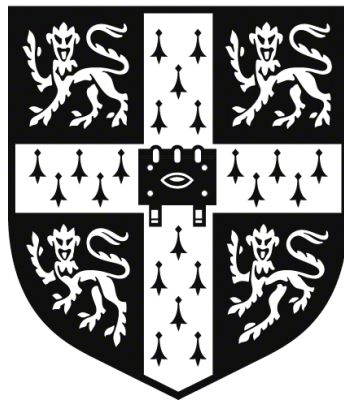


PATHOGEN EVOLUTION WITHIN AND BETWEEN HOSTS WITH
APPLICATIONS TO THE PNEUMOCOCCUS AND SARS-COV-2

GERRY TONKIN-HILL



Wellcome Sanger Institute
University of Cambridge

Wolfson College – December 2020

This dissertation is submitted for the degree of *Doctor of Philosophy*

Gerry Tonkin-Hill: *Pathogen evolution within and between hosts with applications to the pneumococcus and SARS-CoV-2* , This dissertation is submitted for the degree of *Doctor of Philosophy*, © December 2020

DECLARATION

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed 60 000 words in length, as required by the School of Biological Sciences.

University of Cambridge, December 2020

Gerry Tonkin-Hill

ABSTRACT

The tracking of a pathogen's evolution over different timescales is essential: for the design of vaccines, to control antibiotic resistance, for the surveillance of disease, to identify transmission chains and to understand the mechanisms behind the pathogen's evolution. Although the first bacterial genome was sequenced in 1995, techniques and costs have only recently enabled the large scale study of the evolution of pathogens in human populations over short timescales.

This thesis begins by developing methods for the analysis of pathogen genomes, including the identification of population structure, characterisation of gene content and tracking of transmission. These techniques are then applied to the analysis of two large pathogen sequencing studies investigating the evolution of *Streptococcus pneumoniae* and SARS-CoV-2 within the host and during transmission.

Initially, a fast solution to the problem of clustering pathogen genomes is presented. An approximation to a Dirichlet process mixture model is shown to produce accurate genome clusterings 10-100 times faster than existing methods. Next, the problem of identifying the pangenome (the set of all genes that have been found in a species) is considered and a method is presented that is able to reduce by over 10 fold the number of errors introduced by misannotation. Methods for the inference of transmission chains using genomic data are then described. Existing methods are adapted to improve their efficiency and make them applicable to the tracking of endemic disease. In addition, a novel simulation based transmission inference method is introduced.

Following the description of these techniques, a large dataset of nearly 4000 pneumococcal positive nasopharyngeal swabs is considered, comprising samples taken over a period of two years from mothers and their children in the Maela refugee camp in Thailand. After culturing *S. pneumoniae* using selective agar, whole plate sweeps were taken for sequencing allowing for the investigation of the within-host genetic diversity of *S. pneumoniae*. Analysis of the resulting genomic data reveals that the technique has an increased sensitivity to detect instances of multiple pneumococcal carriage of 17.7% over latex sweeps and that different pneumococcal lineages are associated with increased and decreased rates of multiple carriage. The impact of antibiotic resistance and treatment within the host is examined with treatment identified as a risk factor for the transmission of multidrug resistant lineages. The evolution of pneumococcal lineages within the host is explored allowing for the consideration of selection and the identification of a unique mutational spectrum consistent with pre-

vious laboratory experiments in *E. coli*. Using minority variant calls found in each *S. pneumoniae* sample, routes of transmission are then determined, including those between mother and child. This indicated that mothers were more likely to transmit pneumococci to their children during the first year of the child's life than in the subsequent year.

Finally, a similar analysis was performed on a large collection of over 1000 SARS-CoV-2 isolates, sequenced in replicate, from the East of England. In addition to considering the selective and mutational processes acting on the virus within the host, a high rate of independent recurrent within-host mutations is identified which could complicate efforts to infer transmission chains using within-host minority variants.

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications. Including: Chapter 1 [1]; Chapter 2 [3, 4]; Chapter 3 [5]; and Chapters 4 and 9 [2, 6].

- [1] Jukka Corander, Nicholas Croucher, Simon Harris, John Lees, and Gerry Tonkin-Hill. *Handbook of Statistical Genomics*. Ed. by David Balding, Ida Moltke, and John Marioni. 1st ed. Vol. 162. Wiley, July 2019, pp. 997–1020. ISBN: 9781119429142, 9781119487845. DOI: [10.1002/9781119487845.ch36](https://doi.org/10.1002/9781119487845.ch36). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119487845.ch36>.
- [2] William L Hamilton*, Gerry Tonkin-Hill*, et al. “COVID-19 infection dynamics in care homes in the East of England: a retrospective genomic epidemiology study.” en. In: *medRxiv* (Sept. 2020), p. 2020.08.26.20182279. DOI: [10.1101/2020.08.26.20182279](https://doi.org/10.1101/2020.08.26.20182279). URL: <https://www.medrxiv.org/content/10.1101/2020.08.26.20182279v1.abstract>.
- [3] 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.” In: *Wellcome Open Res* 3 (July 2018), p. 93. ISSN: 2398-502X. DOI: [10.12688/wellcomeopenres.14694.1](https://doi.org/10.12688/wellcomeopenres.14694.1). URL: <https://wellcomeopenresearch.org/articles/3-93/v1>.
- [4] Gerry Tonkin-Hill, John A Lees, Stephen D Bentley, Simon D W Frost, and Jukka Corander. “Fast hierarchical Bayesian analysis of population structure.” en. In: *Nucleic Acids Res.* (May 2019). ISSN: 0305-1048, 1362-4962. DOI: [10.1093/nar/gkz361](https://doi.org/10.1093/nar/gkz361). URL: <http://dx.doi.org/10.1093/nar/gkz361>.
- [5] Gerry Tonkin-Hill, Neil MacAlasdair, et al. “Producing polished prokaryotic pangenomes with the Panaroo pipeline.” In: *Genome Biol.* 21.1 (July 2020), p. 180. ISSN: 1465-6906, 1474-760X. DOI: [10.1186/s13059-020-02090-4](https://doi.org/10.1186/s13059-020-02090-4). URL: <https://doi.org/10.1186/s13059-020-02090-4>.
- [6] Gerry Tonkin-Hill*, Inigo Martincorena*, et al. “Patterns of within-host genetic diversity in SARS-CoV-2.” en. Dec. 2020. URL: <https://www.biorxiv.org/content/10.1101/2020.12.23.424229v1>.

*Joint authorship

ACKNOWLEDGEMENTS

Thanks first and foremost go to my supervisors who have all been incredibly supportive over the last four years. Stephen Bentley for introducing me to bacterial genomics and the fascinating Maela dataset and for always being positive, friendly and patient despite the many unplanned projects I became distracted with. Jukka Corander who, in addition to providing invaluable advice on all things statistical genomics, was incredibly welcoming in inviting both Lauren and myself to Oslo for 6 months and introducing us to the delicious Skolebrød. Simon Frost who welcomed me at the Vet School in Cambridge, patiently answered all my questions each week, and who, despite moving to the U.S.A, continued to provide advice even helping to arrange for some of the analysis in this thesis to be run on Microsoft's cloud computing service.

Thanks also go to the many collaborators who have helped me throughout the PhD. John Lees who was instrumental in me joining the lab at the start of the PhD and who has been incredibly generous with his time and advice. Special thanks to Julian Parkhill who invited me into his group meetings and was always willing to answer my bacterial genomics questions. In particular, his extensive knowledge of bacterial pangenomes was essential to the development of Panaroo. Paul Turner and Nick Croucher, who have been critical both in the initial design and subsequent analysis of the Maela deep sequencing study. Claire Ling and Zannah Salter for their Herculean efforts in the processing of thousands of nasopharyngeal swabs. Caroline Colijn for helpful discussions on the TransPhylo algorithm and the analysis of transmission. Dominic Kwiatkowski, Inigo Martincorena, Will Hamilton, Roberto Amato and many others at the Sanger Institute and the University of Cambridge who took time out of their usual research to help with the response to the SARS-CoV-2 pandemic.

To the many people in teams 81 and 264 at the Sanger Institute and in the Biostatistics department at the University of Oslo. In particular, Neil MacAlisdair, my fellow PhD student in the lab, for sharing many Hawaiian pizzas.

Thanks to my fellow PhD students in the Wellcome Trust MGM programme, including Olly, Daniel, Tim and Ivan, who have made the last four years much more enjoyable. Thanks also to Gos Micklem and the other MGM directors and administrators for organising an excellent PhD program.

A big thanks to my family and friends, especially my sister Gina and mother Suzy, who have been incredibly supportive throughout the PhD and were even willing to read parts of this thesis.

Finally, thank you to my wife, Lauren, who managed to stick with me throughout the PhD even after being deported by the U.K. home office, moving between three countries and breaking two arms. Despite all of this she remained a constant source of positivity and fun, even managing to change New Zealand immigration law to get me back into the country.

CONTENTS

Outline	1
I BACTERIAL GENOME ANALYSIS	3
1 INTRODUCTION	5
1.1 A brief history of bacterial genomics	5
1.2 Bacterial genomes	7
1.2.1 Horizontal gene transfer	8
1.3 Bacterial genome analysis	9
1.3.1 Population structure	9
1.3.2 Phylogenetic reconstruction	10
1.3.3 Phylodynamics and dating	12
1.3.4 Recombination	13
1.3.5 The bacterial pangenome	14
1.3.6 Transmission	14
1.3.7 Phenotype association studies (GWAS)	15
2 POPULATION STRUCTURE	17
2.1 Background	17
2.1.1 Linkage disequilibrium	18
2.1.2 Distance-based methods	18
2.1.3 Model-based methods	19
2.1.4 Dirichlet process mixture model	20
2.1.5 Bayesian hierarchical clustering	21
2.1.6 Bayesian analysis of population structure (BAPS) and hierBAPS	21
2.2 Methods	22
2.2.1 Fast hierarchical Bayesian analysis of popula- tion structure	22
2.2.2 Initial clustering	22
2.2.3 Bayesian hierarchical clustering	23
2.2.4 Hyperparameter selection	24
2.2.5 Conditioning on a pre-computed hierarchy	25
2.2.6 Cluster stability using the Bootstrap	26
2.2.7 Implementation	26
2.2.8 Simulation and evaluation	26
2.3 Results	29
2.3.1 Simulations show that fastbaps cluster accuracy is the same as previous methods	29
2.3.2 Fastbaps is efficient and accurate on a diverse range of viral and bacterial datasets	31
2.3.3 Fastbaps accurately clusters previously intractable viral and bacterial datasets containing thousands of samples	35

2.4	Discussion	40
3	PANGENOME INFERENCE	43
3.1	Background	43
3.2	Methods	46
3.2.1	Panaroo algorithm	46
3.2.2	Initial graph creation	46
3.2.3	Contig ends	47
3.2.4	Contamination	47
3.2.5	Mistranslation correction	47
3.2.6	Collapsing gene families	48
3.2.7	Identifying missing genes	48
3.2.8	Thresholds	49
3.2.9	Input and output	49
3.2.10	Structural variation	49
3.2.11	Pre- and post-processing	50
3.2.12	Simulation and method comparison	50
3.3	Results	51
3.3.1	Analysis of a <i>Mycobacterium tuberculosis</i> outbreak in London	51
3.3.2	Performance on simulated populations	54
3.3.3	A diverse <i>Klebsiella pneumoniae</i> collection	56
3.3.4	Pangenome association analysis with Pyseer	59
3.3.5	Pangenome evolutionary dynamics	61
3.3.6	Computational performance	63
3.4	Discussion	64
4	TRANSMISSION INFERENCE	67
4.1	Background	67
4.2	Pairwise transmission inference	68
4.2.1	Calculating SNP distance matrices	69
4.2.2	Transcluster	70
4.3	Transmission network inference	73
4.3.1	Transphylo	75
4.3.2	Results	80
4.4	Transmission inference with deep sequencing	83
4.4.1	A simulation based alternative	86
4.4.2	Likelihood-free inference	87
4.4.3	Synthetic likelihood	88
4.4.4	Bayesian optimisation	88
4.4.5	Importance resampling	90
4.4.6	Simulating transmission	91
4.4.7	A synthetic likelihood approach to transmission inference	92
4.4.8	Results	95
4.4.9	Limitations and further work	100
4.5	Discussion	100

II THE PNEUMOCOCCUS	103
5 INTRODUCTION	105
5.1 Pneumococcal history and biology	105
5.1.1 The pneumococcal capsule	106
5.1.2 Vaccines and immunity	107
5.1.3 Other virulence factors	108
5.1.4 Antibiotic resistance	108
5.1.5 Pneumococcal recombination	109
5.1.6 Transmission	112
5.2 The Maela deep sequencing study	113
5.2.1 Background	113
5.2.2 Sample selection	117
5.2.3 Analysis outline	118
6 LINEAGES, SEROTYPES AND OTHER VIRULENCE FACTORS	119
6.1 Background	119
6.1.1 Strategies in analysing deep sequencing data . .	119
6.1.2 Multiple carriage	120
6.2 Methods	121
6.2.1 Lineage deconvolution	121
6.2.2 Serotype calling	123
6.2.3 Virulence calls	124
6.2.4 Association with multiple infection	125
6.3 Results	126
6.3.1 Quality control	126
6.3.2 Verification of mGEMS/Mash pipeline	130
6.3.3 Verification of serotype calls	130
6.3.4 Deep sequencing increases sensitivity in detecting multiple carriage	132
6.3.5 Differences in carriage rates between mothers and children	136
6.3.6 Prevalence in multiple carriage varies by lineage and serotype which has implications for surveillance and models of pneumococcal strain structure	136
6.3.7 Little evidence for founder advantage in pneumococcal transmission	140
6.3.8 Evidence for preferential co-colonisation	144
6.3.9 Virulence factors are negatively associated with multiple carriage	146
6.4 Discussion	147
7 ANTIBIOTIC RESISTANCE AND TREATMENT	151
7.1 Background	151
7.1.1 Mechanisms of resistance	151
7.1.2 Maintenance of resistance in <i>S. pneumoniae</i> . . .	152
7.1.3 Surveillance of resistance	153
7.1.4 Impacts of antimicrobial treatment	154

7.2	Methods	154
7.2.1	Antimicrobial resistance typing	154
7.3	The impacts of antimicrobial treatment	155
7.3.1	Bacterial GWAS	155
7.3.2	Treatment association study design	157
7.4	Results	160
7.4.1	Verification of the mGEMs/CDC pipeline	160
7.4.2	Whole plate sweeps identify 14.6% more antibiotic resistance calls	161
7.4.3	Resistance elements are less likely to be observed in mixed infections	161
7.4.4	Resistance levels differ between mothers and children	164
7.4.5	Treatment	167
7.4.6	Locus associations	170
7.5	Discussion	176
8	MUTATION, SELECTION AND TRANSMISSION	179
8.1	Background	179
8.1.1	Transmission	179
8.1.2	Within-host mutation	180
8.1.3	Within-host selection	181
8.1.4	Characterising minority variants presents additional challenges	182
8.2	Methods	183
8.2.1	Calling within-host variants	183
8.2.2	Variant calling pipeline	184
8.2.3	Identifying problematic regions	184
8.2.4	Transmission	185
8.2.5	Mutational spectrum	187
8.2.6	Selection	188
8.3	Results	188
8.3.1	Transmission links suggest strong geographic structure	188
8.3.2	Asymmetric transmission between mother and child	190
8.3.3	Close transmission pairs are more likely to share minority variants	191
8.3.4	Despite significant heterogeneity in the diversity of minority variants in carriage there is evidence for parallel within-host evolution	193
8.3.5	The mutational spectrum of <i>S. pneumoniae</i> is heterogeneous across different timescales	195
8.3.6	Evidence for selection acting within the host	197
8.4	Discussion	201

III	SARS-COV-2	203
9	TRANSMISSION AND WITHIN-HOST EVOLUTION OF SARS-COV-2	205
9.1	Background	205
9.1.1	SARS-CoV-2 transmission in care homes	206
9.1.2	Within-host diversity of SARS-CoV-2	207
9.1.3	Data	207
9.2	Methods	208
9.2.1	Variant calling	208
9.2.2	Mutational spectrum and selection	208
9.2.3	Consensus phylogeny construction	209
9.2.4	Transmission model	209
9.2.5	Identification of potential mixed infections	209
9.3	Results	209
9.3.1	Transmission in care homes	209
9.3.2	Healthcare workers	211
9.3.3	Within-host diversity	214
9.3.4	Mutational spectrum	216
9.3.5	Within-host mutational load and selection	217
9.3.6	Longitudinal samples of the same host	219
9.3.7	Shared within-host variants	220
9.3.8	Mixed infections	225
9.4	Discussion	227
10	CONCLUSION	231
10.1	Bacterial genome analysis	231
10.2	Genetic diversity of the pneumococcus within the host and in transmission	232
10.3	Within-host diversity and transmission of SARS-CoV-2	234
10.4	Future work	235
	BIBLIOGRAPHY	237

LIST OF FIGURES

Figure 1	A diagram outlining the fastbaps algorithm. . .	25
Figure 2	A comparison of population structure inference algorithms on simulated data	30
Figure 3	Comparison of phylogenies and inferred clusters	32
Figure 4	A comparison of the different clustering algorithms across a range of species.	33
Figure 5	Time complexity of the different fastbaps modes	35
Figure 6	UMAP plots of Fastbaps clustering on both a large Pneumococcal and HIV dataset	37
Figure 7	Violin plots of fastbaps HIV clusters by country and subtype	38
Figure 8	UMAP plots of the inferred clusters of pneumococcal and HIV isolates using the k-means algorithm	39
Figure 9	Pangenome analysis pipelines	45
Figure 10	Panaroo error correction pipeline	52
Figure 11	Pangenome for 413 <i>M. tuberculosis</i> genomes . .	54
Figure 12	CheckM completeness and contamination scores	55
Figure 13	Pangenome errors on simulated data	57
Figure 14	Pangenome quality control	58
Figure 15	Pangenome of a global <i>K. pneumoniae</i> dataset .	60
Figure 16	Pangenome GWAS results	62
Figure 17	Pneumococcal pangenome evolutionary dynamics	64
Figure 18	Computational requirements for pangenome tools	65
Figure 19	Computational time for pairwise distance calculations	71
Figure 20	Diagram of Transcluster algorithm	72
Figure 21	Adapted TransPhylo algorithm diagram	80
Figure 22	Simulation results of adapted TransPhylo algorithm	82
Figure 23	Pairwise transmission distances in the Maela refugee camp	83
Figure 24	Inferred pneumococcal generation times . . .	84
Figure 25	Correspondence between a phylogeny and transmission chain	85
Figure 26	Three possible transmission scenarios	93
Figure 27	Accuracy of endsim algorithm	98
Figure 28	Within-host population size histograms	99
Figure 29	Global mortality rate of <i>S. pneumoniae</i>	106
Figure 30	Maela refugee camp	116

Figure 31	Proportion of resistance following treatment . . .	127
Figure 32	Coverage and contamination histograms	129
Figure 33	Contamination boxplots	129
Figure 34	Lineage and serotype method verification . . .	131
Figure 35	Serotype calls Venn diagram	133
Figure 36	Serocall and mSWEEP comparison	134
Figure 37	Sensitivity of latex sweeps versus deep sequencing	135
Figure 38	Carriage rates in mothers and children	137
Figure 39	Lineage proportion boxplots	141
Figure 40	Serotype proportion boxplots	142
Figure 41	Lineage competition example	145
Figure 42	Bacterial GWAS diagram	157
Figure 43	Verification of resistance calls	162
Figure 44	Resistance calls single colony vs plate sweep .	163
Figure 45	Resistance lineage frequencies	165
Figure 46	Counts of resistant samples in mothers and children	168
Figure 47	GPSC1 frequencies before and after treatment	169
Figure 48	Presence/absence of unitigs associated with antibiotic treatment	173
Figure 49	Frequency of unitigs associated with antibiotic treatment	174
Figure 50	Minority variant filter example	186
Figure 51	Transmission using minority variants vs geographic distance	189
Figure 52	Mother/child carriage history example	192
Figure 53	Total transmissions by direction between mothers and children	193
Figure 54	Bottleneck inference using shared minority variants	194
Figure 55	Changes in number of minority variants over carriage	195
Figure 56	Histograms of the frequency a minority variant is seen in different hosts	196
Figure 57	Histogram of trinucleotide within and between host mutational frequencies	197
Figure 58	Histogram of within and between host mutational frequencies	198
Figure 59	Global within-host d_N/d_S ratios	200
Figure 60	Significant gene d_N/d_S ratios	201
Figure 61	Pairwise SNP and date distances in close SARS-CoV-2 transmission pairs	212
Figure 62	Inferred SARS-CoV-2 transmission networks .	213
Figure 63	Healthcare worker transmission network . . .	215

Figure 64	Dot plot indicating the concordance of VAFs between replicates	216
Figure 65	SARS-CoV-2 mutational spectrum	218
Figure 66	Selection in SARS-CoV-2	219
Figure 67	Longitudinal within-host variant frequencies in SARS-CoV-2	220
Figure 68	Number of samples that have a within-host variant for each site in the SARS-CoV-2 genome	221
Figure 69	Phylogeny with links indicating shared within-host variants	223
Figure 70	Probability of transmission vs median number of shared within-host variants	224
Figure 71	Within-host VAF heatmaps	225
Figure 72	Dot plot indicating VAFs of mixed infections .	226

LIST OF TABLES

Table 1	Parameters used to simulate genetic population structure using scrm	28
Table 2	Datasets used in the comparison of different clustering methods	29
Table 3	Total CPU time (seconds) of the different population structure algorithms.	34
Table 4	Maximum memory usage of the different population structure algorithms	34
Table 5	Parameters used to simulate pangenomes using the Infinitely Many Genes model	51
Table 6	Outlying pneumococcal lineages by within-host frequency	140
Table 7	Outlying pneumococcal serotypes by within-host frequency	140
Table 8	Outlying pneumococcal lineages by multiple carriage	143
Table 9	Outlying pneumococcal serotypes by multiple carriage	143
Table 10	Significant virulence factors in multiple carriage	148
Table 11	Antibiotic abbreviations and resistance genes .	153
Table 12	Association test results of resistance vs mSWEEP frequencies	164
Table 13	Association test results of resistance vs multiple carriage	166
Table 14	Association test results of resistance vs multiple carriage using Ariba	166
Table 15	Association test results of resistance rates in mothers and children	167
Table 16	Unitig presence/absence vs resistance associations	171
Table 17	Unitig frequency vs resistance associations . .	175

ACRONYMS

GPSC Global Pneumococcal Sequencing Cluster

MCMC Markov Chain Monte Carlo

GWAS Genome Wide Association Study

BAPS Bayesian Analysis of Population Structure

GLM Generalised Linear Model

LMM Linear Mixed Model

CDC Center for Disease Control

SNP Single Nucleotide Polymorphism

PBP Penicillin Binding Protein

PEN Penicillin

CFX Cefuroxime

ERY Erythromycin

TAX Cefotaxime

CFT Ceftriaxone

TET Tetracycline

DOX Doxycycline

COT Cotrimoxazole

CLI Clindamycin

AMO Amoxicillin

MER Meropenem

CHL Chloramphenicol

RIF Rifampin

OUTLINE

The tracking of a pathogen's evolution over different timescales is essential: for the design of vaccines, to control antibiotic resistance, for the surveillance of disease, to identify transmission chains and to understand the mechanisms behind the pathogen's evolution.

