evolving organisms, and this evolution can occur at the same timescale as epidemiological processes, such as host-to-host transmission [304]. This insight sits at the heart of phylodynamics, which seeks to use sequence data and clinical data to explore the transmission and expansion dynamics of pathogenic lineages [305]. Phylodynamic approaches rely on estimating a molecular clock from sequence data, with phylogenies subsequently adapted to be time-calibrated [306]. This allows for temporal estimates of transmission dynamics, and estimates, typically using a coalescent model, of the effective population size, N_e, of a population [304]. These methods have been successfully employed for the last decade in a range of viral pathogens. From estimating the emergence time of HIV in Africa to the 1920s [307], to tracking the spread of the Zika and Ebola epidemics [212, 308], and of course, their ongoing use in teasing apart the dynamics of SARS-CoV2 spread [309].

With the increase of sequencing data through NGS and LRS methods, and tools to detect and filter out recombination from bacterial phylogenies, these phylodynamic methods are becoming applicable to bacterial pathogens [279]. The adaptation of previous phylogeographic methods, used to estimate the location of pathogen emergence as well as the date, to other discrete traits associated with bacterial pathogens is an exciting avenue of research [304, 306, 310]. For instance, these models can be applied to estimate the date and origin of AMR for different bacterial pathogens [311, 312]. While recent approaches have also sought to incorporate a covariate, such as antibiotic consumption rates, to explain the expansion of AMR lineages [313].

1.2.3.3 Insights into bacterial population structure from sequencing

Early models of bacterial evolution and population structure were typically based on the asexual clonal reproduction of cells, with point mutations introducing diversity into a population, but little recombination present [282]. Indeed, early genetic studies using MLEE often predicted low recombination rates in species [314–317]. In contrast to these studies, early efforts of directly sequencing the nucleotides of important genes, such as the *pbp* loci, surface antigens and metabolic loci, depicted the mosaic nature of these loci [318–320]. This, coupled with MLEE studies with a wider sample of isolates, forced a reevaluation of the clonal model of evolution and emphasised a greater role for recombination in the evolution of bacteria [282, 321].

The early use of whole genome sequences and the advent of comparative genomics, further highlighted the important role of recombination in bacterial evolution [201]. Some species were seen to be almost wholly clonal, such as *M. tuberculosis* and *Yersinia pestis*, through multi-locus sequence type (MLST) methods [322–324]. With further WGS comparisons also confirming this clonal structure [325,326]. However, upon the sequencing of the first three *E. coli* genomes, there was extensive evidence for HGT and wide between-strain diversity [201]. Further population level sequencing has led to the concept of the bacterial pan-genome [327, 328].

In the pan-genome representation, genes are split into being either core or accessory in nature. This split follows the typical U-shaped curve for the frequency distribution of genes (Figure 1.3) [330]. Core genes are present in all, or almost all strains, with the cut-off for core genes varying, typically being between 0.95 to 1.00 in terms of proportion of genomes a gene is present within [331, 332]. Accessory genes are those outside of the core. These can be split further into shell genes, those present at intermediate frequencies, corresponding to the trough of the U distribution, and cloud genes, those present in very few isolates representing the first peak of the U-shaped distribution [332]. A pan-genome can also be open in nature, with each additional genome sequenced revealing more previously uncharacterized genes, or closed, where new genomes do not increase the number of novel genes [333]. The relationship between number of genomes sequenced and the number of genes in the pan-genome follows a power law, $n = \kappa N^{\gamma}$,

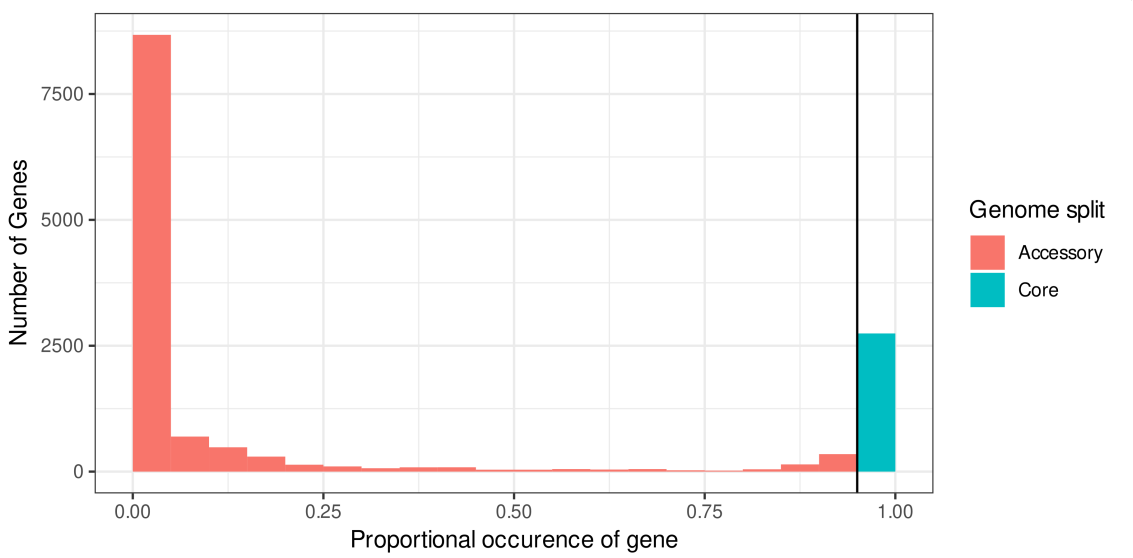

Figure 1.3: Frequency distribution of genes within GC2 lineage of *A. baumannii*. The pan-genome of the 5,092 isolate GC2 strain investigated in chapter 5 was calculated using panaroo [329]. In total 14,173 unique genes were detected in the population. The vertical line occurs at a proportion of 0.95 on the x axis, splitting the core genes, which occur in nearly all the genomes, from the accessory genes.

where the exponent $\gamma > 0$ indicates an open pan-genome [333].

Species with a closed pan-genome tend to be host-restricted, or live in an isolated niche with few interactions with other bacteria [327,334]. These include largely clonal bacteria such as *M. tuberculosis*, *Staphylococcus lugudunensis* and *Chlamydia trachomatis* [327, 335, 336]. In contrast, species with an open pan-genome tend to be able to colonize multiple environments and are known to be recombinogenic or frequently exchange genes through HGT [327, 328]. It is thought the ecology of strains drives their exposure to potential donors of DNA, which in turn affects the rate at which they can gain genes through HGT, which then in turn drives whether they have an open pan-genome (high rates of HGT), or a closed pangenome (low rates of HGT) [337]. Estimates of the total pan-genome size and the core genome of a species are highly sensitive, both to the method used to estimate this and the population of isolates sampled [329, 338]. In general, though, the size of pan-genomes uncovered through bacterial sequencing, particularly with regards to accessory genes, has been surprising and serves to highlight the pervasive nature of HGT in bacterial populations [336].

1.3 Bacterial horizontal gene transfer

In general, resistance genes can enter into a population through either *de novo* mutation or HGT [339]. Mutations are the ultimate source of all resistance genes, with fluoroquinolone resistance, as described above, generally precipitated by two amino acid changes in *parC* in pneumococci [340, 341]. In the main though, HGT offers a more rapid route to adaptation via potential import of a whole host of resistance mechanisms and other potentially adaptive sequence [342, 343]. Within bacteria HGT generally occurs via three different processes: transduction, transformation and conjugation [344]. In recent years, other ”non-canonical” methods for HGT have also been discovered: movement by membrane vesicles [345, 346], sequence movement through nanotubules in the cell membrance [347], and via phage-like gene transfer agents (GTAs) [343, 348] (Figure 1.4). Once foreign DNA has been imported into a cell, there are numerous ways for it to either incorporate into the host chromosome, or reside outside the chromosome (Figure 1.4).

1.3.1 Phage and Transduction

1.3.1.1 Mechanisms of Transduction

MGEs encode their own transportation machinery which can allow them to transfer between cells and within the host chromosome [349]. These selfish elements will spread themselves even if its is deleterious to host cells [350]. Phages are one such selfish element, and are the most abundant life-form on the plant, with an estimated 10^{31} present [351, 352]. They are viruses which infect and kill bacterial cells [353]. Phages can transfer DNA between bacterial cells via transduction (Figure 1.4). Transduction begins with the initial infection of a cell by a phage, which bind to specific receptors on the surface of bacterial cells through their receptor binding proteins (RBPs) [354]. The phage T5 for instance uses the pb5 RBP to bind irreversibly to the fhuA outer-membrane protein of *E. coli* [355, 356]. Upon adsorption onto the cell membrane, phage inject their DNA into the host cell from where it can enter into one of two separate life cycles: lysogenic (also known as temperate) or lytic [357].

In the lysogenic cycle, injected phage DNA will insert within the host chromosome to become a prophage [358]. Prophage sequence typically inserts via site-specific recombi-

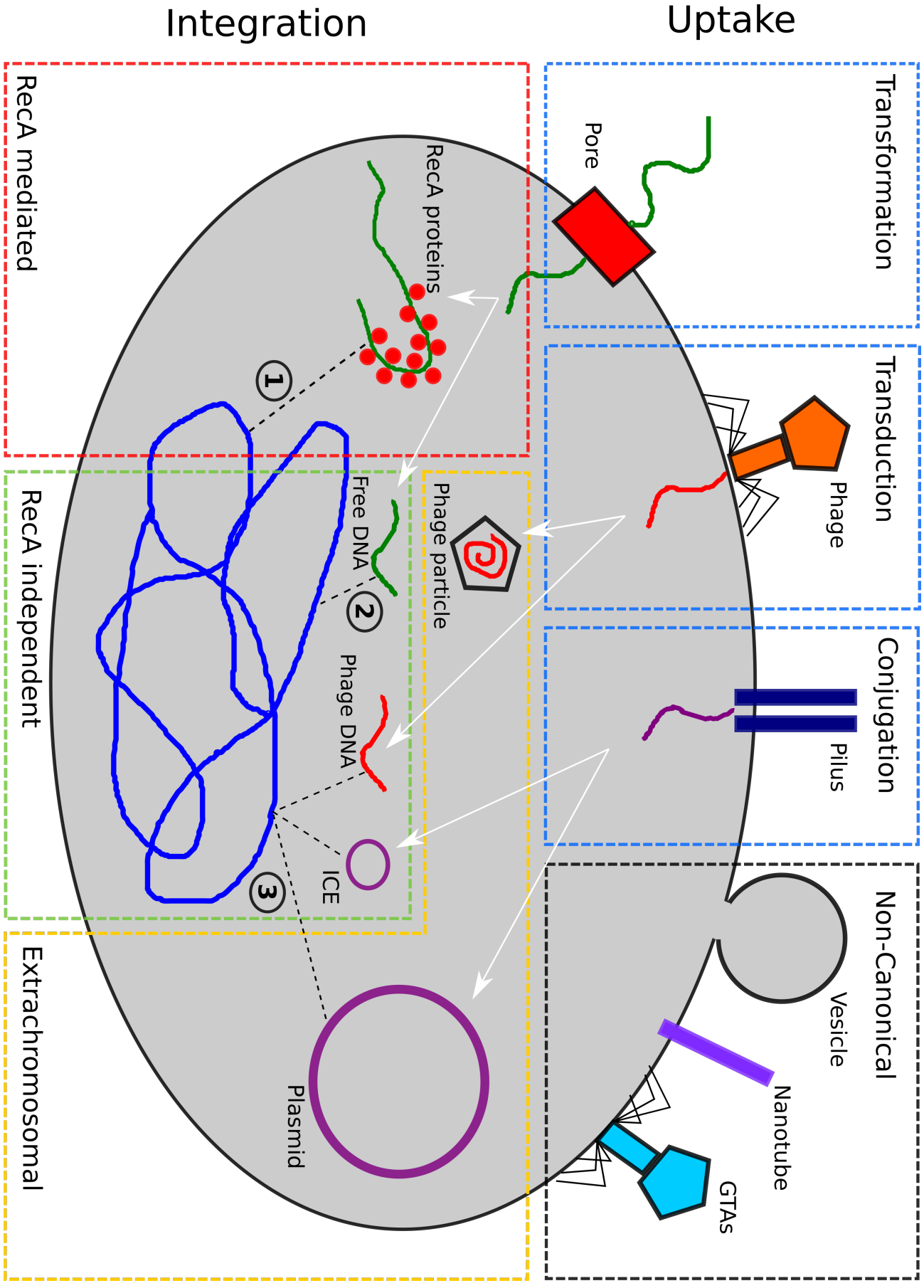

Figure 1.4: Summary of the different HGT methods, split by DNA uptake and integration steps. The three canonical uptake mechanisms are highlighted along the top of the cell, transformation, transduction and conjugation. The different DNA integration steps, be they chromosomal or extrachromosomal are outlined within the cell. **(1)** RecA-mediated recombination, whereby ssDNA uptaken by the cell is bound by RecA before integrating into the chromosome by homologous recombination. **(2)** RecA-independent recombination, this occurs through illegitimate recombination via microhomologies or non-homologous end joining. **(3)** Site-specific recombination, this occurs through integrases, recombinases or transposases that are encoded by MGEs. Figure is adapted from Arnold *et al* 2022 [343]

nation, which can be mediated either by recombinases, such as the tyrosine recombinase used by bacteriophage λ [359], or transposases, such as the MuA DDE transposase used by bacteriophage Mu [360–362]. Once inserted into the host chromosome the prophage lies dormant, replicating with the host cell [363]. The phage can be become induced however, through circumstances that activate the bacterial DNA repair response, such as DNA damage, antibiotic treatment and oxidative stress [364]. This results in the phage switching to the lytic cycle, whereby the phage hijacks the host cell machinery to create more phage capsids, filling these with sequence, before an overburdened host cell goes through cell death and ruptures to release the newly formed phages [357].

In the lytic cycle, the phage DNA goes through episomal circularization followed by theta and then rolling circle replication, before it is packaged into a capsid via a phage terminase enzyme cleaving the DNA at a specific site, either *cos* or *pac* [365]. Transduction occurs when external host DNA can erroneously become packaged in a phage capsid instead [365]. There are three types of transduction: generalized, specific, and lateral. Generalized transduction was the first case of transduction to be discovered, identified in 1952 in *Salmonella* [366]. It occurs when the phage terminase recognises a pseudo-*pac* site in the host genome, packaging this instead of phage DNA [365]. Specialized transduction occurs through improper excision of the prophage from the host DNA, creating a hybrid DNA strand of phage and host DNA [367]. Finally lateral transduction has been discovered most recently in *S. aureus* [368]. Unlike the previous two mechanisms, this appears to be a natural part of the phage lifecycle [365]. DNA packaging begins before a prophage has excised from the host genome, with packing starting at a *pac* locus within the prophage and continuing across the host genome for up to seven headfuls, with this mechanisms capable of transferring much more host DNA than the previous transduction methods [365].

1.3.1.2 DNA transferred by transduction

Prophage have been found to be abundant among sequenced bacterial genomes, with almost half containing at least one prophage [369]. In the prophage form, the interests of phage and hosts often line-up, as the prophage require their hosts to replicate. Hence many phages can encode potentially adaptive genes for their hosts [370]. For instance,

phage are well known in transferring virulence genes among different bacterial species [357, 371]. The important *Vibrio cholerae* virulence factor, cholera toxin (CT), for example was found to be encoded by the CTXϕ phage [372]. In *E. coli* prophage are also seen to upregulate the Shiga toxin genes, with cell lysis from phage infection releasing the toxin, which benefits the closely related neighbour cells [370, 373].

Phage machinery can also be hijacked by other MGEs, which utilize the phage transfer mechanism for the relatively stable, given the sequence is protected within a viral capsid, movement of sequence between cells [365]. Within *S. aureus*, pathogenicity islands (PIs), which can contain superantigens and other virulence factors [374], normally reside within the host chromosome, with their excisions repressed by the master repressor Stl [375]. Helper phages encode anti-repressor proteins which allow the PIs to excise. These PIs then produce new compound terminases from phage terminases and their own, which recognize PI *pac* loci, allowing them to hijack phage transfer machinery for their own ends [365, 376].

However, while phage have been associated with pathogenicity islands and virulence factors, their role in the dissemination of AMR genes spread is still uncertain. Studies *in vitro* have observed transduction spreading AMR genes [377], however the frequency of phages transferring AMR genes appears to be 10^3 times less than the spread via conjugation [378], suggesting transduction may only play a minimal role in AMR spread. Additionally, although virome based studies have found a range of AMR genes in phage sequences [379], these have used the poorly curated Antibiotic Resistance genes Database (ARDB) [380]. A recent study which used the CARD database instead, found that phage genomes rarely carried AMR genes [381].

1.3.1.3 Phage-host conflict

Bacteria and phages are in an evolutionary arms race that began shortly after the emergence of bacteria billions of years ago [382]. As such, bacteria have evolved multiple defence mechanisms, which include: restriction modification (R-M) systems, CRISPR-Cas systems, and Abi abortive infection systems among others [383]. Phage in turn have co-evolved anti-defence systems including: injecting proteins into a host cell that mask the phage DNA to host R-M systems [384, 385], encoding their own CRISPR-Cas systems

to inactivate host defence mechanisms encoded on a pathogencity island [384, 386, 387], and encoding repressors that inhibit the proteases which normally degrade antitoxins and cause abortive cell death [384, 388]. Given the small size of phage genomes, these defence systems are typically located in hotspots of genetic diversity, with a high-turnover of material from a large pool of loci [389]. This means phage tend to be very specific in the strains of hosts that they can successfully infect [390].

Once a phage is within a cell during the lysogenic cycle, its interests and the host's interests align. As such studies have shown that the insertion site of prophages in *Escherichia* and *Salmonella* isolates have evolved to minimize host disruption [370]. With these prophages inserting within intergenic and conserved regions of the host genomes, which are less likely to be deleterious to host cell function [391]. However, transposable phages, which can randomly move around within the host chromosome, can cause genome rearrangement and gene knock-outs [370]. The Mu phage, for instance, replicates by transposition around the host chromosome, which can cause the deletion of genes and chromosome inversion [392], while in group A streptococcus (*Streptococcus pyogenes*) phage cause large scale genome rearrangement [367, 393]. Phage insertions have also been observed to knock-out competence machinery of bacteria [350] and the CRISPR2 loci involved in defence against phage-like sequences [350, 394].

1.3.2 Conjugation and conjugative elements

1.3.2.1 Mechanism of Conjugation

Conjugation is a process that transfers DNA between cells via direct contact, with this generally being mediated by a conjugative pilus [343](Figure 1.4). Conjugation requires sophisticated machinery for this DNA transfer, involving: a relaxase (MOB), a type IV secretion system (T4SS) and a type IV coupling protein (T4CP) [395]. To undergo conjugation, cells must first form a mating pair, with the T4SS essential for this, hence its genes are known as mating pair formation (MPF) genes [396]. T4SSs are the most common secretion systems found in nature, and come in many forms, the simplest of which is encoded by MPF_T systems consisting of 11 proteins [396, 397]. In this system, the pilus used to bind a receptor and a donor cell is first synthesised within the donor cell by the VirB2 and VirB5 proteins of the MPF_T T4SS [396–398]. The translocation appara-

tus, which the DNA sequence of a conjugative element moves through between cells, is created by five proteins: VirB3 and virB6-9 [397, 399].

Once the MPF has been formed with the pilus and transport channel, the MOB relaxase nicks the circularized episomal donor DNA at the *nic* site within the origin of transfer (oriT) [395, 400]. The relaxase then binds to this cleaved, now single-stranded DNA (ssDNA), and the T4CP then attaches to the relaxase [396]. This T4CP MOB complex is then transported, along with the ssDNA of the donor, through the transport chanel formed by the T4SS to the recipient cell [396]. Once all of the DNA is within the recipient cell the relaxase ligates the ends of the ssDNA into a circular molecule [395]. Now the sequence exists in a circular ssDNA form in both donor and recipient and is then complemented into double-stranded DNA (dsDNA) by host machinery [401]. Conjugation represents a stable mechanism of DNA transfer, with the connecting pilus protecting sequence from the external environment [395]. Indeed very large DNA segments, in some cases several megabases in size, have been seen to be transferred between diverse bacterial hosts [402, 403].

1.3.2.2 Plasmids

The term plasmid was first coined by Lederberg in the 1950s as a ”generic term for any extra-chromosomal hereditary determinant” [403–405]. This captures the fundamental nature of plasmids as DNA molecules existing episomally; plasmids can be further divided into conjugative, such that they encode their own conjugation machinery, and non-conjugative, whereby they lack the MPF genes for a T4SS and rely on other conjugative elements to transfer between cells [396]. Plasmids are considered to be parasitic of host cells, with fitness costs associated with their uptake and accessory gene expression [406]. They can also be lost by host cells during cell division, with a further common split in plasmids determined by their approach to counter this [405]. High copy-number plasmids tend to be smaller and produce tens of copies of themselves per cell in order to ensure their stable inheritance during division. Low copy-number plasmids are larger and tend to encode mechanisms such as post-segregational killing in order to remain within a host cell population [405, 407]. Plasmids can also be split into families based on their incompatibility (Inc), which is defined as the inability of two related plasmids to stably co-infect the same

host cell, due to competition between similar partition and replication systems [408, 409].

Plasmids are considered to be modular in nature, with discrete clustering of genes into functional groups, such as the backbone of plasmid-mobility genes and accessory genes [403]. These accessory genes can be used to evade host defence mechanisms, with IncN family plasmids known to encode anti-restriction proteins to block their removal upon entry by host R-M systems for instance [410, 411]. The accessory genes though, can also act to expand a host's niche, most famously with the production of AMR genes [405]. There have been several epidemic plasmids, which are identical to each other but found in a broad range of host taxa, often encoding ESBLs such as CTX-M1 and CTX-M14 and carbapenemases [412–414]. These resistance genes also tend to have higher mutation rates in plasmids, likely as a result of their higher copy number offering a greater chance of mutations occuring in any one copy [415]. The TEM-1 ESBL for instance, was found to evolve additional high-level resistance to ceftazidime (a third generation cephalosporin) when present on a high copy-number plasmid, whereas when TEM-1 was on the chromosome this resistance did not evolve [405, 416].

While plasmids do not tend to integrate into the host chromosome, they can still be hotspots for recombination, with around 40% of sequenced plasmids mosaic in nature [417]. This can affect the backbone of plasmids [418], as well as the accessory genes, playing a major role in the evolution of AMR associated plasmids [405]. Recombination can produce MDR plasmids, with plasmids in India encoding the NDM β-lactamase seen to recombine with plasmids encoding the colistin resistance gene *mcr1*, for instance [419].

1.3.2.3 Integrative conjugative elements

Integrative conjugative elements (ICE) are similar to plasmids in that they they encode their own T4SS that mediates their transfer via conjugation to other cells, however they can also integrate within a host's chromosome [420, 421] (Figure 1.4). ICE are thought to be closely related to conjugative plasmids, with some ICE encoding the partition and replication machinery used by plasmids, which may allow for interconversion between the two MGE types, depending on host conditions [422, 423]. ICE integration is typically mediated by tyrosine recombinases, although other serine recombinases and DDE-transposases are encoded by ICE, whilst they can also integrate through homologous

recombination [395, 422, 424]. ICE are incredibly widespread taxonomically among bacteria, although they can also cause a fitness cost to hosts, either through expression of their gene content or through disruption of host genes upon integration [425]. The activation of the ICE conjugative genes and their excision from a host is also generally detrimental to the bacterium, with ICE tending to be kept as quiescent elements in the host genome [426]. There are a variety of signals that can induce ICE to excise from a host, including the cellular SOS response, stationary phase signals and ICE phenotype dependent induction [421, 427, 428].

The sizes of ICE range from ̃20 kb to greater than 500 kb and they are modular in their layout, with conjugation and integration genes typically conserved, while their cargo genes can vary in number and functions [421,429]. In general cargo gene regions within ICE tend to be hotspots for diversity. A comparative study of the SXT/R391 family of ICEs, originally found in *V. cholerae*, found five hotspot regions that incorporated anywhere between 676 bp to 29,210 bp of variable sequence [430]. Interestingly these regions tended to be mosaic in nature among the 13 ICEs studied, with overlapping distributions of these cargo genes among the family indicating extensive recombination among the ICEs. In contrast to the diversity present among cargo genes, all ICEs studied shared a core set of 52 genes involved in ICE excision, integration and conjugation [430]. Cargo genes in discovered ICE tend to have marked effects on the phenotype, such as providing resistance to a certain antibiotic or enabling utilization of an alternative carbon source [312, 421, 431]. Comparative studies with plasmids depict ICE being less recombinogenic, likely a result of their lower copy number, while also tending to encode more metabolism related cargo genes than plasmids which favour AMR genes [423]. Nevertheless, ICE have also been linked to AMR pandemic strains of *V. cholerae*, *P. aeruginosa* and the pneumococcus [432–435].

Recombination within ICE elements can create both large compound elements and smaller elements lacking some of the essential ICE genes. In *Streptococcus thermophilus* for instance, smaller, truncated, non-mobile versions of the 35 Kb ICE*St1* have been observed at the shared *fda* insertion loci [436, 437]. These fragments are also targets for recombination, with evidence of the gain of efflux pumps and plasmid sequence around

the *fda* locus [436]. Large MDR ICEs are also formed by the insertion of ICEs within each other. The chloramphenicol and tetracycline resistant Tn*5253* element for instance is formed from the insertion of Tn*5251*, which carries the *tet(M)* tetracycline efflux pump gene and is a member of the wider Tn*916* family of ICEs, into the Tn*5252* element which contains the Ω*cat* element encoding chloramphenicol resistance [438, 439]. A Tn*5253*-like element, ICE*Spn*23FST81, is widely maintained in the MDR PMEN1 lineage of the pneumococcus [142, 440].

1.3.3 Transformation and homologous recombination

1.3.3.1 Competence regulation and DNA uptake

Unlike conjugation and transduction described above, which are largely driven by MGEs, DNA uptake through transformation is regulated by the recipient cell and its activation of the competence machinery [441, 442]. Natural competence is a physiological state, considered to be a developmental program in bacteria, that enables cells to uptake DNA from the environment [442, 443]. Only around 80 different bacterial species have been found to develop natural competence in laboratory conditions, although it appears the machinery for competence is widespread among bacteria [343, 443, 444]. Apart from *Helicobacter pylori* and *Neisseria* species, which are constitutively competent, most species exert tight control over their competence machinery [444].

This control initially involves response to an extracellular signalling mechanism [442]. In the pneumococcus this external signal is the 17-residue competence stimulating peptide (CSP), which is formed by the ComC protein, encoded by the *comCDE* operon, being exported and cleaved by the ComAB exporter [444–446]. CSP expression can be induced by environmental cues, such as antibiotic exposure and alterations in pH, and cellular cues, such as an increased rate of translational errors [442, 447–449]. This has led to CSP being described as an alarmone, involved in cell-to-cell signalling [442, 448]. CSP binds to the membrane-embedded histidine kinase ComD, which leads to autophosphorylation and subsequent transfer of a phosphoryl group to the response regulator ComE [448, 450]. ComE then upregulates the transcription of alternative sigma factor SigX (encoded by two copies of *comX*) which is the central regulator for the remaining *com* genes required for transformation [442]. ComE also increases the transcription of

the *comABCDE* operon involved in CSP formation and signalling, creating an autocatalytic feedback loop [442, 444, 451]. In other species, alternative external factors lead to cells becoming competent. In *V. cholerae* for instance, chitin, sensed through ChiS and TfoS, causes an upregulation of the TfoX regulator of competence, while in *A. baumannii* human serum albumin has been seen to enhance the expression of competence genes [442, 452, 453].

In the pneumococcus, the expression of CSP and the ComE upregulation of the *comABCDE* operon is known as the early stage of competence, while SigX controls the late stage, where the DNA uptake machinery is expressed, among other pathways [451]. This DNA uptake machinery includes a type IV pilus, extended from the cell membrane by the ATPase ComGA and consisting of ComGC subunits, which binds to external dsDNA segments [443, 454, 455]. Upon binding of dsDNA, the pilus retracts through the outer peptidoglycan cell wall, binding the DNA to the cell membrane protein comEA, the first component of the DNA translocation system [456]. In the pneumococcus the EndA nuclease then degrades one strand of the dsDNA, with the ssDNA then passing through the comEC pore with a 3' to 5' polarity, with this translocation into the cytosol powered by the ComFA ATPase [444, 455, 456].

As well as the DNA uptake machinery, SigX also upregulates genes involved in the fratricide of neighbouring non-competent pneumococci [442]. Antimicrobial bacteriocin production is upregulated for instance, with the inducer peptide BlpC, similar to CSP, being exported by the ComAB exporter and acting as a signal to other competent cells to also produce bacteriocins, such as CibAB, and their cognate immunity proteins [457, 458]. Other fratricidal elements upregulated include the autolysin LytA, the lysozyme LytC and the amidase CbpD which has muralytic activity and can cause cell rupture [455, 459]. Competent cells become immune to CbpD by arresting their growth through the expression of the late competence gene *comM* [455]. CbpD binds to the septal region of the dividing bacterial cell wall, thus the arresting of growth prevents the septal region from forming and avoids cell rupture from CbpD exposure [460, 461]. This competence-linked upregulation of fratricide proteins, which kill neighbouring pneumococci, is thought to increase the efficiency of DNA transfer by transformation. This is as there is a larger

pool of external DNA now available [442].

1.3.3.2 Incorporation of DNA via homologous recombination

Once the external ssDNA enters into a host pneumococcus through the ComEC channel, it is bound by the recombination mediator protein DprA [462]. DprA stabilises the ssDNA, protecting it from internal nucleases, and loads the recombinase RecA onto the ssDNA [462, 463]. DprA also interacts with ComE, blocking its upregulation of the early *comABCDE* genes, meaning cells are only competent for a short period of time [464,465]. RecA, once bound onto the ssDNA, polymerizes along the strand and promotes a similarity search within the host chromosome [444] (Figure 1.4). Once a similar region to the ssDNA has been found in the host chromosome a heteroduplex of donor and host DNA is formed [444]. This can be fully identical, with the donor DNA matching the host chromosome along its length, or can encompass heterologous sequence flanked by regions of identity to the host chromosome [444]. The minimum efficient processing segment (MEPS), the shortest length of flanking identity required for efficient recombination of heterologous sequence, is estimated at 27bp for pneumococci and *E.coli* [466, 467].

Any sequence that is subsequently bound into the host chromosome can still be targeted by R-M systems, if it is not appropriately methylated [444]. Restriction enzymes in R-M systems can target the newly integrated DNA, degrading the sequence and in the process killing the host cell [468]. To mitigate this, a late competence gene *dpnA* encodes a methylase, a part of the wider DpnII R-M system, which methylates the donor ssDNA before its integration by RecA [469, 470]. This prevents degradation of the sequence by the *Dpn*II restrictase once integrated into the chromosome. This allows for genomically divergent heterologous sequence to be imported into a cell by transformation and homologous recombination [444].

1.3.3.3 The evolution of transformation

Becoming competent incurs a considerable cost to bacterial cells, with the upregulation of the more than 100 genes involved in competence being energetically costly, while cells also arrest their growth during this state, reducing their fitness compared to non-competent cells [471]. Why then would cells engage in this potentially deleterious de-

velopmental program? There have been numerous theories put forward to explain this. These generally encompass: external DNA acting as a nutrient source, DNA uptake to repair double stranded DNA breaks, external DNA as an adaptive evolutionary source, and most recently external DNA integration to remove deleterious MGEs [441, 444, 471, 472].

The DNA as a nutrient source hypothesis has been supported by observations that *H. influenzae* and *Bacillus subtilis* appear to induce competence when purine pools are deplenished [473, 474]. Additionally it has been argued that external DNA may be of too poor a quality for recombination into the host genome, so its use as a source of carbon and other nutrients may be more advantageous [441, 475, 476]. However, other species, such as the pneumococcus, do not appear to induce competence in times of starvation [477]. Furthermore, the expression of the competence pilus and DNA uptake machinery is expensive for cells, plus many species only import ssDNA, surely half the nutritional value of dsDNA, which appears to be a counter-intuitive strategy for enhanced nutrient uptake [444].

The link between competence and fratricide in the pneumococcus is seen as evidence that DNA uptake might be selected for in order to provide genetic material to repair double stranded DNA breaks [441]. Similarly, *H. influenzae* and *N. gonorrhoeae* discriminate uptake towards self-DNA during competence, while *V. cholerae* also produces bacteriocins that can mediate kin-discriminated killing [441, 478]. Here, the uptake of DNA from closely related cells allows for a greater chance of similarity and possible repair of DNA damage. However, while there is some lab evidence for transformation mitigating the effects of UV DNA damage in *B. subtilis*, this was not found in *H. influenzae*, *H. pylori* or *Legionella pneumophila* [350, 479].

Transformation does appear to have driven some major adaptive changes in bacteria. In the pneumococcus for instance, mosaic AMR genes have been formed through transformation and homologous recombination, while recombination around the capsule *cps* locus has led to serotype switching and vaccine escape [140, 142]. In these cases, recombination has eliminated any clonal interference between adaptive mutations, allowing for these mutations to exist in one genome [480]. This could be evidence for the evolution of transformation and recombination as a means to increase adaptive rates. Indeed, sim-

ulation experiments have shown that recombining populations tend to be fitter [481, 482], while lab studies in *E. coli* show recombination speeds up adaptation [343,483,484]. However, as an explanation for the initial evolution of transformation and recombination, this hypothesis relies on the discredited group-selection argument [350]. Transformation is as likely to uptake deleterious as well as advantageous DNA, indeed some argue perhaps it is more likely to uptake deleterious DNA due to this DNA coming from dead cells that are likely not adapted to the local environment [475, 476].

The external DNA uptaken through transformation tends to be short in nature, typically following a geometric length distribution [485,486]. In the pneumococcus the mean size of imported DNA is around 2.3 kb, which is much shorter than the typical MGE inserted within a host chromosome [466, 485, 487]. This observation, coupled with the detection of many successful MGE insertions disrupting host competence machinery [472,488], has led to a new theory that transformation has evolved as a means to delete selfish MGEs from a host genome [350]. Croucher *et al* 2016 used a mathematical model, with transient competence, asymmetrical DNA uptake and deleterious MGE presence, to highlight the fitness benefit of transformation [350]. The authors also analysed pneumococcal genomes and saw transformable isolates had significantly fewer prophage contained within, this analysis was replicated on a smaller scale by Ambur *et al* 2016 [471]. However, much of this theory relies on observations in only a select number of pathogenic species, further work is needed to explore the length distribution of recombination events in other transformable species to ensure this length asymmetry is universal [472].

These four explanations for the evolution of transformation are not necessarily mutually exclusive, indeed it could be argued that the theory of transformation as a means of deleting deleterious MGEs is a sub-component of the wider theory of transformation evolving as a means to import diversity and speed up adaptation [442].

1.4 Summary

In this chapter then we have seen then how AMR has evolved almost in tandem with the growing spread of antibiotics. Bacteria have been targeted in a range of different ways, but have themselves evolved a multitude of different resistance mechanims. How resistance

evolves within bacterial populations is still unclear, as is the true scale of deaths from AMR infections. While methods for predicting resistance from sequence data *in silico* still vary in their accuracy across different species and drug combinations.

HGT mechanisms though, have played an outsized role in the dissemination of key adaptive changes in bacteria, such as AMR, pathogencity alterations and host immune response evasion. Often, the MGEs that confer these adaptations are deleterious to hosts and their insertion within the chromosome leads to intra-genomic conflict that may explain the evolution of these different mechanisms. We have also seen how the improvement in sequencing technologies in the 50 years has greatly expanded our knowledge of bacterial biology. These new techniques have highlighted the role of HGT in bacterial evolution, while also shifting our understanding of the bacterial genome to a more open pan-genome view. This has important clinical implications for the dissemination of resistance, and other virulence-associated genes, between pathogenic species. The era of bacterial genomic epidemiology is also beginning, with sequencing data useful in clinical settings for the understanding of outbreaks, and also playing a role in surveillance and the monitoring of successful clones. Finally the nascent field of bacterial phylodynamics can offer further insight into how clinical decisions impact the dissemination of AMR genes.

Chapter 2

Extending methods to detect recombination in bacterial genomes

Acknowledgements

The Gubbins algorithm was initially published in 2015 in Croucher *et al* [489]. Work on improving and updating the algorithm was carried out jointly by myself and my supervisor Dr. Nicholas Croucher. The sequence simulation code was initially written by Dr. Croucher for the release of Gubbins in 2015. This was adapted for the current chapter by myself. All benchmarking analyses and figures have been created by myself.

Summary

Detecting recombination in bacterial genomes is key to understanding the relationships between isolates. There are numerous approaches to this. One of the most popular, Gubbins, relies on detecting an elevated number of mutations in sequence, indicative of a recombination event. This chapter describes improvements made to the Gubbins algorithm. Thorough benchmarking of the algorithm is performed, including comparing Gubbins to another popular, but slightly different approach, ClonalFrameML. I find that Gubbins is the most accurate method in determining the true relationships between isolates. Improvements in the efficiency of the algorithm are identified that will allow for its continual use with the ever increasing amount of sequencing data.

2.1 Introduction

Bacterial recombination is an often broadly used term that has come to encompass all manner of genetic exchange between cells, be they closely or distantly related. In this work I will consider recombination as a process through which exogenous DNA is integrated into the host chromosome. In this regard, recombination is typically mediated by the three main mechanisms of horizontal gene transfer (HGT): transformation, transduction and conjugation which import foreign DNA into a cell [343]. Once foreign DNA is within a host cell, it can then recombine into the chromosome, or in some cases plasmid DNA, through a variety of homologous, non-homologous and site-specific recombination methods [343, 490]. Initially, recombination was not thought to be common among pathogenic bacteria, with successful clones taken to be indicative of point mutations being the main force driving genomic variation in these species [274, 314]. With the advent of genetic sequencing though, this belief has been overturned, with ample evidence for recombination playing a key role in the emergence of successful MDR lineages across the bacterial domain [142, 491, 492].

In broad terms, recombination events can be split into exchanges and imports [493]. These events are defined with respect to the phylogeny of a sample. Exchanges occur between closely related isolates within the same phylogeny. These exchange events can easily produce homoplasies, which confound the phylogenetic reconstruction of lineages [274, 494]. Imports are typically rarer and represent the integration of highly divergent sequence from an external population. These import events can be quite large and can lead to new lineage formation [495]. While imports tend not to introduce homoplasies, these events do affect branch lengths, which may hamper efforts to identify evidence of a molecular clock among sequences [494]. The estimation of a clock rate is often vital in outbreak analysis [496], while it can also help answer questions around the link between clinical interventions and bacterial population size [497]. In general, both recombination types violate core assumptions of nucleotide substitution models, as SNPs present in an alignment may not be independent of one other if brought into the genome by the same recombination event [498]. Thus, the challenge in reconstructing the evolutionary history of recombinogenic lineages is separating the informative vertically inherited point

mutations from those mutations introduced horizontally by recombination, which, while indicative of the selection pressures an isolate is experiencing, are much less informative of the relationships between taxa.

In this chapter then, I will describe my efforts to improve perhaps the most widely used tool for detecting recombination in bacterial lineages, Gubbins [489]. I will start by outlining the array of methods available for detecting recombination events in sequences and then describe the main Gubbins algorithm, before detailing the improvements made to Gubbins and the results of benchmarking the new package.

2.1.1 Methods of detecting recombination:

2.1.1.1 Detecting exchanges

The first statistical techniques developed to detect recombination were compatibility methods [499]. These methods test for phylogenetic incongruence on a site by site basis, with a site incompatible with a tree, and hence indicative of recombination, if the observed number of characters at a site (c) cannot be explained by c - 1 substitution events [499, 500]. These methods were initially applied to the protein sequences of cytochromes [501], and were also later tested on the DNA sequences of human γ-globin genes [502]. Similar methods based solely on phylogenetic incongruence between gene trees, or similar subset trees and a total sequence tree, have also been developed to detect recombination. These methods are suitable for detecting exchanges within a population. They have been particularly effective when applied to viruses, with their smaller genome sizes more amenable to tree building efforts [503–505].

An early sequence based method to detect exchanges was the Homoplasy test [506]. This measures the rate of homoplasic site mutations in a set of sequences and compares this to a null hypothesis of no recombination, where homoplasies are caused instead by repeat mutation at the same site. It can lead to overall estimates for the recombination rate in a population. However, this does require a suitable outgroup to be used in the analysis, while it can also be confounded by hypermutator sites [493, 506]. The concept of homoplasies as indicative of recombination has been used in earlier studies looking at the emergence of MRSA with WGS data [507, 508]. As well as homoplasies, recent

methods have also looked at the incorporating rates of linkage disequilibrium (LD) within a population to assess recombination rates overall for a population [509, 510]. In general LD is considered to decrease in bacterial populations as recombination rates increase, although this can be confounded by sampling strategies [510]. As such these methods often take into account multiple parameters, including homoplasies and the incompatibility of sites described above, in order to estimate overall population recombination parameters [509, 510]. A recent method from Lin & Kussell [511], uses a similar statistic, measuring the degree of correlation between substitutions differing lengths apart, to also assess recombination parameters for populations.

When trying to detect individual exchange events in a population, and identifying these within an alignment, methods using a sliding window approach have also been implemented. In these approaches alignments are scanned and metrics calculated at each scanning step to assess the likely start and end points of a recombination event. Building from the phylogenetic incongruence tests described above, these tools initially tested windows for differing topologies within a region through parsimony and maximum likelihood methods [512, 513]. For viral genomes, bootstrapping of trees has also been used, with references for donor sequences scanned against the length of a query sequence [514, 515]. These methods can be very effective, but only when reliable sources of donor and recombinant DNA are included in the tests. In addition sliding windows can be prone to multiple testing effects [516], while the window size can affect the resolution of breakpoint determination. Bayesian methods have instead switched to modelling the alignment as a Hidden Markov Model (HMM), with the states the different phylogenetic topologies at each site [516–519].

2.1.1.2 Detecting imports

Some of the earliest attempts to detect imports used sequence based methods, where blocks of high sequence divergence among regions of high sequence similarity are taken as indicative of imports [520, 521]. The χ^2 test developed by Maynard-Smith, which detects imports based on elevated regions of SNPs, is one such example [520]. All these methods typically rely on detecting linked sets of polymorphisms from donors that are highly diverged from recipient strains. However, these blocks of high divergence maybe

rarer in highly recombinogenic species, where LD is more likely to break apart linked polymorphic sites [493].

Extensions of sliding window approaches for detecting exchanges within alignments have also adopted the concept of a bacterial ”Clonal frame” [522]. This represents the portions of a genome at which observed variation arises from vertically inherited sequence and reflects more closely the relationships between taxa [274]. This can be inferred when the donor sequence is not present in the alignment. If recombination detection methods could identify recombinant regions of a genome and remove them, what would remain would be this clonal frame used to infer a phylogeny without the influence of recombination [274]. The ClonalFrame algorithm [523], was one of the first attempts at this sort of analysis. Within a Bayesian framework, the algorithm uses a HMM with the descendent and ancestral nucleotide sequences of a branch being the observed states, while the hidden state is whether a nucleotide was imported by a recombination event [523]. This was initially applied to MLST data and very small collections of WGS data, with the number of parameters estimated in the model leading to slow run times for larger datasets. As such the underlying framework has been adapted to a maximum likelihood approach to estimate parameters, this has been released as ClonalFrameML [524]. In ClonalFrameML the Baum-Welch Expectation-Maximisation (EM) algorithm used to estimate recombination parameters. This is now capable of running on much larger collections of WGS data within reasonable run times.

A faster method incorporating Bayesian parameter estimation for identifying recombination import events is BRATNextGen [525]. This initially splits input data into chunks of 5kb length and identifies SNPs in the chunks, which are then clustered within each chunk. These clusterings then inform a HMM used to infer whether a SNP is present from a recombinant state or a non-recombinant state [525]. This is a reasonably fast method to infer recombination events among a collection of isolates. However, rather than produce a tree representing the clonal relationships between isolates, this method produces a Proportion of Shared Ancestry (PSA) tree, grouping isolates together based on their overall sequence similarity, which is less useful for further phylodynamic analysis. Additionally, both BRATNextGen and ClonalFrame methods rely on identifying recombination events

via the presence of SNP dense regions of the alignment, akin to the early methods applied to bacterial genes described above [520, 521]. In large diverse collections however, especially of highly recombinogenic species, SNPs may appear ubiquitously across an alignment, meaning detecting recombinations via their spatial distribution becomes near impossible [300]. Therefore, these methods are best employed in the analysis of a single well-defined lineage of isolates, such as a recent outbreak, in order to allow for easier identification of SNP dense regions that are representative of a recombination event.

Given the ever increasing amount of sequence data available, comparisons between lineages, especially closely related ones, will become more and more necessary to perform. One method that attempts to tackle this problem is fastGEAR [526]. Given an input alignment fastGEAR will initially attempt to assign lineages within the collection through BAPS which is corrected for recombination events [527]. It then detects recent recombination imports that occur within a lineage, and ancestral events that occur between the ancestors of lineages in the population using a HMM approach that compares sequence both between and within a lineage, leveraging on the diversity of the input collection [526]. This focus on comparing sequences between lineages allows fastGEAR to work effectively on diverse datasets, detangling the mosaic structure of the penicillin binding protein genes for instance [526]. However, the intensive Basyesian approach of fastGEAR means it is difficult to run on larger collections of WGS data, while the initial BAPS clustering method relies on less-robust heuristics to split recombinant lineages apart.

Of all these methods for detecting recombination among bacterial genomes, perhaps the most widely used is Gubbins [489]. Broadly this method is most similar to the ClonalFrameML algorithm, using a sliding window approach to identify regions of increased SNP density and a maximum likelihood framework to assess whether these regions of elevated SNP density constitute a recombination event.

2.1.1.3 Gubbins algorithm

The Gubbins algorithm can be broadly split into three main steps: (1) inferring a phylogeny from a clonal frame, (2) ancestral state reconstruction of the sequence from the phylogeny, (3) identifying and trimming recombination events along each branch of the phylogeny. This algorithm is iterative, with a phylogeny built from the inferred clonal frame produced

from step 3.

For the first step in the algorithm, inference of the tree employs some heuristics for improved computational efficiency. Only polymorphic sites are used for the inference of the phylogeny. For the detection of polymorphic sites Gubbins uses a forerunner of the popular snp-sites tool [528]. Once these polymorphisms have been extracted a phylogeny is formed from the subset alignment. Gubbins v2.4.0 uses either FastTree v2.1.0 [529] or RAxML v8.2.12 [530] for this step, with a GTR model of substitution and a gamma distribution to account for rate heterogeneity between sites. In the first iteration, no recombinations have yet been inferred, hence the whole alignment is treated as the clonal frame. This means the creation of the initial phylogeny is often the slowest step of Gubbins and is the least accurate [489]. Hence a hybrid strategy of using the faster, but less robust, FastTree builder in the initial phylogeny creation, followed by RAxML for later iterations is often used.

Step two of the algorithm is reconstructing the ancestral sequence for each node within the phylogeny produced from step one. In the initial Gubbins release this was run using the FastML joint reconstruction method [531], however this proved to be difficult to maintain and was not open source. Hence, for Gubbins v2.4.0 RAxML marginal ancestral state reconstruction is employed, again with a GTRGAMMA model for substitution rates and heterogeneity between sites. A Marginal reconstruction is optimized to find the state at a single node with highest probability for that node [532]. Joint reconstruction on the other hand, will find the state at each node which maximises the overall probability of a phylogeny [532]. Ancestral sequence reconstruction is run on the phylogeny produced and the polymorhpism alignement, with further processing defining both the substitutions occurring along a branch and the bases that occur between substitutions. Any bases which are ambiguous in the leaf nodes of the tree are considered to be undetermined at internal nodes and are excluded from further processing.

Once sequences have been reconstructed, the scanning of branches to detect elevated areas of SNP density, taken to be indicative of recombination, is performed. As described above (Section 2.1.1.1), this method for detecting recombination events does rely on isolates within a collection being closely related to each other, such that SNPs

would appear to be evenly distributed across a chromosome in the absence of recombination. In diverse populations this assumption is violated [300], as SNP densities may be elevated simply due to the accumulation of point mutations over a longer sampling period. Additionally regions of mutational hotspots, perhaps where loci are under strong positive selection, would also appear as false positives in this approach. The null hypothesis for each branch B, $H_{0,B}$, though assumes that the SNPs which occur along the length of the branch, s_B will be distributed evenly along the chromosome. This is modelled as a binomial distribution, with the number of bases occurring in a window scan s_w, based on the size of a window in bases, w, and the mean density of base substitutions across the bases of the downstream node, $d_{0,B}$ (Equation 2.1). This is tested using a sliding window of varying size w. This size, w, is calculated based on the s_B divided by the minimum number of SNPs in a recombination, l (by default this is set to 3). This gives us the number of possible recombinations on a branch, s_R. The number of bases remaining on a branch is then divided by s_R to give w. The number of SNPs within the window on branch B, N, is at least l.

$$H_{0,B} : N \sim Bin(w, d_{0,B}) \tag{2.1}$$

For each branch on the tree, s_B windows are tested, with each window centred on a unique SNP along the length of the branch. Therefore, to correct for any multiple testing effects, the threshold P-value ($P_{Threshold}$) is a Bonferroni correct value (Equation 2.2).

$$P_{Threshold} = \frac{0.05}{s_B} \tag{2.2}$$

Contiguous regions of one or more windows where $H_{0,B}$ could be rejected, based on the $P_{Threshold}$ value with a one-tailed binomial test, represent loci in the chromosome where there is elevated SNP density and are subsequently labelled as putative recombination events, r. This block r is now proposed to conform to an alternative hypothesis $H_{1,B,r}$. This is also modelled as a binomial distribution, based on the length of the region, l_r and the density of SNPs within the region, $d_{1,B.r}$ (Equation 2.3).

$$H_{1,B,r} : N \sim Bin(l_r, d_{1,B,r}) \tag{2.3}$$

The size of these initial putative recombination events may exceed the length of the recombination event they contain however. The next stage then is trimming these putative events to delineate the likeliest regions of an insertion. The first trimming occurs by reducing the boundaries of r to the outermost SNPs it encompasses. Trimming then commences from the left side, then the right side, measuring until further trimming no longer increases the regions likelihood under $H_{l,B,r}$ relative to its likelihood under $H_{0,B}$. A final inequality must then be satisfied to reject $H_{0,B}$ (Equation 2.4). The left-hand side of this inequality represents another Bonferroni corrected threshold probability for the number of equally sized, l_f, non-overlapping windows which could be observed in the genome of length g. The right-hand side of the inequality estimates the probability, under $H_{0,B}$ that the current trimmed window, of length l_f, would contain at least the same number of SNPs, s_f, which the putative recombination event contains.

$$\frac{0.05}{g/l_f} > 1 - \sum_{i=0}^{i=s_f-1} \binom{l_f}{i} (d_{0,B})^i (1 - d_{0,B})^{l_f - i} \tag{2.4}$$

This final inequality test aims to remove blocks that appear to form with two distinct recombination events at their poles. These events would have a paucity of SNPs in their centre after the trimming process, and would then be removed from consideration as a single recombination block. These stages of scanning across the branches, trimming putative recombination blocks and testing for their significance, will produce a set of blocks for a branch. Of these blocks, the event identified with the smallest value of its probability under the $H_{0,B}$ divided by its probability under $H_{1,B,r}$ ($r_{smallest}$; Equation 2.5), the likelihood ratio, is then converted to missing data for B and all its descendent branches, with this recombination event overwriting any information from the clonal frame. The identification of recombinations is then repeated, with $d_{0,B}$ recalculated after removing individual recombination events so that SNP-dense regions of sequence don't reduce the power to detect other recombination events.

$$r_{smallest} = \frac{H_{0,B}}{H_{1,B,r}} \tag{2.5}$$

Now that recombination events have been identified and removed from the clonal

frame within the alignment, the three steps are then repeated, with FastTree v2.1.0 or RAxML v8.2.12 used to infer the phylogeny. These three steps are repeated either for a default of five iterations or until some degree of convergence in the results is detected. In the algorithm there are three tests for convergence: (i) Identical recombinations produced between two iterations, (ii) Any two trees of the iterations have a weighted Robinson-Foulds distance of 0 [533], (iii) Any two trees have a symmetric distance of 0. The default for convergence testing is the weighted Robinson-Foulds distance of any two trees from the iterations.

That describes the steps in the Gubbins algorithm, now I will detail the extensions I've made to the method

2.1.2 Extensions to the Gubbins algorithm

Gubbins has proven to be a popular method for phylogenetic analysis of bacterial lineages, with other 1,000 citations since its release in 2014, and over 25,000 downloads through the Conda package manager [489, 534]. However, Gubbins v2.4.0 relies on software which is no longer supported by developers. The default phylogeny and ancestral state reconstruction software, RAxML v8.1.2, for instance is no longer actively maintained, with the newer RAxML-NG [535] supplanting the older method. Additionally many users of Gubbins run the method simply to produce a recombination-corrected alignment, then generate a phylogeny with another method, such as the popular IQ-TREE algorithm [536–540]. Extending the number of tree models used by the algorithm would streamline the analysis pipelines for users, with Gubbins able to detect recombination events, produce a recombination-corrected alignment and a phylogeny representing the clonal frame of the lineage.

Gubbins also currently only implements a marginal reconstruction of ancestral sequence states, step two in the algorithm as described in Section 2.1.1.3. Joint reconstruction algorithms have tended to be much more computationally intensive than marginal ones, although there have been attempts to improve these algorithms [532]. Given Gubbins estimates recombination events occurring across the whole tree, and not just for a single branch, joint reconstruction methods may give more accurate predictions of where recombination events have occurred.

When analysing larger lineages, with over 5,000 isolates for instance, Gubbins is less efficient in using memory. Local runs of an *Acinetobacter baumanii* GC2 clade with over 5,000 isolates used over 200 GB of memory to complete (See Chapter 5 for further details of analysis). Heavily sampled species where lineages are harder to define, like *N. gonorrhoeae* [298], have not been able to complete Gubbins analysis. Given the increasing availability of sequence data, in order for Gubbins to remain a useful tool for phylogenetic analysis, improvements need to be made to its memory usage.

We address these limitations in a recent version of Gubbins, v3.2.0. In this version of Gubbins, the tree builders available have been extended, with users now able to choose IQ-TREE [540], RAxML-NG [535] and RapidNJ [541] for full phylogeny creation, while an option to produce a Star phylogeny in the first iteration is also available. In addition the default ancestral state reconstruction has now been switched to a joint reconstruction method. This has been adapted from pyjar [542] which itself is an implementation of the algorithm proposed in Pupko *et al* 2000 [532]. In adapting the code for Gubbins, the memory usage and the speed of the algorithm have also been optimized. Finally the recombination detection algorithm (step 3 in the whole Gubbins algorithm as discussed in Section 2.1.1.3) has also been optimized for memory usage, without changing the core process undertaken by the algorithm. All these changes have led to a faster more flexible Gubbins that can now run efficiently on lineages with over 1,000 isolates.

Now that I have described the rationale for extending Gubbins and the improvements we have made to the algorithm, I will go into more detail about how I benchmarked the new version, starting first by describing the methods used to produce simulated datasets for testing.

2.2 Methods

2.2.1 Simulating artificial sequences

To assess the accuracy of the new Gubbins version, simulated datasets were created using a forward-in-time individual based framework. This was the same framework previously deployed for the initial Gubbins release [489], but expanded upon with an option for GTR model based substitution rates and a Poisson distribution for the number of re-

combination events occurring at each time step. I chose to use this simulation method, as opposed to other forward-in-time simulators like Bacmeta [543] and FastSimBac [544], due to its ability to allow for recombination events from external populations, along with its tailored output for benchmarking against Gubbins.

The starting genotype for the forward-in-time simulation was chosen as the *S. pneumoniae* RMV4 *rpsL** Δ*tvrR* genotype (accession code: ERS1681526) [545], which is the reference isolate for the PMEN3 lineage investigated in later chapters. Recombination donors were selected from 52 different streptococcal genomes (Table 2.1), expanding on those used in the FastGEAR paper [526]. These sequences were aligned to the ancestral genotype using SKA [546], to form an input alignment to act as donor sequence throughout the simulation.

Species	**Count**
S. mitis	30
S. oralis	9
S. pneumoniae	7
S. pseudopneumoniae	6

Table 2.1: Species counts of the 52 *Streptococcus* genomes used as recombination donors in the simulation of artificial sequence.

In the simulation, as described in Croucher *et al* [489], each extant sequence acquires a single point mutation at each timestep from $t = 0$ to $t = t_{max}$, where t_{max} is the time point at which the number of extant sequences reaches a pre-determined limit, n_{seq}, which was set at 100 unless otherwise specified. The default is for this substitution to occur as predicted by a Jukes-Cantor substitution model, with equal probability for each base, irrespective of the starting base. A GTR substitution model has also been implemented in the model, this is parameterised from IQ-TREE [540] fitting the GTR model to the PMEN3 collection, which is described in the later Chapter 3. Sequences also undergo recombination, with the number of events that occur at each timestep drawn from a Poisson distribution, with a mean, μ_{rec}, that I vary for each individual simulation run. The imported sequence comes from a donor chosen randomly among the 52 streptococcal genomes input, while the start of the import is also chosen randomly along the length of the chromosome. The length of the import is modelled from the geometric distribution

of recombination lengths observed in the pneumococcus, with a per base probability of stopping the extension of sequence of 0.0005 bp^{-1}.

Through-out the simulation each sequence also has a probability of being duplicated into two initially identical sequences, p_{branch}. A value, b, is taken randomly from a uniform distribution between 0 and 1 with p_{branch} representing the threshold at which, if the value of b is less than or equal to p_{branch} a new branch is created. Higher values of p_{branch} mean sequences have less time to diverge from one another. This would be akin to either the sequences output from dense sampling of a single lineage, such that the sample would almost or fully encompass all possible diversity of a lineage, or the sequences output from a slowly evolving lineage.

In the benchmarking of Gubbins results against simulated results, recombination and substitution events which occur ancestrally and are subsequently overwritten by more recent events are removed from the accuracy calculations. When comparing the SNP classification of Gubbins to the results from the simulated data, the accuracy was assessed through the positive predictive value (PPV) score and the sensitivity score (Figure 2.11). The PPV score measures the number of true positives divided by the total number of positives produced from the reconstruction. In this case a true positive is defined as a SNP which occurs on the same branch, in the same loci and is classified as the type of SNP. SNPs can be classified into two types: either occurring within recombination events (*r* SNPs), or outside recombination events (*m* SNPs). Sensitivity measures the number of true positives divided by the total number of positives in the simulated data. This is indicative of the level of false negatives produced by the reconstructions.

2.2.2 Assessing differences in phylogenies

The simulated results produce a phylogeny based on the SNPs occurring in the clonal frame of the alignment. This is taken to be the ”true” phylogeny to compare those produced by other models against. To compare these phylogenies I use the the Kendall-Colijn metric [547] as implemented in the treespace R package v1.1.4.1. In this metric each Tree, T, is represented by a vector, $v(T)$, which is composed of two separate vectors, $m(T)$ and $M(T)$, that record the distance between the MRCA of a pair of tips and the root, and the distance from a tip to its antecedent node. The vector $m(T)$ only mea-

sures the distance between the MRCA of two tips and the root in terms of the number of branches, with each branch given a length of 1. This value then only depends on the topology of T. The vector $M(T)$ however, measures the distance in terms of the branch length, both for the distance between a pair of tips and the root, and the distance from a tip to its immediate ancestor node. The parameter $\lambda \in$[0,1] is then used to combine these two separate vectors into $v_\lambda(T)$ (Equation 2.6).

$$v_\lambda(T) = (1 - \lambda)m(T) + \lambda M(T) \tag{2.6}$$

Once this is calculated for T_1 and T_2, the distance between the two trees is then the Euclidean distance between the two vectors $v_\lambda(T_1)$ and $v_\lambda(T_2)$ (Equation 2.7).

$$d_\lambda(T_1, T_2) = ||v_\lambda(T_1) - v_\lambda(T_2)|| \tag{2.7}$$

In this case using $\lambda = 0$ gives the distance between two trees purely in terms of topology, with a score of 0 for this metric indicating the trees have identical topology. When using $\lambda = 1$ however, a score of 0 would indicate not only that the topology of two trees is the same, but that the branch lengths are the same too and thus that these trees are wholly identical.

2.3 Results

2.3.1 Assessing the phylogenies produced from Gubbins

With the updates to Gubbins there are now a possible 60 different combinations of the first iteration phylogeny builder, main iteration phylogeny builder and ancestral state reconstruction method. As noted above (Section 2.1.1.3), typically a hybrid approach, whereby a faster less robust phylogeny builder is employed in the more computationally intensive first iteration of tree building, followed by more robust methods in later iterations, forms the best compromise between speed and accuracy in the detection of recombination events. So, in order to narrow the scope of benchmarking, only hybrid like model combinations have been fully benchmarked for this version of Gubbins. This leaves 18 different combinations tested (Table 2.2).

First iteration	Main Iterations	ASR
FastTree	IQ-TREE	Joint
FastTree	IQ-Tree	Marginal
FastTree	RAxML	Joint
FastTree	RAxML	Marginal
FastTree	RAxML-NG	Joint
FastTree	RAxML-NG	Marginal
RapidNJ	IQ-TREE	Joint
RapidNJ	IQ-TREE	Marginal
RapidNJ	RAxML	Joint
RapidNJ	RAxML	Marginal
RapidNJ	RAxML-NG	Joint
RapidNJ	RAxML-NG	Marginal
Star	IQ-TREE	Joint
Star	IQ-TREE	Marginal
Star	RAxML	Joint
Star	RAxML	Marginal
Star	RAxML-NG	Joint
Star	RAxML-NG	Marginal

Table 2.2: Models benchmarked for Gubbins v3.2.0. ASR refers to Ancestral state reconstruction method

To benchmark these different models and assess the accuracy with which they reconstruct the evolutionary history of samples, ten separate replications were created of simulated data at varying p_{branch} and μ_{rec} values. The output whole genome alignments from these simulations were then analysed using Gubbins, with each of the 18 different models identified in Table 2.2. An IQ-TREE [540] phylogeny produced directly from the output alignment, not correcting for recombination, was used to assess the effects of detecting recombination too. The alignment was also analysed with ClonalFrameML v1.12 [524], run with default parameters, with the simulation alignment and the IQ-TREE produced directly from this alignment used as the starting input for this algorithm. The first method of assessing accuracy was determining the distance of the phylogenies produced by each model from the true phylogeny, based on the clonal frame SNPs alone, produced from the simulation. To measure this distance I used the Kendall-Colijn metric [547] as implemented in the treespace R package v1.1.4.1 (See Section 2.2.2).

When applying this metric to the trees output from Gubbins, ClonalFrameML and directly from IQ-Tree, a few patterns start to emerge (Figures 2.1-2.4). Firstly, when we group the Gubbins models based on their starting tree, we see the phylogenies produced

are remarkably similar in terms of distance from the true tree. The trees have identical topology at different p_{branch} levels, matching the true simulated tree (Figure 2.1B), while they only start to differ in terms of topology at very high levels of recombination in the population, although their distances are all of the same order of magnitude (Figure 2.1A). When incorporating branch lengths too, the trees produced are very similar, across both μ_{rec} and p_{branch}. Overall the Gubbins models tend to match the topology of the simulated phylogeny more closely than the IQ-Tree and ClonalFrameML phylogenies (Figure 2.1A, the Raw and ClonalFrameML tree overlap one another), however at the highest μ_{rec} Gubbins performs slightly worse. When measuring branch lengths though, when $\lambda = 1$, the ClonalFrameML model outperforms all the Gubbins models across μ_{rec} (Figure 2.1C).

When varying the p_{branch} though, where $\mu_{rec} = 0.1$ for all these runs, the Gubbins models tend to get closer to the true tree with increasing p_{branch} than ClonalFrameML. ClonalFrameML assumes that the input ML phylogeny matches the topology of a clonal genealogy of the input data [524, 548], hence why the topology matches that of the tree formed directly by IQ-TREE (Figure 2.1A). Instead, ClonalFrameML re-scales the branch lengths of the input phylogeny based on its estimates of the recombination parameters and an expected number of mutations per site on a branch [524]. This estimation employs a Poisson approximation for the number of substitutions along a branch, similar to how recombination events are modelled in our simulation, which may explain why ClonalFrameML rescales branch lengths to more closely fit the true tree. Gubbins employs a more agnostic approach, removing recombination events and then building a phylogeny directly from this recombination-cured alignment. This approach means Gubbins can avoid the potential confounding of homoplasies introduced by recombination in terms of the topology of the tree, but this approach also appears to lessen the accuracy of Gubbins when inferring the branch lengths of the phylogeny.

When we group the Gubbins results by the phylogeny builder used in the main iterations of the algorithm we observe similar overall patterns to those from the starting iteration phylogeny builder (Figure 2.2). Again the Gubbins trees match the topology more closely over the majority of μ_{rec} values, and are identical to the topology of the true tree across the range of p_{branch} values (Figures 2.2A & 2.2B). When focusing on branch

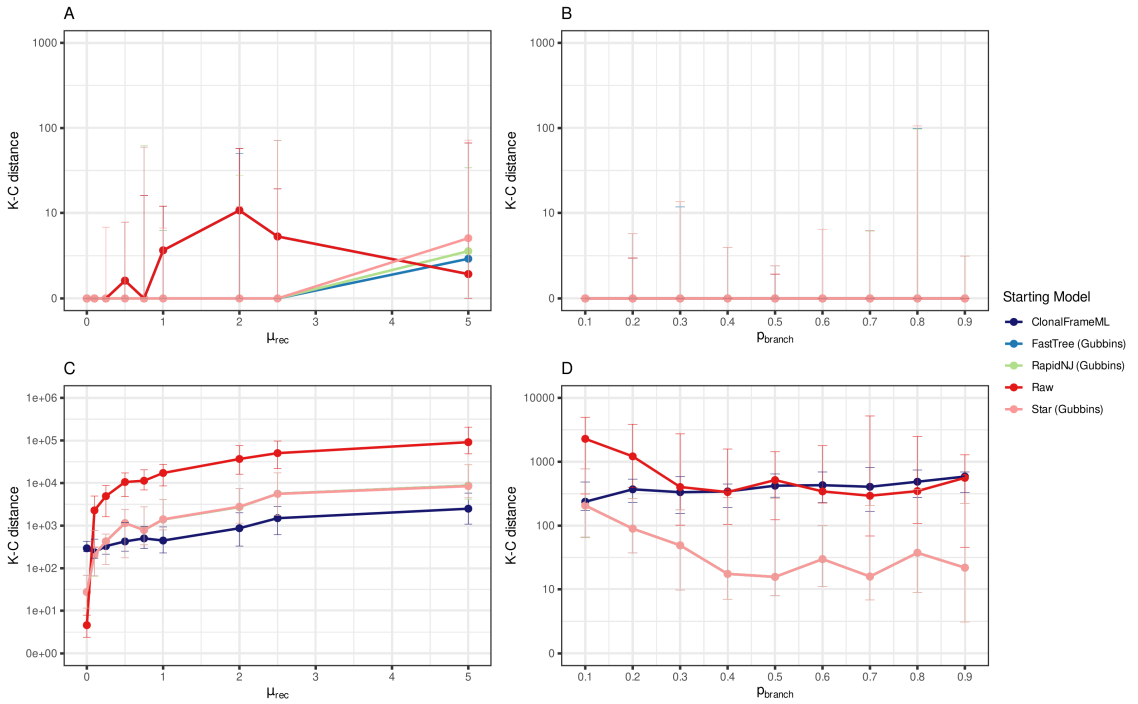

Figure 2.1: Kendall-Colijn distance metric between true simulated tree and output tree, with Gubbins models grouped by first iteration phylogeny builder. Dots represent the median value of the Kendall-Colijn (K-C) distance metric between the model in question and the true simulated tree. Error bars represent the full range of the K-C distance across all the simulations. Raw represents the IQ-TREE phylogeny produced directly from the simulated alignment. **(A)** The K-C distance when $\lambda = 0$ across a range of nine different values for μ_{rec}. The ClonalFrameML and Raw lines overlap one another. **(B)** The K-C distance when $\lambda = 0$ across a range of nine different values for p_{branch}. All median lines overlap one another **(C)** The K-C distance when $\lambda = 1$ across a range of nine different values for μ_{rec}. All Gubbins models overlap one another. **(D)** The K-C distance when $\lambda = 1$ across nine different value for p_{branch}. All Gubbins models also overlap one another.

lengths, ClonalFrameML performs best across μ_{rec}, while Gubbins is closer over the range of p_{branch} values (Figures 2.2C & 2.2D). There is little difference between the main Gubbins models, although at the highest levels of μ_{rec} it does appear that RAxML models are closer matches to the topology of the true tree, performing identically to ClonalFrameML with a median K-C metric of 1.41 (Figure 2.2A), and also slightly closer when measuring based on branch lengths (Figure 2.2C). When grouping by ancestral state reconstruction method we see similar patterns as for grouping by starting and main phylogeny builder (Figure 2.3). In this case the joint reconstruction does appear to perform slightly better than the marginal reconstruction, both in terms of matching topology and branch lengths more closely to the true tree (Figures 2.3A & 2.3C).

Finally, when the results are aggregated by the total model chosen, we can observe

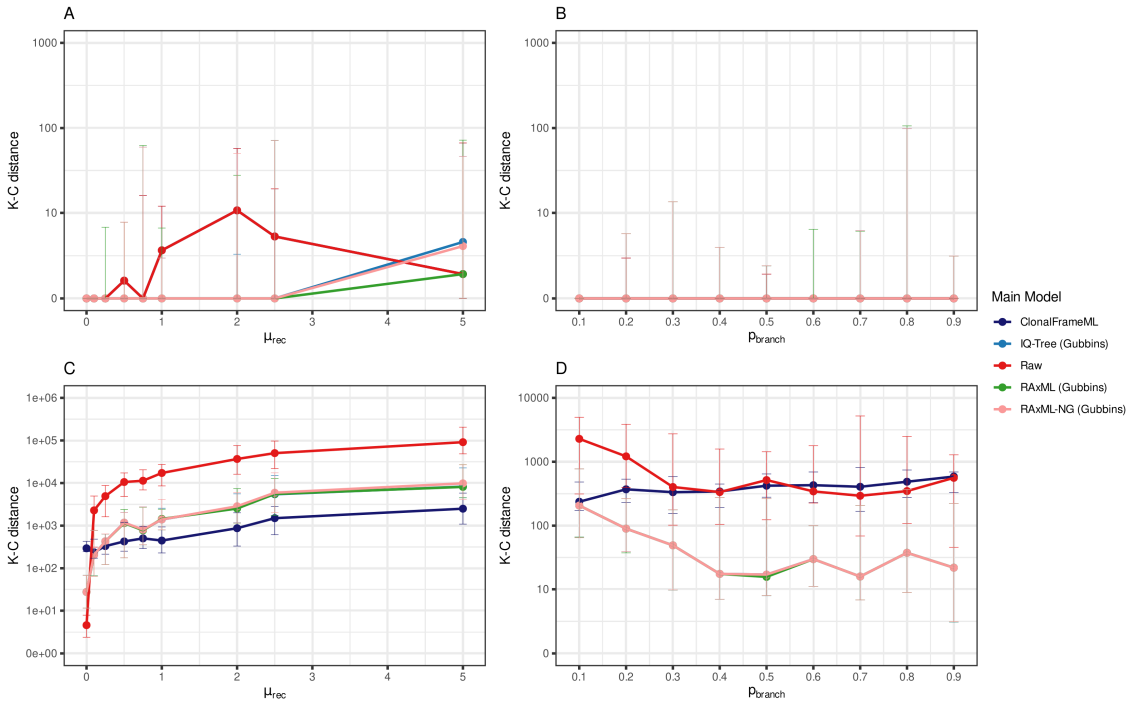

Figure 2.2: Kendall-Colijn distance metric between true simulated tree and output tree, with Gubbins models grouped by main iteration phylogeny builder. Dots represent the median value of the Kendall-Colijn (K-C) distance metric between the model in question and the true simulated tree. Error bars represent the full range of the K-C distance across all the simulations. Raw represents the IQ-TREE phylogeny produced directly from the simulated alignment. **(A)** The K-C distance when $\lambda = 0$ across a range of nine different values for μ_{rec}. The ClonalFrameML and Raw lines overlap one another. **(B)** The K-C distance when $\lambda = 0$ across a range of nine different values for p_{branch}. All lines overlap one another. **(C)** The K-C distance when $\lambda = 1$ across a range of nine different values for μ_{rec}. The Gubbins IQ-Tree and RAxML lines largely overlap with one another. **(D)** The K-C distance when $\lambda = 1$ across nine different value for p_{branch}. All Gubbins models largely overlap with one another.

the differences in their accuracy in approximating the true tree (Figure 2.4). Some models perform worse than others. The [Star, RAxML, Marginal] model for instance starts to diverge from the true tree topology at $\mu_{rec} = 2.5$, being more distant to the true tree's topology than a simple IQ-TREE run on the alignment (Figure 2.4A). Two models though, match the topology of the true tree across the range of μ_{rec}: [FastTree, RAxML, Joint] and [RapidNJ, RAxML, Joint] (Figure 2.4A & Table 2.3). When incorporating branch length, this combination of a RAxML main phylogeny builder and joint ancestral state reconstruction also performs best of the Gubbins models, this time with a star starting phylogeny (Figure 2.4C & Table 2.3). However, as with the previous aggregated results, the ClonalFrameML algorithm appears best at approximating the branch lengths of the true tree. Taken together, though, this suggests that for Gubbins, using RAxML as the main phy-

logeny builder and a joint reconstruction model is optimal for reconstructing a tree closest to that formed by the data, while the starting tree builder appears to make no difference.

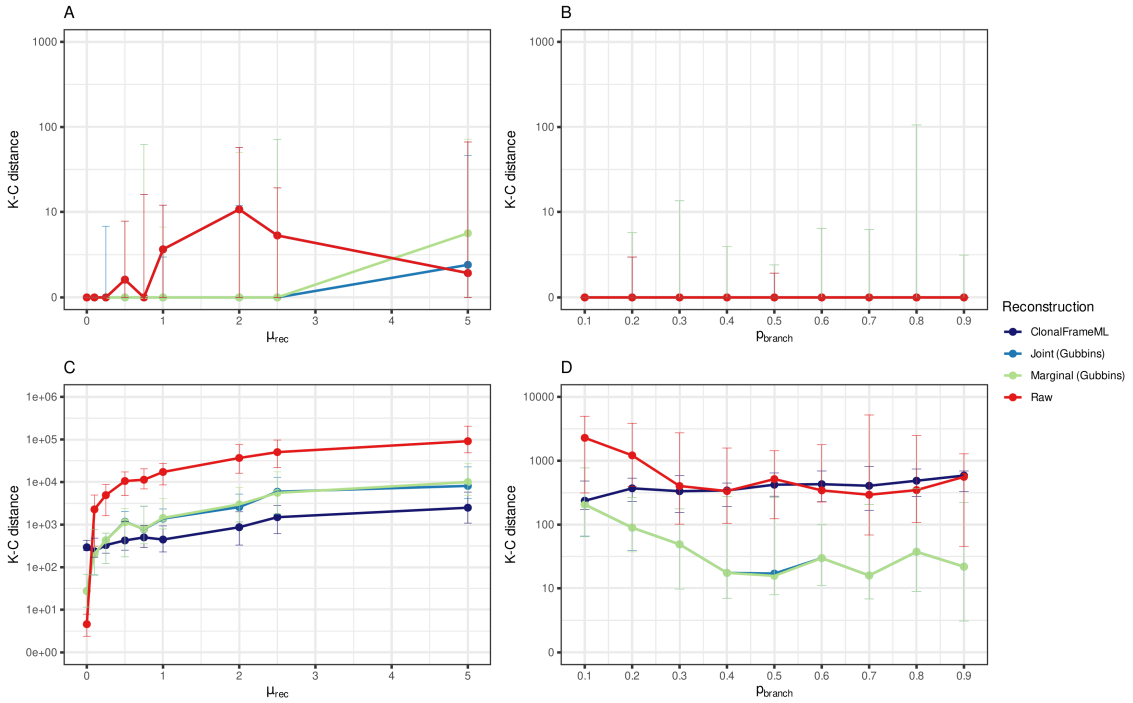

Figure 2.3: Kendall-Colijn distance metric between true simulated tree and output tree, with Gubbins models grouped by ancestral state reconstruction method. Dots represent the median value of the Kendall-Colijn (K-C) distance metric between the model in question and the true simulated tree. Error bars represent the full range of the K-C distance across all the simulations. Raw represents the IQ-TREE phylogeny produced directly from the simulated alignment. **(A)** The K-C distance when $\lambda = 0$ across a range of nine different values for μ_{rec}. The ClonalFrameML and Raw lines overlap with one another **(B)** The K-C distance when $\lambda = 0$ across a range of nine different values for p_{branch}. All lines overlap one another. **(C)** The K-C distance when $\lambda = 1$ across a range of nine different values for μ_{rec}. **(D)** The K-C distance when $\lambda = 1$ across nine different value for p_{branch}. The two Gubbins reconstruction models largely overlap with one another.

2.3.2 Assessing the choice of substitution model on phylogenies produced by Gubbins

As well as extending the array of phylogeny builders employed by Gubbins, we have also added more flexibility in the choice of substitution model used by the phylogeny builders and ancestral state reconstruction methods. In order to evaluate how this wider model choice affects the phylogenies created from Gubbins, data were simulated using a GTR model derived substitution rate across the same range of μ_{rec} and p_{branch} values used in the previous section, with 10 replicates for each value. Gubbins reconstructions were then run with either a JC model or a GTR model. Of the three starting models, only FastTree

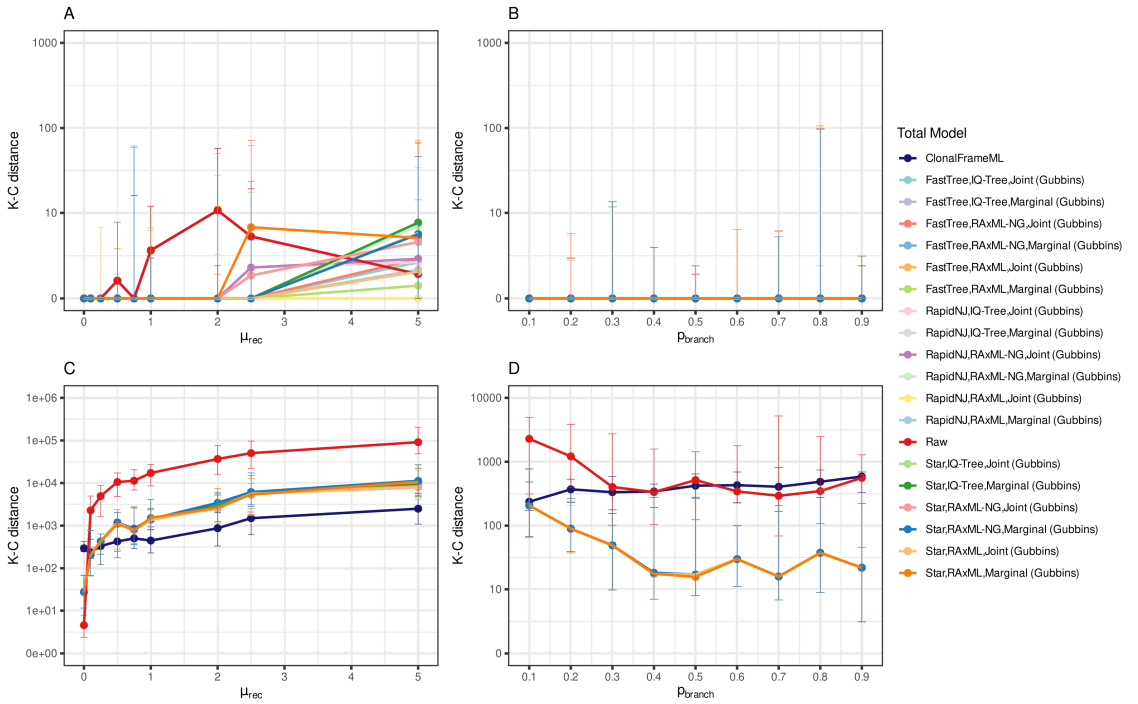

Figure 2.4: Kendall-Colijn distance metric between true simulated tree and output tree, with Gubbins models grouped by total model choice. Dots represent the median value of the Kendall-Colijn (K-C) distance metric between the model in question and the true simulated tree. Error bars represent the full range of the K-C distance across all the simulations. Raw represents the IQ-TREE phylogeny produced directly from the simulated alignment. **(A)** The K-C distance when $\lambda = 0$ across a range of nine different values for μ_{rec}. The ClonalFrameML and Raw lines overlap with one another. **(B)** The K-C distance when $\lambda = 0$ across a range of nine different values for p_{branch}. All models overlap with one another. **(C)** The K-C distance when $\lambda = 1$ across a range of nine different values for μ_{rec}. **(D)** The K-C distance when $\lambda = 1$ across nine different value for p_{branch}. All the Gubbins models largely overlap with one another.

has the option of running both a JC and GTR model, hence this was used as the sole starting phylogeny builder. The three main iteration phylogeny builders, IQ-TREE, RAxML and RAxML-NG, were then run with a joint reconstruction, given its closer topology scores (Figures 2.3A), and either a JC or GTR substition model. For comparison, IQ-TREE was also run directly on the simulation alignment, also with either a GTR or JC substitution model.

At very high recombination rates, the choice of substitution model does appear to affect the topology of the phylogeny produced (Figure 2.5). Apart from the RAxML-NG GTR model at μ_{rec} = 2.5, all the Gubbins models that employ a GTR substitution model remain at a median value of 0 across the range of μ_{rec} (Figure 2.5A). The models that use a JC substitution model however all begin to diverge from the true topology when

Model	$\lambda = 0$	$\lambda = 1$
ClonalFrameML	1.40	2,502
FastTree,IQ-Tree,Joint	2.30	7,956
FastTree,IQ-Tree,Marginal	5.70	9,181
FastTree,RAxML,Joint	0.00	7,953
FastTree,RAxML,Marginal	0.71	10,536
FastTree,RAxML-NG,Joint	2.60	8,563
FastTree,RAxML-NG,Marginal	7.00	11,310
RapidNJ,IQ-Tree,Joint	2.30	7,956
RapidNJ,IQ-Tree,Marginal	5.70	9,181
RapidNJ,RAxML,Joint	0.00	7,953
RapidNJ,RAxML,Marginal	1.70	10,536
RapidNJ,RAxML-NG,Joint	2.60	8,563
RapidNJ,RAxML-NG,Marginal	7.00	11,306
Raw	1.40	91,155
Star,IQ-Tree,Joint	4.40	7,956
Star,IQ-Tree,Marginal	7.60	9,181
Star,RAxML,Joint	1.60	7,953
Star,RAxML,Marginal	4.90	10,704
Star,RAxML-NG,Joint	4.40	8,331
Star,RAxML-NG,Marginal	5.50	11,280

Table 2.3: Kendal-Colijn metric distances between the true tree and each tree formed from the model listed in the table at $\mu_{rec} = 5$ and $p_{branch} = 0.1$. Values are split by the λ value used to measure the distance, values for $\lambda = 0$ have been rounded to 2 decimal places, values for $\lambda = 1$ have been rounded to the nearest integer.

μ_{rec} = 1. For the raw phylogeny formed directly from the alignment, the choice of model does not appear to influence the tree topology. Across the values of p_{branch}, when λ= 0 there appears to be no difference between the models, with all capable of recapitulating the topology of the true tree (Figure 2.5B). When the K-C metric is based on the branch lengths of the phylogenies, at λ= 1, there is little difference between the JC and GTR models (Figures 2.5C). The JC models are slightly closer to the true tree. The IQ-TREE JC model for instance has median score of 7568 at μ_{rec} = 5, whereas the GTR model has a median score of 7710, but at the scale of the results these are of negligible difference. It appears then that a GTR model, when run on data simulated in a GTR manner, is able to match the topology of a phylogeny closer than a simpler JC model.

As well as evaluating Gubbins results based on the output phylogeny, I also sought to understand how they differ in terms of their predictions of recombination events. In the next section I detail how I assessed these models, looking at their key recombination

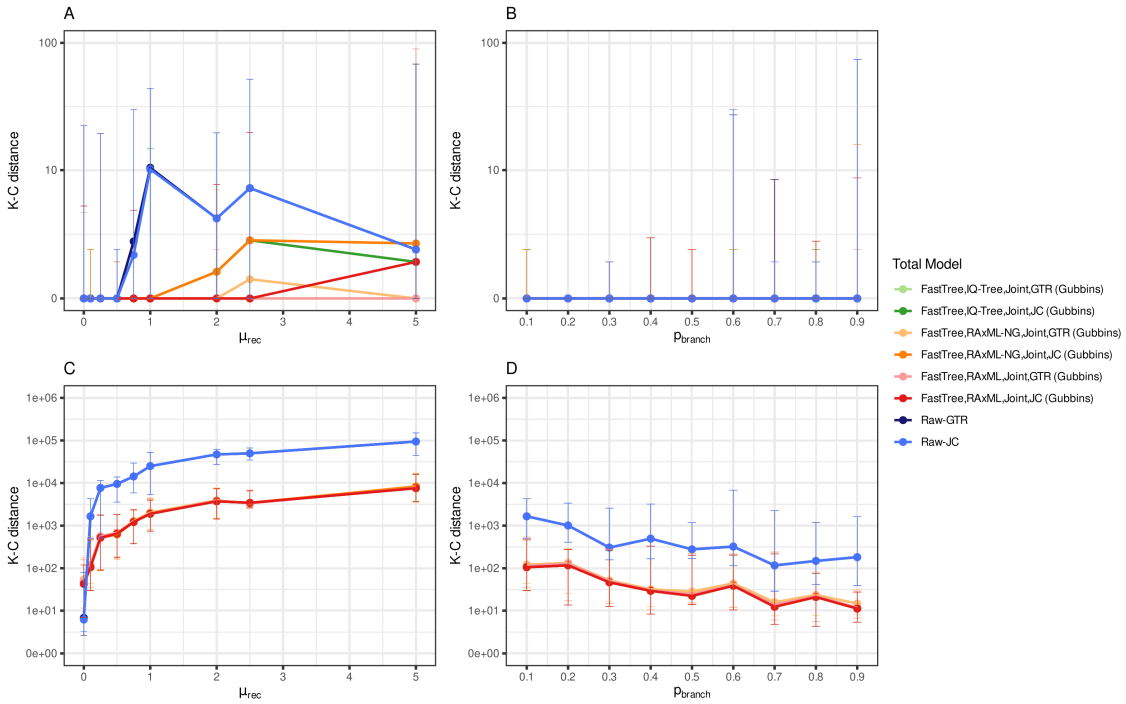

Figure 2.5: Comparison of the Phylogenies produced from a GTR and JC substitution model and the true phylogeny. (A) The Kendall-Colijn (KC) distance when λ= 0, between the true tree and the phylogeny output from the model highlighted. Raw models refer to phylogenies produced by IQ-TREE directly from the simulated alignment. Points represent the median value across 10 simulations, errorbars represent the full range of the values. Values are across a range of μ_{rec}, with p_{branch} set at 0.1. **(B)** The KC distance with λ= 0, across a range of p_{branch} values with μ_{rec} set at 0.1, data are as displayed in **A**. All models largely overlap with one another. **(C)** The KC distance when λ= 1 across a range of μ_{rec} values with p_{branch} = 0.1, data are as displayed in **A**. Both Raw methods overlap with one another, while the Gubbins models largely overlap with themselves too. **(D)** The KC distance when λ= 1 across a range of p_{branch} values with μ_{rec} = 0.1, data are as displayed in **A**. Again both Raw methods largely overlap with one another, while the Gubbins reconstructions also overlap with one another.

values.