This thesis considers the analysis of pathogen evolution over different timescales with a particular emphasis on the within-host evolution of *Streptococcus pneumoniae* and SARS-CoV-2. It is structured in three parts. Part I describes the development of statistical and bioinformatics methods for the analysis of bacterial genome sequencing data. Part II presents the analysis of a large genome sequencing study on the within-host diversity and transmission of *S. pneumoniae*. Using similar techniques, Part III discusses the patterns of within-host diversity in SARS-CoV-2 and the implications for the inference of transmission.

Part I introduces many of the common problems and techniques used to analyse bacterial genomes and then describes novel approaches for solving three major issues. The problem of identifying population structure in large pathogen genomic datasets is discussed and a method presented that substantially improves the speed at which model based clustering of genetic data can be performed. Next, a new approach to inferring the pangenome is presented which substantially reduces the number of errors introduced by automated gene annotation algorithms. Finally, approaches to inferring transmission chains using genomic data are examined, including a novel likelihood-free transmission inference approach.

Part II describes the history and biology of *S. pneumoniae* as well as the Maela deep sequencing study which aims to shed light on the genetic diversity of *S. pneumoniae* within the host. Following this, methods for deconvoluting different lineages of *S. pneumoniae* are verified including approaches for identifying virulence and resistance factors. The improved sensitivity of these approaches is discussed along with the implications of varying rates of virulence and resistance in multiple carriage. Finally, within-host minority variants are considered both in the context of investigating transmission within the Maela camp as well as examining the mutational and selective forces driving the evolution of *S. pneumoniae* within the host.

The outbreak of SARS-CoV-2 in late 2019 led many scientists to shift their attention to the analysis of SARS-CoV-2 and the corresponding disease COVID19. Part III of this thesis describes an analysis of the mutational and selective processes driving SARS-CoV-2

evolution within the host and the implications these results have for the inference of transmission chains.

Part I

BACTERIAL GENOME ANALYSIS

Part I focuses on methods for the analysis of bacterial genome sequencing data with a particular focus on the inference of population structure, pangenomes and transmission.

INTRODUCTION

1.1 A BRIEF HISTORY OF BACTERIAL GENOMICS

The first organism to have its entire genome sequenced was the bacterium *Haemophilus influenzae* [170]. In the following decades, genomics has dramatically increased our understanding of the bacterial evolution, improving our awareness of how they interact with their host, their environment and each other. Recently, bacterial genomics has started to replace many traditional "wet-lab" microbiology techniques for characterising bacterial samples with a notable impact on determination of antibiotic resistance and other important phenotypes for bacterial pathogens [314].

The high costs of sequencing meant initial studies used single genomes to represent a taxonomix group. The first study to consider more than one species from the same genus was conducted in 1996 when multiple species of *Mycoplasma* were sequenced using chain termination method developed by Fred Sanger [223, 435]. It would be another three years before multiple genomes from the same species were sequenced in a single study looking into *Helicobacter pylori* [378]. This was followed in 2006 with the first use of bacterial sequencing to identify within-host evolution of *Burkholderia mallei* in a patient with samples from both the liver and blood [425]. First generation 'Sanger sequencing' enabled the realisation that the process of bacterial evolution varies widely from species to species. Extensive horizontal gene transfer was identified in *Escherichia coli* [214]. The characterisation of horizontal gene transfer in bacterial evolution has subsequently led to the definition of the bacterial pangenome which encompasses both the 'core' genes, found in every member of a species, along with the 'accessory' which are not always present [348]. Reductive evolution was identified in *Rickettsia prowazekii* and *Mycobacterium leprae*, which can occur when isolated populations adapt to a new restricted niche [9, 146]. In contrast to the plasticity in gene content observed in some species, a number of important pathogens were found to have highly stable pangenomes, evolving primarily through smaller mutations including single nucleotide polymorphisms (SNPs), indels and small deletions. These included *Mycobacterium tuberculosis*, a bacterium that currently is responsible for the most bacterial related deaths annually and was the leading cause of death from a single infectious agent in 2019 [79, 171, 518]. The substantial variation in the observed evolutionary characteristics of bacteria is perhaps not so surprising when one considers that bacteria have been evolving for at least 2 billion

years longer than eukaryotes and represent 35 times the biomass of all animals on the planet [18, 269].

Early bacterial genomics studies went beyond describing the evolution of the bacteria themselves to provide insights into human migration, notably in the evolution of *H. pylori* [159]. Other early applications of bacterial genome sequencing include the identification of potential vaccine antigens in the *Neisseria meningitidis* genome, which ultimately aided in the development of a new meningococcal vaccine [401].

With the introduction of 454 and Illumina technologies heralding the advent of second generation sequencing in the mid 2000's, bacterial genomics studies were able to increase markedly in size [326]. The first published application of these new technologies for bacteria was a study of 19 genomes of *Salmonella typhi* [227]. This was followed in 2010 by a pioneering genomic epidemiology study of MRSA, where the authors were able to identify plausible transmission links between continents and within the hospital wards using whole genome sequences of 63 isolates [209]. The following year, genome sequencing was used in a very high profile case that implicated UN peacekeepers in an outbreak of cholera in Haiti [73]. An even larger dataset allowed for the characterisation of vaccine escape in *Streptococcus pneumoniae* where pneumococcal serotypes including in the vaccine were shown to reduce in frequency and be replaced with non-vaccine types through both strain replacement as well as serotype switching via transformation [100]. Subsequently, dense sequencing of over 3000 pneumococcal isolates from a refugee camp in Thailand, identified changes to the pathogens population genomics in response to temporal variation in antibiotic use [71]. This large and well documented dataset led to one of the first bacterial genome wide association studies (GWAS), where novel antibiotic resistance determinants were identified [72]. Bespoke methods have since been developed to investigate association studies in bacterial genomics [144, 293]. These resulted most recently in one of the first joint host-pathogen genome wide association studies looking at meningitis caused by *S. pneumoniae* [288]. Other notable uses of second generation sequencing include the expansion of the field of bacterial genomics to ancient pathogens, starting with an ancient *Yersinia pestis* genome from the black death [38]. Modes of transmission have also been resolved using second generation sequencing with the identification of the transmission of nontuberculosis mycobacterium between cystic fibrosis patients as well as the migration of microbiota from mother to child [46, 166]. Finally, a large-scale second generation sequencing effort was able to shed light on the evolution of *E. coli* over 60,000 generations in a long term evolution experiment, showing that the targets of selection changed over the course of the experiment [195]. The introduction of third generation (rapid long

read) sequencing has led to further discoveries and innovation in bacterial genomics and has opened the way for bacterial genomics to be used in the clinic. This has led some to suggest that we are approaching a ‘singularity’ whereby many tasks commonly performed using traditional microbiological techniques will soon be replaced by bacterial genomics [314]. Early indications of the use of these technologies in the clinic include the development of a rapid sequencing approach to identify resistance determinants in *Staphylococcus aureus* and *M. tuberculosis* [40]. The swift rise of the field of synthetic genomics is also starting to include bacteria with the first entirely synthetic *E. coli* genome generated in 2019 [178]. While there are many parallels between the methods developed for human population genetics and those for bacteria, the distinct biology of bacteria has necessitated the development of bacteria specific approaches.

1.2 BACTERIAL GENOMES

The genomes of bacteria typically consist of a small number of DNA molecules or replicons varying in size from $\sim 1.5 \times 10^5$ nucleotides (nt) in the case of *Endosymbiont Carsonella* to the behemoth *Sorangium cellulosum* with a genome size of $\sim 1.3 \times 10^7$ nt [372, 438]. Essential replicons are known as chromosomes, with the majority of bacterial species consisting of just a single chromosome. Notable pathogenic exceptions include *Vibrio cholerae* and *Burkholderia cenocepacia* which consist of two chromosomes [215, 224]. While bacterial cells often contain multiple copies of their chromosome, the vast majority of bacterial species are haploid organisms. Rare exceptions include the bacterial symbiont of the surgeonfish, *Epulopiscium spp*, a very large bacterium that has been observed to contain tens of thousands of copies of its chromosome allowing it to maintain certain unstable genetic features that would otherwise be fatal [352]. In contrast to chromosomes, non-essential replicons are known as plasmids. The presence of plasmids varies by species. Bacteria are able to accommodate multiple different plasmids, which can each have different copy numbers within the cell. As plasmids are non-essential, they often carry genes that benefit the survival of the bacterium in certain conditions including under antibiotic pressure. Plasmids can be broadly classified into those that are conjugative and are able to independently initiate the process of transfer to new other bacterial cells (conjugation) and non-conjugative, which describes those plasmids that are reliant on other conjugative plasmids to be transferred between cells [455]. Unlike most eukaryotes which consist of linear chromosomes, the majority of bacterial chromosomes and plasmids are circular. However, as with most bacterial genomic characteristics, there are exceptions. These include *Streptomyces coelicolor* and the tick borne bacterium *Borrelia afzelii* which can cause a variant of Lyme

disease and [29, 59]. Like eukaryotes, in addition to varying gene content, bacterial genomes evolve via both point mutations and rearrangement. In contrast to sexually reproducing organisms, bacteria do not recombine at each generation. Instead, some species evolution is nearly entirely clonal whilst others recombine and exchange genes via horizontal gene transfer.

1.2.1 *Horizontal gene transfer*

Initial efforts at reconstructing phylogenetic relationships between organisms focused on using molecular information in different genes and proteins [542]. With the coming of the genomic age it was soon realised that using different genes led to the construction of conflicting phylogenies for the same collection of genomes [221]. Coupled with the well established knowledge that bacteria as well as other organisms could exchange genetic material the concept of horizontal gene transfer between organisms emerged in the 1990's [39, 199]. There are three main mechanisms of horizontal gene transfer in bacteria; transformation, conjugation and transduction. Gene transfer agents can also mediate horizontal gene transfer but are rarer and not discussed here [328].

1.2.1.1 *Transformation*

Natural transformation refers to a bacterium's ability to uptake, integrate and express extracellular DNA in its chromosome. In order for transformation to occur, bacterial cells must first enter a state of competence. This has been found to naturally occur in approximately 1% of bacterial species and has been found to rely on approximately 20 to 50 proteins [479]. Common human pathogens such as *S. pneumoniae*, *Neisseria gonorrhoeae* and *Haemophilus influenzae* are known to be naturally transformable [315]. Bacteria often enter a state of competence in response to stressors such as antimicrobial treatment or starvation. A prerequisite for transformation is the release of extracellular DNA by the donor cell. Some bacteria species such as *S. pneumoniae* have developed fratricidal mechanisms to aid in the lysing of donor cells to increase the rate of transformation [502].

1.2.1.2 *Conjugation*

Conjugation involves the transfer of genetic material through cell to cell junctions and is commonly associated with the transfer of plasmids between cells [479]. Conjugative elements such as plasmids can facilitate the transfer of other genetic material by picking up passenger genes or by integrating themselves into other elements. Transposable elements or insertion sequence (IS) elements that can change their location within a genome can aid in such integrations by provid-

ing regions of high homology between replicons which allows for the homologous recombination of otherwise unrelated genetic material [36].

Multiple mechanisms for conjugation exist. One of the most complex, involves large self transmissible plasmids that include genes that allow for the creation of a pilus, a proteinaceous fibre that mediates the contact between the donor and recipient cells [58]. While multiple other mechanisms exist, some plasmids are not self transmissible. Instead these rely on the presence of a separate self-transmissible plasmid for conjugation [479].

1.2.1.3 *Transduction*

Transduction involves the integration of foreign DNA into a bacterial cell via a viral vector or bacteriophage [479]. Transduction can be both specialised, where the phage can carry only certain parts of the bacterial genomes or generalised where the phage can be a vector for any part of the bacterial chromosome.

Bacteriophages can be grouped into two main categories. Virulent phages are characterised by an infectious cycle that is always lytic, causing the destruction of the host bacterial cell via lysis. Lytic bacteriophages recruit the machinery of the host cell to generate new virions which are then released after the host cell is lysed. In contrast, temperate phages usually exist in a lysogenic cycle in which the phage can integrate within the bacterial genome as a prophage. Temperate phages can follow a lysogenic cycle in some circumstances and can cause lysis when the prophage is activated.

1.3 BACTERIAL GENOME ANALYSIS

While many statistical and bioinformatics methods used in the analysis of human genomes are applicable to bacterial genomes, their unique biology has necessitated the development of novel approaches. In many cases, methods originally developed for use in bacterial genomics have subsequently been used in human genetics. Notable examples include the use of pangenome graphs and the use of kmers in genome wide association studies [143, 293, 348, 414]. In the following section we briefly describe some of the common problems in the analysis of bacterial genomes. The subsequent chapters go into a finer level of detail on three specific areas and describe new approaches to solving these problems.

1.3.1 *Population structure*

Genetic population structure has been a central problem in genetics since the early efforts by Ronald Fisher and Sewall Wright in the

early 20th century. A thorough understanding of population structure is essential for the use of genetics in understanding past migration events, population splits and population bottlenecks. Population structure must also be considered and controlled for in order to accurately infer the association between genetics and various phenotypes of interest in genetic association studies. In general, many of the methods used for analysing population structure in humans and other species are also used in bacterial genetics. However, the sometimes highly clonal evolution of certain bacterial species can require careful consideration.

Initial efforts to investigate genetic structure focused on the use of various indices such as F-statistics [521]. In the following decades more detailed statistical clustering models and computer simulations were developed that described the relatedness of gene pools both backward and forwards over generations through time [407]. While these approaches provided a significant advance over the more simple statistics including being able to identify and describe admixed populations, many of these approaches were computationally expensive. This has led to the development of methods based on pairwise distance metrics, which while heavily dependent on the metric and clustering method chosen provide substantial increases in the speed and size of datasets that can be analysed. [Chapter 2](#), provides a more detailed background on the methods used to infer population structure in bacteria and details recent efforts to improve the performance of statistical model based methods allowing them to be used on very large datasets.

1.3.2 *Phylogenetic reconstruction*

The evolutionary history of a set of bacterial genomes in the absence of recombination can be represented as a phylogenetic tree. Here, the topology of the tree indicates the relationships between isolates based on vertical evolution with the tips of the tree representing the sampled sequences and the internal nodes indicating historical ancestors. The branch lengths indicate the amount of genetic divergence between isolates, or in the case of a time scaled phylogeny, the time between isolates. In a rooted phylogeny the earliest node indicates the most recent common ancestor (MRCA) of the sampled population. Phylogenies can be used to date the emergence of important phenotypes such as antibiotic resistance, to inform transmission inference, to aid in epidemiological studies, and to shed light on past population dynamics and selection. Initial methods to build phylogenetic trees from sequence data relied on pairwise distance based approaches. Subsequently, maximum parsimony and maximum likelihood models and, more recently, fully Bayesian model based approaches have been developed.

1.3.2.1 *Distance based phylogenetics*

One of the earliest and simplest algorithms for generating a hierarchical relationship between sequences was the unweighted pair group method with arithmetic mean (UPGMA) method often attributed to Sokal and Michener [460]. This agglomerative ‘bottom up’ approach assumes that the underlying phylogeny is ultrametric in that the distances from the root to every tip are equal. This effectively assumes a constant molecular clock, that evolves at the same rate in each lineage. The neighbour joining (NJ) method of Saitou and Nei relaxes the assumption of a constant molecular clock and, provided that the underlying distance matrix is additive, guarantees that the resulting phylogeny will be correct [433]. A distance matrix is additive if it satisfies the following extension to the triangle inequality such that given genomes i, j, m and n and distance matrix D

$$D(i, j) + D(m, n) \leq \max\{D(i, n) + D(j, m), D(j, n) + D(i, m)\}$$

The original NJ algorithm had a complexity of $O(N^3)$ where N was the underlying number of sequences. This made it fast for small sized datasets but infeasible for large populations. However, heuristic extensions to the approach have led to dramatic speed improvements making the method applicable to datasets with tens of thousands of sequences [452].

1.3.2.2 *Maximum parsimony*

One of the simplest methods that does not rely on a distance matrix is maximum parsimony. Maximum parsimony attempt to find the phylogeny that minimised the number of character changes that would have to occur along each branch [161, 168]. While it is simple to calculate the number of changes required given a pre-existing phylogeny, finding the optimal tree is an NP hard problem, meaning that an exact algorithm is too computationally expensive for all but very small numbers of isolates [174]. Similar to the NJ algorithm, heuristic search methods have been proposed to help resolve this issue [472].

1.3.2.3 *Maximum likelihood phylogenetics*

Most distance based methods such as neighbour joining do not explicitly model the mechanisms of evolution. However, maximum likelihood methods score a given phylogeny according to an underlying statistical model and attempt to find the phylogeny that maximises the corresponding likelihood or score. This is similar to the maximum parsimony approach, except that the score is based on a specific statistical model of sequence evolution. Like maximum parsimony, heuristic search algorithms are usually used to find a suitable tree. These

approaches typically take a multiple sequence alignment as input and model each column in the alignment independently. Depending upon the underlying evolutionary model, the transition rates between the character states can either be given as known as is the case in the LG matrix for amino acids or estimated as part of the fitting process as with the general time-reversible model of nucleotide evolution [284, 476]. Extensions to these methods include allowing these rates to vary between sites, which can compensate for differences in selection at different locations in the genome as well as modelling invariant sites which are often excluded from multiple sequence alignments leading to biased inferences [297, 526]. Highly used implementations of this style of algorithm include Iqtree and RAXML [376, 464].

1.3.2.4 *Bayesian phylogenetics*

Maximum likelihood models implicitly assume a uniform prior over each phylogeny. Similar to other areas of applied statistics this assumption can be relaxed to combine a prior probability of a given phylogeny with the likelihood of the data. This allows for the inference of a posterior distribution usually inferred via Markov Chain Monte Carlo (MCMC) sampling. Implementations of this approach include MrBayes and BEAST [139, 232].

1.3.3 *Phylogenetics and dating*

As well as providing a picture of the evolutionary history of a set of bacterial isolates, phylogenies can also inform our understanding of the demographic history of the sampled genomes. The demographic history of a population is often modeled in one of two ways. Coalescent theory models the evolution of the underlying population backwards in time, merging lineages into a single ancestor [263] whereas birth-death models evolve forwards in time, where at each continuous time point either a birth or a death or neither can occur [259]. Unlike coalescent models, birth-death approaches account for the proportion of the population being sampled. Both these models have been extended to account for samples taken over varying time points which allows for the divergence between sequences to inform the estimates for the underlying substitution model. This allows for the dating of a phylogeny such that estimates of the dates of MRCAs can be inferred [140]. Dating can be performed using either a ‘strict clock’ where the substitution rate is assumed to be constant across all branches of the phylogeny or a ‘relaxed clock’ which allows the substitution rate to vary. While most bacterial species are better modelled by a relaxed clock there are exceptions including *Acinetobacter baumannii* [85]. Initial algorithms for estimating dated phylogenies were originally designed for analysing small numbers of viruses. The generation of very large data sets of pathogen sequences and the

large genome sizes found in bacteria has motivated the development of more efficient dating methods. This includes the least squares approach of To *et al.*, 2016, which assumes a linear relationship between time and branch length [481]. Subsequently, a number of other fast alternatives have also been developed including treedater and TreeTime [431, 496]. The often high rate of horizontal gene transfer in bacteria has also motivated the development of bacteria specific dating algorithms such as bactdater which explicitly accounts for recombination events inferred using other algorithms [126].

1.3.4 *Recombination*

Large segments of sequence inherited via different recombination mechanisms rather than vertically can substantially influence phylogenetic inference algorithms, often to the point of obscuring the signal provided by point mutations [101]. As a result a number of algorithms have been developed to identify and account for recombinations in pathogen sequence data. The two main strategies employed by these algorithms consist of either attempting to identify cases of sequence exchange within the sample or searching for evidence of divergent imported sequence into a closely related population. The first approach relies on both the donor and recipient sequences being present within the sample which presents a challenge for the analysis of many bacterial datasets where there is often a large amount of unsampled diversity, which is in some cases extinct in the current population. Algorithms that make use of this strategy include RDP3, GARD and the homoplasy test [273, 329, 336]. The second strategy led to the concept of the clonal frame which describes the common regions of a sample of genomes that have not undergone recombination since their MRCA [356]. This concept led to the development of the ClonalFrame algorithm which implements a Bayesian model framework to both detect recombinants and infer a corrected phylogeny [127]. While this algorithm is able to accurately account for recombination in small datasets of closely related isolates, it struggles to scale with larger collections of genomes, motivating the development of more efficient methods including the maximum likelihood version, ClonalFrameML [131]. Alternative approaches for solving this problem on large datasets include Gubbins and BratNextGen [104, 334]. In parts of this thesis we make use of Gubbins. Unlike ClonalFrame, Gubbins is an iterative algorithm that makes use of spatial scanning statistics to identify regions of the genome that have increased densities of substitutions. These are suggestive of horizontal gene transfer and can be excluded prior to running conventional maximum likelihood phylogenetic inference algorithms.

1.3.5 *The bacterial pangenome*

Driven by horizontal gene transfer the collection of genes that can be found across any one species of bacteria is often far greater than the number of genes in an individual isolate. This has led to the definition of the pangenome as the set of all genes present across a given species. Genes that are present in all strains within a species are known as core genes while those present in only a subset are referred to as accessory genes. Obtaining an accurate picture of a species pangenome is often important in understanding clinically significant traits such as virulence and antibiotic resistance [477]. Recent studies have also highlighted the importance of considering the pangenome in understanding the ecological and evolutionary processes driving the evolution of important pathogens such as *S. pneumoniae* and *E. coli*. In these species, negative frequency dependent selection (NFDS) acting on the bacterial pangenome has been suggested to be a major driving force in their evolution including in response to recent vaccine interventions, or the emergence of new clones in populations [86, 347]. A significant challenge in the analysis of bacterial pangenomes is the classification of homologous genes, usually defined by a percentage shared sequence identity, into either orthologous or paralogous clusters. Initial attempts borrowed methods originally developed for identifying orthologs between eukaryotic species [302, 419]. The increasing size of bacterial datasets subsequently led to the development of faster bacteria specific approaches [390]. In Chapter 3, we describe these approaches in a higher level of detail and describe recent efforts to correct for gene annotation errors which can significantly impact pangenome inference efforts.

1.3.6 *Transmission*

Tracking the transmission of a pathogen through a population is crucial to informing control measures and policies to limit spread. Traditionally, this has been achieved through epidemiological methods including contact tracing and case counting. Initial bacterial genomic typing methods such as MLST lacked the genetic resolution to inform estimates of transmission. However, with the reduction in WGS costs, genomics is now starting to be used routinely to infer pathogen transmission. The recent outbreaks of Zika, Ebola and SARS-CoV-2 have highlighted the utility of using genomics in combination with traditional epidemiological techniques to aid in the inference of transmission [160, 162, 189]. Prior to these high profile viral outbreaks, the use of genomics to track bacterial transmission within hospital wards and in foodbourne outbreaks was already starting to gain traction [196, 209].

Various approaches have been developed to infer transmission using genetic data. Simple approaches use the number of substitutions between samples from different hosts to rule out transmission [47]. Phylogenies can be constructed using WGS from multiple samples and used either as a proxy for transmission or as a component of a more complex model that also accounts for the transmission process itself [128, 209]. While a phylogeny built from single pathogen genomes per host can help to inform estimates of transmission, within-host variation can lead to discrepancies between a phylogeny and corresponding transmission network [515]. This has motivated the development of techniques that take advantage of additional genetic information on within-host diversity to infer transmission [115, 523]. Other methodological challenges limiting who-infected-whom studies include sparse or biased sampling, genetic recombination and mixed infections [182]. With the majority of method development in transmission inference focused on outbreaks, the application of these approaches to endemic pathogens has received little attention.

In Chapter 4 we review recent efforts to develop phylogenetic methods to infer transmission networks from whole genome sequencing data. We go on to describe extensions of these approaches to enable their application to endemic diseases rather than outbreaks as well as improvements in their computational efficiency. Finally, Chapter 4 reports a novel method for the inference of transmission using deep sequencing data, which enables the incorporation of additional information on within-host diversity.

1.3.7 *Phenotype association studies (GWAS)*

Identifying associations between genetic variation and important phenotypes such as antibiotic resistance and virulence can be vital in understanding the mechanisms behind these phenotypes. This in turn, can allow for the development of new drugs, vaccines and epidemiological intervention strategies to reduce the public health impact of bacterial pathogens. Genome wide association studies were originally developed in human genetics where single nucleotide polymorphisms (SNPs) are associated with a wide variety of phenotypes including heart disease and susceptibility to cancer. Such studies offer an alternative to performing hundreds of often expensive hypothesis driven wet-lab experiments. Instead these studies enable such wet-lab experiments to focus on the most likely candidate genes for a phenotype of interest. Direct application of GWAS methods developed for human genetics is not ideal for the analysis of bacterial datasets. Strong population structure and the existence of often large accessory genomes present added difficulties that are not addressed by conventional GWAS techniques. Instead, bacterial specific methods have been developed to address these issues. In the place of SNPs,

many bacterial GWAS methods use short sequence words of length k , known as k -mers [293]. The presence or absence of these k -mers can be determined without the need of a reference genome and consequently are able to account for variation in gene content. Extensions to this approach make use of unitigs which are variable length words that reduce the redundancy that can occur when multiple different k -mers represent the same genetic feature [245]. While GWAS methods in human genetics account for population structure, the sometimes highly clonal nature of bacteria can lead to a loss-of-power in such methods abilities to identify association in bacterial populations. This has led to methods that differentiate between ‘lineage-level’ effects and ‘locus-level’ associations [144]. Thus, when it is not possible to identify fine scale locus effect, these methods can instead report association between phenotypes and specific lineages. In bacterial species with high recombination rates such as *S. pneumoniae*, it is possible to identify locus level associations. In this case methods have been developed that account for population structure using a corresponding phylogeny, a kinship matrix, or most recently pangenome spanning regressions that consider all genetic features simultaneously [82, 289, 292]. A subset of these methods are described in more detail in [Part ii](#) where they are applied to the analysis of within-host deep sequencing data of *S. pneumoniae*.

POPULATION STRUCTURE

Declaration of contributions

Jukka Corander, Simon Frost and Stephen Bentley supervised this work. John Lees helped with testing of the rhierBAPS and fastbaps R packages.

2.1 BACKGROUND

Identifying clusters of pathogen genomes that are significantly more similar than would be expected if they had been randomly picked from the circulating population is an important part of the analysis of bacterial genome datasets. Determining sets of similar isolates is essential to the identification of outbreaks as well as the characterisation of bacterial migration patterns both locally and globally [99]. Accurate estimates of population structure are also required in the modelling of natural selection [272], recombination [334], bacterial pangenome evolution [86] and the maintenance of strain structure in some bacterial species [210].

Controlling for the confounding effects of population structure is an important step in association studies which attempt to identify associations between bacterial genetics and important clinically relevant phenotypes such as antibiotic resistance [72], virulence [503], ability to cause outbreaks [1, 288] and host range [503].

Early approaches to identifying genetic population structure in bacterial populations were often based on multilocus sequence typing (MLST) [322]. MLST classifies isolates according to their respective alleles in a number of short fragments in unlinked housekeeping genes. The MLST approach has since been expanded to consider the entire core genome (cgMLST) [350] with some additional extensions also considering accessory genes from the whole genome (wgMLST) [98]. While MLST based approaches have proven to be useful in tracking structure in bacterial populations, the approach relies on an existing database of groups which new genomes can be classified against. As such, it is not as readily applicable to previously unseen population structure. As a result, alternative methods that identify population structure *de novo*, similar to those used in the analysis of human genetics and animal populations, have become increasingly popular [407].

Methods for the *de novo* clustering of genotype data can generally be separated into those based on a statistical model of population

structure, where observations from each underlying cluster are assumed to be drawn from a parametric population genetics model [407], and distance-based methods that rely on more general clustering techniques such as k-means [250].

2.1.1.1 *Linkage disequilibrium*

An important consideration when investigating population structure and admixture in bacteria is the impact of linkage disequilibrium (LD). LD refers to the dependence of alleles at different locations on the same genome. It is usually the result of genetic recombination having less of an impact on loci that are in close proximity. The impact of LD on clustering can vary depending on the rate of recombination in the organism and the time-scale over which the samples have been taken. For very clonal organisms such as *Mycobacterium tuberculosis*, nearly all the sites in the genome will be in high LD. In this case a phylogenetic tree-building approach may be more suitable, such as maximum parsimony [54], maximum likelihood [464] or Bayesian methods [139]. Some bacteria, including *S. pneumoniae* and *Neisseria gonorrhoeae*, have much higher rates of recombination, and consequently constructing accurate phylogenies can be more difficult. The impact of LD on these more recombinant organisms is highly dependent on the time interval over which the samples were taken. If there are considerable gaps between sample dates there may be enough recombination for the impact of LD to be safely ignored, in which case methods such as Structure [407] and BAPS [87] can be used. For more densely sampled datasets clustering methods that account for recombination such as Structure version 2 [158, 473] and fineSTRUCTURE [283] are more suitable. In this chapter we focus on methods that assume independence between sites in the genome and thus a high level of LD. An extension to the fast clustering method proposed which enables conditioning on a pre-existing phylogeny and thus is suitable for more clonal datasets is also discussed.

2.1.1.2 *Distance-based methods*

Pairwise distance-based approaches have gained appeal with the ever-increasing size of genome sequence data sets. The speed of some modern distance-based methods has allowed for the clustering of otherwise unfeasibly large data sets. A notable example was the clustering of all 54,118 NCBI RefSeq genomes [388]. The lack of assumptions on the type of population structure can be a further advantage of distance-based methods. Discriminant analysis of principal components (DAPC) [250] takes advantage of the speed of principal component analysis to reduce the dimensionality of large sequencing data sets before using discriminant analysis [280] to choose a clustering.

While DAPC is considerably faster than model-based approaches, it still relies on a genotype matrix that can require considerable time to compute. With large genomes and high sequencing depth the computation required to produce a consensus sequence alignment or call single nucleotide polymorphisms (SNPs) can become prohibitive when many samples are involved. By adapting the MinHash dimensionality reduction technique [239] the Mash algorithm is able to avoid the alignment stage and allows for the rapid calculation of a distance matrix [388]. This approach has since been extended in the PopPUNK algorithm to model both accessory and core genome distance separately allowing for a more accurate definition of clusters from bacterial genomes [290]. The algorithm can exploit the network properties of the resulting distance matrix to refine strain definitions and can also incorporate additional isolates to an already existing set of clusters.

2.1.3 *Model-based methods*

Whilst distance-based approaches are generally faster, they are not as readily interpretable and typically do not provide group membership probabilities making distinguishing weak separation between clusters and strong population structure difficult [32]. Model-based methods typically indicate the probability that an individual belongs to a certain cluster, allow for different models to be compared, and can provide estimates of uncertainty for the inferred parameters. Generally, either Bayesian or maximum-likelihood based approaches are widely used for inference in population structure models.

A commonly considered model of population structure assumes that there is a fixed number of uncorrelated underlying populations, K . An individual is then assumed to originate either from a single population, the no-admixture model, or to carry alleles from multiple populations, which corresponds to the admixture model [407]. Solutions to this problem include STRUCTURE [158, 407], BAPS [87, 88, 90], Admixture [4], fastStructure [415], sNMF [180] and the method of Anderson and Thompson [8]. More recently, methods that combine model-based techniques with an initial faster distance-based clustering have been proposed including snapclust [32] and hierBAPS [70]. Admixture, fastStructure and sNMF solve for the admixture model and are not considered here. A common theme of most clustering methods is that they require the underlying number of clusters, K , to be provided. In practice, the methods are usually run over a range of K and the ‘best fitting’ model is selected [3, 155]. An alternative Bayesian solution is to put a prior on the number of clusters, usually using a Dirichlet process mixture model (DPM) in an attempt to infer K as part of the model [234, 393]. Whilst this provides a natural method for inferring an appropriate K , the approach is very com-

putationally expensive and does not scale to large numbers of loci and individuals. The greedy stochastic optimisation approach in hierBAPS [70] places a discrete uniform prior on K up to a supplied maximum value and is able to scale to larger datasets. However, as datasets become very large, inferring K using model-based clustering approaches has so far been infeasible. An example of a dataset that is too large for current model-based methods is the set of HIV-1 pol genes, which are routinely generated in clinical settings for the purpose of identifying resistance to antivirals. Despite the relatively short length of this region (typically 1200-1500 base pairs long), the very large number of sequences - in excess of 100,000 in public databases alone - makes choosing an appropriate value for K very challenging.

2.1.4 *Dirichlet process mixture model*

A Dirichlet Process Mixture model (DPM) assumes a joint prior over the cluster allocation vector p , which indicates which cluster each individual is allocated to as well as the overall number of clusters K [234, 393]. More specifically the prior is assumed to follow a Dirichlet process where the prior joint probability of the allocation vector p and the number of clusters K is given by

$$f(p, K | \alpha, n) = \alpha^K \frac{\prod_{i=1}^K (n_i - 1)}{\prod_{i=1}^n (\alpha + i - 1)}$$

The concentration parameter α determines the degree to which individuals group together in clusters. A smaller α value leads to larger clusters. However, it has previously been shown that the choice of α has little impact on the results when the number of loci is large [233]. In addition to using a Dirichlet process prior on the allocation vector and number of clusters K we assume a separate Dirichlet prior distribution on the underlying allele frequencies within each cluster parameterised by β . Thus, we draw alleles frequencies from the model by first sampling the allocation vector and the number of clusters K and then sampling the allele count data D from a Multinomial-Dirichlet distribution. The marginal distribution of counts in each cluster component is given in Section 2.2.3. Unfortunately, methods that rely on a fully Bayesian inference approach using MCMC such as structurama are not feasible for very large datasets [233, 234] with current faster approximations such as hierBAPS also failing to scale to some real world datasets (see Section 2.3.3). Other approximate inference methods for DPM models include variational inference [35], fast search [238], sequential updating and greedy search (SUGS) [94, 500] and expectation propagation [359]. Whilst these methods offer substantial speed improvements over a full MCMC based approach none of them are able to take advantage of the phylogenetic structure present in bacterial genomic datasets.

2.1.5 Bayesian hierarchical clustering

The Bayesian hierarchical clustering (BHC) algorithm [217] presents an alternative approach that approximates a Dirichlet Process Mixture model whilst guaranteeing a solution in polynomial time. The method performs a version of agglomerative bottom up clustering using a Dirichlet process to account for uncertainty in the data and Bayesian model selection to decide which clusters should be merged at each step. Whilst the approach has been shown to succeed in clustering text documents, microarray data and even electrical demand profiles [172, 217, 436], it has yet to be successfully applied to the problem of identifying population structure. A reason for this is that the hyperparameters, β , of the model defining each component of the mixture are set to be proportional to the counts of each allele at each locus in the complete dataset. In practice setting the hyperparameters to be proportional to the entire dataset leads to an overpartitioning of population genetic data whereby the algorithm identifies a very high number of clusters (see Results). A common alternative prior in population structure studies is to use a symmetric Dirichlet prior [407]. However, at the lowest levels of the hierarchy a symmetric prior will lead to different clustering combinations having the same likelihood. This greatly reduces the ability of the algorithm to accurately cluster sequences at the lower levels. In contrast, the hierBAPS algorithm relies on a fast initial clustering using complete-linkage agglomerative clustering to provide an initial partition of the dataset.

2.1.6 Bayesian analysis of population structure (BAPS) and hierBAPS

Similar to other approaches [407], BAPS assumes that alleles are drawn independently from a multinomial distribution with a Dirichlet prior. However, unlike STRUCTURE, which uses Gibbs sampling to estimate the posterior distribution, BAPS attempts to find the partition of the data S that maximises the posterior probability of an allocation over all other possible allocations [87, 89, 90]. The hierBAPS, algorithm extends the BAPS algorithm allowing it to be applied recursively to identify population structure at different resolutions [70]. A partition S is defined as the allocation of each sequence to one of K possible clusters. The maximum possible value of K is given in the hierBAPS algorithm. Here $\mathcal{S}$ indicates the set of all possible partitions with up to $K_{\max}$ clusters. The hierBAPS algorithm attempts to choose S to maximise

$$P(S|\text{data}) = \frac{P(\text{data}|S)P(S)}{\sum_{S \in \mathcal{S}} P(\text{data}|S)P(S)}$$

where $P(\text{data}|S)$ is the marginal likelihood of having the allele frequency parameters analytically integrated out leading to

$$P(\text{data}|S) = \prod_{i=1}^K \prod_{j=1}^{N_l} \left(\frac{\Gamma(\sum_l \alpha_{ijl})}{\Gamma(\sum_l \alpha_{ijl} + n_{ijl})} \prod_{l=1}^{N_{A(j)}} \frac{\Gamma(\alpha_{ijl} + n_{ijl})}{\Gamma(\alpha_{ijl})} \right)$$

where n_{ijl} is the count for allele l at locus j in cluster i and α_{ijl} is the corresponding hyper-parameter for the Dirichlet prior. The BAPS algorithm attempts to find the partition S that maximises the posterior probability using a greedy stochastic search approach. A discretised uniform distribution of the cluster size K ($K = 1, \dots, K_{\max}$) is used in hierBAPS to provide the prior probability of each partition $P(S)$. The Dirichlet hyperparameters are set at $\frac{1}{N_{A(j)}}$ where $N_{A(j)}$ is the number of distinct alleles at locus j .

Previously, hierBAPS was only available as a MATLAB binary but has since been implemented in R, improving the usability and transparency of the algorithm [484]. The R version also includes the ability to make use of multiple cpu cores as well as new plotting functions that both improve the speed of analysis and the interpretability of the results.

2.2 METHODS

2.2.1 Fast hierarchical Bayesian analysis of population structure

Though methods such as hierBAPS and snapclust have allowed for analyses to scale to much larger datasets than previous approaches, they remain in practice inapplicable to the currently emerging datasets comprising tens of thousands or even hundreds of thousands of sequences, where large numbers of underlying clusters may be present. Here, through incorporating ideas from both BHC and hierBAPS, we describe an efficient inference solution to the no-admixture model for very large datasets when the underlying number of clusters may be in the tens or hundreds. By taking advantage of a fast initial clustering approach similar to that used in hierBAPS we are able to place a symmetric or BAPS prior on the mixture components, enabling the BHC algorithm to distinguish between different combinations of clusters.

2.2.2 Initial clustering

Similar to hierBAPS, but unlike the original BHC algorithm, fastbaps begins by generating a fast initial clustering of the data using a value for K much larger than expected (K_{init}). This has the advantage of both reducing the complexity of the Bayesian agglomerative stage whilst allowing for symmetric priors to be used. Starting with a multiple sequence alignment in FASTA format, we first generate a pairwise single nucleotide polymorphism (SNP) distance matrix before clustering using Ward's agglomerative clustering [371]. This hierarchy is

cut to generate K_{init} clusters. The hierarchy is also used to estimate the hyperparameters β of the individual mixture component priors. Very large datasets make the calculation of the initial pairwise distance matrix prohibitive. To counter this limitation, for datasets with more than 10,000 individuals we produce the initial clustering by first performing a Principal Component Analysis. Clustering is then performed using a fast hierarchical algorithm such as *genie* or the memory efficient Ward algorithm in the *fastcluster* package with the first 50 principal components given as input [183, 370]. This removes the need to calculate the complete distance matrix. The results were found to be robust to the number of principal components used with this approach found to provide near identical results to the distance matrix approach on the smaller datasets. This is expected in general as the PCA step is also only used at the lower levels of the clustering hierarchy and thus its impact is minimal on the final clustering solution.

2.2.3 Bayesian hierarchical clustering

Given a collection of small clusters and their respective hierarchies T_i where $i \in \{1..K_{\text{init}}\}$, the BHC algorithm of Heller and Ghahramani [217] proceeds in a similar fashion to traditional agglomerative clustering. In the place of a distance metric, Bayesian hypothesis testing is conducted at each level to decide which clusters to merge. Let $\mathcal{D} = x^{(1)}, \dots, x^{(n)}$ describe the entire dataset of n samples and $\mathcal{D}_i$ the set of points at the leaves of the subtree T_i . We initialise the algorithm with the small subtrees obtained using the fast clustering step as shown in Figure 1. At each subsequent stage we then consider merging all pairs of existing trees. If trees T_i and T_j are merged into a combined tree T_k then the set of points corresponding to the new tree is $\mathcal{D}_k = \mathcal{D}_i \cup \mathcal{D}_j$. At each merge we compare two hypotheses. The merged hypothesis, which we denote $\mathcal{H}_1^k$, is that all data in $\mathcal{D}_k$ were generated identically and independently from the same probabilistic model $p(x|\theta)$ with unknown parameters θ . In our case this model is a multinomial, with parameters $\theta = (n, \phi)$. We also specify a Dirichlet prior over the parameters ϕ with hyperparameters β . Thus, we can write the probability of the data $\mathcal{D}_k$ under $\mathcal{H}_1^k$ as

$$p(\mathcal{D}_k | \mathcal{H}_1^k) = \int p(\mathcal{D}_k | \theta) p(\theta | \beta) d\theta$$

Assuming a multinomial-Dirichlet distribution, this integral is tractable and we can rewrite the above equation as

$$p(\mathcal{D}_k | \mathcal{H}_1^k) = \prod_{j=1}^{N_L} \frac{\Gamma(\sum_l \beta_j)}{\Gamma(\sum_l (\beta_j + n_{jl}))} \prod_{l=1}^{N_{A(j)}} \frac{\Gamma(\beta_j + n_{jl})}{\Gamma(\beta_j)}$$

Where N_L is the number of loci, $N_{A_{(j)}}$ is the number of possible alleles at loci j and n_{jl} is the number of copies of allele l at locus j . β_j is the corresponding Dirichlet prior hyperparameter.