2.3.3 The accuracy of recombination statistic estimation

The accuracy of the reconstructions was also assessed by a comparison of key recombination statistics. These statistics are: r/m, representing the number of SNPs imported via recombination (r) against point mutation (m); ρ/m, the number of recombination events (ρ) against point mutations; the number of recombination events (ρ); and the number of point mutations outside of recombination events (m). These are key statistics which are often used when defining the recombinogencity of a lineage [142, 549, 550]. ClonalFrameML, however, provides no direct estimate of r/m. Instead, multiplying ClonalFrameML's parameter estimates of R/θ, which measures the per site rate of initiation of recombination

relative to mutation, δ, the mean length of DNA imported by homologous recombination, and ν, the divergence rate per site of DNA imported by homologous recombination, represents the ClonalFrameML estimate of *r*/*m* [524]. Gubbins, though, produces precise assignments of the nature of substitutions, detailing whether they are introduced by point mutation or recombination. These statistics are then output into easily interpretable files which can be directly compared to simulation outputs.

2.3.3.1 Accuracy by phylogeny builder

The comparisons of the true simulated values of these statistics to those output from the different models, are not particularly favourable to any model (Figures 2.6-2.9). These simulated results have been narrowed to only encompass recombination events that contain a minimum of 3 SNPs, the default minimum SNP cutoff for Gubbins runs when detecting recombination events. When grouped by the starting phylogeny builder employed by Gubbins, we see, as observed with the tree distances in Section 2.3.1, that there is no appreciable difference between the models' estimates of the metrics (Figure 2.6). The reconstructed *r*/*m* values initially closely follow the true simulated *r*/*m* values at lower μ_{rec} values (Figure 2.6A). However, with increasing μ_{rec} the *r*/*m* for the simulated data increases, but Gubbins reconstructions plateau. ClonalFrameML also underestimates the *r*/*m* values, but this is less pronounced than Gubbins.

This pattern is more pronounced, though, for the ρ/*m* values, for which there is no ClonalFrameML estimate. Here, Gubbins reconstructions plateau at low levels of μ_{rec} (Figure 2.6B). Both these patterns, with plateauing *r*/*m* and ρ/*m* values, could be explained by the trends in the reconstructions of ρ, *m* and *r*(Figures 2.6C-E). From the reconstructions of ρ we can see that the models match the general trend reasonably well, ClonalFrameML in particular is able to detect the increasing number of recombination events as μ_{rec} increases. The reconstructions of the values of *m*, however, are very divergent (Figure 2.6D). The simulated value of *m* stays relatively constant across the range of μ_{rec}, as we'd expect with the p_{branch} set constant at 0.1 for these datapoints. Gubbins, however, reconstructs an increasing number of *m* SNPS as we increase μ_{rec}. This explains the plateauing of the ρ/*m* reconstructions, despite the reconstructions following the same trend for ρ. Similarly for *r*/*m*, the *r* value estimated by the models matches closely

with the simulated value (Figure 2.6C), this though is confounded by the large number of *m* reconstructed by Gubbins. Given the high correlation between ρ and *r* (Figure 2.10), the missed low SNP recombination events, which Gubbins is not powered to detect, are unlikely to be solely driving this increase in the number of *m* SNPs reconstructed.

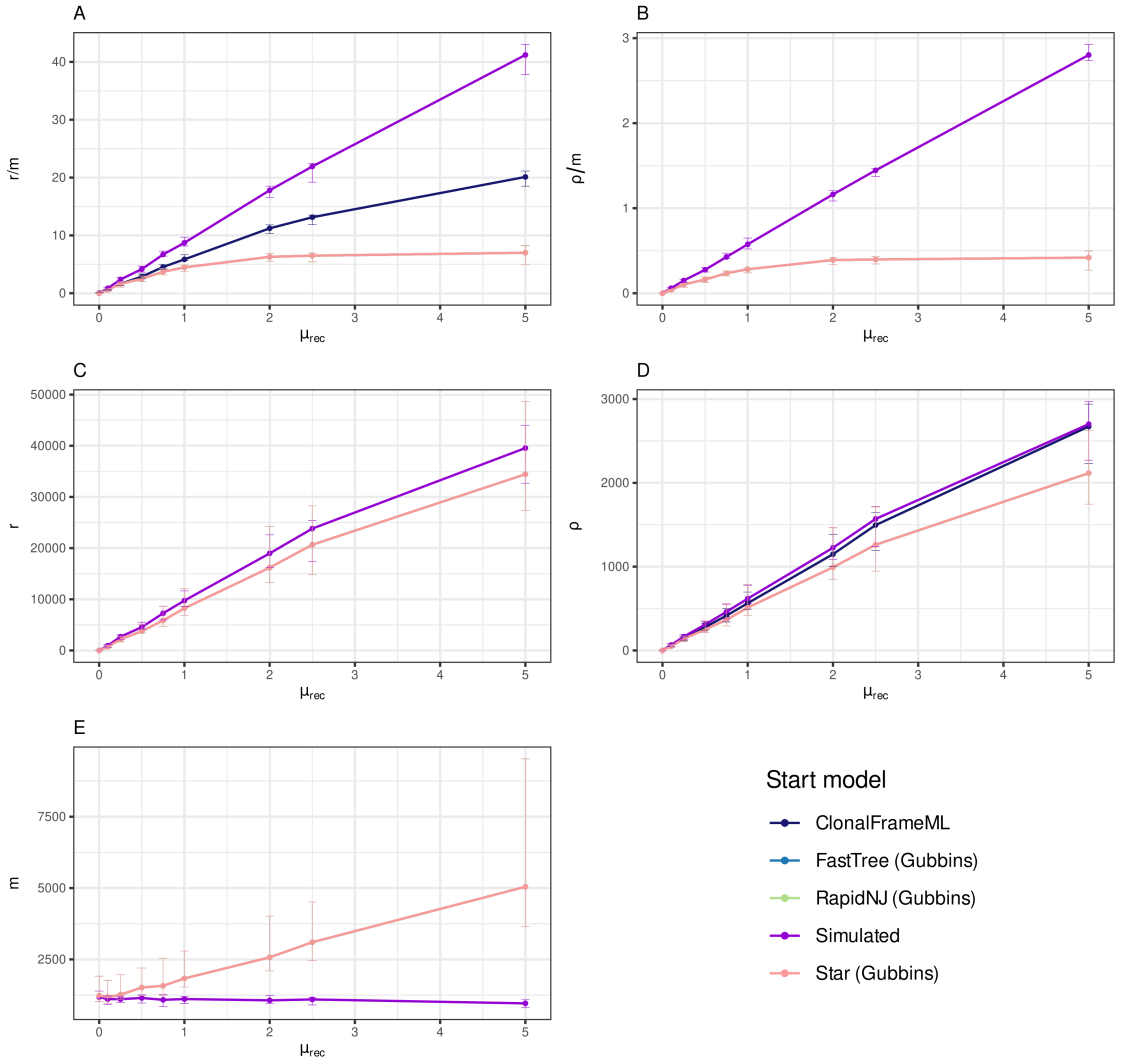

Figure 2.6: Recombination statistics from simulated and reconstructed data, with Gubbins models grouped by first iteration phylogeny builder. Dots represent the median value of the statistic in question. Error bars represent the full range of the statistic across all the simulations. Simulated represents the value produced from the simulation of sequences. For all statistics the Gubbins models largely overlap with one another. **(A)** The *r*/*m* values for the models across a range of μ_{rec} values. **(B)** The ρ/m values across a range of μ_{rec} values. Note ClonalFrameML does not produce ρ/m values. **(C)** The *r* values for the range of start models. ClonalFrameML does not produce estimates of the overall number of *r* SNPs. **(D)** The ρ values for the different models across a range of values for μ_{rec}. **(E)** The *m* values across a range of μ_{rec}. ClonalFrameML does not produce *m* values.

When the Gubbins results are grouped by main iteration phylogeny builder, some slight differences emerge between the models (Figure 2.7). The RAxML-NG model per-

forms best estimating *r*/*m*, having a median *r*/*m* value of 7.26 when μ_{rec} = 5, while IQ-TREE and RAxML have 7.00 and 6.54 respectively (Figure 2.7A). However, the simulated data has a median *r*/*m* of 42.2 when μ_{rec} is 5, while ClonalFrameML has a value of 20.1. For *r* and ρ, RAxML models produce the closest match of the Gubbins results across the range of μ_{rec} (Figure 2.7C-D). Although for *m*, RAxML appears to produce the highest estimates, which explains its lower estimations of *r*/*m* and ρ/*m* compared to the other models (Figure 2.7E).

2.3.3.2 Accuracy by reconstruction

When grouping the results by ancestral reconstruction it appears the joint reconstruction method performs slightly better overall (Figure 2.8). The median *r*/*m* of the joint reconstruction is 7.32 when μ_{rec} is 5, closer than the marginal median of 6.58. However, this is still much lower than the simulated value of 42.2. The joint reconstruction also performs better at estimating ρ/*m*. The median value is 0.432 when μ_{rec} is 5, compared to the marginal reconstructions median value of 0.398. Although this is still lower than the simulated median value of 2.80. As with the differences between the main phylogeny models, this difference between joint and marginal reconstruction appears to be largely driven by joint models reconstructing a lower *m* compared to the marginal reconstructions (Figures 2.8E).

Those models which use joint reconstruction appear to perform best when the results are aggregated by total model choice too (Figure 2.9). The closest *r*/*m* values come from reconstructions using a [Star, RAxML-NG and Joint] model set, with a median *r*/*m* of 7.33 when μ_{rec} is 5. Reassuringly, the [RapidNJ, RAxML, Joint] and [FastTree, RAxML, Joint] models that produced the closest topology matches to the true tree, are also among the highest estimates of *r*/*m* and ρ/*m* (Figure 2.9). The worst performing models are the [RAxML, Marginal] models, which, while matching ρ and *r* closely, overestimate *m* to form the lowest estimates of *r*/*m* and ρ/*m*. Overall, though, no model in particular matches closely to the simulated true statistics.

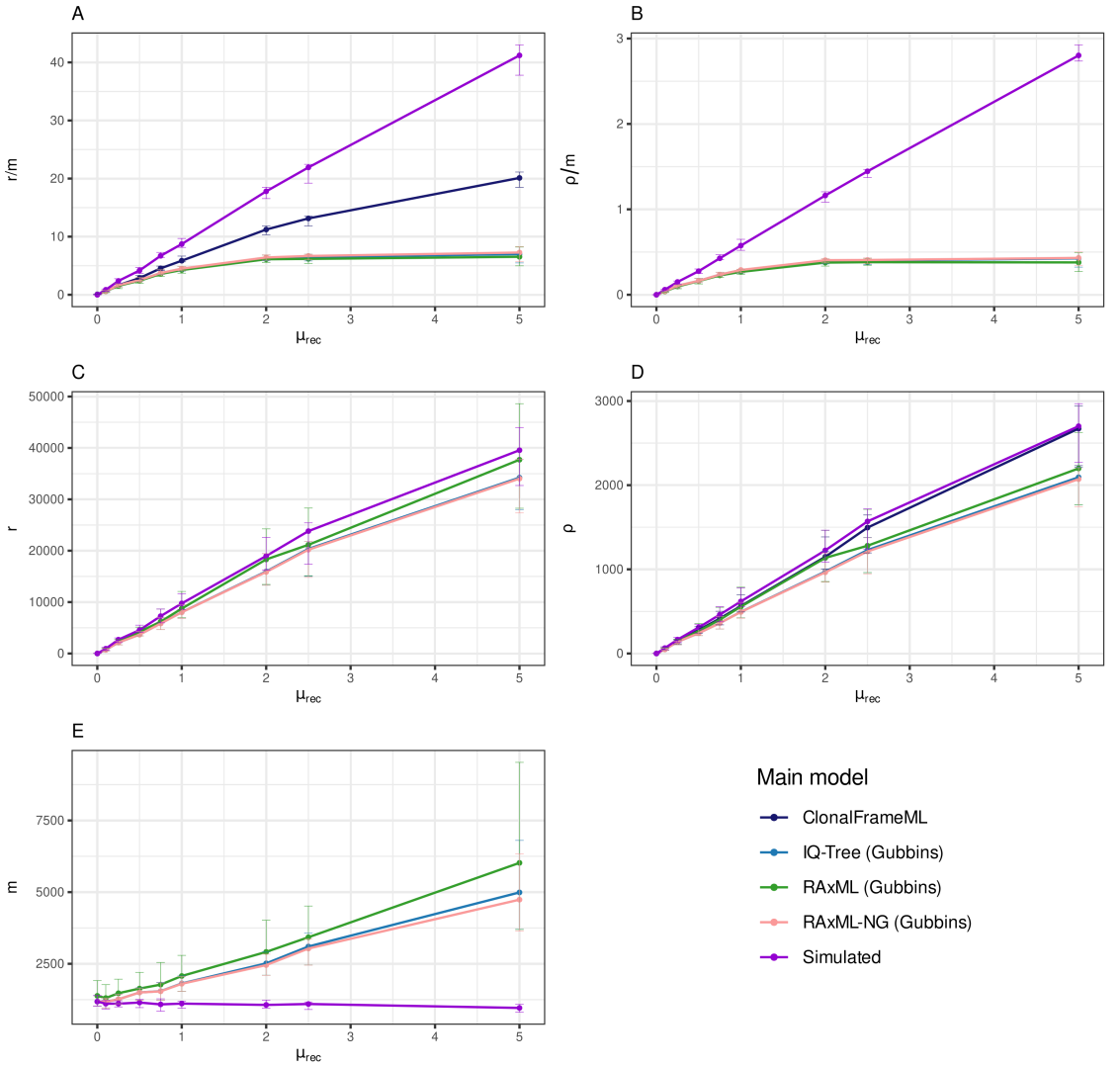

Figure 2.7: Recombination statistics from simulated and reconstructed data, with Gubbins models grouped by main iteration phylogeny builder. Data are as described in Figure 2.6. **(A)** The *r*/*m* values for the models across a range of μ_{rec} values. **(B)** The ρ/m values across a range of μ_{rec} values. Note ClonalFrameML does not produce ρ/m values. **(C)** The *r* values across a range of μ_{rec}. **(D)** The ρ values for the different models across a range of values for μ_{rec}. **(E)** The *m* values across a range of μ_{rec}. ClonalFrameML does not produce *m* values.

2.3.3.3 Assessing the accuracy of SNP classification

To understand why these reconstructions are under-performing, an assessment of the accuracy of the classification of the underlying SNPs, whether they occur within recombination events (r SNPs) or outside recombination events (m SNPs), is also needed. The accuracy of these reconstructions was assessed through the PPV score and the sensitivity score (Figure 2.11). The PPV score measures the number of true positives divided by the total number of positives produced from the reconstruction. In this case a true positive

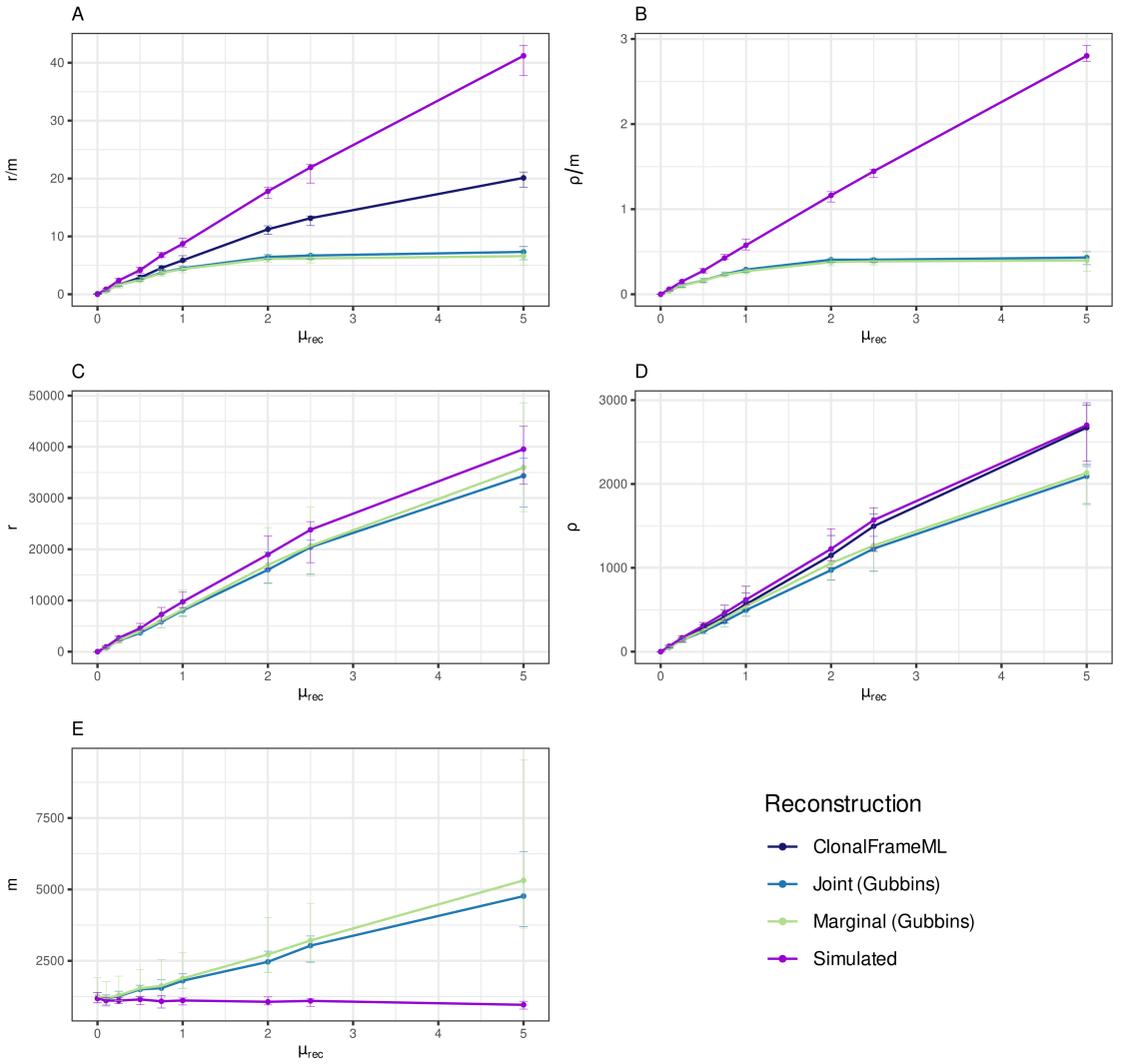

Figure 2.8: Recombination statistics from simulated and reconstructed data, with Gubbins models grouped by ancestral state reconstruction method. Data are as described in Figure 2.6. **(A)** The *r*/*m* values for the models across a range of μ_{rec} values. **(B)** The ρ/m values across a range of μ_{rec} values. Note ClonalFrameML does not produce ρ/m values. **(C)** The *r* values across a range of μ_{rec}. **(D)** The ρ values for the different models across a range of values for μ_{rec}. **(E)** The *m* values across a range of μ_{rec}. ClonalFrameML does not produce *m* values.

is defined as a SNP which occurs on the same branch, at the same locus and is classified as the type of SNP (either *r* or *m* in nature). Sensitivity on the other hand measures the number of true positives divided by the total number of positives in the simulated data. This is indicative of the level of false negatives produced by the reconstructions.

The sensitivity of the different Gubbins models is reasonably consistent and accurate across the range of μ_{rec} values (Figure 2.11A). The [RAxML-NG, Joint] models are the most accurate for *r* SNPs, all three having a median sensitivy greater than 74.5% when μ_{rec} is 5. Although the [RapidNJ, RAxML, Joint] and [FastTree, RAxML, Joint] models

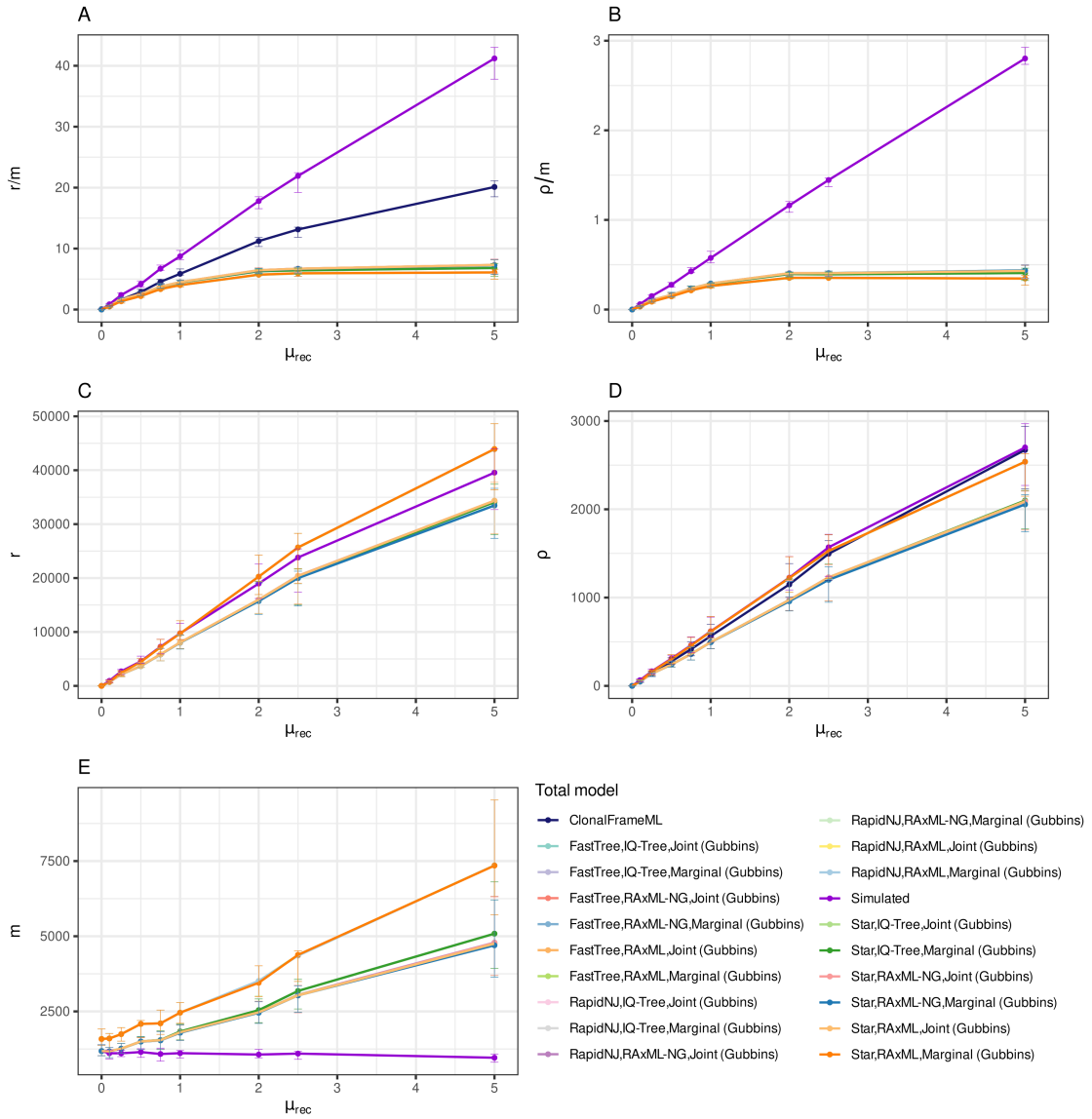

Figure 2.9: Recombination statistics from simulated and reconstructed data, with Gubbins models grouped by full model used. Data are as described in Figure 2.6. **(A)** The *r*/*m* values for the models across a range of μ_{rec} values. **(B)** The ρ/m values across a range of μ_{rec} values. Note ClonalFrameML does not produce ρ/m values. **(C)** The *r* values across a range of μ_{rec}. **(C)** The *r* values across a range of μ_{rec}. **(D)** The ρ values for the different models across a range of values for μ_{rec}. **(E)** The *m* values across a range of μ_{rec}. ClonalFrameML does not produce *m* values.

are again very close with median sensitivities of 74.4% when μ_{rec} is 5, compared to the [RapidNJ, RAxML-NG, Joint] median of 74.7%. For *m* SNPs there appears to be very few false negatives, with [FastTree, RAxML, Joint] model producing a median sensitivity of 98.0%. The three RAxML, Marginal models however are markedly less accurate than the other models, dropping to a median sensitivity of 50.4% when μ_{rec} is 5 for *r* SNPs and 61.5% for *m* SNPs. This likely represents an issue with the RAxML ancestral state

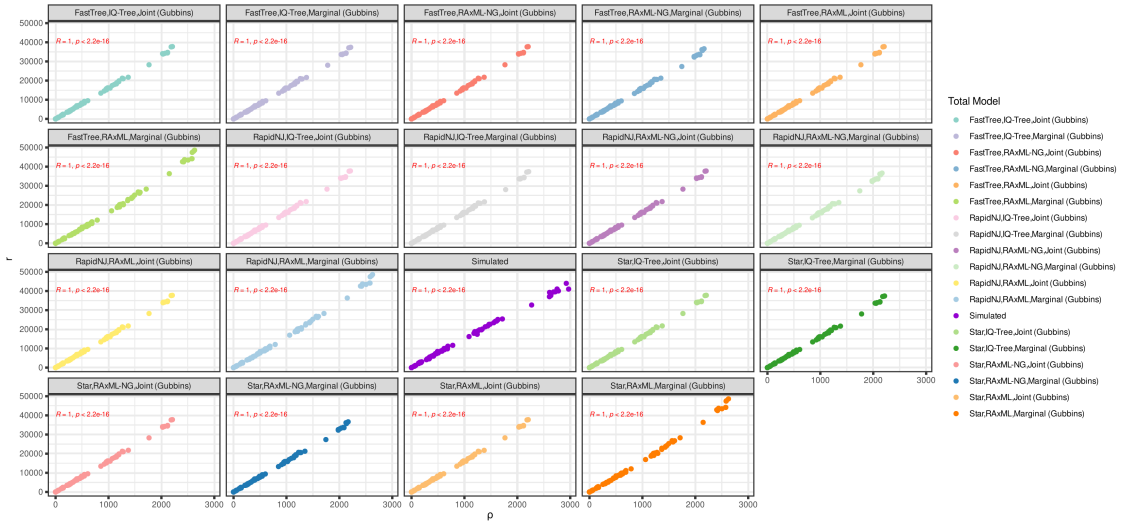

Figure 2.10: Correlation plots of the value of ρ against *r* for different models. Each point represents a single reconstruction, each model has 170 different datapoints. R values displayed are calculated as the Pearson correlation coefficient

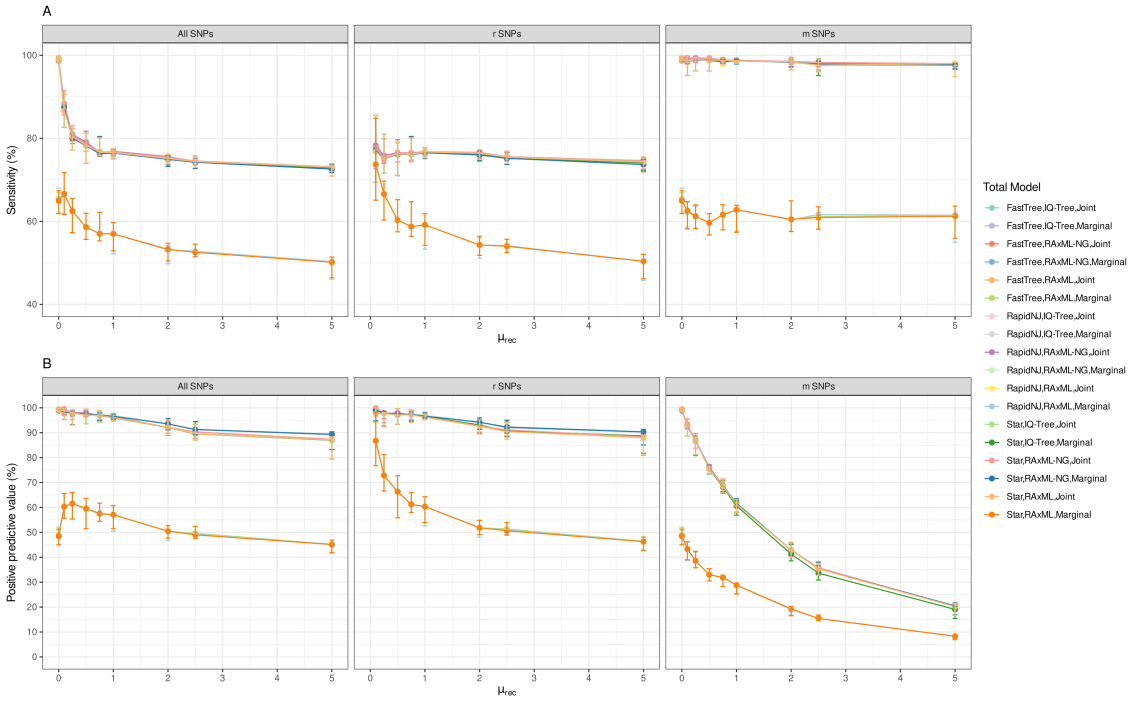

Figure 2.11: Sensitivity and PPV scores for Gubbins reconstructions. (A) The sensitivity scores for all SNPs, *r* SNPS (those reconstructed within recombination events) and *m* SNPs (those reconstructed to occur outside of recombination events) across a range of μ_{rec} values. Lines and points are coloured by total model, points represent the median value across 10 different simulations at each μ_{rec} value, error bars represent the full range of the values at this value of μ_{rec}. **(B)** The PPV scores for the range of Gubbins models, data are as described for **(A)**.

reconstruction step, given the RAxML phylogenies produced for the [RapidNJ, RAxML, Joint] and [FastTree, RAxML, Joint] are among the most accurate of the models. The high

level of consistency of sensitivity across the range of μ_{rec} values, for both *r* and *m* SNPs, indicates the Gubbins algorithm is robust in detecting true events from the input data.

With regards to PPV, when looking at all SNPs the Gubbins models perform consistently well (Figure 2.11B). The majority of models remaining above 85% for the PPV for all SNPS, across the range of μ_{rec}. This is largely driven by the high PPV of *r* SNPs, which vastly outnumber the *m* SNPs at higher μ_{rec} values (Figure 2.9C&E). [RAxML-NG, Marginal] models are the highest scoring, with their median PPV scores remaining above 90% across μ_{rec}. The [RAxML, Joint] models meanwhile score at 88% for the highest μ_{rec} value. These high and consistent scores likely reflect the stringent likelihood checks required for clusters of SNPs to be identified as a recombination block. Once again, however, the three [RAxML, Marginal] models are well below the rest of the models, scoring only 46% at the highest levels of μ_{rec}. For *m* SNPs too, all the models perform poorly. At the lowest levels of μ_{rec} the PPV values are accurate, with the majority of models remaining above 98% when μ_{rec} is 0.1. However, this quickly declines over the range of μ_{rec} values, reaching a nadir of 8.8% for the [RAxML, Marginal] models when μ_{rec} is 5. The large number of false positive *m* SNPs reconstructed by Gubbins at these higher μ_{rec} values explains the plateauing of both *r*/*m* and ρ/*m* for these models (Figure 2.9). It may also explain the worse performance of the Gubbins results when measuring tree distance by branch lengths (Figure 2.4C), with the excess *m* SNPs adversely affecting the branch lengths produced for the reconstructed trees.

These excess SNPs do not appear to be driven by smaller low SNP recombination events missed by Gubbins (Figure 2.12). The difference between the excess number of *m* SNPs reconstructed and the number of SNPs introduced by a recombination event with fewer than 3 SNPs, increases as μ_{rec} increases. This indicates another factor is causing the large number of false positive *m* SNPs.

2.3.4 Time and memory usage of Gubbins models

As well as the accuracy of reconstructions, models must be able to run rapidly and efficiently in terms of memory usage, in order to process the increasingly larger WGS datasets available today. To quantify the time and memory usage of the different recombination detection algorithms, simulations were run to output alignments which contained

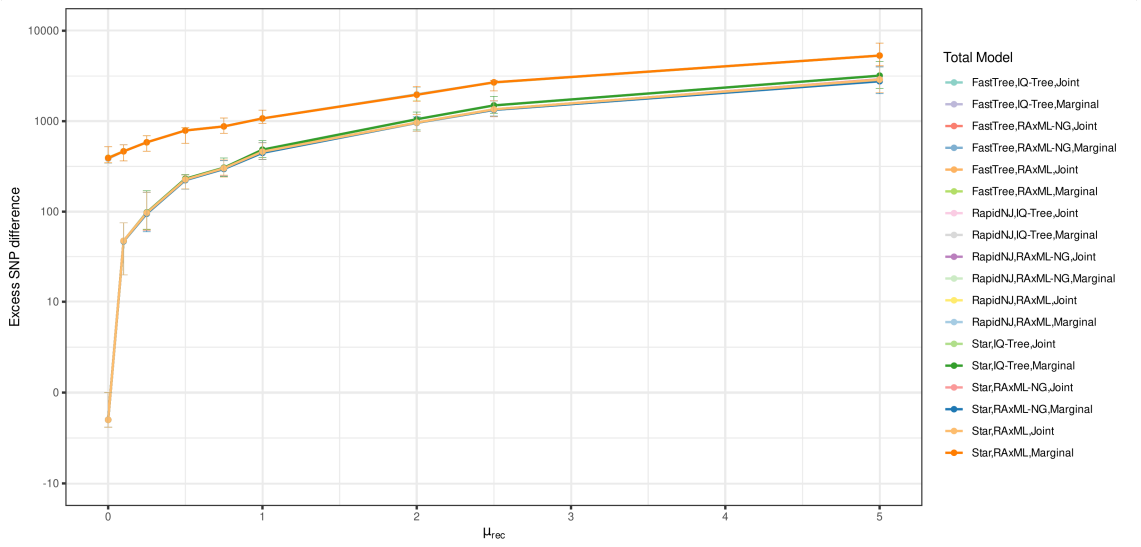

Figure 2.12: The difference between the excess number of *m* SNPs reconstructed by Gubbins models and the number of *r* SNPs present in recombinations with less than three SNPs. Dots represent median values across 10 simulations for each model, errorbars represent the full range of the data. Data are plotted across a range of μ_{rec} values with p_{branch} constant across all datapoints.

a range of 50 to 500 sequences, in increments of 50 sequences. For each alignment sequence number, n_{seq}, 10 sets of replicated simulations were run to give 100 different simulation sets in total. The 18 different model combinations outlined in Table 2.2, along with ClonalFrameML and the hybrid [FastTree, RAXML, Marginal] method from Gubbins v2.4.0, were then used to reconstruct these simulated alignments. For ClonalFrameML, the run-time for the initial IQ-TREE phylogeny is included, as Gubbins also performs this step within its algorithm. This produced 2,000 separate reconstructions in total. All reconstructions were performed on a single core, using AMD EPYC 7742 64-Core Processors.

In terms of run time, the six RAxML-NG models were by far the slowest (Figure 2.13A). The median run time for the slowest model, [FastTree, RAxML-NG, Marginal] was 23.5 hours at an n_{seq} of 500, whereas the equivalent RAxML model had a median run time of 2.2 hours for the same n_{seq}. Outside of the RAxML-NG models, ClonalFrameML and Gubbins v2.4.0 had the next slowest run times. ClonalFrameML taking 5.4 hours on average to run with the initial IQ-TREE forming at an n_{seq} of 500, Gubbins v2.4.0 taking 3.1 hours. The [FastTree, RAxML, Joint] models took 2.1 hours for the same n_{seq}. The relationship between ClonalFrameML runtime, Gubbins runtine and n_{seq} was seen to be quadratic. For ClonalFrameML this was best described by the line: $runtime(s) = 0.05 * n_{seq}^2 + 1.25 *$

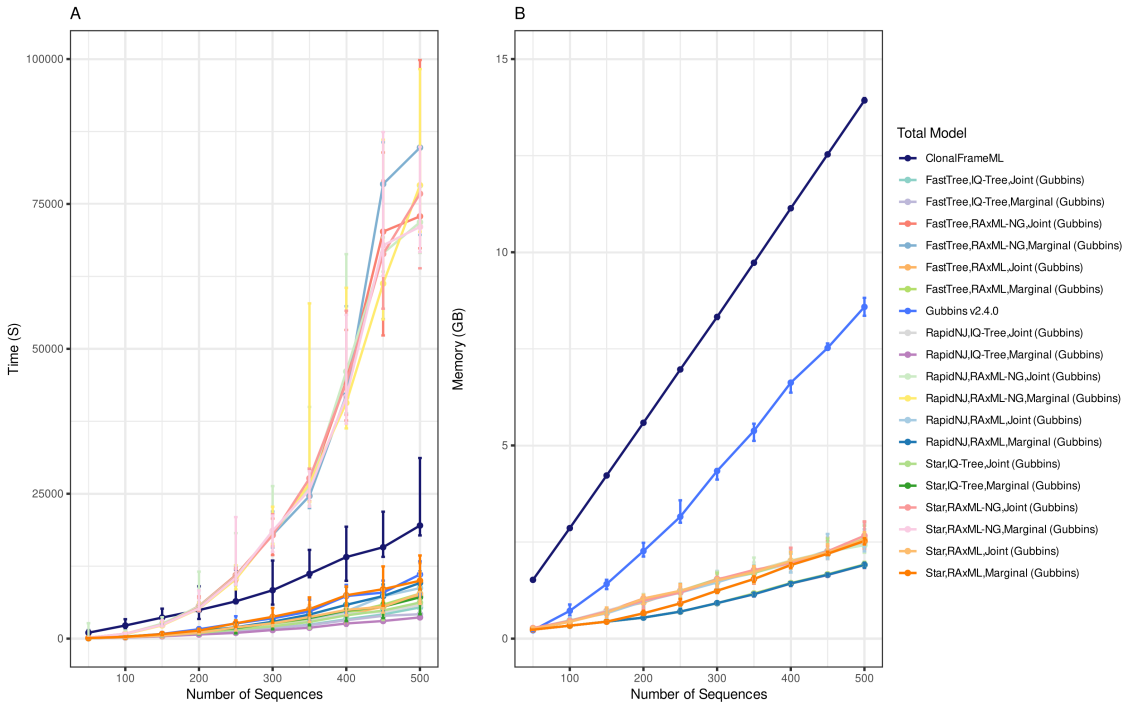

Figure 2.13: The analysis times and memory usages over a range of alignment sizes for different recombination detection methods. (A) The Time taken in seconds against the number of sequences in an alignment. Dots represent median values across 10 simulated datasets, errorbars represent the full range of the values. Dots and lines are coloured by model run. **(B)** The peak memory usage, in gigabytes, during reconstruction across a range of alignment sizes. Data are as described in **(A)**.

$n_{seq} + 435$ (R^2 = 0.9973, F(2,7) = 1297, p = 1.01x10^{-9}). For the [FastTree, RAxML, Joint] model this was best described by the line: $runtime(s) = 0.035 * n_{seq}^2 - 3.94 * n_{seq} + 269$ (R^2 = 0.9935, F(2,7) = 745, p = 7*10^{-9}). ClonalFrameML, has no option to be run in a multi-threaded manner, whereas all Gubbins models can be run in parallel, potentially mitigating any longer runtimes at higher sequence numbers. Indeed, for the [FastTree, RAxML, Joint] model, running on multiple cores does lead to a speed-up in the runtime, while the memory usage remains constant with increasing cores (Table 2.4).

With regards to memory usage, the ClonalFrameML runs used by far the most memory across all values of n_{seq} (Figure 2.13B). ClonalFrameML reached a peak median memory usage of 13.93 GB at an n_{seq} of 500, whereas the [FastTree, RAxML, Joint] model used 2.54 GB of memory at the same n_{seq} value. The updated Gubbins v3.2.0 models also used less memory than Gubbins v2.4.0, with the equivalent [FastTree, RAxML, Marginal] model from v3.2.0 using 1.92 GB when n_{seq} was 500, whereas v2.4.0 used 8.58 GB. Among the v3.2.0 models, the Joint models tended to use slightly more mem-

Cores	Time(s)	Memory(GB)
1	7385.5 [6428-13026]	2.54 [2.46 - 3.01]
4	3282.94 [2848.65-3607]	2.49 [2.24 - 2.96]
8	2722.74 [2146.38-3518.48]	2.5 [2.34 - 2.95]
12	2214.94 [1910.29-2734.07]	2.46 [2.36 - 3.02]

Table 2.4: Runtimes and memory usage for multiple cores. The values shown are the median runtimes and memory usage across 10 replicates of data with $n_{seq} = 500$ for the [FastTree, RAxML, Joint] model. Values within brackets represent the full range of these performance statistics across the 10 replicates.

ory. When grouping the models by reconstruction method, at an n_{seq} of 500 the median memory usage was 2.54 GB for the Joint models compared to 1.93 GB for the marginal models.

All models appeared to have a linear relationship between memory usage and n_{seq}. ClonaFrameML was best described by the line $memory(GB) = 0.027 * n_{seq} + 0.09$ (R^2=0.999, F(1,8) = 151,600, p < 2.2×10^{-16}). ClonalFrameML has a steeper gradient than the [FastTree, RAxML, Joint] model, which is best described by the line $memory(GB) = 0.005 * n_{seq} - 0.042$ (R^2=0.998, F(1,8) = 4946, p = 1.86×10^{-12}).

2.3.5 Conclusion

In this chapter I have outlined how we've expanded the popular Gubbins algorithm in order to improve reconstructions and keep pace with the current level data production in bacterial genomics. I have benchmarked the new models employed against: another commonly used algorithm for detecting recombination, the previous version of Gubbins and the näive reconstruction of evolutionary history without taking into account recombination events. From these results we can see that Gubbins is able to robustly reproduce the topology of the true tree, can infer the number of recombination events and converges on results rapidly with an efficient use of memory.

From this work I can recommend a default combination of a FastTree starting phylogeny, RAxML main phylogeny builder with a joint ancestral state reconstruction and a GTR substitution model. This combination performs among the best output for SNP and recombination event reconstruction accuracy, as well as the best reconstruction method for matching the topology of the true phylogeny. The [RapidNJ, RAxML, Joint] model per-

forms similarly for the topoloy and SNP classification, however it does not have the option to use a GTR model and is slightly slower than the [FastTree, RAxML, Joint] model. That RAxMl is the best performing phylogeny builder is slightly surprising. In previous tests RAxML-NG has outperformed RAxML [535], and IQ-TREE has also been judged more accurate in direct tests with RAxML [551]. IQ-TREE uses a stochastic approach to find the optimal phylogeny among a set of candidates [540]. Given the full range of K-C distances do encompass 0 for the IQ-TREE runs, perhaps this difference was simply a result of having a small sample size of only 10 simulations at a given value to select from. Running a larger number of simulations however would take require more disk space, something which was limited in my analysis.

Crucially though, the Gubbins models do outperform the ClonalFrameML approach in terms of the reconstructing the topology of the true tree. For further phylodynamic analyses, the topology of the tree being correct is key. The time-calibrated phylogeny algorithm BactDating [552] for instance, only models changes in branch lengths when dating a phylogeny. An incorrect topology inferred from a ClonalFrameML model would render any subsequent dating results as spurious.

The Gubbins models though, are outperformed by ClonalFrameML in terms of recreating accurate branch lengths for high recombination rate populations. These inaccurate branch lengths may produce spurious molecular clock estimates that can also alter any future phylodynamic analyses. However, these branch lengths were only tested with a low constant p_{branch} of 0.1. This produces a tree with naturally long branches which represents either a poorly sampled population, or a highly diverse species. With the ever increasing sampling and sequencing of bacteria [553, 554], denser sampling of lineages is becoming the norm and perhaps these low p_{branch} rates are not as applicable. Given that the Gubbins results tended to outperform ClonalFrameML in terms of branch length estimation at higher p_{branch}, perhaps testing at these higher levels would produce Gubbins results that are more accurate.

The altered branch lengths seen at higher μ_{rec} values are driven by increasing numbers of false positive SNPs, predicted to be outside of recombination events, when μ_{rec} is increasing. This excess of false positive events in turn causes r/m and ρ/m values to

plateau at ever increasing values of μ_{rec} for Gubbins reconstructions. The values for *r*/*m* plateau for most models at 7.3, despite the simulated data having an *r*/*m* of 41.2. In the literature however, Gubbins has been used to estimate *r*/*m* values above 20 for pneumococci [155, 555] and other pathogens [549]. This is indicative of there being no upper limit on the *r*/*m* value estimated by Gubbins. Instead, this suggests, perhaps, that the nature of the output simulated data, which can produce many small recombination events with fewer SNPs, may be itself a limit on the *r*/*m* estimated by Gubbins. Further work is needed to fully understand why these simulations produce these skewed estimates.

Now that I have described the Gubbins recombination detection algorithm and the improvements that have been made, in the following chapters we'll look at how these methods can be applied. In particular to understand how AMR emerges in bacterial populations. I'll start with resistance emerging at core genomic loci within MDR populations of the pneumococcus.

Chapter 3

The emergence of resistance among core genes of the pneumococcus

Acknowledgements

The work in this chapter has been previously published in D'Aeth *et al* 2021 [556]. Those isolates used in the study not released through the GPS project, were sequenced at the Wellcome Trust Sanger Institute as described in Section 3.2.1. Gubbins analysis of the individual GPSCs were performed by the GPS team. All other analyses were run by myself.

Summary

This chapter investigates how recombination can drive alterations in core genes that lead to resistance. I describe two multi-drug resistant (MDR) lineages of the pneumococcus, PMEN3 and PMEN9. Using Gubbins, I detect recombination events around the pbp loci leading to the expansion of the PMEN3. Looking at the wider GPS collection of isolates, I also detect numerous instances of interspecies recombination at the pbp loci. These results highlight how selection can drive recombination between diverse donor and recipients at key loci.

3.1 Introduction

3.1.1 Core gene resistance and recombination

Resistance genes can arise within bacterial populations either through modifications to core genes or the movement of genes on MGEs. Common core gene modifications in the pneumococcus include alterations in the penicillin binding protein genes (*pbp1a, pbp2b* and *pbp2x*), encoding proteins targeted by β-lactam antibiotics, and genes encoding enzymes involved in the folate biosynthesis pathway (*dhfR* and *folP*), targeted by co-trimoxazole and its constituent drugs [17, 140]. For the development of resistance to sulfamethoxazole, a component of co-trimoxazole, relatively few mutations are needed, with 3 or 6 bp duplications sufficient for reduced sulfamethoxazole susceptibility [557,558]. While this may spread via clonal expansion of newly resistant lineages, recombination is thought to play a key role in the dissemination of this resistance among the wider pneumococcal population [559].

Early investigations into the nature of β-lactam resistance in pneumococci also revealed recombination was required for the spread of non-susceptibility. Resistant PBP genes were shown to be 'mosaic' in nature, being created from a mixture of sequence from *Streptococcus pneumoniae* and the related oronasopharyngeal commensal streptococcal species, *Streptococcus mitis* and *Streptococcus oralis* [138, 560, 561]. The mosaicism reflected the role of recombination in importing fragments of genes from these other species, with these fragments being much smaller than a typical gene. Although, there was also evidence of these recombinations causing diversification in the flanking regions of the chromosome, particularly the *cps* locus controlling an isolate's serotype [562].

Recombination plays a role in the emergence of these resistance phenotypes at an individual isolate level. In this Chapter I aim to investigate how recombination drove the emergence and spread of core gene-mediated resistance in pneumococcal populations, looking in particular at two multi-drug resistant (MDR) lineages, PMEN3 and PMEN9. This builds on previous work looking at the internationally disseminated PMEN1, PMEN2 and PMEN14 pneumococcal lineages, among others [142, 312, 555]. I expand this by looking at a much larger collection of over 20,000 pneumococcal isolates sampled through the

Global Pneumococcal sequencing project (GPS) [563]. These sequences have split into strains, known as Global pneumococcal sequencing clusters (GPSCs). I will now briefly describe the pneumococcal populations I investigate in this chapter.

3.1.2 PMEN3 population

To understand the effects of recombination on a population wide level, I investigated global collections of pneumococci. The first I shall focus on is the Spain^{9V}-3, or PMEN3 lineage, which is within strain GPSC6 (clonal complex 156 by multi-locus sequence typing, MLST) [563]. PMEN3 was first documented in Spain in 1988 with a serotype 9V capsule, and by the late 1990s it was recognised as one of the major penicillin resistant lineages causing meningitis in the country [564]. By 2000, 55% of all penicillin resistant disease isolates in South America were from the PMEN3 lineage [141]. Isolates with serotype 19F, 19A and 14 capsules have also been observed, all of which are either included in the seven valent pneumococcal polysaccharide conjugate vaccine (PCV7) or the thirteen valent PCV13 [565, 566]. Recent work has also observed PMEN3 with a non-vaccine serotype 11A following the introduction of PCV13 in Spain [567, 568]. While first detected in Spain, PMEN3 was later found in France, the USA and South America [141, 569–571].

3.1.3 PMEN9 collection

The second collection investigated in this chapter is the England^{14}-9, or PMEN9 lineage, of isolates. PMEN9 is within the GPS strain GPSC18 (clonal complex 9 or 15 by MLST). An isolate from the PMEN9 lineage was first described in the UK in 1996 and it has since been detected across Europe, the Americas and Asia [572–574]. Although PMEN9 was originally penicillin susceptible, very high-levels of resistance have emerged among clones in the US [575]. This led to PMEN9 becoming the most common lineage causing penicillin-resistant invasive pneumococcal disease (IPD) in the USA [576]. PMEN9 is also known to display very high levels of resistance to macrolides, with Tn*1207.1*, a defective transposon carrying the *mef(A)* macrolide resistance gene, thought to be a key driver of this [577]. Indeed, prior to the introduction of infant vaccination in Germany, PMEN9 was the most common lineage causing macrolide-resistant IPD in the country [578, 579].

3.1.4 GPS collection

The GPS collection represents data accumulated from large scale genomic surveys, with 33 countries contributing isolates to this database of 20,015 pneumococci sequences and isolates collected from 1991 to 2017 [563]. The majority of isolates were sampled from cases of invasive pneumococcal disease (IPD) in children under the age of five. In total, 49.7% of these isolates were collected from locations before the PCV7 vaccine was introduced. The collection is split into 621 separate GPSC lineages, of which 35 are represented by more than 100 isolates.

3.2 Methods

3.2.1 Isolate collection and sequencing

Isolates belonging to the PMEN3 and PMEN9 lineages were collated from across Europe, the Americas and the Maela refugee camp in Thailand [580] (Table 3.1). These datasets had multi-locus sequence typing (MLST) data [581]. Therefore isolates of sequence type (ST) 156, and single locus variants thereof, were selected as representatives of PMEN3 (or Spain^{9V}-3) [582]. For PMEN9, isolates of the ST9 grouping were selected as representative of the lineage from these collections. This generated collections of 272 isolates for PMEN3 and 325 isolates for PMEN9. Isolates that could be cultured were sequenced as paired-end 24-plex libraries on Illumina HiSeq 2000 machines, generating 75 nt reads. Sample identity was checked through comparing serotype, inferred by seroba v1.0.0 [583], and ST with those determined by sample providers. Samples were checked for contamination through assessing assembly statistics, as described previously [142]. After these tests, 215 PMEN3 isolates and 263 PMEN9 isolates were passed for use in the described analyses.

These datasets was then combined with isolates from the Global Pneumococcal Sequencing (GPS) project [563]. PMEN3 corresponded to strains GPSC6 (454 isolates) in this collection, while PMEN9 GPSC18 (312 isolates) [298]. Hence the final dataset size for PMEN3 was 669 isolates and for PMEN9 575 isolates. WGS data from the 478 isolates not within the GPS collection are publicly available in the EMBL Nucleotide Sequence Database (ENA; Project number PRJEB2255).

Source	PMEN3	PMEN9
ANSORP	1	3
CDC	124	42
GPS	454	312
Herminia de Lencastre, ITQB NOVA	17	9
Maela refugee	5	0
Nationales Referenzzentrum für Streptokokken, Germany	68	208
Timothy Mitchell, Glasgow University	0	1
Totals	669	575

Table 3.1: Isolate sources for PMEN3 and PMEN9

3.2.2 Generation of annotations and alignments

De novo assemblies were generated using an automated pipeline for Illumina sequences [584]. Briefly, reads were assembled using Velvet with parameters selected by VelvetOptimiser. These draft assemblies were then improved by using SSPACE and GapFiller to join contigs [585–587]. The final assemblies were annotated using PROKKA [588].

Whole genome alignments were generated for phylogenetic analysis through mapping of short read data against reference sequences. For the PMEN3 and PMEN9 analyses, the reference genomes were *S. pneumoniae* RMV4 *rpsL** $\triangle$*tvrR* (accession code: ERS1681526) [545] and INV200 (accession code: FQ312029.1) respectively. Mapping was performed using SMALT v0.64, the GATK indel alignment toolkit and SAMtools as described previously [466].

A faster method for generating alignments was applied to GPSCs represented by more than ten isolates. A reference sequence was was chosen as the isolate with the largest N_{50} value (the length of the contig at the midpoint of the assembly, when contigs are ordered by size). Other isolates were mapped to this reference using SKA [546] with default settings.

3.2.3 Phylogenetic and phylodynamic analyses

The phylogeny and recombination patterns within PMEN3, PMEN9 and resistance-associated GPSCs were estimated by running Gubbins v2.3.0 [589]. Starting trees were formed with FastTree 2 [529]. Subsequent iterations generated phylogenies with RAxML v8.2.8 [530], with a generalized time reversible (GTR) model of nucleotide substitution with a discre-

tised gamma distribution of rates across sites. Marginal ancestral state reconstruction was used.

Time-calibrated phylogenies for PMEN3 and PMEN9 were generated from the Gubbins outputs using the BactDating R package v1.0.1 [552]. Isolates without dates of collection were pruned from the phylogeny, and the root-to-tip distances used to test for a molecular clock signal. Where one was detectable, BactDating was run with a relaxed clock model and a MCMC length of 50 million iterations. Chain convergence was checked through visual inspection of trace plots.

Serotype switching among the isolates was assessed by reconstructing the ancestral serotype using the PastML python package [590]. This was run on the phylogenies produced by Gubbins.

3.2.4 Antibiotic resistance analyses

3.2.4.1 Penicillin resistance

The MIC for penicillin had been determined for most isolates in the PMEN3 and PMEN9 collections (65.5 and 80.9%, respectively). The MICs of the remaining 341 isolates, and all GPS isolates, were predicted using an RF approach analogous to that developed in Li et al 2017 [270]. Model choice is elaborated further in section 3.3.4.1. The RF model was implicated using the ranger package v0.12.1 [591] in R.

To assess the emergence time of resistance and its spread among the PMEN3 lineage, the penicillin resistance categories inferred from the metadata and the RF model were reconstructed on the time-calibrated phylogeny using the phytools R package v0.7.7 [592]. The make.simmap function was run using an equal rates model and an MCMC chain sampling every 100 iterations. The input was a matrix of character states for the tips, with their observed or predicted phenotypes were assigned a probability of one.

After the reconstruction, each node's state was assigned as that with the highest posterior probability. Starting at the root, the number of lineages of each state at each coalescent event in the time-calibrated tree was recorded. Every time a node was reached an extra lineage was added to the total. If there was no state change between two nodes the count for the state was increased by one; else if there was a state change not on a terminal branch the count for the new state was increased by two, if there was a state change on a terminal branch the new state count was increased by one. The total number and the proportion of branches in each state were recorded through time.

3.2.4.2 Co-trimoxazole resistance

Resistance to sulfamethoxazole was detected using a hidden Markov model (HMM), constructed using HMMer3 [593], trained to extract the region downstream of S61 in *folP*. If this region contained at least one inserted amino acid, then the isolate was predicted to be resistant. Resistance to trimethoprim used another HMM to identify the amino acid at position 100 in *dhfR* (also known as *dyr* or *folK*). Isolates with an isoleucine at this position were predicted to be sensitive, and isolates with a leucine at this position were assumed to be resistant [594]. If isolates were classified as resistant to both sulfamethoxazole and

trimethoprim, they were also classified as resistant to the combination drug cotrimoxazole [595].

3.2.5 Detecting interspecies recombination events

3.2.5.1 Pipeline for detecting interspecies events at the *pbp* loci

To assess the origin of resistant *pbp* genes, a pipeline was developed to compare sequences against a reference database. The first step in this pipeline was to reconstruct the ancestral resistance state for the isolates in the PMEN3, PMEN9 and GPS collection. This ancestral state reconstruction was performed using PastML [590], as these Gubbins phylogenies were not time-calibrated. The recombination predictions from Gubbins were then searched, in order to detect whether there was a putative recombination event on the branch on which the change in resistance profile spanning any of the *pbp* genes. If such a recombination were identified, this was considered indicative of resistance state alteration via homologous recombination. The descendants of nodes where resistance was acquired or lost with the fewest base substitutions subsequently accumulating in the three *pbp* genes then had their gene sequences extracted.

These gene sequences were compared to a reference collection of 52 streptococcal genomes collated from antimicrobial susceptible *S. pneumoniae* and other *Streptococcus* species, building on the database collated in Mostowy *et al* 2017 [526]. BLASTN v2.5.0 was used to compare each gene region to this database.

The statistic γ was used to determine the species of origin for a gene. This utilised the BLAST bit score, which is a normalized form of the raw score of an alignment [596]. The bit score measure sequence similarity irrespective of query sequence length and database size. The γ statistic was calculated as the bitscore of the top ranked *S. pneumoniae* hit (b) divided by the bitscore of the top ranked hit (B):

$$\gamma = \frac{b}{B} \tag{3.1}$$

Hits where the top match was *S. pneumoniae*, indicating the gene originated from an intraspecies transformation event, had a γ score of 1.0. Any score below 1.0 indicated a potential origin from outside of *S. pneumoniae*.

3.2.5.2 Detecting interspecies recombination at *murM*

For *murM*, where the effects of alterations on resistance levels are less well understood, a different approach was taken. The regions corresponding to the *murM* genes in the annotated references were extracted from the PMEN3 and PMEN9 whole-genome alignments. To enable the detection of possible interspecies recombinations, *murM* sequences from *S. mitis* 21/39 (accession code: AYRR01000000) and *Streptococcus pseudopneumoniae* IS7493 (accession code: CP002925) were added to the dataset. All *murM* sequences were then aligned with Muscle v3.8.31 [597]. Sequences were clustered into lineages, and recombinations inferred, using fastGEAR [526].

3.3 Results

3.3.1 Genomic epidemiology of the PMEN3 and PMEN9 lineages

A recombination corrected phylogeny of the PMEN3 lineage was created using Gubbins v2.3.0. This represents the evolution of the lineage and was constructed from 669 isolates which were collected from 31 countries over 23 years (1992 to 2015; Figure 3.1). This range of collection years was sufficient for the estimation of a molecular clock, performed using the BactDating R package v1.0.1 [552] (Figure 3.2). Six isolates without dates of collection were pruned from the phylogeny, and the root-to-tip distances used to test for a molecular clock signal. This estimated the most recent common ancestor (MRCA) existed in 1942 (95% credible interval of 1910 to 1959) and the lineage had a molecular clock rate of 1.69 x 10^{-6} substitutions per site per year (95% credible interval of 1.52 x 10^{-6} to 1.86 x 10^{-6} substitutions per site per year).

The PMEN3 phylogeny was dominated by the 491 isolate ST156 clade, which was found in 27 countries mainly from Europe (85 isolates), North America (90 isolates) and South America (192 isolates). There is also a smaller, 33 isolate clade of ST143 isolates, which was found in eight countries, primarily being present in Poland (11 isolates) and Belarus (8 isolates).

Most of the PMEN3 isolates were either of the ancestral serotype 9V, or serotype 14, with changes between these two serotypes accounting for 9 of 35 serotype switches

reconstructed within the clade (Figure 3.3). Both of these serotypes were targeted by the heptavalent polysaccharide conjugate vaccine (PCV7) vaccine. However, within ST156 a clade of 26 isolates from the USA of serotype 19A, not included in PCV7, were derived from a MRCA estimated to exist in 2000 (95% credible interval of 1999 to 2001). This coincides with the date of PCV7's introduction into the USA, consistent with these switched isolates evading the vaccine and persisting until the 13-valent conjugate vaccine (PCV13), which includes 19A, was introduced [576]. In total 13 serotypes were found in the PMEN3 lineage, of which seven (11A, 13, 15A, 15B/C, 23A, 23B 35B) are not found in the PCV13 vaccine.

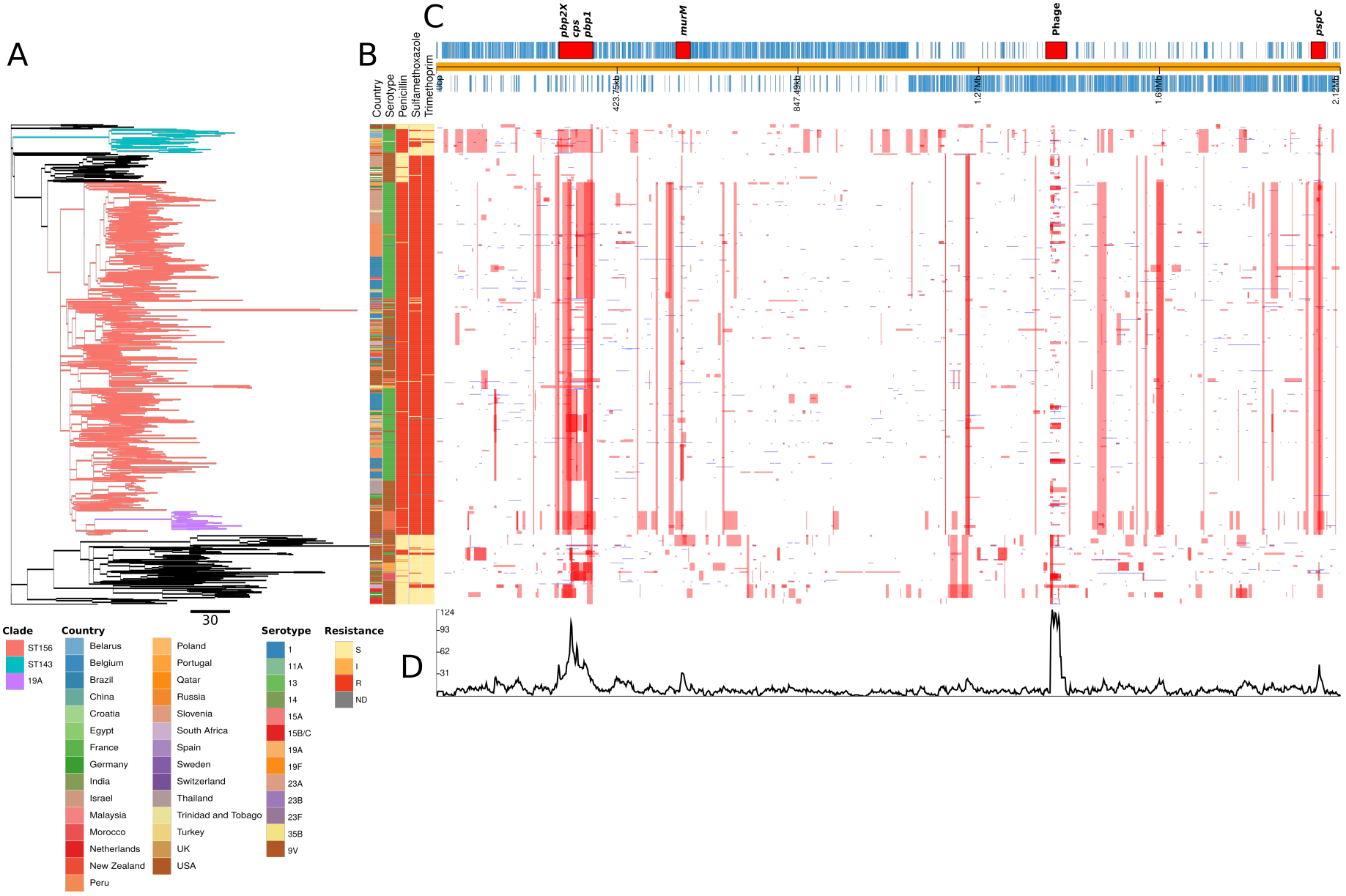

Figure 3.1: Phylogenomic analysis of the PMEN3 lineage. (A) Maximum likelihood phylogeny generated from the non-recombinant regions of the 669 isolates from the PMEN3 lineage. Branches are coloured by clade, as defined in the key. Units for the scale bar are the number of point mutations along a branch. **(B)** Bars highlighting the country of origin, serotype and resistance categories to penicillin, trimethoprim and sulfamethoxazole. Bars map across to isolates on the phylogeny. **(C)** Simplified genome annotation of the PMEN3 reference isolate RMV4. The highlighted regions correspond to peaks of recombination event frequency. Blue bars represent individual genes annotated within the assembly. **(D)** Distribution of recombination events across the PMEN3 lineage. In the upper half of the graph, red bars indicate recombination events occurring on internal nodes in the tree, which were subsequently inherited by multiple descendent isolates. These bars map across to isolates in the phylogeny in section **A** and map to regions in the genome annotated in section **C**. Blue bars indicate recombination events on terminal branches of the tree, occurring in only one isolate. In the bottom half of the graph, the line represents the frequency of recombination events along the genome's length.

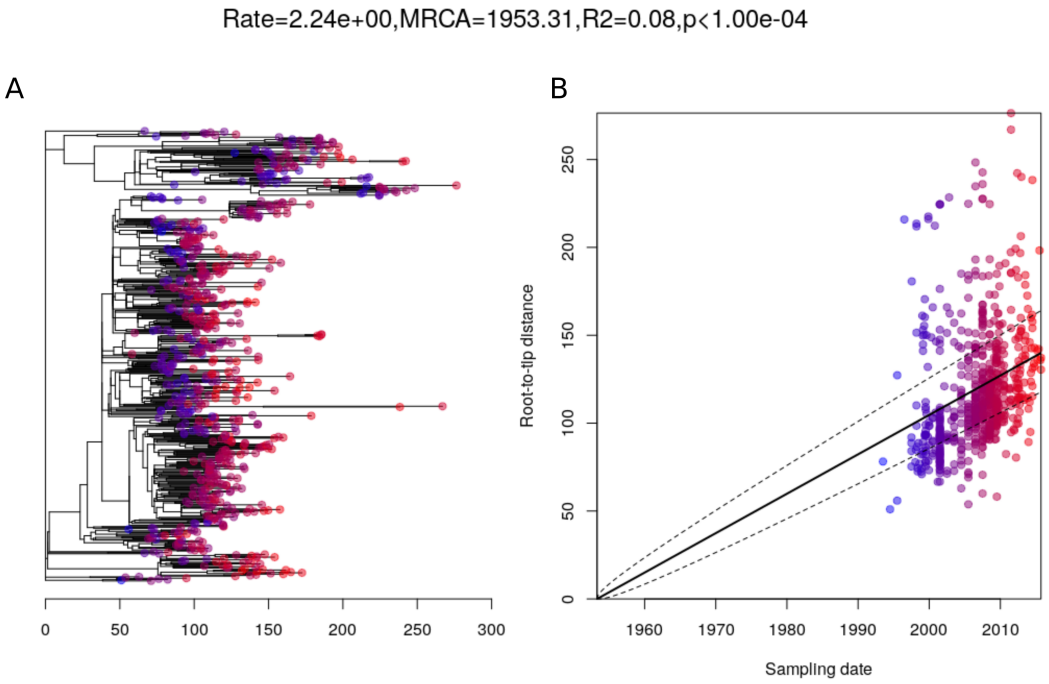

Figure 3.2: Root to tip analysis of PMEN3 lineage. **A** Represents the 663 isolate phylogeny, identical topology to Figure 1, with node tips coloured by date of isolation. **B** Linear regression of root to tip distance against sampling date for Isolates.

In contrast to PMEN3, the phylogeny representing the evolution of PMEN9, constructed from isolates of GPSC18, was split into multiple clades separated by deep branches (Figure 3.4). Even when excluding the outlying serotype 7C isolates, the only discernible molecular clock signal suggested this strain was centuries old (Figure 3.5). Despite this age, the individual clades were generally regionally confined. The largest clade was associated with Germany (accounting for 166 of the 250 isolates), with other representatives from Slovenia and China. Other clades were associated with the USA (accounting for 91 of the 98 isolates), South Africa (accounting for 68 of the 73 isolates), and China (accounting for 18 of 45 isolates).

All the isolates in the three largest clades expressed serotype 14, as did 93% of all isolates in this phylogeny. Only nine serotype switches were identified across PMEN9, including switches to 19F and 23F in the Chinese clade (Figure 3.6). In total there were six serotypes present within the collection, only two of which (16F and 7C) were not found in the PCV13 vaccine. Overall, there was little evidence of frequent intercontinental transmission or serotype diversification with this set of isolates. Hence genomics suggests dif-

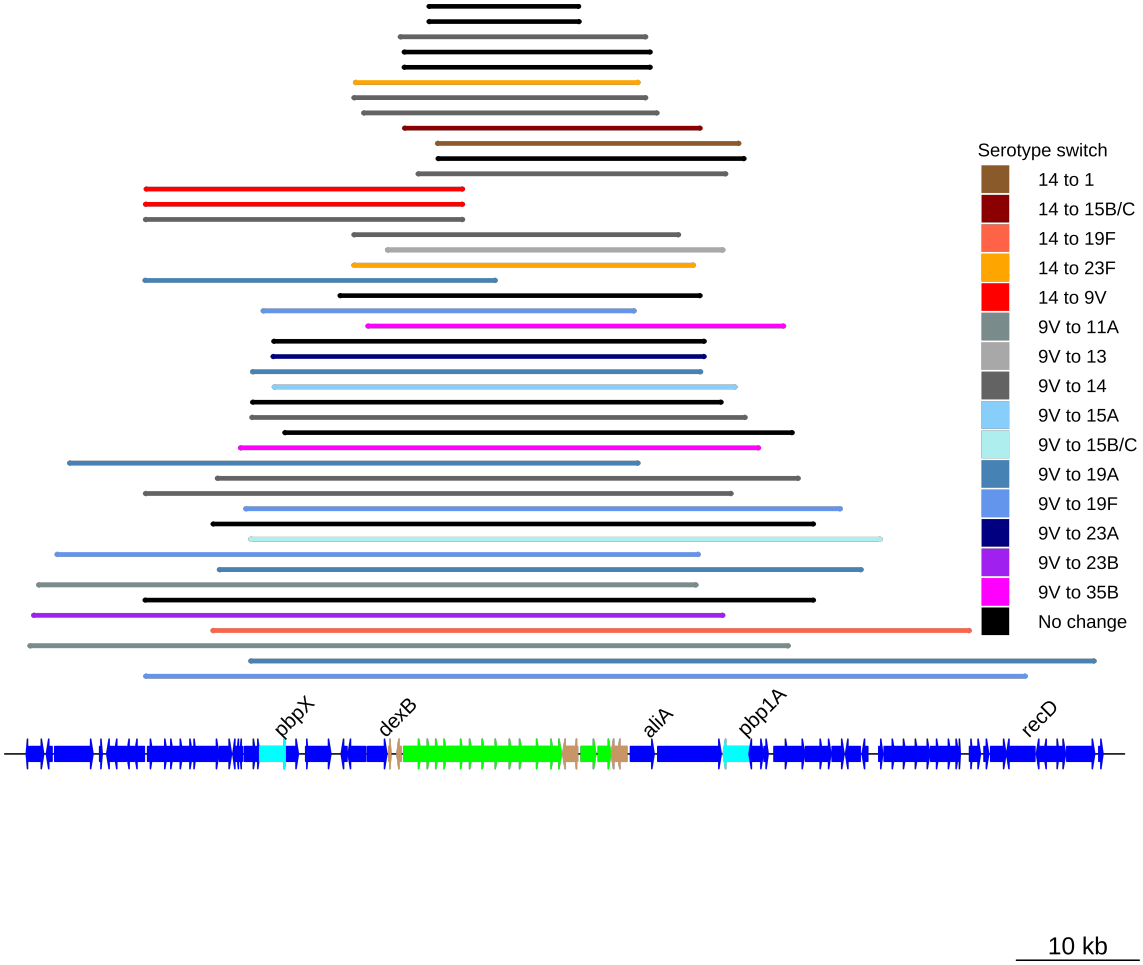

Figure 3.3: Serotype switching events across the PMEN3 lineage. Coloured bars represent the recombination events associated with switches in serotype inferred from the phylogeny. The bars map to the genome annotation below, with the length of the bar indicating the size of the recombination event. The *cps* and surrounding loci are highlighted below, with some recombination events spanning the *pbp1a* and *pbp2x* genes as well. The black bars represent recombinations across the *cps* loci that were not associated with a serotype switch.

fering histories for PMEN3 and PMEN9, despite them both being internationally disseminated antibiotic-resistant *S. pneumoniae*, commonly expressing the invasive serotype 14, with identical sampling approaches.

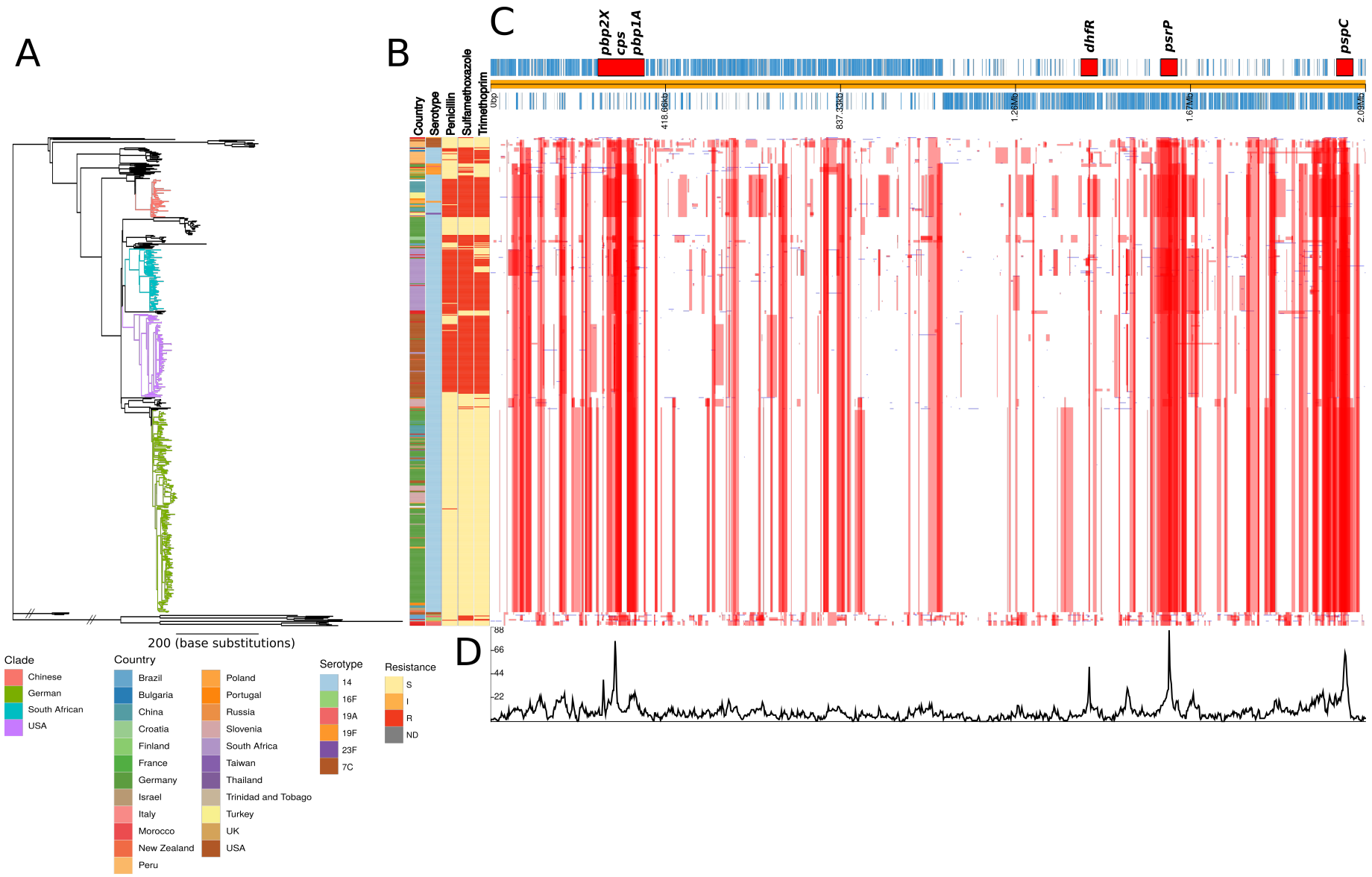

Figure 3.4: Phylogenomic analysis of PMEN9 lineage. (A) Maximum likelihood phylogeny of 575 isolates, generated from the non-recombinant sequences of the PMEN9 lineage. Branches are coloured by clade identified in the key. Units for the scale bar are the number of point mutations along a branch. **(B)** Bars highlighting the country of origin, serotype and resistance categories to penicillin, trimethoprim and sulfamethoxazole. Bars map across to isolates on the phylogeny. **(C)** Simplified annotated genome of the PMEN9 reference isolate INV200. The highlighted regions correspond to peaks of recombination event frequency. Blue bars represent individual genes annotated within the assembly. **(D)** Distribution of recombination events across the PMEN9 lineage. In the upper half of the graph, red bars indicate recombination events occurring on internal nodes in the tree, which are subsequently inherited by multiple descendent isolates. These bars map across to isolates in the phylogeny in section **A** and map to regions in the genome annotated in section **C**. Blue bars indicate recombination events on terminal nodes of the tree, occurring in only one isolate. In the bottom half of the graph, the line represents the frequency of recombination events along the genome's length.

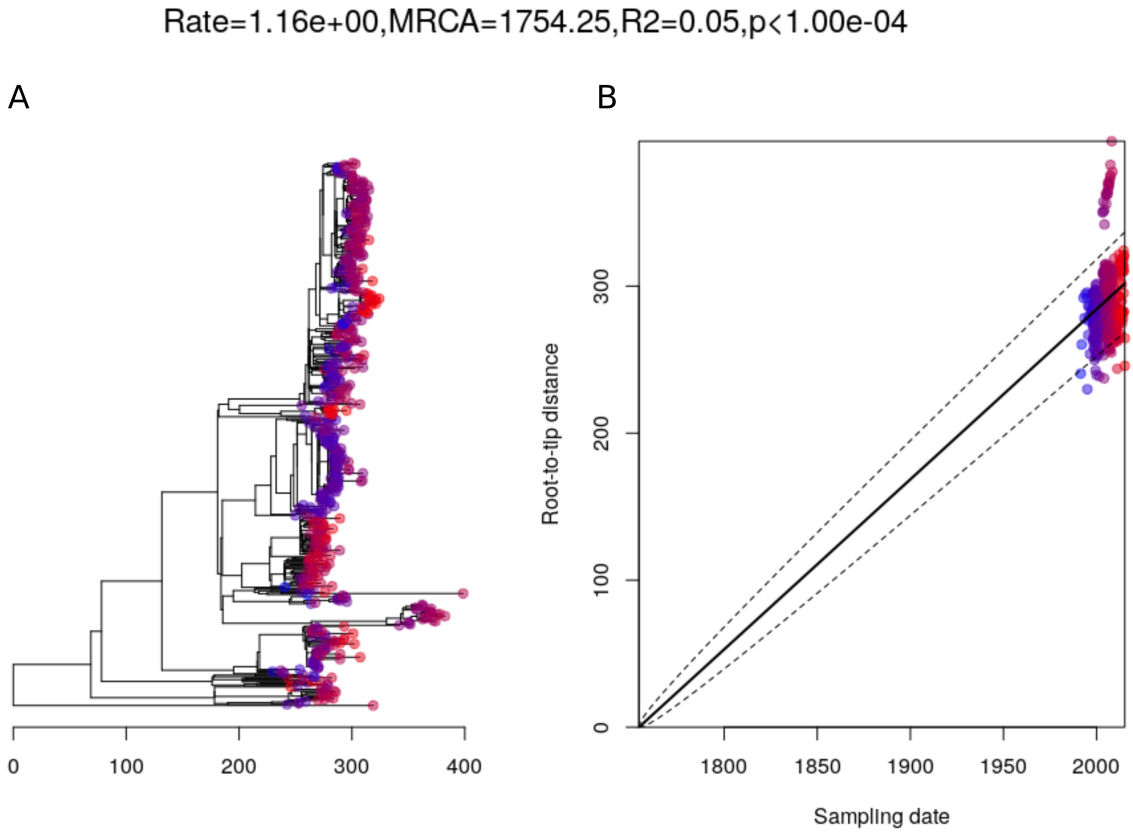

Figure 3.5: Root to tip analysis of the PMEN9 lineage. A Represents the trimmed 529 isolate phylogeny (those isolates with date of isolation data and not present on long branches) with node tips coloured by date of isolation. **B** Linear regression of root to tip distance against sampling date for Isolates

3.3.2 Variable recombination dynamics across the PMEN lineages

The two lineages also differed in the patterns of recombination across their genomes. In the PMEN3 reference genome, there is a high density of recombinations around a 45 kb prophage region, indicating frequent infection by phage. Exclusion of these recombination events allowed estimation of the overall ratios of base substitutions resulting from homologous recombination relative to point mutations (r/m). Consistent with its more rapid serological diversification, r/m was higher in PMEN3 (13.1) than PMEN9 (7.7).

This difference in r/m could be due to the two lineages differing in three ways: (i) in the number of recombinations, (ii) in the length of recombination events or (iii) in the sources of their recombination events, with more divergent sources increasing the r/m.

The first explanation partially accounted for the difference: there were 0.115 recom-

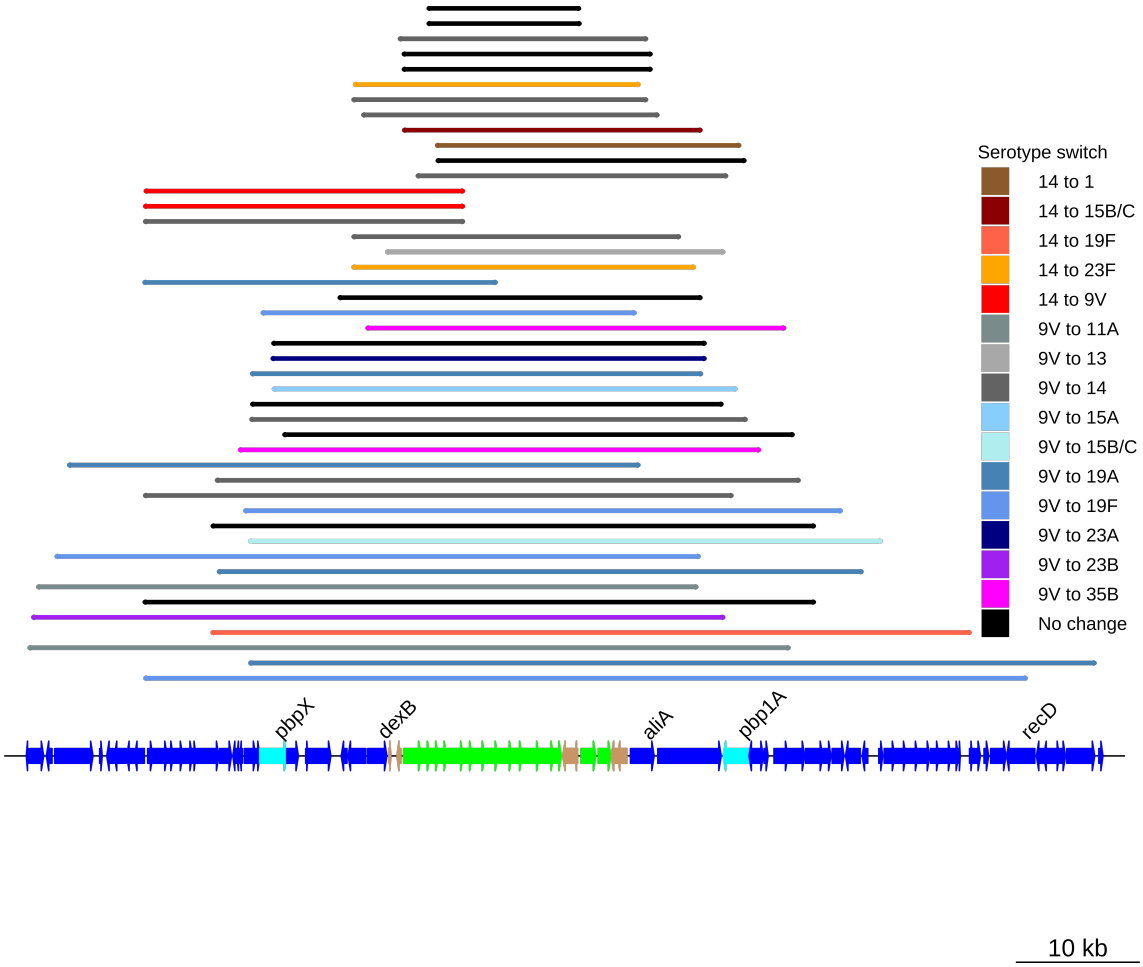

Figure 3.6: Serotype switching events across the PMEN9 lineage. Coloured bars represent the recombination events associated with these switches in serotype. The bars map to the genome annotation below, with the length of the bar indicating the size of the recombination event. The *cps* and surrounding loci are highlighted below, with some recombination events spanning the *pbp1a* and *pbp2x* genes as well. The black bars represent recombinations across the cps loci that were not associated with a serotype switch.

binations per point mutation in the PMEN3 reconstruction, compared to 0.093 per points mutation in PMEN9. Comparing the properties of the recombination events revealed no substantial difference in their length distribution (Figure 3.7). However, PMEN3 generally imported sequences with a significantly higher SNP density, with a median SNP density of 11.8 SNP/kb of sequence imported, compared to PMEN9, which had a median SNP density of 9.2 SNP/kb (MannWhitney U = 4162888, n1 = 2613, n2 = 2823, two-sided, $p < 2.2 \times 10^{-16}$). Therefore, the difference in r/m between the two lineages reflected both the increased frequency of recombination in the PMEN3 lineage, and the increased diversity of the imported sequence.