The alternative hypothesis to merging two clusters ($\mathcal{H}_1^k$) is that the data $\mathcal{D}_k$ have two or more clusters. Summing over all the possible partitions of $\mathcal{D}_k$ into two or more groups is in practice intractable. However, by only considering partitions that are consistent with the subtrees T_i and T_j , we can use recursion to quickly compute this sum. Under the restriction of remaining consistent with the subtrees, the probability of the data under the alternative hypothesis $\mathcal{H}_2^k$, is just the product over the subtrees $p(\mathcal{D}_k|\mathcal{H}_2^k) = p(\mathcal{D}_i|T_i)p(\mathcal{D}_j|T_j)$ (see [Figure 1](#)) where the probability of the data given the tree is given by

$$p(\mathcal{D}_k|T_k) = \pi_k p(\mathcal{D}_k|\mathcal{H}_1^k) + (1 - \pi_k) p(\mathcal{D}_i|T_i) p(\mathcal{D}_j|T_j)$$

Here, the first term for each initial cluster is defined recursively using the fast clustering based hierarchy with initial prior for the merged hypothesis set to $\pi_0 = 1$. The prior for the merged hypothesis $\pi_k := p(\mathcal{H}_1^k)$, can then be computed bottom-up as described in Heller and Ghahramani, [217] and is a deterministic function of the concentration parameter of the DPM. In practice, for the size of datasets we are considering, the actual value of α makes very little difference as the prior π_k is dominated by the factorial on the number of points within a cluster. Similar to Savage *et al.*, [436], we fix the value of α , the concentration parameter of the DPM, rather than keeping it as a user-tunable hyperparameter in the resulting clustering algorithm.

Finally, given the probabilities of each hypothesis described above we can calculate the posterior probability of the merged hypothesis $r_k := p(\mathcal{H}_1^k|\mathcal{D}_k)$ using Bayes' rule

$$r_k = \frac{\pi_k p(\mathcal{D}_k|\mathcal{H}_1^k)}{\pi_k p(\mathcal{D}_k|\mathcal{H}_1^k) + (1 - \pi_k) p(\mathcal{D}_i|T_i) p(\mathcal{D}_j|T_j)}$$

We use this quantity to greedily decide which two subtrees to merge. After generating a final hierarchy we then use the same quantity to identify which merges were justified and cut the tree when $r_k < 0.5$.

2.2.4 Hyperparameter selection

As described in Heller and Ghahramani [217], for any given set of hyperparameters, the root node of the hierarchy approximates the probability of the data given that setting of the hyperparameters. By leveraging this, we use the marginal likelihood of the root node $p(\mathcal{D}|T)$ to perform model comparisons for different settings of the hyperparameters. We found that different settings of the hyperparameter α had a negligible impact on the final clustering due to the dominance of the factorial component of π_k , which provided a strong motivation to fix the hyperparameter at the value $\alpha = 1$. Similar to Heller

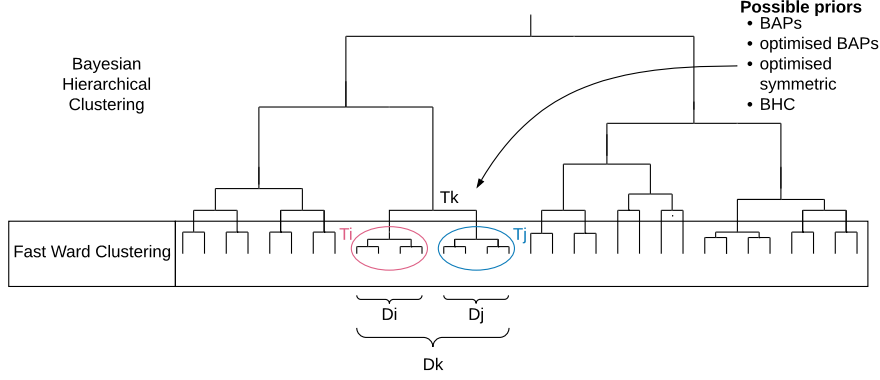


Figure 1: A diagram outlining the fastbaps algorithm. An initial clustering is performed by default using agglomerative clustering with the Ward linkage method. Subsequently the merge with the highest posterior probability is chosen. The final clustering is then determined by only accepting merges with a posterior probability > 0.5

and Ghahramani [217] and Savage *et al.*, [436], we use golden section search as provided in R’s ‘optimise’ function to select the β parameters and thus the variance of the Dirichlet prior. Unlike previous approaches, we did not set the β parameters to be proportional to the discrete total counts for the entire dataset. Instead we used either a symmetric prior, similar to Pritchard *et al.*, [407] or the non-informative prior $\beta_j \propto \frac{1}{N_{A(j)}}$ as used in hierBAPS and suggested in Anderson and Thompson [8]. These priors are referred to as optimised symmetric, and optimised BAPS priors hereafter. These priors were found to generally outperform the prior used in the original BHC algorithm. To increase the efficiency in optimising the variance of the Dirichlet prior, we relied on an initial hierarchy produced using Ward’s agglomerative clustering. The marginal likelihood of the root node was then used to select an appropriate scaling for the β hyperparameters. We subsequently refer to the combination of fast cluster initialisation, Bayesian hierarchical clustering and prior hyperparameter optimisation using a symmetric or BAPS based prior as ‘fastbaps’.

2.2.5 Conditioning on a pre-computed hierarchy

Given a pre-computed hierarchy or phylogeny, we can use the recursion method described previously to decide when merging subclades of the tree is justified according to the DPM model. This provides a parametric model based alternative to previous tree partition algorithms such as Cluster Picker [413], also relaxing the computationally intensive requirement for bootstrap replicates of the tree being partitioned. It allows us to identify a partition of the tree into clades that maximises the marginal likelihood of a DPM model given

a hierarchy. As the hierarchy is pre-calculated, the approach is highly efficient and has a linear computational complexity in the number of nodes in the hierarchy. By leveraging the sparse matrix data structures in fastbaps and coupling them with the efficient implementation of Ward's hierarchical clustering in the fastcluster package [370], we are able to generate clusterings of comparable quality to the full fastbaps algorithm very quickly (see Figure 3 and Figure 5).

2.2.6 Cluster stability using the Bootstrap

In order to determine the sensitivity of the resulting clusters to the alignment, we implemented a simple bootstrap procedure. For each bootstrap replicate, the loci are sampled with replacement and the fastbaps clustering algorithm is run. A binary similarity matrix is then produced as described in Strehl and Ghosh [469] where for each pair of sequences (i, j) , the (i, j) th entry in the matrix is 1 if the i th and j th sequences are clustered together and 0 otherwise. As this matrix is invariant to label switching [421] and its dimension is invariant to the underlying number of clusters, the sum of the resulting matrices can be used to provide an indication of the stability of a given clustering of two isolates. This can be plotted in a heatmap alongside the dendrogram produced using the full algorithm to illustrate the robustness of the final clustering.

2.2.7 Implementation

The fastbaps method is implemented using R and C++ and is available as an R package under the MIT open source licence at <https://github.com/gtonkinhill/fastbaps>. It can be run on Unix, Mac, or Windows operating systems and takes a multiple sequence alignment as input. The results can easily be parsed in R allowing for the generation of informative plots and further processing. The sparse matrix data structure generated by the package can also be used in other analysis types such as Principal Component Analysis (PCA) or k-means.

2.2.8 Simulation and evaluation

To compare the results of different algorithms, population structure was simulated using scrm as part of the Coala R package [462, 463]. The number of underlying populations, recombination rate and migration rate were varied as described in Table 1. Each parameter combination was run for 3 replicates providing 90 test datasets for analysis. Three clustering algorithms were considered in addition to the fastbaps approach. Snapclust and hierBAPS are available as R packages and were compared with fastbaps along with the fully Bayesian

Structure algorithm [32, 70, 407, 484]. Snapclust was run in parallel with the number of underlying clusters, K , varied between 2 and 30. The best fitting model was then chosen using either the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC). HierBAPS was run with 50 initial clusters. Finally, Structure was run with the no admixture model with 100,000 burn in iterations and 500,000 sampled iterations with 3 separate starting conditions. The run with highest likelihood was kept. The underlying number of clusters was fixed to the simulated K when running Structure to obtain a result in a reasonable amount of time, which nevertheless gives the method an unrealistic advantage against the alternatives which are fast enough to identify K . We implement a number of strategies to validate our clustering approach and demonstrate its ability to generate biologically informative clusters. That is clusters that identify broad population structure that allow us to divide large datasets into smaller related subsets for further investigation or to use as a covariate in phenotype association studies. We show that similar to previous programs our clustering approach is able to retrieve the true clusters from simulated data whilst maintaining a considerable advantage in terms of speed. We compare our algorithm to alternative approaches that solve similar problems such as phylogeny reconstruction using Fasttree and dimension reduction using UMAP. We show that our results are consistent with these algorithms whilst using a parametric model of population structure which provides additional insight over clustering on the outputs of either Fasttree or UMAP alone. In addition, using the outputs of Fasttree we compare our algorithm with other competing approaches over a wide range of real world datasets using a distance function which accounts for both within-cluster and between-cluster similarity, penalising solutions that achieve a high within-cluster similarity by producing a large number of very tightly connected clusters. Finally, to investigate our results on the large HIV dataset we consider the concordance of our inferred clusters with the geography and sub-type of the HIV pol genes considered. In summary, our comparisons aimed to provide a wide range of biologically relevant scenarios where the statistical and computational choices made in the different methods may result in biologically relevant variation in the data partitions and thus enable an assessment of the general applicability and robustness of a particular method. To compare the simulated populations with those inferred by the different clustering algorithms, we used the Fowlkes-Mallows index [176]. This accounts for both false positives and false negatives in determining the similarity between two clusterings and is defined as

$$FM = \sqrt{\frac{TP}{TP + FP} \frac{TP}{TP + FN}}$$

Parameter	Value(s)
Number of replicates	3
Number of demes (clusters)	5, 10, 15, 20, 25
Deme level population size	100
Sequence length	100,000
Mutation rate	0.01
Recombination rate	10, 100
Migration rate	0.001, 0.01, 0.02
Proportion of population sampled	0.1

Table 1: Parameters used to simulate genetic population structure using `scrm` [463]

Here TP is the number of pairs of individuals that appear together in both clusterings, FP is the number that appear together in the inferred clusters but not the simulated and FN are those pairs that originated from the same population in the simulated data but are separated in the inferred clusters. Thus a clustering that perfectly matches the simulation will receive a Fowlkes-Mallows index of 1.

To compare with real data, six bacterial and two viral datasets were considered. The sets were chosen to cover a diverse range of pathogen species. The bacterial sets included the enteric bacteria *Escherichia coli*, the gram negative respiratory pathogen *Haemophilus influenzae*, as well as the firmicutes *Streptococcus pneumoniae*, *Listeria monocytogenes* and *Staphylococcus aureus* [1, 71, 270, 275, 291]. The two viral datasets comprised a subset of sequences from the recent Ebola outbreak as well as a global dataset of over 110,000 HIV partial pol genes (positions 1-1497) obtained from the Los Alamos HIV Database [141, 316], with associated metadata on subtype and country of sampling. A summary of the size of each dataset is given in Table 2. As there is no gold standard truth set for the real datasets we instead chose to compare the inferred clusters with a phylogeny built using Fast-tree v2.1.10 [406]. Here, we counted the number of pairs of isolates within each cluster that were inconsistent with the phylogeny in that there was an isolate belonging to another cluster within the clade representing their most recent common ancestor. These were considered ‘false positives’. This count was then divided by the total number of possible pairs within each cluster to give an indication of the error rate of the clustering (assuming the phylogeny to be correct). As the number of possible pairs increases with cluster size, this approach appropriately penalises clustering solutions with too many clusters.

Dataset	Number of sequences	Number of sites	Reference
<i>E. coli</i>	1508	241750	Kallonen <i>et al.</i> 2017
<i>H. influenzae</i>	75	113605	Koelman <i>et al.</i> 2017
<i>L. monocytogenes</i>	128	150759	Kremer <i>et al.</i> 2017
<i>N. meningitidis</i>	882	87730	Lees <i>et al.</i> 2017
<i>S. aureus</i>	284	50104	Aanensen <i>et al.</i> 2016
<i>S. pneumoniae</i>	3156	392524	Chewapreecha <i>et al.</i> 2014
<i>Ebola</i>	1610	2279	Dudas <i>et al.</i> 2017
HIV	118091	1497	http://www.hiv.lanl.gov

Table 2: The datasets used in the comparison of different clustering methods. The HIV and pneumococcal datasets were only compared with k-means and fastbaps as they are too large for other model based methods.

2.3 RESULTS

2.3.1 Simulations show that fastbaps cluster accuracy is the same as previous methods

In order to compare with different methods we made use of simulations where the true clustering solution was known. The Structure, hierBAPS, snapclust and fastbaps algorithms all achieved similar levels of accuracy on the simulated datasets (Figure 2). As expected, an increase in the migration rate between the simulated demes led to a subsequent decrease in the ability of the methods to accurately identify the underlying population structure. Additionally, a lower recombination rate led to poorer results, which is probably due to the algorithms correctly identifying additional phylogenetic structure within each cluster. Snapclust using AIC to select K as well as the fastbaps algorithm with the optimised BAPS prior achieved the best results over the simulated datasets. hierBAPS achieved a similar level of accuracy to the fastbaps algorithms using both the original (unscaled) BAPS prior as well as the optimised symmetric prior. Structure and Snapclust with BIC performed worst over the simulations. This was despite the MCMC chains of each Structure run showing good convergence characteristics. A closer investigation of the Structure results indicated that it tended to allocate empty clusters despite being given the correct K. This suggests that both Snapclust with BIC and Structure tend to be overly conservative in estimating the number of underlying clusters K.

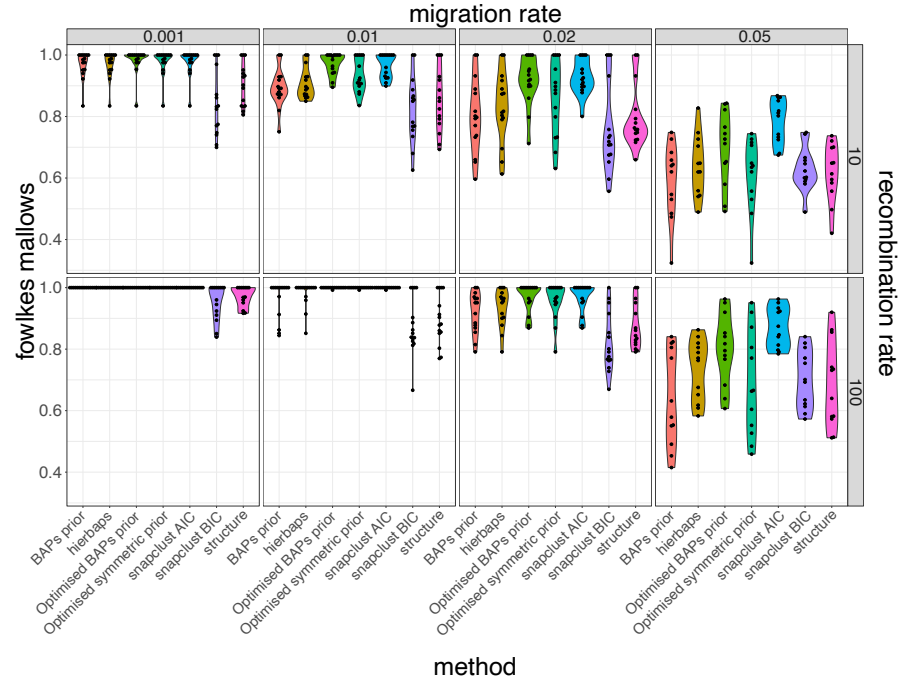


Figure 2: Violin plot indicating the Fowlkes-Mallows distance between the simulated clusters and those inferred by the different algorithms. Results for two distinct recombination rates are shown (10 and 100) as well as four different migration rates. The underlying number of clusters varied between 5-25. The plot indicates that as the migration rate increases the accuracy of the algorithm falls. Additionally, a lower recombination rate leads to lower accuracy, which is likely due to the algorithms identifying additional population structure within each simulated deme.

2.3.2 *Fastbaps is efficient and accurate on a diverse range of viral and bacterial datasets*

To investigate the ability of each algorithm to detect population structure in real datasets, we made use of six additional bacterial and viral datasets of varying sizes and complexity (Table 2). Figure 3 indicates the resulting clusters inferred by the most promising algorithms from the simulation analysis on both a *Neisseria meningitidis* and *Haemophilus influenzae* data set. Here, we also provide comparisons with the prior from the original BHC algorithm and a partition of a Fasttree phylogeny using fastbaps. On the *H. influenzae* dataset, fastbaps using the optimised BAPS prior provided a clustering that was the most consistent with the phylogeny. Both the hierBAPS and snapclust solutions included polyphyletic clusters whilst the Bayesian Hierarchical Clustering (BHC) population mean based prior gave a similar result to the optimised BAPS solution. It should be noted that the initial fast clustering step was the same for both the BHC and optimised BAPS clusterings. The BHC prior failed to perform adequately on the large *N. meningitidis* dataset. Here, the BHC prior led to a highly partitioned solution with over 80 very small clusters. Whilst such a high number of clusters may be valid it does not aid in the biological interpretation of the data set, where we are generally interested in broad population structure that allows us to divide large collections into smaller subsets that can be more closely investigated or controlled for in phenotype association studies. Conversely, the optimised BAPS prior led to a solution of similar quality to both hierBAPS and snapclust. Uncertainty in the Fasttree phylogenies was considered by comparing them with a bootstrapped phylogeny built using IQTREE v1.6.5 with the results found to be very similar. In both of these datasets the partition of the phylogeny using fastbaps provided a solution similar to that found using the optimised BAPS prior indicating it is an appropriate choice if a user's goal is to simply partition a pre-calculated phylogeny. In addition, the uncertainty in the fastbaps clusters could be explored using the bootstrap procedure described in Section 2.2.6.

Figure 4, summarises the results of running the algorithms on all these data sets. The figure indicates the number of pairs of isolates that are clustered together using the population structure algorithms but that are inconsistent with the phylogeny. Assuming the Fasttree phylogeny is accurate, this provides a measurement of error for each of the runs. The fastbaps algorithm with either the optimised symmetric or optimised BAPS priors achieves a superior clustering on five out of six of the datasets. On the *N. meningitidis* dataset the snapclust algorithm using the BIC model selection criteria outperforms the other methods. The BIC model selection technique more heavily penalises solutions with a larger number of clusters and was in the

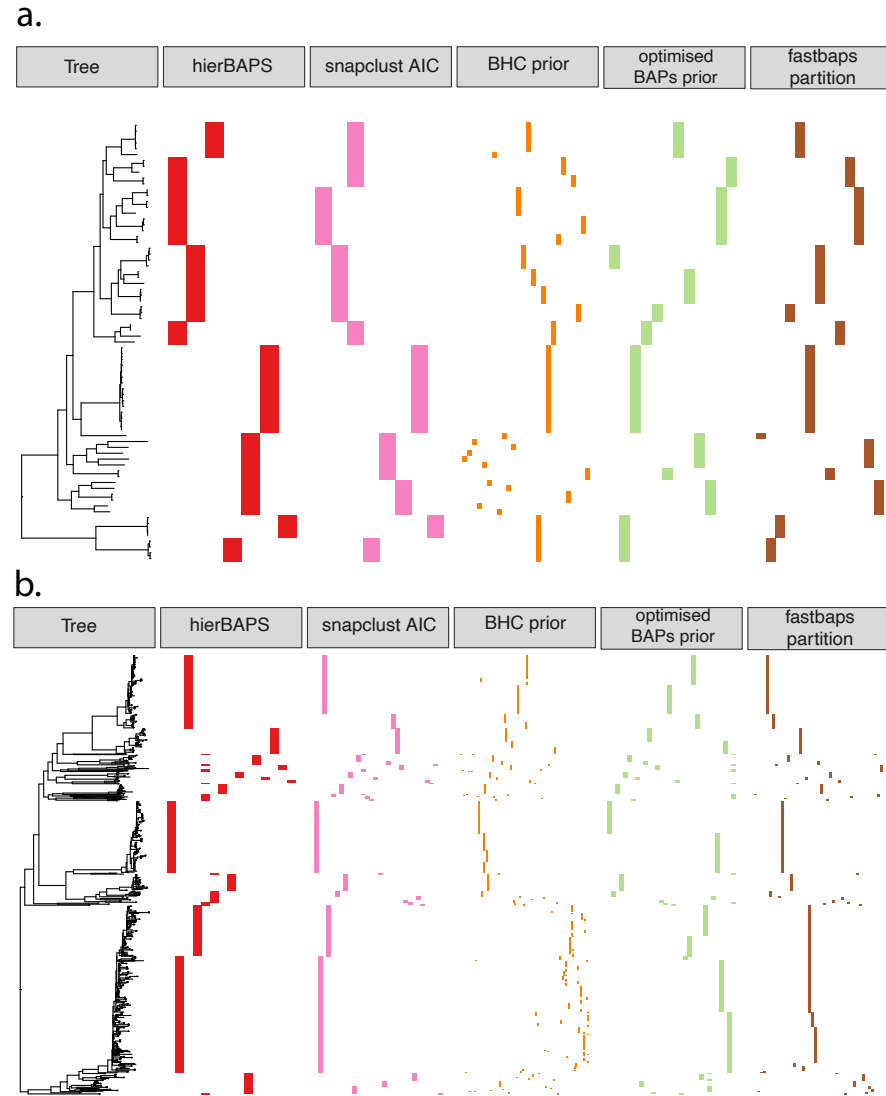


Figure 3: The resulting clustering inferred by hierBAPS, snapclust and fastbaps with both the original BHC prior and the optimised BAPS prior on both *Staphylococcus aureus* (a) and *Neisseria meningitidis* (b). The clusters are shown in comparison to a phylogeny generated using Fasttree. The final clustering indicates a partition of the Fasttree phylogeny using the fastbaps algorithm and thus is constrained to be consistent with the phylogeny.