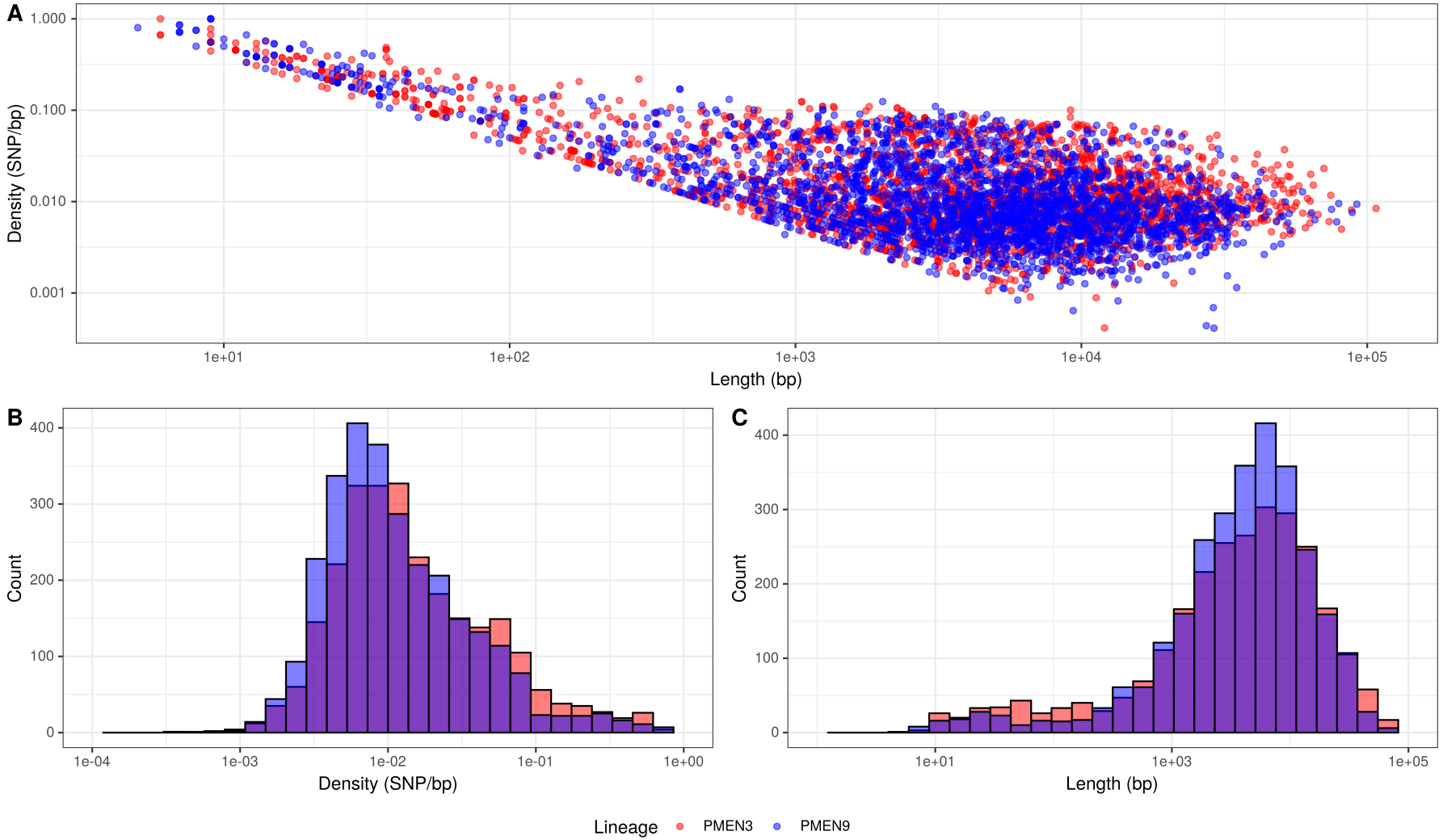

Figure 3.7: Summary of recombination differences between the PMEN3 and PMEN9 lineages. The purple colouration of dots and bars represents the overlap between PMEN3 and PMEN9. **(A)** Plotting the SNP density against the length of recombination events across both lineages. Dots represent individual recombination events and are coloured by the lineage in which they were inferred. **(B)** Overlaid histograms of the SNP density per recombination event in the PMEN collections. **(C)** Overlaid histograms of the length of each recombination event across the PMEN collections.

3.3.3 Hotspots of recombination

The high-level of serotype switching in PMEN3 corresponded to a peak in recombination events detected around the *cps* loci which determine an isolate's serotype (Figure 3.1) [598, 599]. Gubbins predicted that all 35 of the reconstructed serotype switches had an accompanying recombination event covering the *cps* locus (Figure 3.3). The median recombination block size across the 20 kb *cps* locus was 37.5 kb in length for these switch associated events. These recombination blocks, therefore, also frequently encompassed the nearby *pbp1a* and *pbp2x* genes, encoding PBPs involved in penicillin resistance. In total 77% (27 of 35) *cps* switch recombination events also spanned at least one of *pbp2x* or *pbp1a*. Another peak in recombination occured around the *murM* gene in PMEN3. This encodes an enzyme involved in the biosynthesis of branched structured muropeptide components of the pneumococci cell wall [600]. *murMN* has been implicated in mediating penicillin resistance, with deletion of the operon *murMN* containing this gene leading to isolates becoming penicillin susceptible [140, 601].

The *cps* loci and *pbp* genes were also hotspots of recombination for the PMEN9 lineage. In PMEN9, 78% (7 of 9) of serotype switches had accompanying recombination events spanning the *cps* loci. These tended to be smaller than PMEN3 *cps* recombinations, with a median length of 22 kb. As such fewer of these events, only 43% (3 of 7), spanned the *pbp2x* or *pbp1a* genes. PMEN9 also has a peak in recombination around the *dhfR* gene. The sequence of *dhfR* determines resistance to trimethoprim, one of the two components (along with sulfamethoxazole) of co-trimoxazole [17].

Finally for both lineages there are several peaks of recombination within the chromosome correspond to loci likely to be under immune selection. In PMEN9, there was an elevated density of recombinations affecting the *psrP* gene, encoding the antigenic pneumococcal serine-rich repeat surface protein. Serine-rich repeat proteins (SRRPs) are known to act as adhesins in a range of different bacteria, enabling the binding to host tissues and thus playing a key role in the invasiveness of cells [602, 603]. In the pneumococcus, for instance, *psrP* has been shown to increase the virulence of cells within a mouse model [604]. Another peak, this time present in both lineages, was seen around the antigenic Pneumococcal Surface Protein C, encoded by *pspC*. PspC is among the

most variable microbial immune evasion proteins identified and plays a key role in binding several human plasma proteins involved in the immune response [605]. Both *pspC* and *psrP* are highly diverse in pneumococcal populations, and elicit strong immune responses from hosts [606].

3.3.4 Investigating β-lactam resistance emergence in the pneumococcus

3.3.4.1 Determining β-lactam resistance levels

With both lineages containing hotspots of recombination around the *pbp1a* and *pbp2x* genes, I decided to further investigate the evolution of β-lactam resistance in these lineages. The first step in this was to try and accurately determine the levels of resistance in the populations. A majority of isolates in both PMEN3 and PMEN9 (65.5% and 80.9% respectively) had MICs for penicillin determined previously via broth dilution methods. This left the MICs of 342 isolates to be predicted.

In the pneumococcus, given β-lactam resistance is determined largely by the sequence of three *pbp* genes, *in silico* model based predictions directly from the genotype have been implemented successfully [228]. Arguably the most established method for this is the RF approach developed in Li *et al* 2016 and Li *et al* 2017 [228,269,270]. In this protocol, the Transpeptidase domains (TPD) of three PBPs (PBP1A, PBP2B & PBP2X) are extracted and each amino acid position used as a predictor to train an RF model on the continuous $\log_2$ MIC value. Bases not present in the training data are then approximated using the base with the least distance to a training data base in the BLOSUM62 matrix [607]. The output MIC values are then converted into resistance breakpoints, in Li *et al* 2017 these correspond to the CLSI breakpoints for non-meningitis infections [270]. The RF model first used in Li *et al* 2016 [269] has compared favourably to an elastic net model for the prediction of MIC values and, then subsequently, resistant breakpoints.

The training data for the Li *et al* 2017 model came from 2,528 isolates previously characterised by the CDC [608]. These isolates were largely sampled from the Americas, whereas the GPS collections and the PMEN3 and PMEN9 lineages I investigated came from across the world. I therefore set out to test the accuracy of this RF method on the PMEN collections, implementing the protocol from Li *et al* 2017 in R using the Ranger

package (v0.13.1) for RF modelling. I also assessed the accuracy of a linear regression (LR) model as a comparison to the RF protocol. This LR model was also trained on the TPD domains and predicted the $\log_2$ MIC values, following the RF protocol, and was also implemented in R.

In both methods, the models were first validated on an expanded 4,342 isolate CDC dataset [608], with training performed on a random subset of 70% of the data and then testing on the unseen remaining 30% of the data. The models were then trained on the full CDC dataset and tested on the 902 PMEN isolates with MIC values previously measured through broth dilution. Two different breakpoint categories were used for the category agreement (CA) scores: the wider CLSI non-meningitis breakpoints (Susceptible $\leq$0.06 μg/ml, 0.06 μg/ml < Intermediate < 2 μg/ml, Resistant $\geq$2 μg/ml) and the narrower pre-2008 meningitis breakpoints for resistance (Susceptible $\leq$0.06 μg/ml, 0.06 μg/ml < Intermediate < 0.12 μg/ml, Resistant $\geq$0.12 μg/ml) [609].

The implementation of the RF model was able to broadly match the CA scores produced in Li *et al* 2016 for the CLSI non-meningitis breakpoints (Figure 3.8). For the RF model, the CA on the training set of CDC data is 93.7%, while for the unseen testing dataset it is 92.7%. The LR model also has similar scores for its training and testing CA results, at 94% and 91.9% respectively. However both models perform much worse when using the non-meningitis breakpoints on the unobserved PMEN data, dropping to 74.6% CA and 70.5% CA for the RF and LR models respectively. This drop-off is not seen though for the meningitis breakpoints (Figure 3.8). For the LR model the CA on the training dataset is 95.6%, 94.9% on the testing CDC data, and then 92.4% for the PMEN dataset. While for the RF model the CA is 95.4% for the training dataset, 94.7% for the testing CDC data, and then 95.6% for the PMEN dataset.

The reduction in CA for the non-meningitis breakpoints is likely reflective of the diversity of PBP TPBs within the global PMEN3 and PMEN9 datasets not being captured within the CDC dataset of isolates from the Americas. For instance, for the testing dataset, only 27 amino acids at 912 TPB sites are altered via the BLOSUM62 matrix. Whereas for the PMEN testing data 44 amino acids have to be altered through the BLOSUM62 matrix to match those present in the trained model. The narrower meningitis breakpoints on the

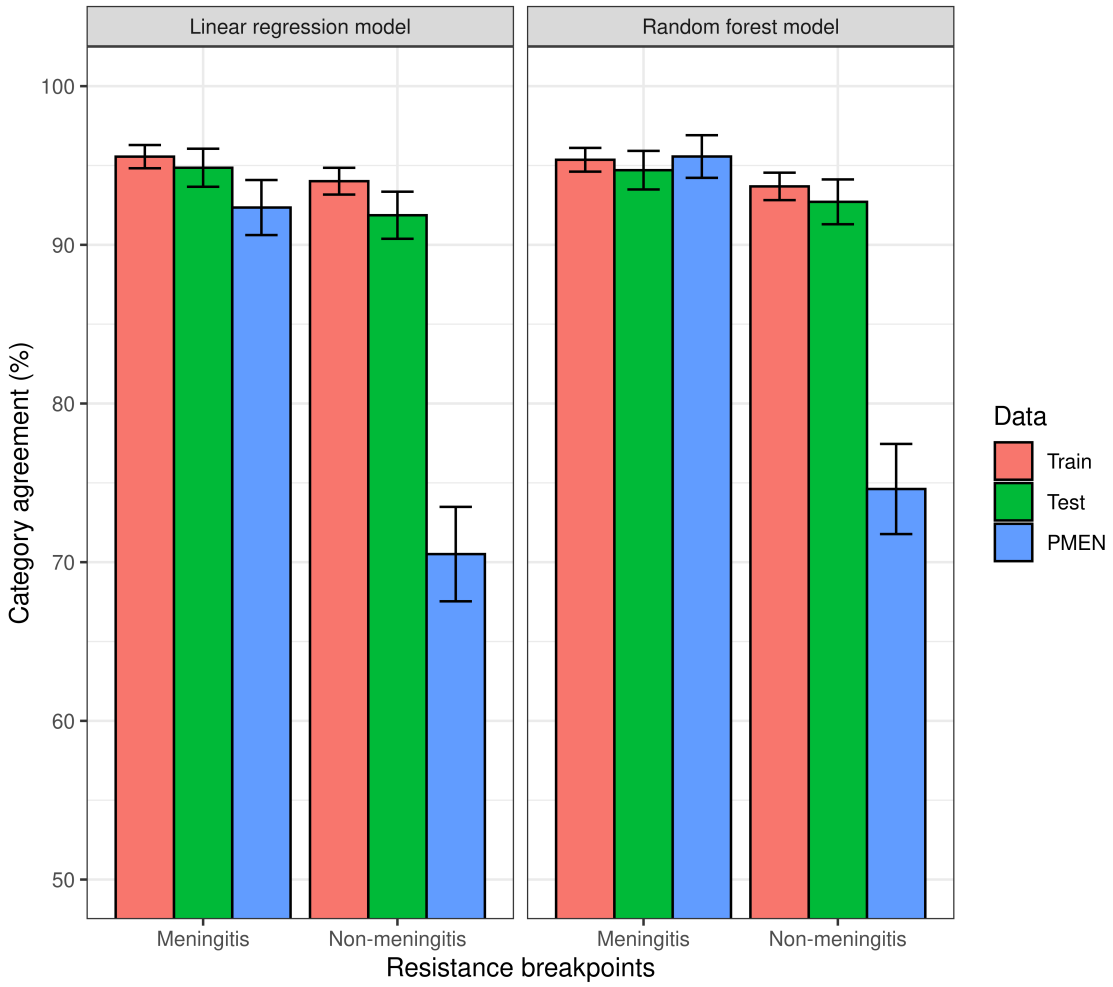

Figure 3.8: Penicillin resistance prediction using different models. Category agreement scores for a linear regression model and a random forest model trained on TPD domains of PBPs. Bars represent the number of categories correctly predicted across the three datasets the model was applied to, with results split by the breakpoints used to assess the model predictions.

other hand greatly reduce the size of the intermediate resistant category, thus reducing the potential for misclassification, while still being clinically relevant.

Given its higher accuracy in predicting the resistance categories of the PMEN datatset, the RF model with the meningitis breakpoints was used to predict the penicillin resistance phenotype of isolates with unknown MIC values. This model predicted the MIC for 341 of the 342 isolates with unknown MIC values in the PMEN dataset, with the one missing isolate having unidentifiable *pbp* genes. This method was then also used on the wider 20,015 GPS collection, with 59 isolates (0.3%) having unidentifiable *pbp* genes.

3.3.4.2 Emergence of β-lactam resistance in PMEN3 and PMEN9

In the PMEN9 collection, 61% of isolates were susceptible to penicillin (recorded or predicted MIC ≤ 0.06 μg/ml; Figure 3.4). However, 79% of the PMEN3 collection was classed as resistant (MIC ≥ 0.12 μg/ml), with 20% susceptible to penicillin, and the remaining 1% classified as intermediately resistant (0.06 μg/ml < MIC < 0.12 μg/ml; Figure 3.1).

Across the two PMEN lineages, there were 35 changes in resistance profile for penicillin. The most common alteration was acquisition of resistance by sensitive isolates, with 16 instances across the two lineages (45% of events). There were also seven instances of resistant isolates reverting to penicillin sensitivity across the collections. In 20 of the 35 alterations in resistance profile, the evolutionary reconstruction identified at least one of the three resistance-associated *pbp* genes was altered by a concomitant recombination event.

In PMEN3, penicillin resistance was common across the ST156 clade, 99% of which was penicillin resistant. Recombinations altered *pbp1a*, *pbp2b* and *pbp2x* at the base of this clade (Figure 3.1). Phylodynamic analysis was performed to assess the likely date of origin of this penicillin resistant clade. This analysis showed the penicillin-resistant proportion of GPSC6 increased throughout the early 1980s (Figure 3.9). This is driven by the expansion of the ST156 clade, which originated around 1984 (95% credible interval 1981 to 1986). This expansion of resistant lineages continued until the early 2000s, when it then plateaued within the strain from roughly 2010 onward.

The highest MICs within the ST156 clade (up to 8 μg/ml) were associated with the

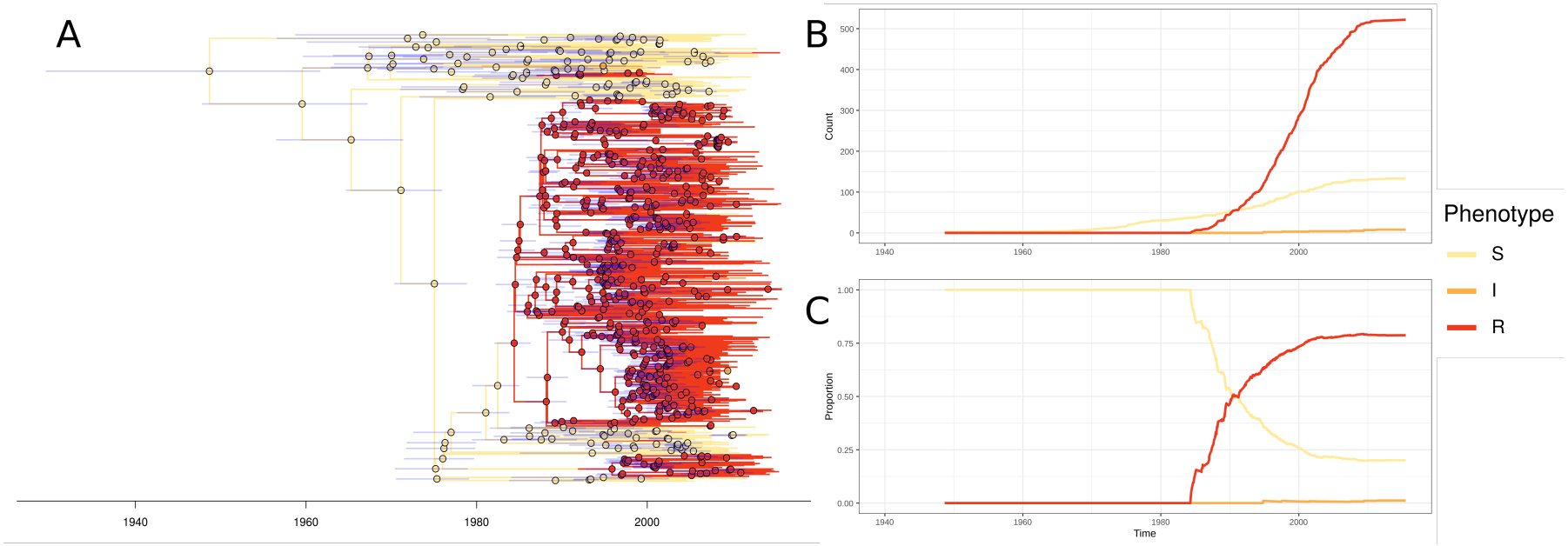

Figure 3.9: PMEN3 resistant lineages through time. (A) Time calibrated phylogeny of PMEN3. Branches are coloured by inferred resistant phenotype. S corresponds to sensitive, I to intermediately resistant and R to resistant. Pie charts present at nodes represent the inferred probability of each state by the ancestral reconstruction. Blue bars across the nodes represent the 95% credible interval for the age of the node. **(B)** The reconstructed absolute number of branches per resistant phenotype through time. **(C)** The proportion of total branch over time in either of the four states.

vaccine escape 19A clade of isolates from the USA (Figure 3.10). This was a consequence of a 53 kb recombination spanning the *cps* locus, which caused the alteration in serotype, also spanning *pbp1a* and *pbp2x* (Figure 3.3). Hence the PCV7-escape recombination also reduced susceptibility to antibiotics. The converse situation was observed for a single ST156 clade member that had reverted to susceptibility. A 53 kb recombination event, causing a switch from serotype 9V to 15B/C (Figure 3.3), restored the ancestral, susceptible versions of *pbp2x* and *pbp1a*.

In contrast to the ST156 clade, in PMEN9, penicillin resistance emerged independently in different locations. The USA and South African clades appear to have both separately gained resistance in a stepwise manner. At the base of the highly resistant USA clade there was a 3.2 kb recombination spanning the *pbp2x* gene (Figure 3.10). Subsequent recombinations modifying the *pbp1a*, then *pbp2b*, genes further increased penicillin resistance. Similarly, within the South African clade, *pbp2b* and *pbp2x* were both modified by recombination in the isolates' most recent common ancestor. Resistance was elevated in a subset of isolates through further modification of *pbp1a* through recombination. Alteration in the *pbp2x* and *pbp2b* genes are the first steps towards resistance, with *pbp1a* modifications required for higher levels of resistance, although isolates with solely a mosaic *pbp2x* gene have been found to be resistant to penicillin [610]. In general, I observed penicillin resistance rapidly emerged and spread worldwide in PMEN3, whereas PMEN9 exhibits repeated, localised stepwise acquisition of modified *pbp* genes.

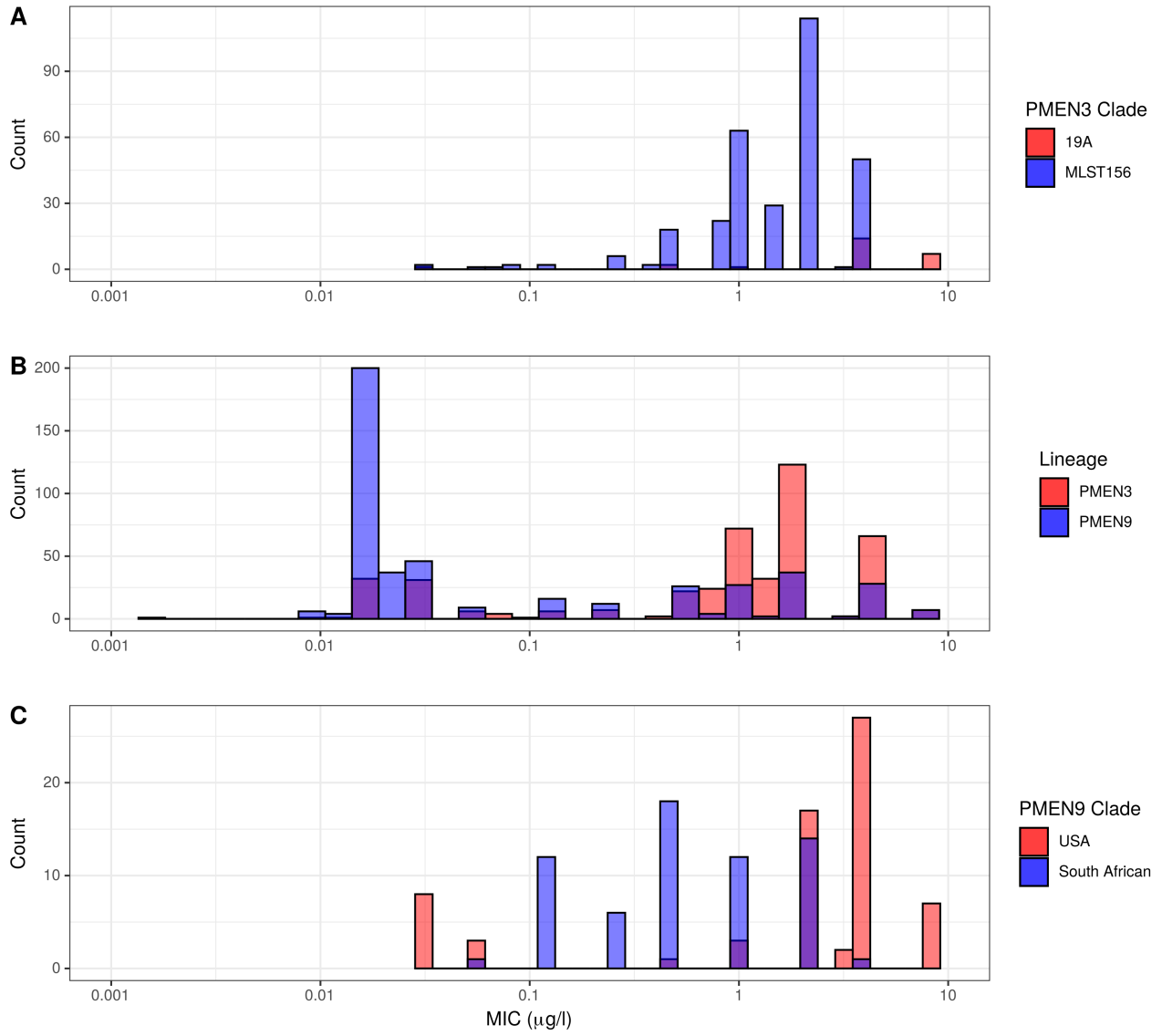

Figure 3.10: Histograms of recorded MIC values for penicillin across the PMEN3 and PMEN9 lineages. A purple colouration through-out represents the overlap between the clades/lineages highlighted in each key. A MIC values for penicillin for the primarily resistant clades ST156 and 19A within the PMEN3 lineage (N = 313 for the ST156 clade and N = 25 for the 19A clade). **B** MIC for penicillin across both PMEN3 and PMEN9 (N = 438 for PMEN3 and N = 465 for PMEN9). **C** MIC values for penicillin for the primarily resistant USA and South African clades within the PMEN9 lineage (N = 68 for the USA clade and N = 64 for the South African clade).

3.3.4.3 Role of interspecies transformation in β-lactam resistance in PMEN3 and PMEN9

As penicillin resistance was originally demonstrated to involve the acquisition of sequence from related commensal streptococci [137, 138, 560], I decided to further investigate the origin of these *pbp* genes within recombination events. To do this an analysis pipeline was developed to extract and assess the likely origin of resistant *pbp* genes (See Methods Section 3.2.5.1).

This pipeline was applied to the PMEN3, PMEN9 and GPS lineages. For gains of resistance from sensitivity across the PMEN3 and PMEN9 lineages, the median γ score for *pbp1a* was 1.0, while for *pbp2b* it was 0.95 and for *pbp2x* it was 0.72. This pattern also applied to the emergence of ST156 (*pbp1a* = 1.0, *pbp2b* = 0.92 & *pbp2x* = 0.62), suggesting the *pbp2x* and *pbp2b* loci were most affected by recombination with non-pneumococcal streptococci. It appears then, that for low-level resistance to emerge (MIC $\geq$0.12 μg/ml), *pbp1a* can remain unaltered in these isolates. When investigating the reversion to penicillin susceptibility within ST156, the γ scores for the *pbp1a* and *pbp2x* genes were 1.0. This is consistent with a reversion to the ancestral susceptible pneumococcal alleles (*pbp2b* was not present within a recombination block for this alteration).

3.3.4.4 The levels and origin of β-lactam resistance in the GPS collection

The origin of resistance-associated *pbp* alleles was also analysed across the species using the GPS collection. Penicillin resistance levels across 621 GPSCs was estimated using the RF method as described for the PMEN3 and PMEN9 lineages. Overall, the RF method generated penicillin resistance phenotype predictions for 19,962 of 20,015 (99%) isolates. The majority of isolates, 64%, were susceptible to penicillin, while 30% were classified as resistant and the remaining 6% intermediately resistant. Ancestral state reconstructions across these strains identified 338 alterations in penicillin resistance levels. The joint most common changes were susceptible-to-resistant, and susceptible-to-intermediately resistant, occurring 117 times (35%) each. In total, 184 of the 338 alterations (54%) were associated with an inferred recombination event affecting at least one of *pbp1a*, *pbp2b* or *pbp2x*. The *pbp2x* gene was most frequently identified as be-

ing altered by recombination, occurring in 100 of the 184 alterations in non-susceptibility associated with a recombination.

As was the case with the PMEN lineages, the emergence of resistance from susceptible genotypes in the GPS collection was often associated with parts of the *pbp2x* and *pbp2b* genes being imported from other species (indicated by $\gamma < 1$; Figure 3.11). The median γ score for *pbp2b* was 0.96, and the gene had a γ score below one in 22 gains of resistance out of 40 associated with a recombination event at the locus. While the median γ score for *pbp2x* was 1.0, there were 15 gains of resistance out of 33 associated with a recombination at the locus where γ was below one. The median γ score of *pbp1a* was also 1.0, with only one instance out of 29 associated with a recombination event at the locus, where its score was below one. This is again consistent with little modification of *pbp1a* in interspecies exchanges. However, where resistant isolates reverted to susceptibility, across all three genes the median γ score was 1.0, indicating within-species recombinations could cause the loss of resistance.

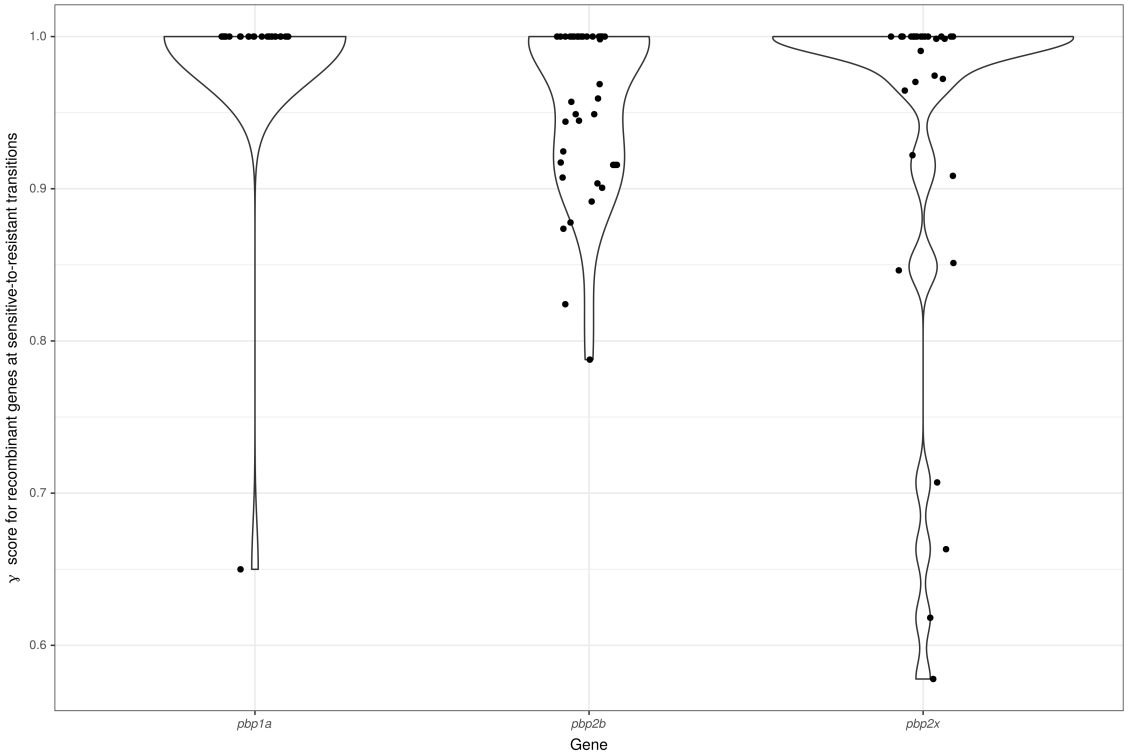

Figure 3.11: Origin of *pbp* genes for penicillin resistant isolates. Scatterplot and violin graph depict the distribution of γ scores for *pbp* gene sequences within the GPS collection where these genes were present within an homologous recombination associated with a gain of resistance to penicillin. Dots represent the individual observations of γ scores, and their overall distribution is summarised as a violin plot.

3.3.5 Evolution of resistance through recombination at other core loci

3.3.5.1 Penicillin resistance and *murM*

In PMEN3, there were further peaks in recombination frequency around the *murM* gene (Figure 3.1), which encodes an enzyme involved in cell wall biosynthesis [600], and has also been implicated in affecting penicillin resistance [140, 601]. Yet compared to the *pbp* genes, the relationship between *murM* modifications and penicillin resistance is much less precisely characterised. Therefore an alignment of the *murM* sequences was analysed with fastGEAR [526] to identify any patterns of sequence import from related species that may be associated with penicillin resistance (Figure 3.12). This revealed evidence of recombination with *S. pseudopneumoniae* and *S. mitis* at *murM* in both lineages. However, only one modification, affecting the region between 946 bp to 1143 bp within *murM*, was associated with high level penicillin resistance. This alteration was observed in both the PMEN9 USA clade and the PMEN3 19A clade, which exhibited the highest penicillin MICs in their respective lineages (Figure 3.10).

3.3.5.2 Co-trimoxazole resistance in PMEN3 and PMEN9

In PMEN9, there was a high density of recombination events affecting the *dhfR* (or *dyr*) gene, the sequence of which determines resistance to trimethoprim, one of the two components (along with sulfamethoxazole) of co-trimoxazole [17]. Overall, PMEN9 was mixed in terms of trimethoprim and sulfamethoxazole resistance. In total 60% and 54% of isolates were predicted to be susceptible to trimethoprim and co-trimoxazole respectively (Figure 3.4). As with penicillin non-susceptibility though, resistance to co-trimoxazole components emerged in parallel across multiple clades. Within the South African clade for instance, 99% of isolates were resistant to sulfamethoxazole, and 77% were resistant to both trimethoprim and sulfamethoxazole. In the USA and Chinese clades, 94% and 100% were resistant to both trimethoprim and sulfamethoxazole, respectively.

Trimethoprim and sulfamethoxazole resistance were much more widespread among the PMEN3 lineage. In total 80% of isolates within PMEN3 were trimethoprim resistant and 81% were sulfamethoxazole resistant (Figure 3.1). This spread was mainly driven by the expansion of the ST156 clade, which inherited alleles conferring these resistance

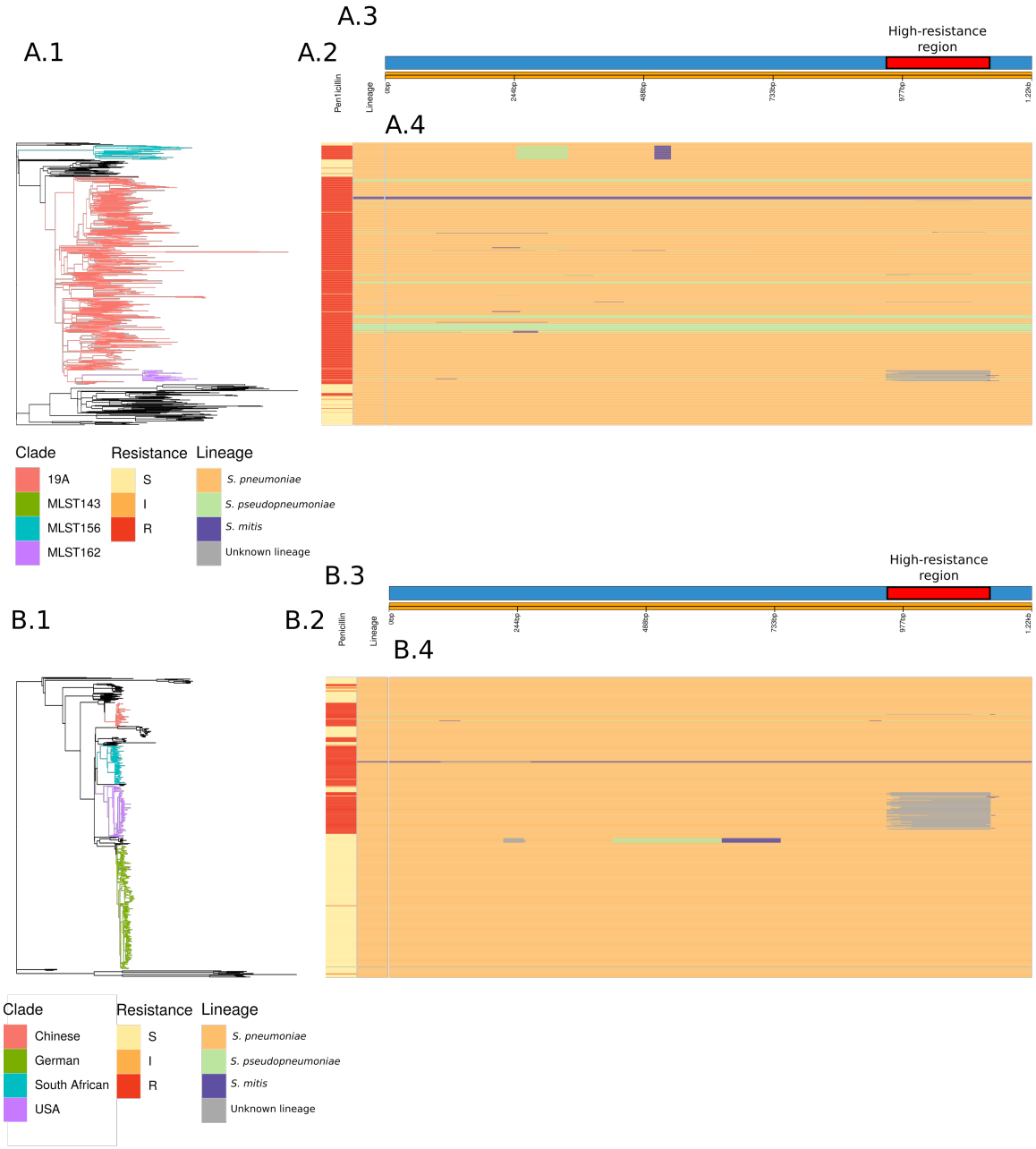

Figure 3.12: Analysis of the origin of the *murM* gene across the PMEN3 and PMEN9 Lineages. **A.1** The whole genome phylogeny of the PMEN3 lineage, as shown in Figure 1, with branches coloured by clade of interest, as described in the main text. **A.2** Bars indicating the penicillin resistance category of isolates, and the lineage inferred for the *murM* gene of each isolate. **A.3** Representation of the *murM* gene. **A.4** This panel shows the inferred lineage for each base of each sequence across the PMEN3 phylogeny. Solid unbroken horizontal bars indicate sequences belonging to a particular lineage, as indicated by the colour. Changes in colour across a bar indicate different *murM* segments were inferred to originate in different lineages, suggesting a mosaic allele generated by recombination. **B.1** Recombination corrected whole genome phylogeny of the PMEN9 lineage, as shown in Figure 2, with branches coloured by clade of interest. **B.2** Bars indicating the penicillin resistance category of isolates and the overall lineage inferred for the *murM* gene of each isolate. **B.3** Representation of the *murM* gene. **B.4** This panel shows the inferred lineage for each base of each sequence across the PMEN9 phylogeny, as for the upper part of the figure.

phenotypes. By contrast, within the ST143 clade, only 12% of isolates were trimethoprim resistant, and 36% were sulfamethoxazole resistant. The 15 isolates from South Africa in the collection were all resistant to both trimethoprim and sulfamethoxazole.The high levels of co-trimoxazole resistance in South Africa across both strains could be driven by widespread co-trimoxazole consumption, as it is commonly used as a prophylactic treatment against secondary infections in HIV positive individuals [611]. In general, co-trimoxazole resistance mirrored the distinctive patterns of penicillin resistance in each lineage.

3.3.5.3 Co-trimoxazole resistance in the GPS collection

Levels of resistance to trimethoprim and sulfamethoxazole were also high across the GPS collection. A majority of isolates were resistant to sulfamethoxazole (11,576 of 20,015; 58%), with fewer isolates resistant to trimethoprim (7,765 of 20,015; 39%). The combination of resistances, conferring full co-trimoxazole resistance, was identified in 7,661 isolates (38%). Among the 4,614 isolates from South Africa in the GPS collection, 2,040 (44%) were resistant to trimethoprim and 2,990 (65%) were resistant to sulfamethoxazole, and there were 1,996 isolates (43%) fully resistant to co-trimoxazole.

3.4 Conclusions

In this chapter I have investigated the role of recombination and transformation in the modification, exchange and dissemination of antimicrobial-resistance genes among globally distributed pneumococcal collections; looking specifically at genes encoded within the core genome of the pneumococcus. Using a mixture of genomic approaches I have sought to first understand the evolutionary histories of the PMEN3 and PMEN9 lineages.

Using Gubbins to produce a phylogeny and incorporating the sampling metadata of these lineages, we can observe the differences in their epidemiology. In the PMEN3 lineage, the phylogeny is dominated by the ST156 clade. This penicillin and co-trimoxazole resistant clade emerged in the 1980s and rapidly spread worldwide. This mirrors the rapid spread of the PMEN1 and PMEN14 lineages [142, 555]. Whereas for PMEN9, the phylogeny is dominated by largely country-specific clades, with penicillin resistance emerging multiple times in different locations. These clades tend not to disseminate as widely as

seen in PMEN3. This is despite the PMEN9 lineage having originated earlier than the PMEN3 clade.

Investigating the data output from Gubbins revealed differences in the lineages' overall rates of recombination too. The PMEN3 lineage tended to be more recombinogenic than PMEN9, not only having a greater *r*/*m*, but also tending to import more diverse sequence, with the SNP density of recombination events significantly higher for PMEN3. This difference helps explain the different levels of serotype switching seen in the data too, with PMEN3 switching capsule locus far more often, 35 times, compared to PMEN9 and its nine switches of its capsule locus. Indeed, the PMEN1 lineage, which had an estimated *r*/*m* of 7.2 similar to PMEN9's 7.7, also only had 8 serotype switch events detected in its collection, although this was with a smaller sample size of 240 isolates [142]. In general both these lineages are within the middle of the range of *r*/*m* values observed for PMEN lineages, with the highest PMEN14 observed at 34.06 *r*/*m* [612] and the lowest PMEN31 observed at 0.07 [613].

Both lineages however shared peaks in recombination around their *pbp* loci. These corresponded with a number of alterations to the resistance level of isolates, with 20 of 35 switches in both lineages (11 in PMEN3 and 9 in PMEN9), having a recombination event affecting at least one of *pbp1a*, *pbp2b* or *pbp2x*. Peaks at these loci are common in pneumococcal MDR lineages, likely reflecting the strong selection pressure placed on bacteria by the use of β-lactam antibiotics [142, 555, 614]. Indeed, we observe the expansion of the ST156 lineage within PMEN3 immediately after gaining resistance to penicillin. More systematic sampling of this clade could've allowed greater resolution on the date of this expansion, as well as information on its likely geographic setting.

This work also consolidates findings from other multidrug resistant lineages, in that I observe PMEN3, PMEN9 and other GPSC lineages acquiring penicillin resistance via the importation of *pbp* gene sequences from other species [137, 138, 560, 561]. This was most frequently observed at the *pbp2b* and *pbp2x* genes. Previous work has suggested alterations at these two loci are usually the first steps required for low-level penicillin resistance to emerge [140]. On the other hand, substantial alterations in *pbp1a* are associated with higher levels of resistance, above the 0.12 μg/ml used as the threshold for defining resis-

tance with the RF model [615]. Indeed, despite the emergence of penicillin resistance at the base of the ST156 clade, there are numerous further recombinations around the *pbp* loci, with sublineages containing high MIC values (Figure 3.10). Hence the lack of strong evidence for *pbp1a* being modified by sequence from other species may be an artefact of how I identified transitions between discrete resistance levels. This may also apply to the alterations to *murM* [616]. Particular imports around *murM* from other streptococci were found in PMEN3 and PMEN9 clades exhibiting high penicillin resistance. These clades also had modified *pbp* genes, suggesting epistatic interactions are likely to be important in fully understanding the role of the *murM* imported segments [617].

This highlights a problem for *in silico* predictions in ensuring training data for models is representative of the diversity within a population. While I was able to broadly recreate the results of the Li *et al* 2016 study [269] for the dataset of resistance levels provided by the CDC, with the RF model they employed also outperforming a simple LR model I tested, the CA scores dropped markedly for the PMEN dataset. This neccessitated the switch to the narrowed meningitis breakpoints used for prediction. Recent work has used logistic ElasticNet models trained on unitigs as input features, to predict resistance for Penicillin, as well as a range of other antibiotics [618]. This could be a more generalisable approach to predicting resistance, although so far this has only been validated on small pneumococcal datasets. In the future, using a method, that can accurately predict individual MICs for novel genotypes would enable a more detailed investigation of the recombination events that can lead to low, intermediate and high levels of resistance emerging in populations.

In the next chapter I will perform a similar analysis on the effects of recombination on the spread of MGEs within pneumococcal populations. This will also look into the likely origin of these AMR encoding elements, and investigate their method of spread in the populations. Together these analyses will give an overall picture of how both the core and accessory genome affect AMR spread and how it is molded by recombination.

Chapter 4

The spread of resistance through mobile genetic elements in pneumococcal populations

Acknowledgements

This work has been previously published in D'Aeth *et al* 2021 [556].

Summary

This chapter investigates how recombination can drive the dissemination of MGEs in pneumococcal populations. I characterise the spread of two families of mobile genetic elements, Tn1207.1 and Tn916, in the populations described in the previous chapter: PMEN3, PMEN9 and the GPS collection. In PMEN9 I show how conflict between the host and Tn1207.1 leads to the rise and subsequent decline of closely related set of isolates in Germany. Additionally I find that these mobile elements are also frequently inserted by homologous recombination. Many of these elements originate in closely related streptococci also common in human carriage. These results further highlight how selection pressure from antibiotic consumption can drive the emergence of lineages with large recombination events.

4.1 Introduction

4.1.1 MGEs, recombination and resistance

In the previous chapter I investigated how modifications of pneumococcal core genes, such as the *pbp* genes, can lead to resistance spreading through populations. The other mechanism through which pneumococcal populations, and bacteria more widely, can gain resistance is via the movement of specialized resistance genes on mobile genetic elements (MGEs). Plasmids are a common vector of AMR genes across species, with many successful pathogenic *Enterobacteriaceae* lineages, such as the ST131 *Escherichia coli* lineage, carrying plasmids encoding extended-spectrum β-lactamases (ESBL) for instance [619, 620]. In the pneumococcus however, the plasmid repertoire is limited to only two types of cryptic elements [621–623]. As such, the MGEs that contribute most of the spread of AMR genes in the pneumococcus are integrative and conjugative elements (ICEs) [435, 487].

These elements integrate within the host chromosome and encode their own conjugation machinery to move between cells [421]. This conjugation machinery enables the elements to move across a broad range of bacterial taxa [561, 624]. Indeed, ICEs such as the common tetracycline resistant element Tn*916*, are present in AMR lineages of pneumococci and the majority of antibiotic-resistant bacterial pathogens considered a priority by the WHO [429, 625, 626]. However, *in vitro* studies looking at how Tn*916* moves between cells appear to show the element is unable to conjugate between pneumococci [627, 628]. Another method these MGEs could be moving between cells is via transformation, although this only tends to import closely related sequences and favours much shorter sequence imports [350, 485].

In order to understand how MGEs may move between pneumococci isolates, and also the likely sources of these imports, in this chapter I investigate the spread of two common MGEs, Tn*916* and the macrolide resistant Tn*1207.1* element. I first determine their distribution in the two PMEN lineages described in the previous chapter and the wider GPS collection. I then investigate the loci where these elements insert into the host chromosome and the effect this has on the host, and finally seek to determine where

these MGEs entered into the pneumococcal population from. Now I will describe the two elements I investigate in more detail.

4.1.2 The Tn*916* element

Tn*916* was the first identified ICE, found in *Enterococcus faecalis*, and is thought to be the shortest at just 18kb in length [421]. Typically Tn*916* carries the *tet(M)* gene conferring tetracycline resistance, with close relatives such as Tn*1545* and Tn*2010* containing the same core gene groups, but adding in the *erm(B)* and *mef(E)* macrolide resistance genes [429, 626]. Tn*916* can also form composite elements that can confer resistance to aminoglycosides, streptogramins and lincosamides through the integration of sequences such as the Omega cassette and Tn*917* elements [142]. The Tn*916* genome is split into two sections, the regulatory region, where cargo genes such as *tet(M)* reside, and the conjugative region (Figure 4.1). When integrated into the host chromosome, typically there are only low levels of transcription of the genes in the regulatory region of the element, with no transcription of the conjugative region (Figure 4.1). Upon excision and circularization though, genes in the conjugative region are expressed [421].

The Tn*916* site-specific integrase (Int) is a member of the transposase subfamily of tyrosine recombinases, and is quite broad in its insertion site preference, favouring sites that are AT rich or bent [425, 429]. As such the element has been detected in over 35 bacterial genera, encompassing the majority of priority pathogens listed by the WHO [429, 625, 626]. This diverse host range means Tn*916* has ample opportunity to acquire new genetic material, with the wider Tn*916* family of transposons displaying a varied array of cargo genes, from mercury resistance to restriction modification systems [626]. Among pneumococci, Tn*916* and Tn*916* related ICE are prevalent among MDR lineages such as CC180, PMEN1 and PMEN2 [142, 312, 629].

4.1.3 The Tn*1207.1* element

Tn*1207.1* is a short 7.2 kb defective transposon, it is closely related to the 5.5kb Mega element that is often found in larger ICEs such as Tn*2010* [630] (Figure 4.2). The Mega element confers macrolide resistance via the presence of the *mef(E)* macrolide efflux pump, while the Tn*1207.1* element contains the closely related *mef(A)* gene [577]. Both

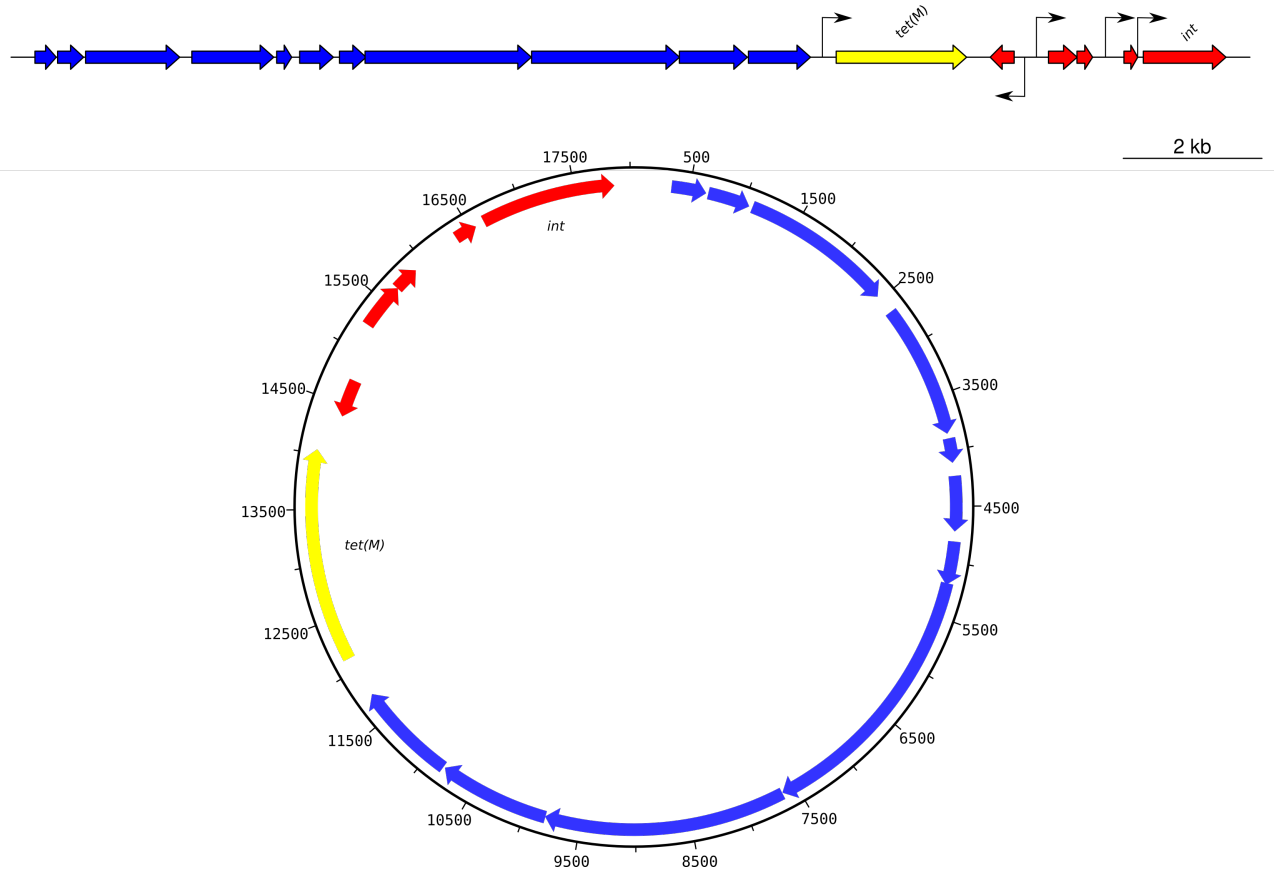

Figure 4.1: Linear and circular representation of the Tn*916* element. Genes are shown as arrows with the direction of the arrow indicative of the direction of transcription they initiate. Some known promoters are shown as bent arrows. Genes in the regulatory region of the element are highlighted in red, while those in the conjugation region of the element are shown in blue, and the resistance gene *tet(M)* is shown in yellow. The circularization of the element upon excision from a host chromosome allows for all genes to be transcribed. Adapted from Johnson & Grossman 2015 [421]

mef(A) and *mef(E)* are classified in the same efflux protein macrolide resistance group and share 90% nucleotide sequence similarity [577, 631]. The *mef(A)* gene was first discovered in *Streptococcus pyogenes* as part of the larger 52 kb Tn*1207.3* conjugative element [632, 633]. The Tn*1207.1* element in turn was found first in the pneumococcus and corresponds to the left-end of the larger Tn*1207.3* element [634, 635]. Within Tn*1207.1* the 5' end of open reading frame (ORF) 8 also appears to be a truncated form of the *umuC* gene encoded by Tn*5252* for UV-resistance [634, 636].

Within pneumococci, Tn*1207.1* has consistently been found inserted at the same loci in the host genome, splitting the competence gene *comEC* [635, 637]. The Tn*1207.3*

element is also found splitting the orthologue of this gene within *S. pyogenes* populations [633]. The majority of the pneumococcal isolates found containing Tn*1207.1* are members of the PMEN9 lineage, suggesting a rare insertion [635, 638]. The Mega element on the other hand, appears to be much more dispersed among diverse bacterial taxa owing to its location on the common ICEs Tn*2010* and Tn*2009*, which are themselves part of the wider Tn*916* family [630,639,640]. A Mega element present within the Tn*2009* ICE is seen at the base of PMEN14 lineage and present in the majority of these MDR isolates [555]. Within the PMEN1 lineage too, Tn*916* elements are modified to Tn*2009/10 in situ* through the acquisition of the Mega element [142]. Outside of these ICE elements though, mega has been seen directly integrated into the pneumococcus' chromosome in at least six different locations [641, 642].

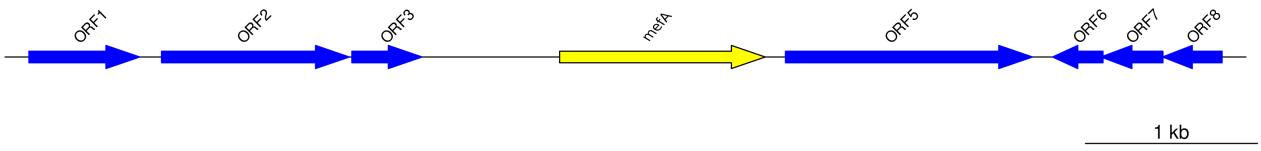

Figure 4.2: Linear representation of the Tn*1207.1* element. Genes are shown as arrows with the direction of transcription initiation indicated by the direction of the arrow. Open reading frames (ORFs) are highlighted and the resistance gene *mef(A)* is annotated and highlighted in yellow.

4.2 Methods

4.2.1 MGE identification

To measure the presence of the MGEs Tn*1207.1* and Tn*916* in the PMEN3, PMEN9 and GPS collection, reference MGE sequences were used to search for intact and partial representatives. For the Tn*916* element, the 18 kb reference given by the transposon registry [643], extracted from *Bacillus subtilis* (accession code: KM516885), was used. For Tn*1207.1*, a 7 kb reference extracted from the *S. pneumoniae* INV200 genome (accession code: FQ312029.1) was used. BLASTN v2.5.0 was used to detect Tn*916* and Tn*1207.1* among the assembled genomes in the collections. Hits were filtered using an empirically determined cutoff alignment length of 7 kb and 2 kb for Tn*916* and Tn*1207.1* respectively. Given the fragmented draft nature of the assemblies in the collections, hits could often span multiple non-adjacent contigs. As such BLASTN results were merged if they represented continuation of an element's sequence split across multiple contigs.

4.2.2 Antibiotic consumption data

Europe-wide macrolide consumption datasets were compiled from the ECDC. Data for Germany, France, Italy, Netherlands, Spain and the UK were collected. For the Italian ECDC data, collection started from 1999, while for the other countries the collection ran from 1997 for macrolide consumption. The rates of macrolide consumption in Germany pre-1997 were sourced from a study looking at macrolide resistance among pneumococci isolates in Germany by Reinert *et al* 2002 [644]. This study recorded data from 1992 to 2000, while the ECDC recorded data from 1997 to the present day. The ECDC data for Germany is from the primary care sector for outpatients, with a population coverage of 90%, while the Reinert *et al* paper takes data from both prescriptions in hospitals and from community general practitioners. These two macrolide usage datasets were combined using the three years of overlap between the datasets as a scaling factor. This was the average transformation that mapped the Reinert *et al* 2002 data to the ECDC data. It was applied to convert the data from 1992 to 1996 into the same units as the ECDC data (defined daily doses / 1000 population). Only the ECDC datasets were used for Europe wide comparisons.

For β-lactam consumption, data was also taken from the ECDC for 1997 to the present day. For the German isolates again, two sources were used with the ECDC data from 1997 to present day combined with β-lactam consumption data from 1992 to 1997 taken from McManus *et al* 1997 [645]. This study has data on the hospital and retail sales of oral antibiotics in West Germany for the years 1989 and 1994 in the same DDD units as the ECDC data. A linear trend between 1989, 1994 and 1997, the first year of data from the ECDC, was used to impute the missing values between 1992 and 1997.

4.2.3 Phylodynamic analysis

The expansion of the German lineage within PMEN9 was further examined using phylodynamic approaches. Initially, the phylogeny of PMEN9 produced by Gubbins (Section 3.2.3) was subset to the 162 German isolates both containing Tn*1207.1* and with a date of isolation. The *roottotip* function in the BactDating R package v1.0.1 [552] was used to test for a molecular clock.

A time-calibrated phylogeny was then created, also using BactDating. This was run with a relaxed clock model and an MCMC length of 100 million iterations. Chain convergence was checked through visual inspection of the trace plots output for the model.

The Skygrowth R package v0.3.1 [313] was then used to formally test the link between antibiotic consumption and population growth rates. The timed phylogeny generated by BactDating was input into Skygrowth, where it was analyzed in combination with the β-lactam, macrolide, and macrolide-to-β-lactam consumption data. Consumption data were each separately rescaled prior to analysis with default settings and priors. Four sets of analysis were run: one for each of the three consumption datasets, and one without consumption data. In each case, the MCMC was run for 100 million iterations, which visual inspection suggested was sufficient for the chains to have converged.

4.2.4 MGE insertion site identification

A pipeline was developed to categorize the insertion points of the two MGEs studied. Figure 4.3 outlines the algorithm. The initial step was the creation of a library of unique hits, with each isolate containing an MGE BLAST searched against the corresponding GPSC's reference, or a global reference if the strain reference also contained the MGE.

This determined the start and end points of an insertion. A hit was defined by three characteristics: (1) the total length of the delineated insertion; (2) the number of genes within the insertion, and (3) the genes within the flanking regions of a hit. Each observed combination of values was considered a unique hit. For instance, if two hits were of similar length and gene content, but differed in where they inserted within the host, they were treated as two unique hits. The unique insertions with the longest flanking matches to the reference, indicating the insert was reconstructed on a large contig, were used as representatives of that insertion within the library.

Hits might not be present in the library either through having shorter flanking sequences, or having MGE insertions spread across contigs. The next step was to allocate such BLAST hits not present in the library to one of the unique library insertion types (the combination of gene number, insertion length and location). Isolates with no matches to the reference either side of the hit, usually when the hit was present in a small contig or within a larger unresolved element, were discarded from the analysis.

Once hits had been allocated an insertion type, the node at which the insertion occurred was reconstructed on the Gubbins phylogeny for each GPSC. This ancestral state reconstruction was performed using PastML v1.9.15 [590]. The recombination predictions were then searched to detect whether there was a putative recombination event, on the branch on which acquisition was estimated to occur, spanning the insertion site within the reference for a GPSC. If such a recombination were identified, this was considered indicative of element insertion via homologous recombination. The flanking regions of the isolate with the fewest reconstructed SNPs around the insertion site of the element since its insertion, as inferred from the Gubbins base reconstruction, were then extracted to test for the origin of this element.

As with the origin of the *pbp* gene analyses presented in the previous chapter, these flanking regions were then compared to a reference collection of 52 streptococcal genomes collated from antimicrobial susceptible *S. pneumoniae* and other *Streptococcus* species, building on the database collated in Mostowy *et al* 2017 [526]. Again, BLASTN v2.5.0 was used to compare each flanking region to this database. The orthologous regions to these flanks were also extracted from isolates not containing the insertion, to act as a control.

The statistic γ (Equation 3.1; Defined in Section 3.2.5.1) was also used to determine the species of origin for an insertion. The code for this pipeline is available at `https://github.com/jdaeth274/ISA`.

4.3 Results

4.3.1 MGE distribution among PMEN3 and PMEN9

Searching for both Tn*916* and Tn*1207.1* revealed their widespread distribution among the MDR lineages PMEN3 and PMEN9. Tn*916*-type elements were present in 70 representatives of PMEN3 (Figure 4.4). An ancestral state reconstruction using PastML v1.9.15 [590] identified 17 independent insertions. Only two spread to a notable extent: one was a clade of 22 isolates within the ST156 clade of which 18 isolates contained Tn*916*, and the other was within the 33 isolate ST143 clade, of which 25 isolates contained Tn*916*. In both clades it appears there were multiple instances of Tn*916*-type elements being lost, with the isolates without Tn*916* sporadically appearing in both clades (Figure 4.4).

In PMEN9, Tn*916*-type elements were present in 150 isolates (Figure 4.5). They were most commonly present in the South African clade, where Tn*916*-type elements were found in 71 of the 73 isolates in this clade, with likely deletion in two isolates. Similarly, 40 of the 45 isolates within the Chinese clade had also acquired Tn*916*-type elements, with five isolates without an element appearing to have lost these independently, as with ST143 in PMEN3.

The Tn*1207.1*-type elements were more common in both strains. In PMEN3, 108 isolates contained a Tn*1207.1*-type element, resulting from 27 independent insertions (Figure 4.4). The two insertions associated with the largest clonal expansions were one within the 19A subclade (26 isolates), and a second in another subclade of ST156 (25 isolates). The other 25 insertions were less successful, appearing sporadically around the phylogeny.

In PMEN9, Tn*1207.1*-type elements were present in 341 isolates (Figure 4.5). The elements were present in 92 isolates of a subclade of the USA clade, and ubiquitous in the 238 isolates of the German clade, the most widespread insertion observed in the collection.

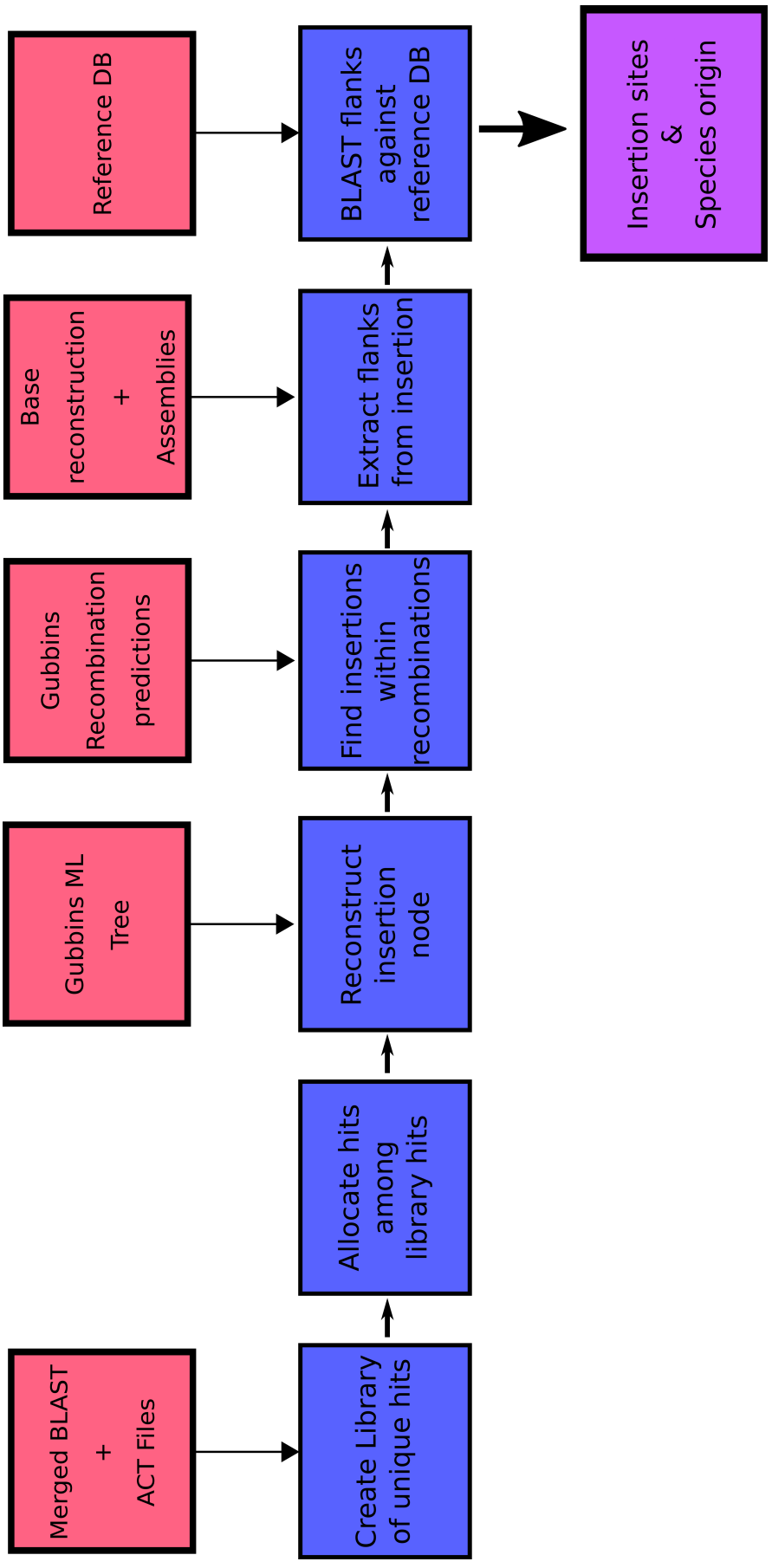

Figure 4.3: Outline of insertion point pipeline. Red boxes represent data input into the pipeline, blue boxes the individual analysis steps within the pipeline and purple the pipeline's output.

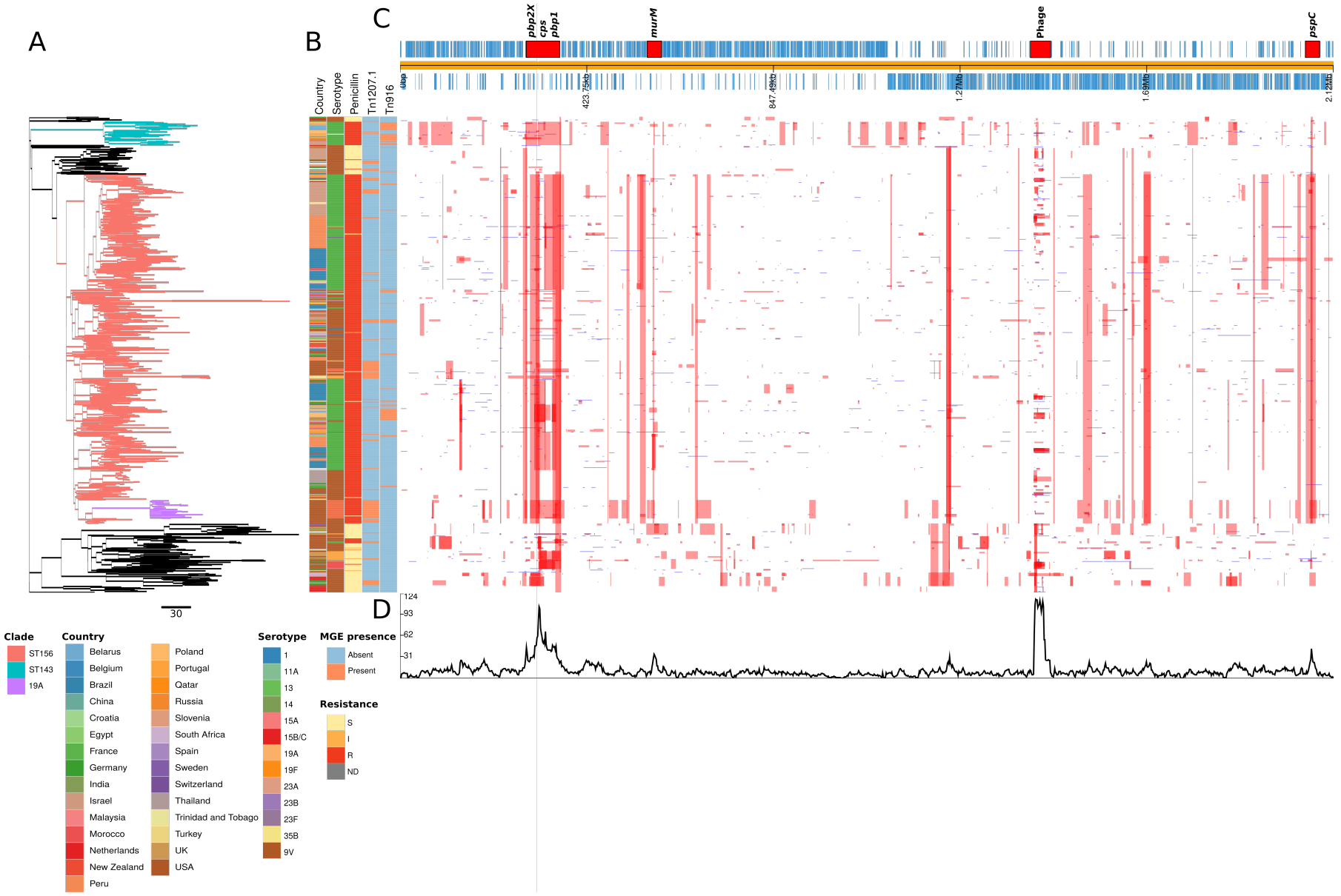

Figure 4.4: Phylogenomic analysis of PMEN3 lineage. **(A)** Maximum likelihood phylogeny generated from the non-recombinant regions of the 669 isolates from the PMEN3 lineage. Branches are coloured by clade identified in the key. Units for the scale bar are the number of point mutations along a branch. **(B)** Bars highlighting the country of origin, serotype, resistance categories to penicillin and the presence or absence of the Tn*1207.1*-type and Tn*916*-type elements. Bars map across to isolates on the phylogeny. **(C)** Simplified genome annotation of the PMEN3 reference isolate RMV4. The highlighted regions correspond to peaks of recombination event frequency. Blue bars represent individual genes annotated within the assembly. **(D)** Distribution of recombination events across the PMEN3 lineage. In the upper half of the graph, red bars indicate recombination events occurring on internal nodes in the tree, which were subsequently inherited by multiple descendent isolates. These bars map across to isolates in the phylogeny in section **A** and map to regions in the genome annotated in section **C**. Blue bars indicate recombination events on terminal branches of the tree, occurring in only one isolate. In the bottom half of the graph, the line represents the frequency of recombination events along the genome's length.

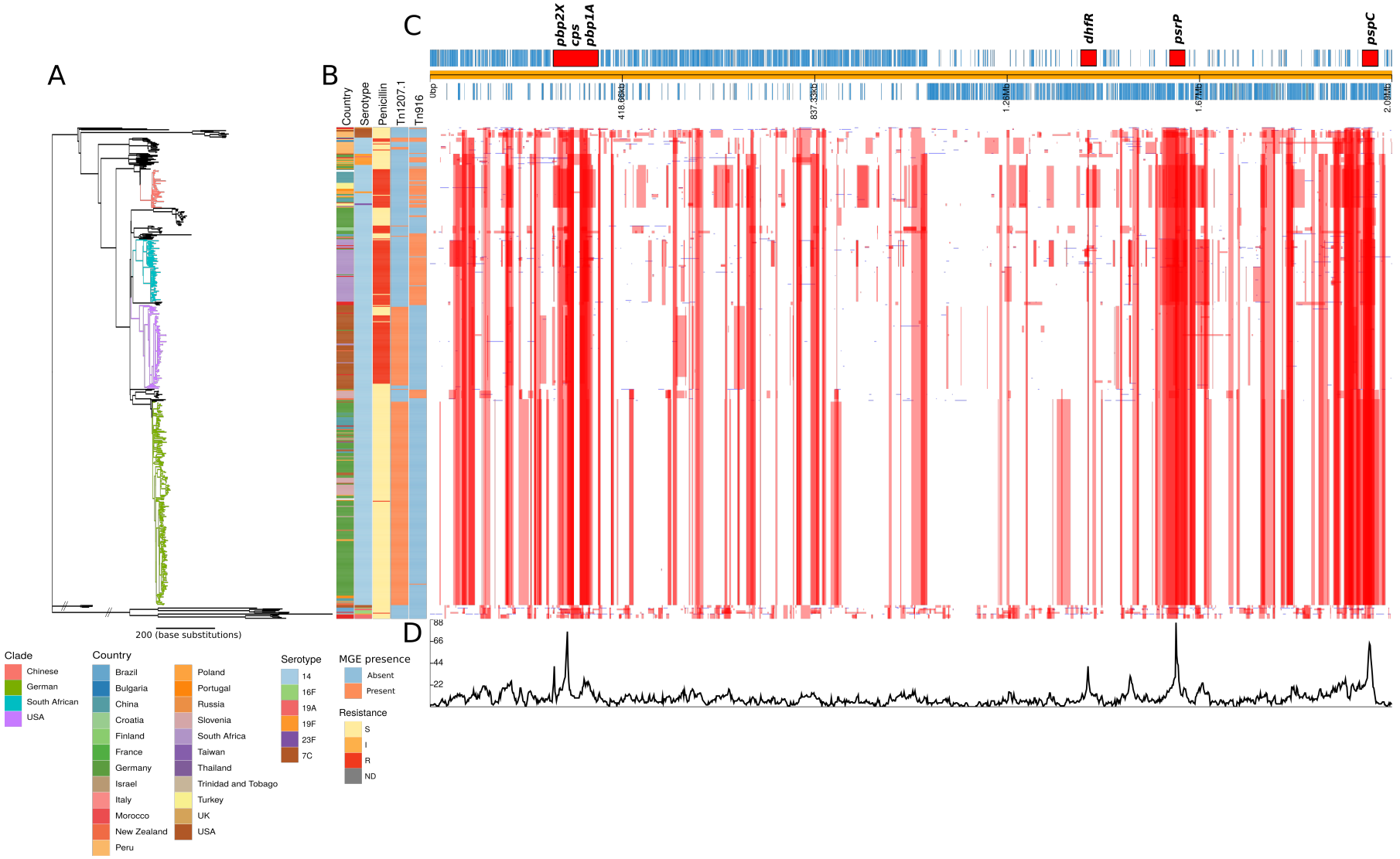

Figure 4.5: Phylogenomic analysis of PMEN9 lineage. **(A)** Maximum likelihood phylogeny generated from the non-recombinant of the 575 isolates in the PMEN9 lineage. Branches are coloured by clade identified in the key. Units for the scale bar are the number of point mutations along a branch. **(B)** Bars highlighting the country of origin, serotype, resistance categories to penicillin and the presence absence of the Tn*1207.1*-type and Tn*916*-type elements. Bars map across to isolates on the phylogeny. **(C)** Simplified annotated genome of the PMEN9 reference isolate INV200. he highlighted regions correspond to peaks of recombination event frequency. Blue bars represent individual genes annotated within the assembly. **(D)** Distribution of recombination events across the PMEN9 lineage. In the upper half of the graph, red bars indicate recombination events occurring on internal nodes in the tree, which are subsequently inherited by multiple descendent isolates. These bars map across to isolates in the phylogeny in section *A* and map to regions in the genome annotated in section **C**. Blue bars indicate recombination events on terminal nodes of the tree, occurring in only one isolate. In the bottom half of the graph, the line represents the frequency of recombination events along the genome's length.

Both Tn*916*-type and Tn*1207.1*-type elements were frequently acquired by PMEN3 and PMEN9. However, while in PMEN9 some of these acquisitions appear at the base of highly successful clades, in general few of these insertions resulted in internationally disseminated MDR genotypes.

4.3.2 Expansion of macrolide resistance in German pneumococci

The expansion of the German clade carrying Tn*1207.1* represented an unusual case of an MGE insertion being associated with a successful genotype. This suggested strong selection for a macrolide resistant genotype in Germany in recent years. In order to assess the degree of selection pressure on this clade, antibiotic consumption data was collected from across Europe for both β-lactams, which this clade is sensitive to (Figure 4.5), and macrolides, to which this clade is resistant to. From these datasets, we can see that German antibiotic consumption is generally low relative to the rest of Europe [646] (Figure 4.6). However, Germany has a high ratio of macrolide-to-β-lactam usage relative to other European countries (Figure 4.7).

Given these atypical antibiotic consumption trends, I decided to further investigate whether consumption patterns may have aided the dissemination of this lineage in Germany. I performed a phylodynamic analysis of the 162 German isolates in this clade with available date of isolation data (excluding two isolates without date of isolation data). The first step in this was checking for evidence of a molecular clock. The 162 isolates had been collected between 1992 to 2008. There was significant evidence of a molecular clock, based on the correlation between the root-to-tip distance and the date of isolation for this clade (n = 162, Pearson's correlation coefficient R^2 = 0.15, p value < 1 x 10^{-4}) (Figure 4.8). This root-to-tip analysis estimated the clade's most recent common ancestor (MRCA) existed in 1970.

With sufficient evidence for a molecular clock within the data, the next step in the analysis was the generation of a time calibrated phylogeny. The time-calibrated phylogeny output from BactDating [552] suggested a relatively slow clock rate of 5.5 x 10^{-7} substitutions per site per year (95% credibility interval of 4.5 x 10^{-7} to 6.6 x 10^{-7} substitutions per site per year).

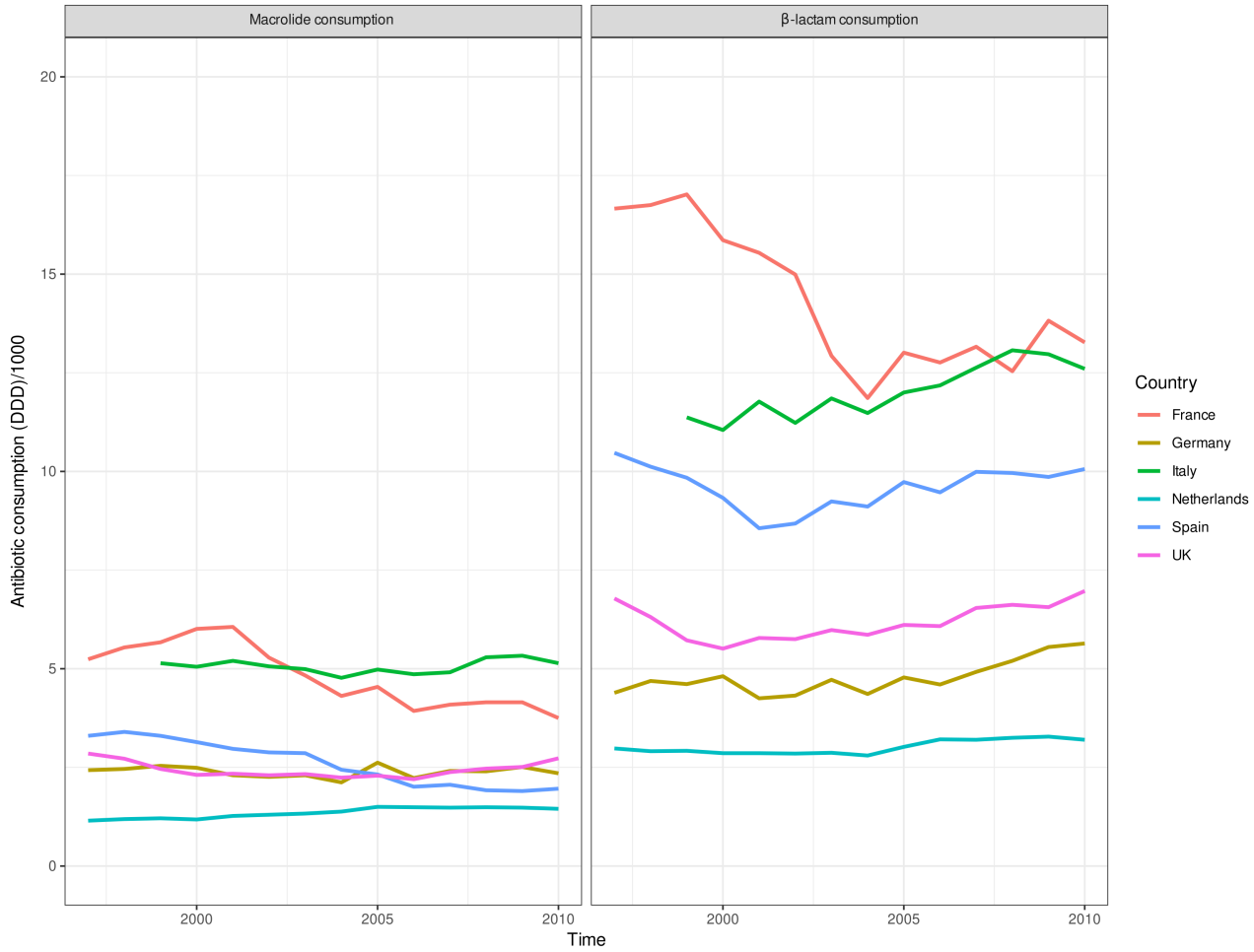

Figure 4.6: Consumption of macrolide and β-lactam antibiotics across Europe. The DDD / 1000 consumption rates for macrolides and β-lactam antibiotics for six major European countries across a 13 year period from 1997 to 2010

The Skygrowth R package v0.3.1 [313] was then used to reconstruct the effective population size, N_e, and the growth rate of N_e through time of this clade. The antibiotic usage data over this period was used as a covariate, to test for evidence of selection by changing consumption (Figure 4.9). All these isolates were serotype 14, which is included in the PCV7 vaccine that was introduced into the universal vaccination programme for children under two years of age in Germany in 2006 [647]. As such, only isolates collected pre-2006 were included in this analysis, to minimise any effect of the vaccine on N_e. This left 103 isolates from Germany, sampled between 1992 and 2005 for further analysis.

From the reconstruction without using the macrolide and penicillin consumption data, it is evident this lineage expanded rapidly during the late 1990s and early 2000s, with its peak in growth rate around 1997 preceding a peak in N_e around 2003. Both N_e and

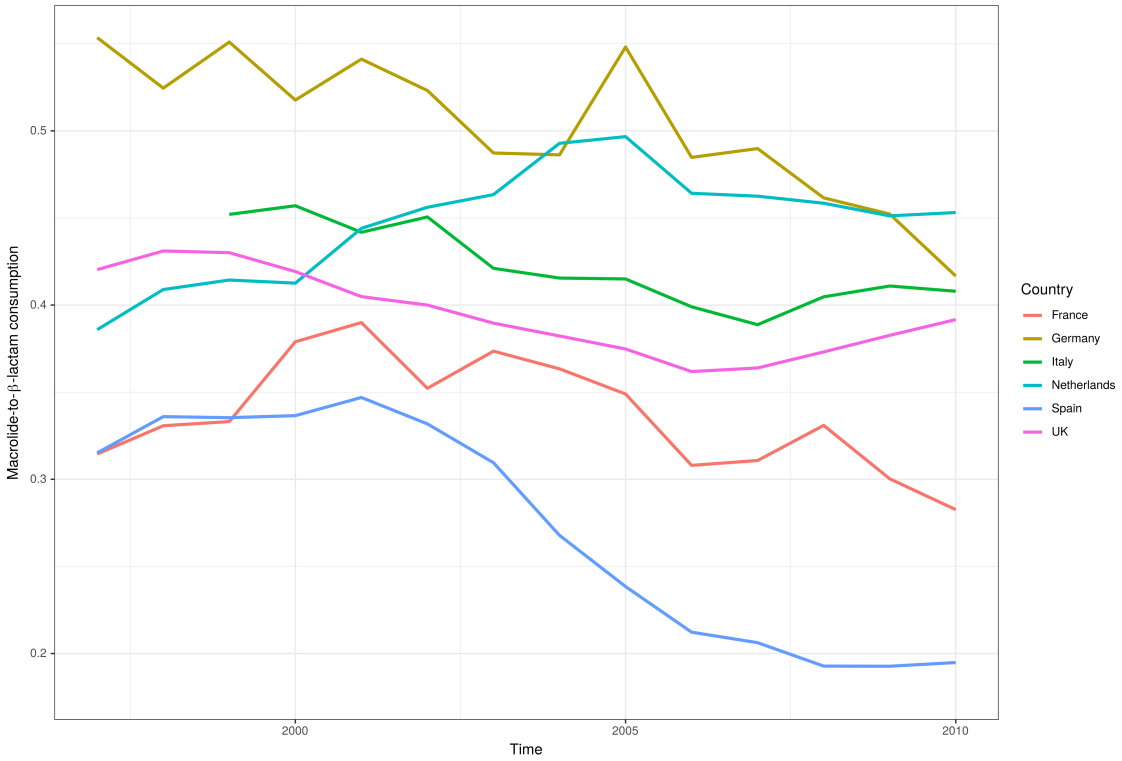

Figure 4.7: Ratio of macrolide to β-lactam consumption in Europe. The ratios of macrolide use, in DDD, per β-lactam use for six major European countries across a 13 year period from 1997 to 2010.

growth rate subsequently declined. The peak of the macrolide-to-β-lactam consumption ratio was in the mid-to-late 1990s, whereas the peak N_e was not reached until after 2000. This peak N_e growth rate, rather than maximum N_e itself, coincided with the timing of the highest consumption ratio. A similar observation was made for methicillin consumption and the spread of MRSA, with increased consumption leading to the expansion of the USA300 clade [313].

Correspondingly, when incorporating the macrolide-to-β-lactam consumption ratio into the reconstruction, the credible intervals for the growth rate estimation narrowed. Additionally, the macrolide-to-penicillin consumption ratio had a significant positive mean posterior effect of +0.24 [95% credible interval 0.06 to 0.48] on the growth rate of the clade. Both results supported the hypothesis that growth rate was correlated with the contemporary patterns of antibiotic consumption in Germany, consistent with selection pressures from national level prescribing practices driving the expansion of this clade in the late 1990s.

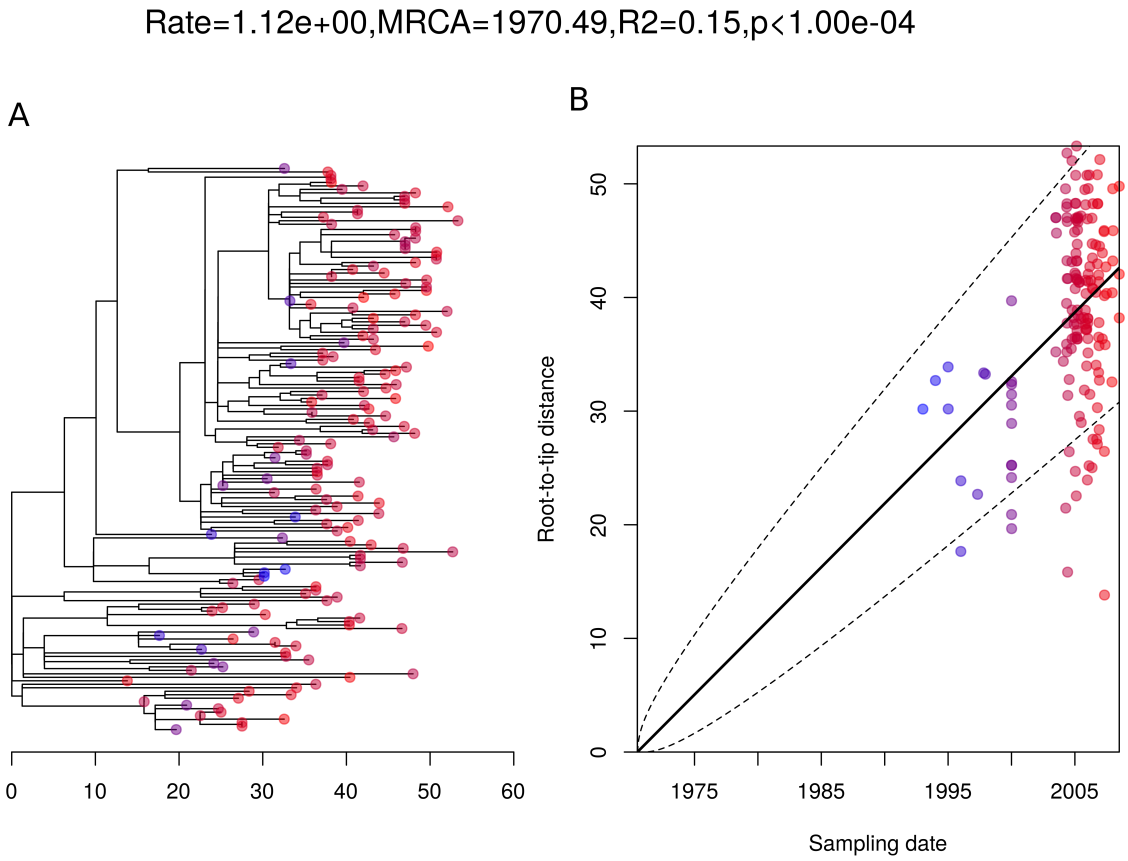

Figure 4.8: Root to tip analysis of 162 German isolates within PMEN9. A Represents the 162 isolate phylogeny with node tips coloured by date of isolation. **B** Linear regression of root to tip distance against sampling date for Isolates.

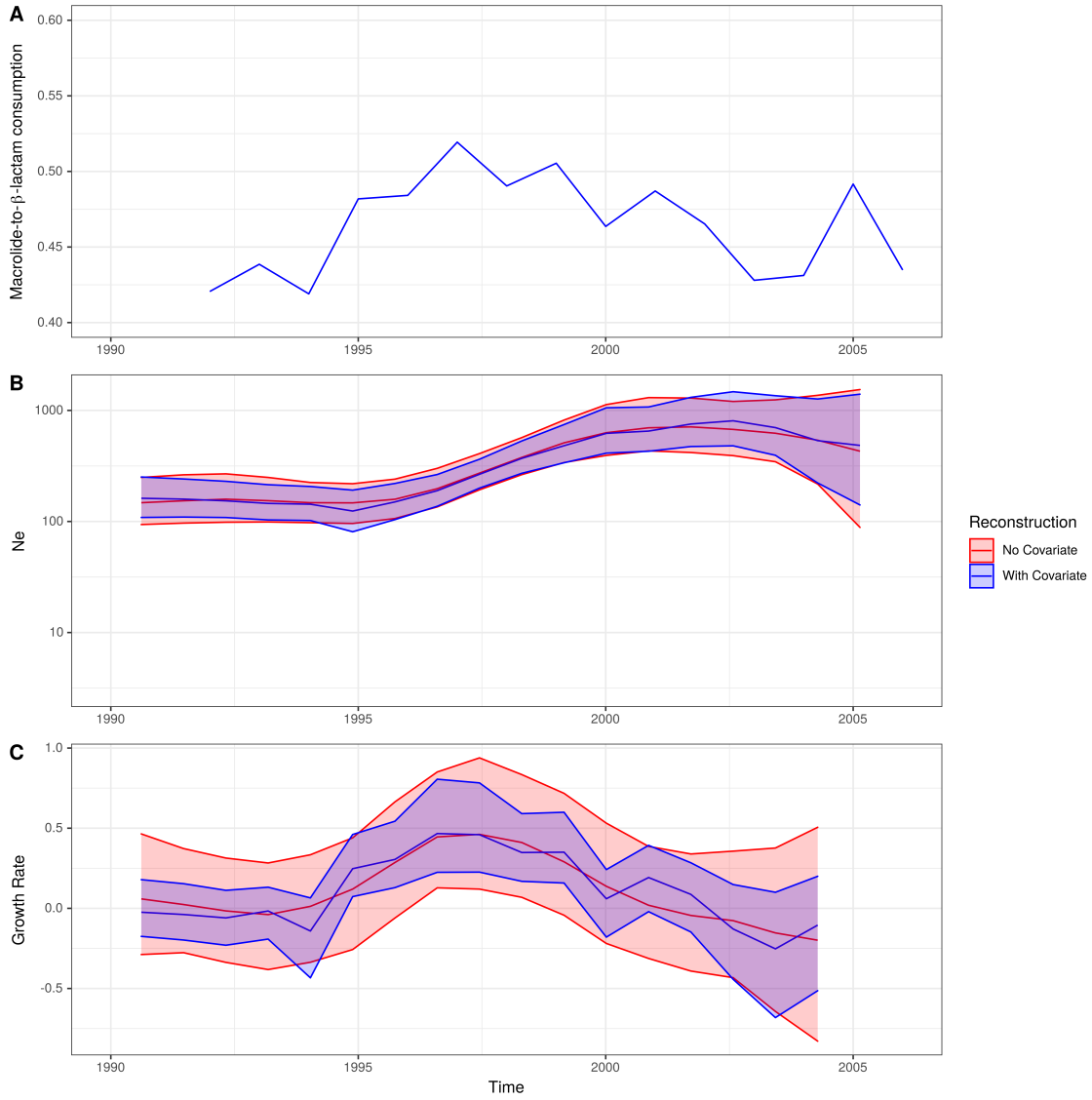

Figure 4.9: Expansion of a macrolide resistant clade in Germany pre-vaccine. (A) The ratio of macrolide-to-penicillin consumption in Germany. **(B)** The change in N_e through time inferred by Skygrowth, with the red line representative of when no covariates are incorporated and the blue line when the macrolide-to-β-lactam ratio is incorporated into the reconstruction. Shaded regions represent the 95% credible intervals. **(C)** The reconstruction of the growth rate of N_e through time. The red line represents the result of model fitting without covariates, and the blue line when macrolide-to-β-lactam ratio data were incorporated. Shaded regions represent the 95% credible interval for the reconstruction.

To investigate whether this relationship holds for the individual antibiotic consumption rates, as opposed to the ratio of rates, further analyses incorporating just the macrolide consumption and the β-lactam consumption were also run (Figures 4.10 & 4.11). For the β-lactam consumption data, there was no significant effect on the growth rate of the clade (mean = +0.19, 95% credible interval -0.09 to 0.57). However for the macrolide consumption data, there was a significant effect (mean = +0.22, 95% credible interval 0.02 to 0.55), albeit to a lesser extent than the macrolide-to-β-lactam ratio.

The absence of penicillin resistance, or vaccine evasion through serotype switching [648], may be a consequence of the Tn*1207.1* element itself. This MGE inserted into, and split, the gene *comEC* (Figure 4.12), which encodes a membrane channel protein integral to extracellular DNA uptake during competence [649]. Therefore these cells were unable to import DNA for transformation, necessary for serotype switching and the acquisition of penicillin resistance alleles of the *pbp* genes [561]. The impact of this insertion is evident from the absence of ongoing transformation within the German clade (Figure 4.5).

Further analysis of the origin of this Tn*1207.1* insertion, via analysing the flanking regions as for the *pbp* genes in the previous chapter, revealed a probable interspecies origin. The flanking regions immediately adjacent to the insertion have a low percent identity matched to other pneumococci, ranging between 92% and 94% (Figure 4.12). The immediate upstream 500 bp region most closely matched to a *S. mitis* reference genome (accession code AFQV00000000).

Given the likely interspecies origin of this Tn*1207.1* element, I further investigated the origins of this element and Tn*916*-type elements across the GPS collection. I looked firstly at the diversity of these elements, in terms of insertion loci and length, in the wider GPS collection and the likely number of independent insertions occurring here.

4.3.3 Multiple acquisitions of the Tn*1207.1* and Tn*916* elements across the GPS collection

4.3.3.1 Defining unique hits

In order to determine the number of element insertions, I first had to identify all the unique insertions of the elements within the collections. Initial BLAST search results often give

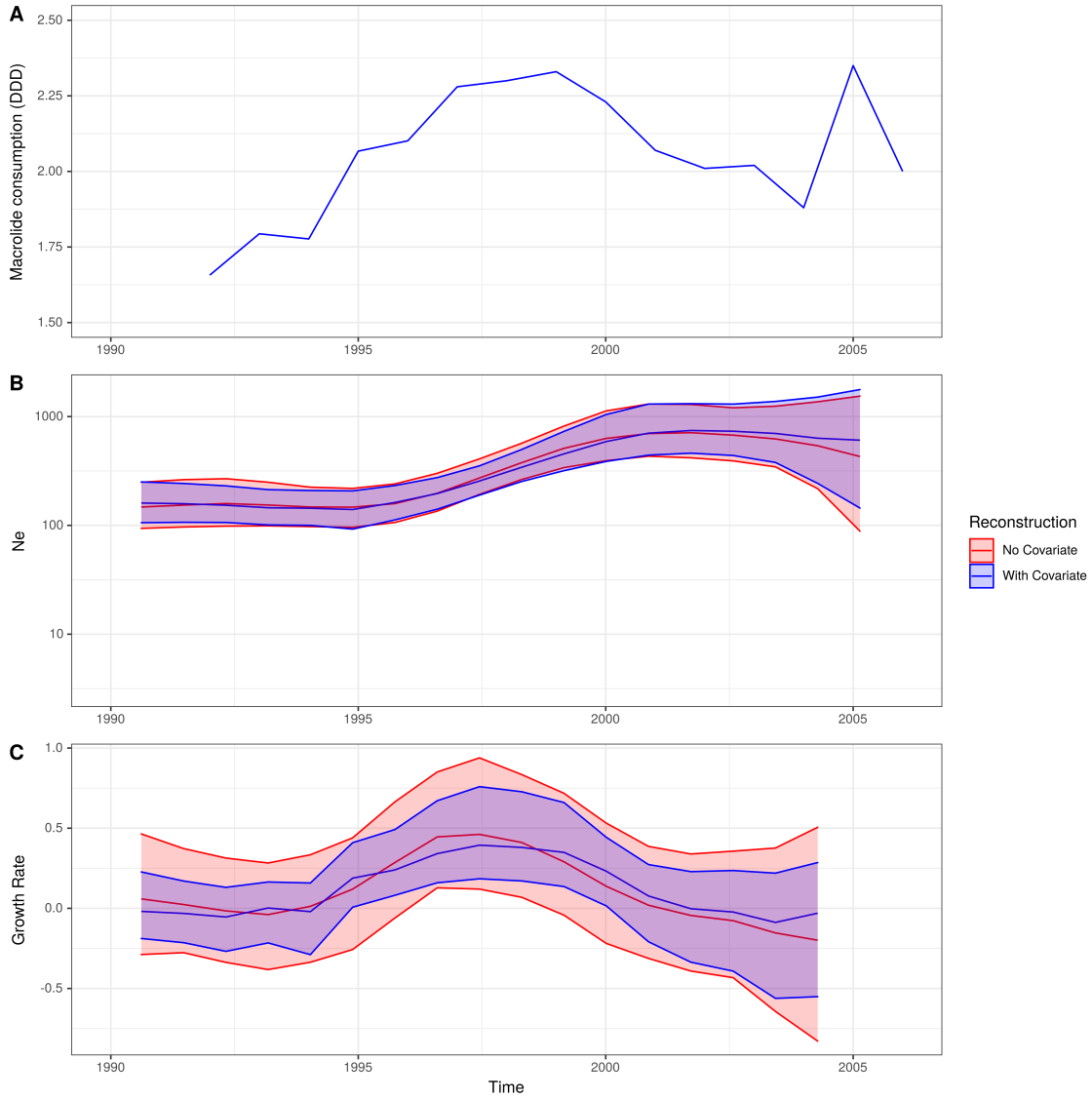

Figure 4.10: Skygrowth analysis incorporating macrolide consumption data (A) The rate of macrolide consumption in Germany in DDD. **(B)** The change in N_e through time inferred by Skygrowth, with the red line representative of when no covariates are incorporated, and the blue line is representative of when the macrolide consumption rate is incorporated into the reconstruction. Shaded regions represent the 95% credible intervals. **(C)** The reconstruction of the growth rate of N_e through time. The red line represents the result of model fitting without covariates, and the blue line when macrolide consumption data were incorporated. Shaded regions represent the 95% credible interval for the reconstruction.

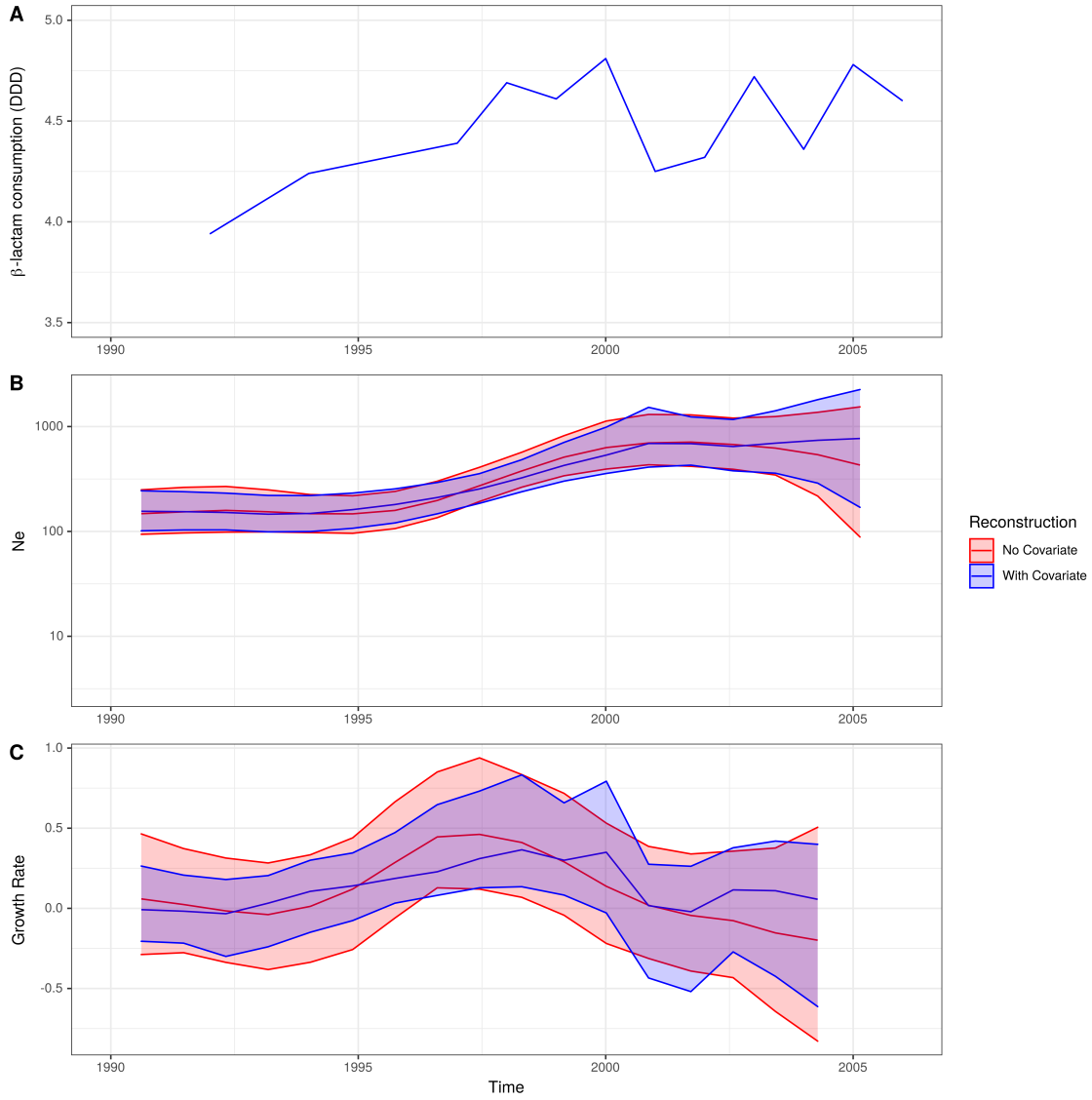

Figure 4.11: Skygrowth analysis incorporating β-lactam consumption data (A) The rate of macrolide consumption in Germany in DDD. **(B)** The change in N_e through time inferred by Skygrowth, with the red line representative of when no covariates are incorporated, and the blue line when the β-lactam consumption rate is incorporated into the reconstruction. Shaded regions represent the 95% credible intervals. **(C)** The reconstruction of the growth rate of N_e through time. The red line represents the result of model fitting without covariates, and the blue line when β-lactam consumption data were incorporated. Shaded regions represent the 95% credible interval for the reconstruction.

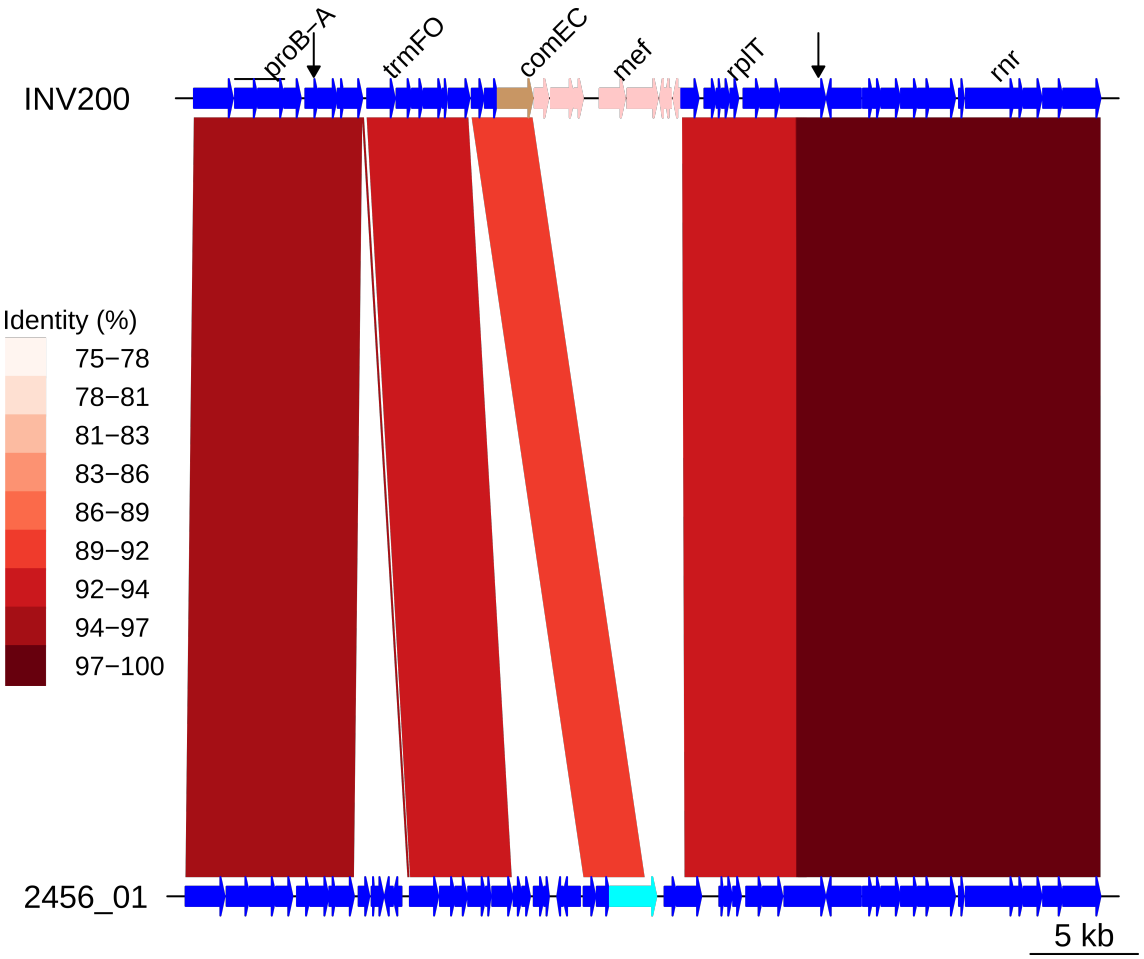

Figure 4.12: Insert of Tn*1207.1* within the PMEN9 reference genome. Comparison of the Tn*1207.1* element insertion, highlighted in pink within the INV200 genome, with the orthologous unmodified locus in sample 2456_01. The red bands between the genomes represent sequence matches identified by BLASTN, with these bars shaded by percentage identity between the sequences. The intact *comEC* gene is coloured cyan within 2456_01, where the fragments of this gene generated by the Tn*1207.1* insertion are coloured brown in INV200. Arrows along the INV200 genome mark the start and end of the recombination event inferred to have imported Tn*1207.1* by the phylogenetic analyses

fragmented hits, due to the nature of these elements and their ability to gain and lose cargo genes to form distinct elements from their Tn*1207.1* and Tn*916* backbones. Therefore, simply taking the presence/absence of a reconstructed element from these BLAST results may give a misleading picture of the number of times an element has inserted. For instance, isolates within a clade may all have a Tn*916* element present, but this may be located at different loci in each isolate and in different forms. Only using the BLAST results would lead to a parsimonious reconstruction of a single insertion, perhaps underestimating the true promiscuity of the element in question.

To identify the unique insertion types and infer the node where this insertion type is likely to have entered into the population, I developed a pipeline to categorise these MGEs (Section 4.2.4; Figure 4.3). Broadly, this categorises the initial BLAST hits based on their length, gene number and flanking region identity, to create a library of unique insertion types. The insertion of these hits are then reconstructed within a strain's phylogeny and cross-referenced with Gubbins predictions of recombination events to detect whether insertion is likely to have occurred through transformation and homologous recombination. If this an insertion is found to be within a putative recombination event, its flanking sequences are compared to a reference streptococcal database to identify the likely species of origin.

4.3.3.2 MGE distribution across the GPS collection

Applying the above pipeline to the GPS collection of 20,015 isolates revealed a wide diversity of insertion sites, and frequent insertion by each of Tn*916*-type and Tn*1207.1*-type elements. At least one of the elements was found in 5,860 isolates (29% of the GPS collection) across 146 resistance-associated GPSCs. Of these, 1,304 isolates contained both Tn*1207.1*-type and Tn*916*-type elements (7%).

The Tn*1207.1*-type element was found across 64 GPSCs in 1,940 isolates (10% of the GPS collection). The mean prevalence of Tn*1207.1*-type elements in GPSCs in which it was present was 16%. Of the 1,940 isolates containing the element, 1,800 (93%) had their insertion point reconstructed. The majority of the 140 isolates where the insertion point was not reconstructed had the Tn*1207.1*-type element present within a small contig with no flanking hits to the reference (74 of 140). There were 50 unique reconstructed

insertion types of the Tn*1207.1*-type element, distributed across 27 different insertion loci (Figure 4.13). Some insertion loci were targeted by multiple insertion types. The loci surrounding the *rlmCD* gene, encoding a 23S rRNA methyltransferase, was the most common target, with 9 different Tn*1207.1* insertion types targeting this region. The most common insertion type was as a Mega-type cassette within a Tn*916*-type element, which occurred in 1,033 (57%) of the isolates. Hence the diversity of Tn*1207.1* insertion types was relatively low, with a Simpson's diversity index of 0.64.

Tn*1207.1*-type elements sometimes disrupted the host cell's machinery upon their integration. For instance, the second most common insertion type for Tn*1207.1* was the 5.5 kb Mega version of the element inserting into, and splitting, *tag*. The *tag* gene encodes a methyladenine glycosylase, involved in DNA base excision repair [635]. This insertion was present in 260 isolates (14% of identified hits) across 30 different GPSCs. This was also common in the PMEN collections, with Tn*1207.1* within the USA clade of PMEN9 being in the form of Mega splitting *tag* (Figure 4.14). The insertion of the 7.2 kb Tn*1207.1* element into *comEC*, as in the German PMEN9 clade, was the third most common, accounting for 5% of insertions (92 isolates) in the GPS collection and appearing in four different GPSCs.

Contrary to the results for PMEN3 and PMEN9, Tn*916*-type elements were more widespread than Tn*1207.1*-types among the collection, being present in 5,230 isolates across 134 of the 146 GPSCs. The mean prevalence for Tn*916* was 41% among GPSCs in which it was present. Of these hits, the insertion sites of 1,895 (36%) were not classifiable. This was primarily due to elements being present in contigs with no, or very short, matches back to the reference (1,496 isolates). For the classifiable 3,335 (64% of insertions) isolates, there were 407 unique reconstructed insertion types, distributed across 102 different insertion sites (Figure 4.15). The insertion sites harbouring the joint greatest number of Tn*916*-type integrations were adjacent to the *rbgA* gene, which encodes a ribosomal biogenesis GTPase and is a common insertion site for Tn*5253* [440], and the *spxA* gene, which encodes a transcriptional regulator for the *comCDE* competence operon [650], with 42 different insertion types proximate to both. The Simpson's diversity index for Tn*916* insertion types was 0.97, indicating Tn*916* is much more variable in how

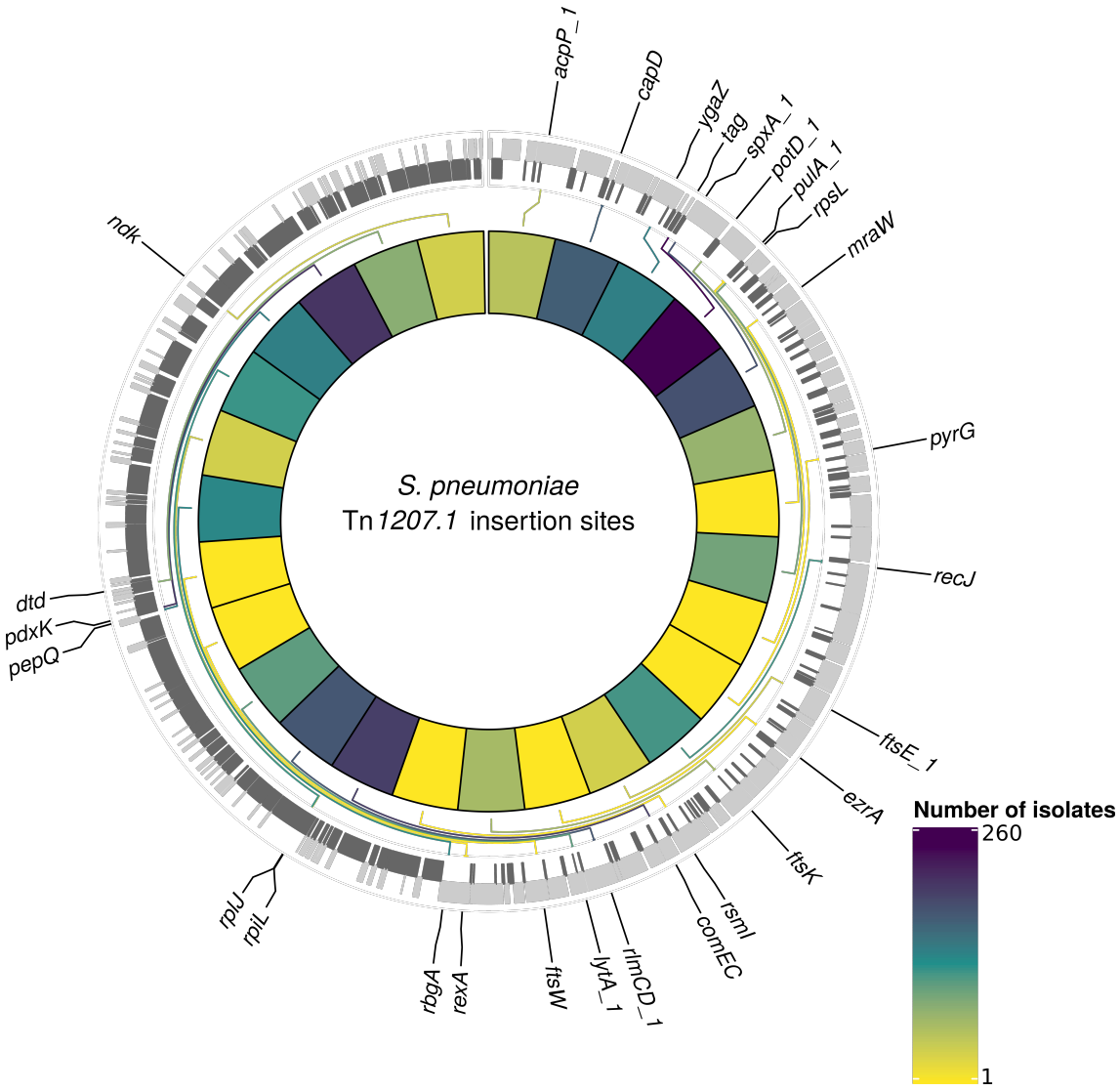

Figure 4.13: Insertion points of classified Tn*1207.1* hits within *S. pneumoniae*. Annotated genome of the reference *Streptococcus pneumoniae* RMV4 isolate (ENA accession number: ERS1681526) showing genes that Tn*1207.1* has inserted either within, or adjacent to, among the collection. Only genes present within this element free reference are annotated. Grey bars represent coding sequences (CDS): lighter grey regions are CDS annotated in the forward strand, darker grey in the reverse. The inner heat map represents the number of isolates that have Tn*1207.1*-type elements inserted within or adjacent to the annotated genes. The colour scale is logarithmically transformed.

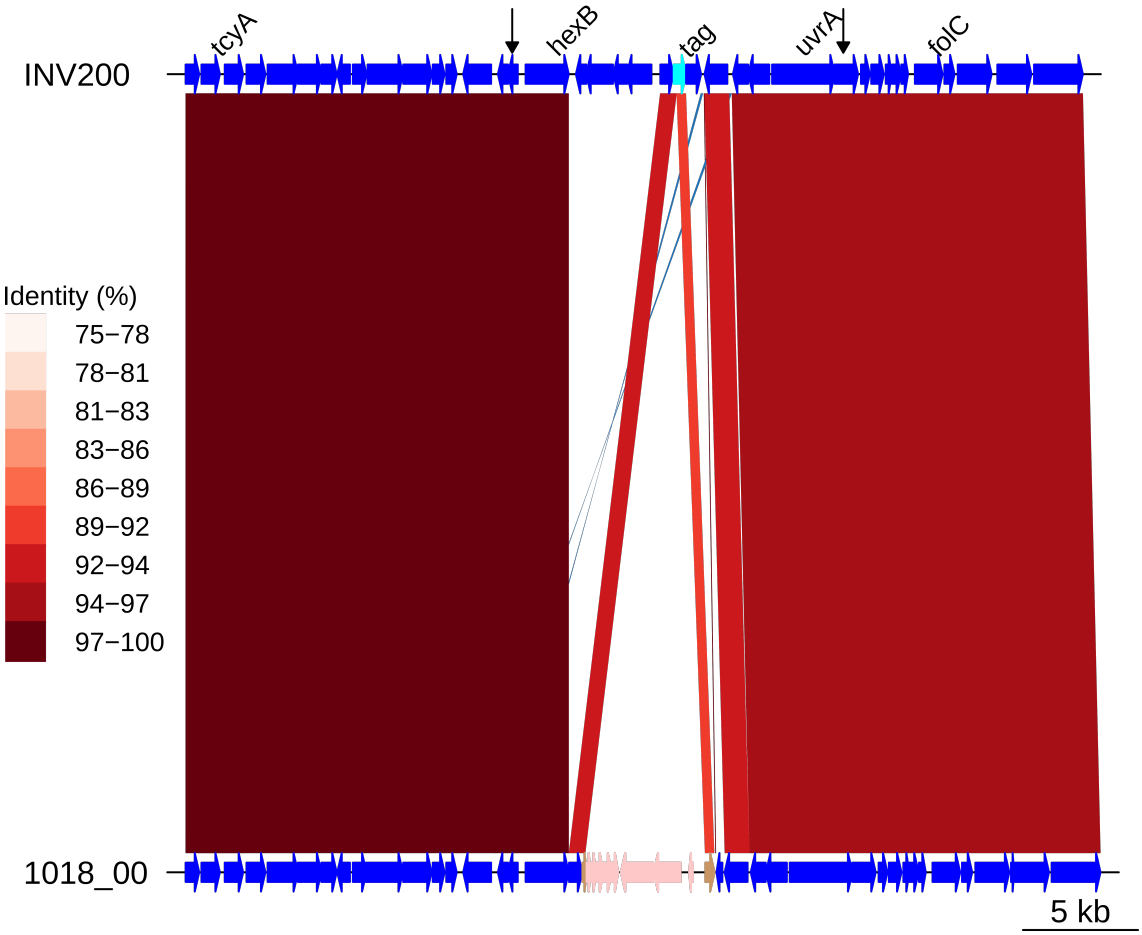

Figure 4.14: Insert of Tn*1207.1* as Mega within *tag*. Comparison of the Tn*1207.1* Mega element insertion, highlighted in pink within the 1018_00 genome, with the orthologous unmodified locus in INV200. Data are shown as described in Figure 4.12

it integrates into the *S. pneumoniae* genome than Tn*1207.1*.

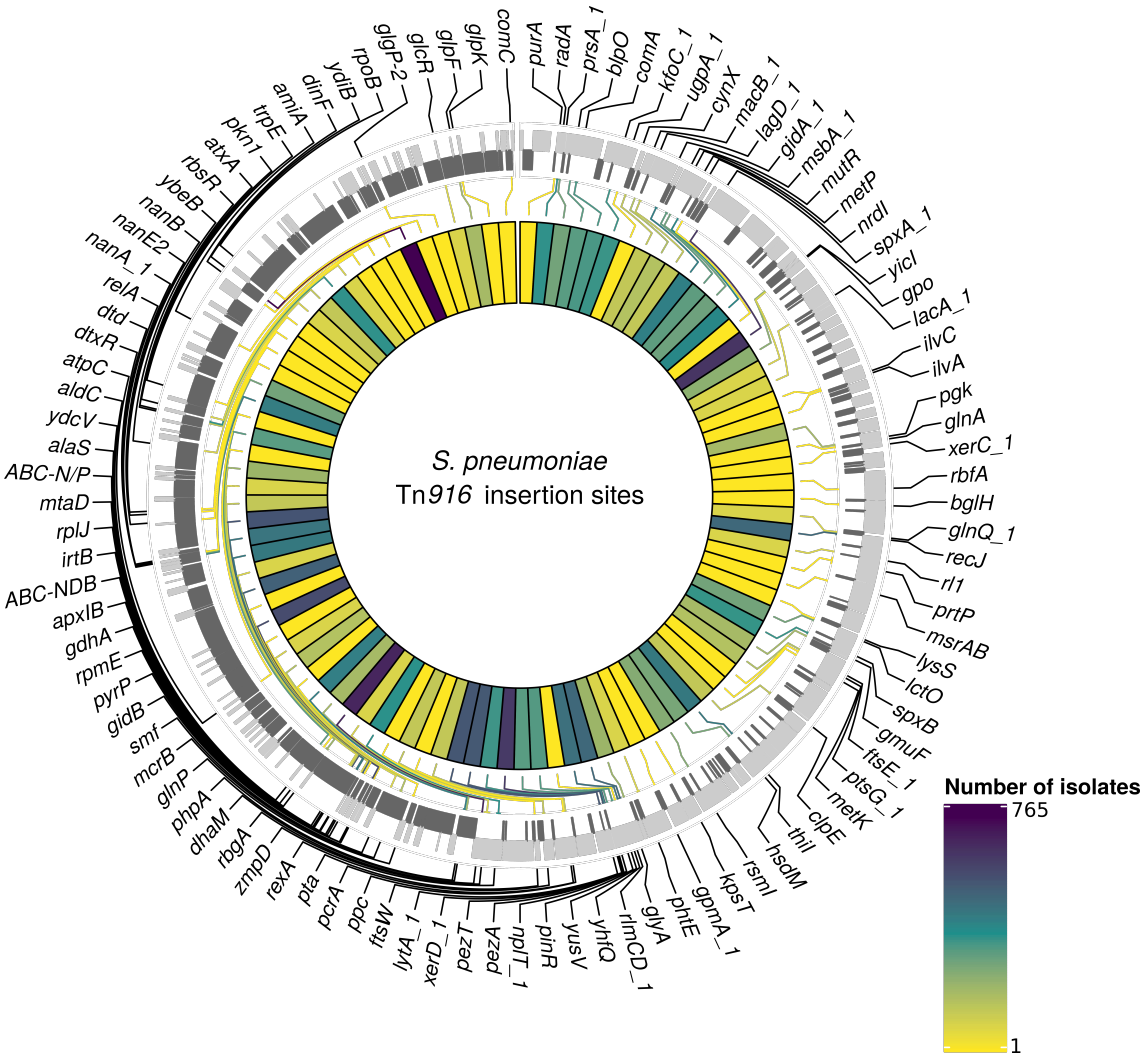

Figure 4.15: Insertion points of classified Tn*916* hits within *S. pneumoniae*. Annotated genome of the reference *Streptococcus pneumoniae* RMV4 isolate (ENA accession number: ERS1681526) with genes where Tn*916*-type elements have inserted either within, or adjacent to, among the collection. Grey bars represent coding sequences (CDS): lighter grey regions are CDS annotated in the forward strand, darker grey in the reverse. The inner heat map represents the number of isolates that have hits inserted within or adjacent to each of the annotated genes. The colour scale is logarithmically transformed.

The most common unique reconstructed insertion types for Tn*916*-type elements were insertions of the Tn*2010* and Tn*2009* elements, which also contain the Mega form of Tn*1207.1*. The majority of both these hits occurred within the globally-distributed GPSC1 lineage [651] (containing the PMEN14 lineage): 364 of the 372 of Tn*2009* examples, and 354 of the 360 of Tn*2010* examples.

Tn*916* was often present in the complete 64.5 kb Tn*5253* element, or with remnants

of the Tn*5253* backbone. The next most common insertion type after Tn*2009* and Tn*2010* was Tn*916* present in 177 isolates as part of a 66kb insertion between the immunoglobulin A protease *zmpA* [652] and *rbgA*. This contained the majority of the Tn*5253* backbone, although the Ω*cat* cassette was missing, and the Tn*916* was present in the form of Tn*2009*. The next most common composite ICE insertion was then an 84 kb element containing Tn*2009* and a Ω*cat* element, present in 59 isolates in GPSC16. The diversity in both Tn*5253*-type insertion sites and cassette content makes accurately reconstructing Tn*916* insertion types difficult. In general, Tn*916* is present in elements over 50 kb in length in 943 isolates (28% of classifiable hits).

4.3.3.3 Frequent insertion of MGEs via recombination

Given this diversity of insertion types, as outlined in section 4.2.4 ancestral state reconstruction was then used to identify the insertions of Tn*916*-type and Tn*1207.1*-type elements across the GPS collection. For Tn*1207.1*-type elements, the 50 unique reconstructed insertion types were found to have inserted 222 times across 59 GPSCs. The most frequent insertion type was the short 4.5 kb element splitting the *tag* gene. This occurred 72 times, representing 32% of all acquisitions of the cassette across *S. pneumoniae*.

Tn*916*-type elements appeared to insert much more frequently across the collection: 1023 times across 128 different GPSCs. While 134 GPSCs contained Tn*916*-type elements, in six of these the insertion types of elements were not able to be reconstructed, hence no insertion events were reconstructed. In total, 163 of the 407 Tn*916* insertion types appeared to insert multiple times across the collection. The most frequent insertion (29 times across 8 different GPSCs) was a 42 kb Tn*5253*-like element, containing only *tetM* as a resistance gene inserted upstream of *zmpA*.

The proportion of these insertions occurring within putative recombinations differed between the two elements. For Tn*1207.1*-type elements, 55% of insertions were within recombination blocks (123 of 222), compared with only 8% of the insertions for Tn*916*-type elements (81 of 1023). This difference could have multiple explanations. Tn*916* encodes for its own conjugative machinery, and is often present within larger conjugative elements, and therefore may frequently move independently of transformation.

Alternatively, Tn*916* may be imported through transformation, but then transpose between loci once in a cell, thus moving away from its site of insertion. The median Simpson's diversity index for within-GPSC Tn*916*-type element insertion site diversity was 0.54, whereas for Tn*1207.1*-type elements it was 0.25. This suggests Tn*916*-type elements, once inserted, might excise and transpose within the chromosome at a higher rate than Tn*1207.1*-type elements.

To assess the properties of these recombination events importing MGEs, they were labelled and compared against the non-MGE importing recombination events from all 146 GPSCs where either of the elements were found. From this, most recombination events mediated by transformation appear to be much shorter than the lengths of these elements (Figure 4.16). Such exchanges will generally favour deletion of elements rather than their insertion [485]. Comparisons of the length distribution and SNP density for recombination events that import one of the MGEs (with the length of the MGE seqeuence itself excluded), against other recombination events, suggested these MGE importation recombinations were atypical (Figure 4.16). MGE recombinations were significantly longer, with a median length (excluding the length of the element itself) of 10.9 kb, compared to a median length of 7.4 kb for non-MGE recombinations (Mann-Whitney U test; U = 7775792, n1 = 183, n2 = 66419, two-sided p = 6.18 x 10^{-11}). Additionally, the median SNP density was significantly higher for MGE recombinations, at 4.41 SNPs per kb, compared to non-MGE recombinations with a median of 3.49 SNPs per kb (Mann-Whitney U test; U = 7643988, n1 = 183, n2 = 66419, two-sided p = 1.62 x 10^{-9}). Given the pneumococcus tends to be conserved at core genome loci, the higher SNP density of these transformation events inserting MGEs suggested they may arise from donors of other species [298].

4.3.4 The interspecies origin of MGE importation events

To test whether these recombination events were likely to be incorporating DNA from other species, the flanking regions of these insertions were compared against the reference Streptococcal database, as outlined in section 4.2.4 and Figure 4.3. This analysis of the flanking regions of those MGE insertions occurring within homologous recombinations, revealed that many sequences matched more closely to non-pneumococcal streptococci than pneumococcal references. For Tn*1207.1*-type elements, the median γ score was

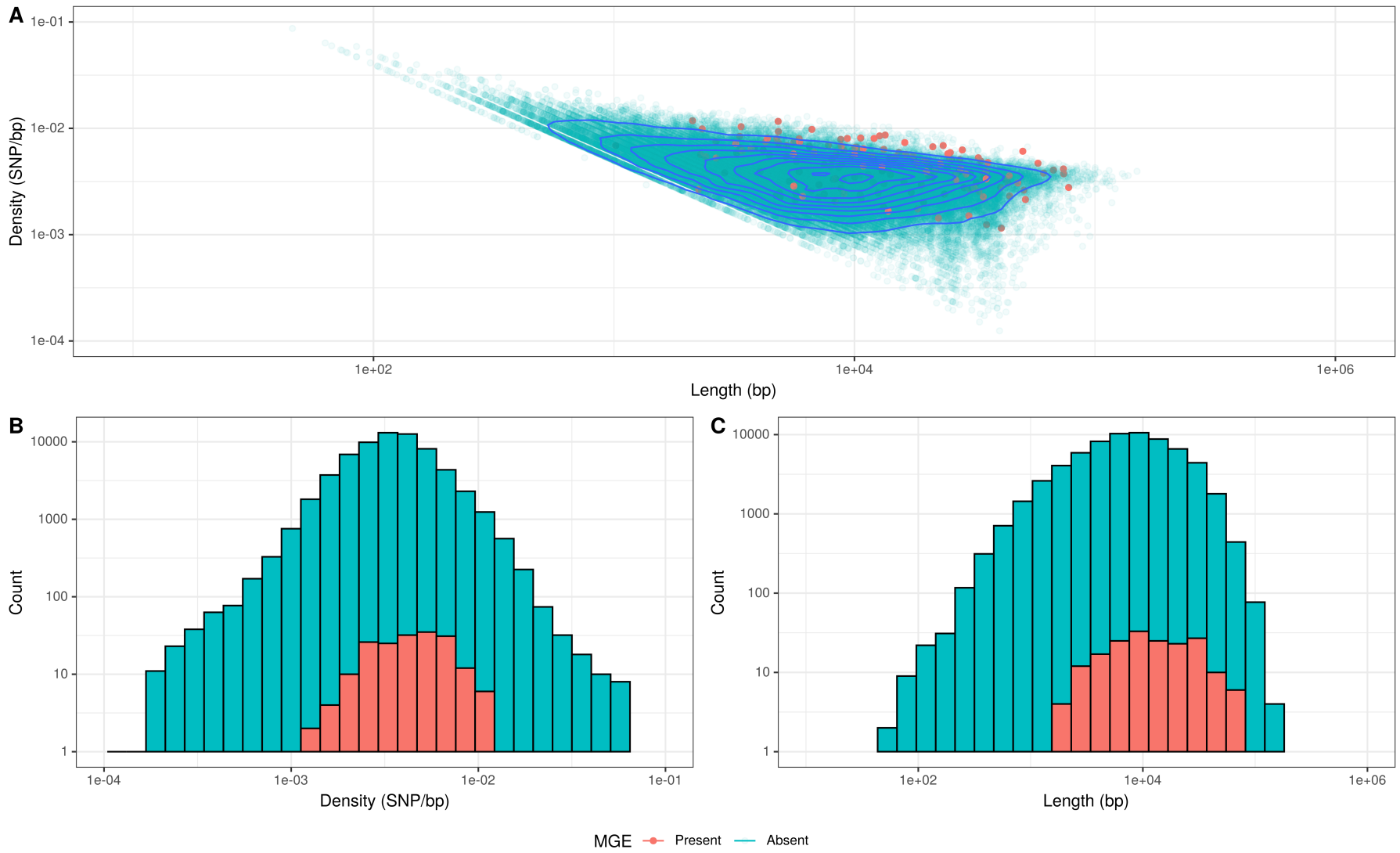

Figure 4.16: Comparison of length and SNP density of recombination events. (A) Joint plot of the SNP density and number of bases within homologous recombinations, with events classified based on whether or not they imported an MGE. Blue contour lines represent the density of points. **(B)** Overlaid histogram comparing the SNP density in recombination events for MGE import vs non-MGE import. **(C)** Overlaid histogram of the length of recombinations, comparing MGE import vs non-MGE import events.

0.88 for insertions across the flanking lengths and insertion types. For control isolates, where the element was not inserted and the orthologous flanking regions were extracted, the median γ score was 1.0. The overall distribution of γ scores was significantly lower between the control and MGE isolates (Mann-Whitney U = 2860659, n1 = n2 = 3690, two-sided, $p < 2.2 \times 10^{-16}$). This lower score for MGE flanks, relative to orthologous regions in isolates without the MGE, likely represents MGEs being acquired from other species.

For Tn*916*-type elements' insertions within recombination blocks, the median γ scores for both control and MGE isolates was 1.0. However, a Mann-Whitney U test revealed significant difference between the control and MGE isolates γ scores, with Tn*916*-type insertions scoring lower (U = 1642284, n1 = 2260, n2 = 2250, two-sided, $p < 2.2 \times 10^{-16}$). Hence there is evidence that some of the Tn*916*-type insertions occurred through inter-species recombinations.

The trends in the most closely-matching species to the flanking regions, over increasing distance from the MGE, follow the expectation for interspecies transfers (Figure 4.17). The control flanking regions matched most closely to pneumococci at all tested lengths. For MGE insertion flanks, non-pneumococcal species matches were much more frequent closer to the insertion. As the flank length increased from 500 bp to 7500 bp, and linkage to the integrated resistance genes decreased, the recombinant isolates were more likely to match pneumococcal DNA.

For Tn*1207.1*-type elements, it appeared *S. mitis* was the most likely donor (Figure 4.17). In the regions upstream of the Tn*1207.1* insertion, *S. mitis* was the top match for 92% of 500 bp long flanks. Even at longer flank lengths, *S. mitis* was still the leading match for upstream regions, although for downstream regions the pneumococcus tended to become the predominant match to flanks by 4000 bp outside of the insertion.

The most common Tn*1207.1*-type insertion, that splitting the *tag* gene, can be used to illustrate the local import of sequence from another species (Figure 4.18). For the downstream flanks (Figure 4.18 B) this score trend appeared roughly linear with increasing flank length, with the evidence for imported *S. mitis* sequence disappearing 4000 bp from the insertion site. However, for the upstream flanking regions (Figure 4.18 A), the median γ score remained low with increasing flank length, with a median of 0.83 at 7500 bp

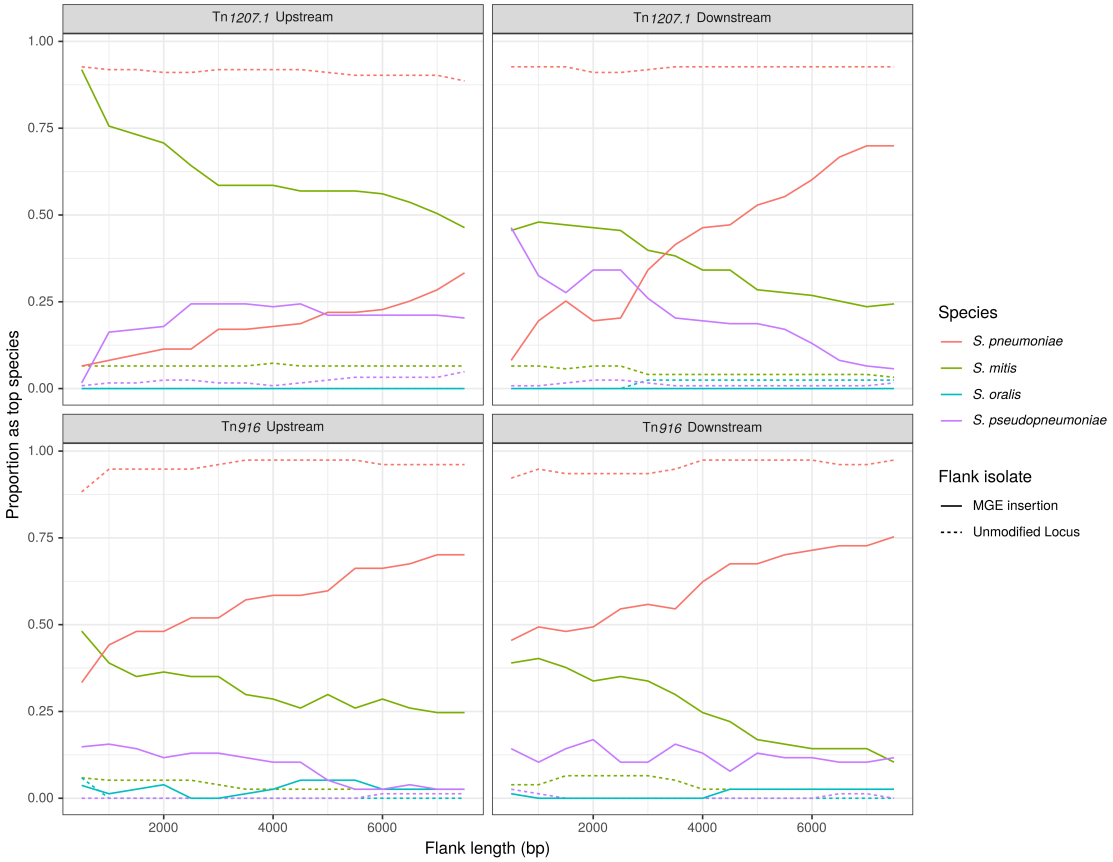

Figure 4.17: The likely origin of MGE insertions. Flanking regions upstream and downstream from MGE insertions sites were compared to a reference streptococcal database. Solid Lines represent the proportion of matches, across all insertions reconstructed to have occurred in homologous recombinations, that correspond to each of the four species present in the reference streptococcal database. The dashed unmodified locus lines represent data from the orthologous regions of isolates without the MGE insertion. These proportions are calculated over a range of flanking region lengths around the insertion site.

upstream of the insertion. This upstream region, replaced by *S. mitis* sequence in many isolates, extended into the *uvrA* gene, another component, like *tag*, of the nucleotide excision repair machinery within the pneumococcus. The consistency of top matches for this *tag* Tn*1207.1* insertion type was high across the 66 independent acquisitions within recombination events (Table 4.1). For the 500 bp upstream region, 85% of the insertions (56 of the 66 within recombination blocks) had the *S. mitis* 21/39 (accession code AYRR00000000) reference as their top hit. In total 97% of these upstream regions of *tag* insertions (64 of 66) had their top hit as an *S. mitis* sequence. Such consistency suggests these imports originated from a single insertion in the *S. mitis* population. These imports are then likely to have moved between pneumococci multiple times.

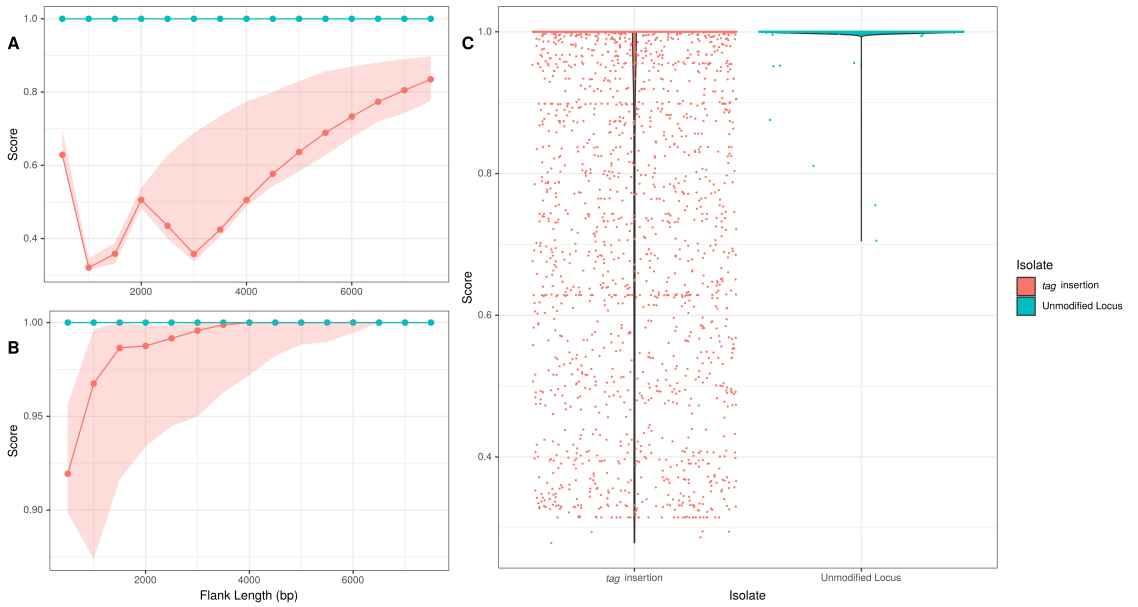

Figure 4.18: Flanking region origin for Tn*1207.1 tag* insertions. **A** The median γ score of upstream flanking regions of the Tn*1207.1* insertions into the *tag* gene, coloured in coral. The median γ score for the orthologous regions in isolates without the MGE, not expected to be modified by interspecies homologous recombination, are coloured in cyan. Shaded regions represent the inter-quartile range (IQR) of the γ score. **B** The γ score for regions extracted downstream of the insertion. Shaded regions represent the IQR of the γ score. **C** The distribution of γ scores across both flanks, upstream and downstream, of the Tn*1207.1 tag* insertions. Data are shown separately for isolates with this particular insert, and those without Tn*1207.1* integrated in this orthologous region.

The signal for the interspecies origins of Tn*916*-type element insertions was less pronounced (Figure 4.17). While the pneumococcus tends to be the most frequent match for the regions flanking Tn*916*-type element insertions, the proportion of matches to the pneumococcus is still lower than seen in the control isolates, suggesting a detectable contribution of interspecies transformation. Within the PMEN lineages 13 of the 61 insertions (21%) of Tn*916*-type elements were detected to have inserted within putative recombination events. All 13 of these had either their immediate upstream or downstream (or both) regions matching most closely to non-pneumococcal species. To verify these insertions were from interspecies recombination events, a selection were also investigated manually. This applied to independent insertions near *recJ* (Figure 4.19); *gmuF*, which encodes mannose-6-phosphate isomerase (also known as *manA*; Figure 4.20), and *gidB* (Figure 4.21). Of these, the upstream regions of the *recJ* and *gmuF* insertions were identified by the algorithm to match most closely to *S. mitis*, while the upstream region of the *gidB* insertion matched most closely to *S. pseudopneumoniae*. The relatively low percent

identity scores in the flanking regions inspected manually also indicates these elements were likely imported from another species.

Insertions detected outside of putative recombination events were also inspected. Tn*916* inserted near *rplL*, with flanking remnants of Tn*5252*, on three independent occasions within PMEN3 (Figures 4.22:4.24). Given the sequence divergence in the flanking regions from the reference, these insertions were also likely to be interspecies in origin. These were likely missed due to inaccurate reconstruction of the insertion node of these elements. The algorithmic results missing these isolates, suggests that these results may underestimate the overall contribution of interspecies homologous recombination in the spread of Tn*916*-type elements.

Species	ENA accession	Frequency
S.mitis	AYRR00000000	56
S.mitis	AYRS00000000	2
S.mitis	JPFW00000000	2
S.mitis	AEDU00000000	1
S.mitis	AEDV00000000	1
S.mitis	AJJL00000000	1
S.mitis	AQTU00000000	1
S.pseudopneumoniae	AYRN00000000	1
S.pneumoniae	D39	1

Table 4.1: Closest species match to 500 bp region upstream of *tag* disrupting Tn*1207.1* insertion. The species, ENA accessions code and frequency of the matches to the 500 bp region upstream of the 66 Tn*1207.1 tag* insertions reconstructed to have occurred within inferred homologous recombination events.

4.4 Conclusions

In this chapter I have investigated the spread, insertion loci and likely origins of two common families of MGEs, related to Tn*1207.1* and Tn*916* elements. Within the PMEN3 and PMEN9 lineages, these elements are widespread, although their distribution follows different patterns. In PMEN3 elements appear to insert often, but do not tend to lead to successful clade expansion, whereas in PMEN9 elements insert less frequently but they insert at the bases of highly successful clades. This echos the results of the previous chapter, depicting how different evolutionary histories can lead to the international dissemination of MDR clones.

When I further investigate the success of the German lineage in PMEN9, we can

see how the effects of local selection pressures can enable this seemingly deleterious MGE insertion to spread widely. The splitting of the *comEC* gene here represents an insertion site that is beneficial to the MGE, as the abrogation of competence in this clade would prevent its deletion through subsequent recombination with donors lacking the insertion [350]. This insertion can expand though due to the unique local antibiotic consumption patterns that favour this macrolide resistant cassette. However, the loss of transformability in this lineage likely prevented vaccine escape through transformation mediated serotype switching. These dynamics have also been observed in Iceland with the PMEN2 lineage, where the ΦIC1 prophage disrupted the competence gene *comYC* and resulted in a decline in the lineage [312]. These phylodynamic results though, are based on the combination of two antibiotic consumption datasets with slightly different coverage populations. Current reporting of antibiotic consumption is of much higher standards, as such future phylodynamic analyses will be able to test the relationship between clade expansion and antibiotic consumption more rigorously.

In the wider GPS collection these elements were found to be common too. Tn*916*-type elements are more frequently observed in the collection, present in 134 different GPSCs in over 400 different forms at over 100 different loci within the pneumococcal chromosome. This diversity in insertion site is likely driven by the *int* gene of Tn*916* with its low insertion site specificity [425, 429].

For Tn*1207.1*, as in the literature [634, 635], we only observe the 7.2 kb element inserting into the *comEC* gene across the collection. Instead the majority of this elements spread is via the 5.5 kb Mega form that is often found within larger elements such as Tn*2010* or Tn*2009*. The most frequent insertion type for this element outside of these larger constructs was splitting the *tag* gene. This encodes a methyl-adenine glycosylase involved in DNA base repair. The splitting of this gene may disrupt DNA base repair, therefore leading to a hyper-mutator phenotype. This would most likely be detrimental to the host cell. Outside of the pneumococcus there are examples of other mobile elements inserting into mutation repair machinery, causing mutator phenotypes. For instance, in group B streptococcus and *Vibrio splendidus* MGEs insert between the *mutS* and *mutL* genes involved in mismatch base repair [653–655]. These MGEs however, appear to

excise during different stages of cell growth, only to re-enter and disrupt the genes in question, producing mutator phenotypes, during later phases of growth. In both these species, these MGEs appear functionally under the control of the host cell [654, 655]. It is unclear if the Tn*1207.1* Mega element can similarly excise and reinsert under host cell control.

The number of interspecies transformation events linked to the spread of both elements is striking within the GPS collection. Typically longer recombination events, such as those importing these elements, tend to be much rarer due to transformation's biases against the import of longer sequences [485]. Previous work by Chancey *et al* 2015 [656] has looked at the movement of Tn*916*-type elements, including larger Tn*5253* constructs, among pneumococcal isolates in Atlanta. They concluded that these elements initially inserted into pneumococcal populations via conjugation, but then transformation would facilitate their intraspecies spread. My work also highlights intra-species transformation events, but reveals that these elements can also move via interspecies transformation events. There is growing evidence of the importance of interspecies transformation in creating MDR bacterial lineages, with recent work in *Acinetobacter baumannii* also suggesting large elements moving between species within the *Acinetobacter* genus [657].

In this work though, I make no attempt to definitively determine the species of origin for these recombination events. While our reference database is sufficient to split likely non-pneumococcal from pneumococcal DNA, it is not fine-grained enough to fully delineate the networks through which AMR genes spread. Much greater sampling of commensal streptococcal species is needed in order to assess the most likely donor species for these interspecies transformations.

One further caveat with our results is the large number of isolates that were unable to be classified to a hit type, especially concerning the diverse Tn*916*-type hits. Typically pipelines for detecting MGEs in genomic data rely on finding specific motifs or repeats that demarcate DNA foreign to the rest of the genome [658–661]. These methods are accurate at defining the bounds of an MGE that can insert via site specific recombination. However, in order to assess the origin of an insert moving via transformation and homologous recombination, it is necessary to find the flanking regions that match to a within

species reference. Poor assembly quality and a lack of mapping to a reference drives the large number of unassigned hits with Tn*916*-type elements. Accurate long-read sequencing methods, from which better assemblies would be produced, could enable more elements to be fully realised within the genome output.

In the next chapter I will investigate the recombination dynamics of two other species, *Acinetobacter baumanii* and *Legionella pneumophila*, using the updates in the Gubbins software presented in Chapter 2. These species also contain AMR encoding MGEs that can disrupt the transformation machinery, providing further evidence of the intra-genomic conflict between MGEs and host bacteria.

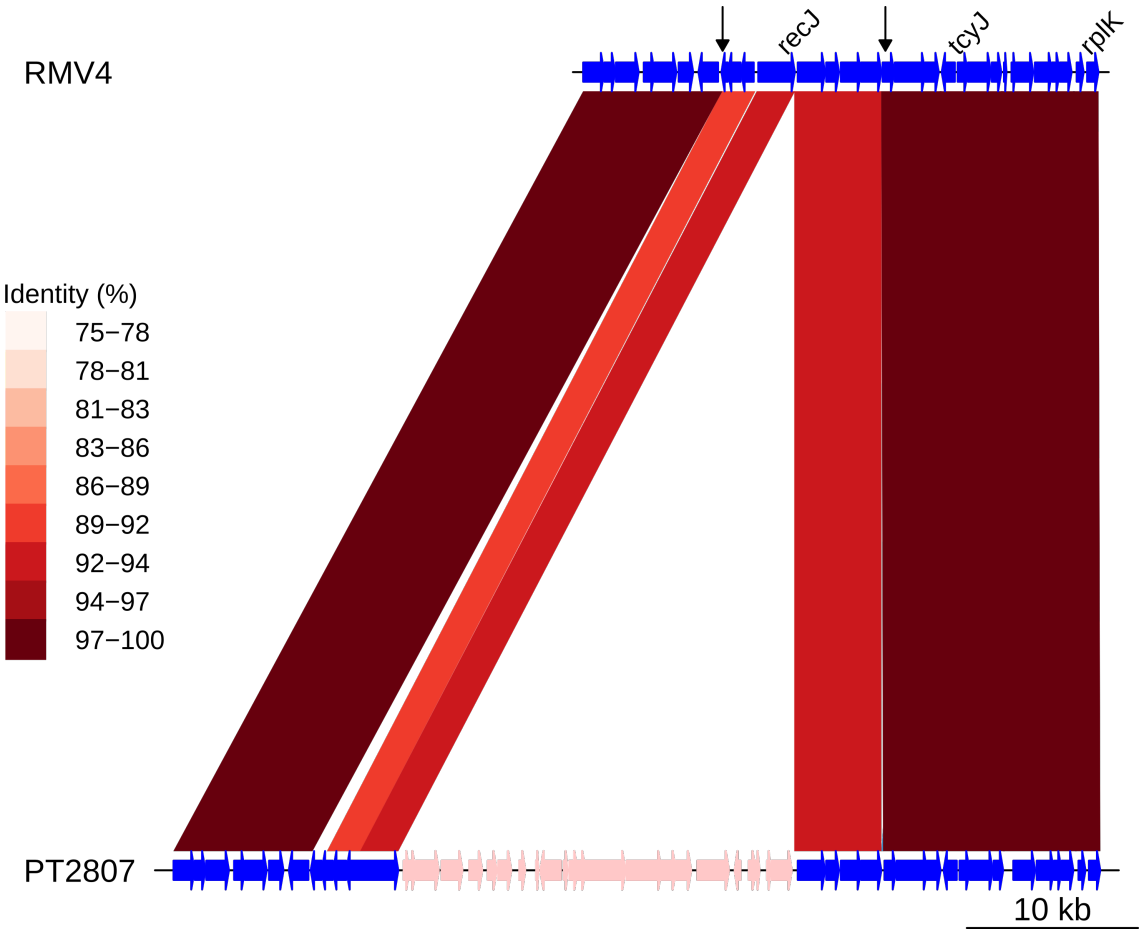

Figure 4.19: Insert of Tn*916* downstream of *recJ*. Comparison of the Tn*916* element insertion, highlighted in pink within the PT2807 genome, with the orthologous unmodified locus in the RMV4 sample. Data shown are as described in Figure 4.12.

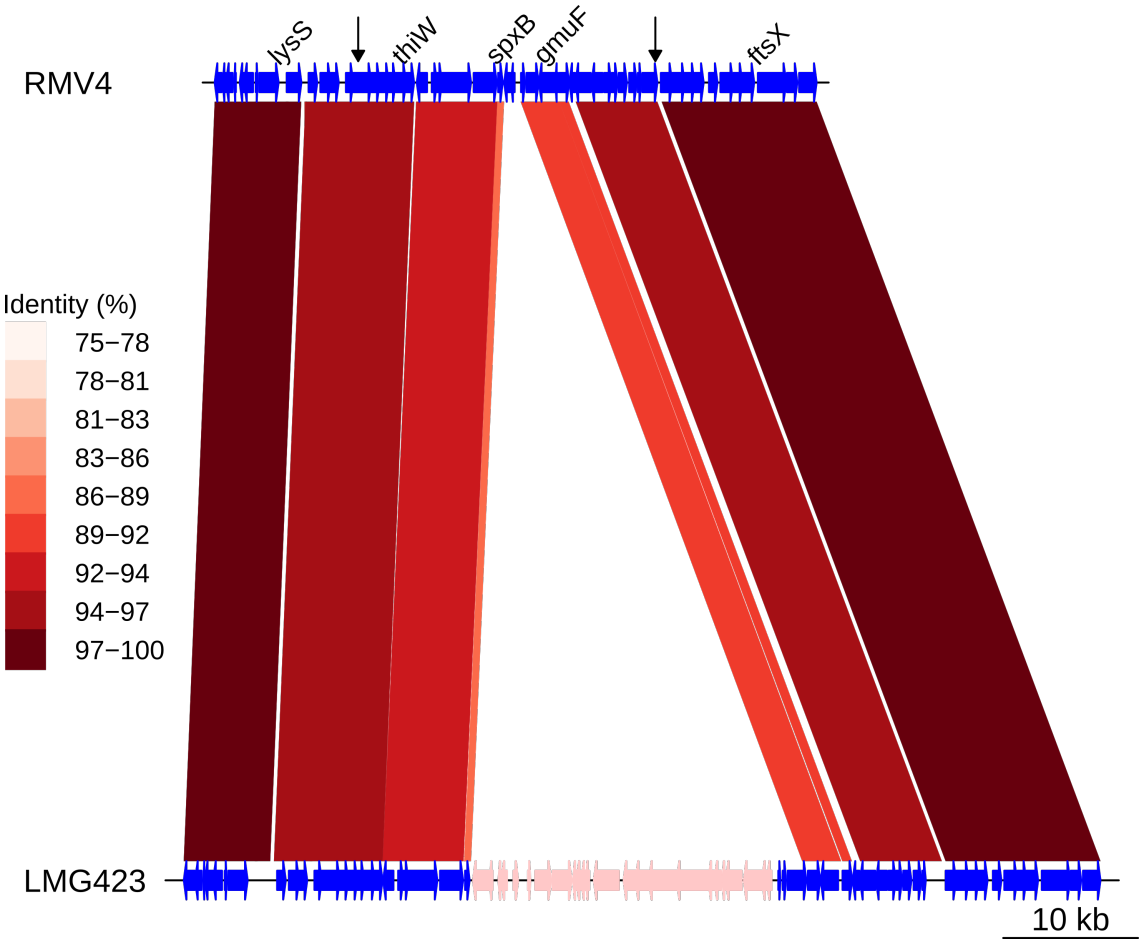

Figure 4.20: Insert of Tn*916* downstream of *gmuF*. Comparison of the Tn*916* element insertion, highlighted in pink within the LMG423 genome, with the orthologous unmodified locus in the RMV4 sample. Data shown are as described in Figure 4.12.

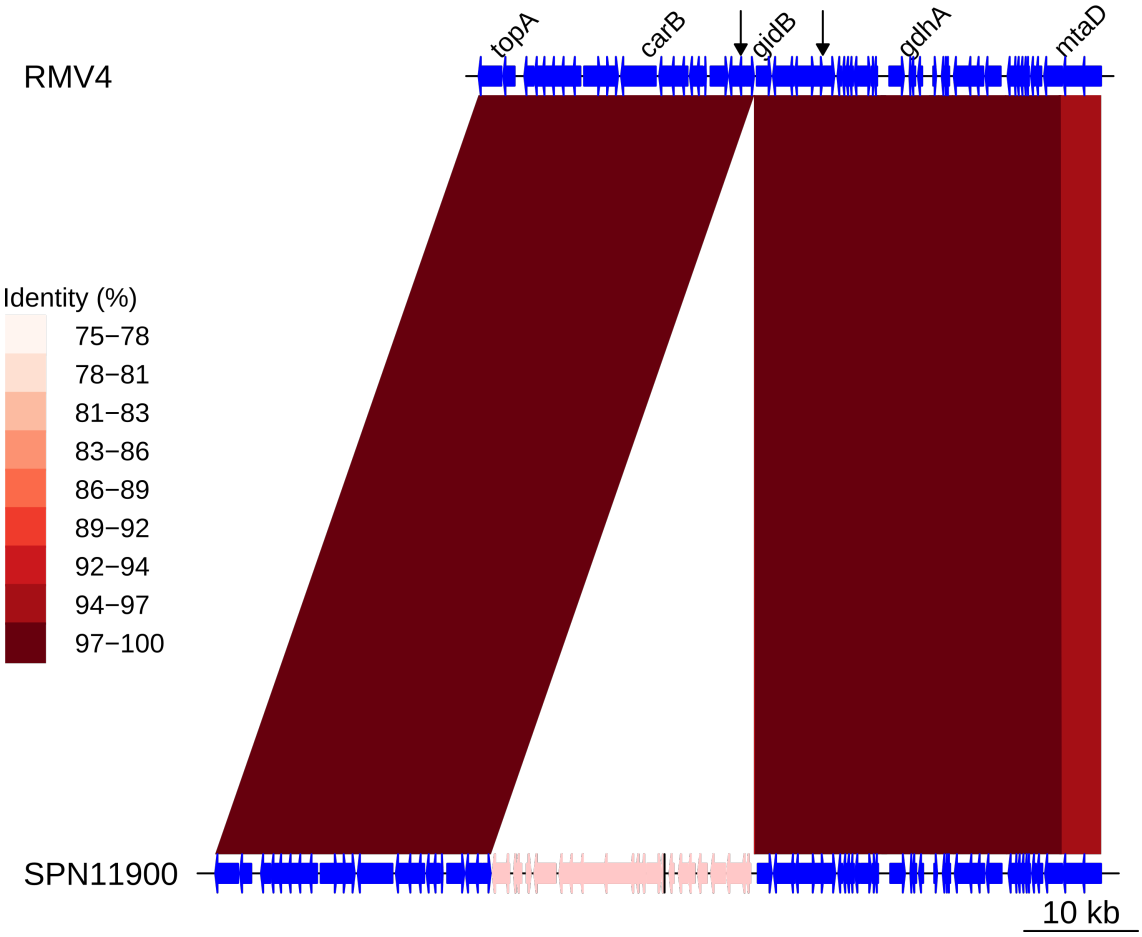

Figure 4.21: Insert of Tn*916* upstream of *gidB*. Comparison of the Tn*916* element insertion, highlighted in pink within the SPN11900 genome, with the ortholgous unmodified locus in the RMV4 sample. Data shown are as described in Figure 4.12.

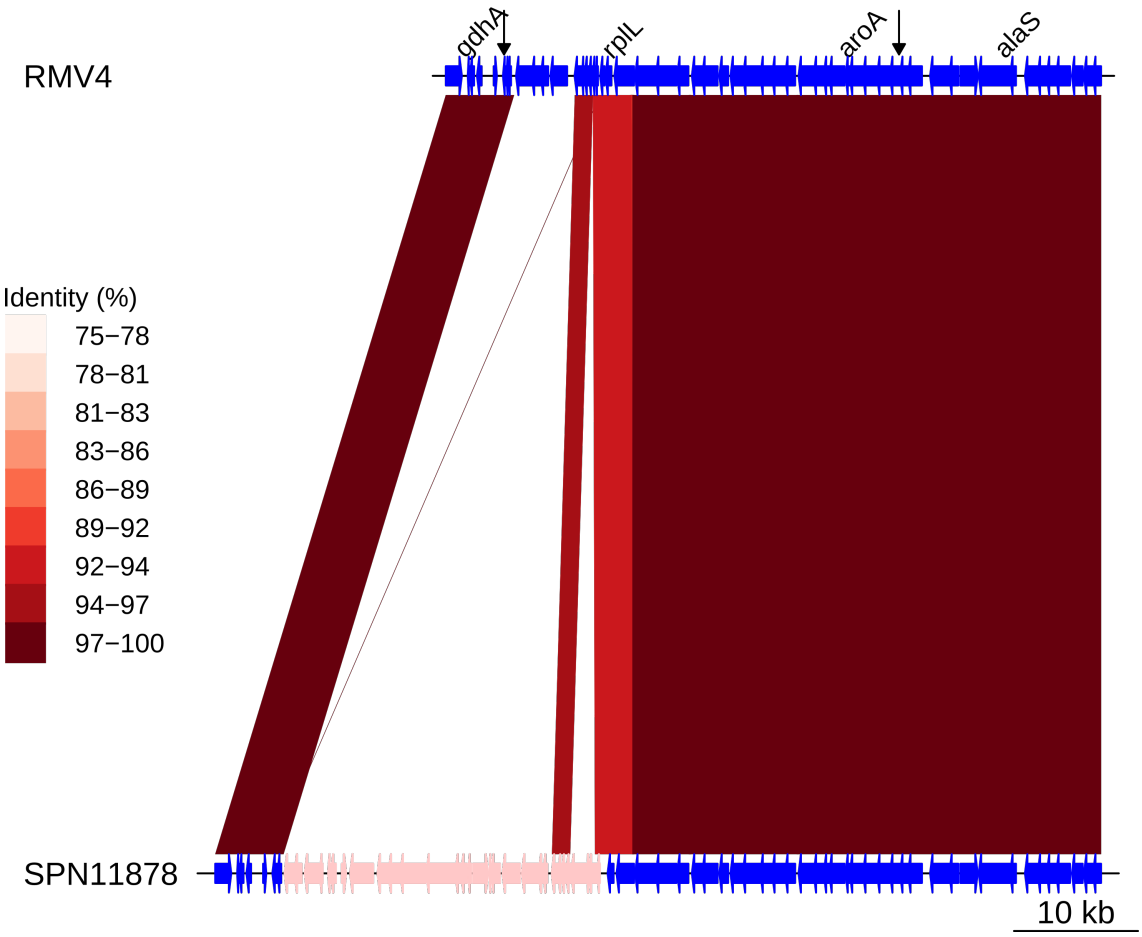

Figure 4.22: Insert of Tn*916* upstream of *rplL*.Comparison of the Tn*916* element insertion, highlighted in pink within the SPN11878 genome, with the orthologous unmodified locus in the RMV4 sample. Data shown are as described in Figure 4.12.

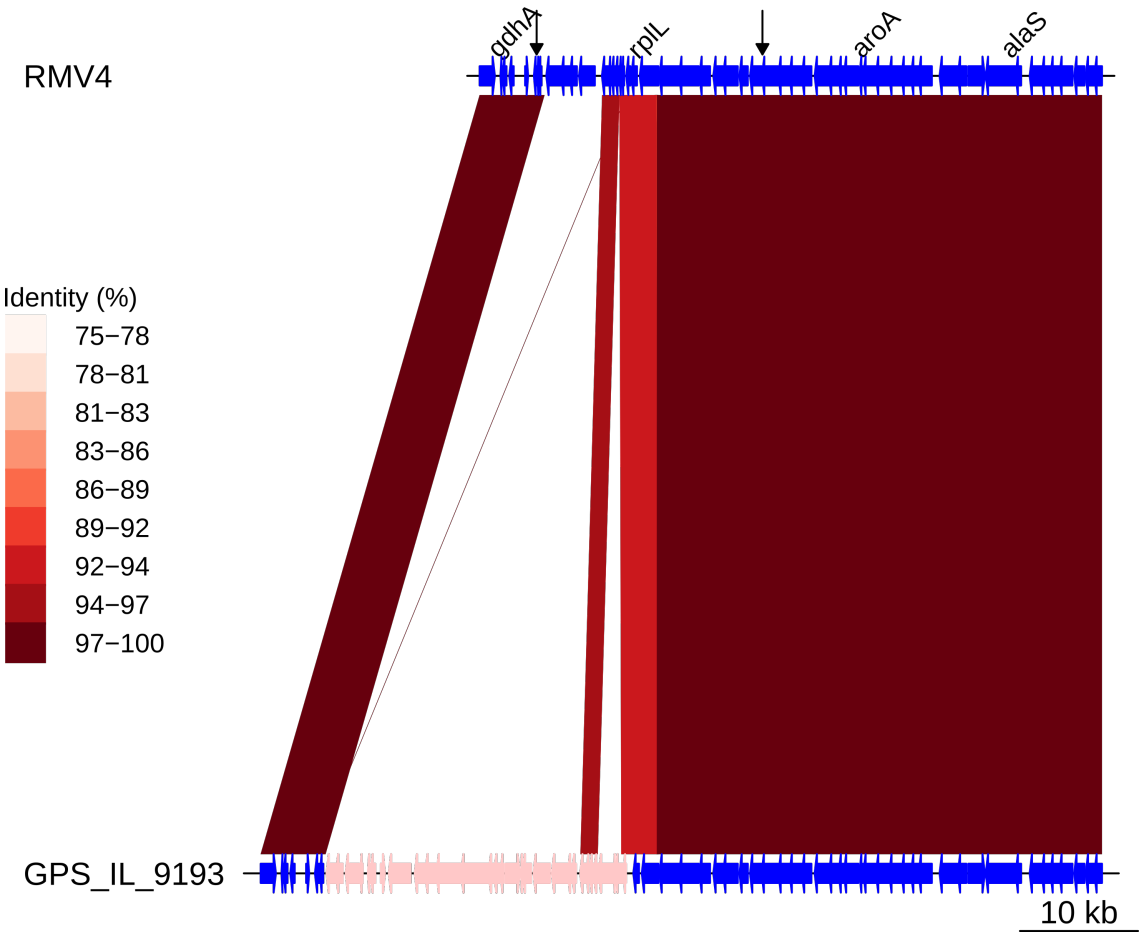

Figure 4.23: Insert of Tn*916* upstream of *rplL*. Comparison of the Tn*916* element insertion, highlighted in pink within the GPS_IL_9193 genome, with the orthologous unmodified locus in the RMV4 sample with no insertion. Data are as described in Figure 4.12.

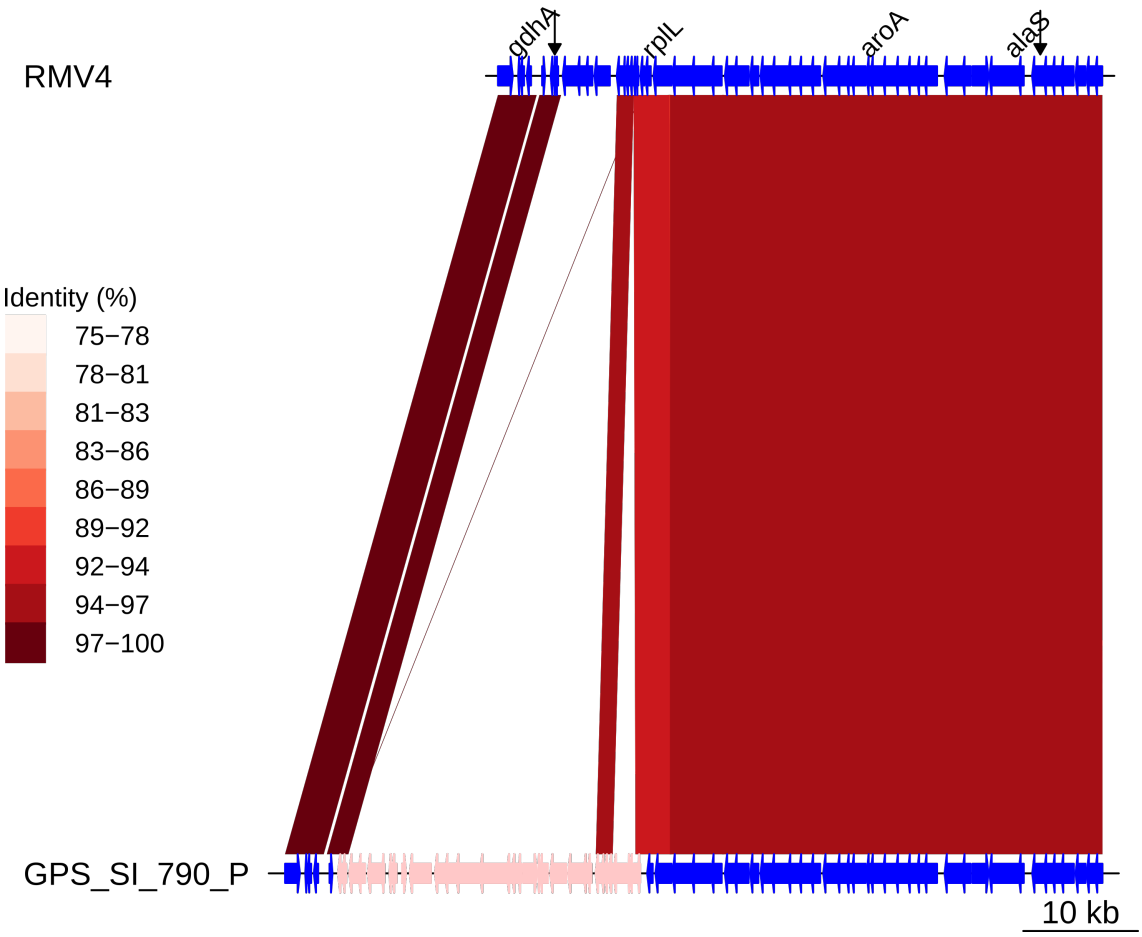

Figure 4.24: Insert of Tn*916* upstream of *rplL*. Comparison of the Tn*916* element insertion, highlighted in pink within the GPS_SI_790_P genome, and the orthologous unmodified locus in the RMV4 sample with no insertion. Data are as described in Figure 4.12.

Chapter 5

Investigating host MGE conflict in gram-negative bacterial species

Summary

Transformation may have evolved as a strategy to cure bacterial genomes of large selfish mobile genetic elements (MGEs). Within the gram-negative pathogens A. baumannii and L. pneumophila, recent in vitro work has shown mobile genetic elements targeting the competence machinery involved in transformation. In this chapter I collate publicly available genomes from both species to assess these MGEs effects on recombination dynamics at a population level. I find evidence for the disruption of competence machinery is widespread in A. baumannii. However there appears to be no consistent significant effect on recombination dynamics. In L. pneumophila on the other hand, disruption is rarer, but has the consistent effect of reducing recombination frequency. Despite this, recombinations around immunogenic loci are still common. These results highlight how selection can drive recombination events, even with a reduced frequency of transformation.