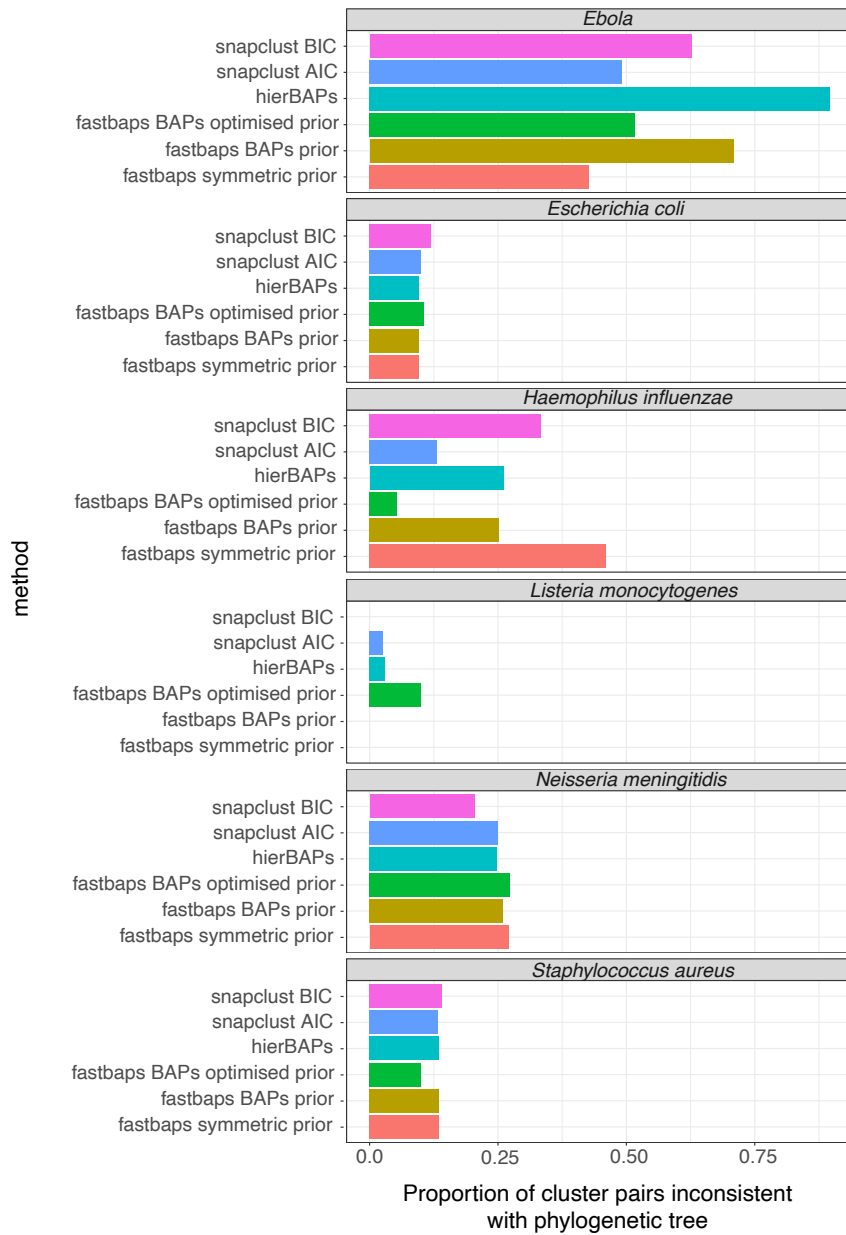


Figure 4: The proportion of isolate pairs that appear in the same inferred cluster but have isolates from a separate cluster in the clade represented by their most recent common ancestor in the Fasttree phylogeny. This provides a indication of the error of each algorithm assuming the phylogeny generated by Fasttree is correct. The fastbaps algorithm using either the optimised BAPS or symmetric prior outperforms the other methods on 5 out of 6 datasets.

Dataset	N	L	fastbaps BAPS prior	fastbaps optimised BAPS prior	fastbaps optimised symmetric prior	rhierbaps	snapclust	snapclust single K
<i>Ebola</i>	1610	2279	9.93	16.52	16.24	11556.36	686.18	45.8
<i>E. coli</i>	1508	241750	282.97	1210.3	1155.3	515913.93	498433.4	15007.32
<i>H. influenzae</i>	75	113605	9.36	30.85	33.4	9018.08	13472.79	277.64
<i>HIV</i>	118091	1497	39718.89	55012.17	55067.16	NA	NA	NA
<i>L. monocytogenes</i>	128	150759	22.86	78.39	68.36	18787.03	29257.58	392.6
<i>N. meningitidis</i>	882	87730	77.71	278.62	283.58	110312.24	76361.75	1900.13
<i>S. aureus</i>	284	50104	12.18	39.82	41.26	6514.21	35322.14	212.02
<i>S. pneumoniae</i>	3156	392524	794.99	3822.11	3998.63	NA	NA	NA

Table 3: Total CPU time (seconds) of the different algorithms. Only fastbaps was able to run in under a week for the Pneumococcal and HIV datasets.

Dataset	N	L	fastbaps BAPS prior	fastbaps optimised BAPS prior	fastbaps optimised symmetric prior	rhierBAPS	snapclust	snapclust single K
<i>Ebola</i>	1610	2279	411.64	468.64	487.32	1102.51	519.71	498.44
<i>E. coli</i>	1508	241750	4290.98	4559.68	5178.42	15194.38	54714.39	45680.39
<i>H. influenzae</i>	75	113605	394.56	375.84	382.09	852.5	4148.99	2113.79
<i>HIV</i>	118091	1497	3465.45	6511.35	7409.47	NA	NA	NA
<i>L. monocytogenes</i>	128	150759	791.76	842.88	823.48	1400.69	5814.35	5194.14
<i>N. meningitidis</i>	882	87730	1439.4	1477.06	1382.16	4231.05	13779.11	10822.19
<i>S. aureus</i>	284	50104	384.85	387.04	396.44	742	3152.37	2423.9
<i>S. pneumoniae</i>	3156	392524	3915.3	5790.93	5710.96	NA	NA	NA

Table 4: Maximum memory usage of the different algorithms (Mb)

simulations shown to frequently result in too few clusters. It is unlikely that the other methods are over-partitioning these data relative to snapclust with BIC. Hence, the outlying result for *N. meningitidis* could rather be due to inaccuracies in the Fasttree phylogeny, since *N. meningitidis* has a high recombination rate.

The computational run times and maximum memory requirements of each of the algorithms are shown in Tables 3 & 4 respectively. The tables indicate that fastbaps often performs more than an order of magnitude faster than snapclust and hierBAPS which themselves significantly outperform Structure in computational efficiency [32, 282]. Part of this speed is due to the use of highly optimised sparse matrix libraries in fastbaps. As snapclust must be run for all values of K that are to be considered, its computational performance is directly tied to the maximum K chosen. Here, snapclust was run for a maximum K = 50 as this corresponded to a sensible upper bound for the *E. coli* dataset. Reducing the maximum K to use would improve the timing of the snapclust algorithm at the expense of failing to explore higher dimension solutions. However, Table 3 indicates that the fastbaps algorithm achieves a superior runtime compared with snapclust even for a single value of K, whilst automatically determining an appro-

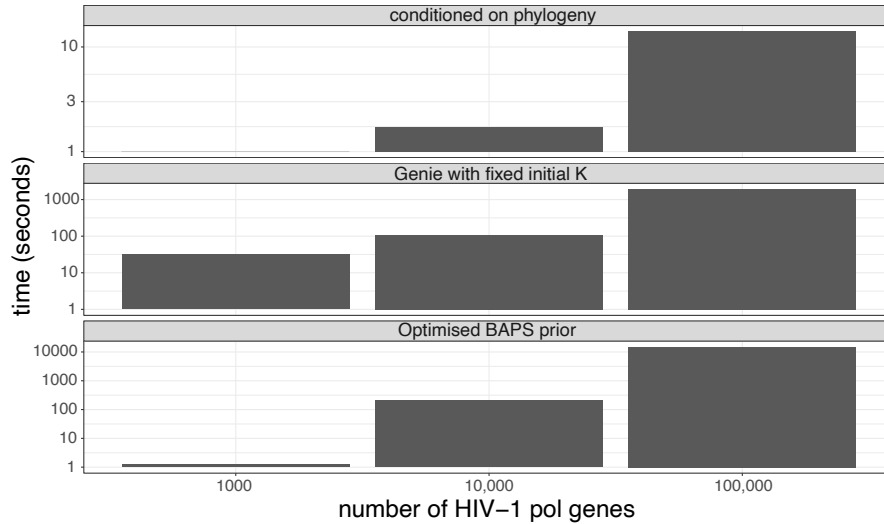


Figure 5: A bar chart indicating the time complexity of the different modes of running fastbaps. The phylogeny conditioned algorithm scales linearly with number of samples whilst the full mode scales quadratically. If the initial number of clusters is fixed and a sub-quadratic initial clustering algorithm is used the method scales sub-quadratically as indicated by the Genie mode.

priate number of clusters. Depending on the dataset, the prior optimisation step in fastbaps can take longer than running the complete algorithm. Thus if a short run-time is a necessity, running the method with a fixed prior is available as an option. Finally, if a hierarchy is already available, either through the use of a traditional phylogeny reconstruction algorithm such as Fasttree or via agglomerative clustering a partition using the fastbaps algorithm can be achieved with a complexity linear in the number of sequences. Fastbaps was able to partition a phylogeny of the full HIV dataset built using Fasttree [406] including reading in the data in 45.8 seconds. Figure 5 indicates the running time of the different modes of fastbaps excluding loading each data set. The phylogeny conditioned mode can be seen to scale linearly, whilst the full fastbaps mode scales quadratically with the number of samples. Genie is a sub-quadratic hierarchical clustering method and by fixing the number of initial clusters Figure 5 indicates we are able to achieve a sub-quadratic clustering algorithm. However, as the initial number of sub-clusters is fixed this mode would be most useful in cases where the user expected a large amount of redundancy in the data set.

2.3.3 *Fastbaps accurately clusters previously intractable viral and bacterial datasets containing thousands of samples*

To investigate the performance of fastbaps on very large datasets, we analysed a dataset of over 110,000 HIV partial pol genes downloaded

from the Los Alamos National Laboratories HIV Sequence Database (<http://www.hiv.lanl.gov>). The large number of sequences in this set presents a significant challenge to most model-based techniques as determining even the range of values of K is not straightforward. Additionally, we investigated the success of the algorithm on over 3,100 pneumococcal genomes from the Maela refugee camp in Thailand [71]. Whilst this dataset has fewer genomes, the large number of variable sites (284,194) makes this dataset very challenging for all current methods.

Figure 6 illustrates the resulting clusters inferred by fastbaps using the optimised BAPS prior on both the pneumococcal and HIV datasets. We made use of the UMAP dimensionality reduction method as implemented in the umap-learn python package to project the isolates onto a 2D plane [346]. UMAP is a non-linear dimensionality reduction technique that has been shown to perform well on large biological datasets including visualizing large genetic population datasets as well as in the analysis of single cell RNAseq data [26, 125]. As the other model based-methods considered here were unable to produce results on either of these datasets in under a week, we compared the resulting clusters to those inferred using k-means. We used both the K identified by fastbaps as well as the K found using the ‘elbow method’ [255] when running the k-means algorithm. Figure 6a indicates that fastbaps mostly corresponds with the groupings formed in the UMAP projection. The visualization also demonstrates the utility of investigating the variability of the level of clonality across the clusters, as the layout of some clusters is more scattered spatially, reflecting the varying influence of recombination on the relatedness of pneumococcal strains. Whilst the k-means algorithm using the fastbaps inferred K ($K = 79$), provides a similar result, the K inferred using the elbow method leads to many inferred clusters being separated across the projection (Figure 8a). This highlights the advantages of using model-based techniques when investigating genetic datasets.

The fastbaps algorithm provides a cleaner clustering of the HIV dataset as seen in Figure 6b. The large group of isolates in the centre of the UMAP projection has been assigned to one cluster. Conversely, k-means is unable to distinguish this group, and instead partitions it into many smaller overlapping clusters (Figure 8b). Similar to the pneumococcal dataset, when the lower value of K identified using the elbow method is used ($K = 21$), the resulting clusters are separated out over the UMAP projection.

As HIV is highly recombinant, comparing the clustering of such a large global dataset to a phylogeny is unlikely to be informative. Instead, to further investigate the quality of the fastbaps clustering we examined the HIV subtype composition (inferred using a variety of methods, and including circulating recombinant forms as well as

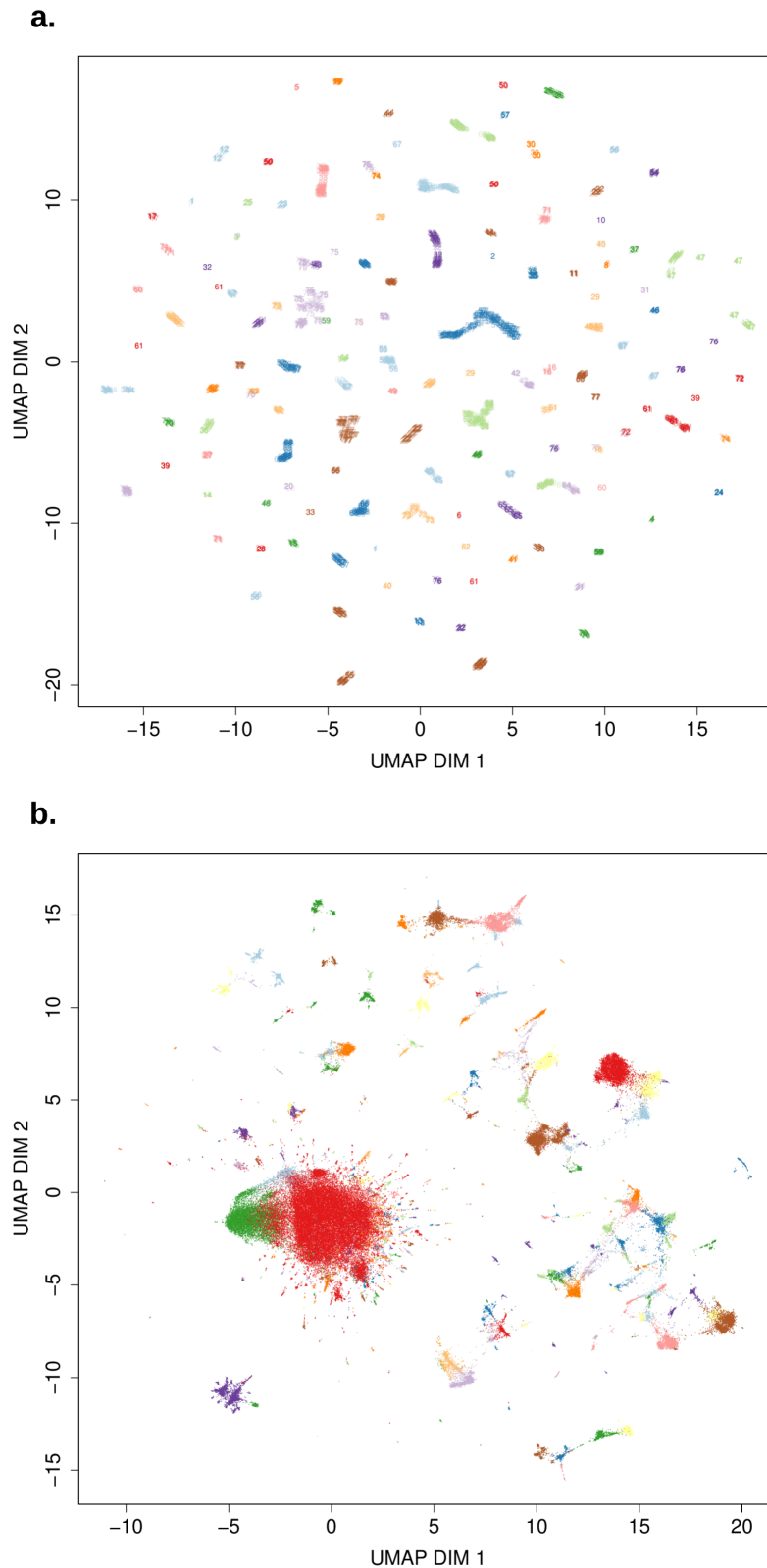


Figure 6: (a) A UMAP projection of over 3,100 pneumococcal genomes from Maela Thailand. The points are labeled and coloured by the cluster inferred using fastbaps with the optimised BAPS prior. In general the labeling appears to agree with the grouping observed in the UMAP projection. (b) Over 110,000 HIV-1 pol gene sequences from the Los Alamos public database. Sequences are coloured by their fastbaps inferred cluster. The large red cluster in the center contrasts with the noisy clustering observed using k-means (Figure 8b).

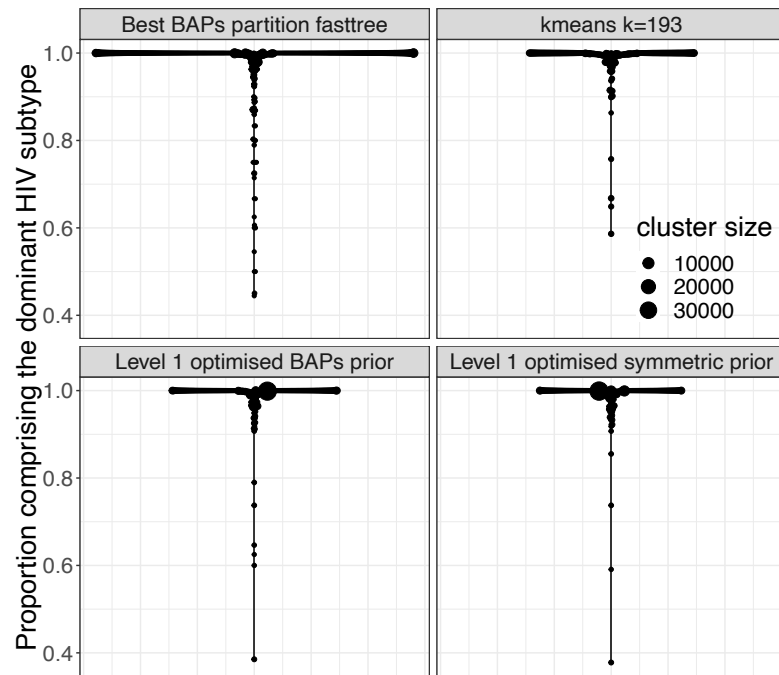
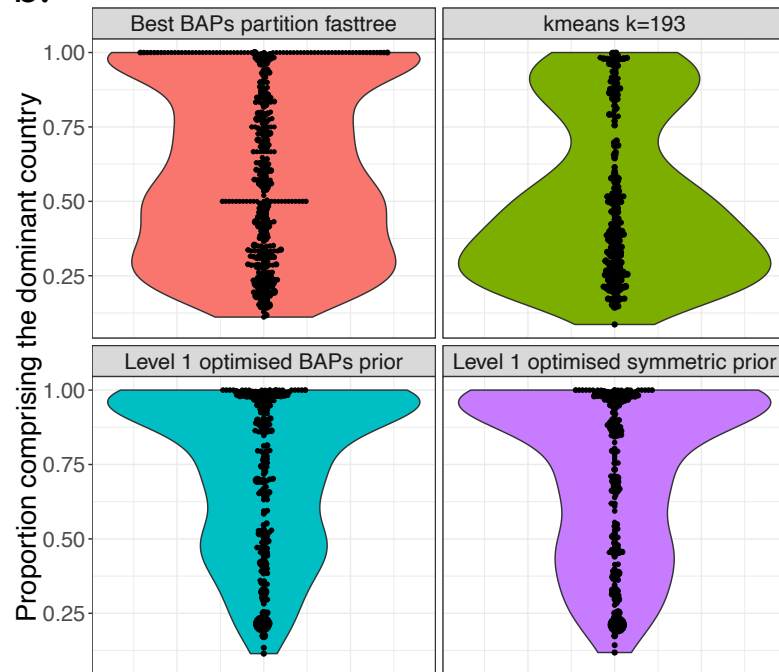
a.**b.**

Figure 7: (a) Violin plots indicating the proportion of each inferred cluster that make up its most dominant subtype. Cluster size is represented by the size of a point and the width of the violin plot is proportional to the total number of clusters inferred. The fastbaps algorithm with the optimised BAPS prior provides a clustering that is most consistent with the underlying HIV subtypes whilst not overly segmenting the dataset. (b)) Violin plots indicating the proportion of each inferred cluster that make up its most dominant country. Interestingly, the fastbaps approach is able to provide clusters that are more consistent with the underlying geography than alternative approaches such as k-means.