5.1 Introduction

In the previous chapter we have seen how an MGE, Tn*1207.1*, can insert within and disrupt the competence machinery of a host. A recent theory [350] has been proposed to

explain why events of this nature, where the host transformation machinery is disrupted by a MGE, can increase the fitness of MGEs. Croucher *et al* 2016 posit that, given the length asymmetry of transformation imports, transformation is able to cure selfish MGEs more efficiently than it can insert them [350]. Any disruption to the transformation machinery, therefore, can prevent an MGE's deletion and allow these selfish elements to spread further in a population [350]. Aside from Tn*1207.1* inserting within *comEC* in pneumococci, there are numerous other examples of MGEs disrupting transformation machinery in order to prevent their deletion [472, 662]. Prophage have been seen to insert within the competence pilus structural gene *comYC* [142] and its orthologs among streptococci [350, 394, 663]. Additionally, within *Staphylococcus pseudintermedius*, prophage have been seen to disrupt the *comG* operon, also involved in the formation of competence pilins [664]. Prophage also insert into the *comK* gene, a transcription activator for all known genes related to competence [665], among *Listeria* species [350, 666].

Recent experimental work has identified MGEs that inhibit transformation in gram-negative pathogens too. Within *A. baumannii* work has examined how AbaR-type genomic islands (AbaRs) insert within and disrupt the *comM* gene, which encodes an ATPase involved in transformation [667]. MGEs which do not insert within the chromosome also appear to disrupt transformation machinery. In *Legionella pneumophila* isolates a plasmid, pLPL, encodes RocRp a small RNA (sRNA) which blocks isolates from expressing their DNA uptake machinery [668]. Both these systems, AbaR and RocRp disruption of transformation, have been extensively studied in the lab [657, 667, 668]. In this chapter I investigate the impact of these disruptions on the global population structure of *A. baumannii* and *L. pneumophila* isolates. Using large publicly available WGS datasets I investigate the distribution of these disruption events and their effect on recombination rates within lineages. Now I will go into more detail about both the species and systems investigated and then describe my methods.

5.1.1 *Acinetobacter baumannii* and AbaRs

A. baumannii is an opportunistic nosocomial pathogen that can cause ventillator-associated pneumonia, burn infections, wound infections and bloodstream infections, among other pathologies [669–671]. It is an incredibly hardy pathogen, able to survive extremes of pH

and desiccation, while its ability to form a biofilm is linked to resistance to disinfectants [670,672]. This means *A. baumannii* is readily able to cause outbreaks in hospitals, where very stringent infection control methods required to mitigate its further spread [673–677]. Outbreaks caused by *A. baumannii* have been recorded around the world, with the majority of these caused by only two successful major clones: global clone 1 (GC1) & global clone 2 (GC2) [670, 678]. Worryingly, both these clones are associated with high levels of resistance, with an increasing number of carbapenem resistant outbreaks a particular concern [549, 670, 679]. This has led to *A. baumannii* being classified as one of the six ESKAPE pathogens, identified by the WHO as species which cause the majority of AMR nosocomial infections [161, 680].

One class of elements which drives the levels of resistance seen in *A. baumannii* clones are the AbaRs. AbaR1 was the first large cluster of AMR genes sequenced from an MDR *A. baumannii* GC1 strain, a 86 kb resistance island that contained 18 different, in some cases redundant, AMR genes [673, 681, 682]. Since this first detection there has been a wide diversity of AbaR elements detected, with GC1 & GC2 tending to have distinct AbaR elements [683]. The AbaR3-type islands are commonly found in GC1 and have a core backbone based on the Tn*6019* transposon, and tend to be present within the host chromosome [683–685]. In contrast GC2 tends to contain the AbaR4-type islands. These are based on a backbone of the Tn*6022* transposon, commonly carry the $bla_{OXA\text{-}23}$ carbapenem resistance gene and have been found in plasmids [683,686]. AbaRs encode their own transposition machinery, allowing for movement within host DNA, be it plasmidic or chromosomal in nature [687].

When inserted within the host chromosome however, a common target site for all AbaRs is within the *comM* gene. Insertion here results in a 5 bp target site duplication at both ends of the AbaR element, which is indicative of insertion through transposition [673, 688]. The *comM* gene encodes for hexameric helicase that can promote the integration of DNA through its role in branch migration [689]. It is not necessary for transformation in *V. cholerae* and *H. influenzae*, although its deletion does result in at least a 100 fold reduction in transformation efficiency [668, 689, 690]. Recent work in *A. baumannii* by Godeux *et al* looked at the *in vitro* effect of a 19.7 kbp AbaR11 and a 86 kbp of AbaR1

insertion into *comM* [667]. They tested the frequency of transformation events in these isolates using a 3.8 kbp long PCR product, finding that isolates with either AbaR insertion also had at least a 10 fold reduction in transformation frequency compared to those with a repaired *comM* gene [657,667]. Work on 45 WGSs of GC1 *A. baumannii*, while not directly investigating this, appears to show AbaR insertion does not appear to affect acquisition of different capsule gene clusters [549]. This analysis can now be undertaken species-wide with the thousands of genomes available in the International Nucleotide Sequence Database Collaboration (INSDC).

5.1.2 *Legionella pneumophila* and pLPL

The genus *Legionella* was first described following an outbreak of severe pneumonia among attendees at an American Legion convention in Philadelphia in 1976 [691]. This severe pneumonia pathology became known as Legionnaire's disease (LeD), and is caused by the inhalation of contaminated aerosols containing these aquatic bacteria [691, 692]. Of the over 60 different *Legionella* species described so far, the majority are only known to infect aquatic protozoans and arthropods, with only accidental spillover into human hosts [693, 694]. Among these species, *L. pneumophila* is responsible for up to 95% of the LeD diagnosed worldwide [695, 696]. Within *L. pneumophila* only a few successful clones also account for the majority of LD, with five sequence types (STs) accounting for almost half of the LeD cases in northwest Europe [692, 697]. While antibiotic resistance has been observed in *L. pneumophila* isolates, particularly to macrolides, in general most clinical relevant clones appear to be susceptible to common frontline antibiotics [698].

Early studies looking into the recombination dynamics within *L. pneumophila* have shown the species to undergo homologous recombination frequently, with events transferring segments up to 200 kb in length [699,700]. More recent work by David *et al* 2017, focusing on clinically relevant clones, has calculated a very high r/m values for lineages [699]. For instance the ST23 lineage had an r/m value of 93.8, detected using Gubbins, with particular hotspots around the outer membrane TolC-like proteins and the Dot/Icm effectors that are vital for intracellular growth within a host eukaryotic cell [699,701].

For *L. pneumophila* cells, and bacteria more generally, to acquire DNA via transformation they must reach a competent state [441]. *L. pneumophila* appears to lack a

transcriptional activator that can activate its competence machinery. This is unlike well-studied transformable bacteria like the pneumococcus and *H. influenzae*, which use σ^X factors or TfpX/Sxy respectively [702, 703]. Instead, recent work by Attaiech *et al* has shown that competence is repressed during the exponential phase of *L. pneumophila* by the sRNA RocR and the RNA chaperone RocC [702]. A competent state is then only reached during the midlog and stationary phase with the reduced expression of RocR, which is unlike other transformable species where competence is upregulated by a transcriptional activator.

In *L. pneumophila* it appears that certain conjugative elements have also hijacked this competence regulatory system. Work from Durieux *et al* revealed that a plasmid, pLPL, encodes a homolog to the RocR sRNA, RocRp [668]. When expression of RocR decreases during the transition and stationary phases, the RocRp sRNA instead binds to RocC and represses the expression of the competence machinery [668]. Unlike the disruption of *comM* in *A. baumannii* by AbaR elements, RocRp expression appears to effectively silence transformation, causing reductions of 10^3 or 10^4 fold in the transformation frequency [668]. In this Chapter I will analyse whether these changes in transformation affect the epidemiology and evolution of *L. pneumophila*.

5.2 Methods

5.2.1 Isolate collections

For both *A. baumannii* and *L. pneumophila*, all available genomes assemblies from the NCBI GenBank database were downloaded. For *A. baumannii* all isolates identified as *A. baumannii* and available on 31/03/2021 were downloaded. This encompassed a total of 8,431 WGS assemblies. Four sequences less than 1 Mbp long were removed from further analysis, leaving a total of 8,427 isolates.

All isolates identified as *L. pneumophila* in the NCBI GenBank database and available on 17/01/2022 were downloaded. In total 3,373 genomes were downloaded, one of which contained less than 1Mbp of sequence and was removed, leaving a total collection of 3,372 for further analysis. For both collections the assemblies ranged from full chromosome assemblies to contig and scaffold based assemblies (Table 5.1).

Assembly level	*Acinetobacter baumannii*	*Legionella pneumophila*
Chromosome	26	0
Complete Genome	266	104
Contig	6341	3117
Scaffold	1798	152
Total	8431	3373

Table 5.1: Assembly levels of the collections for *A. baumannii* and *L. pneunmophila* from GenBank.

5.2.2 Population structure and quality control of assemblies

PopPUNK v2.4.5 was run on the two collections, to define their population structure and to perform quality control (QC) on the assemblies [298]. PopPUNK calculates the core and accessory distance between isolates, assigning them to strains based on these distances. The default QC of PopPUNK is to test for an excess of ambiguous base calls among isolates and to detect outliers in overall genome length. For *A. baumannii* this QC was extended by using a type isolate to compare against. The ATCC 19606 strain sequence (ENA accession SRR10295884), assembled as a complete chromosome from PacBio machines was chosen as the type isolate [704]. Isolates outside of two standard deviations of the length of the collection were removed, as well as those with a core distance greater than 0.03 and an accessory distance greater than 0.8 from the type isolate. For *A. baumannii* this results in a loss of 154 isolates. For *L. pneumophila* a total of six isolates were removed due to initial QC methods based on ambiguous base calls and overal genome length being five standard deviations from the population mean. Visual inspection of the distance plots produced from PopPUNK was also used to further remove outlier sequences from both collections. For *A. baumannii*, three isolates that appeared frequently in a cluster of distances that had 0 core distance but greater than 0.3 accessory distance were removed. This left a total collection of 8,270 isolates. For *L. pneumophila*, isolates over-represented in a cluster of distances with greater than 0.5 accessory distance were removed. This led to 250 isolates being removed, leaving a collection of 3,116 isolates for further analysis.

Once these collections had been narrowed through the above QC steps, a model was fitted and refined using PopPUNK to assign strains. Both models were fit using the dbscan

option, which employs the HDBSCAN clustering model [705]. The default parameters for the maximum number of clusters produced and the minimum proportion of samples in a cluster were used. A 2D linear boundary separating within and between strain distances was identified using the refine mode of PopPUNK.

In order to compare the strains produced by PopPUNK on the *A. baumannii* collection, these isolates were also typed with the Pasteur MLST profile from PubMLST [706]. For *L. pneumophila* isolates, there is no robust typing method for assemblies [245]. With this collection, therefore, I adapted the algorithm used by Gordon *et al* [245] to run solely on assemblies. The code for this is available at `http://github.com/jdaeth274/lp_mlst`.

5.2.3 Detecting disrupted competence machinery

Within the *A. baumannii* collection, the presence of a disrupted *comM* was detected using a BLAST search approach. The complete 1,488 bp *comM* gene sequence from isolate XH856 (GenBank accession number CP014541.1) was used as a reference. Isolates were taken to have a complete *comM* gene if at least one BLAST hit had a length > 1450bp matching to the reference *comM* sequence.

AbaR sequences were also identified in the *A. baumannii* collection in a similar manner. The conserved regions of AbaRs identified in Bi *et al* [683], were used to identify the outer regions of a putative AbaR element. The left end region was 2,891 bp long, while the right end region was 1871 bp long. If an element had a BLAST hit to both the conserved left end and conserved right end, regardless of contig position or orientation of the conserved ends, this was considered evidence for the presence of an AbaR. For the left end region, a BLAST hit was considered a hit with an align length of 2880 bp or greater, for the right end region a hit was considered to be an align length of 1860 bp or longer. This appeared to be a slightly more liberal approach to identifying AbaRs than employed in Bi *et al*, who also included a minimum distance between left and right end hits [683].

For the detection of RocRp within the *L. pneumophila* collections, a BLAST-based search method was also used. The 70 nt form of RocRp was extracted from the 59,832 bp pLPL plasmid of the Lens strain of *L. pneumophila* (plasmid accession: CR628339).

The RocRp sequence was extracted from bases 38,417 to 38,348 of the pLPL plasmid as outlined in Durieux *et al* [668]. This was then used as a query to BLAST search the *L. pneumophila* collection for the presence of RocRp. An evalue threshold of 0.001 was set, 301 of the 302 BLAST hits were of the exact length and 100% sequence identity. The 302nd hit was included, this was 63 bp long and had 98.4% identity, due to RocRp also existing in 65, 68 and 70 bp forms.

5.2.4 Detecting recombination dynamics

For the *A. baumannii* collection the six most common strains identified in the above PopPUNK analysis were chosen for further recombination investigation of their recombination dynamics. These six strains represented 6,409 total isolates. For each lineage, a reference isolate was chosen as an isolate with a whole chromosome assembly level, with, if possible, no AbaR present. SKA v1.0.0 [546] was then used to created a mapped alignment to the reference isolate for each lineage. Gubbins v3.2.0 [489] was then run on these alignments to detect recombination dynamics. FastTree was used as the initial phylogeny builder, and RAxML the main iteration tree builder. Both of these were run with a GTR model, and a joint ancestral state reconstruction was used.

For the *L. pneumophila* collections, the same protocol was followed. The three largest strains, representing 1854 isolates, were selected for further recombination dynamic analysis. Again the references for each strain were chosen as isolates with a complete chromosome. The strain 1 reference, isolate Flint 2 (D-7477) (ENA accession: CP021281), also contained a plasmid sequence which was removed prior to the mapping process. This mapped alignment was also created with SKA v1.0.0 [546]. Gubbins v3.2.0 was then run on these mapped alignments, with a FastTree initial phylogeny builder, RAxML main iteration builder, both run with a GTR substitution model, and a joint ancestral state reconstruction model.

5.3 Results

5.3.1 Population structure of *Acinetobacter baumannii*

In order to assess the impact of two separate MGEs: AbaRs and pLPL, on the recombination dynamics of isolates, two collections of bacteria were curated. For AbaR elements,

with their frequent insertion loci of *comM*, a collection of 8,431 *Acinetobacter baumannii* assemblies were downloaded from the NCBI GenBank database. Of these 8,270 sequences passed QC for further investigation. These were then analysed using PopPUNK v2.4.5 [298] to determine the population structure of the collection. Looking at the distance plots produced by the collection, there appears to be rapid gene content diversification within *A. baumannii* strains (Figure 5.1A). For instance, at low levels of core genome distance, there is a wide variation in the accessory distance, from 0 to 0.2. This could be driven by high rates of recombination or rapid MGE movement between strains. The between-strain distances appear to be primarily concentrated in one dense cluster which centers around an accessory distance of 0.3 and a core distance of 0.0175. This clustering is also consistent with simulations performed in Lees *et al* [298] of populations with high recombination rates.

PopPUNK stratifies the *A. baumannii* species into 403 separate strains, these range in size from a single isolate (of which there are 255 strains) to the largest strain, strain 1, with 5,092 isolates (Figure 5.2). The two largest strains, strain 1 with 5,092 isolates and strain 2 with 505 isolates, correspond to MLST types 2 and 1 respectively. MLSTs 1 and 2 are commonly referred to as the GC1 and GC2 clones respectively, and cause the majority of detected *A. baumannii* outbreaks around the world [670, 678]. That these two clones represent the majority of isolates (67.7%) within the GenBank database likley reflects a bias towards sequencing outbreak isolates.

More broadly, the strains produced by PopPUNK appear to closely match the MLST results. The similarity between the two typing methods was measured through the adjusted Rand index, which gives a score of zero for completely different clustering and one for identical clustering [298, 707, 708]. This is also adjusted for the chance overlap of strains. The adjusted Rand index score for the PopPUNK and MLST strains was 0.91, indicating very similar clustering. However, there were 198 isolates untypeable through the Pasteur MLST scheme available on PubMLST [706]. The majority of these (145 of 198, 73.2%) were due to loci combinations not previously seen within *A. baumannii*. Here the use of PopPUNK is advantageous in that the addition of previously unseen genotypes can be automated.

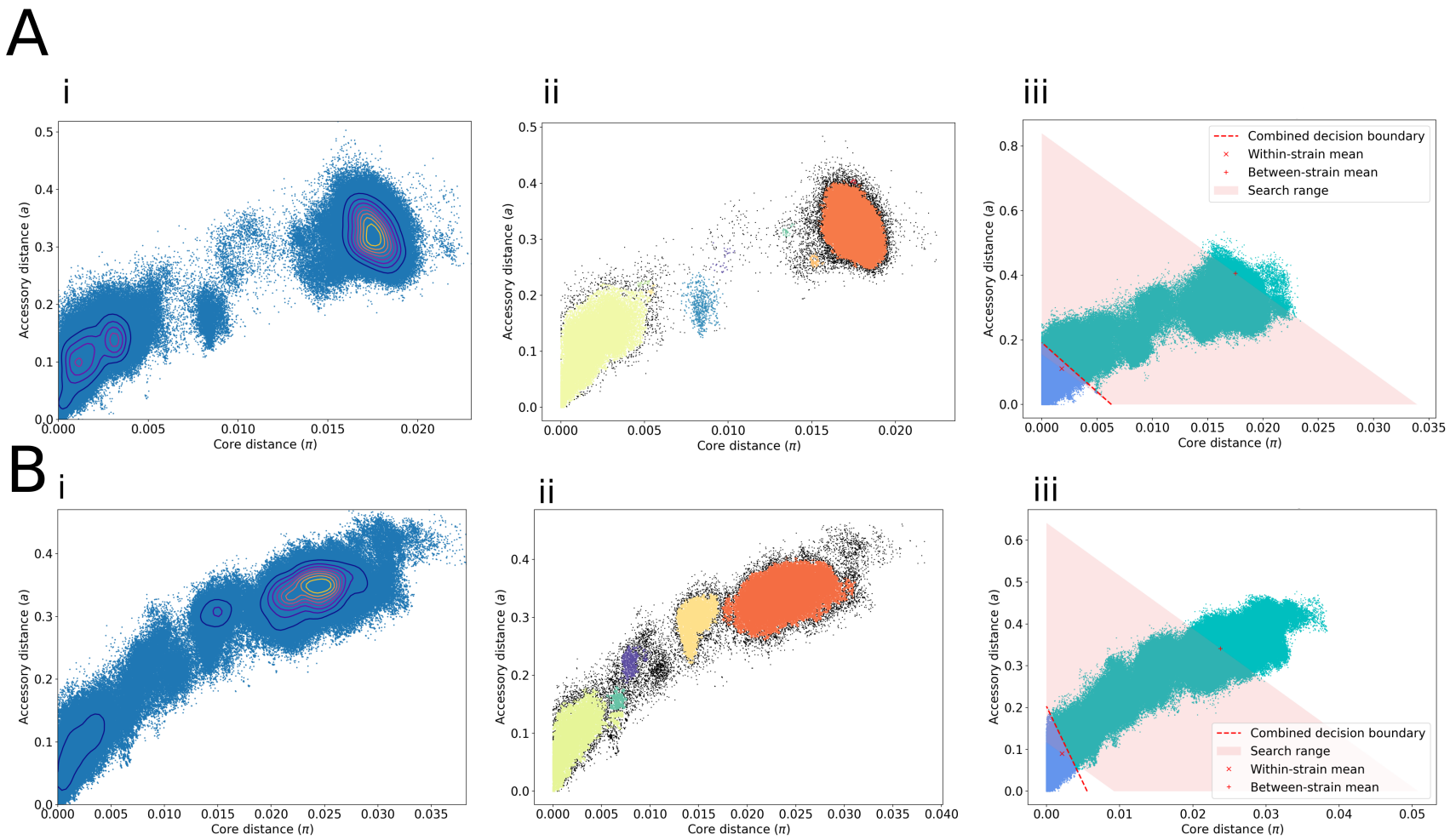

Figure 5.1: Distance calculations for *Acinetobacter baumannii* and *Legionella pneumophila* populations. **A** The results from the *Acinetobacter baumannii* collection. **A(i)** The initial pairwise distance plots between all isolates for the core and accessory distances calculated for the collection. **A(ii)** The HDBSCAN model fit to the distances. The figure displays only a subsample of 10^4 distances. This estimated a total of 9 different spatial clusters. Unclassified noise points are shown in black. **A(iii)** The refined model fit maximising the gradient and intercept of the line separating within strain from between strain differences. **B** The results for the *Legionella pneumophila* collection, data are as displayed in **A**. Within **B(ii)**, the HDBSCAN model estimated 5 separate spatial clusters.

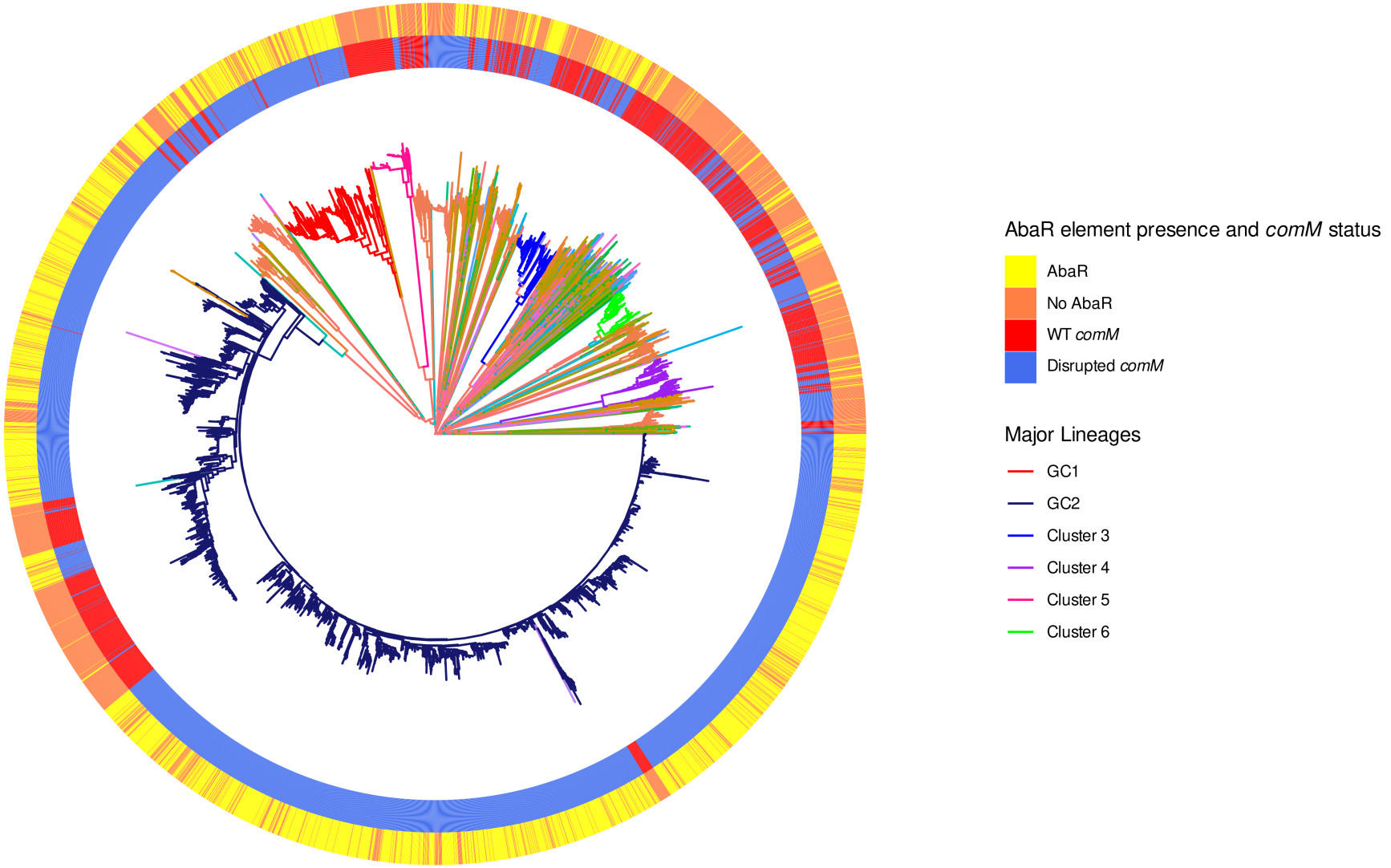

Figure 5.2: Population structure of *Acinetobacter baumannii* collection. A phylogeny formed for the 8,270 isolate *A. baumannii* collection. The phylogeny is a neighbour-joining tree created from the core distances output by PopPUNK. Branches are coloured by the PopPUNK strain for the isolate, with 403 separate strains. The major strains which are investigated further are highlighted in the key. The inner annotation ring represents the status of the *comM* gene for isolates. The outer annotation ring represents the detected presence of an AbaR element in each isolate.

The large GC2 strain, within the NJ tree formed from the core distances calculated by PopPUNK, is paraphyletic. The clade containing paraphyletic GC2 also contains five separate diverged GC2-like strains, four of which constitute single isolates, appearing sporadically among the GC2 isolates (Figure 5.2). The core distances between the GC2 isolates and the GC2-like strain 39, the only one of five strains to contain more than isolate with eight, were much larger than any distances within a strain that shares the MRCA of the GC2 strain (Figure 5.3). The accessory distances follow a similar pattern, although the range of within strain diversity of GC2 clustered isolates appears to match the range of between strain distances. Hence strain 39 appears to be distinct and has emerged from a GC2-like progenitor.

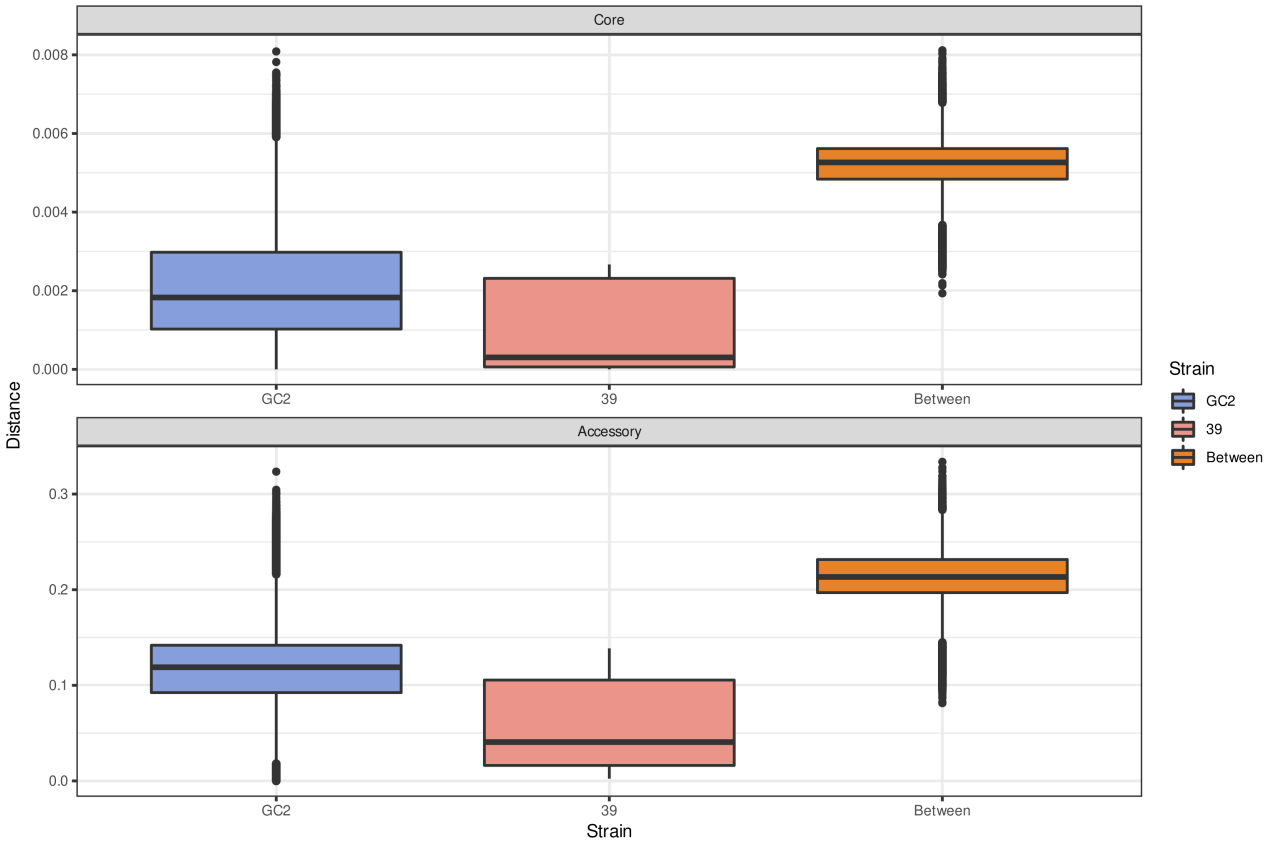

Figure 5.3: Distribution of the within strain and between strain distances for isolates within the *A. baumannii* GC2 clade. Boxplots represent the distribution of the pairwise distances between isolates of the same strain and between isolates of different strains.

5.3.2 Population struction of the *Legionella pnuemophila* collection

To investigate the effect of the pLPL plasmid, with its transformation blocking sRNA RocRp, on recombination dynamics, a collection of 3,373 *L. pneumophila* assemblies

were curated from the NCBI GenBank database. Of these 3,373 assemblies, a total of 3,116 passed QC and were analysed. As with the *A. baumannii* collection, PopPUNK v2.4.5 [298] was used to determines the population structure in the collection. The initial distance plots again revealed large accessory diversity within closely related isolates (Figure 5.1B). The between strain distances however, were not as uniformly distributed as those for *A. baumannii* (Figure 5.1B). There are two modal peaks in the pairwise comparisons for isolates with > 0.01 core distances. This could be indicative of a deeper strain structure within the population, where there is an ancestral split followed by two long branches leading to divergent populations. This is consistent with the NJ phylogeny formed from the core distances between isolates (Figure 5.4). The population can be split in two, with 2,669 isolates in the larger subtype 1 clade, and 447 in the smaller subtype 2 clade. The distances between strains within a subtype and between these two subtypes also closely follows the distance plot produced (Figure 5.5). Here the median core distance for between strain comparisons in subtype 2 is similar to the smaller modal peak in distances at 0.017. The between subtype and within subtype 1 comparisons align more closer with the larger modal peak in distances at 0.024 core distance. This is evidence for a hierarchial clustering of the *L. pneumophila* population.

PopPUNK further stratifies the population into 179 separate strains, with 93 single isolate strains and the largest strain being 834 isolates in size (Figure 5.6). The two largest strains are of similar size, with strain 1 containing 834 isolates and strain 2 820. The five major disease associated *L. pneumophila* STs all reside in single strains: ST1 (135 isolates) within strain 1, ST23 (4 isolates) within strain 15, ST37 (131 isolates) within strain 2, ST47 (105 isolates) within strain 4 and ST62 (13 isolates) within strain 5 [692]. However, in general the *in silico* sequence based typing (SBT) scheme is not effective. A large proportion of the collection (1406 isolates, 45%) are processed as un-typable. This is primarily driven by variation in the *mompS* loci. For instance, in total 242 isolates have at least one of the seven ST loci missing, with *mompS* missing in 240 of these isolates, whereas *flaA* and *pilE*, the only other missing loci, are only missing in one isolate each. Additionally, 153 of the un-typable isolates have more than one complete copy of *mompS*. Furthermore, for untypable isolates with at least one copy of every locus (1165 of 1406),

Figure 5.4: Core distance phylogeny of the *L. pneumophila* collection, with ancestral branch highlighted. Subtype 1 contains 2,669 isolates and Subtype 2 447 isolates.

removing the *mompS* sequences from the typing scheme reduces the number of isolates with an unknown type to 288. For the other six loci in the typing scheme (*flaA*, *pilE*, *asd*, *mip*, *proA* & *neuA*), removing each of these and leaving the remaining six loci for typing always produced at least 1102 unknown type isolates. The problem of *mompS* variability in the typing of *L. pneumophila* has been noted previously by Gordon *et al* [245].

This variability leads to a low similarity score between the PopPUNK strains and the SBT method. For the 1,710 isolates with both a PopPUNK strain and an ST, the adjusted Rand index score was 0.259. While the major disease STs tended to remain within single PopPUNK strains, often these strains contained multiple different STs. For instance, while strain 2 contains all 131 identified ST37 isolates, this is not the largest ST within the strain, with both ST30 (179 isolates) and ST36 (142 isolates) more frequently present. In general though, as for the *A. baumannii* collection, PopPUNK's use of all available

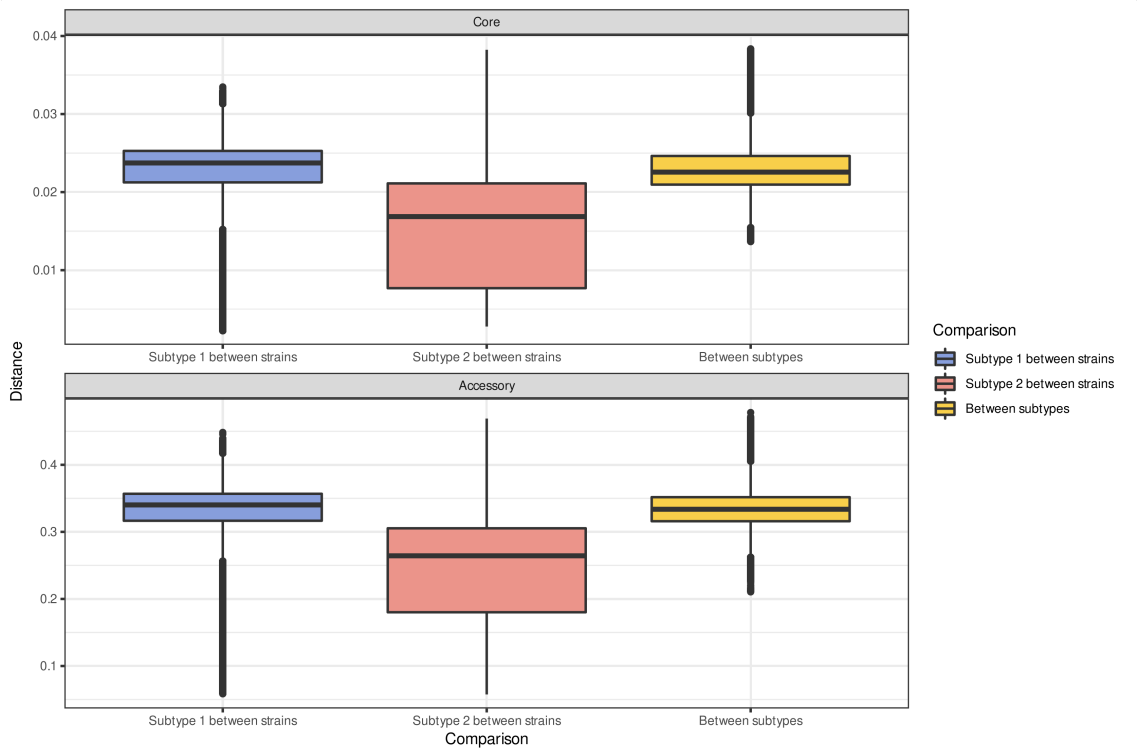

Figure 5.5: Distribution of the between strain pairwise distances based on *L. pneumophila* subtype. Boxplots represent the between strain distances of isolates within each larger subtype and then also the distances between strains in a different subtype to one another.

information within WGSs for clustering a collection, allows for the exploration of the wider collection. That PopPUNK can also separate the major disease clones into separate strains, highlights the biological feasibility of the method.

As with the *A. baumannii* collection, the largest strains appear to be paraphyletic in the *L. pneumophila* collection (Figure 5.6). This is especially the case for the strain 2 lineage, where there are 5 strains within the larger clade that is primarily associated with strain 2, including the 75 isolate strain 8 lineage. Comparisons of the distances within this 903 isolate clade reveal much smaller within strain distances than those between strains, indicative of this being a suitable clustering of isolates (Figure 5.7). The strain 8 distances are very narrow, even for the accessory distances between genomes, this is reflected in the shallow nature of the strain 8 clade within strain 2. This could be indicative of recent expansion of the lineage, or a more biased sampling strategy within an outbreak.

Now that I have described the population structure of both collections of isolates, I will move on to characterise the distribution of MGEs which can block the transformation

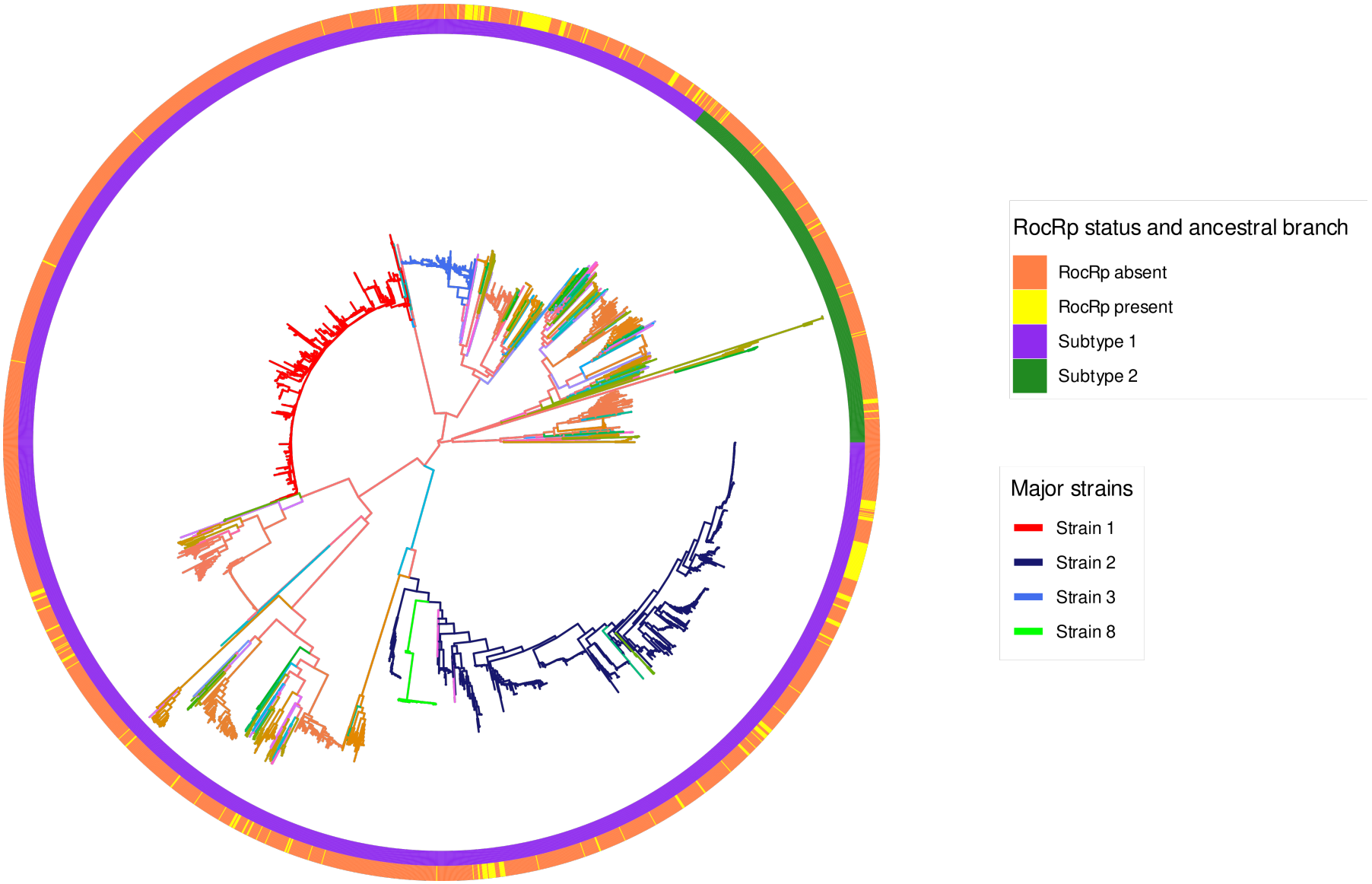

Figure 5.6: Population structure of *Legionella pneumophila* collection. A phylogeny formed for the 3,116 isolate *L. pneumophila* collection. The phylogeny is a neighbour-joining tree created from the core distances output by PopPUNK. Branches are coloured by the PopPUNK strain for the isolate, with 179 separate strains. The major strains mentioned in the text are further highlighted in the key. The inner annotation ring represents the represents the subtype the isolates form a part of, as highlighted in Figures 5.4. The outer annotation ring represents the detected presence of the RocRp sRNA in each isolate.

machinery in these two species.

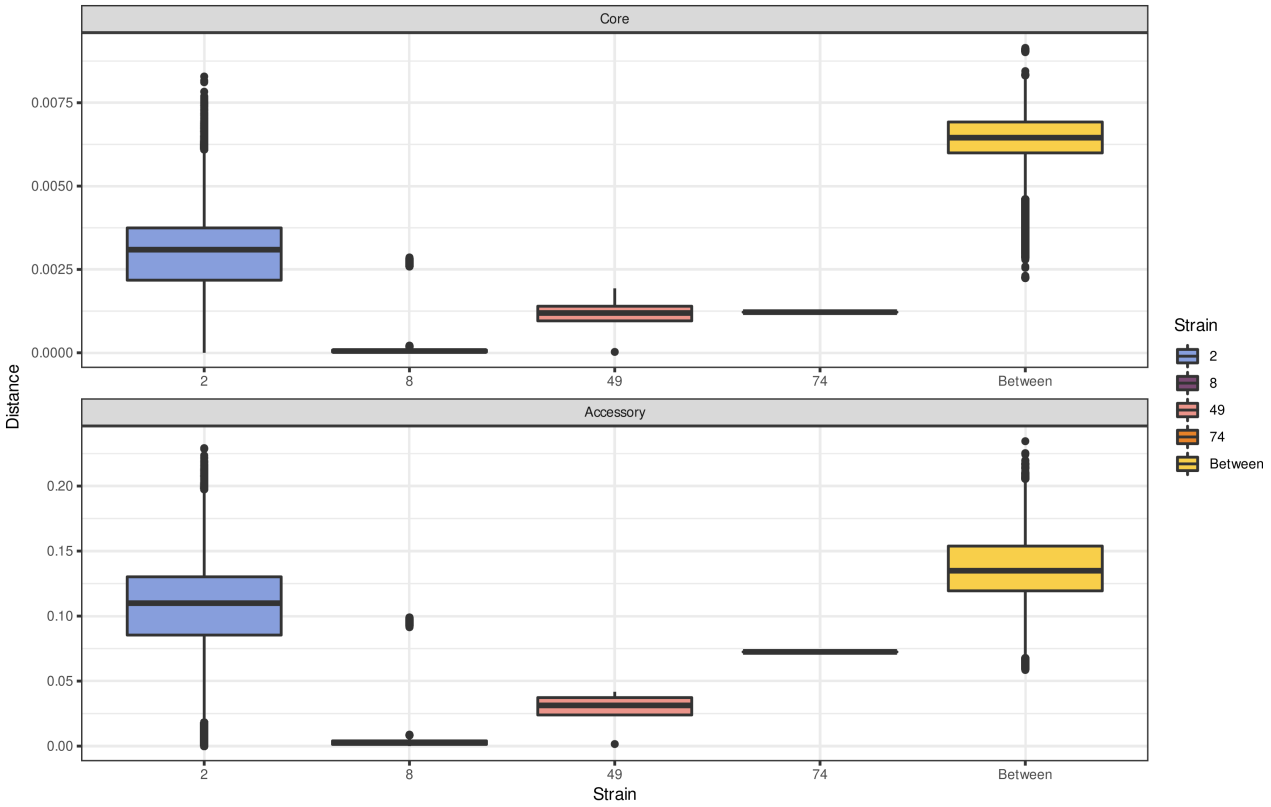

Figure 5.7: Within and between-strain distances for *L. pneumophila* strain 2 and strain 2-like strains. Boxplots represent the distribution of the pairwise distances between isolates within the same strain, strains 2 and the strain 2-like strains 8, 49 and 74, and between isolates of the different strains.

5.3.3 The distribution of MGEs across the collections

Within the *A. baumannii* collection the presence of AbaR elements was identified using the conserved left and right ends of the elements previously identified in Bi *et al* [683]. If both elements were present within an assembly an AbaR element was considered to be present within the isolate. In total this led to 4,826 of the 8,270 (58%) isolates within the collection containing an AbaR element (Figures 5.2). Both the GC1 and GC2 strains had particularly high numbers of AbaR elements. In GC2, 3,701 of the 5,092 (73%) isolates contained an AbaR, while in GC1, 376 of the 505 (74%) isolates contained some form of the element. Of the strains with > 100 isolates, strain 4 had the lowest number of isolates with AbaRs, with only 15 of the 214 isolate strain (7%) containing the element.

AbaR elements commonly insert within the competence gene *comM*. To fully account

for this method of transformation blocking by MGEs, isolates were also searched for the presence of at least one complete *comM* gene. This was found to be widely disrupted across the collection, with 6,169 isolates in total (75%) not having at least one complete *comM* locus (Figure 5.2). The levels of disruption in GC1 and GC2 were again particularly high, with 487 of 505 isolates (96%) in GC1 and 4,489 of 5092 isolates (88%) in GC2 having no complete *comM* locus. The vast majority of isolates with an AbaR element also had a disrupted *comM*, with 4,699 of the 4,826 (97%) AbaR containing isolates having a disrupted *comM*. AbaRs have been known to insert within diverse loci apart from the *comM* gene [683]. However, 1,470 isolates appear to have a disrupted *comM* gene with no AbaR detected. Of these 1,470, 955 have either the left or right conserved end of the AbaR element, but not both to fully qualify as an AbaR insertion. Here the quality of the assemblies could be affecting the ability to infer the insertion of these elements.

The *L. pneumophila* collection was searched for the presence of the sRNA RocRp, which blocks the expression of the competence machinery and is commonly found on the pLPL plasmid. RocRp was much less prevalent that the AbaR elements in the *A. baumannii* collection: only 302 of the 3116 (10%) isolates in the *L. pneumophila* collection contained the sRNA (Figure 5.6). Of the major lineages, strain 2 contained the most isolates with RocRp present, with 130 of 820 isolates (16%) found to have RocRp, strain 3 had the same proportion of 16% with RocRp (33 of 200 isolates). Strain 1 only had five isolates that contained RocRp (1%).

Given the high proportion of isolates with disrupted competence machinery in the *A. baumannii* GC2 clade, I will now move on to discuss the recombination dynamics in this important pathogenic clone.

5.3.4 Recombination dynamics in the GC2 strain of *Acinetobacter baumannii*

To assess the recombination dynamics in the largest *A. baumannii* lineage, GC2, Gubbins v3.2.0 was run with a FastTree starting phylogeny builder, a RAxML main phylogeny builder and a joint ancestral state reconstruction method. This identified very large recombination events occurring across the lineage, in particular with respect to the capsule K locus and phage regions (Figure 5.8). When removing these phage regions from the

recombination events, in order to more accurately assess the rate of sequence import through homologous recombination, the ratio of SNPs introduced by recombination compared with mutation (r/m) for this lineage is 1.48. This is relatively low, especially compared with a previous estimate of the r/m for the GC1 lineage of 22 [549]. The ratio of recombination events to substitution mutations (ρ/m) was also lower compared to the past GC1 lineage, at 0.016 for GC2 compared to a previous estimate of 0.1 for GC1.

Gubbins detected a number of very large recombination events across the lineage, particularly around the capsule K locus. For instance, there are events of up to 200 kb in length encompassing the K locus and at the base of a 4,095 isolate clade. This also covers the *parC* gene involved in fluoroquinolone resistance. When the circular nature of the genome is taken into account, this event appears to span 400 kb, an extremely large recombination event. This is similar in size to hybridisation events previously seen in *K. pneumoniae* [709]. Kaptive v0.7.3 [710] was used to assess the K locus types of the collection, which controls the synthesis and the export of the CPS for *A. baumannii* isolates. In total there were 49 different K locus strains detected in GC2 (Figure 5.8). Using PastML [590] to reconstruct the ancestral K locus types, there were an estimated 325 switches in the K locus across the phylogeny. From the recombination reconstruction it appears the large recombination blocks spanning the K locus, coincide with a shift from a largely KL22/KL13 locus in the less derived isolates, to a KL2 & KL3 capsule type among the more derived group of isolates. Indeed, when incorporating the reconstructed recombination events from Gubbins, the 200 kb event spanning 4,095 isolates coincides with a switch from the KL9 locus to KL2 locus. Selection pressures from the host immune response likely drive this loci to become a recombination hotspot as depicted here.

Prophages also appear as peaks in recombination across the genome, reflecting their frequent interchange between hots. The Phaster tool [272] and manual inspection of the reference genome was used to detect the occurrences of phage. In total three separate regions were identified as possible prophages. Two of these were identified as the lytic bacteriophage Bϕ-B1251 a member of the *Siphoviridae* family [711, 712] (Figures 5.8). The Bϕ-B1251 insertion between 1.52 Mb to 1.58 Mb within the GC2 reference is marked as a complete intact phage by Phaster and appears to recombine often, the other insertion

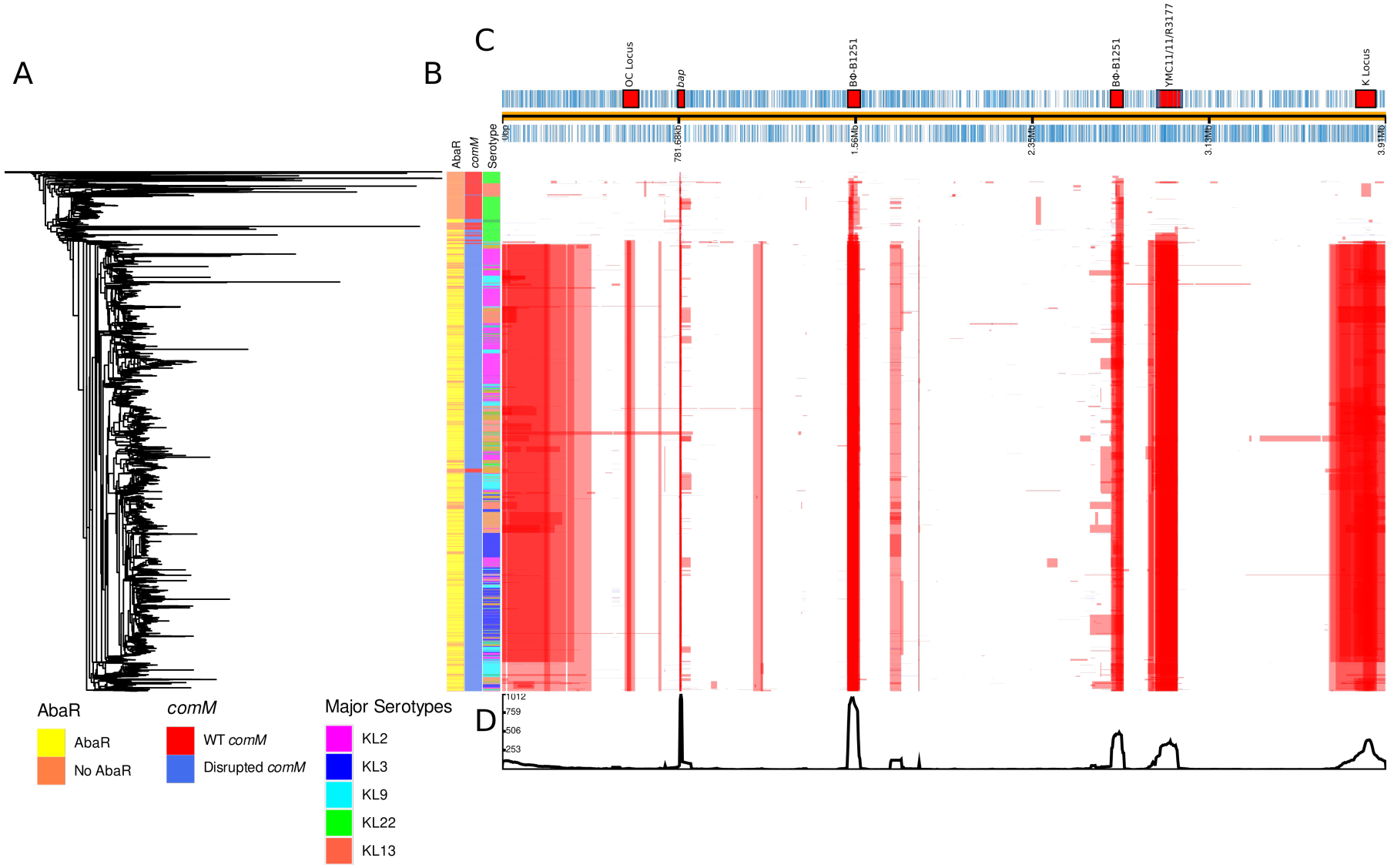

Figure 5.8: Recombination dynamics in the *A. baumannii* GC2 lineage. A The ML phylogeny produced by Gubbins of the non-recombinant regions of the GC2 alignment. There are a total of 5,092 isolates in the phylogeny. **B** Bars describing the AbaR presence, *comM* status and serotype of individual isolates. Bars map across to individual isolates. For the serotype, 49 different loci are coloured, for ease of viewing only the five most common serotypes are highlighted in the key. **C** Simplified annotation genome of the reference strain XH856 (ENA accession number SRR2119615). The red bars represent loci with hotspots of recombination. Blue bars represent individual genes. **D** Distribution of recombination events across the GC2 lineage. In the upper half of the graph, red bars indicate recombination events occurring on internal nodes in the tree, which were subsequently inherited by multiple descendent isolates. These bars map across to isolates in the phylogeny in section **A** and map to regions in the genome annotated in section **C**. Blue bars indicate recombination events on terminal branches of the tree, occurring in only one isolate. In the bottom half of the graph, the line represents the frequency of recombination events along the genome's length.

at 2.59 Mb to 2.75 Mb is marked as a questionable insertion. The other phage inserted is YMC11/11/R3177, also a member of the *Siphoviridae* viruses [711, 713]. This insertion is only classified as an incomplete phage by Phaster. This region also contains two ESBL genes, both within the $bla_{OXA\text{-}23}$ family of genes. This region could therefore also represent a larger ICE element. Although, the ICEfinder tool, which uses the ICEberg2 database [714], found no putative ICE within the reference genome.

Another hotspot for recombination in the GC2 collections was a large 15 kb gene, annotated as a hypothetical protein by Prokka v1.14.6 [588] (Figure 5.8). Further searching revealed this to be the biofilm-associated protein *bap* gene. In total there were 1,601 recombination events spanning some portion of this locus. The majority of these events (1496 of 1601) spanned only the first 10 kb of this gene. Running the gene sequence through InterProScan [715] found an Immunoglobulin-like (Ig-like) fold from AA 6 to 3473, corresponding to an end base of 10,419. These motifs are commonly found in bacterial adhesins involved in host tissue colonization [716]. The *bap* gene is necessary for both biofilm formation and the binding to host epithelial cells [717]. Hence the high levels of recombination, particularly around the Ig-like fold region, could be driven by selection pressure from the host immune response. The *bap* gene is also a recombination hotspot in the GC1 lineages. In this case the gene is 25.9 kb long, also with a putative Ig-like domain detected by InterProScan as the primary region affected by recombination.

Now that I have described the general recombination dynamics in the large GC2 lineage, I will look at how the disruption of *comM* affects recombination dynamics, both in GC2 and in the six largest strains detected in the *A. baumannii* collection.

5.3.5 The effect of *comM* disruption on recombination dynamics in *Acinetobacter baumannii*

The distribution of disrupted *comM* varied across the six largest strains within the *A. baumannii* collection (Table 5.2). While, as noted above, GC1 & GC2 had high numbers of isolates without a complete *comM*, strains 4 & 5 had the opposite with by far the majority of isolates containing a complete *comM*. To assess the effect of these disruptions on overall recombination dynamics, Gubbins v3.2.0, with a FastTree initial phylogeny builder, a RAxML main iteration builder and a joint ancestral state reconstruction, was run on these

six strains. PastML was then run on the resulting phylogenies to reconstruct the ancestral *comM* state for each branch, in order to classify recombination events as occuring with a complete or disrupted *comM*. Recombination events spanning putative prophage regions, detected using Phaster and manual inspection, were also removed from further analysis.

Strain	WT *comM*	Disrupted *comM*	Total isolates
GC1	18 (4%)	487 (96%)	505
GC2	603 (12%)	4489 (88%)	5092
Strain 3	137 (56%)	106 (44%)	243
Strain 4	190 (89%)	24 (11%)	214
Strain 5	167 (89%)	21 (11%)	188
Strain 6	117 (66%)	59 (34%)	176

Table 5.2: Numbers of *A. baumannii* isolates split by *comM* status across the six largest strains

When looking at the length distribution of recombination events, it appears the disruption of *comM* has no effect on the length of sequence imported (Figure 5.9). Unexpectedly, in the GC1 strain the median length of import is higher for those with a disrupted *comM*, 33,552 bp compared to 15,737 for isolates with a WT *comM*. Although, these differences are not statistically significant. In terms of the SNP density of recombination events, however, there were some significant differences, although these were not consistent across the strains (Figure 5.10). For the GC2 strain, the WT *comM* isolates import sequence with a significantly higher SNP density (Mann-Whitney U test; U = 529088, n1 = 412, n2 = 2908, two-sided p = 1.22×10^{-4}), with a median SNP density of 3.6×10^{-3} SNPs/bp, whereas those with a WT *comM* have a median density of 3.0×10^{-3} SNPs/bp. Similarly in strain 3, WT *comM* isolates import sequences with a significantly higher SNP density (Mann-Whitney U test; U = 29659, n1 = 438, n2 = 156, two-sided p = 0.01439), with a median density of 5.9×10^{-3} SNPs/bp compared to isolates with a disrupted *comM* importing sequence with a SNP density of 5.1×10^{-3} SNPs/bp. However, in the other strains there are no significant differences in terms of the SNP density of imports. These inconsistent results suggested that *comM* disruption had a minimal effect on the length of divergent sequence imported through transformation and recombination. In some strains though , there is evidence of a marginal effect on the ability to acquire highly divergent sequence through transformation.

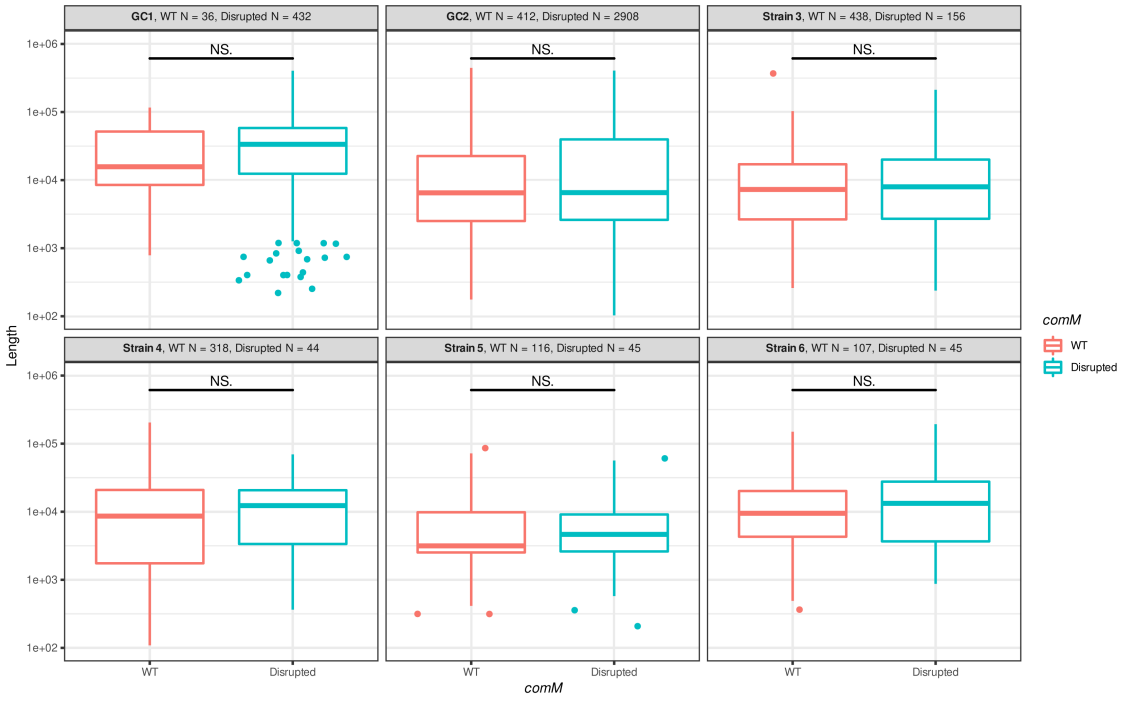

Figure 5.9: Distributions of the lengths of recombination events, split by *comM* status across the six largest *A. baumanii* strains. Within the title for each plot, N represents the number of recombination events the boxplot describes. Recombinations spanning putative prophage regions have been removed. Significance tests have been performed using the Mann-Whitney test.

When looking at the key recombination statistics r/m and ρ/m I also observed a mixed picture in terms of the effect of *comM* disruption (Figure 5.11). Statistics were calculated through bootstrap sampling (n = 100,000) of the branches on the phylogenies whose *comM* state had been reconstructed through PastML. The difference between the r/m values calculated for each *comM* state was then plotted, with the 95% interval of the differences compared to zero, to test for significance. For the majority of clades, GC1 & strains 4 to 6, *comM* disruption had no significant effect on the r/m values. For strain 5 it does appear that the r/m for isolates with a WT *comM* was significantly higher, with a median difference of 1.55 (95% interval 0.10 to 2.59). For the GC2 lineage however, it appears that isolates with a disrupted *comM* actually have a higher r/m, with WT isolates having a median difference to disrupted *comM* isolates of -2.14 (95% interval -2.54 to -1.77). This is odd, considering previous lab work has shown those isolates with an AbaR element inserted into *comM* have up to 10^3 less transformation frequency that WT *comM* isolates [667].

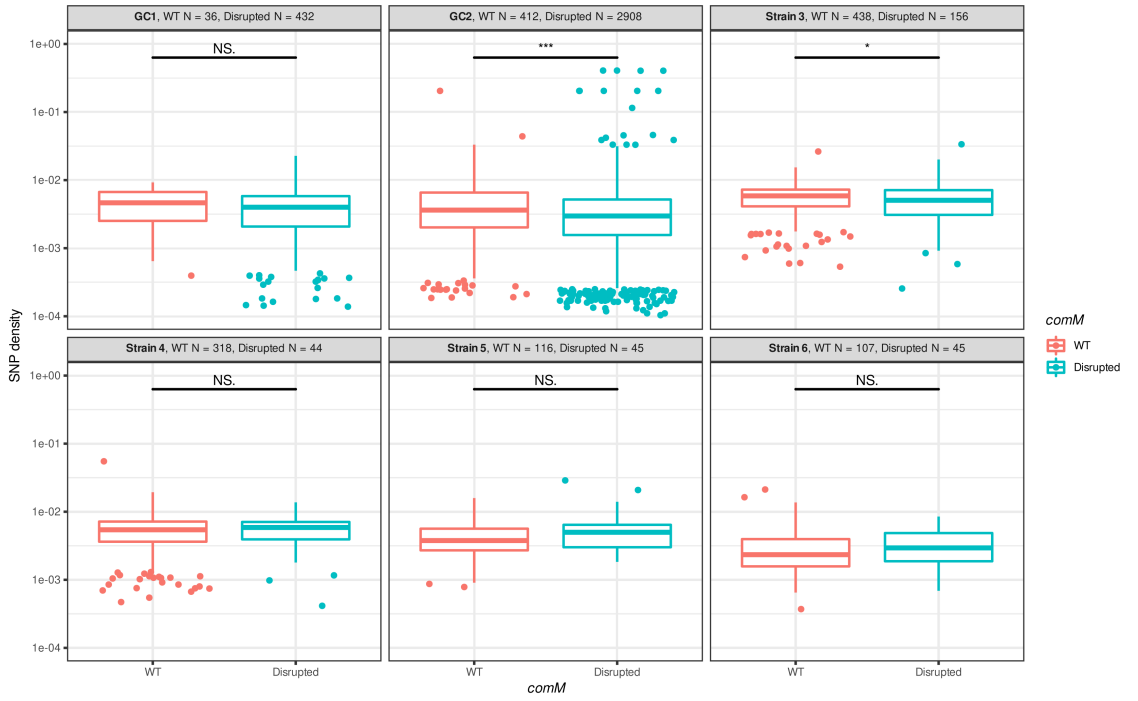

Figure 5.10: Distributions of the SNP density of recombination events, split by *comM* status across the six largest *A. baumanii* strains. Within the title for each plot, N represents the number of recombination events each boxplot describes. Recombinations spanning putative prophage regions have been removed. Significance tests have been performed using the Mann-Whitney test.

A similar trend appears for the ρ/m metric too (Figure 5.12). Again half the strains appear to have their ρ/m values unaffected by *comM* disruption, GC1 & strains 4 & 6. For ρ/m however, strain 3, along with strain 5, does appear to have a slighter higher ρ/m when having a complete *comM* (median = +0.026, 95% interval 0.0033 to 0.0495). GC2 again however has a lower ρ/m for isolates with a complete *comM* compared to those with a disrupted *comM*. The median difference was -2.144 (95% interval -2.54 to -1.77), which indicated that disrupted *comM* actually recombine more frequently, again this is against expectations from previous lab results.

Now that I have gone through the recombination dynamics of the *A. baumannii* collection, I will move on to the *L. pneumophila* collection. Starting with the recombination results from the largest strain, strain 1 which corresponds to the prominent disease-associated clone ST1.

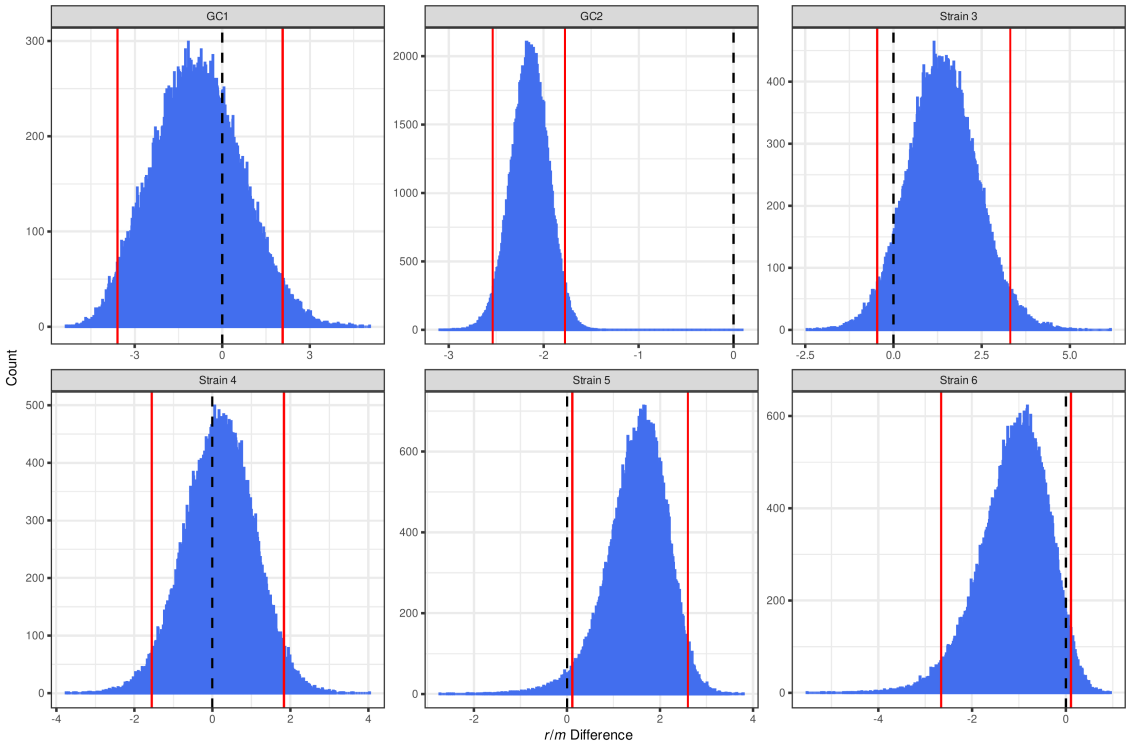

Figure 5.11: Distributions of the differences between *A. baumannii* WT*comM* isolates and disrupted *comM* isolates *r*/*m* values. Histograms represent the differences between WT *comM* isolates and disrupted *comM* isolates *r*/*m*. The histogram is calculated from 100,000 bootstrap samples of the branches of each strain's phylogeny with *r*/*m* calculated from the subset branches for both *comM* states. The red vertical lines represent the region where 95% of the difference values lie. The black dashed line represent zero difference between the *r*/*m* values.

5.3.6 Recombination dynamics in *L. pneumophila* strain 1.

The recombination dynamics for the largest *L. pneumophila* strain, strain 1, which contains the ST1 clone, were analysed using Gubbins v3.2.0. This was run with the same FastTree, RAxML joint ancestral reconstruction model choice as previously used for the *A. baumannii* lineages. Overall there were fewer large recombination events in this lineage compared to GC2, with most of these larger events occuring around prophage regions (Figure 5.13). However, the *r*/*m* value for this lineage was much higher than that for the *A. baumannii* lineages. Gubbins estimated the *r*/*m* at 8.91 when excluding prophage regions. Similarly the ρ/m was higher at 0.05 recombination events per point mutation, with a total of 543 events outside of prophage regions.

Genes involved in the production of the lipopolysaccharide capsule (LPS), were frequent targets of recombination (Figure 5.13). In total 76 recombination events span a

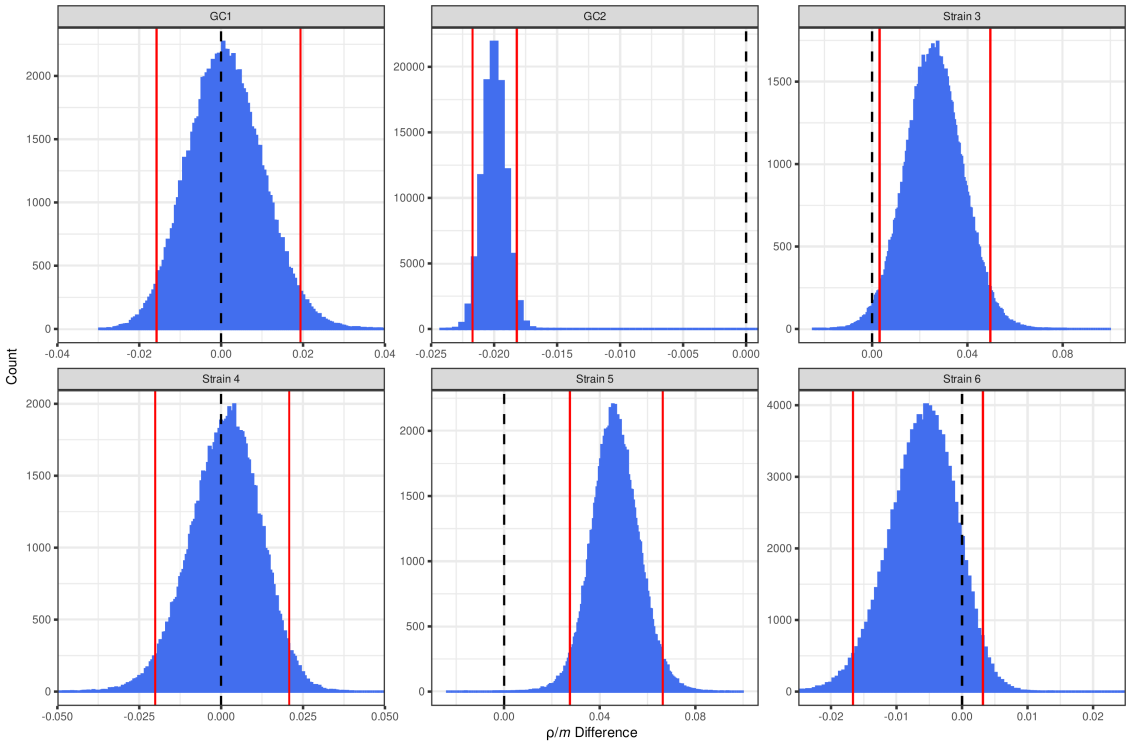

Figure 5.12: Distributions of the differences between *A. baumannii* WT*comM* isolates and disrupted *comM* isolates ρ/m values. Histograms represent the differences between WT *comM* isolates and disrupted *comM* isolates ρ/m. Histograms are calculated from 100,000 bootstrap samples of the branches of each strain's phylogeny with ρ/m calculated from the subset branches for both *comM* states. The red vertical lines represent the region where 95% of the difference values lie. The black dashed line represent zero difference between the ρ/m values.

part of this LPS 35 kb locus. Furthermore, within the LPS locus the *lag-1* gene was a particular hotspot for recombination, with 42 events spanning this locus. This gene has previously been identified as having a strong association with invasive clinical isolates, aiding cells in avoiding complement-mediated killing [718]. There is also a peak in recombination around the *mutS* gene, which is involved in DNA mismatch repair. Other genes close to this locus include *recA*, which binds to ssDNA facillitating strand invasion during homologous recombination, and *recX*, which acts as a regulator of *recA*. Why recombination would peak around these loci is unclear. Previously an MGE has targetted *mutS* in marine *Vibrio splendidus* to create a hypermutator phenotype [655]. However, no MGEs appear to insert around this locus within the *L. pneumophila* collection.

MGE regions are also hotspots for recombination in this lineage (Figure 5.13). There is a probable ICE element within the reference sequence, from 2.25 Mb to 2.32 Mb. This region contains the *virB4* gene encoding a component of the type IV secretion system

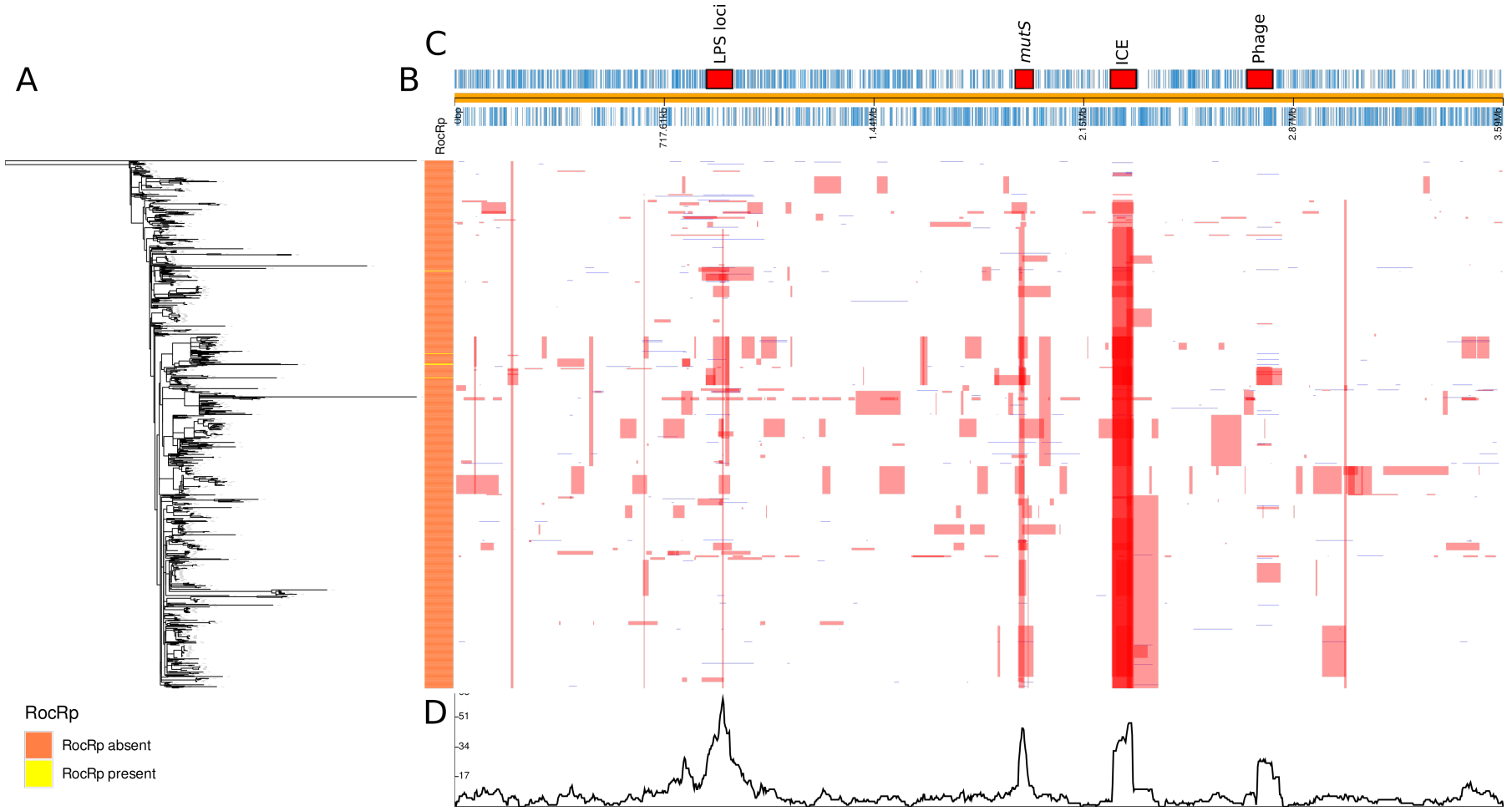

Figure 5.13: Recombination dynamics in strain 1 of *L. pneumophila*. **A** The ML phylogeny produced by Gubbins from the non-recombinant regions of the strain 1 alignment. There are a total of 834 isolates in the phylogeny. **B** Bars describing the RocRp presence of individual isolates. Bars map across to individual isolates. **C** Simplified annotation genome of the reference strain Flint 2 (D-7477) (ENA accession: CP021281). The red bars represent loci with hotspots of recombination. Blue bars represent individual genes. **D** Distribution of recombination events across the GC2 lineage. In the upper half of the graph, red bars indicate recombination events occurring on internal nodes in the tree, which were subsequently inherited by multiple descendent isolates. These bars map across to isolates in the phylogeny in section **A** and map to regions in the genome annotated in section **C**. Blue bars indicate recombination events on terminal branches of the tree, occurring in only one isolate. In the bottom half of the graph, the line represents the frequency of recombination events along the genome's length.

used in conjugation, as well as the *tra* operon involved in DNA transfer [719, 720]. It also contains a suite of heavy metal resistance efflux pumps, such as *cusA* and *czcA*, which are part of the resistance-nodulation-division (RND) drug resistance efflux pumps [721]. A phage region from 2.74 Mb to 2.84 Mb was also identified by Phaster. This has only been identified as an incomplete phage, with its closest match to the ST147-VIM1ϕ7.1 *Klebsiella pneumoniae* phage [722]. Both these MGE regions have been removed from the overall *r*/*m* analysis due to their ability to move between genomes via mechanims other than homologous recombination.

Now that I have described the overall recombination dynamics for the strain 1, I will look into the effect of RocRp on the recombination dynamics in the three largest strains of *L. pneumophila*.

5.3.7 RocRp disrupts recombination dynamics in *L. pneumophila*

RocRp is not as widely distributed as AbaR elements and *comM* disruptions are in the *A. baumannii* collection (Table 5.3). Strains 2 & 3 have similar proportions of RocRp present, whereas strain 1 has almost no isolates with RocRp present. As with the *comM* disruptions in *A. baumannii*, Gubbins v3.2.0 with a FastTree, RAxML and joint ancestral state reconstruction model was run to assess the impact of RocRp on recombination dynamics in these three lineages. PastML was also run to assess the likely ancestral state of RocRp presence and recombination events spanning known MGE regions were excluded from further analysis.

Strain	WT	RocRp present	Total isolates
Strain 1	829 (99%)	5 (1%)	834
Strain 2	690 (84%)	130 (16%)	820
Strain 3	167 (84%)	33 (16%)	200

Table 5.3: Distribution of the transformation inhibiting sRNA RocRp across the three largest *L. pneumophila* strains.

In terms of the length of recombination events, RocRp presence does not appear to affect the length of sequence imported (Figure 5.14). For strain 1, there is only one recombination event along a RocRp branch that is included in the analysis, so comparisons are of statistical power. For strains 2 & 3, the median length is similar for isolates with

RocRp (46,245 bp for strain 2 and 18,160 for strain 3) to those WT isolates (40,392 for strain 2 and 18,160 for strain 3). In terms of the SNP density of recombinations too, RocRp appeared to have a limited effect (Figure 5.15). While in this case WT isolates in strains 2 & 3 had a higher median SNP density (6.1x10^{-3} for strain 2 and 3.4x10^{-3} for strain 3) compared to isolates with RocRp (4.7x10^{-3} for strain 2 and 2.9x10^{-3} for strain 3). The difference, however, was not significant.

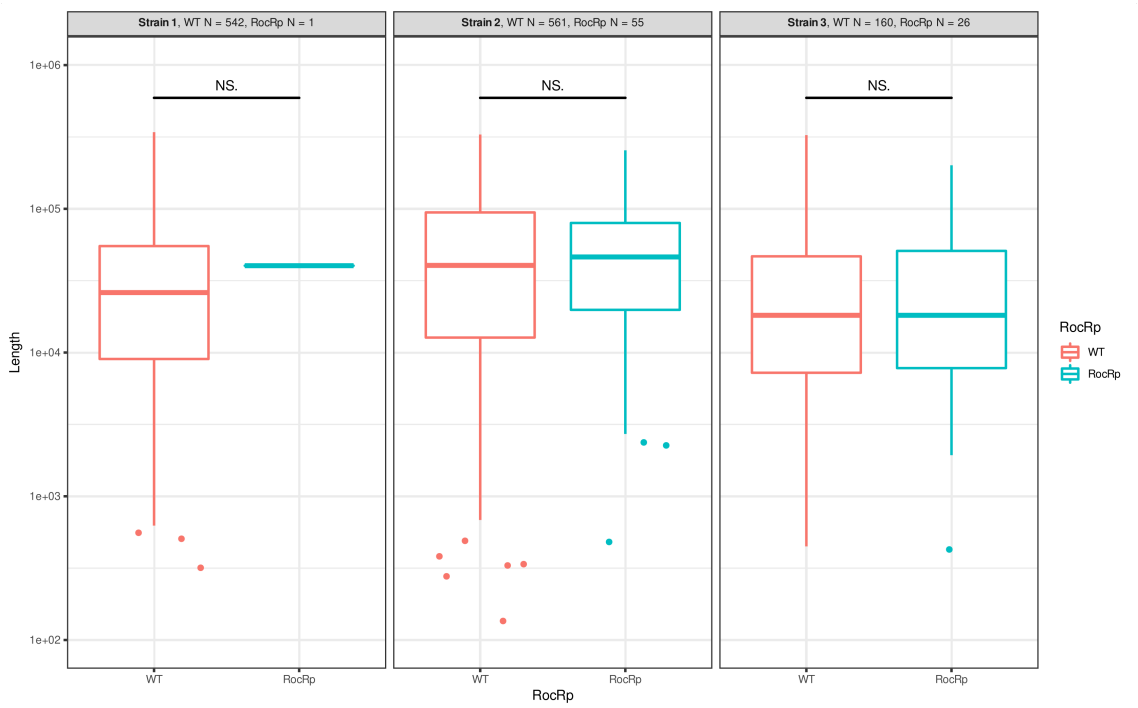

Figure 5.14: Distribution of the lengths of recombination events, split by RocRp possession across the three largest *L. pneumophila* strains. Within each plot, the title N values represent the number of recombination events each boxplots represents the distribution for. Significance tests were performed using the Mann-Whitney U test.

For the key recombination metrics, r/m and ρ/m, RocRp does however appear to have an effect. These statistics were again calculated from a subset of branches, split by RocRp presence, chosen through Bootstrap sampling (n = 100,000). . For strains 1 & 2 the r/m value appears to be significantly higher for WT isolates over isolates found to contain RocRp (Figure 5.16). For strain 1 there was a median increase in r/m of 8.75 for isolates without RocRp compared to those with RocRp (95% interval 5.86 to 11.45). While for strain 2 this increase was even higher, at a median increase of 12.55 in r/m (95% interval 1.73 to 22.11). For strain 3, while the difference in r/m between the RocRp and WT isolates was not significant, the median value was still positive with an increase

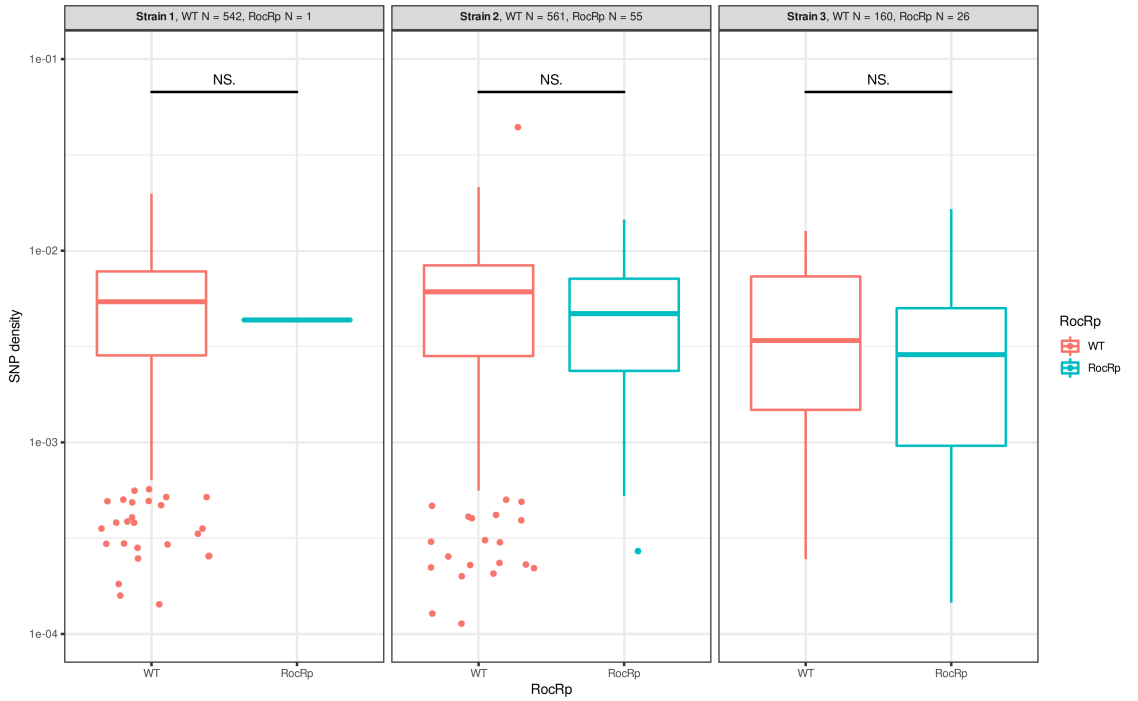

Figure 5.15: Distributions of the SNP density of recombination events, split by RocRp presence across the three largest *L. pneumophila* strains. Within each plot, the title N values represent the number of recombination events each boxplots represents the distribution for. Significance tests were performed using the Mann-Whitney U test.

of 16.98 (95% interval -13.44 to 43.91). Across the three strains, RocRp presence lead to a median reduction in r/m of 58.4%.

RocRp was also seen to lower the ρ/m metrics for these lineages (Figure 5.17). Again for strains 1 and 2 this difference was significant. For strain 1 WT isolates had a ρ/m value on average a higher ρ/m value by 0.051 more recombination events per point mutation SNP than RocRp isolates (95% interval 0.035 to 0.065). For strain 2 this time the increase was slightly lower, with an average increase in ρ/m of 0.039 (95% interval 0.002 to 0.071). For strain 3 the difference in ρ/m was not significant between WT and RocRp isolates, although the median difference was again greater than zero (median = 0.074, 95% interval -0.126 to 0.219). From these metrics then, we can see that RocRp, unlike *comM* disruption in *A. baumannii*, does appear to negatively impact the frequency of recombination in large collections of *L. pneumophila*. Across the three strains, RocRp presence was associated with a 47.4% decrease in ρ/m

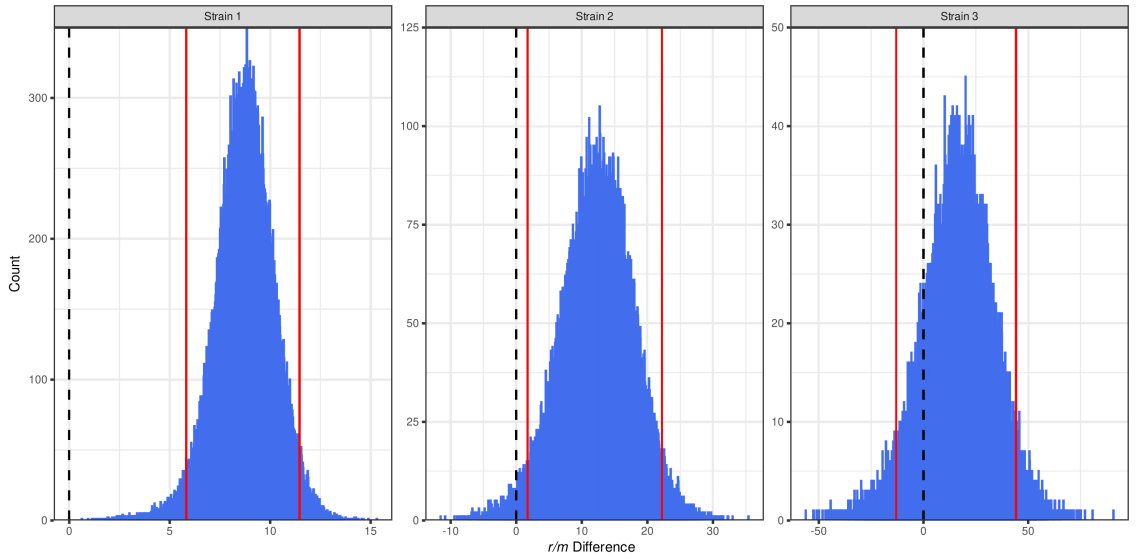

Figure 5.16: Distributions of the differences between *L. pneumophila* WT and RocRp present isolates' *r*/*m* values. Histograms represent the differences between WT and RocRp isolates' *r*/*m*. The difference is calculated from 100,000 bootstrap samples of the branches of each strain's phylogeny with *r*/*m* calculated from the subset branches for both RocRp states. The red vertical lines represent the region where 95% of the difference values lie. The black dashed line represent 0 difference.

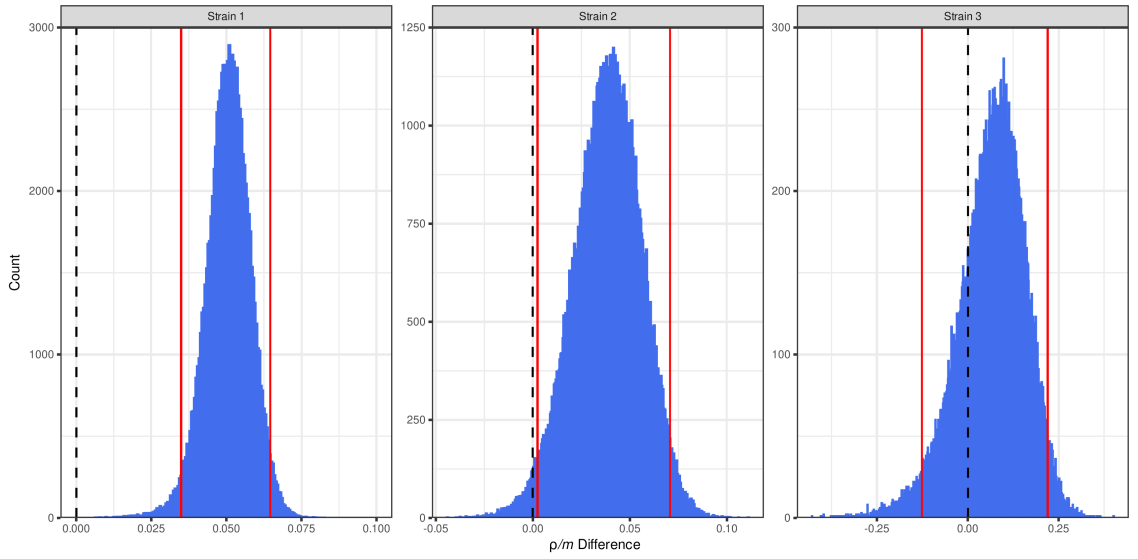

Figure 5.17: Distributions of the differences between WT and RocRp present isolates' ρ/*m* values. Histograms represent the differences between WT and RocRp isolates' ρ/*m*. The difference is calculated from 100,000 bootstrap samples of the branches of each strain's phylogeny with ρ/*m* calculated from the subset branches for both RocRp states. The red vertical lines represent the region where 95% of the difference values lie. The black dashed line represent 0 difference.

5.4 Conclusions

In this chapter I have focused on two important gram-negative pathogenic species, using genomic epidemiological techniques to investigate their recombination dynamics. Across

both species, PopPUNK was far more consistent than MLST schemes in clustering isolates into strains. This difference was most pronounced in the *L. pneumophila* collections, where assembly based SBT methods struggled to classify isolates. *L. pneumophila* isolates often carry two copies of the *mompS* used as one of the seven loci in the typing scheme. To account for this variability, previous typing schemes have leveraged data from raw reads as well as assemblies, showing an increase in the accuracy of typing compared to an assembly only approach [245]. With the advent of LRS technologoies, which promise both an increase in data availability and assembly quality, typing based on WGS assembly data promises to be more robust and consistent.

We have seen how population level recombination dynamics can be affected by the presence of MGEs within isolates. For *A. baumannii*, despite evidence of AbaR insertion and *comM* disruption leading to a reduction in transformation efficiency *in vitro* [667], at a population level we observe that there is no consistent significant effect. In GC2, it even appears that *comM* disruption is associated with more recombinogenic populations. The remarkable spread of the AbaR elements in this lineage, present in 73% of isolates, may be driven by the resistance cargo genes these elements often carry [657, 679, 687]. While recombination may be less effective in these isolates, their success provides a large enough population for recombination at key immunogenic loci to occur and subsequently be selected for.

In *L. pneumophila* however, these transformation restricting elements are much less widespread. Unlike AbaR elements, the pLPL plasmid is not associated with any resistance genes. Furthermore, estimates of the effect of RocRp on transformation frequency are much higher than AbaR insertion, at a reduction of 10^3 compared to AbaR insertion at *comM* which has been estimated as low as a 10x reduction in transformation frequency [657, 668]. These factors combine to make the effect of RocRp on a population much more apparent, with all lineages having a reduction in r/m and ρ/m comparing to WT isolates. Why exactly a plasmid would block transformation is less clear. Plasmids are thought to be closely related to ICE elements, which can insert within a host chromosome and as such could avoid deletion by disrupting the transformation machinery [350, 723]. Perhaps in this case an ICE encoding RocRp has converted into a plasmid. Further work

is needed to explore the diversity of elements which RocRp is found in.

Both species shared loci with similar peaks in recombination events. The capsule locus for instance, as has been observed in other species [142, 724–726], is a hotspot for recombination. Phage and other MGE regions appear frequently too, likely as a result of their ability to excise and move between hosts. For *A. baumannii* the *r*/*m* for lineages was lower than predicted in the past. Previous estimates of *r*/*m* for GC1 using Gubbins have recorded values of 22 [549], while in this collection I calculate it at 4.42. In the original GC1 study by Holt *et al* [549], an earlier version of Gubbins was used, while the smaller number of isolates were selected to be representative of the diversity of GC1.This may have introduced a larger number of SNPs than present in my collection of all sequenced isolates, hence increasing the *r*/*m*.

This concludes the results sections of this thesis. I will now move on to discuss these results further, detailing the limitations of this work and how future research may be focused.

Chapter 6

Discussion

6.1 Summary of results

The interaction between host bacteria and selfish MGEs has driven a range of important evolutionary processes. From the emergence of bacterial defence systems like R-M and CRISPR-Cas, to the spread of clinically important resistance and virulence genes. All are driven by this fundamental conflict between MGEs and hosts. How this conflict drives recombination dynamics in pathogenic species, and how in turn these dynamics can affect the spread of AMR among these species, has been the focus of this thesis.

The first step in assessing these dynamics is detecting recombination events between isolates. This is not a trivial problem [506]. The work I have presented here includes an update to the popular recombination detection method Gubbins. This has included a thorough benchmarking on simulated data, assessing a range of potential models used in the detection of recombination, and the first in-depth comparison of the tool to the ClonalFrameML method. From these simulated datatsets we have seen that Gubbins is the most accurate method in reconstructing the topology of the phylogeny among sequences, across a range of recombination and branching rates. Furthermore, I have highlighted the optimal set of model parameters for accurate phylogenetic reconstruction: a joint ancestral state reconstruction with a FastTree and RAxML phylogeny builder working with a GTR substitution model. Finally, the run-time and memory efficiency of Gubbins has been greatly improved, both over previous iterations and the ClonalFrameML method. This en-

sures that Gubbins can continue to be a useful tool with the ever increasing amount of sequence data available moving forward.

The evolutionary processes driving the spread of AMR and MDR lineages remain only partially realized. While the question of how ICEs move between pneumococci is also open to debate. To answer these questions, in Chapters 3 and 4 Gubbins was applied to understand the recombination dynamics underlying the spread of two globally distributed MDR lineages of pneumococci, PMEN3 and PMEN9. I found that resistance evolved repeatedly within these lineages, however, only in select cases did this lead to the expansion of the clade. The ST156 clade in PMEN3, for instance, expanded from the early 1980s with the gain of penicillin resistance via recombination at the *pbp* loci. In Germany too, a macrolide resistant but penicillin sensitive clade expanded in line with the increase in macrolide to penicillin consumption in the late 1990s. Both these clades gained resistance through recombination with non-pneumococcal DNA.

The German clade expansion was also an example of the conflict between MGEs and their host cell. Tn*1207.1*, the transposon conferring macrolide resistance, disrupted the competence gene *comEC* preventing further recombination. While this led to the initial spread of the lineage within Germany, the lack of recombination prevented both serotype switching and the adaptation of this isolate to the PCV7 vaccine, introduced in 2006 to Germany. The scale of MGE insertion site diversity was elucidated through searching for Tn*1207.1*-type and Tn*916*-type elements in the wider GPS collection. Tn*916*-type elements in particular were widespread across the 20,000 isolate collection, with over 100 different sites targeted in the pneumococccal genome. Both these elements were also found to move between cells via recombination, with frequent interspecies spread too. This offers a new perspective on how these important AMR genes can disseminate within a population.

Finally the effect of host MGE conflict was also investigated at a population level, in the important pathogens *A. baumannii* and *L. pneumophila*. MGEs that had disrupted competence machinery, the AbaR resistance element and the *L. pneumophila* plasmid encoded RocRp sRNA, were identified across large public collections of assembled genomes. These results show that the recent PopPUNK tool for genomic epidemiology can effec-

tively cluster the populations into biologically plausible strains, outperforming traditional MLST methods. Despite laboratory studies finding a reduction in transformation efficiency from *comM* disruption by AbaRs, this is not observed in the pattern of exchange between strains inferred with Gubbins. However, recombination at key immunogenic loci, such as the capsule locus, still occurred in the *A. baumannii* population. In *L. pneumophila* on the other hand, there does appear to be a significant reduction in the rate of exchange between lineages caused by RocRp . The explanation for why a extra-chromosomal plasmid would cause this, though is still unclear.

6.2 Implications of research

6.2.1 Selection and recombination

Gubbins, due to its approach of detecting recombinations via an elevated SNP density, will only find recombination events that introduce diversity into a strain. However, the majority of recombination events likely occur between isogenic cells that grow in close proximity to one another [350]. These events would leave no signature in the host genome, and the frequency of their occurrence could likely only be derived through extensive laboratory work. However, if the diversifying recombinations that Gubbins detects were selectively neutral we might expect that they would be randomly distributed across the genome, with no particular locus having abundant recombination events. Instead, what we observe are hotspots of recombination across the three species that I investigate. This is in line with previous studies looking at recombination dynamics, where there were hotspots around loci associated with pathogenicity and resistance [142, 549, 699, 727, 728]. Selection then must be driving the spread of these divergent recombination events.

6.2.2 Adaptive evolution to public health interventions

Selection may be preserving these diversifying recombination as they are adaptive to clinical interventions. For instance, building on previous work in the pneumococcus, and other species, [318, 560] I observe interspecies recombination events around the *pbp* loci leading to gains of resistance across both the pneumococcal MDR lineages and in the wider GPS collection. Typically, these previously identified transformation events generating mosaic *pbp* and *murM* gene structures were short, and confined to few, specific

loci [729, 730]. My results though, extend the importance of interspecies transformation to show its role in importing long stretches of DNA. These events can cause structural variation through the integration of antibiotic resistance cassettes at many sites around the pneumococcal chromosome. Such transformations are atypical in two regards. Firstly, in the number of SNPs they introduce into the recipient, as the efficiency of exchanges decreases exponentially with sequence divergence [731]. Secondly, insertion of large loci is rare because the efficiency of transformation also decreases exponentially with the length of the imported donor locus [485]. Correspondingly, the recombinations importing resistance loci, such as the Tn*916*-type and Tn*1207.1*-type elements, from other species are clearly atypical in their properties among all detected homologous recombinations.

Other detected recombination events around important loci are also atypical. The serotype switching recombinations importing the *cps* loci required to escape vaccine induced immunity in the pneumococcus, for instance, were much larger than most detected around the genome. In *A. baumannii* and *L. pneumophila* too, very large recombination events around the capsular K and LPS loci respectively, are observed and likely driven by host immune selection [732, 733]. Within the pneumococcus, these large *cps* locus recombinations often encompassed the *pbp* loci as well. For instance, the emergence of the 19A clade in PMEN3, which had greatly increased penicillin MIC values, coincided with a large 54 kb recombination spanning the *cps* loci and *pbp2x* and *pbp1a* at the base of this clade. Hence, the adaptation of these pathogens to clinical and public health interventions, selects for recombinations that are either rare, or disruptive enough to not typically persist in populations.

These important adaptive changes mediated by recombination likely reflect the concept underlying Milkman's hypothesis. This states exchanges between divergent genotypes will only become common in the recipient where there exists an atypically strong selection pressure [734, 735]. In our data, the longer recombination events around capsule loci are also an example of this. Were these events more common, genotypes would routinely converge through recombination [736]. In particular, the interspecies recombinations observed in pneumococcal populations suggests the mechanistic or selective barriers to recombinations between streptococcal species in the human oronasopharynx

are not absolute. This is in keeping with the concept of “fuzzy species”, where recombination has created imperfect separation between strains [736–738].

Milkman’s hypothesis could also explain the limited population-level effect of a disrupted *comM* gene on recombination dynamics in the *A. baumannii* lineages. Here, despite previous evidence that *comM* disruption lowers the efficiency of transformation [657, 667, 689], there is no consistent drop in recombination rates through *comM* disruption. Indeed, within GC2, large recombination events around the K locus are observed, these also encompass the *parC* gene involved in fluoroquinolone resistance [739]. Hence, the selection pressure from host immune response and antibiotic consumption may be strong enough to overcome the effect of *comM* disruption. Within GC2, *comM* disruption appears to actually lead to an increase in both the r/m and ρ/m, a surprising feature. This itself may be driven by the frequent recombination around the K locus in those isolates with a disrupted *comM*, with 49 different K types detected within these 4,489 number of isolates, compared to 6 different K types in the 603 isolates without the *comM* disruption.

6.2.3 Comparisons of recombination properties across species

Overall though, among the three pathogens studied, *A. baumannii* tended to import the least diverse sequence through recombination. Across the six largest clades the r/m for *A. baumannii* was 2.08 compared to the value of 17.46 across the three largest *L. pneumophila* clades, while for PMEN3 and PMEN9 this was 13.1 and 7.7 respectively. This could be due to the fact that *A. baumannii* is able to exist as a biofilm on both biotic and harsh abiotic surfaces [740, 741]. This may reduce its exposure to other bacterial species, offering fewer opportunities for diverse sequence uptake. Indeed, the median SNP density of recombination events in the GC2 lineage was low, at 0.003 SNPs/bp. It was interesting to observe the *bap* gene, which is neccessary for biofilm formation [742], being a particular hotspot for recombination in *A. baumannii*. Previous studies have highlighted the variability of *bap* within pathogenic isolates [743]. As well as biofilm formation, its role in binding to human epithelial cells suggests some of this recombination may be driven by host immune pressure [717]. The *bap* gene though, is also highly repetitive in nature, which may confound genome assemblers and be driving some of the variation I observe at this locus [744].

The high *r/m* values observed across the largest *L. pneumophila* clades do broadly follow previous estimates in disease-associated STs [699, 745]. David *et al* 2017 [699] found ST1, which is contained within strain 1 of my results, had an *r/m* of 56.2 compared to my estimate for strain 1 of 9.15 overall, while ST37, contained within strain 2, had an *r/m* of 20.8 compared to the estimate for strain 2 of 26.74 overall. These high values are in contrast to the ρ/m values of 0.053 and 0.092 for strains 1 and 2 respectively. These ρ/m values are lower than the estimates for the PMEN3 lineage at 0.115, and comparable to the PMEN9 lineage, 0.093, which has Tn*1207.1* blocking transformation in over 200 isolates. The SNP density of recombinations within the *L. pneumophila* strains also tends to be lower than observed in the pneumococcus, at a median of 0.0054 SNPs/bp and 0.0059 SNPs/bp for *L. pneumophila* strains 1 and 2 respectively, compared to the medians of 0.0118 SNPs/bp and 0.0092 SNPs/bp for PMEN3 and PMEN9 respectively.

The length of recombination events detected by Gubbins however, is much larger in the *L. pneumophila* strains. The median length for strain 1 is 26,149 bp and for strain 2 is 41,197 bp, this is much higher than seen in PMEN3, at a median of 4,314 bp, and PMEN9, at a median length of 4582 bp. This increased length, which ensures that while SNP density is low the number of imported SNPs overall is high, could be coupled with the low mutation rate of *L. pneumophila* to increase the r/m. Indeed, previous studies have estimated a mutation rate of 1.39 x 10^{-7} substitutions per site per year for *L. pneumophila* [745], an order of magnitude lower than the 1.69 x 10^{-6} substitutions per site per year estimated for PMEN3. In *L. pneumophila* then, selection is preserving large recombination events within strains.

6.2.4 Conflict between hosts and MGEs

The presence of RocRp, and its significant effect of reducing transformation frequency in *L. pneumophila* strains, offers further evidence for the hypothesis that transformation evolved as a means to cure genomes of selfish MGEs [350]. Similarly the insertion of Tn*1207.1* into *comEC*, effectively preventing transformation in the German clade of PMEN9, also highlights how MGEs can target the cell competence machinery to prevent their deletion through transformation. The disruption of *comM* by AbaR elements, while having no consistent effect on transformation at a population level, is still seen to reduce

transformation events *in vitro* [657, 667]. As described above, the recombination events detected by Gubbins tends to be more SNP dense. In this case, regions under selection, such as immunogenic loci, will have recombination events more easily detected. Hence, recombinations from closely related or isogenic cells, which is likely the origin of most donor DNA, are less likely to be detected. These events could drive the deletion of AbaR elements from the host genome. However, within the clades that a disrupted *comM* is detected, this disruption is rarely lost among isolates. In GC2 for instance, where a disrupted *comM* is detected in over 4,000 isolates, there are only three detected reversions to a WT *comM*, as opposed to 16 instances where *comM* has become disrupted. Similarly in GC1, where 96% of the strain has a disrupted *comM*, there is only one detected reversion to a WT *comM*. This could indicate that the rate of curing of the AbaR elements likely causing this disruption is low, possibly due to the lowering of transformation efficiency. Cells could also be selected to retain the resistance cargo that AbaR elements carry, preventing cured isolates from expansion.

Indeed, the MGEs that can disrupt host cell machinery can often also be beneficial to hosts via their cargo genes, especially resistance genes. There is a high selection pressure on pathogenic and human commensal bacteria to evolve resistance [746]. Hence, more reliable mechanisms of spread such as conjugation, which can disseminate large sequences safely through a pilus [395], may be expect to drive much of the spread of these resistance elements. However, in this work I've identified how transformation can also lead to dissemination of these elements. The Tn*916* and Tn*5253*-type ICEs, which can often contain the smaller Tn*1207.1*-type elements, do encode their own conjugation machinery. While these large ICEs may impose a burden on the host cell, their site-specific integration machinery is under selection to minimise the disruption of their insertion to the host cell [747]. By contrast, transformation's extensive import of sequence from another species flanking the insertion is likely to be deleterious to the recipient. This is especially likely when a host gene is disrupted, as in the example of frequent insertions in the *tag* DNA repair gene, and the integration of Tn*1207.1* into *comEC* in the German PMEN9 clade. However, this work shows how transformation and homologous recombination may be a mechanism of further spread of ICE insertions that are not deleterious to the

host cell. Given that antibiotic exposure is one of the many activators of the competence system in bacteria [447, 561], this exposure, as well as selecting for resistance, may lead to further dissemination of resistance genes.

6.2.5 Impact of vaccines on resistance

Finally, it is interesting to observe an instance of vaccination driving the removal of an AMR lineage from the population in this work. The decline and extinction of the PMEN9 German clade post PCV7 introduction in 2007 in Germany, represents the removal of macrolide-resistance from the invasive pneumococcal population. Vaccines targeted toward bacterial pathogens reduce the prevalence of resistant pathogens by inducing host immunity. This both limits the number of resistant pathogens in a population and leads to lower consumption patterns, which may also decrease the evolution of resistance [748]. The PCV vaccines have been an excellent case study in the reduction of AMR, with the introduction of PCV7 in the US followed by an 84% reduction in MDR IPD cases for instance [749, 750]. However, I also observe non-PCV7 serotypes gaining high levels of resistance in the US post introduction of the vaccine, with the 19A clade in PMEN3 an example of this. This is mirrored by findings overall from the US, where 19A linked resistance eroded the initial decrease in resistance from PCV7 introduction [750]. With the subsequent introduction of the PCV-13 vaccine though, which includes the 19A serotype, the burden of resistant infections was again lowered [751]. Vaccines can be a powerful tool in combating the spread of AMR. Genomic epidemiology studies, such as the work presented here in Chapters 3, 4 and 5, will be vital in understanding the effect of vaccines on bacterial populations.

6.3 Limitations and Future work

6.3.1 Detecting MGEs

One of the key drawbacks to the work presented here has been the inability to consistently detect MGEs within sequence data. This has limited the conclusions drawn from Chapter 4, concerning the spread of Tn*916*-type and Tn*1207.1*-type elements, and Chapter 5, with regards to the presence of AbaR elements within isolates. For Chapter 4, it is possible I am underestimating the true diversity of insertion loci of both pneumococcal

MGEs investigated. For instance, 1,895 detected BLAST matches of Tn*916* being non-classifiable, mainly due to poor assembly of the insert and flanking regions. Additionally, given the modular nature of ICEs it could be that I am missing the scale of the diversity of these elements present within pneumococcal populations, as the division between host and MGE genes can be ambiguous. ICE elements have been known to leave ”scars” in the genome upon excision, which could be causing further confusion in the delimiting of ICEs [487].

AbaR elements appear to be similarly diverse in their modular configuration [683,687]. This may be driving some of the disparity in the numbers of isolates with disrupted *comM* genes in the *A. baumannii* strains and the number of isolates with an AbaR element. Other elements than AbaRs could also be causing the observed disruption in *comM* among isolates, although no other element is known to insert within the gene. In general, detecting the presence of MGEs and their insertion sites is a difficult problem. MGEs are diverse in nature and can be flanked by repeat seqeuences that are difficult to assemble. For all the MGEs mentioned here, improved assembly of genomes would allow for a more accurate assessment of their insertion loci. The improvements in accuracy of LRS methods, offers a route through which future studies could create more representative assemblies that allow for the insertion sites of elements to be understood. These LRS methods also promise to be more accurate in identifying plasmids, something NGS short-read methods have struggled to do [752]. Additionally, the CONJscan software, released in late 2019, could be a more appropriate method to detect the conjugation systems of ICE [753]. This would be a more agnostic approach than taken here, looking for all types of ICEs as opposed to specific elements. However, this analysis would still require an extension of the boundaries of ICE to any potential homologous arms. This would then allow for detection of whether an MGE was imported via transformation.

6.3.2 Assembly consistency

The detection of recombination dynamics among the populations of *A. baumannii* and *L. pneumophila* in Chapter 5, also requires good assemblies to ensure adequate SNP detection for Gubbins. However, for these populations I have taken assemblies directly from the GenBank site. These are not actively maintained, as sequences from the Ref-

Seq database are. Additionally there is no consistent tool used for the assembly of these genomes, with this dependent on the date of sequencing and the sequencing method used. The assembler used can have a large effect on the nature of the genome assembly [754]. Hence, future studies should look to use a dataset with a consistent assembly pipeline. Recently, Blackwell *et al* 2021 [554] compiled a dataset of over 600,000 genomes, with 5,162 *A. baumanii* and 2,296 *L. pneumophila* assemblies. They used the same assembly pipeline for all the reads downloaded, Shovil v1.0.4, which gives greater confidence in any downstream analyses on the assemblies.

6.3.3 Gubbins improvements

It should also be noted that Gubbins, on the simulated datasets, struggled to properly categorise SNPs external to recombination. This occured with lower p_{branch} values, highlighted by the increased accuracy of the phylogeny branch length reconstruction at higher p_{branch} values. At lower p_{branch} values, there will be more SNPs on each branch and more recombination events along the branch. With the increase in μ_{rec} the number of recombination events will increase too. Also, given that this is determined by a Poisson distribution, the variation in the number of recombination events will increase between branches. This may drive the poor classification of SNPs at low p_{branch} and high μ_{rec} values. Improvements to the core Gubbins algorithm for detecting higher SNP density may alleviate this issue. The reason ClonalFrameML performs better than Gubbins in terms of branch length estimation may be due to its expectation that recombination lengths follow an exponential distribution. The simulated datasets were created with an exponential distribution of recombination lengths. This is a reasonable prior, given the distribution of recombination events observed in Chapters 3, 4 and 5. However longer recombination events may be enriched in real-world populations. This could be due to the movement of MGEs, which have a minimum length, or selection favouring events targeting larger loci such as the *cps* locus. These are not explicitly modelled in our neutral simulations. Hence, perhaps also incorporating a Pareto distribution with a minimum length cut-off to simulate these enriched longer recombination events, may better model sequence movement between cells via transformation.

Gubbins also requires the mapping of a collection of isolates to a single reference

sequence in order to initially detect SNPs. Depending on how diverse the strain under investigation is, this limits the detection of recombination events primarily to those genes present in the core genome. This is key for determining the clonal frame of isolates, which represents their relationships to each other over generations of evolution. However, recombination events in accessory loci are missed. These loci may play an important role in determining the success of sublineages within a strain [755]. Recent approaches have attempted to call SNPs in the accessory genome using comparisons to a set of reference graphs formed of representative sequences from a collection [240]. Adapting the Gubbins algorithm to also focus on SNPs detected within accessory loci could allow for a greater understanding of the role of recombination in the dissemination of lineages.

6.3.4 Genomic epidemiology

The protocol of steps outlined in Chapters 3, 4 & 5 for the detection of recombination dynamics within strains, offers a reliable suite of methods for genomic epidemiology in the pathogen surveillance era. Assemblies from large populations can be split into closely related strains by PopPUNK, these strains can be mapped to a reference genome through SKA, and finally a recombination free phylogeny of isolates can be inferred by Gubbins (Figure 6.1). Further work combining these steps into one pipeline tool, similar to the PopPIPE tool (`https://github.com/bacpop/PopPIPE`), would enable a wide-range of users access to more fine-scale genomic epidemiological analysis. This could also be further extended with initial assembly steps, which might be dependent on the sequencing methodology deployed. Further analysis steps, such as the prediction of AMR genes, MGE detection and phylodynamic analyses of strains, may also be incorporated.

6.4 Conclusion

The dissemination of AMR genes among pathogenic bacterial species represents a pressing concern for public health. Understanding how and why resistance spreads, seemingly aggregating in individual clones among the vast diversity of some bacterial species, will be important in preventing this rise in resistance. Genomic epidemiology studies, such as the work presented here, represent a powerful tool in tracking and understanding these dynamics. Allied with improved sequencing methods, enhanced surveillance networks

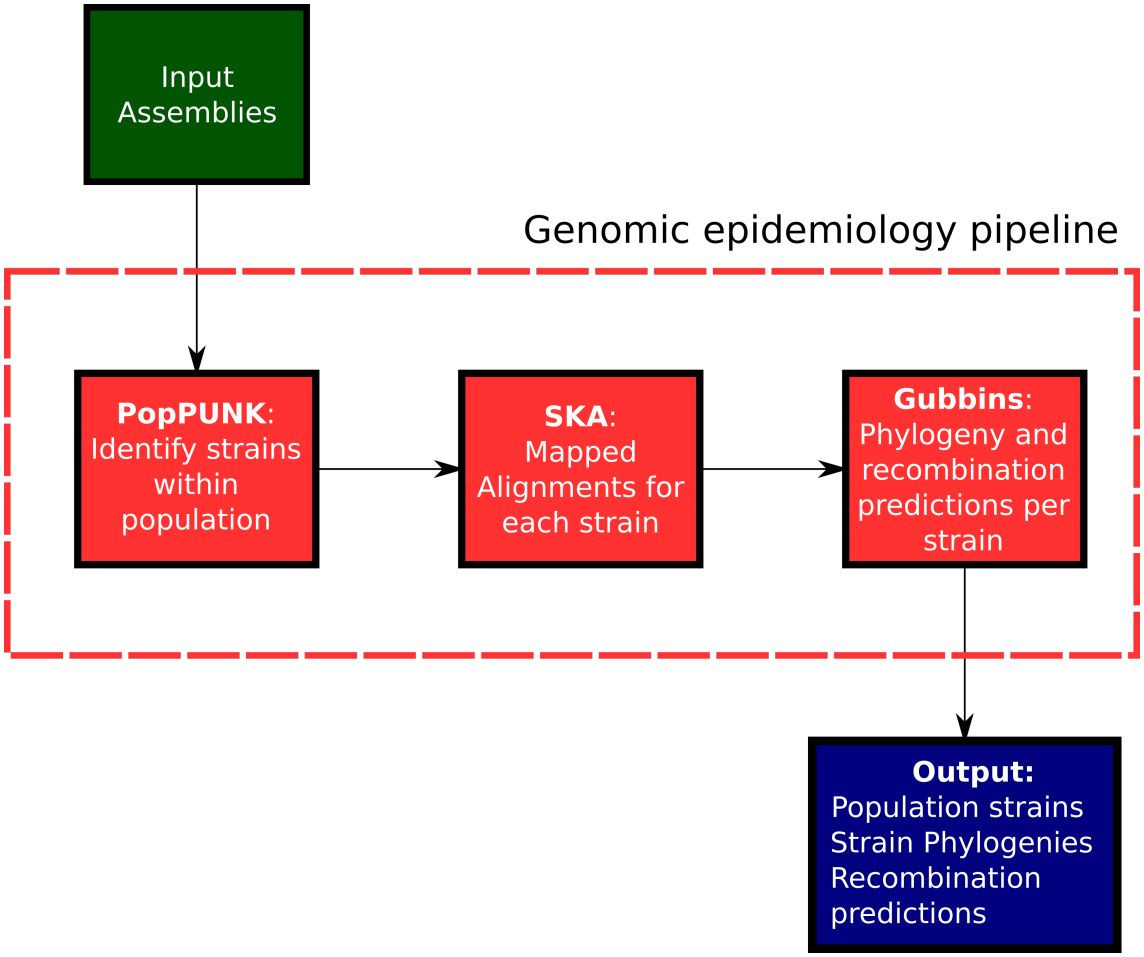

Figure 6.1: Outline of the main steps in moving from assemblies to strain level phylogenies, for a collection of sequences.

and verstaile bioinformatic tools, these studies will become ever more informative in answering these key questions. Apart from the clinical implications however, the spread of resistance is a fascinating evolutionary case study occurring within the scale of human lifetimes. One which happens to have profound implications for the practice of medicine.

Bibliography

[1] Department or Myths and Fables: The Churchill-Fleming Non-Connection: International Churchill Society - https://winstonchurchill.org/publications/finest-hour/finest-hour-102/department-or-myths-and-fables-the-churchill-fleming-non-connection/ Date accessed: 28/05/2022.

[2] Diana Davenport. The war against bacteria: how were sulphonamide drugs used by Britain during World War II? *Medical humanities*, 38(1):55–58, 6 2012.

[3] Gerhard Domagk. Chemotherapie der bakteriellen Infektionen. *Angewandte Chemie*, 48(42):657–667, 10 1935.

[4] Richard J Henry. The mode of action of sulfonamides. *Bacteriological Reviews*, 7(4):175, 3 1943.

[5] Ola Sköld. Sulfonamide resistance: mechanisms and trends. *Drug Resistance Updates*, 3(3):155–160, 6 2000.

[6] Aniruddha Achari, Donald O. Somers, John N. Champness, Patrick K. Bryant, Jane Rosemond, and David K. Stammers. Crystal structure of the anti-bacterial sulfonamide drug target dihydropteroate synthase. *Nature Structural Biology 1997 4:6*, 4(6):490–497, 6 1997.

[7] G. M. Brown. The Biosynthesis of Folic Acid: II. Inhibition by sulfonamides. *Journal of Biological Chemistry*, 237(2):536–540, 2 1962.

[8] Anna Białk-Bielińska, Stefan Stolte, Jürgen Arning, Ute Uebers, Andrea Böschen, Piotr Stepnowski, and Marianne Matzke. Ecotoxicity evaluation of selected sulfonamides. *Chemosphere*, 85(6):928–933, 10 2011.

[9] Cameron R. Strachan and Julian Davies. The Whys and Wherefores of Antibiotic Resistance. *Cold Spring Harbor Perspectives in Medicine*, 7(2), 2 2017.

[10] George A Jacoby and G A Jacoby. History of Drug-Resistant Microbes. *Antimicrobial Drug Resistance*, pages 3–8, 2017.

[11] Douglas S. Damrosch. Chemoprophylaxis and Sulfonamide resistant streptococci. *Journal of the American Medical Association*, 130(3):124–128, 1 1946.

[12] H. A. Feldman. Sulfonamide-resistant meningococci. *Annual review of medicine*, 18:495–506, 1967.

[13] Magnus Unemo and William M. Shafer. Antimicrobial Resistance in Neisseria gonorrhoeae in the 21st Century: Past, Evolution, and Future. *Clinical Microbiology Reviews*, 27(3):587, 2014.

[14] M Landy, N W Larkum, E J Oswald, and F Streightoff. Increase synthesis of p-aminobenzoic acid associated with the development of sulfonamide resistance in Staphylococcus aureus. *Science (New York, N.Y.)*, 97(2516):265–267, 3 1943.

[15] E. M. Wise and M. M. Abou Donia. Sulfonamide resistance mechanism in Escherichia coli: R plasmids can determine sulfonamide-resistant dihydropteroate synthases. *Proceedings of the National Academy of Sciences of the United States of America*, 72(7):2621–2625, 1975.

[16] G. Swedberg, S. Castensson, and O. Skold. Characterization of mutationally altered dihydropteroate synthase and its ability to form a sulfonamide-containing dihydrofolate analog. *Journal of bacteriology*, 137(1):129–136, 1979.

[17] J P Maskell, A M Sefton, and L M Hall. Multiple mutations modulate the function of dihydrofolate reductase in trimethoprim-resistant Streptococcus pneumoniae. *Antimicrobial agents and chemotherapy*, 45(4):1104–8, 4 2001.

[18] Julian Davies and Dorothy Davies. Origins and Evolution of Antibiotic Resistance. *Microbiology and Molecular Biology Reviews : MMBR*, 74(3):417, 2010.

[19] Kathrin I Mohr. History of Antibiotics Research. In Marc Stadler and Petra Dersch, editors, *How to Overcome the Antibiotic Crisis : Facts, Challenges, Technologies and Future Perspectives*, pages 237–272. Springer International Publishing, Cham, 2016.

[20] E. P. Abraham and E. Chain. An Enzyme from Bacteria able to Destroy Penicillin. *Nature 1940 146:3713*, 146(3713):837–837, 1940.

[21] Mariya Lobanovska and Giulia Pilla. Focus: Drug Development: Penicillin's Discovery and Antibiotic Resistance: Lessons for the Future? *The Yale Journal of Biology and Medicine*, 90(1):135, 3 2017.

[22] Charles H. Rammelkamp and Thelma Maxon. Resistance of Staphylococcus aureus to the Action of Penicillin. *Exp Biol Med*, 51(3):386–389, 11 1942.

[23] Kate Gould. Antibiotics: from prehistory to the present day. *Journal of Antimicrobial Chemotherapy*, 71(3):572–575, 3 2016.

[24] Michael Rawlins. The disputed discovery of streptomycin. *The Lancet*, 380(9838):207, 7 2012.

[25] Matt Hutchings, Andrew Truman, and Barrie Wilkinson. Antibiotics: past, present and future. *Current Opinion in Microbiology*, 51:72–80, 10 2019.

[26] Essaid Ait Barka, Parul Vatsa, Lisa Sanchez, Nathalie Gaveau-Vaillant, Cedric Jacquard, Hans-Peter Klenk, Christophe Clément, Yder Ouhdouch, and Gilles P. van Wezel. Taxonomy, Physiology, and Natural Products of Actinobacteria. *Microbiology and molecular biology reviews : MMBR*, 80(1):1–43, 3 2015.

[27] Selman A. Waksman - Facts - https://www.nobelprize.org/prizes/medicine/1952/waksman/facts/ Date accessed 30/05/2022.

[28] H. Corwin Hinshaw, Marjorie M. Pyle, and William H. Feldman. Streptomycin in tuberculosis. *The American Journal of Medicine*, 2(5):429–435, 5 1947.

[29] Yōko Takahashi and Takuji Nakashima. Actinomycetes, an Inexhaustible Source of Naturally Occurring Antibiotics. *Antibiotics*, 7(2):45, 5 2018.

[30] D. P. Labeda. Transfer of the type strain of Streptomyces erythraeus (Waksman 1923) Waksman and Henrici 1948 to the genus Saccharopolyspora Lacey and Goodfellow 1975 as Saccharopolyspora erythraea sp. nov., and designation of a neotype strain for Streptomyces erythraeus. *International Journal of Systematic Bacteriology*, 37(1):19–22, 1 1987.

[31] Patrick Radden Keefe. *Empire of pain: The secret history of the Sackler family*. Picador, London, 2021.

[32] T. M. Embley and E. Stackebrandt. The molecular phylogeny and systematics of the actinomycetes. *Annual review of microbiology*, 48:257–289, 1994.

[33] Gerald B. Pier. On the Greatly Exaggerated Reports of the Death of Infectious Diseases. *Clinical Infectious Diseases*, 47(8):1113–1114, 10 2008.

[34] Henry F. Chambers and Frank R. DeLeo. Waves of Resistance: Staphylococcus aureus in the Antibiotic Era. *Nature reviews. Microbiology*, 7(9):629, 2009.

[35] M. P. Jevons and M. T. Parker. The evolution of new hospital strains of Staphylococcus aureus. *Journal of clinical pathology*, 17(3):243–250, 1964.

[36] M Patricia Jevons. “Celbenin” - resistant Staphylococci. *British Medical Journal*, 1(5219):124–125, 1 1961.

[37] M. Barber. Methicillin-resistant staphylococci. *Journal of clinical pathology*, 14(4):385–393, 1961.

[38] David J. Payne, Linda Federici Miller, David Findlay, James Anderson, and Lynn Marks. Time for a change: addressing R&D and commercialization challenges for antibacterials. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1670), 2015.

[39] Ruben Tommasi, Dean G. Brown, Grant K. Walkup, John I. Manchester, and Alita A. Miller. ESKAPEing the labyrinth of antibacterial discovery. *Nature Reviews Drug Discovery 2015 14:8*, 14(8):529–542, 7 2015.

[40] Victoria L. Simpkin, Matthew J. Renwick, Ruth Kelly, and Elias Mossialos. Incentivising innovation in antibiotic drug discovery and development: progress, challenges and next steps. *The Journal of Antibiotics 2017 70:12*, 70(12):1087–1096, 11 2017.

[41] Javier Santos-Aberturas and Natalia M. Vior. Beyond Soil-Dwelling Actinobacteria: Fantastic Antibiotics and Where to Find Them. *Antibiotics (Basel, Switzerland)*, 11(2), 2 2022.

[42] Zhiwei Qin, John T. Munnoch, Rebecca Devine, Neil A. Holmes, Ryan F. Seipke, Karl A. Wilkinson, Barrie Wilkinson, and Matthew I. Hutchings. Formicamycins, antibacterial polyketides produced by Streptomyces formicae isolated from African Tetraponera plant-ants. *Chemical science*, 8(4):3218–3227, 2017.

[43] Kyuho Moon, Fei Xu, Chen Zhang, and Mohammad R. Seyedsayamdost. Bioactivity-HiTES Unveils Cryptic Antibiotics Encoded in Actinomycete Bacteria. *ACS chemical biology*, 14(4):767–774, 4 2019.

[44] Garima Kapoor, Saurabh Saigal, and Ashok Elongavan. Action and resistance mechanisms of antibiotics: A guide for clinicians. *Journal of Anaesthesiology, Clinical Pharmacology*, 33(3):300, 7 2017.

[45] Marcio V. Bertacine Dias, Jademilson C. Santos, Gerardo A. Libreros-Zúñiga, João A. Ribeiro, and Sair M. Chavez-Pacheco. Folate biosynthesis pathway: mechanisms and insights into drug design for infectious diseases. *Future medicinal chemistry*, 10(8):935–959, 4 2018.

[46] Krista S. Crider, Thomas P. Yang, Robert J. Berry, and Lynn B. Bailey. Folate and DNA Methylation: A Review of Molecular Mechanisms and the Evidence for Folate's Role. *Advances in Nutrition*, 3(1):21, 1 2012.

[47] Magdalena Uzlikova and Eva Nohynkova. The effect of metronidazole on the cell cycle and DNA in metronidazole-susceptible and -resistant Giardia cell lines. *Molecular and biochemical parasitology*, 198(2):75–81, 2014.

[48] Nitrofurantoin. *Meyler's Side Effects of Drugs*, pages 210–218, 1 2016.

[49] F. M. Barnard and A. Maxwell. Interaction between DNA Gyrase and Quinolones: Effects of Alanine Mutations at GyrA Subunit Residues Ser83 and Asp87. *Antimicrobial Agents and Chemotherapy*, 45(7):1994, 2001.

[50] Bumduuren Tuvshintulga, Mahmoud Aboulaila, Thillaiampalam Sivakumar, Dickson Stuart Tayebwa, Sambuu Gantuya, Khandsuren Naranbaatar, Aki Ishiyama, Masato Iwatsuki, Kazuhiko Otoguro, Satoshi Omura, Mohamad Alaa Terkawi, Azirwan Guswanto, Mohamed Abdo Rizk, Naoaki Yokoyama, and Ikuo Igarashi. Chemotherapeutic efficacies of a clofazimine and diminazene aceturate combination against piroplasm parasites and their AT-rich DNA-binding activity on Babesia bovis. *Scientific Reports 2017 7:1*, 7(1):1–10, 10 2017.

[51] Heinz G. Floss and Tin Wein Yu. Rifamycin - Mode of action, resistance, and biosynthesis. *Chemical Reviews*, 105(2):621–632, 2 2005.

[52] Aashish Srivastava, Meliza Talaue, Shuang Liu, David Degen, Richard Y. Ebright, Elena Sineva, Anirban Chakraborty, Sergey Y. Druzhinin, Sujoy Chatterjee, Jayanta Mukhopadhyay, Yon W. Ebright, Alex Zozula, Juan Shen, Sonali Sengupta, Rui Rong Niedfeldt, Cai Xin, Takushi Kaneko, Herbert Irschik, Rolf Jansen, Stefano Donadio, Nancy Connell, and Richard H. Ebright. New target for inhibition of bacterial RNA polymerase: 'switch region'. *Current Opinion in Microbiology*, 14(5):532–543, 10 2011.

[53] Andrew P. Carter, William M. Clemons, Ditlev E. Brodersen, Robert J. Morgan-Warren, Brian T. Wimberly, and V. Ramakrishnan. Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature 2000 407:6802*, 407(6802):340–348, 9 2000.

[54] Frank Schluenzen, Ante Tocilj, Raz Zarivach, Joerg Harms, Marco Gluehmann, Daniela Janell, Anat Bashan, Heike Bartels, Ilana Agmon, François Franceschi, and Ada Yonath. Structure of Functionally Activated Small Ribosomal Subunit at 3.3 Å Resolution. *Cell*, 102(5):615–623, 9 2000.

[55] Maria M. Anokhina, Andrea Barta, Knud H. Nierhaus, Vera A. Spiridonova, and Alexei M. Kopylov. Mapping of the second tetracycline binding site on the ribosomal small subunit of E.coli. *Nucleic Acids Research*, 32(8):2594–2597, 4 2004.

[56] W. S. Champney and R. Burdine. Macrolide antibiotics inhibit 50S ribosomal subunit assembly in Bacillus subtilis and Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, 39(9):2141–2144, 1995.

[57] N. Ban, P. Nissen, J. Hansen, P. B. Moore, and T. A. Steitz. The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution. *Science*, 289(5481):905–920, 8 2000.

[58] Corinna Kehrenberg, Stefan Schwarz, Lene Jacobsen, Lykke H. Hansen, and Birte Vester. A new mechanism for chloramphenicol, florfenicol and clindamycin resistance: methylation of 23S ribosomal RNA at A2503. *Molecular Microbiology*, 57(4):1064–1073, 8 2005.

[59] Stefan Schwarz, Jianzhong Shen, Kristina Kadlec, Yang Wang, Geovana Brenner Michael, Andrea T. Feßler, and Birte Vester. Lincosamides, Streptogramins, Phenicols, and Pleuromutilins: Mode of Action and Mechanisms of Resistance. *Cold Spring Harbor Perspectives in Medicine*, 6(11):a027037, 11 2016.

[60] Karen L. Leach, Steven M. Swaney, Jerry R. Colca, William G. McDonald, James R. Blinn, Lisa M M. Thomasco, Robert C. Gadwood, Dean Shinabarger, Liqun Xiong, and Alexander S. Mankin. The Site of Action of Oxazolidinone Antibiotics in Living Bacteria and in Human Mitochondria. *Molecular Cell*, 26(3):393–402, 5 2007.

[61] Daniel J. Diekema and Ronald N. Jones. Oxazolidinone antibiotics. *The Lancet*, 358(9297):1975–1982, 12 2001.

[62] Nora Vázquez-Laslop and Alexander S. Mankin. How Macrolide Antibiotics Work. *Trends in Biochemical Sciences*, 43(9):668–684, 9 2018.

[63] David J. Farrell, Mariana Castanheira, and Ian Chopra. Characterization of Global Patterns and the Genetics of Fusidic Acid Resistance. *Clinical Infectious Diseases*, 52(suppl_7):S487–S492, 6 2011.

[64] Christopher M. Thomas, Joanne Hothersall, Christine L. Willis, and Thomas J. Simpson. Resistance to and synthesis of the antibiotic mupirocin. *Nature Reviews Microbiology 2010 8:4*, 8(4):281–289, 3 2010.

[65] Saeed Khoshnood, Mohsen Heidary, Arezoo Asadi, Saleh Soleimani, Moloudsadat Motahar, Mohammad Savari, Morteza Saki, and Mahtab Abdi. A review on mechanism of action, resistance, synergism, and clinical implications of mupirocin against Staphylococcus aureus. *Biomedicine & Pharmacotherapy*, 109:1809–1818, 1 2019.

[66] Michael J. Trimble, Patrik Mlynárčik, Milan Kolář, and Robert E.W. Hancock. Polymyxin: Alternative Mechanisms of Action and Resistance. *Cold Spring Harbor Perspectives in Medicine*, 6(10):a025288, 10 2016.

[67] Hamza Olleik, Cendrine Nicoletti, Mickael Lafond, Elise Courvoisier-Dezord, Peiwen Xue, Akram Hijazi, Elias Baydoun, Josette Perrier, and Marc Maresca. Comparative Structure-Activity Analysis of the Antimicrobial Activity, Cytotoxicity, and Mechanism of Action of the Fungal Cyclohexadepsipeptides Enniatins and Beauvericin. *Toxins*, 11(9), 9 2019.

[68] Thomas M. Wood and Nathaniel I. Martin. The calcium-dependent lipopeptide antibiotics: structure, mechanism, & medicinal chemistry. *MedChemComm*, 10(5):634–646, 2019.

[69] Je Wen Liou, Yu Jiun Hung, Chin Hao Yang, and Yi Cheng Chen. The Antimicrobial Activity of Gramicidin A Is Associated with Hydroxyl Radical Formation. *PLOS ONE*, 10(1):e0117065, 1 2015.

[70] Karen Bush and Patricia A. Bradford. β-Lactams and β-Lactamase Inhibitors: An Overview. *Cold Spring Harbor Perspectives in Medicine*, 6(8):a025247, 8 2016.

[71] Leticia I. Llarrull, Sebastian A. Testero, Jed F. Fisher, and Shahriar Mobashery. The future of the β-lactams. *Current Opinion in Microbiology*, 13(5):551–557, 10 2010.

[72] Gareth A. Prosser and Luiz Pedro S. De Carvalho. Reinterpreting the mechanism of inhibition of Mycobacterium tuberculosis D-alanine:D-alanine ligase by D-cycloserine. *Biochemistry*, 52(40):7145–7149, 10 2013.

[73] Lynn L. Silver. Fosfomycin: Mechanism and Resistance. *Cold Spring Harbor Perspectives in Medicine*, 7(2):a025262, 2 2017.

[74] Catherine Vilchèze and William R. Jacobs. The Mechanism of Isoniazid Killing: Clarity Through the Scope of Genetics. *http://dx.doi.org/10.1146/annurev.micro.61.111606.122346*, 61:35–50, 9 2007.

[75] Chen Zhu, Yu Liu, Lihua Hu, Min Yang, and Zheng Guo He. Molecular mechanism of the synergistic activity of ethambutol and isoniazid against Mycobacterium tuberculosis. *Journal of Biological Chemistry*, 293(43):16741–16750, 10 2019.

[76] Daina Zeng, Dmitri Debabov, Theresa L. Hartsell, Raul J. Cano, Stacy Adams, Jessica A. Schuyler, Ronald McMillan, and John L. Pace. Approved Glycopeptide Antibacterial Drugs: Mechanism of Action and Resistance. *Cold Spring Harbor Perspectives in Medicine*, 6(12), 2016.

[77] T. J. Pollock, L. Thorne, M. Yamazaki, M. J. Mikolajczak, and R. W. Armentrout. Mechanism of bacitracin resistance in gram-negative bacteria that synthesize exopolysaccharides. *Journal of bacteriology*, 176(20):6229–6237, 1994.

[78] Jessica M.A. Blair, Mark A. Webber, Alison J. Baylay, David O. Ogbolu, and Laura J.V. Piddock. Molecular mechanisms of antibiotic resistance. *Nature Reviews Microbiology 2014 13:1*, 13(1):42–51, 12 2014.

[79] Christopher P. Randall, Katherine R. Mariner, Ian Chopra, and Alex J. O'Neill. The target of daptomycin is absent from Escherichia coli and other gram-negative pathogens. *Antimicrobial agents and chemotherapy*, 57(1):637–639, 1 2013.

[80] Mathieu Brochet, Elisabeth Couvé, Mohamed Zouine, Claire Poyart, and Philippe Glaser. A Naturally Occurring Gene Amplification Leading to Sulfonamide and Trimethoprim Resistance in Streptococcus agalactiae. *Journal of Bacteriology*, 190(2):672, 1 2008.

[81] Martin Antonio, Neil McFerran, and Mark J. Pallen. Mutations affecting the Rossman fold of isoleucyl-tRNA synthetase are correlated with low-level mupirocin resistance in Staphylococcus aureus. *Antimicrobial agents and chemotherapy*, 46(2):438–442, 2002.

[82] Andie S. Lee, Yann Gizard, Joanna Empel, Eve Julie Bonetti, Stephan Harbarth, and Patrice François. Mupirocin-Induced Mutations in ileS in Various Genetic Backgrounds of Methicillin-Resistant Staphylococcus aureus. *Journal of Clinical Microbiology*, 52(10):3749–3754, 10 2014.

[83] Thomas Bataillon. Shaking the 'deleterious mutations' dogma? *Trends in Ecology & Evolution*, 18(7):315–317, 7 2003.

[84] Dan I. Andersson, Sophie Maisnier Patin, Annika I. Nilsson, and Elisabeth Kugelberg. The Biological Cost of Antibiotic Resistance. In Robert A. Bonomo and Marcelo Tolmansky, editors, *Enzyme-Mediated Resistance to Antibiotics*, pages 339–348. John Wiley & Sons, Ltd, 4 2014.

[85] Wilhelm Paulander, Sophie Maisnier-Patin, and Dan I. Andersson. Multiple mechanisms to ameliorate the fitness burden of mupirocin resistance in Salmonella typhimurium. *Molecular microbiology*, 64(4):1038–1048, 5 2007.

[86] Wilhelm Paulander, Dan I. Andersson, and Sophie Maisnier-Patin. Amplification of the Gene for Isoleucyl–tRNA Synthetase Facilitates Adaptation to the Fitness Cost of Mupirocin Resistance in Salmonella enterica. *Genetics*, 185(1):305–312, 5 2010.

[87] Angela M. Starks, Aysel Gumusboga, Bonnie B. Plikaytis, Thomas M. Shinnick, and James E. Posey. Mutations at embB Codon 306 Are an Important Molecular Indicator of Ethambutol Resistance in Mycobacterium tuberculosis. *Antimicrobial Agents and Chemotherapy*, 53(3):1061–1066, 3 2009.

[88] Hassan Safi, Subramanya Lingaraju, Anita Amin, Soyeon Kim, Marcus Jones, Michael Holmes, Michael McNeil, Scott N. Peterson, Delphi Chatterjee, Robert Fleischmann, and David Alland. Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-β-D-arabinose biosynthetic and utilization pathway genes. *Nature Genetics 2013 45:10*, 45(10):1190–1197, 9 2013.

[89] Sebastien Gagneux, Clara Davis Long, Peter M. Small, Tran Van, Gary K. Schoolnik, and Brendan J.M. Bohannan. The competitive cost of antibiotic resistance in Mycobacterium tuberculosis. *Science*, 312(5782):1944–1946, 6 2006.

[90] Amel Kevin Alame Emane, Xujun Guo, Howard E. Takiff, and Shengyuan Liu. Drug resistance, fitness and compensatory mutations in Mycobacterium tuberculosis. *Tuberculosis*, 129:102091, 7 2021.

[91] S. K. Parida, R. Axelsson-Robertson, M. V. Rao, N. Singh, I. Master, A. Lutckii, S. Keshavjee, J. Andersson, A. Zumla, and M. J. Maeurer. Totally drug-resistant tuberculosis and adjunct therapies. *Journal of Internal Medicine*, 277(4):388–405, 4 2015.

[92] Matthias Merker, Maxime Barbier, Helen Cox, Jean Philippe Rasigade, Silke Feuerriegel, Thomas Andreas Kohl, Roland Diel, Sonia Borrell, Sebastien Gagneux, Vladyslav Nikolayevskyy, Sönke Andres, Ulrich Nübel, Philip Supply, Thierry Wirth, and Stefan Niemann. Compensatory evolution drives multidrug-resistant tuberculosis in central Asia. *eLife*, 7, 10 2018.

[93] Maxim S. Svetlov, Egor A. Syroegin, Elena V. Aleksandrova, Gemma C. Atkinson, Steven T. Gregory, Alexander S. Mankin, and Yury S. Polikanov. Structure of Erm-modified 70S ribosome reveals the mechanism of macrolide resistance. *Nature Chemical Biology 2021 17:4*, 17(4):412–420, 1 2021.

[94] Adam J. Schaenzer and Gerard D. Wright. Antibiotic Resistance by Enzymatic Modification of Antibiotic Targets. *Trends in Molecular Medicine*, 26(8):768–782, 8 2020.

[95] Peter A. Lambert. Bacterial resistance to antibiotics: Modified target sites. *Advanced Drug Delivery Reviews*, 57(10):1471–1485, 7 2005.

[96] F. Depardieu and P. Courvalin. Mutation in 23S rRNA responsible for resistance to 16-membered macrolides and streptogramins in Streptococcus pneumoniae. *Antimicrobial Agents and Chemotherapy*, 45(1):319–323, 2001.

[97] Lucía Fernández and Robert E.W. Hancock. Adaptive and Mutational Resistance: Role of Porins and Efflux Pumps in Drug Resistance. *Clinical Microbiology Reviews*, 25(4):661, 10 2012.

[98] Seiji Kojima and Hiroshi Nikaido. Permeation rates of penicillins indicate that Escherichia coli porins function principally as nonspecific channels. *Proceedings of the National Academy of Sciences of the United States of America*, 110(28):E2629, 7 2013.

[99] Aniela Wozniak, Nicolás A. Villagra, Agustina Undabarrena, Natalia Gallardo, Nicole Keller, Marcela Moraga, Juan C. Román, Guido C. Mora, and Patricia García. Porin alterations present in non-carbapenemase-producing Enterobacteriaceae with high and intermediate levels of carbapenem resistance in Chile. *Journal of medical microbiology*, 61(Pt 9):1270–1279, 9 2012.

[100] M. A. Arbing, J. W. Hanrahan, and J. W. Coulton. Altered channel properties of porins from Haemophilus influenzae: isolates from cystic fibrosis patients. *The Journal of membrane biology*, 189(2):131–141, 9 2002.

[101] Bryan D. Schindler and Glenn W. Kaatz. Multidrug efflux pumps of Gram-positive bacteria. *Drug Resistance Updates*, 27:1–13, 7 2016.

[102] Paula Blanco, Sara Hernando-Amado, Jose Antonio Reales-Calderon, Fernando Corona, Felipe Lira, Manuel Alcalde-Rico, Alejandra Bernardini, Maria Blanca Sanchez, and Jose Luis Martinez. Bacterial Multidrug Efflux Pumps: Much More Than Antibiotic Resistance Determinants. *Microorganisms*, 4(1), 3 2016.

[103] Keith Poole. Efflux pumps as antimicrobial resistance mechanisms. *Annals of medicine*, 39(3):162–176, 2007.

[104] K. Ubukata, N. Itoh-Yamashita, and M. Konno. Cloning and expression of the norA gene for fluoroquinolone resistance in Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, 33(9):1535–1539, 1989.

[105] Sofia Santos Costa, Benjamin Sobkowiak, Ricardo Parreira, Jonathan D. Edgeworth, Miguel Viveiros, Taane G. Clark, and Isabel Couto. Genetic diversity of norA, coding for a main efflux pump of Staphylococcus aureus. *Frontiers in Genetics*, 10(JAN):710, 2019.

[106] Aurélie A. Huet, Jose L. Raygada, Kabir Mendiratta, Susan M. Seo, and Glenn W. Kaatz. Multidrug efflux pump overexpression in Staphylococcus aureus after single and multiple in vitro exposures to biocides and dyes. *Microbiology*, 154(10):3144–3153, 10 2008.

[107] Xin Deng, Fei Sun, Quanjiang Ji, Haihua Liang, Dominique Missiakas, Lefu Lan, and Chuan He. Expression of multidrug resistance efflux pump gene norA is iron responsive in Staphylococcus aureus. *Journal of bacteriology*, 194(7):1753–1762, 4 2012.

[108] Laura J.V. Piddock. Clinically relevant chromosomally encoded multidrug resistance efflux pumps in bacteria. *Clinical Microbiology Reviews*, 19(2):382–402, 4 2006.

[109] João Anes, Matthew P. McCusker, Séamus Fanning, and Marta Martins. The ins and outs of RND efflux pumps in Escherichia coli. *Frontiers in Microbiology*, 6(JUN):587, 2015.

[110] Hidetada Hirakawa, Yoshihiko Inazumi, Takeshi Masaki, Takahiro Hirata, and Akihito Yamaguchi. Indole induces the expression of multidrug exporter genes in Escherichia coli. *Molecular Microbiology*, 55(4):1113–1126, 2 2005.

[111] P. A. Bradford. Extended-Spectrum β-Lactamases in the 21st Century: Characterization, Epidemiology, and Detection of This Important Resistance Threat. *Clinical Microbiology Reviews*, 14(4):933, 2001.

[112] Naomi Datta and Polyxeni Kontomichalou. Penicillinase synthesis controlled by infectious R factors in Enterobacteriaceae. *Nature*, 208(5007):239–241, 1965.

[113] Merijn L.M. Salverda, J. Arjan G.M. de Visser, and Miriam Barlow. Natural evolution of TEM-1 β-lactamase: experimental reconstruction and clinical relevance. *FEMS Microbiology Reviews*, 34(6):1015–1036, 11 2010.

[114] Mariana Castanheira, Patricia J Simner, and Patricia A Bradford. Extended-spectrum β-lactamases: an update on their characteristics, epidemiology and detection. *JAC-Antimicrobial Resistance*, 3(3), 7 2021.

[115] Teemu Kallonen, Hayley J. Brodrick, Simon R. Harris, Jukka Corander, Nicholas M. Brown, Veronique Martin, Sharon J. Peacock, and Julian Parkhill. Systematic longitudinal survey of invasive Escherichia coli in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome Research*, 2017.

[116] Alan McNally, Teemu Kallonen, Christopher Connor, Khalil Abudahab, David M. Aanensen, Carolyne Horner, Sharon J. Peacock, Julian Parkhill, Nicholas J. Croucher, and Jukka Corander. Diversification of colonization factors in a multidrug-resistant escherichia coli lineage evolving under negative frequency- dependent selection. *mBio*, 10(2), 3 2019.

[117] D. M. Livermore and Peter M. Hawkey. CTX-M: changing the face of ESBLs in the UK. *Journal of Antimicrobial Chemotherapy*, 56(3):451–454, 9 2005.

[118] Maria S. Ramirez and Marcelo E. Tolmasky. Aminoglycoside modifying enzymes. *Drug Resistance Updates*, 13(6):151–171, 12 2010.

[119] Kuastros Mekonnen Belaynehe, Seung Won Shin, Park Hong-Tae, and Han Sang Yoo. Occurrence of aminoglycoside-modifying enzymes among isolates of Escherichia coli exhibiting high levels of aminoglycoside resistance isolated from Korean cattle farms. *FEMS Microbiology Letters*, 364(14), 8 2017.

[120] Johan Bengtsson-Palme and D. G.Joakim Larsson. Antibiotic resistance genes in the environment: prioritizing risks. *Nature Reviews Microbiology 2015 13:6*, 13(6):396–396, 4 2015.

[121] Swati Sharma, Tuhina Banerjee, Ashok Kumar, Ghanshyam Yadav, and Sriparna Basu. Extensive outbreak of colistin resistant, carbapenemase (bla OXA-48, bla NDM) producing Klebsiella pneumoniae in a large tertiary care hospital, India. *Antimicrobial Resistance and Infection Control*, 11(1):1–9, 12 2022.

[122] Simon Lax and Jack A. Gilbert. Hospital-associated microbiota and implications for nosocomial infections. *Trends in Molecular Medicine*, 21(7):427–432, 7 2015.

[123] Ana Paula Christoff, Aline Fernanda Rodrigues Sereia, Giuliano Netto Flores Cruz, Daniela Carolina De Bastiani, Vanessa Leitner Silva, Camila Hernandes, Ana Paula Metran Nascente, Ana Andrea Dos Reis, Renata Gonçalves Viessi, Andrea Dos Santos Pereira Marques, Bianca Silva Braga, Telma Priscila Lovizio Raduan, Marines Dalla Valle Martino, Fernando Gatti De Menezes, and Luiz Felipe Valter De Oliveira. One year cross-sectional study in adult and neonatal intensive care units reveals the bacterial and antimicrobial resistance genes profiles in patients and hospital surfaces. *PLOS ONE*, 15(6):e0234127, 6 2020.

[124] C. Urbaniak, A. Checinska Sielaff, K. G. Frey, J. E. Allen, N. Singh, C. Jaing, K. Wheeler, and K. Venkateswaran. Detection of antimicrobial resistance genes associated with the International Space Station environmental surfaces. *Scientific Reports 2018 8:1*, 8(1):1–13, 1 2018.

[125] Kirandeep Bhullar, Nicholas Waglechner, Andrew Pawlowski, Kalinka Koteva, Eric D. Banks, Michael D. Johnston, Hazel A. Barton, and Gerard D. Wright. Antibiotic Resistance Is Prevalent in an Isolated Cave Microbiome. *PLOS ONE*, 7(4):e34953, 2012.

[126] Nicholas Waglechner and Gerard D. Wright. Antibiotic resistance: it's bad, but why isn't it worse? *BMC biology*, 15(1), 9 2017.

[127] Gerard D. Wright and Hendrik Poinar. Antibiotic resistance is ancient: implications for drug discovery. *Trends in Microbiology*, 20(4):157–159, 4 2012.

[128] Abiola Olumuyiwa Olaitan and Jean-Marc Rolain. Ancient Resistome. In Michel Dranourt and Didier Raoult, editors, *Paleomicrobiology of Humans*, pages 75–80. John Wiley & Sons, Ltd, 9 2016.

[129] Kate S. Baker, Edward Burnett, Hannah McGregor, Ana Deheer-Graham, Christine Boinett, Gemma C. Langridge, Alexander M. Wailan, Amy K. Cain, Nicholas R. Thomson, Julie E. Russell, and Julian Parkhill. The Murray collection of pre-antibiotic era Enterobacteriacae: A unique research resource. *Genome Medicine*, 7(1):1–7, 9 2015.

[130] P. M. Hawkey. The origins and molecular basis of antibiotic resistance. *BMJ : British Medical Journal*, 317(7159):657, 9 1998.

[131] Gerard D. Wright. The antibiotic resistome: the nexus of chemical and genetic diversity. *Nature Reviews Microbiology 2007 5:3*, 5(3):175–186, 3 2007.

[132] Laurent Poirel, Peter Kämpfer, and Patrice Nordmann. Chromosome-encoded Ambler class A beta-lactamase of Kluyvera georgiana, a probable progenitor of a subgroup of CTX-M extended-spectrum beta-lactamases. *Antimicrobial agents and chemotherapy*, 46(12):4038–4040, 12 2002.

[133] Marie Frédérique Lartigue, Laurent Poirel, and Patrice Nordmann. Diversity of genetic environment of bla(CTX-M) genes. *FEMS microbiology letters*, 234(2):201–207, 5 2004.

[134] Christel Humeniuk, Guillaume Arlet, Valerie Gautier, Patrick Grimont, Roger Labia, and Alain Philippon. β-lactamases of Kluyvera ascorbata, probable progenitors of some plasmid-encoded CTX-M types. *Antimicrobial Agents and Chemotherapy*, 46(9):3045–3049, 2002.

[135] Brian Wahl, Katherine L. O'Brien, Adena Greenbaum, Anwesha Majumder, Li Liu, Yue Chu, Ivana Lukšić, Harish Nair, David A. McAllister, Harry Campbell, Igor

Rudan, Robert Black, and Maria Deloria Knoll. Burden of Streptococcus pneumoniae and Haemophilus influenzae type b disease in children in the era of conjugate vaccines: global, regional, and national estimates for 2000–15. *The Lancet Global Health*, 6(7):e744–e757, 7 2018.

[136] Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, 388(10053):1459–1544, 10 2016.

[137] G Laible, B G Spratt, and R Hakenbeck. Interspecies recombinational events during the evolution of altered PBP 2x genes in penicillin-resistant clinical isolates of Streptococcus pneumoniae. *Molecular Microbiology*, 5(8):1993–2002, 8 1991.

[138] C G Dowson, T J Coffey, C Kell, and R A Whiley. Evolution of penicillin resistance in Streptococcus pneumoniae; the role of Streptococcus mitis in the formation of a low affinity PBP2B in S. pneumoniae. *Molecular Microbiology*, 9(3):635–643, 8 1993.

[139] Christopher G. Dowson, Tracey J. Coffey, and Brian G. Spratt. Origin and molecular epidemiology of penicillin-binding-protein-mediated resistance to beta-lactam antibiotics. *Trends in microbiology*, 2(10):361–366, 1994.

[140] Tamsin C M Dewé, Joshua C D'Aeth, and Nicholas J Croucher. Genomic epidemiology of penicillin-non-susceptible Streptococcus pneumoniae. *Microbial genomics*, 10 2019.

[141] Alexander Tomasz, Alejandra Corso, Elena P. Severina, Gabriela Echániz-Aviles, Maria Cristina De Cunto Brandileone, Teresa Camou, Elizabeth Castañeda, Oscar Figueroa, Alicia Rossi, and José Luis Di Fabio. Molecular epidemiologic characterization of penicillin-resistant Streptococcus pneumoniae invasive pediatric isolates recovered in six Latin-American countries: an overview. PAHO/Rockefeller University Workshop. Pan American Health Organization. *Microbial Drug Resistance*, 4(3):195–207, 1998.

[142] N J Croucher, S R Harris, C Fraser, M A Quail, J Burton, M van der Linden, L Mcgee, A von Gottberg, J H Song, K S Ko, B Pichon, S Baker, C M Parry, L M Lambertsen, D Shahinas, D R Pillai, T J Mitchell, G Dougan, A Tomasz, K P Klugman, J Parkhill, W P Hanage, and S D Bentley. Rapid Pneumococcal Evolution in Response to Clinical Interventions. *Science*, 331(6016):430–434, 2011.

[143] Herman Goossens, Matus Ferech, Robert Vander Stichele, and Monique Elseviers. Outpatient antibiotic use in Europe and association with resistance: a cross-national database study. *Lancet (London, England)*, 365(9459):579–587, 2 2005.

[144] Herman Goossens. Antibiotic consumption and link to resistance. *Clinical Microbiology and Infection*, 15(SUPPL. 3):12–15, 4 2009.

[145] Nicholas G. Davies, Stefan Flasche, Mark Jit, and Katherine E. Atkins. Within-host dynamics shape antibiotic resistance in commensal bacteria. *Nature Ecology & Evolution*, 2 2019.

[146] Caroline Colijn, Ted Cohen, Christophe Fraser, William Hanage, Edward Goldstein, Noga Givon-Lavi, Ron Dagan, and Marc Lipsitch. What is the mechanism for persistent coexistence of drug-susceptible and drug-resistant strains of Streptococcus pneumoniae? *Journal of the Royal Society, Interface*, 7(47):905–19, 6 2010.

[147] Sonja Lehtinen, François Blanquart, Nicholas J Croucher, Paul Turner, Marc Lipsitch, and Christophe Fraser. Evolution of antibiotic resistance is linked to any genetic mechanism affecting bacterial duration of carriage. *Proceedings of the National Academy of Sciences of the United States of America*, 114(5):1075–1080, 1 2017.

[148] Anand Manoharan, Vikas Manchanda, Sundaram Balasubramanian, Sanjay Lalwani, Meera Modak, Sushama Bai, Ajith Vijayan, Anita Shet, Savitha Nagaraj, Sunil Karande, Gita Nataraj, Vijay N Yewale, Shrikrishna A Joshi, Ranganathan N Iyer, Mathuram Santosham, Geoffrey D Kahn, and Maria Deloria Knoll. Invasive pneumococcal disease in children aged younger than 5 years in India: a surveillance study. *The Lancet Infectious Diseases*, 17(3):305–312, 3 2017.

[149] José-María López-Lozano, Timothy Lawes, César Nebot, Arielle Beyaert, Xavier Bertrand, Didier Hocquet, Mamoon Aldeyab, Michael Scott, Geraldine Conlon-Bingham, David Farren, Gábor Kardos, Adina Fésűs, Jesús Rodríguez-Baño, Pilar Retamar, Nieves Gonzalo-Jiménez, and Ian M. Gould. A nonlinear time-series analysis approach to identify thresholds in associations between population antibiotic use and rates of resistance. *Nature Microbiology*, page 1, 4 2019.

[150] Nicholas G. Davies, Stefan Flasche, Mark Jit, and Katherine E. Atkins. Modeling the effect of vaccination on selection for antibiotic resistance in Streptococcus pneumoniae. *Science Translational Medicine*, 13(606):8690, 8 2021.

[151] Marc Lipsitch, Caroline Colijn, Ted Cohen, William P Hanage, and Christophe Fraser. No coexistence for free: Neutral null models for multistrain pathogens. *Epidemics*, 1(1):2–13, 2009.

[152] Sarah Cobey, Edward B Baskerville, Caroline Colijn, William Hanage, Christophe Fraser, and Marc Lipsitch. Host population structure and treatment frequency maintain balancing selection on drug resistance. *Journal of the Royal Society, Interface*, 14(133), 2017.

[153] François Blanquart, Sonja Lehtinen, Marc Lipsitch, and Christophe Fraser. The evolution of antibiotic resistance in a structured host population. *Journal of the Royal Society, Interface*, 15(143):20180040, 6 2018.

[154] Sonja Lehtinen, Sonja Lehtinen, Claire Chewapreecha, John Lees, William P. Hanage, Marc Lipsitch, Nicholas J. Croucher, Paul Turner, Stephen D. Bentley, Christophe Fraser, and Rafał J. Mostowy. Horizontal gene transfer rate is not the primary determinant of observed antibiotic resistance frequencies in streptococcus pneumonia. *Science Advances*, 6(21), 5 2020.

[155] Chrispin Chaguza, Cheryl P. Andam, Simon R. Harris, Jennifer E. Cornick, Marie Yang, Laura Bricio-Moreno, Arox W. Kamng'ona, Julian Parkhill, Neil French, Robert S. Heyderman, Aras Kadioglu, Dean B. Everett, Stephen D. Bentley, and William P. Hanage. Recombination in Streptococcus pneumoniae lineages increase with carriage duration and size of the polysaccharide capsule. *mBio*, 7(5), 9 2016.

[156] Sonja Lehtinen, Claire Chewapreecha, John Lees, William P Hanage, Marc Lipsitch, Nicholas J Croucher, Stephen D Bentley, Paul Turner, Christophe Fraser, and Rafał J Mostowy. Horizontal gene transfer rate is not the primary determinant of observed antibiotic resistance frequencies in Streptococcus pneumoniae. *Science Advances*, 6(21):eaaz6137, 5 2020.

[157] I. M. HASTINGS. Complex dynamics and stability of resistance to antimalarial drugs. *Parasitology*, 132(05):615–24, 5 2006.

[158] Michael Baym, Laura K. Stone, and Roy Kishony. Multidrug evolutionary strategies to reverse antibiotic resistance. *Science*, 351(6268), 1 2016.

[159] François Blanquart. Evolutionary epidemiology models to predict the dynamics of antibiotic resistance. *Evolutionary Applications*, 12(3):365–383, 3 2019.

[160] Hsiao-Han Chang, Ted Cohen, Yonatan H. Grad, William P. Hanage, Thomas F. O'Brien, and Marc Lipsitch. Origin and proliferation of multiple-drug resistance in bacterial pathogens. *Microbiology and molecular biology reviews : MMBR*, 79(1):101–116, 3 2015.

[161] David M.P. De Oliveira, Brian M. Forde, Timothy J. Kidd, Patrick N.A. Harris, Mark A. Schembri, Scott A. Beatson, David L. Paterson, and Mark J. Walker. Antimicrobial Resistance in ESKAPE Pathogens. *Clinical Microbiology Reviews*, 33(3), 2020.

[162] Mansura S. Mulani, Ekta E. Kamble, Shital N. Kumkar, Madhumita S. Tawre, and Karishma R. Pardesi. Emerging Strategies to Combat ESKAPE Pathogens in the Era of Antimicrobial Resistance: A Review. *Frontiers in Microbiology*, 10(APR):539, 2019.

[163] Raspail Carrel Founou, Luria Leslie Founou, and Sabiha Yusuf Essack. Clinical and economic impact of antibiotic resistance in developing countries: A systematic review and meta-analysis. *PLoS ONE*, 12(12), 12 2017.

[164] Sonja Lehtinen, François Blanquart, Marc Lipsitch, and Christophe Fraser. On the evolutionary ecology of multidrug resistance in bacteria. *PLOS Pathogens*, 15(5):e1007763, 5 2019.

[165] Laura J.V. Piddock. Multidrug-resistance efflux pumps ? not just for resistance. *Nature Reviews Microbiology 2006 4:8*, 4(8):629–636, 8 2006.

[166] David L. Paterson and Robert A. Bonomo. Extended-Spectrum β-Lactamases: a Clinical Update. *Clinical Microbiology Reviews*, 18(4):657, 10 2005.

[167] M. Rozwandowicz, M. S.M. Brouwer, J. Fischer, J. A. Wagenaar, B. Gonzalez-Zorn, B. Guerra, D. J. Mevius, and J. Hordijk. Plasmids carrying antimicrobial resistance genes in Enterobacteriaceae. *Journal of Antimicrobial Chemotherapy*, 73(5):1121–1137, 5 2018.

[168] Christopher B. Ford, Rupal R. Shah, Midori Kato Maeda, Sebastien Gagneux, Megan B. Murray, Ted Cohen, James C. Johnston, Jennifer Gardy, Marc Lipsitch, and Sarah M. Fortune. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nature genetics*, 45(7):784–790, 7 2013.

[169] Suzanne M. Hingley-Wilson, Rosalyn Casey, David Connell, Samuel Bremang, Jason T. Evans, Peter M. Hawkey, Grace E. Smith, Annette Jepson, Stuart Philip, Onn Min Kon, and Ajit Lalvani. Undetected Multidrug-Resistant Tuberculosis Amplified by First-line Therapy in Mixed Infection. *Emerging Infectious Diseases*, 19(7):1138, 7 2013.

[170] Eliott Jacopin, Sonja Lehtinen, Florence Débarre, and François Blanquart. Factors favouring the evolution of multidrug resistance in bacteria. *Journal of the Royal*

Society Interface, 17(168), 7 2020.

[171] David V. McLeod and Sylvain Gandon. Understanding the evolution of multiple drug resistance in structured populations. *eLife*, 10, 6 2021.

[172] M. S. Niederman. Impact of antibiotic resistance on clinical outcomes and the cost of care. *Critical care medicine*, 29(4 Suppl), 2001.

[173] Andrew J. Stewardson, A. Allignol, J. Beyersmann, N. Graves, M. Schumacher, R. Meyer, E. Tacconelli, G. De Angelis, C. Farina, F. Pezzoli, X. Bertrand, H. Gbaguidi-Haore, J. Edgeworth, O. Tosas, J. A. Martinez, M. P. Ayala-Blanco, A. Pan, A. Zoncada, C. A. Marwick, D. Nathwani, H. Seifert, N. Hos, S. Hagel, M. Pletz, S. Harbarth, Cristina Masuet-Aumatell, Marta Banqué Navarro, and Chiara Falcone. The health and economic burden of bloodstream infections caused by antimicrobial-susceptible and non-susceptible Enterobacteriaceae and Staphylococcus aureus in European hospitals, 2010 and 2011: a multicentre retrospective cohort study. *Eurosurveillance*, 21(33), 8 2016.

[174] Brian Brown and Paul Crawford. 'Post antibiotic apocalypse': discourses of mutation in narratives of MRSA. *Sociology of Health & Illness*, 31(4):508–524, 5 2009.

[175] Gabriela Capurro. "Superbugs" in the Risk Society: Assessing the Reflexive Function of North American Newspaper Coverage of Antimicrobial Resistance. *SAGE Open*, 10(1):2158244020901800, 2020.

[176] Cédric Abat, Pierre Edouard Fournier, Marie Thérèse Jimeno, Jean Marc Rolain, and Didier Raoult. Extremely and pandrug-resistant bacteria extra-deaths: myth or reality? *European Journal of Clinical Microbiology and Infectious Diseases*, 37(9):1687–1697, 9 2018.

[177] Diamantis P. Kofteridis, Angeliki M. Andrianaki, Sofia Maraki, Anna Mathioudaki, Marina Plataki, Christina Alexopoulou, Petros Ioannou, George Samonis, and Antonis Valachis. Treatment pattern, prognostic factors, and outcome in patients with infection due to pan-drug-resistant gram-negative bacteria. *European Journal of Clinical Microbiology and Infectious Diseases*, 39(5):965–970, 5 2020.

[178] Sara E. Cosgrove, George Sakoulas, Eli N. Perencevich, Mitchell J. Schwaber, Adolf W. Karchmer, and Yehuda Carmeli. Comparison of mortality associated with methicillin-resistant and methicillin-susceptible Staphylococcus aureus bacteremia: A meta-analysis. *Clinical Infectious Diseases*, 36(1):53–59, 1 2003.

[179] Carlos A. DiazGranados, Shanta M. Zimmer, Mitchel Klein, and John A. Jernigan. Comparison of mortality associated with vancomycin-resistant and vancomycin-susceptible enterococcal bloodstream infections: A meta-analysis. *Clinical Infectious Diseases*, 41(3):327–333, 8 2005.

[180] Wouter C. Rottier, Heidi S.M. Ammerlaan, and Marc J.M. Bonten. Effects of confounders and intermediates on the association of bacteraemia caused by extended-spectrum β-lactamase-producing Enterobacteriaceae and patient outcome: a meta-analysis. *The Journal of antimicrobial chemotherapy*, 67(6):1311–1320, 6 2012.

[181] E. V. Lemos, F. P. de la Hoz, T. R. Einarson, W. F. Mcghan, E. Quevedo, C. Castañeda, and K. Kawai. Carbapenem resistance and mortality in patients with Acinetobacter baumannii infection: systematic review and meta-analysis. *Clinical Microbiology and Infection*, 20(5):416–423, 5 2014.

[182] Galo Peralta, M. Blanca Sánchez, J. Carlos Garrido, Inés De Benito, M. Eliecer Cano, Luis Martínez-Martínez, and M. Pía Roiz. Impact of antibiotic resistance and of adequate empirical antibiotic treatment in the prognosis of patients with Escherichia coli bacteraemia. *Journal of Antimicrobial Chemotherapy*, 60(4):855–863, 10 2007.

[183] Eun Jeong Joo, Cheol In Kang, Young Eun Ha, Seung Ji Kang, So Yeon Park, Doo Ryeon Chung, Kyong Ran Peck, Nam Yong Lee, and Jae Hoon Song. Risk factors for mortality in patients with Pseudomonas aeruginosa bacteremia: clinical impact of antimicrobial resistance on outcome. *Microbial drug resistance*, 17(2):305–312, 6 2011.

[184] Raúl Recio, Mikel Mancheño, Esther Viedma, Jennifer Villa, Mar a.Angeles Orellana, Jaime Lora-Tamayo, and Fernando Chaves. Predictors of mortality in bloodstream infections caused by pseudomonas aeruginosa and impact of antimicrobial resistance and bacterial virulence. *Antimicrobial Agents and Chemotherapy*, 64(2), 1 2020.

[185] Marlieke E.A. De Kraker, Martin Wolkewitz, Peter G. Davey, and Hajo Grundmann. Clinical impact of antimicrobial resistance in European hospitals: Excess mortality and length of hospital stay related to methicillin-resistant Staphylococcus aureus bloodstream infections. *Antimicrobial Agents and Chemotherapy*, 55(4):1598–1605, 4 2011.

[186] David R. Cameron, Benjamin P. Howden, and Anton Y. Peleg. The Interface Between Antibiotic Resistance and Virulence in Staphylococcus aureus and Its Impact Upon Clinical Outcomes. *Clinical Infectious Diseases*, 53(6):576–582, 9 2011.

[187] Martin E. Stryjewski, Lynda A. Szczech, Daniel K. Benjamin, Jula K. Inrig, Zeina A. Kanafani, John J. Engemann, Vivian H. Chu, Maria J. Joyce, L. Barth Reller, G. Ralph Corey, and Vance G. Fowler. Use of vancomycin or first-generation cephalosporins for the treatment of hemodialysis-dependent patients with methicillin-susceptible Staphylococcus aureus bacteremia. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 44(2):190–196, 1 2007.

[188] Daozhou Gao, Thomas M. Lietman, and Travis C. Porco. Antibiotic resistance as collateral damage: the tragedy of the commons in a two-disease setting. *Mathematical biosciences*, 263:121, 5 2015.

[189] Alex J. McCarthy and Jodi A. Lindsay. The distribution of plasmids that carry virulence and resistance genes in Staphylococcus aureus is lineage associated. *BMC microbiology*, 12, 2012.

[190] Jim O ' Neill. Antimicrobial Resistance: Tackling a crisis for the health and wealth of nations The Review on Antimicrobial Resistance Chaired b by J O'Neil, and supported by the Wellcome Trust and the UK Government. (December), 2014.

[191] Marlieke E A de Kraker, Andrew J Stewardson, and Stephan Harbarth. Will 10 Million People Die a Year due to Antimicrobial Resistance by 2050? *PLoS medicine*, 13(11):e1002184, 11 2016.

[192] Andrew F. Read and Silvie Huijben. PERSPECTIVE: Evolutionary biology and the avoidance of antimicrobial resistance. *Evolutionary Applications*, 2(1):40–51, 1 2009.