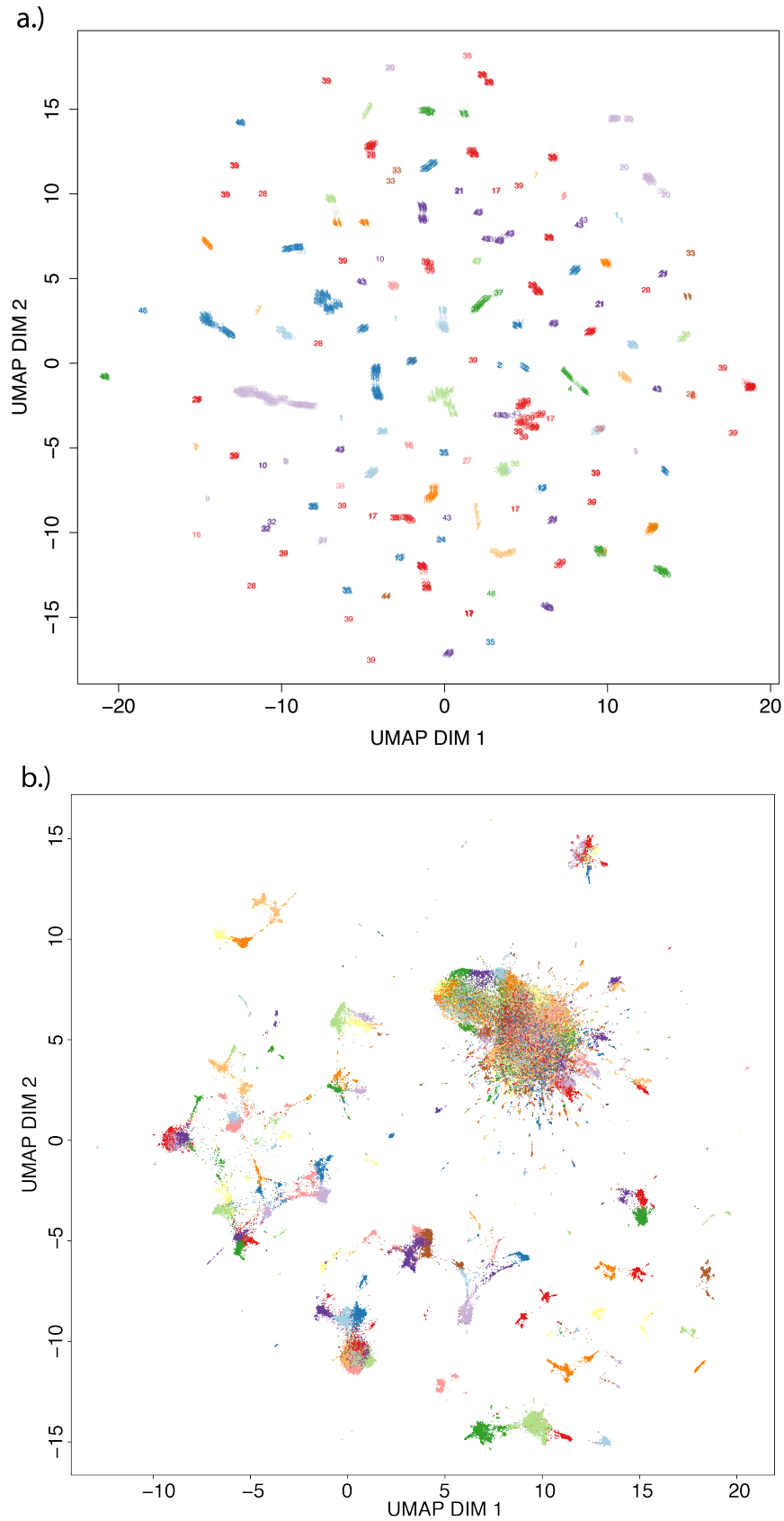


Figure 8: a) A UMAP plot of Pneumococcal isolates coloured and numbered by their inferred clustering using k-means with the number of underlying clusters found using the elbow method ($k = 48$). (b) A UMAP plot of HIV isolates coloured by their inferred clustering using k-means with the number of underlying clusters matching that found using fastbaps ($k = 193$)

non-recombinant subtypes) as well as the distribution of the countries of origin of the isolates found in each cluster. Figure 7 indicates the proportion of each cluster that was dominated by a single subtype or any of its recombinant forms. Thus a cluster that is composed entirely of one subtype will have a proportion of 1. The figure indicates that despite having the same number of clusters as k-means, the fastbaps algorithm with both the optimised BAPS and optimised symmetric prior identifies larger clusters of purely one subtype.

This suggests the algorithm is identifying legitimate clusters within the dataset without over-partitioning the data. A similar comparison was made to investigate the proportion of each cluster that was dominated by a single country. Figure 7 indicates that fastbaps provides clusters that are more consistent with the underlying geography where as the k-means is dominated by many clusters that comprise isolates from many countries. Fastbaps was also used to partition a phylogeny. The resulting clustering produced a larger number of clusters due to the restriction of the phylogeny but the resulting clusters were more consistent by subtype and geography than those inferred using k-means.

2.4 DISCUSSION

Finding clusters of related sequences present in a genetic alignment is a critical first step for many genetic and ecological analyses, allowing targeted sub-analysis and determination of the structure of the population when external classifications are not present. Model-based clustering methods are attractive due to their ability to hierarchically find highly specific clusters, estimate meaningful clustering parameters incorporating uncertainties, compare different model fits and produce probabilities of cluster assignment. Typical cohort sizes now range between thousands and tens of thousands of samples, growing even larger if combined with previous cohorts. However, the ability of methods to fit a clustering model to these alignments has not increased in step with these increases in alignment size [32, 484], with even modestly sized datasets requiring gene-by-gene or distance based approaches to determine clusters [250, 290, 537]. Model-based approaches have also proved impractical in surveillance settings, where the continuous addition of small numbers of new sequences would require refitting the entire model. By leveraging ideas from both the Bayesian Hierarchical Clustering [217] and the hierBAPS [70] algorithms, along with a fast sparse matrix based implementation, fastbaps enables efficient model-based clustering of large alignments that were previously infeasible to analyse using existing methods. Our approach enjoys all of the advantages inherent in model-based approaches and can produce comparable or higher quality clusters than previous methods. Additionally, our algorithm

is able to rapidly partition pre-computed phylogenies. This provides an attractive alternative approach for clustering genetic data when a phylogeny is available. The significant acceleration of inference provided by fastbaps enables the use of bootstrap replicates, allowing for the sensitivity of the resulting clusters to the alignment to be investigated in manner that has traditionally not been possible for model-based clustering methods on large alignments. We verified our new approach by comparing its performance with other algorithms on both simulated data and eight real datasets. We show that as well as a considerable increase in the speed of the algorithm, we are able to achieve comparable accuracy, often outperforming previous model-based approaches. The speed of our new approach allowed us to cluster a large HIV sequence dataset containing over 100,000 sequences. The resulting clusters have higher concordance with HIV subtypes and geography than an alternative approach using k-means, whilst providing a principled method for selecting the underlying number of clusters; a major limitation of k-means. Whilst our method enables the clustering of very large alignments, its complexity is still tied to the initial hierarchy generation and thus is $O(n^2)$ or possibly $O(n \log(n))$ if single linkage hierarchical clustering is used. After the initial hierarchy is generated, the remaining Bayesian hierarchical clustering is of the order $O(lm^2)$ where m is the number of clusters generated in the first stage and l is the number of variable sites. Thus, in the future, as alignments begin to comprise tens of millions of sequences, additional improvements will need to be explored. Possible options include exploring the randomised extension to the BHC algorithm [256] or other machine learning based approaches such as those based on matrix decomposition [527]. Fastbaps therefore offers two new ways to rapidly find high-quality clusters from a genetic alignment or phylogeny. The significant speed increase our software provides over previous approaches enables fitting a clustering model to previously intractable large alignments, and has the potential to allow continuous model refitting in surveillance settings. Fastbaps is provided as an open source package with a clearly documented interface.

Declaration of contributions

Jukka Corander, Simon Frost, Stephen Bentley and Julian Parkhill supervised this work. Neil MacAlasdair interpreted the results of the *N. gonorrhoeae* association analysis. Neil MacAlasdair, Christopher Ruis, Aaron Weimann, Gal Horesh and John A. Lees contributed to method development. Rebecca A. Gladstone, Stephanie Lo, Christopher Beaudoin and R. Andres Floto aided in the interpretation of the *S. pneumoniae* and *M. tuberculosis* results.

3.1 BACKGROUND

Prokaryotic genome evolution is driven by both the transfer of genetic material vertically from parent to offspring as well as by horizontal gene transfer between organisms [204]. Large population sequencing studies of bacteria have confirmed that this results in large scale differences in intraspecies genome content [228]. This has led to the description of the pangenome, the set of all genes that have been found in a species as a whole [349]. Within the pangenome, genes are often then described as being part of the ‘core’ genome, the set of genes present in all members of a species, or the non-core (‘accessory’) genome. Throughout this chapter we refer to the problem of correctly identifying all the gene families that are present in a collection of annotated assemblies interchangeably as both inferring and determining the pangenome.

A common problem when inferring the pangenome of bacterial genomic datasets is the classification of homologous genes, usually defined by a percentage shared sequence identity, into either orthologous or paralogous clusters. Orthologs trace their most recent common ancestor to a speciation event whereas paralogs trace their most recent common ancestor to a gene duplication event. When analysing bacterial pangenomes we are often interested in identifying paralogs as genes with near identical sequence that may perform a different function or be under differential regulation at different locations in the genome. Many programs for pangenome analysis therefore use location information to further identify paralogs, as well as xenologs which occur when gene duplications are acquired through horizontal gene transfer.

Previous approaches which aid in the inference of the pangenome of a collection of bacterial isolates include Roary, OrthoMCL, PanOCT, PIRATE, PanX, PGAP, COGsoft, MultiParanoid, PPanGGoLiN and MetaPGN [24, 133, 175, 185, 276, 302, 390, 394, 539]. The majority of methods for determining the pangenome tend to make use of one of two similar approaches (see Figure 9). Most start by inferring similarity between predefined gene sequences using a homology search tool such as CD-HIT, BLAST or DIAMOND [6, 49, 304]. Using this output a pairwise distance matrix is created and genes are then clustered into orthologous groups using either the popular Markov Clustering algorithm (MCL) or by looking at triangles of pairwise best hits (BeTs) [151, 474]. A subset of these methods then use gene adjacency information to build a graphical representation of the pangenome. This graph is then used to further split orthologous clusters into paralogs. Roary, PIRATE, PPanGGoLiN and MetaPGN also provide this graphical representation as an output file. A final step of some pipelines is to classify the resulting clusters into core and accessory categories based upon their prevalence within the data set. This is usually done using predefined thresholds, however, more recently, model based extensions to this approach have been suggested [185].

As bacterial genomic population studies have grown larger there has not been a corresponding increase in genome annotation accuracy or genome assembly contiguity. Thus as these databases have grown, so has the number of erroneous gene annotations. This can have profound implications for the resulting estimates of the number of gene families present, whereby a higher number genomes leads to a higher number of errors [122, 434]. Such errors can cause difficulties in any downstream modelling of the pangenome, such as the modelling of negative frequency-dependent selection (NFDS) acting through the loci in the accessory genome [86, 347]. Errors can be introduced into pangenome analyses by fragmented assemblies, mis-annotation, contamination and mis-assembly. Denton *et al.*, have shown that fragmented assemblies were the major cause of inflated gene numbers in draft eukaryotic genomes [122]. Whilst errors often lead to inflations in the estimates of the size of the accessory genome they can also lead to missing genes when the annotation software fails to identify a gene or where the gene is fragmented by a break in the assembly, which reduces the estimated size of the core genome. Many current pangenome inference algorithms have not been subjected to rigorous verification using simulated data. Consequently, their ability to deal with the errors occurring in the initial genome annotations has received limited attention.

In this chapter, we present an alternative approach to inferring the pangenome, Panaroo, which makes use of a graph based algorithm to share information between genomes, allowing us to correct for many of the sources of annotation error. Panaroo leverages the additional

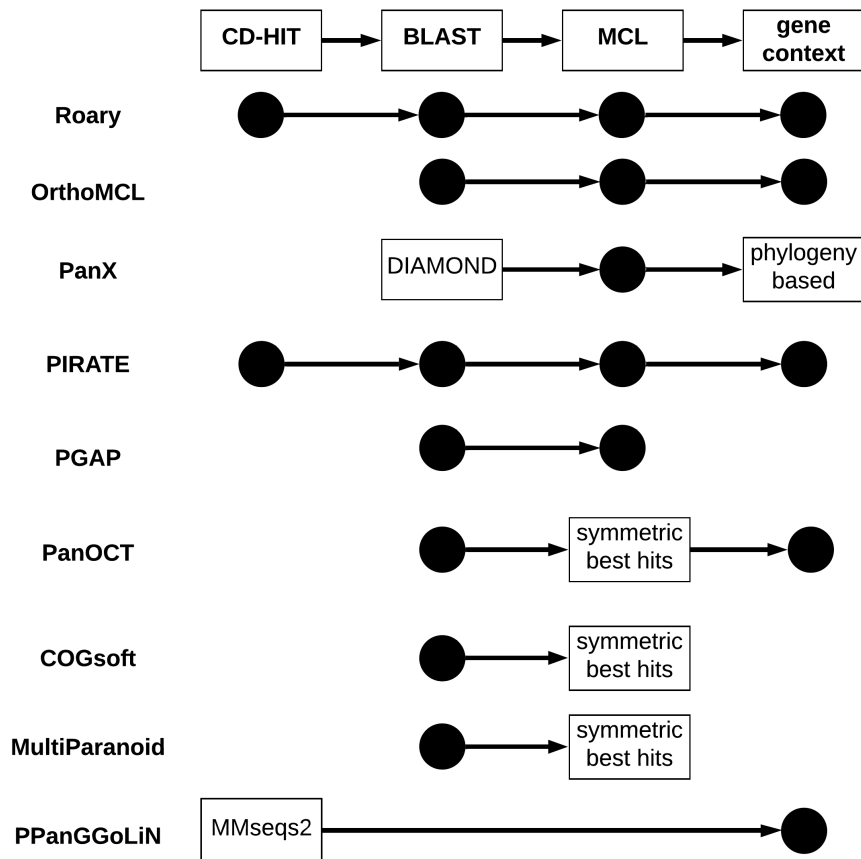


Figure 9: A comparison of the pipelines used by different pangenome analysis tools.

information provided by each genome in a dataset to improve annotation calls, and as a result, the clustering of orthologs and paralogs within the pangenome. The package also provides a number of pre- and post-analysis scripts, allowing integrated data quality control, gene association analysis, and for the comparison of pangenomes between species. As Panaroo constructs a full graph representation of the pangenome, we are able to investigate structural variations within the resulting graph, allowing for associations between structural variations and phenotypes to be called. We demonstrate the success of the algorithm through extensive simulation using the Infinitely Many Genes model [21] and by analysing a diverse array of large bacterial genomic datasets including the major clades of the Global Pneumococcal Sequencing (GPS) project [190]. We compare the output of Panaroo with the previous gold standard methods for analysing the pangenome and show that Panaroo produces superior ortholog clusters, often leading to both significant reductions in the size of the estimated accessory genome and increases in the size of the core genome.

3.2 METHODS

3.2.1 *Panaroo algorithm*

The Panaroo algorithm builds a graphical representation of the pangenome where nodes are genes and edges connect nodes if two genes appear adjacent to one another on at least one contig. The algorithm then uses this initial graph structure to perform a number of cleaning steps which correct for many of the problems encountered in genome annotation. Panaroo accepts annotated assemblies in GFF3 format as output by the popular annotation pipeline Prokka [443]. Unlike similar pangenome software, Panaroo attempts to preserve the full global context of each gene in the graph. This is in contrast to other programs such as Roary [24, 276, 390] which uses only the local context surrounding genes to build the graph.

3.2.2 *Initial graph creation*

To first build the graph, Panaroo runs CD-HIT (v4.8.1) at a high sequence identity threshold (98%) [304]. The resulting clusters are then either classified as non-paralogous gene clusters, if they contain at most one instance of each genome, or paralogous clusters if they contain more than one gene from any single genome. Initially, non-paralogous gene clusters are represented by a single node in the graph whilst paralogous clusters are split into a single node for every occurrence of that cluster in the dataset. For instance, if a paralogous gene appears twice in two genomes and once in another, the initial

graph will contain five nodes representing that paralog. The graph is then built by connecting cluster nodes with edges between them if the two clusters appear adjacent to one another on any contig. Paralogous nodes are collapsed back into the maximum number of nodes in which those genes appear in a single genome using the global context of the graph. In the above example, this would result in the final graph having two instances of the paralog node.

3.2.3 *Contig ends*

Fragmented assemblies can cause issues for gene annotation software, whereby genes are often mis-annotated near contig breaks [122]. These spurious annotations appear as short paths of low support edges and nodes that end in a node of degree one that splits off from the main graph. To deal with this, Panaroo recursively removes nodes of degree one that are below a given support threshold as indicated in [Figure 10](#).

3.2.4 *Contamination*

Contigs originating from sample contamination are generally significantly diverged from the target species pangenome. Thus, contaminating contigs tend to appear as disconnected components from the main graph with low support. To remove these, Panaroo uses the same approach as described for contig ends to recursively delete low supported nodes with less than or equal to one degree (see [Figure 10](#)). This approach has the advantage of retaining rare genes which are present in the main graph whilst removing likely contaminants. Whilst this has in general been found to be very successful it can occasionally lead to rare plasmids being removed. We have found that the benefits of removing unwanted noise far exceed the small loss in sensitivity that this approach provides. However, we also provide three settings for the algorithm with the most sensitive retaining such rare calls which can be useful when one is interested in rare plasmids.

3.2.5 *Mistranslation correction*

Many annotation algorithms rely on an initial training phase where their parameters are adapted to the dataset at hand [118, 119, 237]. Often this training is performed separately on each genome. This is the case in the Prokka pipeline, which makes use of Prodigal to perform the initial gene annotation [237, 443]. This can result in an identical sequence being annotated differently in different genomes. To correct for this Panaroo checks genes that are within close proximity in the pangenome graph to determine if any are likely to be mistranslations, frame shift mutations, or pseudogenised gene copies by comparing

their sequence at the nucleotide level. If two gene sequence matches at a high coverage and identity, typically 95% and 99% respectively a mistranslation is called and the gene node with the lower support is collapsed into the node with higher support.

3.2.6 *Collapsing gene families*

Gene families diversify at different rates due to the influence of positive and purifying selection. This makes choosing a strict sequence identity threshold for defining orthologous clusters difficult. Most pangenome analysis software rely on either a pairwise sequence identity or BLAST e-value threshold. This reliance can lead to both over clustering, where separate gene families are incorrectly merged, and over splitting where a single gene family is incorrectly split into several smaller clusters. Many approaches attempt to deal with the former of these problems by utilising contextual information to split apart clusters that have different gene neighbourhoods [175, 390]. More recently, alternatives that make use of clustering at lower thresholds followed by more involved splitting techniques have been proposed [24, 133]. As an alternative to these approaches we extend the idea of using gene context to the over splitting problem. Panaroo utilises gene contextual information to collapse diverse gene families that have been incorrectly split into multiple clusters during the initial pangenome graph creation. Initial gene clusters that share a common neighbour in the graph are compared at a lower pairwise sequence threshold (default 70%). If they fall within this threshold the two nodes are collapsed and the resulting node is annotated to indicate it consists of a more diverse family. We have found that utilising this additional contextual information leads to more robust clusters.

3.2.7 *Identifying missing genes*

Previous pangenome clustering software tools are unable to identify missing annotations. Gene annotations can be lost due to variability in model training, fragmented assemblies and mis-assemblies. Panaroo remedies this issue by identifying pairs of nodes in the pangenome graph where one node is present in a genome and its neighbour is not. The potentially missing node is then searched for in the sequence surrounding the neighbouring node. If a match of sufficient coverage and identity is found, the graph is corrected to include an annotation for this missing gene in that genome. The alignment tool edlib (v1.3.4) is used to perform these searches which enables millions of checks in a reasonable time frame [461].

3.2.8 *Thresholds*

Panaroo uses a number of predefined thresholds to construct the pangenome graph. These can all be adjusted by the user but we provide a number of modes for common use cases. In the ‘strict’ mode Panaroo takes a more aggressive approach to contamination and erroneous annotation removal. This is most useful when investigating genomes where rare plasmids are not expected or when phylogenetic parameters such as gene gain and loss rates are of interest. In these cases erroneous gene clusters can quickly dominate the estimated parameters. In its ‘sensitive’ mode Panaroo does not remove any gene clusters. This is useful if a researcher is interested in rare plasmids which may be hard to distinguish from contamination. When running Panaroo in sensitive mode it is important to be aware of the possibility of a higher number of erroneous clusters. In the following analyses we run Panaroo in both its ‘strict’ and ‘sensitive’ modes with Panaroo generally outperforming all other tools even without contamination removal in the ‘sensitive’ mode.