[193] Alessandro Cassini, Liselotte Diaz Högberg, Diamantis Plachouras, Annalisa Quattrocchi, Ana Hoxha, Gunnar Skov Simonsen, Mélanie Colomb-Cotinat, Mirjam E. Kretzschmar, Brecht Devleesschauwer, Michele Cecchini, Driss Ait Ouakrim, Tiago Cravo Oliveira, Marc J. Struelens, Carl Suetens, Dominique L. Monnet, Reinhild Strauss, Karl Mertens, Thomas Struyf, Boudewijn Catry, Katrien Latour, Ivan N. Ivanov, Elina G. Dobreva, Arjana Tambic Andraševic, Silvija Soprek, Ana Budimir, Niki Paphitou, Helena Žemlicková, Stefan Schytte Olsen, Ute Wolff Sönksen, Pille Märtin, Marina Ivanova, Outi Lyytikäinen, Jari Jalava, Bruno Coignard, Tim Eckmanns, Muna Abu Sin, Sebastian Haller, George L. Daikos, Achilleas Gikas, Sotirios Tsiodras, Flora Kontopidou, Ákos Tóth, Ágnes Hajdu, Ólafur Guólaugsson, Karl G. Kristinsson, Stephen Murchan, Karen Burns, Patrizio Pezzotti, Carlo Gagliotti, Uga Dumpis, Agne Liuimiene, Monique Perrin, Michael A. Borg, Sabine C. de Greeff, Jos CM Monen, Mayke BG Koek, Petter Elstrøm, Dorota Zabicka, Aleksander Deptula, Waleria Hryniewicz, Manuela Caniça, Paulo Jorge Nogueira, Paulo André Fernandes, Vera Manageiro, Gabriel A. Popescu, Roxana I. Serban, Eva Schréterová, Slavka Litvová, Mária Štefkovicová, Jana Kolman, Irena Klavs, Aleš Korošec, Belén Aracil, Angel Asensio, María Pérez-Vázquez, Hanna Billström, Sofie Larsson, Jacqui S. Reilly, Alan Johnson, and Susan Hopkins. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. *The Lancet Infectious Diseases*, 19(1):56–66, 1 2019.

[194] Christopher JL Murray, Kevin Shunji Ikuta, Fablina Sharara, Lucien Swetschinski, Gisela Robles Aguilar, Authia Gray, Chieh Han, Catherine Bisignano, Puja Rao, Eve Wool, Sarah C. Johnson, Annie J. Browne, Michael Give Chipeta, Frederick Fell, Sean Hackett, Georgina Haines-Woodhouse, Bahar H. Kashef Hamadani, Emmanuelle A.P. Kumaran, Barney McManigal, Ramesh Agarwal, Samuel Akech, Samuel Albertson, John Amuasi, Jason Andrews, Aleskandr Aravkin, Elizabeth Ashley, Freddie Bailey, Stephen Baker, Buddha Basnyat, Adrie Bekker, Rose Bender, Adhisivam Bethou, Julia Bielicki, Suppawat Boonkasidecha, James Bukosia, Cristina Carvalheiro, Carlos Castañeda-Orjuela, Vilada Chansamouth, Suman Chaurasia, Sara Chiurchiù, Fazle Chowdhury, Aislinn J. Cook, Ben Cooper, Tim R. Cressey, Elia Criollo-Mora, Matthew Cunningham, Saffiatou Darboe, Nicholas P.J. Day, Maia De Luca, Klara Dokova, Angela Dramowski, Susanna J. Dunachie, Tim Eckmanns, Daniel Eibach, Amir Emami, Nicholas Feasey, Natasha Fisher-Pearson, Karen Forrest, Denise Garrett, Petra Gastmeier, Ababi Zergaw Giref, Rachel Claire Greer, Vikas Gupta, Sebastian Haller, Andrea Haselbeck, Simon I. Hay, Marianne Holm, Susan Hopkins, Kenneth C. Iregbu, Jan Jacobs, Daniel Jarovsky, Fatemeh Javanmardi, Meera Khorana, Niranjan Kissoon, Elsa Kobeissi, Tomislav Kostyanev, Fiorella Krapp, Ralf Krumkamp, Ajay Kumar, Hmwe Hmwe Kyu, Cherry Lim, Direk Limmathurotsakul, Michael James Loftus, Miles Lunn, Jianing Ma, Neema Mturi, Tatiana Munera-Huertas, Patrick Musicha, Marisa Marcia Mussi-Pinhata, Tomoka Nakamura, Ruchi Nanavati, Sushma Nangia, Paul Newton, Chanpheaktra Ngoun, Amanda Novotney, Davis Nwakanma, Christina W. Obiero, Antonio Olivas-Martinez, Piero Olliaro, Ednah Ooko, Edgar Ortiz-Brizuela, Anton Yariv Peleg, Carlo Perrone, Nishad Plakkal, Alfredo Ponce-de Leon, Mathieu Raad, Tanusha Ramdin, Amy Riddell, Tamalee Roberts, Julie Victoria Robotham, Anna Roca, Kristina E. Rudd, Neal Russell, Jesse Schnall, John Anthony Gerard Scott, Madhusudhan Shivamallappa, Jose Sifuentes-Osornio, Nico-

las Steenkeste, Andrew James Stewardson, Temenuga Stoeva, Nidanuch Tasak, Areerat Thaiprakong, Guy Thwaites, Claudia Turner, Paul Turner, H. Rogier van Doorn, Sithembiso Velaphi, Avina Vongpradith, Huong Vu, Timothy Walsh, Seymour Waner, Tri Wangrangsimakul, Teresa Wozniak, Peng Zheng, Benn Sartorius, Alan D. Lopez, Andy Stergachis, Catrin Moore, Christiane Dolecek, and Mohsen Naghavi. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, 399(10325):629–655, 2 2022.

[195] Cristiana Abbafati, Kaja M. Abbas, Mohsen Abbasi-Kangevari, Foad Abd-Allah, Ahmed Abdelalim, Mohammad Abdollahi, Ibrahim Abdollahpour, Kedir Hussein Abegaz, Hassan Abolhassani, Victor Aboyans, ..., Kairat Davletov, Arash Ziapour, Stefania Mondello, Stephen S. Lim, Christopher J.L. Murray, Taweewat Wiangkham, and Saeed Amini. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, 396(10258):1204–1222, 10 2020.

[196] Anne Mills. Health Care Systems in Low- and Middle-Income Countries. *New England Journal of Medicine*, 370(6):552–557, 2 2014.

[197] Gaëtan Gavazzi, Francois Herrmann, and Karl Heinz Krause. Aging and infectious diseases in the developing world. *Clinical Infectious Diseases*, 39(1):83–91, 7 2004.

[198] Marlee Tichenor, Devi Sridhar, Peter Byass, and Carla Abouzahr. Metric partnerships: global burden of disease estimates within the World Bank, the World Health Organisation and the Institute for Health Metrics and Evaluation. *Wellcome Open Research 2020 4:35*, 4:35, 1 2020.

[199] Linsey McGoey. Philanthrocapitalism and the Separation of Powers. *Annual Review of Law and Social Science*, 17(1):391–409, 2021.

[200] Robert D. Fleischmann, Mark D. Adams, Owen White, Rebecca A. Clayton, Ewen F. Kirkness, Anthony R. Kerlavage, Carol J. Bult, Jean Francois Tomb, Brian A. Dougherty, Joseph M. Merrick, Keith McKenney, Granger Sutton, Will FitzHugh, Chris Fields, Jeannine D. Gocayne, John Scott, Robert Shirley, Li Ing Liu, Anna Glodek, Jenny M. Kelley, Janice F. Weidman, Cheryl A. Phillips, Tracy Spriggs, Eva Hedblom, Matthew D. Cotton, Teresa R. Utterback, Michael C. Hanna, David T. Nguyen, Deborah M. Saudek, Rhonda C. Brandon, Leah D. Fine, Janice L. Fritchman, Joyce L. Fuhrmann, N. S.M. Geoghagen, Cheryl L. Gnehm, Lisa A. McDonald, Keith V. Small, Claire M. Fraser, Hamilton O. Smith, and J. Craig Venter. Whole-Genome Random Sequencing and Assembly of Haemophilus influenzae Rd. *Science*, 269(5223):496–512, 1995.

[201] Nicholas J. Loman and Mark J. Pallen. Twenty years of bacterial genome sequencing. *Nature Reviews Microbiology 2015 13:12*, 13(12):787–794, 11 2015.

[202] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463, 1977.

[203] Eric S. Donkor. Sequencing of Bacterial Genomes: Principles and Insights into Pathogenesis and Development of Antibiotics. *Genes*, 4(4):556, 2013.

[204] Claire M. Fraser and Robert D. Fleischmann. Strategies for whole microbial genome sequencing and analysis. *Electrophoresis*, 18(8):1207–1216, 8 1997.

[205] Michael L. Metzker. Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1):31–46, 12 2009.

[206] Erwin L. van Dijk, Yan Jaszczyszyn, Delphine Naquin, and Claude Thermes. The Third Revolution in Sequencing Technology. *Trends in Genetics*, 34(9):666–681, 9 2018.

[207] Shanika L. Amarasinghe, Shian Su, Xueyi Dong, Luke Zappia, Matthew E. Ritchie, and Quentin Gouil. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology 2020 21:1*, 21(1):1–16, 2 2020.

[208] Miten Jain, Hugh E. Olsen, Benedict Paten, and Mark Akeson. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology 2016 17:1*, 17(1):1–11, 11 2016.

[209] Eric S. Tvedte, Mark Gasser, Benjamin C. Sparklin, Jane Michalski, Carl E. Hjelmen, J. Spencer Johnston, Xuechu Zhao, Robin Bromley, Luke J. Tallon, Lisa Sadzewicz, David A. Rasko, and Julie C. Dunning Hotopp. Comparison of long-read sequencing technologies in interrogating bacteria and fly genomes. *G3: Genes|Genomes|Genetics*, 11(6), 6 2021.

[210] Aaron M. Wenger, Paul Peluso, William J. Rowell, Pi Chuan Chang, Richard J. Hall, Gregory T. Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D. Olson, Armin Töpfer, Michael Alonge, Medhat Mahmoud, Yufeng Qian, Chen Shan Chin, Adam M. Phillippy, Michael C. Schatz, Gene Myers, Mark A. DePristo, Jue Ruan, Tobias Marschall, Fritz J. Sedlazeck, Justin M. Zook, Heng Li, Sergey Koren, Andrew Carroll, David R. Rank, and Michael W. Hunkapiller. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature biotechnology*, 37(10):1155–1162, 10 2019.

[211] Jörn Petersen, John Vollmers, Victoria Ringel, Henner Brinkmann, Claire Ellebrandt-Sperling, Cathrin Spröer, Alexandra M. Howat, J. Colin Murrell, and Anne-Kristin Kaster. A marine plasmid hitchhiking vast phylogenetic and geographic distances. *Proceedings of the National Academy of Sciences*, page 201905878, 9 2019.

[212] N. R. Faria, J. Quick, I. M. Claro, J. Thézé, J. G. De Jesus, M. Giovanetti, M. U.G. Kraemer, S. C. Hill, A. Black, A. C. Da Costa, L. C. Franco, S. P. Silva, C. H. Wu, J. Raghwani, S. Cauchemez, L. Du Plessis, M. P. Verotti, W. K. De Oliveira, E. H. Carmo, G. E. Coelho, A. C.F.S. Santelli, L. C. Vinhal, C. M. Henriques, J. T. Simpson, M. Loose, K. G. Andersen, N. D. Grubaugh, S. Somasekar, C. Y. Chiu, J. E. Muñoz-Medina, C. R. Gonzalez-Bonilla, C. F. Arias, L. L. Lewis-Ximenez, S. A. Baylis, A. O. Chieppe, S. F. Aguiar, C. A. Fernandes, P. S. Lemos, B. L.S. Nascimento, H. A.O. Monteiro, I. C. Siqueira, M. G. De Queiroz, T. R. De Souza, J. F. Bezerra, M. R. Lemos, G. F. Pereira, D. Loudal, L. C. Moura, R. Dhalia, R. F. França, T. Magalhães, E. T. Marques, T. Jaenisch, G. L. Wallau, M. C. De Lima, V. Nascimento, E. M. De Cerqueira, M. M. De Lima, D. L. Mascarenhas, J. P.Moura Neto, A. S. Levin, T. R. Tozetto-Mendoza, S. N. Fonseca, M. C. Mendes-Correa, F. P. Milagres, A. Segurado, E. C. Holmes, A. Rambaut, T. Bedford, M. R.T. Nunes, E. C. Sabino, L. C.J. Alcantara, N. J. Loman, and O. G. Pybus. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature 2017 546:7658*, 546(7658):406–410, 5 2017.

[213] Joshua Quick, Nathan D. Grubaugh, Steven T. Pullan, Ingra M. Claro, Andrew D. Smith, Karthik Gangavarapu, Glenn Oliveira, Refugio Robles-Sikisaka, Thomas F. Rogers, Nathan A. Beutler, Dennis R. Burton, Lia Laura Lewis-Ximenez, Jaqueline Goes De Jesus, Marta Giovanetti, Sarah C. Hill, Allison Black, Trevor Bedford, Miles W. Carroll, Marcio Nunes, Luiz Carlos Alcantara, Ester C. Sabino, Sally A. Baylis, Nuno R. Faria, Matthew Loose, Jared T. Simpson, Oliver G. Pybus, Kristian G. Andersen, and Nicholas J. Loman. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nature protocols*, 12(6):1261, 6 2017.

[214] Nuno R. Faria, Thomas A. Mellan, Charles Whittaker, Ingra M. Claro, Darlan Da S. Candido, Swapnil Mishra, Myuki A.E. Crispim, Flavia C.S. Sales, Iwona Hawryluk, John T. McCrone, Ruben J.G. Hulswit, Lucas A.M. Franco, Mariana S. Ramundo, Jaqueline G. De Jesus, Pamela S. Andrade, Thais M. Coletti, Giulia M. Ferreira, Camila A.M. Silva, Erika R. Manuli, Rafael H.M. Pereira, Pedro S. Peixoto, Moritz U.G. Kraemer, Nelson Gaburo, Cecilia Da C. Camilo, Henrique Hoeltgebaum, William M. Souza, Esmenia C. Rocha, Leandro M. De Souza, Mariana C. De Pinho, Leonardo J.T. Araujo, Frederico S.V. Malta, Aline B. De Lima, Joice Do P. Silva, Danielle A.G. Zauli, Alessandro C. Alessandro, Ricardo P. Schnekenberg, Daniel J. Laydon, Patrick G.T. Walker, Hannah M. Schlüter, Ana L.P. Dos Santos, Maria S. Vidal, Valentina S. Del Caro, Rosinaldo M.F. Filho, Helem M. Dos Santos, Renato S. Aguiar, José L. Proença-Modena, Bruce Nelson, James A. Hay, Mélodie Monod, Xenia Miscouridou, Helen Coupland, Raphael Sonabend, Michaela Vollmer, Axel Gandy, Carlos A. Prete, Vitor H. Nascimento, Marc A. Suchard, Thomas A. Bowden, Sergei L.K. Pond, Chieh Hsi Wu, Oliver Ratmann, Neil M. Ferguson, Christopher Dye, Nick J. Loman, Philippe Lemey, Andrew Rambaut, Nelson A. Fraiji, Maria Do P.S.S. Carvalho, Oliver G. Pybus, Seth Flaxman, Samir Bhatt, and Ester C. Sabino. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science*, 372(6544), 5 2021.

[215] Fabienne Antunes Ferreira, Karin Helmersen, Tina Visnovska, Silje Bakken Jørgensen, and Hege Vangstein Aamot. Rapid nanopore-based DNA sequencing protocol of antibiotic-resistant bacteria for use in surveillance and outbreak investigation. *Microbial genomics*, 7(4), 2021.

[216] Kim Judge, Simon R. Harris, Sandra Reuter, Julian Parkhill, and Sharon J. Peacock. Early insights into the potential of the Oxford Nanopore MinION for the detection of antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy*, 70(10):2775, 10 2015.

[217] James H. Jorgensen and Mary Jane Ferraro. Antimicrobial susceptibility testing: General principles and contemporary practices. *Clinical Infectious Diseases*, 26(4):973–980, 4 1998.

[218] Alex Van Belkum and W. Michael Dunne. Next-generation antimicrobial susceptibility testing. *Journal of Clinical Microbiology*, 51(7):2018–2024, 2013.

[219] EUCAST: Clinical breakpoints and dosing of antibiotics - https://www.eucast.org/clinical_breakpoints/ Date accessed: 14/01/2021.

[220] Alex van Belkum, Carey Ann D. Burnham, John W.A. Rossen, Frederic Mallard, Olivier Rochas, and William Michael Dunne. Innovative and rapid antimicrobial susceptibility testing systems. *Nature Reviews Microbiology 2020 18:5*, 18(5):299–311, 2 2020.

[221] Nicole E. Wheeler, Leonor Sánchez-Busó, Silvia Argimón, and Benjamin Jeffrey. Lean, mean, learning machines. *Nature Reviews Microbiology 2020 18:5*, 18(5):266–266, 3 2020.

[222] Emma Jonasson, Erika Matuschek, and Gunnar Kahlmeter. The EUCAST rapid disc diffusion method for antimicrobial susceptibility testing directly from positive blood culture bottles. *The Journal of antimicrobial chemotherapy*, 75(4):968–978, 4 2020.

[223] Iain Dickson. Sequencing the unculturable majority. *Nature Research 2021*, 2 2021.

[224] Jay Shendure, Shankar Balasubramanian, George M. Church, Walter Gilbert, Jane Rogers, Jeffery A. Schloss, and Robert H. Waterston. DNA sequencing at 40: past, present and future. *Nature*, 550(7676), 10 2017.

[225] Manish Boolchandani, Alaric W. D'Souza, and Gautam Dantas. Sequencing-based methods and resources to study antimicrobial resistance. *Nature reviews. Genetics*, 20(6):356–370, 6 2019.

[226] Allison L. Hicks, Nicole Wheeler, Leonor Sánchez-Busó, Jennifer L. Rakeman, Simon R. Harris, and Yonatan H. Grad. Evaluation of parameters affecting performance and reliability of machine learning-based antibiotic susceptibility testing from whole genome sequencing data. *PLOS Computational Biology*, 15(9):e1007349, 2019.

[227] Yang Yang, Katherine E. Niehaus, Timothy M. Walker, Zamin Iqbal, A. Sarah Walker, Daniel J. Wilson, Tim E.A. Peto, Derrick W. Crook, E. Grace Smith, Tingting Zhu, and David A. Clifton. Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. *Bioinformatics*, 34(10):1666–1671, 5 2018.

[228] Michelle Su, Sarah W. Satola, and Timothy D. Read. Genome-Based Prediction of Bacterial Antibiotic Resistance. *Journal of Clinical Microbiology*, 57(3), 3 2019.

[229] Jean Pierre Flandrois, Gérard Lina, and Oana Dumitrescu. MUBII-TB-DB: a database of mutations associated with antibiotic resistance in Mycobacterium tuberculosis. *BMC bioinformatics*, 15(1), 4 2014.

[230] Saurav B. Saha, Vishwas Uttam, and Vivek Verma. u-CARE: user-friendly Comprehensive Antibiotic resistance Repository of Escherichia coli. *Journal of Clinical Pathology*, 68(8):648–651, 8 2015.

[231] Quan K. Thai, Fabian Bös, and Jürgen Pleiss. The Lactamase Engineering Database: a critical survey of TEM sequences in public databases. *BMC genomics*, 10:390, 8 2009.

[232] Brian P. Alcock, Amogelang R. Raphenya, Tammy T.Y. Lau, Kara K. Tsang, Mégane Bouchard, Arman Edalatmand, William Huynh, Anna Lisa V. Nguyen, Annie A. Cheng, Sihan Liu, Sally Y. Min, Anatoly Miroshnichenko, Hiu Ki Tran, Rafik E. Werfalli, Jalees A. Nasir, Martins Oloni, David J. Speicher, Alexandra Florescu, Bhavya Singh, Mateusz Faltyn, Anastasia Hernandez-Koutoucheva, Arjun N. Sharma, Emily Bordeleau, Andrew C. Pawlowski, Haley L. Zubyk, Damion Dooley, Emma Griffiths, Finlay Maguire, Geoff L. Winsor, Robert G. Beiko, Fiona S.L. Brinkman, William W.L. Hsiao, Gary V. Domselaar, and Andrew G. McArthur. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic acids research*, 48(D1):D517–D525, 1 2020.

[233] Valeria Bortolaia, Rolf S. Kaas, Etienne Ruppe, Marilyn C. Roberts, Stefan Schwarz, Vincent Cattoir, Alain Philippon, Rosa L. Allesoe, Ana Rita Rebelo, Alfred Ferrer Florensa, Linda Fagelhauer, Trinad Chakraborty, Bernd Neumann, Guido Werner, Jennifer K. Bender, Kerstin Stingl, Minh Nguyen, Jasmine Coppens, Basil Britto Xavier, Surbhi Malhotra-Kumar, Henrik Westh, Mette Pinholt, Muna F. Anjum, Nicholas A. Duggett, Isabelle Kempf, Suvi Nykäsenoja, Satu Olkkola, Kinga Wieczorek, Ana Amaro, Lurdes Clemente, Joël Mossong, Serge Losch, Catherine Ragimbeau, Ole Lund, and Frank M. Aarestrup. ResFinder 4.0 for predictions of phenotypes from genotypes. *The Journal of antimicrobial chemotherapy*, 75(12):3491–3500, 12 2020.

[234] Alfred Ferrer Florensa, Rolf Sommer Kaas, Philip Thomas Lanken Conradsen Clausen, Derya Aytan-Aktug, and Frank M. Aarestrup. ResFinder – an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. *Microbial Genomics*, 8(1), 2022.

[235] Norhan Mahfouz, Inês Ferreira, Stephan Beisken, Arndt von Haeseler, and Andreas E. Posch. Large-scale assessment of antimicrobial resistance marker databases for genetic phenotype prediction: a systematic review. *Journal of Antimicrobial Chemotherapy*, 75(11):3099–3108, 11 2020.

[236] Michael Inouye, Harriet Dashnow, Lesley Ann Raven, Mark B. Schultz, Bernard J. Pope, Takehiro Tomita, Justin Zobel, and Kathryn E. Holt. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome medicine*, 6(11), 11 2014.

[237] Amy Mason, Dona Foster, Phelim Bradley, Tanya Golubchik, Michel Doumith, N. Claire Gordon, Bruno Pichon, Zamin Iqbal, Peter Staves, Derrick Crook, A. Sarah Walker, Angela Kearns, and Tim Peto. Accuracy of Different Bioinformatics Methods in Detecting Antibiotic Resistance and Virulence Factors from Staphylococcus aureus Whole-Genome Sequences. *Journal of Clinical Microbiology*, 56(9), 9 2018.

[238] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods 2012 9:4*, 9(4):357–359, 3 2012.

[239] Philip T.L.C. Clausen, Frank M. Aarestrup, and Ole Lund. Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics*, 19(1), 8 2018.

[240] Rachel M. Colquhoun, Michael B. Hall, Leandro Lima, Leah W. Roberts, Kerri M. Malone, Martin Hunt, Brice Letcher, Jane Hawkey, Sophie George, Louise Pankhurst, and Zamin Iqbal. Pandora: nucleotide-resolution bacterial pangenomics with reference graphs. *Genome Biology*, 22(1):1–30, 12 2021.

[241] Phelim Bradley, N. Claire Gordon, Timothy M. Walker, Laura Dunn, Simon Heys, Bill Huang, Sarah Earle, Louise J. Pankhurst, Luke Anson, Mariateresa De Cesare, Paolo Piazza, Antonina A. Votintseva, Tanya Golubchik, Daniel J. Wilson, David H. Wyllie, Roland Diel, Stefan Niemann, Silke Feuerriegel, Thomas A. Kohl, Nazir Ismail, Shaheed V. Omar, E. Grace Smith, David Buck, Gil McVean, A. Sarah Walker, Tim E.A. Peto, Derrick W. Crook, and Zamin Iqbal. Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nature Communications*, 6, 12 2015.

[242] Will P.M. Rowe and Martyn D. Winn. Indexed variation graphs for efficient and accurate resistome profiling. *Bioinformatics*, 34(21):3601, 11 2018.

[243] Michael Feldgarden, Vyacheslav Brover, Narjol Gonzalez-Escalona, Jonathan G. Frye, Julie Haendiges, Daniel H. Haft, Maria Hoffmann, James B. Pettengill, Arjun B. Prasad, Glenn E. Tillman, Gregory H. Tyson, and William Klimke. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Scientific reports*, 11(1), 12 2021.

[244] Sean R. Eddy. Profile hidden Markov models. *Bioinformatics (Oxford, England)*, 14(9):755–763, 1998.

[245] M. Gordon, E. Yakunin, L. Valinsky, V. Chalifa-Caspi, and J. Moran-Gilad. A bioinformatics tool for ensuring the backwards compatibility of Legionella pneumophila typing in the genomic era. *Clinical Microbiology and Infection*, 23(5):306–310, 5 2017.

[246] M. J. Ellington, O. Ekelund, F. M. Aarestrup, R. Canton, M. Doumith, C. Giske, H. Grundman, H. Hasman, M. T.G. Holden, K. L. Hopkins, J. Iredell, G. Kahlmeter, C. U. Köser, A. MacGowan, D. Mevius, M. Mulvey, T. Naas, T. Peto, J. M. Rolain, Samuelsen, and N. Woodford. The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. *Clinical Microbiology and Infection*, 23(1):2–22, 1 2017.

[247] Sarah Goldstein, Lidia Beka, Joerg Graf, and Jonathan L. Klassen. Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. *BMC Genomics*, 20(1):1–17, 1 2019.

[248] Jay S. Ghurye, Victoria Cepeda-Espinoza, and Mihai Pop. Focus: Microbiome: Metagenomic Assembly: Overview, Challenges and Applications. *The Yale Journal of Biology and Medicine*, 89(3):353, 9 2016.

[249] Timothy M. Walker, Thomas A. Kohl, Shaheed V. Omar, Jessica Hedge, Carlos Del Ojo Elias, Phelim Bradley, Zamin Iqbal, Silke Feuerriegel, Katherine E. Niehaus, Daniel J. Wilson, David A. Clifton, Georgia Kapatai, Camilla L.C. Ip, Rory Bowden, Francis A. Drobniewski, Caroline Allix-Béguec, Cyril Gaudin, Julian Parkhill, Roland Diel, Philip Supply, Derrick W. Crook, E. Grace Smith, A. Sarah Walker, Nazir Ismail, Stefan Niemann, Tim E.A. Peto, Jim Davies, Charles Crichton, Milind Acharya, Laura Madrid-Marquez, David Eyre, David Wyllie, Tanya Golubchik, and Melinda Munang. Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. *The Lancet. Infectious diseases*, 15(10):1193–1202, 10 2015.

[250] Leen Rigouts, Mourad Gumusboga, Willem Bram De Rijk, Elie Nduwamahoro, Cécile Uwizeye, Bouke De Jong, and Armand Van Deun. Rifampin resistance missed in automated liquid culture system for Mycobacterium tuberculosis isolates with specific rpoB mutations. *Journal of clinical microbiology*, 51(8):2641–2645, 8 2013.

[251] N. C. Gordon, J. R. Price, K. Cole, R. Everitt, M. Morgan, J. Finney, A. M. Kearns, B. Pichon, B. Young, D. J. Wilson, M. J. Llewelyn, J. Paul, T. E.A. Peto, D. W. Crook, A. S. Walker, and T. Golubchik. Prediction of Staphylococcus aureus antimicrobial resistance by whole-genome sequencing. *Journal of clinical microbiology*, 52(4):1182–1191, 2014.

[252] Alimuddin Zumla, Jaffar A. Al-Tawfiq, Virve I. Enne, Mike Kidd, Christian Drosten, Judy Breuer, Marcel A. Muller, David Hui, Markus Maeurer, Matthew Bates, Peter Mwaba, Rafaat Al-Hakeem, Gregory Gray, Philippe Gautret, Abdullah A. Al-Rabeeah, Ziad A. Memish, and Vanya Gant. Rapid point of care diagnostic tests for viral and bacterial respiratory tract infections—needs, advances, and future prospects. *The Lancet. Infectious Diseases*, 14(11):1123, 11 2014.

[253] Magali Jaillard, Mattia Palmieri, Alex van Belkum, and Pierre Mahé. Interpreting k-mer–based signatures for antibiotic resistance prediction. *GigaScience*, 9(10):1–16, 10 2020.

[254] Sebastian M. Gygli, Sonia Borrell, Andrej Trauner, and Sebastien Gagneux. Antimicrobial resistance in Mycobacterium tuberculosis: mechanistic and evolutionary perspectives. *FEMS Microbiology Reviews*, 41(3):354–373, 5 2017.

[255] Johan Pensar, Santeri Puranen, Brian Arnold, Neil MacAlasdair, Juri Kuronen, Gerry Tonkin-Hill, Maiju Pesonen, Yingying Xu, Aleksi Sipola, Leonor Sánchez-Busó, John A. Lees, Claire Chewapreecha, Stephen D. Bentley, Simon R. Harris, Julian Parkhill, Nicholas J. Croucher, and Jukka Corander. Genome-wide epistasis and co-selection study using mutual information. *Nucleic Acids Research*, 47(18):e112–e112, 10 2019.

[256] Jee In Kim, Finlay Maguire, Kara K. Tsang, Theodore Gouliouris, Sharon J. Peacock, Tim A. McAllister, Andrew G. McArthur, and Robert G. Beiko. Machine Learning for Antimicrobial Resistance Prediction: Current Practice, Limitations, and Clinical Perspective. *Clinical Microbiology Reviews*, 5 2022.

[257] Courtney Hebert, Courtney Hebert, Yuan Gao, Protiva Rahman, Courtney Dewart, Mark Lustberg, Preeti Pancholi, Kurt Stevenson, Nirav S. Shah, Erinn M. Hadeh, and Erinn M. Hadeh. Prediction of antibiotic susceptibility for urinary tract infection in a hospital setting. *Antimicrobial Agents and Chemotherapy*, 64(7), 7 2020.

[258] Finlay Maguire, Muhammad Attiq Rehman, Catherine Carrillo, Moussa S. Diarra, and Robert G. Beiko. Identification of Primary Antimicrobial Resistance Drivers in Agricultural Nontyphoidal Salmonella enterica Serovars by Using Machine Learning. *mSystems*, 4(4), 8 2019.

[259] Danesh Moradigaravand, Martin Palm, Anne Farewell, Ville Mustonen, Jonas Warringer, and Leopold Parts. Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data. *PLoS computational biology*, 14(12), 12 2018.

[260] Marcus Nguyen, Robert Olson, Maulik Shukla, Margo VanOeffelen, and James J. Davis. Predicting antimicrobial resistance using conserved genes. *PLOS Computational Biology*, 16(10):e1008319, 10 2020.

[261] Pieter Jan Van Camp, David B. Haslam, and Aleksey Porollo. Prediction of Antimicrobial Resistance in Gram-Negative Bacteria From Whole-Genome Sequencing Data. *Frontiers in Microbiology*, 11:1013, 5 2020.

[262] Katherine E Niehaus, Timothy M Walker, Derrick W Crook, Tim E A Peto, and David A Clifton. Machine learning for the prediction of antibacterial susceptibility in Mycobacterium tuberculosis. In *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 618–621, 2014.

[263] Abu Sayed Chowdhury, Douglas R. Call, and Shira L. Broschat. Antimicrobial Resistance Prediction for Gram-Negative Bacteria via Game Theory-Based Feature

Evaluation. *Scientific Reports 2019 9:1*, 9(1):1–9, 10 2019.

[264] Alexandre Drouin, Sébastien Giguère, Maxime Déraspe, Mario Marchand, Michael Tyers, Vivian G. Loo, Anne Marie Bourgault, François Laviolette, and Jacques Corbeil. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC genomics*, 17(1), 9 2016.

[265] Zhichang Liu, Dun Deng, Huijie Lu, Jian Sun, Luchao Lv, Shuhong Li, Guanghui Peng, Xianyong Ma, Jiazhou Li, Zhenming Li, Ting Rong, and Gang Wang. Evaluation of Machine Learning Models for Predicting Antimicrobial Resistance of Actinobacillus pleuropneumoniae From Whole Genome Sequences. *Frontiers in Microbiology*, 11:48, 2 2020.

[266] Yang Yang, Timothy M. Walker, A. Sarah Walker, Daniel J. Wilson, Timothy E.A. Peto, Derrick W. Crook, Farah Shamout, Tingting Zhu, David A. Clifton, Irena Arandjelovic, Iñaki Comas, Maha R. Farhat, Qian Gao, Vitali Sintchenko, Dickvan Soolingen, Sarah Hoosdally, Ana L.Gibertoni Cruz, Joshua Carter, Clara Grazian, Sarah G. Earle, Samaneh Kouchaki, Philip W. Fowler, Zamin Iqbal, Martin Hunt, E. Grace Smith, Priti Rathod, Lisa Jarrett, Daniela Matias, Daniela M. Cirillo, Emanuele Borroni, Simone Battaglia, Arash Ghodousi, Andrea Spitaleri, Andrea Cabibbe, Sabira Tahseen, Kayzad Nilgiriwala, Sanchi Shah, Camilla Rodrigues, Priti Kambli, Utkarsha Surve, Rukhsar Khot, Stefan Niemann, Thomas Kohl, Matthias Merker, Harald Hoffmann, Nikolay Molodtsov, Sara Plesnik, Nazir Ismail, Shaheed Vally Omar, Guy Thwaites, Thuong Nguyen Thuy Thuong, Nhung Hoang Ngoc, Vijay Srinivasan, David Moore, Jorge Coronel, Walter Solano, George F. Gao, Guangxue He, Yanlin Zhao, Aijing Ma, Chunfa Liu, Baoli Zhu, Ian Laurenson, Pauline Claxton, Anastasia Koch, Robert Wilkinson, Ajit Lalvani, James Posey, Jennifer Gardy, Jim Werngren, Nicholas Paton, Ruwen Jou, Mei Hua Wu, Wan Hsuan Lin, Lucilaine Ferrazoli, and Rosangela Siqueira de Oliveira. DeepAMR for predicting co-occurrent resistance of Mycobacterium tuberculosis. *Bioinformatics (Oxford, England)*, 35(18):3240–3249, 9 2019.

[267] Michael L. Chen, Akshith Doddi, Jimmy Royer, Luca Freschi, Marco Schito, Matthew Ezewudo, Isaac S. Kohane, Andrew Beam, and Maha Farhat. Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in Mycobacterium tuberculosis resistance prediction. *EBioMedicine*, 43:356–369, 5 2019.

[268] Jinhong Shi, Yan Yan, Matthew G. Links, Longhai Li, Jo Anne R. Dillon, Michael Horsch, and Anthony Kusalik. Antimicrobial resistance genetic factor identification from whole-genome sequence data using deep feature selection. *BMC Bioinformatics*, 20(15):1–14, 12 2019.

[269] Yuan Li, Benjamin J. Metcalf, Sopio Chochua, Zhongya Li, Robert E. Gertz, Hollis Walker, Paulina A. Hawkins, Theresa Tran, Cynthia G. Whitney, Lesley McGee, and Bernard W. Beall. Penicillin-binding protein transpeptidase signatures for tracking and predicting β-lactam resistance levels in Streptococcus pneumoniae. *mBio*, 7(3), 2016.

[270] Yuan Li, Benjamin J. Metcalf, Sopio Chochua, Zhongya Li, Robert E. Gertz, Hollis Walker, Paulina A. Hawkins, Theresa Tran, Lesley McGee, and Bernard W. Beall. Validation of β-lactam minimum inhibitory concentration predictions for pneumococcal isolates with newly encountered penicillin binding protein (PBP) sequences. *BMC Genomics*, 18(1):621, 12 2017.

[271] Yonatan H. Grad, Simon R. Harris, Robert D. Kirkcaldy, Anna G. Green, Debora S. Marks, Stephen D. Bentley, David Trees, and Marc Lipsitch. Genomic Epidemiology of Gonococcal Resistance to Extended-Spectrum Cephalosporins, Macrolides, and Fluoroquinolones in the United States, 2000–2013. *The Journal of Infectious Diseases*, 214(10):1579–1587, 11 2016.

[272] David Arndt, Jason R. Grant, Ana Marcu, Tanvir Sajed, Allison Pon, Yongjie Liang, and David S. Wishart. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic acids research*, 44(W1):W16–W21, 2016.

[273] Georgia Kapatai, Carmen L. Sheppard, Ali Al-Shahib, David J. Litt, Anthony P. Underwood, Timothy G. Harrison, and Norman K. Fry. Whole genome sequencing of Streptococcus pneumoniae: development, evaluation and verification of targets for serogroup and serotype prediction using an automated pipeline. *PeerJ*, 4(9), 2016.

[274] Nicholas J. Croucher, Simon R. Harris, Yonatan H. Grad, and William P. Hanage. Bacterial genomes in epidemiology—present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1614):20120202, 3 2013.

[275] Nicholas J. Croucher and Xavier Didelot. The application of genomics to tracing bacterial pathogen transmission. *Current Opinion in Microbiology*, 23:62–67, 2 2015.

[276] Chin Yih Ou, Carol A. Ciesielski, Gerald Myers, Claudiu I. Bandea, Chi Cheng Luo, Bette T.M. Korber, James I. Mullins, Gerald Schochetman, Ruth L. Berkelman, A. Nikki Economou, John J. Witte, Lawrence J. Furman, Glen A. Satten, Kersti A. MacInnes, James W. Curran, Harold W. Jaffe, J. Moore, Y. Villamarzo, C. Schable, E. G. Shpaer, T. Liberti, S. Lieb, R. Scott, J. Howell, R. Dumbaugh, A. Lasch, B. Kroesen, L. Ryan, K. Bell, V. Munn, D. Marianos, and B. Gooch. Molecular Epidemiology of HIV Transmission in a Dental Practice. *Science*, 256(5060):1165–1171, 5 1992.

[277] Stuart T. Nichol, Christina F. Spiropoulou, Sergey Morzunov, Pierre E. Rollin, Thomas G. Ksiazek, Heinz Feldmann, Anthony Sanchez, James Childs, Sherif Zaki, and Clarence J. Peters. Genetic Identification of a Hantavirus Associated with an Outbreak of Acute Respiratory Illness. *Science*, 262(5135):914–917, 1993.

[278] Edward C. Holmes, Lin Qi Zhang, Pamela Robertson, Alexander Cleland, Elizabeth Harvey, Peter Simmonds, and Andrew J. Leigh Brown. The molecular epidemiology of human immunodeficiency virus type 1 in Edinburgh. *The Journal of infectious diseases*, 171(1):45–53, 1995.

[279] Jennifer L. Gardy and Nicholas J. Loman. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nature Reviews Genetics 2017 19:1*, 19(1):9–20, 11 2017.

[280] Kyle J. Popovich and Evan S. Snitkin. Whole Genome Sequencing—Implications for Infection Prevention and Outbreak Investigations. *Current Infectious Disease Reports*, 19(4):1–7, 4 2017.

[281] Mark Achtman, James Hale, Ronan A. Murphy, E. Fidelma Boyd, and Steffen Porwollik. Population structures in the SARA and SARB reference collections of Salmonella enterica according to MLST, MLEE and microarray hybridization. *Infection, Genetics and Evolution*, 16:314–325, 6 2013.

[282] Martin C J Maiden. Multilocus Sequence Typing of Bacteria. *Annual Review of Microbiology*, 60(1):561–588, 2006.

[283] Hui min Neoh, Xin Ee Tan, Hassriana Fazilla Sapri, and Toh Leong Tan. Pulsed-field gel electrophoresis (PFGE): A review of the ”gold standard” for bacteria typing and current alternatives. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 74, 10 2019.

[284] B. Swaminathan, T. J. Barrett, S. B. Hunter, and R. V. Tauxe. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerging Infectious Diseases*, 7(3):382, 2001.

[285] Martin C.J. Maiden, Jane A. Bygraves, Edward Feil, Giovanna Morelli, Joanne E. Russell, Rachel Urwin, Qing Zhang, Jiaji Zhou, Kerstin Zurth, Dominique A. Caugant, Ian M. Feavers, Mark Achtman, and Brian G. Spratt. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America*, 95(6):3140–3145, 3 1998.

[286] Roy R. Chaudhuri and Ian R. Henderson. The evolution of the Escherichia coli phylogeny. *Infection, Genetics and Evolution*, 12(2):214–226, 3 2012.

[287] Ziheng Yang and Bruce Rannala. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics 2012 13:5*, 13(5):303–314, 3 2012.

[288] David S. Horner and Graziano Pesole. Phylogenetic analyses: a brief introduction to methods and their application. *Expert review of molecular diagnostics*, 4(3):339–350, 5 2004.

[289] Thomas H Jukes and Charles R Cantor. CHAPTER 24 - Evolution of Protein Molecules. In H N Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, 1969.

[290] Simon Tavaré. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences*, 17(2):57–86, 1986.

[291] Roy D. Sleator. Phylogenetics. *Archives of Microbiology*, 193(4):235–239, 4 2011.

[292] Grace P. McCormack and Jonathan P. Clewley. The application of molecular phylogenetics to the analysis of viral genome diversity and evolution. *Reviews in Medical Virology*, 12(4):221–238, 7 2002.

[293] Joseph Felsenstein. Cases in which Parsimony or Compatibility Methods Will be Positively Misleading. *Systematic Zoology*, 27(4):401, 12 1978.

[294] Ruslan Kalendar and Alan H Schulman. Molecular Plant Taxonomy. *Molecular Plant Taxonomy: Methods and Protocols, Methods in Molecular Biology*, 1115(January 2014):233–255, 2014.

[295] Fabrícia F. Nascimento, Mario Dos Reis, and Ziheng Yang. A biologist’s guide to Bayesian phylogenetic analysis. *Nature Ecology & Evolution 2017 1:10*, 1(10):1446–1454, 9 2017.

[296] Edward J. Feil, Martin C.J. Maiden, Mark Achtman, and Brian G. Spratt. The relative contributions of recombination and mutation to the divergence of clones of Neisseria meningitidis. *Molecular Biology and Evolution*, 16(11):1496–1502, 11 1999.

[297] Marcos Pérez-Losada, Patricia Cabezas, Eduardo Castro-Nallar, and Keith A. Crandall. Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. *Infection, Genetics and Evolution*, 16:38–53, 6 2013.

[298] John A. Lees, Simon R. Harris, Gerry Tonkin-Hill, Rebecca A. Gladstone, Stephanie W. Lo, Jeffrey N. Weiser, Jukka Corander, Stephen D. Bentley, and Nicholas J. Croucher. Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Research*, 29(2):304–316, 2 2019.

[299] Simon R. Harris, Edward J.P. Cartwright, M. Estée Török, Matthew T.G. Holden, Nicholas M. Brown, Amanda L. Ogilvy-Stuart, Matthew J. Ellington, Michael A. Quail, Stephen D. Bentley, Julian Parkhill, and Sharon J. Peacock. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. *The Lancet. Infectious Diseases*, 13(2):130, 2 2013.

[300] Thomas Sakoparnig, Chris Field, and Erik van Nimwegen. Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species. *eLife*, 10:1–61, 1 2021.

[301] A. Friães, R. Mamede, M. Ferreira, J. Melo-Cristino, and M. Ramirez. Annotated Whole-Genome Multilocus Sequence Typing Schema for Scalable High-Resolution Typing of Streptococcus pyogenes. *Journal of Clinical Microbiology*, 6 2022.

[302] Laura Uelze, Josephine Grützke, Maria Borowiak, Jens Andre Hammerl, Katharina Juraschek, Carlus Deneke, Simon H. Tausch, and Burkhard Malorny. Typing methods based on whole genome sequencing data. *One Health Outlook 2020 2:1*, 2(1):1–19, 2 2020.

[303] Gerry Tonkin-Hill, John A. Lees, Stephen D. Bentley, Simon D.W. Frost, and Jukka Corander. RhierBAPS: An R implementation of the population clustering algorithm hierBAPS. *Wellcome Open Research*, 3, 2018.

[304] Danielle J. Ingle, Benjamin P. Howden, and Sebastian Duchene. Development of Phylodynamic Methods for Bacterial Pathogens. *Trends in Microbiology*, 29(9):788–797, 9 2021.

[305] Bryan T Grenfell, Oliver G Pybus, Julia R Gog, James L N Wood, Janet M Daly, Jenny A Mumford, and Edward C Holmes. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science (New York, N.Y.)*, 303(5656):327–332, 1 2004.

[306] Guy Baele, Marc A. Suchard, Andrew Rambaut, and Philippe Lemey. Emerging Concepts of Data Integration in Pathogen Phylodynamics. *Systematic Biology*, 66(1):e47–e65, 1 2017.

[307] Nuno R Faria, Andrew Rambaut, Marc A Suchard, Guy Baele, Trevor Bedford, Melissa J Ward, Andrew J Tatem, João D Sousa, Nimalan Arinaminpathy, Jacques Pépin, David Posada, Martine Peeters, Oliver G Pybus, and Philippe Lemey. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. *Science*, 346(6205):56–61, 11 2014.

[308] Edward C. Holmes, Gytis Dudas, Andrew Rambaut, and Kristian G. Andersen. The evolution of Ebola virus: Insights from the 2013-2016 epidemic. *Nature*, 538(7624):193–200, 10 2016.

[309] Louis du Plessis, John T. McCrone, Alexander E. Zarebski, Verity Hill, Christopher Ruis, Bernardo Gutierrez, Jayna Raghwani, Jordan Ashworth, Rachel Colquhoun,

Thomas R. Connor, Nuno R. Faria, Ben Jackson, Nicholas J. Loman, Áine O'Toole, Samuel M. Nicholls, Kris V. Parag, Emily Scher, Tetyana I. Vasylyeva, Erik M. Volz, Alexander Watts, Isaac I. Bogoch, Kamran Khan, David M. Aanensen, Moritz U.G. Kraemer, Andrew Rambaut, and Oliver G. Pybus. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science*, 371(6530):708–712, 2 2021.

[310] Oliver G. Pybus, Marc A. Suchard, Philippe Lemey, Flavien J. Bernardin, Andrew Rambaut, Forrest W. Crawford, Rebecca R. Gray, Nimalan Arinaminpathy, Susan L. Stramer, Michael P. Busch, and Eric L. Delwart. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, 109(37):15066–15071, 9 2012.

[311] M. J. Ward, C. L. Gibbons, P. R. McAdam, B. A.D. van Bunnik, E. K. Girvan, G. F. Edwards, J. R. Fitzgerald, and M. E.J. Woolhouse. Time-scaled evolutionary analysis of the transmission and antibiotic resistance dynamics of Staphylococcus aureus clonal complex 398. *Applied and Environmental Microbiology*, 80(23):7275–7282, 2014.

[312] Nicholas J Croucher, William P Hanage, Simon R Harris, Lesley McGee, der Linden van, Herminia de Lencastre, Raquel Sá-Leão, Jae-Hoon Song, Kwan Soo Ko, Bernard Beall, Keith P Klugman, Julian Parkhill, Alexander Tomasz, Karl G Kristinsson, and Stephen D Bentley. Variable recombination dynamics during the emergence, transmission and 'disarming' of a multidrug-resistant pneumococcal clone. *BMC Biology*, 12(1):49, 2014.

[313] Erik M Volz and Xavier Didelot. Modeling the Growth and Decline of Pathogen Effective Population Size Provides Insight into Epidemic Dynamics and Drivers of Antimicrobial Resistance. *Systematic Biology*, 2 2018.

[314] R K Selander and J M Musser. Molecular Basis of Bacterial Infections. In B H Iglewski and V L Clark, editors, *Molecular Basis of Bacterial Infections*, pages 11–36. Academic, San Diego, 1990.

[315] Thomas S. Whittam, Kaye Wachsmuth, Richard A. Wilson, and Thomas S. Whittam. Genetic evidence of clonal descent of Escherichia coli O157:H7 associated with hemorrhagic colitis and hemolytic uremic syndrome. *The Journal of infectious diseases*, 157(6):1124–1130, 1988.

[316] James M. Musser, Alan R. Hauser, Michael H. Kim, Patrick M. Schlievert, Kimberlyn Nelson, and Robert K. Selander. Streptococcus pyogenes causing toxic-shock-like syndrome and other invasive diseases: clonal diversity and pyrogenic exotoxin expression. *Proceedings of the National Academy of Sciences of the United States of America*, 88(7):2668–2672, 4 1991.

[317] Robert K. Selander and Bruce R. Levin. Genetic Diversity and Structure in Escherichia coli Populations. *Science*, 210(4469):545–547, 10 1980.

[318] Brian G. Spratt, Lucas D. Bowler, Qian Yun Zhang, Jiaji Zhou, and John Maynard Smith. Role of interspecies transfer of chromosomal genes in the evolution of penicillin resistance in pathogenic and commensal Neisseria species. *Journal of Molecular Evolution*, 34(2):115–125, 1992.

[319] I. M. Feavers, A. B. Heath, J. A. Bygraves, and M. C.J. Maiden. Role of horizontal genetic exchange in the antigenic variation of the class 1 outer membrane protein

of Neisseria meningitidis. *Molecular Microbiology*, 6(4):489–495, 2 1992.

[320] Edward Feil, Gill Carpenter, and Brian G. Spratt. Electrophoretic variation in adenylate kinase of Neisseria meningitidis is due to inter- and intraspecies recombination. *Proceedings of the National Academy of Sciences of the United States of America*, 92(23):10535–10539, 11 1995.

[321] J. M. Smith, N. H. Smith, M. O'Rourke, and B. G. Spratt. How clonal are bacteria? *Proceedings of the National Academy of Sciences of the United States of America*, 90(10):4384–4388, 1993.

[322] Mark Achtman. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annual review of microbiology*, 62:53–70, 2008.

[323] Srinand Sreevatsan, Xi Pan, Kathryn E. Stockbauer, Nancy D. Connell, Barry N. Kreiswirth, Thomas S. Whittam, and James M. Musser. Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination. *Proceedings of the National Academy of Sciences of the United States of America*, 94(18):9869–9874, 9 1997.

[324] Mark Achtman, Kerstin Zurth, Giovanna Morelli, Gabriela Torrea, Annie Guiyoule, and Elisabeth Carniel. Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis. *Proceedings of the National Academy of Sciences of the United States of America*, 96(24):14043–14048, 11 1999.

[325] Luca Freschi, Roger Vargas, Ashaque Husain, S. M.Mostofa Kamal, Alena Skrahina, Sabira Tahseen, Nazir Ismail, Anna Barbova, Stefan Niemann, Daniela Maria Cirillo, Anna S. Dean, Matteo Zignol, and Maha Reda Farhat. Population structure, biogeography and transmissibility of Mycobacterium tuberculosis. *Nature Communications*, 12(1), 12 2021.

[326] Yujun Cui, Chang Yu, Yanfeng Yan, Dongfang Li, Yanjun Li, Thibaut Jombart, Lucy A. Weinert, Zuyun Wang, Zhaobiao Guo, Lizhi Xu, Yujiang Zhang, Hancheng Zheng, Nan Qin, Xiao Xiao, Mingshou Wu, Xiaoyi Wang, Dongsheng Zhou, Zhizhen Qi, Zongmin Du, Honglong Wu, Xianwei Yang, Hongzhi Cao, Hu Wang, Jing Wang, Shusen Yao, Alexander Rakin, Yingrui Li, Daniel Falush, Francois Balloux, Mark Achtman, Yajun Song, Jun Wang, and Ruifu Yang. Historical variations in mutation rate in an epidemic pathogen, Yersinia pestis. *Proceedings of the National Academy of Sciences of the United States of America*, 110(2):577–582, 1 2013.

[327] Duccio Medini, Claudio Donati, Hervé Tettelin, Vega Masignani, and Rino Rappuoli. The microbial pan-genome. *Current Opinion in Genetics & Development*, 15(6):589–594, 12 2005.

[328] L. Rouli, V. Merhej, P. E. Fournier, and D. Raoult. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections*, 7:72, 9 2015.

[329] Gerry Tonkin-Hill, Neil MacAlasdair, Christopher Ruis, Aaron Weimann, Gal Horesh, John A. Lees, Rebecca A. Gladstone, Stephanie Lo, Christopher Beaudoin, R. Andres Floto, Simon D.W. Frost, Jukka Corander, Stephen D. Bentley, and Julian Parkhill. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology*, 21(1):1–21, 7 2020.

[330] Alexander E. Lobkovsky, Yuri I. Wolf, and Eugene V. Koonin. Gene Frequency Distributions Reject a Neutral Model of Genome Evolution. *Genome Biology and*

Evolution, 5(1):233–242, 1 2013.

[331] Andrew J Page, Carla A Cummins, Martin Hunt, Vanessa K Wong, Sandra Reuter, Matthew T G Holden, Maria Fookes, Daniel Falush, Jacqueline A Keane, and Julian Parkhill. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31(22):3691–3693, 11 2015.

[332] Guillaume Gautreau, Adelme Bazin, Mathieu Gachet, Rémi Planel, Laura Burlot, Mathieu Dubois, Amandine Perrin, Claudine Médigue, Alexandra Calteau, Stéphane Cruveiller, Catherine Matias, Christophe Ambroise, Eduardo P.C. Rocha, and David Vallenet. PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLOS Computational Biology*, 16(3):e1007732, 2020.

[333] Hervé Tettelin, David Riley, Ciro Cattuto, and Duccio Medini. Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology*, 11(5):472–477, 10 2008.

[334] Gal Horesh, Alyce Taylor-Brown, Stephanie McGimpsey, Florent Lassalle, Jukka Corander, Eva Heinz, and Nicholas R. Thomson. Different evolutionary trends form the twilight zone of the bacterial pan-genome. *Microbial Genomics*, 7(9):670, 11 2021.

[335] Xavier Argemi, Dorota Matelska, Krzysztof Ginalski, Philippe Riegel, Yves Hansmann, Jochen Bloom, Martine Pestel-Caron, Sandrine Dahyot, Jérémie Lebeurre, and Gilles Prévost. Comparative genomic analysis of Staphylococcus lugdunensis shows a closed pan-genome and multiple barriers to horizontal gene transfer. *BMC Genomics*, 19(1):1–16, 8 2018.

[336] James O. McInerney, Alan McNally, and Mary J. O'Connell. Why prokaryotes have pangenomes. *Nature Microbiology 2017 2:4*, 2(4):1–5, 3 2017.

[337] Michael A. Brockhurst, Ellie Harrison, James P.J. Hall, Thomas Richards, Alan McNally, and Craig MacLean. The Ecology and Evolution of Pangenomes. *Current Biology*, 29(20):R1094–R1103, 10 2019.

[338] N. Luisa Hiller and Raquel Sá-Leão. Puzzling Over the Pneumococcal Pangenome. *Frontiers in Microbiology*, 9:2580, 3 2018.

[339] Christopher M. Thomas and Kaare M. Nielsen. Mechanisms of and Barriers to, Horizontal Gene Transfer between Bacteria. *Nature Reviews Microbiology*, 3(9):711–721, 9 2005.

[340] Samir N Patel, Roberto Melano, Allison McGeer, Karen Green, and Donald E Low. Characterization of the quinolone resistant determining regions in clinical isolates of pneumococci collected in Canada. *Annals of clinical microbiology and antimicrobials*, 9:3, 1 2010.

[341] Pieter-Jan Ceyssens, Françoise Van Bambeke, Wesley Mattheus, Sophie Bertrand, Frédéric Fux, Eddie Van Bossuyt, Sabrina Damée, Henry-Jean Nyssen, Stéphane De Craeye, Jan Verhaegen, The Belgian Streptococcus pneumoniae Study Group, Paul M. Tulkens, and Raymond Vanhoof. Molecular Analysis of Rising Fluoroquinolone Resistance in Belgian Non-Invasive Streptococcus pneumoniae Isolates (1995-2014). *PLoS ONE*, 11(5):e0154816, 2016.

[342] Shannon M. Soucy, Jinling Huang, and Johann Peter Gogarten. Horizontal gene transfer: Building the web of life, 2015.

[343] Brian J. Arnold, I. Ting Huang, and William P. Hanage. Horizontal gene transfer and adaptive evolution in bacteria. *Nature reviews. Microbiology*, 20(4):206–218, 4 2022.

[344] Michael T Madigan, Kelly S Bender, Daniel H Buckley, W Matthew Sattley, and David A Stahl. *Brock biology of microorganisms*. Pearson, Boston, 15th edition, 2018.

[345] Kimihiro Abe, Nobuhiko Nomura, and Satoru Suzuki. Biofilms: hot spots of horizontal gene transfer (HGT) in aquatic environments, with a focus on a new HGT mechanism. *FEMS Microbiology Ecology*, 96(5):31, 5 2020.

[346] Hiroshi Xavier Chiura, Kazuhiro Kogure, Sylvia Hagemann, Adolf Ellinger, and Branko Velimirov. Evidence for particle-induced horizontal gene transfer and serial transduction between bacteria. *FEMS microbiology ecology*, 76(3):576–591, 6 2011.

[347] Gyanendra P. Dubey and Sigal Ben-Yehuda. Intercellular nanotubes mediate bacterial communication. *Cell*, 144(4):590–600, 2 2011.

[348] Pavol Bárdy, Tibor Füzik, Dominik Hrebík, Roman Pantůček, J. Thomas Beatty, and Pavel Plevka. Structure and mechanism of DNA delivery of a gene transfer agent. *Nature communications*, 11(1), 12 2020.

[349] Laura S. Frost, Raphael Leplae, Anne O. Summers, and Ariane Toussaint. Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology*, 3(9):722–732, 9 2005.

[350] Nicholas J Croucher, Rafal Mostowy, Christopher Wymant, Paul Turner, Stephen D Bentley, and Christophe Fraser. Horizontal DNA Transfer Mechanisms of Bacteria as Weapons of Intragenomic Conflict. *PLoS Biology*, 14(3):e1002394, 3 2016.

[351] Curtis A. Suttle. Viruses in the sea. *Nature*, 437(7057):356–361, 9 2005.

[352] Roger W. Hendrix, Margaret C.M. Smith, R. Neil Burns, Michael E. Ford, and Graham F. Hatfull. Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *Proceedings of the National Academy of Sciences of the United States of America*, 96(5):2192, 3 1999.

[353] Moïra B. Dion, Frank Oechslin, and Sylvain Moineau. Phage diversity, genomics and phylogeny. *Nature Reviews Microbiology 2020 18:3*, 18(3):125–138, 2 2020.

[354] Athina Zampara, Martine C.Holst Sørensen, Dennis Grimon, Fabio Antenucci, Amira Ruslanovna Vitt, Valeria Bortolaia, Yves Briers, and Lone Brøndsted. Exploiting phage receptor binding proteins to enable endolysins to kill Gram-negative bacteria. *Scientific Reports 2020 10:1*, 10(1):1–12, 7 2020.

[355] L. Plançon, C. Janmot, M. Le Maire, M. Desmadril, M. Bonhivers, L. Letellier, and P. Boulanger. Characterization of a High-affinity Complex Between the Bacterial Outer Membrane Protein FhuA and the Phage T5 Protein pb5. *Journal of Molecular Biology*, 318(2):557–569, 4 2002.

[356] M. Bonhivers, L. Plançon, A. Ghazi, P. Boulanger, M. Le Maire, O. Lambert, J. L. Rigaud, and L. Letellier. FhuA, an Escherichia coli outer membrane protein with a dual function of transporter and channel which mediates the transport of phage DNA. *Biochimie*, 80(5-6):363–369, 1998.

[357] José R. Penadés, John Chen, Nuria Quiles-Puchalt, Nuria Carpena, and Richard P. Novick. Bacteriophage-mediated spread of bacterial virulence genes. *Current Opinion in Microbiology*, 23:171–178, 2 2015.

[358] Zack Hobbs and Stephen T. Abedon. Diversity of phage infection types and associated terminology: the problem with 'Lytic or lysogenic'. *FEMS Microbiology Letters*, 363(7):47, 4 2016.

[359] Nigel D.F. Grindley, Katrine L. Whiteson, and Phoebe A. Rice. Mechanisms of site-specific recombination. *Annual review of biochemistry*, 75:567–605, 2006.

[360] Bernard Hallet and David J. Sherratt. Transposition and site-specific recombination: adapting DNA cut-and-paste mechanisms to a variety of genetic rearrangements. *FEMS Microbiology Reviews*, 21(2):157–178, 9 1997.

[361] Tania A. Baker and Li Luo. Identification of residues in the Mu transposase essential for catalysis. *Proceedings of the National Academy of Sciences of the United States of America*, 91(14):6654–6658, 7 1994.

[362] Sherwin P. Montaño, Ying Z. Pigli, and Phoebe A. Rice. The Mu transpososome structure sheds light on DDE recombinase evolution. *Nature 2012 491:7424*, 491(7424):413–417, 11 2012.

[363] Anne Chevallereau, Benoît J. Pons, Stineke van Houte, and Edze R. Westra. Interactions between bacterial and phage communities in natural environments. *Nature Reviews Microbiology 2021 20:1*, 20(1):49–62, 8 2021.

[364] Arun M. Nanda, Kai Thormann, and Julia Frunzke. Impact of spontaneous prophage induction on the fitness of bacterial populations and host-microbe interactions. *Journal of Bacteriology*, 197(3):410–419, 2015.

[365] Yin Ning Chiang, José R. Penadés, and John Chen. Genetic transduction by phages and chromosomal islands: The new and noncanonical. *PLOS Pathogens*, 15(8):e1007878, 2019.

[366] N. D. Zinder and J. Lederberg. Genetic exchange in Salmonella. *Journal of bacteriology*, 64(5):679–699, 11 1952.

[367] George P.C. Salmond and Peter C. Fineran. A century of the phage: past, present and future. *Nature Reviews Microbiology 2015 13:12*, 13(12):777–786, 11 2015.

[368] John Chen, Nuria Quiles-Puchalt, Yin Ning Chiang, Rodrigo Bacigalupe, Alfred Fillol-Salom, Melissa Su Juan Chee, J. Ross Fitzgerald, and José R. Penadés. Genome hypermobility by lateral transduction. *Science*, 362(6411):207–212, 10 2018.

[369] Marie Touchon, Aude Bernheim, and Eduardo P.C. Rocha. Genetic and life-history traits associated with the distribution of prophages in bacteria. *The ISME Journal*, 10(11):2744, 11 2016.

[370] Ellie Harrison and Michael A. Brockhurst. Ecological and Evolutionary Benefits of Temperate Phage: What Does or Doesn't Kill You Makes You Stronger. *BioEssays*, 39(12):1700112, 12 2017.

[371] Richard P Novick. Mobile genetic elements and bacterial toxinoses: the superantigen-encoding pathogenicity islands of Staphylococcus aureus. *Plasmid*, 49(2):93–105, 3 2003.

[372] Matthew K. Waldor and John J. Mekalanos. Lysogenic Conversion by a Filamentous Phage Encoding Cholera Toxin. *Science*, 272(5270):1910–1913, 6 1996.

[373] Alison D. O'Brien, John W. Newland, Steven F. Miller, Randall K. Holmes, H. Williams Smith, and Samuel B. Formal. Shiga-Like Toxin-Converting Phages from Escherichia coli Strains That Cause Hemorrhagic Colitis or Infantile Diarrhea. *Science*, 226(4675):694–696, 1984.

[374] Alexey Ruzin, Jodi Lindsay, and Richard P. Novick. Molecular genetics of SaPI1 – a mobile pathogenicity island in Staphylococcus aureus. *Molecular Microbiology*, 41(2):365–377, 7 2001.

[375] Richard P. Novick and Geeta Ram. The Floating (Pathogenicity) Island: A Genomic Dessert. *Trends in Genetics*, 32(2):114–126, 2 2016.

[376] John Chen and Richard P. Novick. Phage-mediated intergeneric transfer of toxin genes. *Science*, 323(5910):139–141, 1 2009.

[377] Kinga I. Stanczak-Mrozek, Anusha Manne, Gwenan M. Knight, Katherine Gould, Adam A. Witney, and Jodi A. Lindsay. Within-host diversity of MRSA antimicrobial resistances. *The Journal of antimicrobial chemotherapy*, 70(8):2191–2198, 8 2015.

[378] Victoriya V. Volkova, Zhao Lu, Thomas Besser, and Yrjö T. Gröhn. Modeling the infection dynamics of bacteriophages in enteric Escherichia coli: estimating the contribution of transduction to antimicrobial gene spread. *Applied and environmental microbiology*, 80(14):4350–4362, 2014.

[379] Sheetal R. Modi, Henry H. Lee, Catherine S. Spina, and James J. Collins. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature*, 499(7457):219–222, 6 2013.

[380] Willem van Schaik. The human gut resistome. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1670):20140087, 6 2015.

[381] François Enault, Arnaud Briet, Léa Bouteille, Simon Roux, Matthew B. Sullivan, and Marie Agnès Petit. Phages rarely encode antibiotic resistance genes: A cautionary tale for virome analyses. *ISME Journal*, 11(1):237–247, 2017.

[382] Eugene V. Koonin, Tatiana G. Senkevich, and Valerian V. Dolja. The ancient Virus World and evolution of cells. *Biology Direct*, 1:29, 9 2006.

[383] Aude Bernheim and Rotem Sorek. The pan-immune system of bacteria: antiviral defence as a community resource. *Nature Reviews Microbiology 2019 18:2*, 18(2):113–119, 11 2019.

[384] Julie E. Samson, Alfonso H. Magadán, Mourad Sabri, and Sylvain Moineau. Revenge of the phages: Defeating bacterial defences. *Nature Reviews Microbiology*, 11(10):675–687, 2013.

[385] Shigeru Iida, Markus B. Streiff, Thomas A. Bickle, and Werner Arber. Two DNA antirestriction systems of bacteriophage P1, darA, and darB: characterization of darA- phages. *Virology*, 157(1):156–166, 1987.

[386] Kimberley D. Seed, David W. Lazinski, Stephen B. Calderwood, and Andrew Camilli. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature 2013 494:7438*, 494(7438):489–491, 2 2013.

[387] Geeta Ram, John Chen, Krishan Kumar, Hope F. Ross, Carles Ubeda, Priyadarshan K. Damle, Kristin D. Lane, José R. Penaḍes, Gail E. Christie, and Richard P.

Novick. Staphylococcal pathogenicity island interference with helper phage reproduction is a paradigm of molecular parasitism. *Proceedings of the National Academy of Sciences of the United States of America*, 109(40):16300–16305, 10 2012.

[388] Hila Sberro, Azita Leavitt, Ruth Kiro, Eugene Koh, Yoav Peleg, Udi Qimron, and Rotem Sorek. Discovery of functional toxin/antitoxin systems in bacteria by shotgun cloning. *Molecular cell*, 50(1):136, 4 2013.

[389] François Rousset, Florence Depardieu, Solange Miele, Julien Dowding, Anne Laure Laval, Erica Lieberman, Daniel Garry, Eduardo P.C. Rocha, Aude Bernheim, and David Bikard. Phages and their satellites encode hotspots of antiviral systems. *Cell Host & Microbe*, 30(5):740–753, 5 2022.

[390] Matthieu Haudiquet, Amandine Buffet, Olaya Rendueles, and Eduardo P.C. Rocha. Interplay between the cell envelope and mobile genetic elements shapes gene flow in populations of the nosocomial pathogen Klebsiella pneumoniae. *PLOS Biology*, 19(7):e3001276, 7 2021.

[391] Louis Marie Bobay, Eduardo P.C. Rocha, and Marie Touchon. The Adaptation of Temperate Bacteriophages to Their Host Genomes. *Molecular Biology and Evolution*, 30(4):737–751, 4 2013.

[392] Ariane Toussaint and Phoebe A. Rice. Transposable phages, DNA reorganization and transfer. *Current Opinion in Microbiology*, 38:88–94, 8 2017.

[393] Harald Brüssow, Carlos Canchaya, and Wolf-Dietrich Hardt. Phages and the Evolution of Bacterial Pathogens: from Genomic Rearrangements to Lysogenic Conversion. *Microbiology and Molecular Biology Reviews*, 68(3):560, 9 2004.

[394] Maria José Lopez-Sanchez, Elisabeth Sauvage, Violette Da Cunha, Dominique Clermont, Elisoa Ratsima Hariniaina, Bruno Gonzalez-Zorn, Claire Poyart, Isabelle Rosinski-Chupin, and Philippe Glaser. The highly dynamic CRISPR1 system of Streptococcus agalactiae controls the diversity of its mobilome. *Molecular Microbiology*, 85(6):1057–1071, 9 2012.

[395] Julien Guglielmini, Leonor Quintais, Maria Pilar Garcillán-Barcia, Fernando de la Cruz, and Eduardo P. C. Rocha. The Repertoire of ICE in Prokaryotes Underscores the Unity, Diversity, and Ubiquity of Conjugation. *PLoS Genetics*, 7(8):e1002222, 8 2011.

[396] Elena Cabezón, Jorge Ripoll-Rozada, Alejandro Peña, Fernando de la Cruz, and Ignacio Arechaga. Towards an integrated model of bacterial conjugation. *FEMS Microbiology Reviews*, 39(1):n/a–n/a, 9 2014.

[397] Tiago R.D. Costa, Catarina Felisberto-Rodrigues, Amit Meir, Marie S. Prevost, Adam Redzej, Martina Trokter, and Gabriel Waksman. Secretion systems in Gram-negative bacteria: structural and mechanistic insights. *Nature Reviews Microbiology 2015 13:6*, 13(6):343–359, 5 2015.

[398] Martina Trokter, Catarina Felisberto-Rodrigues, Peter J. Christie, and Gabriel Waksman. Recent advances in the structural and molecular biology of type IV secretion systems. *Current Opinion in Structural Biology*, 27(1):16–23, 8 2014.

[399] Peter J. Christie, Neal Whitaker, and Christian González-Rivera. BBA Review Revised Mechanism and Structure of the Bacterial Type IV Secretion Systems. *Biochimica et biophysica acta*, 1843(8):1578, 2014.

[400] Michael Chandler, Fernando De La Cruz, Fred Dyda, Alison B. Hickman, Gabriel Moncalian, and Bao Ton-Hoang. Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. *Nature reviews. Microbiology*, 11(8):525, 8 2013.

[401] Fernando De La Cruz, Laura S. Frost, Richard J. Meyer, and Ellen L. Zechner. Conjugative DNA metabolism in Gram-negative bacteria. *FEMS microbiology reviews*, 34(1):18–40, 1 2010.

[402] Mathieu Brochet, Christophe Rusniok, Elisabeth Couvé, Shaynoor Dramsi, Claire Poyart, Patrick Trieu-Cuot, Frank Kunst, and Philippe Glaser. Shaping a bacterial genome by large chromosomal replacements, the evolutionary history of Streptococcus agalactiae. *Proceedings of the National Academy of Sciences of the United States of America*, 105(41):15961–15966, 10 2008.

[403] Anders Norman, Lars H. Hansen, and Soren J. Sørensen. Conjugative plasmids: vessels of the communal gene pool. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1527):2275, 8 2009.

[404] J. Lederberg. Cell genetics and hereditary symbiosis. *Physiological reviews*, 32(4):403–430, 10 1952.

[405] Jerónimo Rodríguez-Beltrán, Javier DelaFuente, Ricardo León-Sampedro, R. Craig MacLean, and Álvaro San Millán. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nature Reviews Microbiology 2021 19:6*, 19(6):347–359, 1 2021.

[406] Alvaro San Millan and R. Craig MacLean. Fitness Costs of Plasmids: a Limit to Plasmid Transmission. *Microbiology spectrum*, 5(5), 9 2017.

[407] Ana María Hernández-Arriaga, Wai Ting Chan, Manuel Espinosa, and Ramón Díaz-Orejas. Conditional Activation of Toxin-Antitoxin Systems: Postsegregational Killing and Beyond. *Microbiology Spectrum*, 2(5), 9 2014.

[408] Masaki Shintani, Zoe K. Sanchez, and Kazuhide Kimbara. Genomics of microbial plasmids: Classification and identification based on replication and transfer systems and host taxonomy. *Frontiers in Microbiology*, 6(MAR):242, 2015.

[409] Michelle M.C. Buckner, Maria Laura Ciusa, and Laura J.V. Piddock. Strategies to combat antimicrobial resistance: anti-plasmid and plasmid curing. *FEMS Microbiology Reviews*, 42(6):781–804, 11 2018.

[410] Mark R. Tock and David T.F. Dryden. The biology of restriction and anti-restriction. *Current Opinion in Microbiology*, 8(4):466–472, 8 2005.

[411] Brian M. Wilkins. Plasmid promiscuity: meeting the challenge of DNA immigration control. *Environmental microbiology*, 4(9):495–500, 9 2002.

[412] Alessandra Carattoli. Plasmids and the spread of resistance. *International Journal of Medical Microbiology*, 303(6-7):298–304, 8 2013.

[413] Jennifer L. Cottell, Mark A. Webber, Nick G. Coldham, Dafydd L. Taylor, Anna M. Cerdeño-Tárraga, Heidi Hauser, Nicholas R. Thomson, Martin J. Woodward, and Laura J.V. Piddock. Complete Sequence and Molecular Epidemiology of IncK Epidemic Plasmid Encoding blaCTX-M-14. *Emerging Infectious Diseases*, 17(4):645, 4 2011.

[414] Monika Dolejska and Costas C. Papagiannitsis. Plasmid-mediated resistance is going wild. *Plasmid*, 99:99–111, 9 2018.

[415] Alejandro Couce, Alexandro Rodríguez-Rojas, and Jesús Blázquez. Bypass of genetic constraints during mutator evolution to antibiotic resistance. *Proceedings of the Royal Society B: Biological Sciences*, 282(1804), 2015.

[416] Alvaro San Millan, Jose Antonio Escudero, Danna R. Gifford, DIdier Mazel, and R. Craig MacLean. Multicopy plasmids potentiate the evolution of antibiotic resistance in bacteria. *Nature Ecology and Evolution*, 1(1), 11 2016.

[417] Mitchell W. Pesesky, Rayna Tilley, and David A.C. Beck. Mosaic plasmids are abundant and unevenly distributed across prokaryotic taxa. *Plasmid*, 102:10–18, 3 2019.

[418] Peter Norberg, Maria Bergström, Vinay Jethava, Devdatt Dubhashi, and Malte Hermansson. The IncP-1 plasmid backbone adapts to different host bacterial species and evolves through homologous recombination. *Nature communications*, 2(1), 2011.

[419] Miaomiao Xie, Ruichao Li, Zhonghua Liu, Edward Wai Chi Chan, and Sheng Chen. Recombination of plasmids in a carbapenem-resistant NDM-5-producing clinical Escherichia coli isolate. *Journal of Antimicrobial Chemotherapy*, 73(5):1230–1234, 5 2018.

[420] Xavier Bellanger, Sophie Payot, Nathalie Leblond-Bourget, and Gérard Guédon. Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. *FEMS Microbiology Reviews*, 38(4):720–760, 7 2014.

[421] Christopher M Johnson and Alan D Grossman. Integrative and Conjugative Elements (ICEs): What They Do and How They Work. *Annual Review of Genetics*, 49(1):577–601, 1 2015.

[422] Jean Cury, Marie Touchon, and Eduardo P.C. Rocha. Integrative and conjugative elements and their hosts: composition, distribution and organization. *Nucleic acids research*, 45(15):8943–8956, 9 2017.

[423] Jean Cury, Pedro H. Oliveira, Fernando De La Cruz, and Eduardo P.C. Rocha. Host Range and Genetic Plasticity Explain the Coexistence of Integrative and Extrachromosomal Mobile Genetic Elements. *Molecular Biology and Evolution*, 35(9):2230, 9 2018.

[424] J. Casey, C. Daly, and G. F. Fitzgerald. Chromosomal integration of plasmid DNA by homologous recombination in Enterococcus faecalis and Lactococcus lactis subsp. lactis hosts harboring Tn919. *Applied and environmental microbiology*, 57(9):2677–2682, 1991.

[425] Rachel A. F. Wozniak and Matthew K. Waldor. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nature Reviews Microbiology*, 8(8):552–563, 8 2010.

[426] John W Beaber, Bianca Hochhut, and Matthew K Waldor. SOS response promotes horizontal dissemination of antibiotic resistance genes. *Nature*, 427(6969):72, 11 2004.

[427] Xavier Bellanger, Catherine Morel, Bernard Decaris, and Gérard Guédon. Derepression of excision of integrative and potentially conjugative elements from Streptococcus thermophilus by DNA damage response: implication of a cI-related repressor. *Journal of bacteriology*, 189(4):1478–81, 2 2007.

[428] Ryo Miyazaki, Marco Minoia, Nicolas Pradervand, Sandra Sulser, Friedrich Reinhard, and Jan Roelof van der Meer. Cellular Variability of RpoS Expression Underlies Subpopulation Activation of an Integrative and Conjugative Element. *PLoS Genetics*, 8(7):e1002818, 7 2012.

[429] Adam P Roberts and Peter Mullany. A modular master on the move: the Tn916 family of mobile genetic elements. *Trends in Microbiology*, 17(6):251–258, 11 2009.

[430] Rachel A. F. Wozniak, Derrick E. Fouts, Matteo Spagnoletti, Mauro M. Colombo, Daniela Ceccarelli, Geneviève Garriss, Christine Déry, Vincent Burrus, and Matthew K. Waldor. Comparative ICE Genomics: Insights into the Evolution of the SXT/R391 Family of ICEs. *PLoS Genetics*, 5(12):e1000786, 12 2009.

[431] Chioma C. Obi, Shivangi Vayla, Vidya de Gannes, Mark E. Berres, Jason Walker, Derek Pavelec, Joshua Hyman, and William J. Hickey. The Integrative Conjugative Element clc (ICEclc) of Pseudomonas aeruginosa JB2. *Frontiers in Microbiology*, 9:1532, 7 2018.

[432] Michel A. Marin, Erica L. Fonseca, Bruno N. Andrade, Adriana C. Cabral, and Ana Carolina P. Vicente. Worldwide Occurrence of Integrative Conjugative Element Encoding Multidrug Resistance Determinants in Epidemic Vibrio cholerae O1. *PLOS ONE*, 9(9):e108728, 9 2014.

[433] Matteo Spagnoletti, Daniela Ceccarelli, Adrien Rieux, Marco Fondi, Elisa Taviani, Renato Fani, Mauro M. Colombo, Rita R. Colwell, and François Balloux. Acquisition and evolution of SXT-R391 integrative conjugative elements in the seventh-pandemic Vibrio cholerae lineage. *mBio*, 5(4), 8 2014.

[434] Erica L. Fonseca, Michel A. Marin, Fernando Encinas, and Ana Carolina P. Vicente. Full characterization of the integrative and conjugative element carrying the metallo-β-lactamase blaSPM-1 and bicyclomycin bcr1 resistance genes found in the pandemic Pseudomonas aeruginosa clone SP/ST277. *Journal of Antimicrobial Chemotherapy*, 70(9):2547–2550, 9 2015.

[435] Nicholas J Croucher, Danielle Walker, Patricia Romero, Nicola Lennard, Gavin K Paterson, Nathalie C Bason, Andrea M Mitchell, Michael A Quail, Peter W Andrew, Julian Parkhill, Stephen D Bentley, and Tim J Mitchell. Role of Conjugative Elements in the Evolution of the Multidrug-Resistant Pandemic Clone Streptococcus pneumoniaeSpain23F ST81. *Journal of Bacteriology*, 191(5):1480–1489, 3 2009.

[436] Helena Seth-Smith and Nicholas J. Croucher. Breaking the ICE. *Nature Reviews Microbiology 2009 7:5*, 7(5):328–329, 2009.

[437] Guillaume Pavlovic, Vincent Burrus, Brigitte Gintz, Bernard Decaris, and Gérard Guédon. Evolution of genomic islands by deletion and tandem accretion by site-specific recombination: ICESt1-related elements from Streptococcus thermophilus. *Microbiology*, 150(4):759–774, 4 2004.

[438] P. Ayoubi, A. O. Kilic, and M. N. Vijayakumar. Tn5253, the pneumococcal Ω (cat tet) BM6001 element, is a composite structure of two conjugative transposons, Tn5251 and Tn5252. *Journal of Bacteriology*, 173(5):1617–1622, 1991.

[439] Francesco Iannelli, Francesco Santoro, Marco R. Oggioni, and Gianni Pozzi. Nucleotide Sequence Analysis of Integrative Conjugative Element Tn 5253 of Streptococcus pneumoniae. *Antimicrobial Agents and Chemotherapy*, 58(2):1235–1239, 2 2014.

[440] Francesco Santoro, Alessandra Romeo, Gianni Pozzi, and Francesco Iannelli. Excision and Circularization of Integrative Conjugative Element Tn5253 of Streptococcus pneumoniae. *Frontiers in Microbiology*, 9:1779, 7 2018.

[441] Patrick Seitz and Melanie Blokesch. Cues and regulatory pathways involved in natural competence and transformation in pathogenic and environmental Gram-negative bacteria. *FEMS Microbiology Reviews*, 37(3):336–363, 5 2013.

[442] Jan Willem Veening and Melanie Blokesch. Interbacterial predation as a strategy for DNA acquisition in naturally competent bacteria. *Nature Reviews Microbiology 2017 15:10*, 15(10):621–629, 7 2017.

[443] Melanie Blokesch. Natural competence for transformation. *Current Biology*, 26(21):R1126–R1130, 11 2016.

[444] Calum Johnston, Bernard Martin, Gwennaele Fichant, Patrice Polard, and Jean Pierre Claverys. Bacterial transformation: distribution, shared mechanisms and divergent control. *Nature Reviews Microbiology 2014 12:3*, 12(3):181–196, 2 2014.

[445] Marc Prudhomme, Mathieu Berge, Bernard Martin, and Patrice Polard. Pneumococcal Competence Coordination Relies on a Cell-Contact Sensing Mechanism. *PLOS Genetics*, 12(6):e1006113, 6 2016.

[446] Ola Johnsborg and Leiv Sigve Håvarstein. Regulation of natural genetic transformation and acquisition of transforming DNA in Streptococcus pneumoniae. *FEMS microbiology reviews*, 33(3):627–642, 5 2009.

[447] Marc Prudhomme, Laetitia Attaiech, Guillaume Sanchez, Bernard Martin, and Jean-Pierre Claverys. Antibiotic stress induces genetic transformability in the human pathogen Streptococcus pneumoniae. *Science (New York, N.Y.)*, 313(5783):89–92, 7 2006.

[448] Jean Pierre Claverys, Marc Prudhomme, and Bernard Martin. Induction of competence regulons as a general response to stress in gram-positive bacteria. *Annual review of microbiology*, 60:451–475, 2006.

[449] Kathleen E. Stevens, Diana Chang, Erin E. Zwack, and Michael E. Sebert. Competence in Streptococcus pneumoniae is regulated by the rate of ribosomal decoding errors. *mBio*, 2(5), 2011.

[450] Bernard Martin, Anne Lise Soulet, Nicolas Mirouze, Marc Prudhomme, Isabelle Mortier-Barrière, Chantal Granadel, Marie Françoise Noirot-Gros, Philippe Noirot, Patrice Polard, and Jean Pierre Claverys. ComE/ComE~P interplay dictates activation or extinction status of pneumococcal X-state (competence). *Molecular microbiology*, 87(2):394–411, 1 2013.

[451] Erin Shanker and Michael J. Federle. Quorum Sensing Regulation of Competence and Bacteriocins in Streptococcus pneumoniae and mutans. *Genes*, 8(1), 1 2017.

[452] German Matias Traglia, Brettni Quinn, Sareda T.J. Schramm, Alfonso Soler-Bistue, and Maria Soledad Ramirez. Serum Albumin and Ca2+ Are Natural Competence Inducers in the Human Pathogen Acinetobacter baumannii. *Antimicrobial agents and chemotherapy*, 60(8):4920–4929, 8 2016.

[453] Casin Le, Camila Pimentel, Marisel R. Tuttobene, Tomas Subils, Brent Nishimura, German M. Traglia, Federico Perez, Krisztina M. Papp-Wallace, Robert A. Bonomo,

Marcelo E. Tolmasky, and Maria Soledad Ramirez. Interplay between meropenem and human serum albumin on expression of carbapenem resistance genes and natural competence in Acinetobacter baumannii. *Antimicrobial Agents and Chemotherapy*, 65(10), 10 2021.

[454] Calum Johnston, Nathalie Campo, Matthieu J. Bergé, Patrice Polard, and Jean Pierre Claverys. Streptococcus pneumoniae, le transformiste. *Trends in Microbiology*, 22(3):113–119, 3 2014.

[455] David Dubnau and Melanie Blokesch. Mechanisms of DNA Uptake by Naturally Competent Bacteria. *Annual review of genetics*, 53:217–237, 2019.

[456] Sandra Muschiol, Murat Balaban, Staffan Normark, and Birgitta Henriques-Normark. Uptake of extracellular DNA: Competence induced pili in natural transformation of Streptococcus pneumoniae. *BioEssays*, 37(4):426–435, 4 2015.

[457] Morten Kjos, Eric Miller, Jelle Slager, Frank B. Lake, Oliver Gericke, Ian S. Roberts, Daniel E. Rozen, and Jan Willem Veening. Expression of Streptococcus pneumoniae Bacteriocins Is Induced by Antibiotics via Regulatory Interplay with the Competence System. *PLoS Pathogens*, 12(2), 2 2016.

[458] Charles Y. Wang and Suzanne Dawid. Mobilization of Bacteriocins during Competence in Streptococci. *Trends in microbiology*, 26(5):389, 5 2018.

[459] Sébastien Guiral, Tim J. Mitchell, Bernard Martin, and Jean Pierre Claverys. Competence-programmed predation of noncompetent cells in the human pathogen Streptococcus pneumoniae: genetic requirements. *Proceedings of the National Academy of Sciences of the United States of America*, 102(24):8710–8715, 6 2005.

[460] Vegard Eldholm, Ola Johnsborg, Daniel Straume, Hilde Solheim Ohnstad, Kari Helene Berg, Juan A. Hermoso, and Leiv Sigve Håvarstein. Pneumococcal CbpD is a murein hydrolase that requires a dual cell envelope binding specificity to kill target cells during fratricide. *Molecular microbiology*, 76(4):905–917, 2010.

[461] Matthieu J. Bergé, Chryslène Mercy, Isabelle Mortier-Barrière, Michael S. Vannieuwenhze, Yves V. Brun, Christophe Grangeasse, Patrice Polard, and Nathalie Campo. A programmed cell division delay preserves genome integrity during natural genetic transformation in Streptococcus pneumoniae. *Nature Communications 2017 8:1*, 8(1):1–13, 11 2017.

[462] Jean Pierre Claverys, Bernard Martin, and Patrice Polard. The genetic transformation machinery: composition, localization, and mechanism. *FEMS Microbiology Reviews*, 33(3):643–656, 5 2009.

[463] Calum Johnston, Isabelle Mortier-Barriere, Vanessa Khemici, and Patrice Polard. Fine-tuning cellular levels of DprA ensures transformant fitness in the human pathogen Streptococcus pneumoniae. *Molecular Microbiology*, 109(5):663–675, 9 2018.

[464] Nicolas Mirouze, Mathieu A. Bergé, Anne Lise Soulet, Isabelle Mortier-Barrière, Yves Quentin, Gwennaele Fichant, Chantal Granadel, Marie Françoise Noirot-Gros, Philippe Noirot, Patrice Polard, Bernard Martin, and Jean Pierre Claverys. Direct involvement of DprA, the transformation-dedicated RecA loader, in the shut-off of pneumococcal competence. *Proceedings of the National Academy of Sciences of the United States of America*, 110(11):E1035, 3 2013.