3.2.9 *Input and output*

Panaroo takes annotated assemblies in GFF3 format as input and generates a variety of output formats including a fully annotated graph in GML format for viewing in Cytoscape or other graph visualisation software [447]. Each gene node and edge is annotated with the genomes to which it belongs as well as the gene annotations given by Prokka, gene sequence and whether or not the node has been classified as being a paralog. This graph format provides a valuable tool for visually inspecting the results of Panaroo. As Panaroo attempts to build the full pangenome graph rather than only using local context, this graph is able to provide insights hidden in many of the outputs of similar tools such as Roary [390]. To allow for simple integration with existing bioinformatics pipelines Panaroo also outputs many of the same file formats as Roary. This includes the same gene presence/absence file format as well as core and accessory genome alignments created using either MAFFT, Prank or Clustal Omega [257, 318, 451]. Panaroo interfaces easily with many other pangenome analysis packages including the latest version of pyseer allowing for associations between phenotypes and gene presence/absence as well as structural variation in the graph to be investigated [289].

3.2.10 *Structural variation*

As Panaroo constructs the full pangenome graph, it is possible to go beyond gene presence/absence and look at the underlying structure of the graph. To facilitate the analysis of this structure, Panaroo gen-

erates a gene triplet presence/absence matrix, indicating when three genes are present in a path along a genome. This is demonstrated in [Figure 16a](#) and the resulting presence/absence matrix can be used in association studies to investigate differences in rearrangements between genomes in a species. The context of each triplet can then be analysed by looking at the full graph in Cytoscape.

3.2.11 *Pre- and post-processing*

To aid in the quality control of input data as well as in the analysis of the resulting homologous gene clusters we have created a number of pre and post processing scripts for analysing pangenomes. We have included a wrapper for the popular Mash and Mash screen algorithms which generates diagnostic plots for quality control prior to running the Panaroo pipeline [387, 388]. These plots include a Multidimensional scaling (MDS) projection of pairwise Mash distances, interactive bar charts to investigate contamination, as well as gene and contig counts.

In addition we have included post processing scripts for estimating gene gain and loss rates using both the infinitely many genes (IMG) model [22, 531] and the finitely many genes (FMG) model of [133, 531]. These are preferable to the common practice of plotting accumulation curves to indicate pangenome size as they account for the diversity and timescale of the sampled isolates. This allows for a clearer comparison between the pangenomes of different species or clades. Panaroo also includes an implementation of the Spydrpick algorithm which allows for the identification of gene presence/absence patterns that are either highly correlated or anti-correlated whilst accounting for population structure [395]. Such correlations can indicate that the genes involved have epistatic effects on fitness or that their presence or absence is a result of similar selective pressures. Finally, the output of Panaroo seamlessly interfaces with pyseer (v1.3.0), a bacterial GWAS package [289, 293]. Pyseer includes a wide range of methods for performing association studies allowing for phenotypic associations to be found with gene or structural presence/absence patterns.

3.2.12 *Simulation and method comparison*

Using the *E. coli* reference genome ASM584v2 as a starting point, we simulated variation in the accessory genome by varying the rates of gene gain and loss using a phylogeny simulated with the Kingman coalescent in dendropy (v4.4.0) [470]. In addition, we simulated various degrees of sequence variation by varying the within gene codon substitution rate. Three replicate datasets of 100 sampled genomes were created for each set of model parameters outlined in [Table 5](#). Re-

Gain/Loss Ratio	Mutation Rate	Source of Error
0.1	1.00E-14	NA
1	1.00E-14	NA
10	1.00E-14	NA
10	1.00E-15	NA
10	1.00E-16	NA
10	1.00E-14	Fragmentation
10	1.00E-14	Contamination with <i>S. epidermidis</i>

Table 5: Parameters used to simulate pangenomes using the Infinitely Many Genes model.

alistic sequence assemblies were generated by first simulating NGS sequencing reads using either Mason (v2.0.9) or ART (v2.5.8) [229, 231]. These were assembled using SPAdes (v3.13.0) [17]. The resulting assemblies were annotated using Prokka (v1.13.3) with a custom BLAST database containing the correctly assigned proteins from the simulation prior to assembly. This extensive simulation pipeline provided more realistic data and included many of the sources of error encountered in pangenome analyses. To simulate the problems that fragmentation can bring to the analysis of pangenomes we also simulated a fragmented assembly by breaking the simulated whole genomes into fragments prior to simulating the NGS reads. This resulted in highly fragmented final assemblies. Contamination was also simulated by randomly adding 10kb segments of the *S. epidermidis* reference genome ASM764v1 a common lab contaminant to the simulated genomes prior to NGS simulation. These segments were added by sampling from a Poisson distribution with mean 1. The gene presence/absence matrix was then generated for PanX (v1.5.1), Roary (v1.007002), PIRATE (v1.0), COGsoft (v201204), PPanGGolIN (v1.1.72) and Panaroo (v1.0.0). These were compared with the simulated matrix and the number of inferred orthologous clusters that contained an error was counted and is shown in Figure 13.

3.3 RESULTS

3.3.1 Analysis of a *Mycobacterium tuberculosis* outbreak in London

To assess the effectiveness of Panaroo and the impact of annotation errors on other pangenome inference methods we analysed a large outbreak of highly clonal, isoniazid resistant *Mycobacterium tuberculosis* in London [57]. *M. tuberculosis* exhibits a very low mutation rate and is understood to have a ‘closed’ pangenome. Due to the short timescale of the outbreak, the maximum pairwise SNP distance

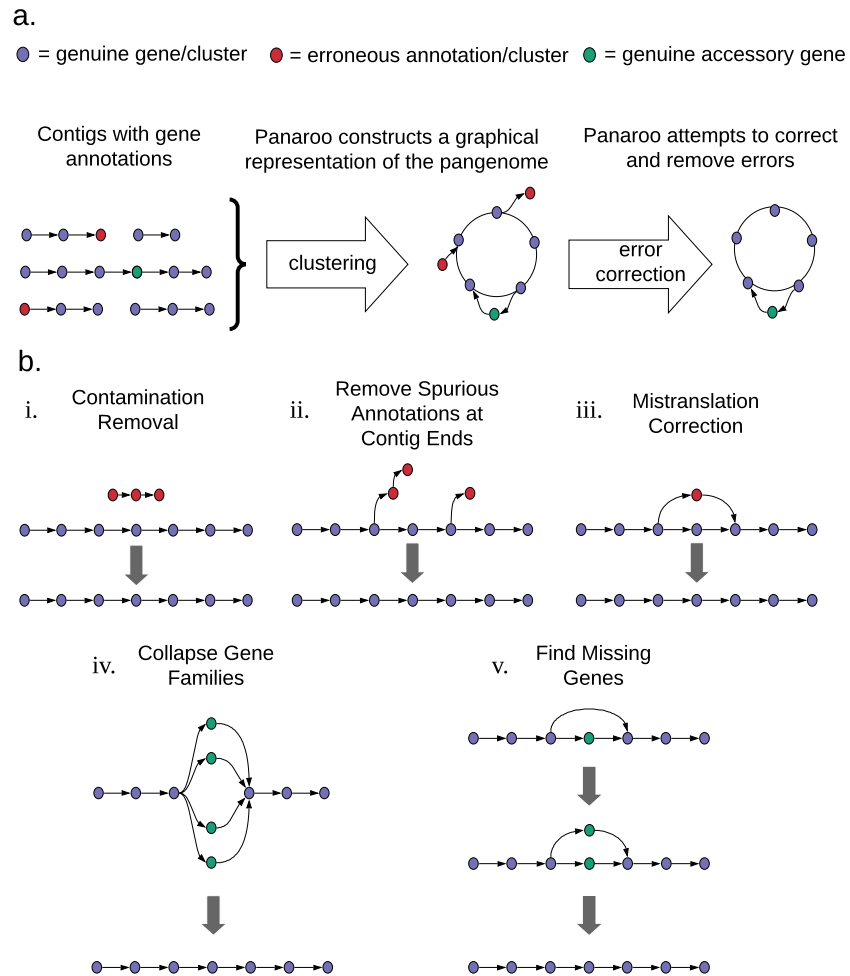


Figure 10: a.) An overview conceptualizing the problem with current gene annotation methods and the stages Panaroo uses to correct for annotation errors. b.) Expanded specific stages in the process. i.) Contamination appears in the graph as poorly supported components. In the default mode, Panaroo removes contamination by recursively removing poorly supported nodes of degree 1. ii.) Genes are often mis-annotated near contig breaks [122]. Panaroo corrects such mis-annotations by recursively removing poorly supported nodes of degree 1. iii.) Panaroo corrects cases where the same DNA sequence has been translated in multiple reading frames into a single gene by clustering concomitant genes at the DNA level. iv.) Panaroo uses context and a lower clustering threshold to combine diverse gene families into a single gene. v.) Annotation algorithms may predict a gene in some but not all samples, even when the samples share exactly the same DNA sequence. Panaroo finds missing genes by searching for the gene sequence in the surrounding DNA.

within this dataset was 9. As we would expect to find no pangenome variation, this dataset provides a useful control to compare the different pangenome tools.

We ran each of the pangenome inference methods on all 413 *M. tuberculosis* genome assemblies after first annotating them using Prokka [443]. Panaroo identified both the highest number of core genes and the smallest accessory genome (Figure 11), consistent with the established biology of *M. tuberculosis* and a highly clonal dataset [79, 192]. In contrast, PanX, PIRATE, PPanGGoLiN, COGsoft and Roary all reported inflated accessory genomes ranging in size from 2584 to 3670 genes representing a nearly ten-fold increase to that reported by Panaroo. In its default mode PPanGGoLiN reported over 10,000 gene clusters giving it the highest error rate. However, this reduced to 7,131 after enabling the `-defrag` parameter. We thus ran PPanGGoLiN using this parameter in all analyses. The small number of accessory genes that Panaroo did predict mostly consisted of core genes where the algorithm was unable to refine the genes in a subset of the assemblies. The majority of the difference between the methods was driven by genes being fragmented during assembly (59%). A smaller subset of genes were only called in a small minority of the isolates despite the underlying sequence being nearly identical (10%). Whilst some of these differences could be due to frame shifts in the PE/PPE genes, 27.9% of the isolates were indistinguishable with only one isolate being more than 5 SNPs from this major clone. We found that the majority of the difference was due to the annotation algorithm optimising for each isolate individually, leading to inconsistent gene calls. However Panaroo's consensus approach helps to resolve these discrepancies. The magnitude of the difference observed in this dataset suggests that failing to account for annotation errors can have profound impacts on the resulting estimates of the pangenome.

An alternative to correcting gene annotations is to perform strict quality control checks on assemblies prior to running pangenome inference tools. For very highly contaminated assemblies or those of very low quality this can be the best option and Panaroo includes a quality control pre-processing script for this purpose. However, in many cases low level contamination and fragmented assemblies are common and thus filtering out assemblies with minor errors can lead to large data losses. In large collections even very low annotation error rates will eventually compound pangenome inference results. To investigate such a strategy on the *M. tuberculosis* dataset we ran CheckM, which is a common assembly quality control pipeline [391]. CheckM produces completeness and contamination scores by using a reference gene data set to compare with assemblies. The resulting scores on the *M. tuberculosis* dataset are given in Figure 12. As we know this data set should contain highly similar assemblies it is possible to identify a number of problematic genomes with slightly lower

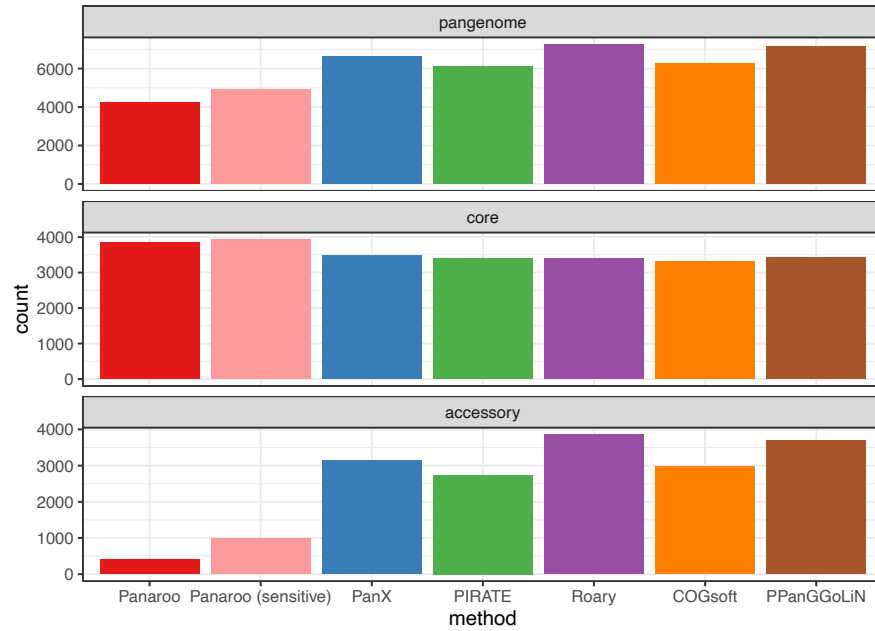


Figure 11: Pangenome counts for 413 *Mycobacterium tuberculosis* genomes from an outbreak in London [57]. The maximum pairwise SNP distance between these isolates was 9, suggesting extremely limited variation. Consequently, we would expect a very limited accessory genome and a core genome of approximately 4000 genes. All tools with the exception of Panaroo found in excess of 2500 accessory genes, which can be attributed to annotation errors.

completeness scores. If we were to remove these genomes it would result in a loss of 12% of the data set which could potentially have a large impact on downstream analysis. Instead, using Panaroo we are able to retain these assemblies whilst controlling the error rate.

3.3.2 Performance on simulated populations

To further assess the ability of the different methods to identify the correct gene presence/absence matrix, we simulated pangenomes using the *E. coli* reference genome ASM584v2 (accession number NC000913) and the Infinitely Many Genes model [21, 83]. To more accurately simulate the kind of errors that typical annotation pipelines produce, we simulated short read assemblies from these pangenomes using Mason, ART and SPADes [17, 229, 231]. We conducted five simple and two more complicated simulations, each with three replicates (Table 5). In the simple simulations, the gene gain/loss rate was varied with lower rates corresponding to a larger core genome and smaller accessory genome, whereas higher rates corresponded to a larger accessory genome. The mutation rate of the accessory genome was also varied. In addition, we simulated two more complicated datasets, one of which had an increased level of fragmentation of the assembly by

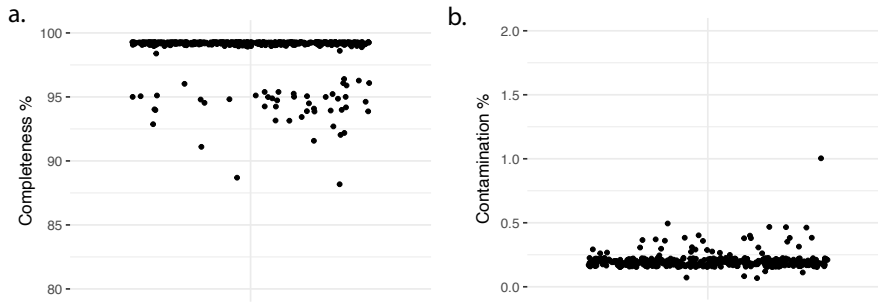


Figure 12: Completeness (a.) and contamination (b.) results after running the CheckM pipeline on the 413 *M. tuberculosis* genomes

fragmenting the input genome prior to the NGS simulation. This resulted in a mean N50 of 23,112. The second more complex simulation included contamination by randomly adding in short fragments of the *S. epidermidis* reference genome, which is a common contaminant. The more complicated simulations represent data sets with unusually high error rates but help to clarify the large impact that these sources of error can have on pangenome inference as was previously demonstrated in the analysis of the highly clonal *M. tuberculosis* data set.

Figure 13a indicates the number of gene clusters which contained errors for each of the scenarios. Such errors included genes that were missing, incorrectly annotated or had incorrectly clustered together. Most methods performed fairly well when applied to the output from the simple simulation. All methods include some errors due to genes never being annotated except in the original reference. As each method relied on the same input files this was consistent between methods.

For the simple simulations, PanX and Panaroo produced the fewest errors, followed by PIRATE, Roary, PPanGGoLin and COGsoft. Roary was the most sensitive method to the substitution rate, with higher rates leading to more errors. This can be attributed to its reliance on a strict BLAST e-value threshold. COGsoft gave variable results, performing poorly on pangenomes with larger accessories suggesting it may over collapse genes. This interpretation was further supported in our analysis of a diverse *K. pneumoniae* dataset (see Section 3.3.3).

Whilst most methods were able to perform adequately on relatively error-free simulated data, the introduction of more realistic significant sources of annotation error had a large impact. Figure 13b indicates the resulting error counts after simulating both contamination and highly fragmented assemblies. Here, the importance of Panaroo's multiple annotation error correction approaches becomes apparent. As expected, when small amounts of contaminating *S. epidermidis* DNA were added to the simulated NGS data all methods except Panaroo and COGsoft incorrectly called a larger accessory genome. This is due to their inability to account for and remove contaminating con-

tigs. In contrast, Panaroo achieved error rates similar to that found for the clean assemblies. As expected Panaroo's sensitive mode did not correct for the additional contamination as potential contamination is not removed in this mode. COGsoft had a similar number of total errors to the other programs but rather than calling a larger accessory genome tended to incorrectly merge the contamination with other genes.

The highly fragmented assemblies led to the largest error rates in each pangenome analysis tool. Fragmentation can lead to gene annotation software such as Prokka miscalling genes near the ends of contigs. It can also impact on the consistency of the training step in some annotation algorithms. This resulted in a large increase in the estimated accessory genome size for all methods except Panaroo. Similarly, miscalling can lead to genes being left unannotated resulting in smaller estimates of the core genome. In both cases Panaroo's error correction and refinding steps were able to accurately recover the true pangenome, while PanX, COGsoft, PIRATE, PPanGGoLiN and Roary all produced nearly an order of magnitude higher error rates. Whilst Panaroo's sensitive mode produced far cleaner results than the other tools the fact that it does not delete clusters prevents it from removing some spurious annotations. These results mirror that observed in the analysis of the highly clonal *M. tuberculosis* outbreak, helping to confirm the impact that such errors can have on pangenome estimates.

3.3.3 A diverse *Klebsiella pneumoniae* collection

We then went on to compare each method on a more complex real dataset – 328 globally sourced *K. pneumoniae* genomes from both human and animal hosts [228]. *K. pneumoniae* is a highly diverse gram-negative bacterium that can colonise both plants and animals and has previously been found to have a large pangenome [228]. The high recombination rate and often multiple plasmids per bacterium complicates analysis of the *K. pneumoniae* pangenome. Nine of the 328 isolates were identified as outliers by the Panaroo quality control script due to the number of contigs or number of genes they contained (see Figure 14). These isolates were removed before running each algorithm. Figure 4a indicates the resulting total, core and accessory gene counts inferred by each method, using the 99% presence threshold for core genes as used in Roary [390].

As species such as *K. pneumoniae* are known to have many rare plasmids which are difficult to distinguish from contamination, Panaroo's sensitive mode is of particular relevance here. Panaroo identified the highest number of core genes in both its default and sensitive modes, 3372 and 3376 respectively. Hence for these genomes there was only a minor difference in the estimated core between the two pipeline options. Roary identified the smallest core genome of 1800 genes. Given

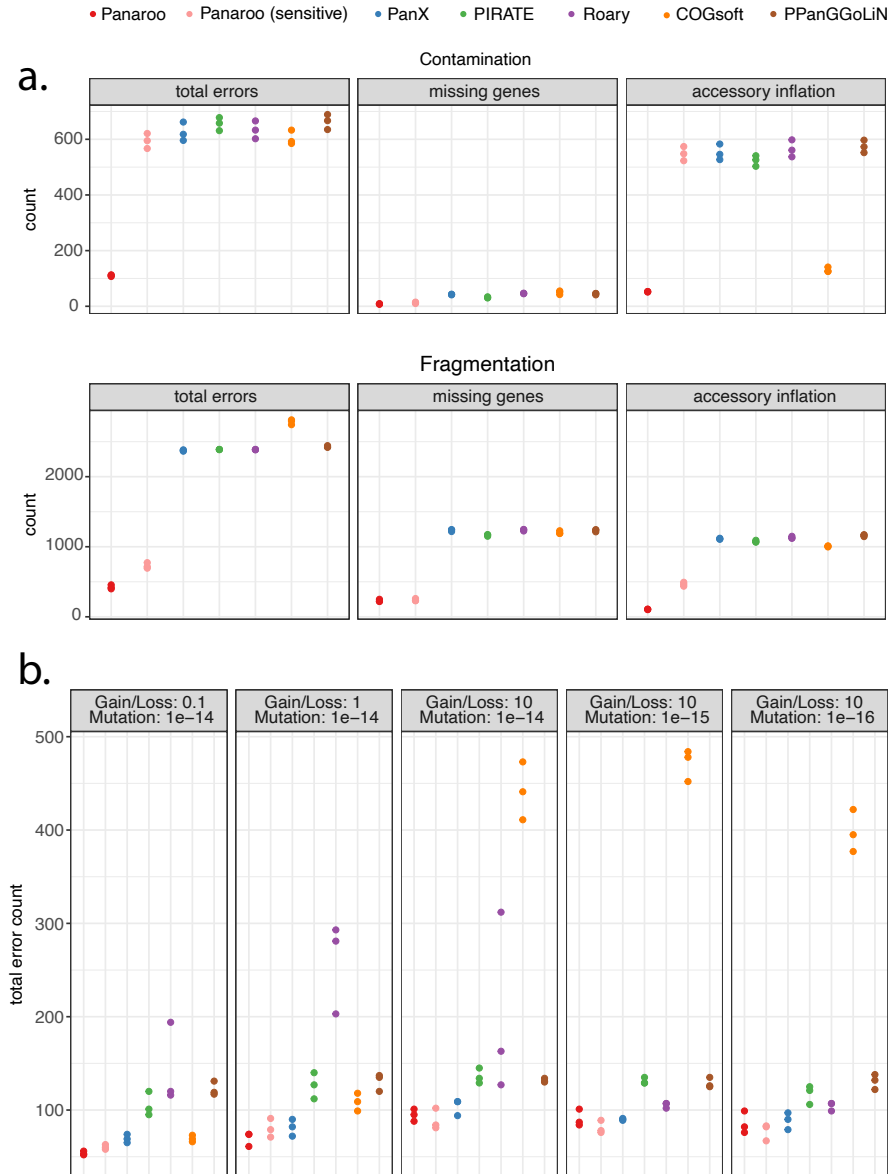


Figure 13: Error counts for the different algorithms after comparing with simulated data on different scenarios. Accessory genome inflation refers to the number of erroneous clusters that do not correspond to any simulated gene cluster. Missing genes refer to false negative gene calls where the annotation is not present in the final pangenome. Even in simulations of pangenome variation from a single *E. coli* reference with only relatively simple sources of error, a), panaroo outperforms other methods across a variety of gene gain/loss rates and mutation rates. In more realistic simulations of sequencing data, b), the only method with reasonable control of the error rate is panaroo.