[465] Jingjun Lin, Luchang Zhu, and Gee W. Lau. Disentangling competence for genetic transformation and virulence in Streptococcus pneumoniae. *Current Genetics*, 62(1):97–103, 2 2016.

[466] Nicholas J Croucher, Simon R Harris, Lars Barquist, Julian Parkhill, and Stephen D Bentley. A High-Resolution View of Genome-Wide Pneumococcal Transformation. *PLoS Pathogens*, 8(6):e1002745, 6 2012.

[467] William P. Hanage. Not So Simple After All: Bacteria, Their Population Genetics, and Recombination. *Cold Spring Harbor Perspectives in Biology*, 8(7):a018069, 7 2016.

[468] Calum Johnston, Bernard Martin, Patrice Polard, and Jean Pierre Claverys. Postreplication targeting of transformants by bacterial immune systems? *Trends in Microbiology*, 21(10):516–521, 10 2013.

[469] Sanford A. Lacks, Sahlu Ayalew, Adela G. De La Campa, and Bill Greenberg. Regulation of competence for genetic transformation in Streptococcus pneumoniae: expression of dpnA, a late competence gene encoding a DNA methyltransferase of the DpnII restriction system. *Molecular microbiology*, 35(5):1089–1098, 2000.

[470] Calum Johnston, Bernard Martin, Chantal Granadel, Patrice Polard, and Jean Pierre Claverys. Programmed protection of foreign DNA from restriction allows pathogenicity island exchange during pneumococcal transformation. *PLoS pathogens*, 9(2), 2 2013.

[471] Ole Herman Ambur, Jan Engelstädter, Pål J Johnsen, Eric L Miller, and Daniel E Rozen. Steady at the wheel: conservative sex and the benefits of bacterial transformation. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 371(1706), 2016.

[472] Melanie Blokesch. In and out—contribution of natural transformation to the shuffling of large genomic regions. *Current Opinion in Microbiology*, 38:22–29, 8 2017.

[473] R. J. Redfield. Evolution of natural transformation: testing the DNA repair hypothesis in Bacillus subtilis and Haemophilus influenzae. *Genetics*, 133(4):755–761, 4 1993.

[474] Sunita Sinha, Joshua Mell, and Rosemary Redfield. The availability of purine nucleotides regulates natural competence by controlling translation of the competence activator Sxy. *Molecular Microbiology*, 88(6):1106–1119, 6 2013.

[475] R J Redfield, M R Schrag, and A M Dean. The evolution of bacterial transformation: sex with poor relations. *Genetics*, 146(1):27–38, 5 1997.

[476] R. J. Redfield. Evolution of bacterial transformation: is sex with dead cells ever better than no sex at all? *Genetics*, 119(1):213–221, 5 1988.

[477] Jean Pierre Claverys and Leiv S. Håvarstein. Cannibalism and fratricide: mechanisms and raisons d'être. *Nature Reviews Microbiology 2007 5:3*, 5(3):219–229, 3 2007.

[478] Sandrine Borgeaud, Lisa C. Metzger, Tiziana Scrignari, and Melanie Blokesch. The type VI secretion system of Vibrio cholerae fosters horizontal gene transfer. *Science*, 347(6217):63–67, 1 2015.

[479] Marion S. Dorer, Jutta Fero, and Nina R. Salama. DNA damage triggers genetic exchange in Helicobacter pylori. *PLoS pathogens*, 6(7):1–10, 7 2010.

[480] Michiel Vos. Why do bacteria engage in homologous recombination? *Trends in Microbiology*, 2009.

[481] Elisheva Cohen, David A. Kessler, and Herbert Levine. Recombination dramatically speeds up evolution of finite populations. *Physical Review Letters*, 94(9):098102, 3 2005.

[482] Bruce R. Levin and Omar E. Cornejo. The population and evolutionary dynamics of homologous gene recombination in bacterial populations. *PLoS genetics*, 5(8), 8 2009.

[483] Tim F. Cooper. Recombination Speeds Adaptation by Reducing Competition between Beneficial Mutations in Populations of Escherichia coli. *PLOS Biology*, 5(9):e225, 9 2007.

[484] James Winkler and Katy C. Kao. Harnessing recombination to speed adaptive evolution in Escherichia coli. *Metabolic Engineering*, 14(5):487–495, 9 2012.

[485] Katinka J. Apagyi, Christophe Fraser, and Nicholas J. Croucher. Transformation asymmetry and the evolution of the bacterial accessory genome. *Molecular Biology and Evolution*, 2018.

[486] Joshua Chang Mell, Jae Yun Lee, Marlo Firme, Sunita Sinha, and Rosemary J. Redfie. Extensive cotransformation of natural variation into chromosomes of naturally competent Haemophilus influenzae. *G3: Genes, Genomes, Genetics*, 4(4):717–731, 4 2014.

[487] Nicholas J. Croucher, Paul G. Coupland, Abbie E. Stevenson, Alanna Callendrello, Stephen D. Bentley, and William P. Hanage. Diversification of bacterial genome content through distinct mechanisms over different timescales. *Nature Communications*, 2014.

[488] Ankur B. Dalia, Kimberley D. Seed, Stephen B. Calderwood, and Andrew Camilli. A globally distributed mobile genetic element inhibits natural transformation of Vibrio cholerae. *Proceedings of the National Academy of Sciences of the United States of America*, 112(33):10485–10490, 8 2015.

[489] N J Croucher, A J Page, T R Connor, A J Delaney, J A Keane, S D Bentley, J Parkhill, and S R Harris. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research*, 43(3), 2015.

[490] Xavier Didelot and Martin C.J. Maiden. Impact of recombination on bacterial evolution, 7 2010.

[491] Simon R. Harris, Ian N. Clarke, Helena M.B. Seth-Smith, Anthony W. Solomon, Lesley T. Cutcliffe, Peter Marsh, Rachel J. Skilton, Martin J. Holland, David Mabey, Rosanna W. Peeling, David A. Lewis, Brian G. Spratt, Magnus Unemo, Kenneth Persson, Carina Bjartling, Robert Brunham, Henry J.C. De Vries, Servaas A. Morré, Arjen Speksnijder, Cécile M. Bébéar, Maïté Clerc, Bertille De Barbeyrac, Julian Parkhill, and Nicholas R. Thomson. Whole genome analysis of diverse Chlamydia trachomatis strains identifies phylogenetic relationships masked by current clinical typing. *Nature genetics*, 44(4):413, 4 2012.

[492] Edward J. Feil, Edward C. Holmes, Debra E. Bessen, Man Suen Chan, Nicholas P.J. Day, Mark C. Enright, Richard Goldstein, Derek W. Hood, Awdhesh Kalia, Catrin E. Moore, Jiaji Zhou, and Brian G. Spratt. Recombination within natural

populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proceedings of the National Academy of Sciences of the United States of America*, 98(1):182–187, 1 2001.

[493] J Maynard Smith. The detection and measurement of recombination from sequence data. *Genetics*, 153(2):1021–1027, 10 1999.

[494] Philip Awadalla. The evolutionary genomics of pathogen recombination. *Nature Reviews Genetics 2003 4:1*, 4(1):50–60, 1 2003.

[495] Kelly L. Wyres and Kathryn E. Holt. Klebsiella pneumoniae Population Genomics and Antimicrobial-Resistant Clones. *Trends in Microbiology*, 24(12):944–956, 12 2016.

[496] Xavier Didelot, David W. Eyre, Madeleine Cule, Camilla L.C. Ip, M. Azim Ansari, David Griffiths, Alison Vaughan, Lily O'Connor, Tanya Golubchik, Elizabeth M. Batty, Paolo Piazza, Daniel J. Wilson, Rory Bowden, Peter J. Donnelly, Kate E. Dingle, Mark Wilcox, A. Sarah Walker, Derrick W. Crook, Tim E.A. Peto, and Rosalind M. Harding. Microevolutionary analysis of Clostridium difficile genomes to investigate transmission. *Genome biology*, 13(12):R118, 2012.

[497] Xavier Didelot and Julian Parkhill. A scalable analytical approach from bacterial genomes to epidemiology. *bioRxiv*, page 2021.11.19.469232, 11 2021.

[498] Darren P. Martin, Philippe Lemey, and David Posada. Analysing recombination in nucleotide sequences. *Molecular Ecology Resources*, 11(6):943–955, 11 2011.

[499] David Posada, Keith A. Crandall, and Edward C. Holmes. Recombination in Evolutionary Genomics. *http://dx.doi.org/10.1146/annurev.genet.36.040202.111115*, 36:75–97, 11 2003.

[500] Carsten Wiuf, Thomas Christensen, and Jotun Hein. A Simulation Study of the Reliability of Recombination Detection Methods. *Molecular Biology and Evolution*, 18(10):1929–1939, 10 2001.

[501] P. H.A. Sneath, M. J. Sackin, and R. P. Ambler. Detecting Evolutionary Incompatibilities From Protein Sequences. *Systematic Biology*, 24(3):311–332, 9 1975.

[502] J C Stephens. Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. *Molecular Biology and Evolution*, 2(6):539–556, 11 1985.

[503] Paul L Bollyky, Andrew Rambaut, Paul H Harvey, and Edward C Holmes. Recombination between sequences of hepatitis B virus from different genotypes. *Journal of Molecular Evolution*, 42(2):97–102, 1996.

[504] David L Robertson, Paul M Sharp, Francine E McCutchan, and Beatrice H Hahn. Recombination in HIV-1. *Nature*, 374(6518):124–126, 1995.

[505] E C Holmes, M Worobey, and A Rambaut. Phylogenetic evidence for recombination in dengue virus. *Molecular Biology and Evolution*, 16(3):405–409, 3 1999.

[506] J. Maynard Smith and N. H. Smith. Detecting recombination from gene trees. *Molecular biology and evolution*, 15(5):590–599, 1998.

[507] Ulrich Nübel, Philippe Roumagnac, Mirjam Feldkamp, Jae Hoon Song, Kwan Soo Ko, Yhu Chering Huang, Geoffrey Coombs, Margaret Ip, Henrik Westh, Robert Skov, Marc J. Struelens, Richard V. Goering, Birgit Strommenger, Annette Weller, Wolfgang Witte, and Mark Achtman. Frequent emergence and limited geographic

dispersal of methicillin- resistant Staphylococcus aureus. *Proceedings of the National Academy of Sciences of the United States of America*, 105(37):14130–14135, 9 2008.

[508] Simon R. Harris, Edward J. Feil, Matthew T.G. Holden, Michael A. Quail, Emma K. Nickerson, Narisara Chantratita, Susana Gardete, Ana Tavares, Nick Day, Jodi A. Lindsay, Jonathan D. Edgeworth, Hermínia De Lencastre, Julian Parkhill, Sharon J. Peacock, and Stephen D. Bentley. Evolution of MRSA During Hospital Transmission and Intercontinental Spread. *Science (New York, N.Y.)*, 327(5964):469, 1 2010.

[509] M. Azim Ansari and Xavier Didelot. Inference of the properties of the recombination process from whole bacterial genomes. *Genetics*, 196(1):253–265, 1 2014.

[510] Brian J. Arnold, Michael U. Gutmann, Yonatan H. Grad, Samuel K. Sheppard, Jukka Corander, Marc Lipsitch, and William P. Hanage. Weak epistasis may drive adaptation in recombining bacteria. *Genetics*, 208(3):1247–1260, 3 2018.

[511] Mingzhi Lin and Edo Kussell. Inferring bacterial recombination rates from large-scale sequencing datasets. *Nature Methods 2019 16:2*, 16(2):199–204, 1 2019.

[512] David H.A. Fitch and Morris Goodman. Phylogenetic scanning: a computer-assisted algorithm for mapping gene conversions and other recombinational events. *Bioinformatics*, 7(2):207–215, 4 1991.

[513] Nicholas C. Grassly and Edward C. Holmes. A likelihood method for the detection of selection and recombination using nucleotide sequences. *Molecular Biology and Evolution*, 14(3):239–247, 3 1997.

[514] D. P. Martin, D. Posada, K. A. Crandall, and C. Williamson. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS research and human retroviruses*, 21(1):98–102, 1 2005.

[515] Darren P. Martin, Philippe Lemey, Martin Lott, Vincent Moulton, David Posada, and Pierre Lefeuvre. RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics*, 26(19):2462, 8 2010.

[516] Marc A. Suchard, Robert E. Weiss, Karin S. Dorman, and Janet S. Sinsheimer. Oh Brother, Where Art Thou? A Bayes Factor Test for Recombination with Uncertain Heritage. *Systematic Biology*, 51(5):715–728, 9 2002.

[517] Dirk Husmeier and Gráinne McGuire. Detecting recombination with MCMC. *Bioinformatics*, 18(suppl_1):S345–S353, 7 2002.

[518] Vladimir N. Minin, Karin S. Dorman, Fang Fang, and Marc A. Suchard. Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics*, 21(13):3034–3042, 7 2005.

[519] Sergei L.Kosakovsky Pond, David Posada, Michael B. Gravenor, Christopher H. Woelk, and Simon D.W. Frost. Automated Phylogenetic Detection of Recombination Using a Genetic Algorithm. *Molecular Biology and Evolution*, 23(10):1891–1901, 10 2006.

[520] John Maynard Smith. Analyzing the mosaic structure of genes. *Journal of molecular evolution*, 34(2):126–129, 1992.

[521] C. G. Dowson, A. Hutchison, J. A. Brannigan, R. C. George, D. Hansman, J. Linares, A. Tomasz, J. M. Smith, and B. G. Spratt. Horizontal transfer of penicillin-

binding protein genes in penicillin-resistant clinical isolates of Streptococcus pneumoniae. *Proceedings of the National Academy of Sciences of the United States of America*, 86(22):8842–8846, 1989.

[522] R. Milkman and M. M. Bridges. Molecular evolution of the Escherichia coli chromosome. III. Clonal frames. *Genetics*, 126(3):505–517, 11 1990.

[523] Xavier Didelot and Daniel Falush. Inference of bacterial microevolution using multilocus sequence data. *Genetics*, 175(3):1251–1266, 3 2007.

[524] Xavier Didelot and Daniel J Wilson. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS computational biology*, 11(2):e1004041, 2 2015.

[525] Pekka Marttinen, William P. Hanage, Nicholas J. Croucher, Thomas R. Connor, Simon R. Harris, Stephen D. Bentley, and Jukka Corander. Detection of recombination events in bacterial genomes from large population samples. *Nucleic acids research*, 40(1), 1 2012.

[526] Rafal Mostowy, Nicholas J. Croucher, Cheryl P. Andam, Jukka Corander, William P. Hanage, and Pekka Marttinen. Efficient Inference of Recent and Ancestral Recombination within Bacterial Populations. *Molecular Biology and Evolution*, 34(5):1167–1182, 5 2017.

[527] Jukka Corander, Patrik Waldmann, and Mikko J. Sillanpää. Bayesian analysis of genetic differentiation between populations. *Genetics*, 163(1):367–374, 1 2003.

[528] Andrew J. Page, Ben Taylor, Aidan J. Delaney, Jorge Soares, Torsten Seemann, Jacqueline A. Keane, and Simon R. Harris. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial genomics*, 2(4):e000056, 4 2016.

[529] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 5(3):e9490, 3 2010.

[530] Alexandros Stamatakis. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 5 2014.

[531] Haim Ashkenazy, Osnat Penn, Adi Doron-Faigenboim, Ofir Cohen, Gina Cannarozzi, Oren Zomer, and Tal Pupko. FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Research*, 40(W1):W580–W584, 7 2012.

[532] Tal Pupko, Itsik Pe'er, Ron Shamir, and Dan Graur. A Fast Algorithm for Joint Reconstruction of Ancestral Amino Acid Sequences. *Molecular Biology and Evolution*, 17(6):890–896, 6 2000.

[533] D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2):131–147, 2 1981.

[534] Gubbins :: Anaconda.org.

[535] Alexey M. Kozlov, Diego Darriba, Tomáš Flouri, Benoit Morel, and Alexandros Stamatakis. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21):4453, 11 2019.

[536] Carla Rodrigues, Siddhi Desai, Virginie Passet, Devarshi Gajjar, and Sylvain Brisse. Genomic evolution of the globally disseminated multidrug-resistant Klebsiella pneumoniae clonal group 147. *Microbial Genomics*, 8(1):737, 1 2022.

[537] Joseph Elikem Efui Acolatse, Edward A.R. Portal, Ian Boostrom, George Akafity, Mavis Puopelle Dakroah, Victoria J. Chalker, Kirsty Sands, and Owen B. Spiller. Environmental surveillance of ESBL and carbapenemase-producing gram-negative bacteria in a Ghanaian Tertiary Hospital. *Antimicrobial Resistance and Infection Control*, 11(1):1–15, 12 2022.

[538] Samiratu Mahazu, Wakana Sato, Alafate Ayibieke, Isaac Prah, Takaya Hayashi, Toshihiko Suzuki, Shiroh Iwanaga, Anthony Ablordey, and Ryoichi Saito. Insights and genetic features of extended-spectrum beta-lactamase producing Escherichia coli isolates from two hospitals in Ghana. *Scientific Reports 2022 12:1*, 12(1):1–11, 2 2022.

[539] Sophie Hoffman, Zena Lapp, Joyce Wang, and Evan S. Snitkin. regentrans: a framework and R package for using genomics to study regional pathogen transmission. *Microbial genomics*, 8(1):000747, 1 2022.

[540] Lam-Tung Nguyen, Heiko A Schmidt, Arndt von Haeseler, and Bui Quang Minh. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1):268–274, 1 2015.

[541] Martin Simonsen, Thomas Mailund, and Christian N S Pedersen. Rapid Neighbour-Joining. In Keith A Crandall and Jens Lagergren, editors, *Algorithms in Bioinformatics*, pages 113–122, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[542] simonrharris/pyjar: A python implementation of the joint ancestral state reconstruction algorithm of Pupko et al.

[543] Aleksi Sipola, Pekka Marttinen, and Jukka Corander. Bacmeta: simulator for genomic evolution in bacterial metapopulations. *Bioinformatics (Oxford, England)*, 34(13):2308–2310, 7 2018.

[544] Nicola De Maio and Daniel J. Wilson. The bacterial sequential markov coalescent. *Genetics*, 206(1):333–343, 5 2017.

[545] Min Jung Kwun, Marco R Oggioni, Megan De Ste Croix, Stephen D Bentley, and Nicholas J Croucher. Excision-reintegration at a pneumococcal phase-variable restriction-modification locus drives within- and between-strain epigenetic differentiation and inhibits gene acquisition. *Nucleic Acids Research*, 2018.

[546] Harris S R. SKA: Split Kmer Analysis Toolkit for Bacterial Genomic Epidemiology. *bioRxiv*, page 453142, 10 2018.

[547] Michelle Kendall and Caroline Colijn. Mapping Phylogenetic Trees to Reveal Distinct Patterns of Evolution. *Molecular Biology and Evolution*, 33(10):2735–2743, 10 2016.

[548] Jessica Hedge and Daniel J. Wilson. Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. *mBio*, 5(6), 11 2014.

[549] Kathryn Holt, Johanna J. Kenyon, Mohammad Hamidian, Mark B. Schultz, Derek J. Pickard, Gordon Dougan, and Ruth Hall. Five decades of genome evolution in the globally distributed, extensively antibiotic-resistant Acinetobacter baumannii global clone 1. *Microbial Genomics*, 2(2):1–16, 2 2016.

[550] Pedro González-Torres, Francisco Rodríguez-Mateos, Josefa Antón, and Toni Gabaldón. Impact of Homologous Recombination on the Evolution of Prokaryotic Core

Genomes. *mBio*, 10(1), 1 2019.

[551] Xiaofan Zhou, Xing Xing Shen, Chris Todd Hittinger, and Antonis Rokas. Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets. *Molecular Biology and Evolution*, 35(2):486, 2 2018.

[552] Xavier Didelot, Nicholas J Croucher, Stephen D Bentley, Simon R Harris, and Daniel J Wilson. Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Research*, 2018.

[553] Blanca M. Perez-Sepulveda, Darren Heavens, Caisey V. Pulford, Alexander V. Predeus, Ross Low, Hermione Webster, Gregory F. Dykes, Christian Schudoma, Will Rowe, James Lipscombe, Chris Watkins, Benjamin Kumwenda, Neil Shearer, Karl Costigan, Kate S. Baker, Nicholas A. Feasey, Jay C.D. Hinton, Neil Hall, Blanca M. Perez-Sepulveda, Darren Heavens, Caisey V. Pulford, María Teresa Acuña, Dragan Antic, Martin Antonio, Kate S. Baker, Johan Bernal, Hilda Bolaños, Marie Chattaway, John Cheesbrough, Angeziwa Chirambo, Karl Costigan, Saffiatou Darboe, Paula Díaz, Pilar Donado, Carolina Duarte, Francisco Duarte, Dean Everett, Séamus Fanning, Nicholas A. Feasey, Patrick Feglo, Adriano M. Ferreira, Rachel Floyd, Ronnie G. Gavilán, Melita A. Gordon, Neil Hall, Rodrigo T. Hernandes, Gabriela Hernández-Mora, Jay C.D. Hinton, Daniel Hurley, Irene N. Kasumba, Benjamin Kumwenda, Brenda Kwambana-Adams, James Lipscombe, Ross Low, Salim Mattar, Lucy Angeline Montaño, Cristiano Gallina Moreira, Jaime Moreno, Dechamma Mundanda Muthappa, Satheesh Nair, Chris M. Parry, Chikondi Peno, Jasnehta Permala-Booth, Jelena Petrović, Alexander V. Predeus, José Luis Puente, Getenet Rebrie, Martha Redway, Will Rowe, Terue Sadatsune, Christian Schudoma, Neil Shearer, Claudia Silva, Anthony M. Smith, Sharon Tennant, Alicia Tran-Dien, Chris Watkins, Hermione Webster, François Xavier Weill, Magdalena Wiesner, and Catherine Wilson. An accessible, efficient and global approach for the large-scale sequencing of bacterial genomes. *Genome Biology*, 22(1), 12 2021.

[554] Grace A. Blackwell, Martin Hunt, Kerri M. Malone, Leandro Lima, Gal Horesh, Blaise T.F. Alako, Nicholas R. Thomson, and Zamin Iqbal. Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. *PLOS Biology*, 19(11):e3001421, 11 2021.

[555] Nicholas J Croucher, Claire Chewapreecha, William P Hanage, Simon R Harris, Lesley McGee, Mark van der Linden, Jae-Hoon Song, Kwan Soo Ko, Herminia de Lencastre, Claudia Turner, Fan Yang, Raquel Sá-Leão, Bernard Beall, Keith P Klugman, Julian Parkhill, Paul Turner, and Stephen D Bentley. Evidence for soft selective sweeps in the evolution of pneumococcal multidrug resistance and vaccine escape. *Genome biology and evolution*, 6(7):1589–1602, 7 2014.

[556] Joshua C D'aeth, Mark P G Van Der Linden, Lesley Mcgee, Herminia De Lencastre, Paul Turner, Jaehoon Song, Stephanie W Lo, Rebecca A Gladstone, Raquel Sáleão, Kwan Soo Ko, William P Hanage, Bernard Beall, Stephen D Bentley, Nicholas J Croucher, and The Gps Consortium. The Role of Interspecies recombinations in the evolution of antibioticresistant pneumococci. *bioRxiv*, page 2021.02.22.432219, 2 2021.

[557] Jeffrey P. Maskell, Armine M. Sefton, and Lucinda M.C. Hall. Mechanism of sulfonamide resistance in clinical isolates of Streptococcus pneumoniae. *Antimicrobial Agents and Chemotherapy*, 41(10):2121–2126, 1997.

[558] Y. Haasum, K. Ström, R. Wehelie, V. Luna, M. C. Roberts, J. P. Maskell, L. M.C. Hall, and G. Swedberg. Amino acid repetitions in the dihydropteroate synthase of Streptococcus pneumoniae lead to sulfonamide resistance with limited effects on substrate Km. *Antimicrobial Agents and Chemotherapy*, 45(3):805–809, 2001.

[559] Jennifer E. Cornick, Simon R. Harris, Christopher M. Parry, Michael J. Moore, Chikondi Jassi, Arox Kamng'ona, Benard Kulohoma, Robert S. Heyderman, Stephen D. Bentley, and Dean B. Everett. Genomic identification of a novel cotrimoxazole resistance genotype and its prevalence amongst Streptococcus pneumoniae in Malawi. *Journal of Antimicrobial Chemotherapy*, 69(2):368–374, 2 2014.

[560] C. G. Dowson, A. Hutchison, N. Woodford, A. P. Johnson, R. C. George, and B. G. Spratt. Penicillin-resistant viridans streptococci have obtained altered penicillin-binding protein genes from penicillin-resistant strains of Streptococcus pneumoniae. *Proceedings of the National Academy of Sciences of the United States of America*, 87(15):5858–5862, 1990.

[561] Christian J.H. Von Wintersdorff, John Penders, Julius M. Van Niekerk, Nathan D. Mills, Snehali Majumder, Lieke B. Van Alphen, Paul H.M. Savelkoul, and Petra F.G. Wolffs. Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer, 2 2016.

[562] M C Enright and B G Spratt. Extensive variation in the ddl gene of penicillin-resistant Streptococcus pneumoniae results from a hitchhiking effect driven by the penicillin-binding protein 2b gene. *Molecular Biology and Evolution*, 16(12):1687–1695, 12 1999.

[563] Rebecca A. Gladstone, Stephanie W. Lo, John A. Lees, Nicholas J. Croucher, Andries J. van Tonder, Jukka Corander, Andrew J. Page, Pekka Marttinen, Leon J. Bentley, Theresa J. Ochoa, Pak Leung Ho, Mignon du Plessis, Jennifer E. Cornick, Brenda Kwambana-Adams, Rachel Benisty, Susan A. Nzenze, Shabir A. Madhi, Paulina A. Hawkins, Dean B. Everett, Martin Antonio, Ron Dagan, Keith P. Klugman, Anne von Gottberg, Lesley McGee, Robert F. Breiman, and Stephen D. Bentley. International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. *EBioMedicine*, 43:338–346, 5 2019.

[564] Mark C. Enright, Asunción Fenoll, David Griffiths, and Brian G. Spratt. The Three Major Spanish Clones of Penicillin-Resistant Streptococcus pneumoniae Are the Most Common Clones Recovered in Recent Cases of Meningitis in Spain. *Journal of Clinical Microbiology*, 37(10):3210, 1999.

[565] Tracey J. Coffey, Margaret Daniels, Mark C. Enright, and Brian G. Spratt. Serotype 14 variants of the Spanish penicillin-resistant serotype 9V clone of Streptococcus pneumoniae arose by large recombinational replacements of the cpsA-pbp1a region. *Microbiology*, 145(8):2023–2031, 8 1999.

[566] Johanna MC Jefferies, Emily Macdonald, Saul N Faust, and Stuart C Clarke. 13-valent pneumococcal conjugate vaccine (PCV13). *Human Vaccines*, 7(10):1012–1018, 10 2011.

[567] Jordi Càmara, Meritxell Cubero, Antonio J Martín-Galiano, Ernesto García, Imma Grau, Jesper B Nielsen, Peder Worning, Fe Tubau, Román Pallarés, M Ángeles Domínguez, Mogens Kilian, Josefina Liñares, Henrik Westh, and Carmen Ardanuy. Evolution of the β-lactam-resistant Streptococcus pneumoniae PMEN3 clone over

a 30□year period in Barcelona, Spain. *Journal of Antimicrobial Chemotherapy*, 73(11):2941–2951, 11 2018.

[568] Aida González-Díaz, Miguel P. Machado, Jordi Càmara, José Yuste, Emmanuelle Varon, Miriam Domenech, María Del Grosso, José María Marimón, Emilia Cercenado, Nieves Larrosa, María Dolores Quesada, Dionisia Fontanals, Assiya El-Mniai, Meritxell Cubero, João A. Carriço, Sara Martí, Mario Ramirez, and Carmen Ardanuy. Two multi-fragment recombination events resulted in the β-lactam-resistant serotype 11A-ST6521 related to Spain9V-ST156 pneumococcal clone spreading in south-western Europe, 2008 to 2016. *Eurosurveillance*, 25(16), 4 2020.

[569] T J Coffey, C G Dowson, M Daniels, J Zhou, C Martin, B G Spratt, and J M Musser. Horizontal transfer of multiple penicillin-binding protein genes, and capsular biosynthetic genes, in natural populations of Streptococcus pneumoniae. *Molecular Microbiology*, 5(9):2255–2260, 9 1991.

[570] J C Lefèvre, M A Bertrand, and G Faucon. Molecular analysis by pulsed-field gel electrophoresis of penicillin-resistant Streptococcus pneumoniae from Toulouse, France. *European Journal of Clinical Microbiology & Infectious Diseases: Official Publication of the European Society of Clinical Microbiology*, 14(6):491–497, 6 1995.

[571] A Corso, E P Severina, V F Petruk, Y R Mauriz, and A Tomasz. Molecular characterization of penicillin-resistant Streptococcus pneumoniae isolates causing respiratory disease in the United States. *Microbial Drug Resistance (Larchmont, N.Y.)*, 4(4):325–337, 1998.

[572] G Gherardi, C G Whitney, R R Facklam, and B Beall. Major related sets of antibiotic-resistant Pneumococci in the United States as determined by pulsed-field gel electrophoresis and pbp1a-pbp2b-pbp2x-dhf restriction profiles. *The Journal of Infectious Diseases*, 181(1):216–229, 1 2000.

[573] R Sá-Leão, A Tomasz, I S Sanches, A Brito-Avô, S E Vilhelmsson, K G Kristinsson, and H de Lencastre. Carriage of internationally spread clones of Streptococcus pneumoniae with unusual drug resistance patterns in children attending day care centers in Lisbon, Portugal. *The Journal of Infectious Diseases*, 182(4):1153–1160, 10 2000.

[574] Maria N Tsolia, George Stamos, Sophia Ioannidou, Ronit Trefler, Maria Foustoukou, Dimitris Kafetzis, and Nurith Porat. Genetic relatedness of resistant and multiresistant Streptococcus pneumoniae strains, recovered in the Athens area, to international clones. *Microbial Drug Resistance (Larchmont, N.Y.)*, 8(3):219–226, 2002.

[575] Stephanie J Schrag, Lesley McGee, Cynthia G Whitney, Bernard Beall, Allen S Craig, Miriam E Choate, James H Jorgensen, Richard R Facklam, Keith P Klugman, and the Active Bacterial Core Surveillance Team. Emergence of Streptococcus pneumoniae with Very-High-Level Resistance to Penicillin. *Antimicrobial Agents and Chemotherapy*, 48(8):3016–3023, 8 2004.

[576] Lindsay Kim, Lesley McGee, Sara Tomczyk, and Bernard Beall. Biological and Epidemiological Features of Antibiotic-Resistant Streptococcus pneumoniae in Pre- and Post-Conjugate Vaccine Eras: a United States Perspective. *Clinical microbiology reviews*, 29(3):525–552, 7 2016.

[577] M Del Grosso, F Iannelli, C Messina, M Santagati, N Petrosillo, S Stefani, G Pozzi, and A Pantosti. Macrolide efflux genes mef(A) and mef(E) are carried by different genetic elements in Streptococcus pneumoniae. *Journal of clinical microbiology*, 40(3):774–778, 3 2002.

[578] Christine Bley, Mark van der Linden, and Ralf René Reinert. mef(A) is the predominant macrolide resistance determinant in Streptococcus pneumoniae and Streptococcus pyogenes in Germany. *International Journal of Antimicrobial Agents*, 37(5):425–431, 2011.

[579] Matthias Imöhl, Ralf René Reinert, Christina Mutscher, and Mark van der Linden. Macrolide susceptibility and serotype specific macrolide resistance of invasive isolates of Streptococcus pneumoniae in Germany from 1992 to 2008. *BMC Microbiology*, 10(1):299, 2010.

[580] Paul Turner, Claudia Turner, Auscharee Jankhot, Naw Helen, Sue J. Lee, Nicholas P. Day, Nicholas J. White, Francois Nosten, and David Goldblatt. A Longitudinal Study of Streptococcus pneumoniae Carriage in a Cohort of Infants and Their Mothers on the Thailand-Myanmar Border. *PLoS ONE*, 7(5):e38271, 5 2012.

[581] Mark C. Enright and Brian G. Spratt. A multilocus sequence typing scheme for Streptococcus pneumoniae: Identification of clones associated with serious invasive disease. *Microbiology*, 144(11):3049–3060, 11 1998.

[582] L McGee, L McDougal, J Zhou, B G Spratt, F C Tenover, R George, R Hakenbeck, W Hryniewicz, J C Lefévre, A Tomasz, and K P Klugman. Nomenclature of major antimicrobial-resistant clones of Streptococcus pneumoniae defined by the pneumococcal molecular epidemiology network. *Journal of Clinical Microbiology*, 39(7):2565–2571, 7 2001.

[583] Lennard Epping, Andries J. van Tonder, Rebecca A. Gladstone, Stephen D. Bentley, Andrew J. Page, and Jacqueline A. Keane. SeroBA: Rapid high-throughput serotyping of Streptococcus pneumoniae from whole genome sequence data. *Microbial Genomics*, 4(7), 7 2018.

[584] Andrew J. Page, Nishadi De Silva, Martin Hunt, Michael A. Quail, Julian Parkhill, Simon R. Harris, Thomas D. Otto, and Jacqueline A. Keane. Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. *Microbial genomics*, 2(8):e000083, 8 2016.

[585] Daniel R. Zerbino and Ewan Birney. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821–829, 5 2008.

[586] Marten Boetzer and Walter Pirovano. SSPACE-LongRead: Scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics*, 15(1):211, 6 2014.

[587] Marten Boetzer and Walter Pirovano. Toward almost closed genomes with GapFiller. *Genome Biology*, 13(6):R56, 6 2012.

[588] Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, 7 2014.

[589] Nicholas J. Croucher, Andrew J. Page, Thomas R. Connor, Aidan J. Delaney, Jacqueline A. Keane, Stephen D. Bentley, Julian Parkhill, and Simon R. Harris. Rapid phylogenetic analysis of large samples of recombinant bacterial whole

genome sequences using Gubbins. *Nucleic Acids Research*, 43(3):e15–e15, 2 2015.

[590] Sohta A. Ishikawa, Anna Zhukova, Wataru Iwasaki, Olivier Gascuel, and Tal Pupko. A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios. *Molecular Biology and Evolution*, 36(9):2069–2085, 9 2019.

[591] Marvin N. Wright and Andreas Ziegler. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 3 2017.

[592] Liam J. Revell. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2):217–223, 4 2012.

[593] Sean R Eddy. Accelerated Profile HMM Searches. *PLOS Computational Biology*, 7(10):e1002195, 10 2011.

[594] Peter V. Adrian and Keith P. Klugman. Mutations in the dihydrofolate reductase gene of trimethoprim-resistant isolates of Streptococcus pneumoniae. *Antimicrobial Agents and Chemotherapy*, 41(11):2406–2413, 1997.

[595] B J Metcalf, S Chochua, R E Jr Gertz, Z Li, H Walker, T Tran, P A Hawkins, A Glennen, R Lynfield, Y Li, L McGee, and B Beall. Using whole genome sequencing to identify resistance determinants and predict antimicrobial resistance phenotypes for year 2015 invasive pneumococcal disease isolates recovered in the United States. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 22(12):1–1002, 12 2016.

[596] Raúl A. González-Pech, Timothy G. Stephens, and Cheong Xin Chan. Commonly misunderstood parameters of NCBI BLAST and important considerations for users. *Bioinformatics*, 35(15):2697–2698, 8 2019.

[597] Robert C Edgar. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5(1):113, 2004.

[598] Rafal J Mostowy, Nicholas J Croucher, Nicola De Maio, Claire Chewapreecha, Susannah J Salter, Paul Turner, David M Aanensen, Stephen D Bentley, Xavier Didelot, and Christophe Fraser. Pneumococcal Capsule Synthesis Locus cps as Evolutionary Hotspot with Potential to Generate Novel Serotypes by Recombination. *Molecular biology and evolution*, 34(10):2537–2554, 2017.

[599] S. J. Salter, J. Hinds, K. A. Gould, L. Lambertsen, W. P. Hanage, M. Antonio, P. Turner, P. W. M. Hermans, H. J. Bootsma, K. L. O'Brien, and S. D. Bentley. Variation at the capsule locus, cps, of mistyped and non-typable Streptococcus pneumoniae isolates. *Microbiology*, 158(Pt 6):1560, 2012.

[600] S R Filipe, E Severina, and A Tomasz. Distribution of the mosaic structured murM genes among natural populations of Streptococcus pneumoniae. *Journal of bacteriology*, 182(23):6798–805, 12 2000.

[601] S. R. Filipe and A. Tomasz. Inhibition of the expression of penicillin resistance in Streptococcus pneumoniae by inactivation of cell wall muropeptide branching genes. *Proceedings of the National Academy of Sciences*, 97(9):4891–4896, 4 2000.

[602] Carlos J. Sanchez, Pooja Shivshankar, Kim Stol, Samuel Trakhtenbroit, Paul M. Sullam, Karin Sauer, Peter W.M. Hermans, and Carlos J. Orihuela. The Pneumo-

coccal Serine-Rich Repeat Protein Is an Intra-Species Bacterial Adhesin That Promotes Bacterial Aggregation In Vivo and in Biofilms. *PLoS Pathogens*, 6(8):33–34, 8 2010.

[603] Dimitrios Latousakis, Donald A. MacKenzie, Andrea Telatin, and Nathalie Juge. Serine-rich repeat proteins from gut microbes. *Gut Microbes*, 11(1):102, 1 2020.

[604] Pooja Shivshankar, Carlos Sanchez, Lloyd F Rose, and Carlos J Orihuela. The Streptococcus pneumoniae adhesin PsrP binds to Keratin 10 on lung cells. *Molecular Microbiology*, 73(4):663, 8 2009.

[605] Shanshan Du, Cláudia Vilhena, Samantha King, Alfredo Sahagún-Ruiz, Sven Hammerschmidt, Christine Skerka, and Peter F. Zipfel. Molecular analyses identifies new domains and structural differences among Streptococcus pneumoniae immune evasion proteins PspC and Hic. *Scientific Reports*, 11(1), 12 2021.

[606] Nicholas J. Croucher, Joseph J. Campo, Timothy Q. Le, Xiaowu Liang, Stephen D. Bentley, William P. Hanage, and March Lipsitch. Diverse evolutionary patterns of pneumococcal antigens identified by pangenome-wide immunological screening. *Proceedings of the National Academy of Sciences of the United States of America*, 114(3):E357–E366, 1 2017.

[607] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22):10915–10919, 1992.

[608] Streptococcus Lab | StrepLab | MIC tables and sequences | CDC - https://www.cdc.gov/streplab/pneumococcus/mic.html Date accessed: 14/01/2021.

[609] Effects of New Penicillin Susceptibility Breakpoints for Streptococcus pneumoniae United States, 2006-2007 https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5750a2.htm Date accessed: 26/01/2021.

[610] André Zapun, Carlos Contreras-Martel, and Thierry Vernet. Penicillin-binding proteins and β-lactam resistance. *FEMS Microbiology Reviews*, 32(2):361–385, 3 2008.

[611] Brodie Daniels, Anna Coutsoudis, Eshia Moodley-Govender, Helen Mulol, Elizabeth Spooner, Photini Kiepiela, Shabashini Reddy, Linda Zako, Nhan T. Ho, Louise Kuhn, and Gita Ramjee. Effect of co-trimoxazole prophylaxis on morbidity and mortality of HIV-exposed, HIV-uninfected infants in South Africa: a randomised controlled, non-inferiority trial. *The Lancet Global Health*, 7(12):e1717–e1727, 12 2019.

[612] Nicholas J Croucher, Jonathan A Finkelstein, Stephen I Pelton, Patrick K Mitchell, Grace M Lee, Julian Parkhill, Stephen D Bentley, William P Hanage, and Marc Lipsitch. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nature genetics*, 45(6):656–663, 11 2013.

[613] Nicholas J Croucher, Andrea M Mitchell, Katherine A Gould, Donald Inverarity, Lars Barquist, Theresa Feltwell, Maria C Fookes, Simon R Harris, Janina Dordel, Susannah J Salter, Sarah Browall, Helena Zemlickova, Julian Parkhill, Staffan Normark, Birgitta Henriques-Normark, Jason Hinds, Tim J Mitchell, and Stephen D Bentley. Dominant role of nucleotide substitution in the diversification of serotype 3 pneumo-

cocci over decades and during a single infection. *PLoS genetics*, 9(10):e1003868, 2013.

[614] Cheryl P. Andam, Patrick K. Mitchell, Alanna Callendrello, Qiuzhi Chang, Jukka Corander, Chrispin Chaguza, Lesley McGee, Bernard W. Beall, and William P. Hanage. Genomic epidemiology of penicillin- nonsusceptible pneumococci with nonvaccine serotypes causing invasive disease in the United States. *Journal of Clinical Microbiology*, 55(4):1104–1115, 4 2017.

[615] Mignon du Plessis, Edouard Bingen, and Keith P Klugman. Analysis of penicillin-binding protein genes of clinical isolates of Streptococcus pneumoniae with reduced susceptibility to amoxicillin. *Antimicrobial agents and chemotherapy*, 46(8):2349–57, 8 2002.

[616] A. M. Smith and K. P. Klugman. Alterations in MurM, a cell wall muropeptide branching enzyme, increase high-level penicillin and cephalosporin resistance in Streptococcus pneumoniae. *Antimicrobial Agents and Chemotherapy*, 45(8):2393–2396, 2001.

[617] Marcin J. Skwark, Nicholas J. Croucher, Santeri Puranen, Claire Chewapreecha, Maiju Pesonen, Ying Ying Xu, Paul Turner, Simon R. Harris, Stephen B. Beres, James M. Musser, Julian Parkhill, Stephen D. Bentley, Erik Aurell, and Jukka Corander. Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis. *PLOS Genetics*, 13(2):e1006508, 2 2017.

[618] AMIMA - https://amr.poppunk.net/ Date accessed: 08/07/2022.

[619] Alessandra Carattoli. Resistance Plasmid Families in Enterobacteriaceae. *Antimicrobial Agents and Chemotherapy*, 53(6):2227, 6 2009.

[620] Gisele Peirano and Johann D.D. Pitout. Molecular epidemiology of Escherichia coli producing CTX-M β-lactamases: the worldwide emergence of clone ST131 O25:H4. *International Journal of Antimicrobial Agents*, 35(4):316–321, 4 2010.

[621] M. D. Smith and W. R. Guild. A plasmid in Streptococcus pneumoniae. *Journal of Bacteriology*, 137(2):735–739, 1979.

[622] Patricia Romero, Daniel Llull, Ernesto García, Tim J. Mitchell, Rubens López, and Miriam Moscoso. Isolation and characterization of a new plasmid pSpnP1 from a multidrug-resistant clone of Streptococcus pneumoniae. *Plasmid*, 58(1):51–60, 7 2007.

[623] C Schuster, M van der Linden, and R Hakenbeck. Small cryptic plasmids of Streptococcus pneumoniae belong to the pC194/pUB110 family of rolling circle plasmids. *FEMS microbiology letters*, 164(2):427–431, 7 1998.

[624] Sanin Musovic, Gunnar Oregaard, Niels Kroer, and Søren J. Sørensen. Cultivation-independent examination of horizontal transfer and host range of an IncP-1 plasmid among gram-positive and gram-negative bacteria indigenous to the barley rhizosphere. *Applied and Environmental Microbiology*, 72(10):6687–6692, 10 2006.

[625] WHO list of bacterial species for which research is urgently needed - http://www.who.int/news-room/detail/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed Date accessed: 30/11/2018.

[626] Adam P Roberts and Peter Mullany. Tn916-like genetic elements: a diverse group of modular mobile elements conferring antibiotic resistance. *FEMS microbiology reviews*, 35(5):856–871, 9 2011.

[627] Marina Mingoia, Emily Tili, Esther Manso, Pietro E Varaldo, and Maria Pia Montanari. Heterogeneity of Tn5253-Like Composite Elements in Clinical Streptococcus pneumoniae Isolates. *Antimicrobial Agents and Chemotherapy*, 55(4):1453–1459, 11 2011.

[628] Ileana Cochetti, Emily Tili, Manuela Vecchi, Aldo Manzin, Marina Mingoia, Pietro E Varaldo, and Maria Pia Montanari. New Tn916-related elements causing erm(B)-mediated erythromycin resistance in tetracycline-susceptible pneumococci. *Journal of Antimicrobial Chemotherapy*, 60(1):127–131, 7 2007.

[629] Taj Azarian, Patrick K. Mitchell, Maria Georgieva, Claudette M. Thompson, Amel Ghouila, Andrew J. Pollard, Anne von Gottberg, Mignon du Plessis, Martin Antonio, Brenda A. Kwambana-Adams, Stuart C. Clarke, Dean Everett, Jennifer Cornick, Ewa Sadowy, Waleria Hryniewicz, Anna Skoczynska, Jennifer C. Moïsi, Lesley McGee, Bernard Beall, Benjamin J. Metcalf, Robert F. Breiman, PL Ho, Raymond Reid, Katherine L. O'Brien, Rebecca A. Gladstone, Stephen D. Bentley, and William P. Hanage. Global emergence and population dynamics of divergent serotype 3 CC180 pneumococci. *PLOS Pathogens*, 14(11):e1007438, 11 2018.

[630] Max R. Schroeder, Sarah Lohsen, Scott T. Chancey, and David S. Stephens. High-Level Macrolide Resistance Due to the Mega Element [mef(E)/mel] in Streptococcus pneumoniae. *Frontiers in Microbiology*, 10(APR), 2019.

[631] Marilyn C. Roberts, Joyce Sutcliffe, Patrice Courvalin, Lars Bogo Jensen, Julian Rood, and Helena Seppala. Nomenclature for Macrolide and Macrolide-Lincosamide-Streptogramin B Resistance Determinants. *Antimicrobial Agents and Chemotherapy*, 43(12):2823, 1999.

[632] Joanna Clancy, Joan Petitpas, Fadia Dib-Hajj, Wei Yuan, Melissa Cronan, Ajith V. Kamath, Jay Bergeron, and James A. Retsema. Molecular cloning and functional analysis of a novel macrolide-resistance determinant, mefA, from Streptococcus pyogenes. *Molecular microbiology*, 22(5):867–879, 1996.

[633] Maria Santagati, Francesco Iannelli, Carmela Cascone, Floriana Campanile, Marco R. Oggioni, Stefania Stefani, and Gianni Pozzi. The Novel Conjugative Transposon Tn1207.3 Carries the Macrolide Efflux Gene mef(A) in Streptococcus pyogenes. *https://home.liebertpub.com/mdr*, 9(3):243–247, 7 2004.

[634] Maria Santagati, Francesco Iannelli, Marco R Oggioni, Stefania Stefani, and Gianni Pozzi. Characterization of a Genetic Element Carrying the Macrolide Efflux Gene mef(A) in Streptococcus pneumoniae. *Antimicrobial Agents and Chemotherapy*, 44(9):2585–2587, 9 2000.

[635] Maria Del Grosso, Romina Camilli, Francesco Iannelli, Gianni Pozzi, and Annalisa Pantosti. The mef(E)-carrying genetic element (mega) of Streptococcus pneumoniae: insertion sites and association with other genetic elements. *Antimicrobial Agents & Chemotherapy*, 50(10):3361–3366, 10 2006.

[636] Ursula Munoz-Najar and Moses N. Vijayakumar. An Operon That Confers UV Resistance by Evoking the SOS Mutagenic Response in Streptococcal Conjugative Transposon Tn5252. *Journal of Bacteriology*, 181(9):2782, 1999.

[637] M Del Grosso, F Iannelli, C Messina, M Santagati, N Petrosillo, S Stefani, G Pozzi, and A Pantosti. Macrolide Efflux Genes mef(A) and mef(E) Are Carried by Different Genetic Elements in Streptococcus pneumoniae. *Journal of Clinical Microbiology*, 40(3):774–778, 3 2002.

[638] Aleksandra K. Wierzbowski, Dean Swedlo, Dave Boyd, Michael Mulvey, Kim A. Nichol, Daryl J. Hoban, and George G. Zhanel. Molecular Epidemiology and Prevalence of Macrolide Efflux Genes mef(A) and mef(E) in Streptococcus pneumoniae Obtained in Canada from 1997 to 2002. *Antimicrobial Agents and Chemotherapy*, 49(3):1257, 3 2005.

[639] Li Lin Liu, Shu Juan Ji, Zhi Ruan, Ying Fu, Yi Qi Fu, Yan Fei Wang, and Yun Song Yu. Dissemination of blaOXA-23 in Acinetobacter spp. in China: Main roles of conjugative plasmid pAZJ221 and transposon Tn2009. *Antimicrobial Agents and Chemotherapy*, 59(4):1998–2005, 4 2015.

[640] Y. Li, H. Tomita, Y. Lv, J. Liu, F. Xue, B. Zheng, and Y. Ike. Molecular characterization of erm(B)- and mef(E)-mediated erythromycin-resistant Streptococcus pneumoniae in China and complete DNA sequence of Tn2010. *Journal of Applied Microbiology*, 110(1):254–265, 1 2011.

[641] K. Gay and D. S. Stephens. Structure and Dissemination of a Chromosomal Insertion Element Encoding Macrolide Efflux in Streptococcus pneumoniae. *The Journal of Infectious Diseases*, 184(1):56–65, 7 2001.

[642] Max R Schroeder and David S Stephens. Macrolide Resistance in Streptococcus pneumoniae. *Frontiers in Cellular & Infection Microbiology*, 6:98, 2016.

[643] Supathep Tansirichaiya, Md. Ajijur Rahman, and Adam P Roberts. The Transposon Registry. *Mobile DNA*, 10(1):40, 2019.

[644] Ralf René Reinert, Adnan Al-Lahham, Maria Lemperle, Christoph Tenholte, Claudia Briefs, Stefan Haupts, Hans Hubert Gerards, and Rudolf Lütticken. Emergence of macrolide and penicillin resistance among invasive pneumococcal isolates in Germany. *Journal of Antimicrobial Chemotherapy*, 49(1):61–68, 11 2002.

[645] P McManus, M L Hammond, S D Whicker, J G Primrose, A Mant, and S R Fairall. Antibiotic use in the Australian community, 1990-1995. *The Medical journal of Australia*, 167(3):124–127, 8 1997.

[646] Sonja Hansen, Dorit Sohr, Brar Piening, Luis Pena Diaz, Alexander Gropmann, Rasmus Leistner, Elisabeth Meyer, Petra Gastmeier, and Michael Behnke. Antibiotic usage in German hospitals: results of the second national prevalence study. *Journal of Antimicrobial Chemotherapy*, 68(12):2934–2939, 12 2013.

[647] Mark van der Linden, Gerhard Falkenhorst, Stephanie Perniciaro, Christina Fitzner, and Matthias Imöhl. Effectiveness of Pneumococcal Conjugate Vaccines (PCV7 and PCV13) against Invasive Pneumococcal Disease among Children under Two Years of Age in Germany. *PLoS One*, 11(8):e0161257, 8 2016.

[648] Nicholas J. Croucher, Lisa Kagedan, Claudette M. Thompson, Julian Parkhill, Stephen D. Bentley, Jonathan A. Finkelstein, Marc Lipsitch, and William P. Hanage. Selective and Genetic Constraints on Pneumococcal Serotype Switching. *PLOS Genetics*, 11(3):e1005095, 3 2015.

[649] Mathieu Bergé, Miriam Moscoso, Marc Prudhomme, Bernard Martin, and Jean Pierre Claverys. Uptake of transforming DNA in Gram-positive bacteria: A

view from Streptococcus pneumoniae. *Molecular Microbiology*, 45(2):411–421, 2002.

[650] Catherine Turlan, Marc Prudhomme, Gwennaele Fichant, Bernard Martin, and Claude Gutierrez. SpxA1, a novel transcriptional regulator involved in X-state (competence) development in Streptococcus pneumoniae. *Molecular Microbiology*, 73(3):492–506, 8 2009.

[651] Stephanie W. Lo, Rebecca A. Gladstone, Andries J. van Tonder, John A. Lees, Mignon du Plessis, Rachel Benisty, Noga Givon-Lavi, Paulina A. Hawkins, Jennifer E. Cornick, Brenda Kwambana-Adams, Pierra Y. Law, Pak Leung Ho, Martin Antonio, Dean B. Everett, Ron Dagan, Anne von Gottberg, Keith P. Klugman, Lesley McGee, Robert F. Breiman, Stephen D. Bentley, Abdullah W. Brooks, Alejandra Corso, Alexander Davydov, Alison Maguire, Andrew Pollard, Anmol Kiran, Anna Skoczynska, Benild Moiane, Bernard Beall, Betuel Sigauque, David Aanensen, Deborah Lehmann, Diego Faccone, Ebenezer Foster-Nyarko, Ebrima Bojang, Ekaterina Egorova, Elena Voropaeva, Eric Sampane-Donkor, Ewa Sadowy, Godfrey Bigogo, Helio Mucavele, Houria Belabbès, Idrissa Diawara, Jennifer Moïsi, Jennifer Verani, Jeremy Keenan, Jyothish N. Nair Thulasee Bhai, Kedibone M. Ndlangisa, Khalid Zerouali, K. L. Ravikumar, Leonid Titov, Linda De Gouveia, Maaike Alaerts, Margaret Ip, Maria Cristina de Cunto Brandileone, Md Hasanuzzaman, Metka Paragi, Michele Nurse-Lucas, Mushal Ali, Naima Elmdaghri, Nicholas Croucher, Nicole Wolter, Nurit Porat, Özgen Köseoglu Eser, Patrick E. Akpaka, Paul Turner, Paula Gagetti, Peggy Estelle Tientcheu, Philip E. Carter, Rafal Mostowy, Rama Kandasamy, Rebecca Ford, Rebecca Henderson, Roly Malaker, Sadia Shakoor, Samanta Cristine Grassi Almeida, Samir K. Saha, Sanjay Doiphode, Shabir A. Madhi, Shamala Devi Sekaran, Somporn Srifuengfung, Stephen Obaro, Stuart C. Clarke, Susan A. Nzenze, Tamara Kastrin, Theresa J. Ochoa, Veeraraghavan Balaji, Waleria Hryniewicz, and Yulia Urban. Pneumococcal lineages associated with serotype replacement and antibiotic resistance in childhood invasive pneumococcal disease in the post-PCV13 era: an international whole-genome sequencing study. *The Lancet Infectious Diseases*, 19(7):759–769, 7 2019.

[652] K Poulsen, J Reinholdt, and M Kilian. Characterization of the Streptococcus pneumoniae immunoglobulin A1 protease gene (iga) and its translation product. *Infection and immunity*, 64(10):3957–3966, 10 1996.

[653] Scott V. Nguyen and William M. McShan. Chromosomal islands of Streptococcus pyogenes and related streptococci: molecular switches for survival and virulence. *Frontiers in Cellular and Infection Microbiology*, 4:109, 8 2014.

[654] Julie Scott, Prestina Thompson-Mayberry, Stephanie Lahmamsi, Catherine J King, and W Michael McShan. Phage-associated mutator phenotype in group A streptococcus. *Journal of bacteriology*, 190(19):6290–301, 10 2008.

[655] Nathaniel D Chu, Sean A Clarke, Sonia Timberlake, Martin F Polz, Alan D Grossman, and Eric J Alm. A Mobile Element in mutS Drives Hypermutation in a Marine Vibrio. *mBio*, 8(1):02045–16, 3 2017.

[656] Scott T Chancey, Sonia Agrawal, Max R Schroeder, Monica M Farley, Herve Tettelin, and David S Stephens. Composite mobile genetic elements disseminating macrolide resistance in Streptococcus pneumoniae. *Frontiers in Microbiology*, 6:26, 2015.

[657] Anne-Sophie Godeux, Elin Svedholm, Samuel Barreto, Anaïs Potron, Samuel Venner, Xavier Charpentier, and Maria-Halima Laaberki. Interbacterial Transfer of Carbapenem Resistance and Large Antibiotic Resistance Islands by Natural Transformation in Pathogenic Acinetobacter. *mBio*, 13(1), 2 2022.

[658] Claire Bertelli, Keith E Tilley, and Fiona S L Brinkman. Microbial genomic island discovery, visualization and analysis. *Briefings in Bioinformatics*, 2018.

[659] Georgios S. Vernikos and Julian Parkhill. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. *Bioinformatics*, 22(18):2196–2203, 9 2006.

[660] Dongsheng Che Han Wang, Han Wang, John Fazekas, and Bernard Chen. An Accurate Genomic Island Prediction Method for Sequenced Bacterial and Archaeal Genomes. *Journal of Proteomics & Bioinformatics*, 07(08), 7 2014.

[661] Dongsheng Che, Cory Hockenbury, Robert Marmelstein, and Khaled Rasheed. Classification of genomic islands using decision trees and their ensemble algorithms. *BMC genomics*, 11 Suppl 2(Suppl 2):S1, 11 2010.

[662] Maud Fléchard, Céline Lucchetti-Miganeh, Bernard Hallet, Pascal Hols, and Philippe Gilot. Intensive targeting of regulatory competence genes by transposable elements in streptococci. *Molecular genetics and genomics : MGG*, 294(3):531–548, 6 2019.

[663] Matthew T.G. Holden, Zoe Heather, Romain Paillot, Karen F. Steward, Katy Webb, Fern Ainslie, Thibaud Jourdan, Nathalie C. Bason, Nancy E. Holroyd, Karen Mungall, Michael A. Quail, Mandy Sanders, Mark Simmonds, David Willey, Karen Brooks, David M. Aanensen, Brian G. Spratt, Keith A. Jolley, Martin C.J. Maiden, Michael Kehoe, Neil Chanter, Stephen D. Bentley, Carl Robinson, Duncan J. Maskell, Julian Parkhill, and Andrew S. Waller. Genomic Evidence for the Evolution of Streptococcus equi: Host Restriction, Increased Virulence, and Genetic Exchange with Human Pathogens. *PLOS Pathogens*, 5(3):e1000346, 3 2009.

[664] Michael R. Brooks, Lyan Padilla-Vélez, Tarannum A. Khan, Azaan A. Qureshi, Jason B. Pieper, Carol W. Maddox, and Md Tauqeer Alam. Prophage-Mediated Disruption of Genetic Competence in Staphylococcus pseudintermedius. *mSystems*, 5(1), 2 2020.

[665] Calum Johnston, Bernard Martin, Gwennaele Fichant, Patrice Polard, and Jean-Pierre Claverys. Bacterial transformation: distribution, shared mechanisms and divergent control. *Nature Reviews Microbiology*, 12(3):181–196, 3 2014.

[666] Carmen Buchrieser, Christophe Rusniok, Frank Kunst, Pascale Cossart, Philippe Glaser, L. Frangeul, A. Amend, F. Baquero, P. Berche, H. Bloecker, P. Brandt, T. Chakaborty, A. Charbit, F. Chétouani, E. Couvé, A. De Daruvar, P. Dehoux, E. Domann, G. Domínguez-Bernal, E. Duchaud, L. Durand, O. Dusurget, K. D. Entian, H. Fsihi, P. Garcia-Del Portillo, P. Garrido, L. Gautier, W. Goebel, N. Gómez-López, T. Hain, J. Hauf, D. Jackson, L. M. Jones, U. Kärst, J. Kreft, M. Kuhn, G. Kurapkat, E. Madueño, A. Maitournam, J. Mata Vicente, E. Ng, G. Nordsiek, B. De Pablos, J. C. Pérez-Diaz, B. Remmel, M. Rose, T. Schlueter, N. Simoes, J. A. Vázquez-Boland, H. Voss, and J. Wehland. Comparison of the genome sequences of Listeria monocytogenes and Listeria innocua: clues for evolution and pathogenicity. *FEMS Immunology & Medical Microbiology*, 35(3):207–213, 4 2003.

[667] Anne Sophie Godeux, Elin Svedholm, Agnese Lupo, Marisa Haenni, Samuel Venner, Maria Halima Laaberki, and Xavier Charpentier. Scarless Removal of Large Resistance Island AbaR Results in Antibiotic Susceptibility and Increased Natural Transformability in Acinetobacter baumannii. *Antimicrobial Agents and Chemotherapy*, 64(10), 10 2020.

[668] Isabelle Durieux, Christophe Ginevra, Laetitia Attaiech, Kévin Picq, Pierre Alexandre Juan, Sophie Jarraud, and Xavier Charpentier. Diverse conjugative elements silence natural transformation in Legionella species. *Proceedings of the National Academy of Sciences of the United States of America*, 116(37):18613–18618, 9 2019.

[669] Carole Ayoub Moubareck and Dalal Hammoudi Halat. Insights into Acinetobacter baumannii: A Review of Microbiological, Virulence, and Resistance Traits in a Threatening Nosocomial Pathogen. *Antibiotics*, 9(3), 2020.

[670] Mohammad Hamidian and Steven J. Nigro. Emergence, molecular mechanisms and global spread of carbapenem-resistant Acinetobacter baumannii. *Microbial Genomics*, 5(10), 2019.

[671] Luísa C.S. Antunes, Paolo Visca, and Kevin J. Towner. Acinetobacter baumannii: evolution of a global pathogen. *Pathogens and Disease*, 71(3):292–301, 8 2014.

[672] Alemu Gedefie, Wondmagegn Demsis, Melaku Ashagrie, Yeshimebet Kassa, Melkam Tesfaye, Mihret Tilahun, Habtye Bisetegn, and Zenawork Sahle. Acinetobacter baumannii Biofilm Formation and Its Role in Disease Pathogenesis: A Review. *Infection and Drug Resistance*, 14:3711, 2021.

[673] Pierre Edouard Fournier and Hervé Richet. The epidemiology and control of Acinetobacter baumannii in health care facilities. *Clinical Infectious Diseases*, 42(5):692–699, 3 2006.

[674] Ghayda Al-Hashem, Vincent O. Rotimi, and M. John Albert. Genetic relatedness of serial rectal isolates of Acinetobacter baumannii in an adult intensive care unit of a tertiary hospital in Kuwait. *PloS one*, 15(4), 2020.

[675] Patricia Cornejo-Juárez, Miguel Angel Cevallos, Semiramis Castro-Jaimes, Santiago Castillo-Ramìrez, Consuelo Velázquez-Acosta, David Martìnez-Oliva, Angeles Pérez-Oseguera, Frida Rivera-Buendìa, and Patricia Volkow-Fernández. High mortality in an outbreak of multidrug resistant Acinetobacter baumannii infection introduced to an oncological hospital by a patient transferred from a general hospital. *PloS one*, 15(7), 7 2020.

[676] G. Molter, H. Seifert, F. Mandraka, G. Kasper, B. Weidmann, B. Hornei, M. Öhler, P. Schwimmbeck, P. Kröschel, P. G. Higgins, and S. Reuter. Outbreak of carbapenem-resistant Acinetobacter baumannii in the intensive care unit: a multi-level strategic management approach. *The Journal of hospital infection*, 92(2):194–198, 2 2016.

[677] Yongxin Zhao, Kewang Hu, Jisheng Zhang, Yuhang Guo, Xuecai Fan, Yong Wang, Sedzro Divine Mensal, and Xiaoli Zhang. Outbreak of carbapenem-resistant Acinetobacter baumannii carrying the carbapenemase OXA-23 in ICU of the eastern Heilongjiang Province, China. *BMC infectious diseases*, 19(1), 5 2019.

[678] Grace A. Blackwell, Mohammad Hamidian, and Ruth M. Hall. IncM Plasmid R1215 Is the Source of Chromosomally Located Regions Containing Multiple Antibiotic

Resistance Genes in the Globally Disseminated Acinetobacter baumannii GC1 and GC2 Clones. *mSphere*, 1(3):117–133, 6 2016.

[679] Dae Hun Kim, Ji Young Choi, Hae Won Kim, So Hyun Kim, Doo Ryeon Chung, Kyong Ran Peck, Visanu Thamlikitkul, Thomas Man Kit So, Rohani M.D. Yasin, Po Ren Hsueh, Celia C. Carlos, Li Yang Hsu, Latre Buntaran, M. K. Lalitha, Jae Hoon Song, and Kwan Soo Ko. Spread of carbapenem-resistant Acinetobacter baumannii global clone 2 in Asia and AbaR-type resistance islands. *Antimicrobial Agents and Chemotherapy*, 57(11):5239–5246, 11 2013.

[680] Louis B. Rice. Federal Funding for the Study of Antimicrobial Resistance in Nosocomial Pathogens: No ESKAPE. *The Journal of Infectious Diseases*, 197(8):1079–1081, 4 2008.

[681] Mohammad Hamidian and Ruth M. Hall. The AbaR antibiotic resistance islands found in Acinetobacter baumannii global clone 1 – Structure, origin and evolution. *Drug Resistance Updates*, 41:26–39, 11 2018.

[682] Jane F. Turton, S. N. Gabriel, C. Valderrey, M. E. Kaufmann, and T. L. Pitt. Use of sequence-based typing and multiplex PCR to identify clonal lineages of outbreak strains of Acinetobacter baumannii. *Clinical Microbiology and Infection*, 13(8):807–815, 8 2007.

[683] Dexi Bi, Ruting Xie, Jiayi Zheng, Huiqiong Yang, Xingchen Zhu, Hong Yu Ou, and Qing Wei. Large-Scale Identification of AbaR-Type Genomic Islands in Acinetobacter baumannii Reveals Diverse Insertion Sites and Clonal Lineage-Specific Antimicrobial Resistance Gene Profiles. *Antimicrobial agents and chemotherapy*, 63(4), 4 2019.

[684] Mohammad Hamidian, Kathryn E. Holt, Derek Pickard, Gordon Dougan, and Ruth M. Hall. A GC1 Acinetobacter baumannii isolate carrying AbaR3 and the aminoglycoside resistance transposon TnaphA6 in a conjugative plasmid. *Journal of Antimicrobial Chemotherapy*, 69(4):955–958, 4 2014.

[685] Mohammad Hamidian, Jane Hawkey, Ryan Wick, Kathryn E. Holt, and Ruth M. Hall. Evolution of a clade of Acinetobacter baumannii global clone 1, lineage 1 via acquisition of carbapenem- and aminoglycoside-resistance genes and dispersion of ISAba1. *Microbial Genomics*, 5(1), 1 2019.

[686] Mohammad Hamidian, Johanna J. Kenyon, Kathryn E. Holt, Derek Pickard, and Ruth M. Hall. A conjugative plasmid carrying the carbapenem resistance gene blaOXA-23 in AbaR4 in an extensively resistant GC1 Acinetobacter baumannii isolate. *Journal of Antimicrobial Chemotherapy*, 69(10):2625, 10 2014.

[687] Dexi Bi, Jiayi Zheng, Ruting Xie, Yin Zhu, Rong Wei, Hong-Yu Ou, Qing Wei, and Huanlong Qin. Comparative Analysis of AbaR-Type Genomic Islands Reveals Distinct Patterns of Genetic Features in Elements with Different Backbones. *mSphere*, 5(3), 6 2020.

[688] Rémy A. Bonnin, Laurent Poirel, and Patrice Nordmann. AbaR-type transposon structures in Acinetobacter baumannii. *Journal of Antimicrobial Chemotherapy*, 67(1):234–236, 1 2012.

[689] Thomas M Nero, Triana N Dalia, Joseph Che-Yen Wang, David T Kysela, Matthew L Bochman, and Ankur B Dalia. ComM is a hexameric helicase that

promotes branch migration during natural transformation in diverse Gram-negative species. *bioRxiv*, page 147660, 1 2018.

[690] M. L. Gwinn, R. Ramanathan, H. O. Smith, and J. F. Tomb. A new transformation-deficient mutant of Haemophilus influenzae Rd with normal DNA uptake. *Journal of Bacteriology*, 180(3):746–748, 1998.

[691] Hayley J. Newton, Desmond K.Y. Ang, Ian R. Van Driel, and Elizabeth L. Hartland. Molecular Pathogenesis of Infections Caused by Legionella pneumophila. *Clinical Microbiology Reviews*, 23(2):274, 4 2010.

[692] Sophia David, Christophe Rusniok, Massimo Mentasti, Laura Gomez-Valero, Simon R. Harris, Pierre Lechat, John Lees, Christophe Ginevra, Philippe Glaser, Laurence Ma, Christiane Bouchier, Anthony Underwood, Sophie Jarraud, Timothy G. Harrison, Julian Parkhill, and Carmen Buchrieser. Multiple major disease-associated clones of Legionella pneumophila have emerged recently and independently. *Genome Research*, 26(11):1555–1564, 11 2016.

[693] Olivier Duron, Patricia Doublet, Fabrice Vavre, and Didier Bouchon. The Importance of Revisiting Legionellales Diversity. *Trends in Parasitology*, 34(12):1027–1037, 12 2018.

[694] David K. Boamah, Guangqi Zhou, Alexander W. Ensminger, and Tamara J. O'Connor. From many hosts, one accidental pathogen: The diverse protozoan hosts of Legionella. *Frontiers in Cellular and Infection Microbiology*, 7(NOV):477, 11 2017.

[695] A. Khodr, E. Kay, L. Gomez-Valero, C. Ginevra, P. Doublet, C. Buchrieser, and S. Jarraud. Molecular epidemiology, phylogeny and evolution of Legionella. *Infection, Genetics and Evolution*, 43:108–122, 9 2016.

[696] J. Beauté. Legionnaires' disease in Europe, 2011 to 2015. *Eurosurveillance*, 22(27):30566, 7 2017.

[697] Nick Phin, Frances Parry-Ford, Timothy Harrison, Helen R. Stagg, Natalie Zhang, Kartik Kumar, Olivier Lortholary, Alimuddin Zumla, and Ibrahim Abubakar. Epidemiology and clinical management of Legionnaires' disease. *The Lancet Infectious Diseases*, 14(10):1011–1021, 10 2014.

[698] Marine Vandewalle-Capo, Clémence Massip, Ghislaine Descours, Joséphine Charavit, Joelle Chastang, Pierre Alain Billy, Sandrine Boisset, Gerard Lina, Christophe Gilbert, Max Maurin, Sophie Jarraud, and Christophe Ginevra. Minimum inhibitory concentration (MIC) distribution among wild-type strains of Legionella pneumophila identifies a subpopulation with reduced susceptibility to macrolides owing to efflux pump genes. *International Journal of Antimicrobial Agents*, 50(5):684–689, 11 2017.

[699] Sophia David, Leonor Sánchez-Busó, Simon R. Harris, Pekka Marttinen, Christophe Rusniok, Carmen Buchrieser, Timothy G. Harrison, and Julian Parkhill. Dynamics and impact of homologous recombination on the evolution of Legionella pneumophila. *PLOS Genetics*, 13(6):e1006855, 2017.

[700] Laura Gomez-Valero, Christophe Rusniok, Sophie Jarraud, Benoit Vacherie, Zoé Rouy, Valerie Barbe, Claudine Medigue, Jerome Etienne, and Carmen Buchrieser. Extensive recombination events and horizontal gene transfer shaped the Legionella pneumophila genomes. *BMC Genomics*, 12:536, 11 2011.

[701] Alexander W. Ensminger and Ralph R. Isberg. Legionella pneumophila Dot/Icm translocated substrates: a sum of parts. *Current Opinion in Microbiology*, 12(1):67–73, 2 2009.

[702] Laetitia Attaiech, Aïda Boughammoura, Céline Brochier-Armanet, Omran Allatif, Flora Peillard-Fiorente, Ross A. Edwards, Ayat R. Omar, Andrew M. MacMillan, Mark Glover, and Xavier Charpentier. Silencing of natural transformation by an RNA chaperone and a multitarget small RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 113(31):8813–8818, 8 2016.

[703] R. J. Redfield. sxy-1, a Haemophilus influenzae mutation causing greatly enhanced spontaneous competence. *Journal of bacteriology*, 173(18):5612–5618, 1991.

[704] Yan Zhu, Jing Lu, Jinxin Zhao, Xinru Zhang, Heidi H. Yu, Tony Velkov, and Jian Li. Complete genome sequence and genome-scale metabolic modelling of Acinetobacter baumannii type strain ATCC 19606. *International Journal of Medical Microbiology*, 310(3):151412, 4 2020.

[705] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 3 2017.

[706] Keith A. Jolley, James E. Bray, and Martin C.J. Maiden. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome open research*, 3, 2018.

[707] William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[708] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification 1985 2:1*, 2(1):193–218, 12 1985.

[709] Liang Chen, Barun Mathema, Johann D.D. Pitout, Frank R. DeLeo, and Barry N. Kreiswirth. Epidemic Klebsiella pneumoniae ST258 Is a Hybrid Strain. *mBio*, 5(3), 6 2014.

[710] Kelly L. Wyres, Sarah M. Cahill, Kathryn E. Holt, Ruth M. Hall, and Johanna J. Kenyon. Identification of acinetobacter baumannii loci for capsular polysaccharide (KL) and lipooligosaccharide outer core (OCL) synthesis in genome assemblies using curated reference databases compatible with kaptive. *Microbial Genomics*, 6(3):e000339, 3 2020.

[711] Dann Turner, Hans Wolfgang Ackermann, Andrew M. Kropinski, Rob Lavigne, J. Mark Sutton, and Darren M. Reynolds. Comparative Analysis of 37 Acinetobacter Bacteriophages. *Viruses*, 10(1), 1 2018.

[712] Jongsoo Jeon, Jae-won Kim, Dongeun Yong, Kyungwon Lee, and Yunsop Chong. Complete genome sequence of the podoviral bacteriophage YMC/09/02/B1251 ABA BP, which causes the lysis of an OXA-23-producing carbapenem-resistant Acinetobacter baumannii isolate from a septic patient. *Journal of virology*, 86(22):12437–12438, 11 2012.

[713] Jongsoo Jeon, Roshan D'Souza, Naina Pinto, Choong Min Ryu, Jong hwan Park, Dongeun Yong, and Kyungwon Lee. Complete genome sequence of the siphoviral bacteriophage B□-R3177, which lyses an OXA-66-producing carbapenem-resistant Acinetobacter baumannii isolate. *Archives of virology*, 160(12):3157–3160, 10 2015.

[714] Meng Liu, Xiaobin Li, Yingzhou Xie, Dexi Bi, Jingyong Sun, Jun Li, Cui Tai, Zixin Deng, and Hong Yu Ou. ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. *Nucleic Acids Research*, 47(D1):D660–D665, 1 2019.

[715] Philip Jones, David Binns, Hsin Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F. Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew Yit Yong, Rodrigo Lopez, and Sarah Hunter. InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, 5 2014.

[716] Shruti Chatterjee, Aditya J. Basak, Asha V. Nair, Kheerthana Duraivelan, and Dibyendu Samanta. Immunoglobulin-fold containing bacterial adhesins: molecular and structural perspectives in host tissue colonization and infection. *FEMS microbiology letters*, 368(2), 1 2021.

[717] Kari A. Brossard and Anthony A. Campagnari. The Acinetobacter baumannii biofilm-associated protein plays a role in adherence to human epithelial cells. *Infection and Immunity*, 80(1):228–233, 1 2012.

[718] Bryan A. Wee, Joana Alves, Diane S.J. Lindsay, Ann Brit Klatt, Fiona A. Sargison, Ross L. Cameron, Amy Pickering, Jamie Gorzynski, Jukka Corander, Pekka Marttinen, Bastian Opitz, Andrew J. Smith, and J. Ross Fitzgerald. Population analysis of Legionella pneumophila reveals a basis for resistance to complement-mediated killing. *Nature Communications 2021 12:1*, 12(1):1–13, 12 2021.

[719] Günther Koraimann and Maria A. Wagner. Social behavior and decision making in bacterial conjugation. *Frontiers in Cellular and Infection Microbiology*, 4(APR):54, 2014.

[720] Rodrigo Flores-Ríos, Ana Moya-Beltrán, Claudia Pareja-Barrueto, Mauricio Arenas-Salinas, I. Sebastián Valenzuela, Omar Orellana, and Raquel Quatrini. The type IV secretion system of ICEAfe1: Formation of a conjugative pilus in Acidithiobacillus ferrooxidans. *Frontiers in Microbiology*, 10(FEB), 2019.

[721] Dinesh M. Fernando and Ayush Kumar. Resistance-Nodulation-Division Multidrug Efflux Pumps in Gram-Negative Bacteria: Role in Virulence. *Antibiotics*, 2(1):163, 3 2013.

[722] Ines Bleriot, Rocío Trastoy, Lucia Blasco, Felipe Fernández-Cuenca, Antón Ambroa, Laura Fernández-García, Olga Pacios, Elena Perez-Nadales, Julian Torre-Cisneros, Jesús Oteo-Iglesias, Ferran Navarro, Elisenda Miró, Alvaro Pascual, German Bou, Luis Martínez-Martínez, and Maria Tomas. Genomic analysis of 40 prophages located in the genomes of 16 carbapenemase-producing clinical strains of Klebsiella pneumoniae. *Microbial Genomics*, 6(5):1–18, 2020.

[723] Jean Cury, Pedro H Oliveira, Fernando de la Cruz, and Eduardo P C Rocha. Host Range and Genetic Plasticity Explain the Coexistence of Integrative and Extrachromosomal Mobile Genetic Elements. *Molecular Biology and Evolution*, 35(9):2230–2239, 9 2018.

[724] Claire E. Turner, Matthew T.G. Holden, Beth Blane, Carolyne Horner, Sharon J. Peacock, and Shiranee Sriskandan. The emergence of successful Streptococcus pyogenes lineages through convergent pathways of capsule loss and recombination directing high toxin expression. *mBio*, 10(6), 11 2019.

[725] Stephen C. Watts and Kathryn E. Holta. HICAP: In silico serotyping of the haemophilus influenzae capsule locus. *Journal of Clinical Microbiology*, 57(6), 6 2019.

[726] Kelly L. Wyres, Claire Gorrie, David J. Edwards, Heiman F.L. Wertheim, Li Yang Hsu, Nguyen Van Kinh, Ruth Zadoks, Stephen Baker, and Kathryn E. Holt. Extensive Capsule Locus Variation and Large-Scale Genomic Recombination within the Klebsiella pneumoniae Clonal Group 258. *Genome Biology and Evolution*, 7(5):1267, 5 2015.

[727] Kelly L. Wyres, Ryan R. Wick, Louise M. Judd, Roni Froumine, Alex Tokolyi, Claire L. Gorrie, Margaret M.C. Lam, Sebastián Duchêne, Adam Jenney, and Kathryn E. Holt. Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of Klebsiella pneumoniae. *PLOS Genetics*, 15(4):e1008114, 2019.

[728] Koji Yahara, Xavier Didelot, Keith A. Jolley, Ichizo Kobayashi, Martin C.J. Maiden, Samuel K. Sheppard, and Daniel Falush. The Landscape of Realized Homologous Recombination in Pathogenic Bacteria. *Molecular Biology and Evolution*, 33(2):456–471, 2 2016.

[729] Regine Hakenbeck, Andrea König, Izabella Kern, Mark van der Linden, Wolfgang Keck, Danielle Billot-Klein, Raymond Legrand, Bernard Schoot, and Laurent Gutmann. Acquisition of Five High-Mr Penicillin-Binding Protein Variants during Transfer of High-Level β-Lactam Resistance from Streptococcus mitis to Streptococcus pneumoniae. *Journal of Bacteriology*, 180(7):1831 LP – 1840, 4 1998.

[730] Regine Hakenbeck, Nadège Balmelle, Beate Weber, Christophe Gardès, Wolfgang Keck, and Antoine de Saizieu. Mosaic Genes and Mosaic Chromosomes: Intra- and Interspecies Genomic Variation of Streptococcus pneumoniae. *Infection and Immunity*, 69(4):2477 LP – 2486, 4 2001.

[731] Jacek Majewski, Piotr Zawadzki, Paul Pickerill, Frederick M Cohan, and Christopher G Dowson. Barriers to Genetic Exchange between Bacterial Species: Streptococcus pneumoniae Transformation. *Journal of Bacteriology*, 182(4):1016–1023, 2 2000.

[732] Thomas A. Russo, Janet M. Beanan, Ruth Olson, Ulrike MacDonald, Andrew D. Cox, Frank St. Michael, Evgeny V. Vinogradov, Brad Spellberg, Nicole R. Luke-Marshall, and Anthony A. Campagnari. The K1 Capsular Polysaccharide from Acinetobacter baumannii Is a Potential Therapeutic Target via Passive Immunization. *Infection and Immunity*, 81(3):915, 3 2013.

[733] Marta Palusinska-Szysz, Rafal Luchowski, Wieslaw I. Gruszecki, Adam Choma, Agnieszka Szuster-Ciesielska, Christian Lück, Markus Petzold, Anna Sroka-Bartnicka, and Bozena Kowalczyk. The Role of Legionella pneumophila Serogroup 1 Lipopolysaccharide in Host-Pathogen Interaction. *Frontiers in Microbiology*, 10:2890, 12 2019.

[734] Roger Milkman, Erich Jaeger, and Ryan D McBride. Molecular Evolution of the Escherichia coli Chromosome. VI. Two Regions of High Effective Recombination. *Genetics*, 163(2):475 LP – 483, 2 2003.

[735] B. Jesse Shapiro, Lawrence A. David, Jonathan Friedman, and Eric J. Alm. Looking for Darwin's footprints in the microbial world, 5 2009.

[736] Christophe Fraser, William P. Hanage, and Brian G. Spratt. Recombination and the nature of bacterial speciation, 1 2007.

[737] William P. Hanage, Christophe Fraser, and Brian G. Spratt. Fuzzy species among recombinogenic bacteria. *BMC Biology*, 3(1):6, 3 2005.

[738] Samuel K. Sheppard, Noel D. McCarthy, Daniel Falush, and Martin C.J. Maiden. Convergence of Campylobacter species: Implications for bacterial evolution. *Science*, 320(5873):237–239, 4 2008.

[739] Anton Y. Peleg, Harald Seifert, and David L. Paterson. Acinetobacter baumannii: Emergence of a successful pathogen. *Clinical Microbiology Reviews*, 21(3):538–582, 7 2008.

[740] Mahrokh Saadati, Leila Rahbarnia, Safar Farajnia, Behrooz Naghili, and Reza Mohammadzadeh. The prevalence of biofilm encoding genes in multidrug-resistant Acinetobacter baumannii isolates. *Gene Reports*, 23:101094, 6 2021.

[741] Jennifer A. Gaddy and Luis A. Actis. Regulation of Acinetobacter baumannii biofilm formation. *Future Microbiology*, 4(3):273–278, 3 2009.

[742] H. M. Sharon Goh, Scott A. Beatson, Makrina Totsika, Danilo G. Moriel, Minh Duy Phan, Jan Szubert, Naomi Runnegar, Hanna E. Sidjabat, David L. Paterson, Graeme R. Nimmo, Jeffrey Lipman, and Mark A. Schembri. Molecular analysis of the Acinetobacter baumannii biofilm-associated protein. *Applied and Environmental Microbiology*, 79(21):6535–6543, 2013.

[743] Arianna Pompilio, Daniela Scribano, Meysam Sarshar, Giovanni Di Bonaventura, Anna Teresa Palamara, and Cecilia Ambrosi. Gram-Negative Bacteria Holding Together in a Biofilm: The Acinetobacter baumannii Way. *Microorganisms*, 9(7), 7 2021.

[744] Christian M. Harding, Seth W. Hennon, and Mario F. Feldman. Uncovering the mechanisms of Acinetobacter baumannii virulence. *Nature Reviews Microbiology 2017 16:2*, 16(2):91–102, 12 2017.

[745] Leonor Sánchez-Busó, Iñaki Comas, Guillermo Jorques, and Fernando González-Candelas. Recombination drives genome evolution in outbreak-related Legionella pneumophila isolates. *Nature Genetics 2014 46:11*, 46(11):1205–1211, 10 2014.

[746] Johan Bengtsson-Palme, Erik Kristiansson, and D. G.Joakim Larsson. Environmental factors influencing the development and spread of antibiotic resistance. *FEMS Microbiology Reviews*, 42(1):68–80, 1 2018.

[747] Marie Touchon, Louis Marie Bobay, and Eduardo P.C. Rocha. The chromosomal accommodation and domestication of mobile genetic elements. *Current Opinion in Microbiology*, 22:22–29, 12 2014.

[748] Francesca Micoli, Fabio Bagnoli, Rino Rappuoli, and Davide Serruto. The role of vaccines in combatting antimicrobial resistance. *Nature Reviews Microbiology 2021 19:5*, 19(5):287–302, 2 2021.

[749] Moe H Kyaw, Ruth Lynfield, William Schaffner, Allen S Craig, James Hadler, Arthur Reingold, Ann R Thomas, Lee H Harrison, Nancy M Bennett, Monica M Farley, Richard R Facklam, James H Jorgensen, John Besser, Elizabeth R Zell, Anne Schuchat, and Cynthia G Whitney. Effect of introduction of the pneumococcal con-

jugate vaccine on drug-resistant Streptococcus pneumoniae. *The New England Journal of Medicine*, 354(14):1455–1463, 4 2006.

[750] Marc Lipsitch and George R. Siber. How Can Vaccines Contribute to Solving the Antimicrobial Resistance Problem? *mBio*, 7(3), 2016.

[751] Matthew R. Moore, Ruth Link-Gelles, William Schaffner, Ruth Lynfield, Catherine Lexau, Nancy M. Bennett, Susan Petit, Shelley M. Zansky, Lee H. Harrison, Arthur Reingold, Lisa Miller, Karen Scherzinger, Ann Thomas, Monica M. Farley, Elizabeth R. Zell, Thomas H. Taylor, Tracy Pondo, Loren Rodgers, Lesley McGee, Bernard Beall, James H. Jorgensen, and Cynthia G. Whitney. Impact of 13-Valent Pneumococcal Conjugate Vaccine Used in Children on Invasive Pneumococcal Disease in Children and Adults in the United States: Analysis of Multisite, Population-based Surveillance. *The Lancet. Infectious diseases*, 15(3):301, 3 2015.

[752] Alex Orlek, Nicole Stoesser, Muna F. Anjum, Michel Doumith, Matthew J. Ellington, Tim Peto, Derrick Crook, Neil Woodford, A. Sarah Walker, Hang Phan, and Anna E. Sheppard. Plasmid Classification in an Era of Whole-Genome Sequencing: Application in Studies of Antibiotic Resistance Epidemiology. *Frontiers in Microbiology*, 8(FEB):182, 2 2017.

[753] Jean Cury, Sophie S. Abby, Olivia Doppelt-Azeroual, Bertrand Néron, and Eduardo P.C. Rocha. Identifying Conjugative Plasmids and Integrative Conjugative Elements with CONJscan. *Methods in Molecular Biology*, 2075:265–283, 2020.

[754] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 19(5):455, 5 2012.

[755] Gabrielle L. Harrow, John A. Lees, William P. Hanage, Marc Lipsitch, Jukka Corander, Caroline Colijn, and Nicholas J. Croucher. Negative frequency-dependent selection and asymmetrical transformation stabilise multi-strain bacterial population structures. *ISME Journal*, 15(5):1523–1538, 5 2021.