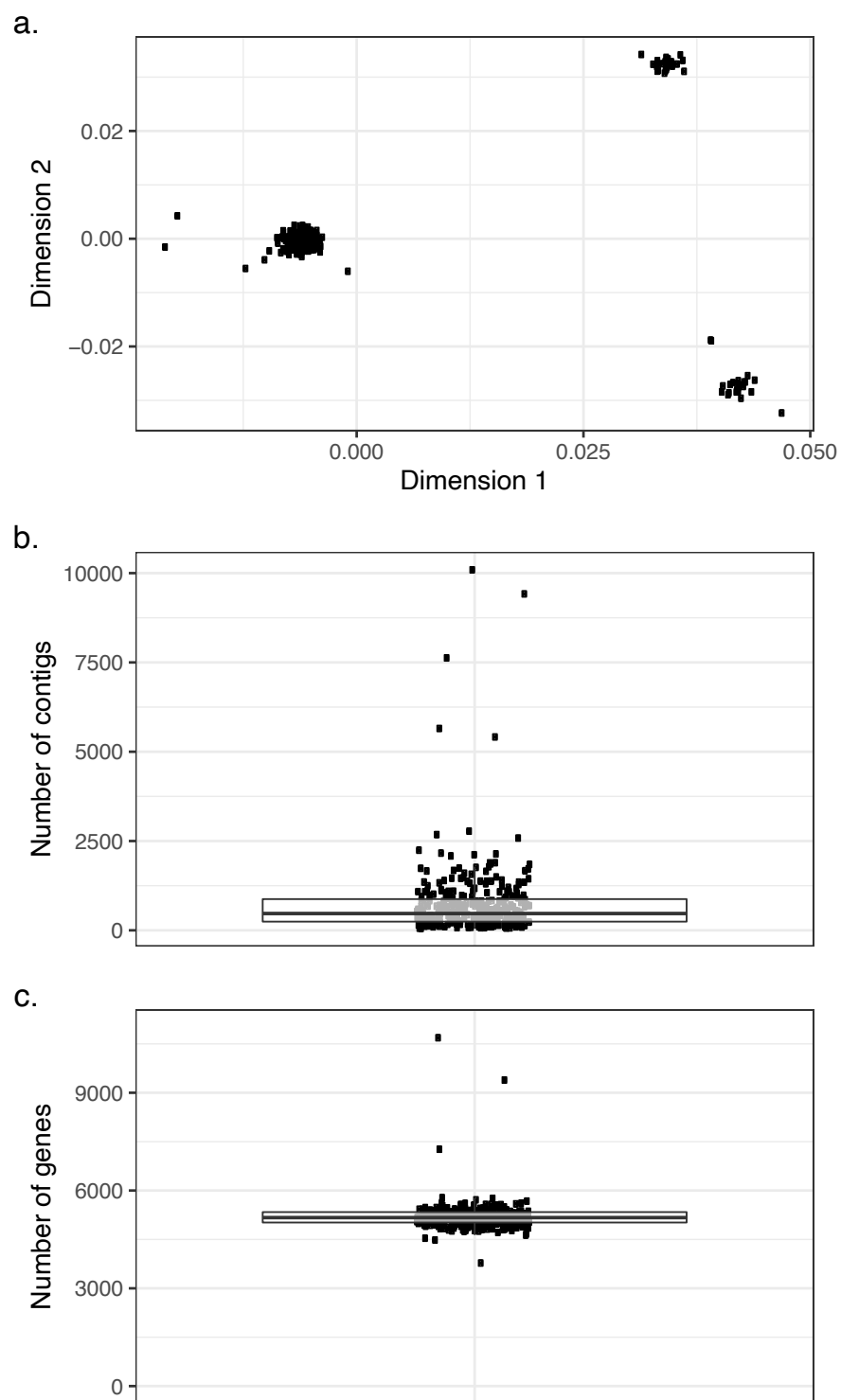


Figure 14: (a) Multi Dimensional Scaling (MDS) plot of pairwise mash distances between isolates in the global *K. pneumoniae* dataset. (b & c) Boxplots indicating the number of contigs and genes in each of the *K. pneumoniae* assemblies.

the result of the simulations, this is likely due to gene clusters being incorrectly split into multiple smaller clusters, as the default Roary pairwise identity threshold of 95% is too stringent for such a diverse dataset. PIRATE relaxes the strict threshold required in Roary and it identified a similar number of core genes to Panaroo (3318) but a smaller number of accessory genes than both the Panaroo (sensitive) and PanX methods which agreed more closely with the original estimates in Holt *et al.* [228].

Whilst there is no gold standard with which to compare these results, we can look at the gene annotations within clusters to identify cases where a gene cluster contains multiple different annotations, which would suggest separate gene clusters have been incorrectly merged. Figure 15b indicates the number of conflicting annotations in the clusters of each method. As gene fragments and genes annotated as “hypothetical” are often the result of errors and thus can have erroneous annotations, we did not consider conflicts that involved these. Panaroo in both its default and sensitive modes had the lowest number of conflicting annotations. PPanGGoLiN had the second lowest number whilst COGsoft recorded the highest number of conflicts which is consistent with the tendency of its method to over-cluster genes. The lower number of conflicting annotations found by PPanGGoLiN is consistent with it favouring splitting gene clusters over merging them as we found earlier. Overall, Panaroo identified a larger core genome and fewer conflicting annotations than any other method showing that its error correction approach is also suitable for diverse datasets of highly recombinogenic bacteria.

3.3.4 Pangenome association analysis with Pyseer

Panaroo provides a number of outputs as well as post processing scripts for analysing the cleaned pangenome graph. Panaroo outputs both a gene presence/absence matrix as well as structural variation presence/absence matrix that can be used as input to pyseer or Scoary for association analyses [48, 289]. We refer the the gene presence-absence and structural variant association pipelines as pan-GWAS and sv-pan-GWAS respectively. Panaroo generates structural variation calls by identifying distinct consecutive triplets of gene families in the graph that describe different paths through a node (see Figure 16a). As larger insertion and deletion events will only be represented once in the structural presence/absence matrix rather than repeatedly for each gene, this approach increases the power of such association analyses. The approach also identifies associations with large structural re-arrangements although these are often more difficult to interpret. Once a significant association between a gene triplet and a phenotype of interest have been identified, the context of the structural rearrangement can be investigated manually by interrogat-

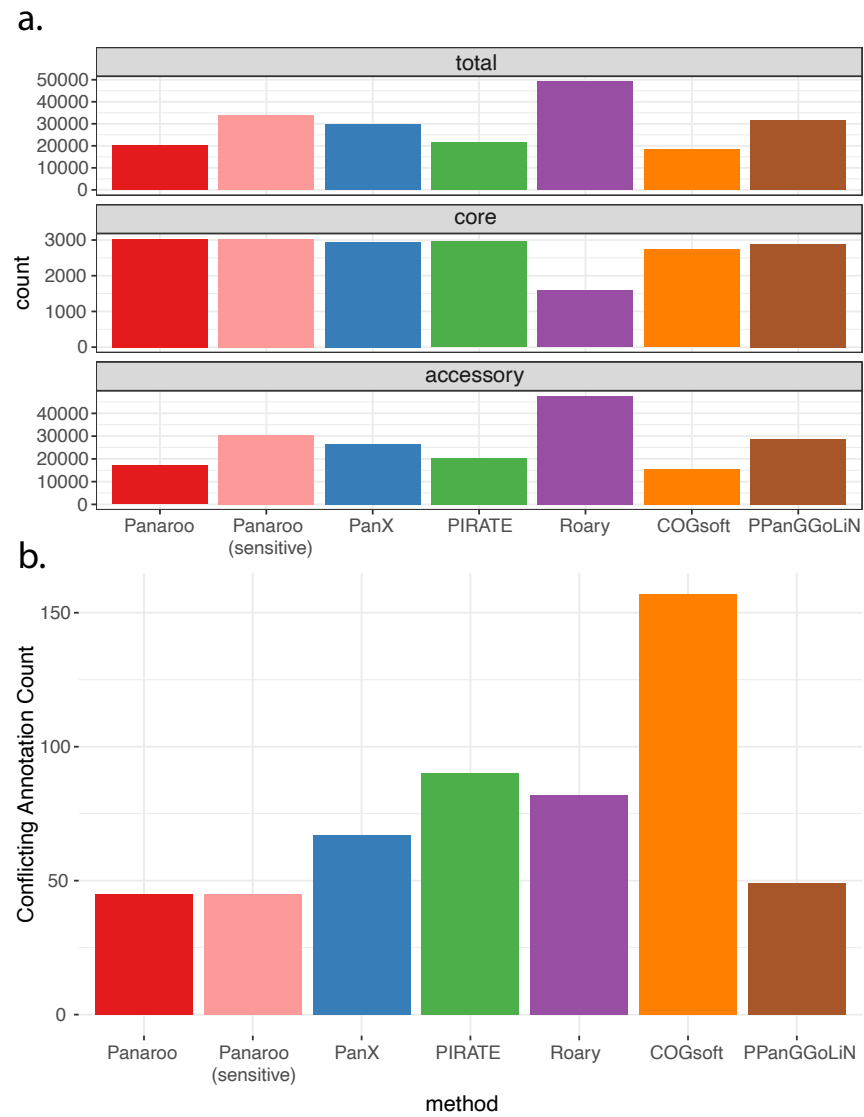


Figure 15: (a) the estimated pangenome, core and accessory sizes from the different algorithms in the global *K. pneumoniae* dataset. (b) The number of conflicting gene annotations in the inferred clusters of the different algorithms.

ing the pangenome graph in Cytoscape [447]. Panaroo is only able to call large structural rearrangements that result in genes being re-located within the genome. Finer scale structural variants are better called using assembly graph based approaches such as Cortex [242].

To verify the pangenome associations pipelines, we investigated the Euro-GASP collection of 1054 *N. gonorrhoeae* isolates collected from 20 countries across Europe from September to November 2013 [208]. The minimum inhibitory concentrations (MICs) evaluated for seven antibiotics in the original study were used along with the outputs of Panaroo as input to pyseer [289].

The pan-GWAS pipeline identified 67 genes with significant associations to elevated MIC values (adjusted p-value ≤ 0.05). These included an uncharacterised ABC transporter (group_464), for penicillin resistance. ABC transporters are known to function as efflux pumps providing a resistance mechanism to ribosome-targeting antimicrobials [510].

Using the same MIC values the sv-pan-GWAS pipeline identified 138 triplets with significant associations with resistance (adjusted p-value ≤ 0.05). These included many linked with known mobile elements included phage-associated, transposase, or pilin genes. The full list of significant genes and triplets can be found in the supplementary material of Tonkin-Hill *et al.*, [485].

Four triplets were linked with either the insertion or deletion of the gene cluster 'group_1138', a transmembrane protein. Insertion of this protein was associated with ceftriaxone resistance. Efflux and permeability have been suggested as potentially playing a role in the resistance of *N. gonorrhoeae* to ceftriaxone although the exact mechanism is not well understood [194]. The sv-pan-gwas approach was also able to identify that a common plasmid present in 430 of the *N. gonorrhoeae* genomes actually consisted of a family of several related plasmids. While these plasmids tend to include very similar sets of genes, one version (highlighted in red in Figure 16b) is associated with susceptibility to tetracycline treatment. The alternative versions include the gene cluster 'group_1999', a tetM tetracycline resistance gene which explains the differences in the resistance profiles and suggests that multiple versions of this plasmid with differing resistance profile are circulating within Europe. These results demonstrate the benefits of the sv-pan-GWAS method.

3.3.5 Pangenome evolutionary dynamics

The higher accuracy obtained by Panaroo allows for the comparison of gene gain and loss rates between lineages and species as well as the more accurate inference of pangenome size. Whilst it is common practice to plot gene accumulation curves in the analysis of pangenomes, these are not robust to errors and fail to account for sampling biases

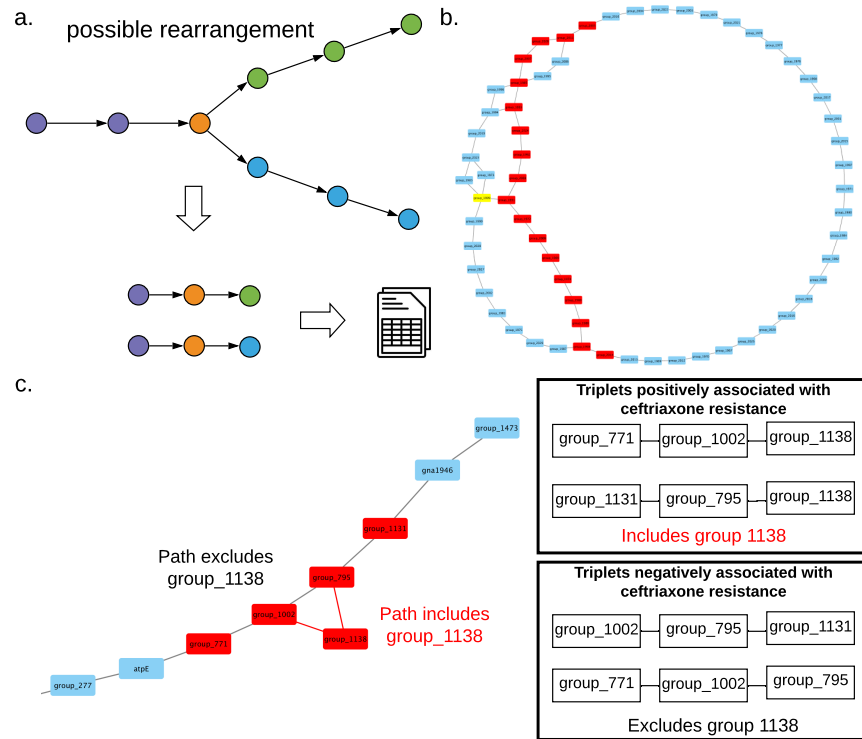


Figure 16: **a)** A diagram indicating how gene triplets are called in the graph. A single genome can only pass through a node once; thus, variations in the arrangement of genes in different genomes can be called using triplets. These triplets are summarised as a binary presence absence matrix. **b)** A family of related plasmids present in the *N. gonorrhoeae* pangenome gene network. The path highlighted in red contained 4 structural variant gene triplets significantly negatively associated with tetracycline resistance, or associated with tetracycline susceptibility by a structural variant pan-GWAS (all adjusted p-value < 0.05). The gene highlighted in yellow, group_1999, was found to be a tetM resistance gene. **c)** A subsection of the *N. gonorrhoeae* pangenome gene network of the region surrounding gene group_1138. The presence of gene triplets (group_771-group_1002-group_1138) and (group_1131-group_795-group_1138) is positively associated with tetracycline resistance while the triplets (group_1002-group_795-group_1131) and (group_771-group_1002-group_795) are negatively associated with tetracycline resistance (all adjusted p-value < 0.05).

and population structure. Thus, accumulation curves should not be used to compare pangenome characteristics of different lineages or species. Recently, a number of phylogenetically informed methods for investigating pangenome dynamics have been published, including the Infinitely Many Genes (IMG) model and the Finitely Many Genes (FMG) model [21, 22, 133]. Both of these approaches account for the diversity of the sample and have been implemented as post-processing scripts in Panaroo.

To demonstrate the utility of using the corrected pangenome graph to infer gene gain/loss rates and pangenome size, we used the FMG model to investigate 51 of the major Global Pneumococcal Sequence Clusters (GPSCs) for which reliable dated phylogenies could be constructed [190]. The major clades of the pneumococcus have distinct accessory gene profiles [98]. We ran Panaroo on each GPSC separately and used the resulting gene presence/absence matrix with the corresponding dated phylogeny to infer gene gain and loss rates for each cluster. We compared the inferred parameters with other variables of interest calculated by Gladstone *et al.* [190], including the inferred recombination rate (r/m), odds ratio of invasive disease and the number of distinct serotypes for each cluster. We found that the estimated effective pangenome size correlated positively with the recombination rate of a cluster (Spearman correlation coefficient 0.53, $p < 0.001$) and the number of serotypes present in the cluster (Spearman coefficient 0.51, $p = 0.001$). This is consistent with biological understanding of the genome diversification and gives confidence to our results, as a higher recombination rate would allow for a clade to more easily gain and lose genes, including serotype-defining gene clusters, resulting in a larger pangenome. Interestingly, GPSCs that have lower gene gain rates were more likely to have a significant odds ratio for invasive disease ($p = 0.04$) (see Figure 17). The association with gene loss rate was weaker, although the effect was in the same direction ($p = 0.08$). Genome reduction has previously been associated with increasingly obligate interactions with the host in multiple unrelated bacterial pathogens [368].

3.3.6 Computational performance

Panaroo uses a similar level of computational resources to competing methods. Figure 18 7 indicates the memory and cpu time required for the analysis of 10, 100 and 1000 *N. gonorrhoeae* isolates subsampled from the Euro-GASP collection. PanX and COGsoft used the most resources with COGsoft not completing the largest dataset in under a week. Roary, PIRATE and Panaroo all performed similarly.

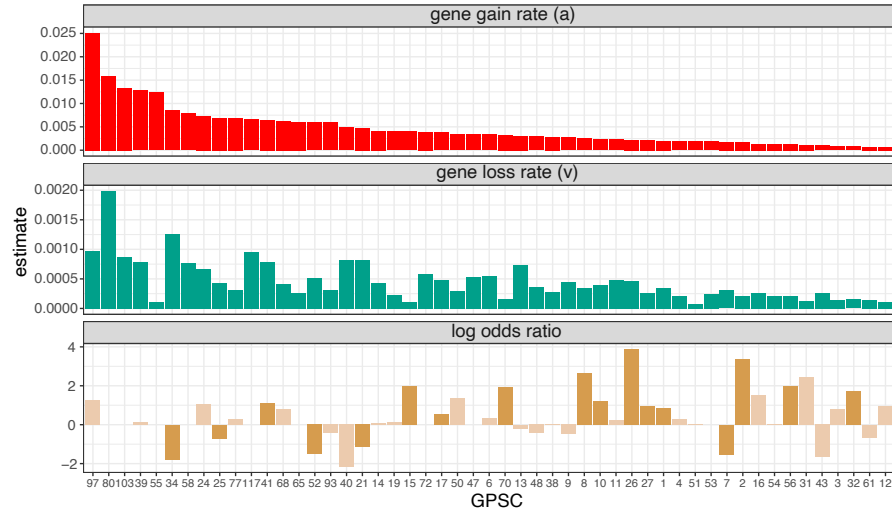


Figure 17: The inferred gene gain and loss rates of each of the 51 major clades of the Global Pneumococcal Sequencing project plotted above the respective log odds ratio of invasive disease in that clade. Clades which had significant odds ratios in Gladstone *et al.* [190] are represented in dark yellow.

3.4 DISCUSSION

Annotation errors, fragmented assemblies and contamination represent a major challenge for pangenome analysis. We have designed Panaroo to tackle these challenges using a sophisticated framework for error-correction that leverages information across strains through a population graph-based pangenome representation. Using both simulations and well-characterised real world datasets, we demonstrated that many commonly used methods greatly inflated the size of the accessory genome while reducing the estimated size of the core genome. In contrast, Panaroo exhibited far lower error rates and reconstructed highly accurate core and accessory genomes for simulated datasets that included contamination and genome fragmentation. Analysis of both a highly conserved *M. tuberculosis* dataset and a highly diverse *K. pneumoniae* dataset indicated that Panaroo provides superior solutions in challenging real world population genomics applications.

Panaroo also includes a number of pre- and post-processing scripts for the analysis of bacterial pangenomes that assist in quality control of the input data and facilitate down-stream processing of the pangenome. We used the Panaroo pre-processing QC scripts to identify nine *K. pneumoniae* samples that were outliers based on the number of contigs or genes and excluded these samples from our analysis. We recommend that such pre-processing QC be carried out on all datasets to identify potentially erroneous samples.

We used the output from Panaroo as input to pyseer to run pan-GWAS and sv-pan-GWAS analyses on *N. gonorrhoeae*. Through this

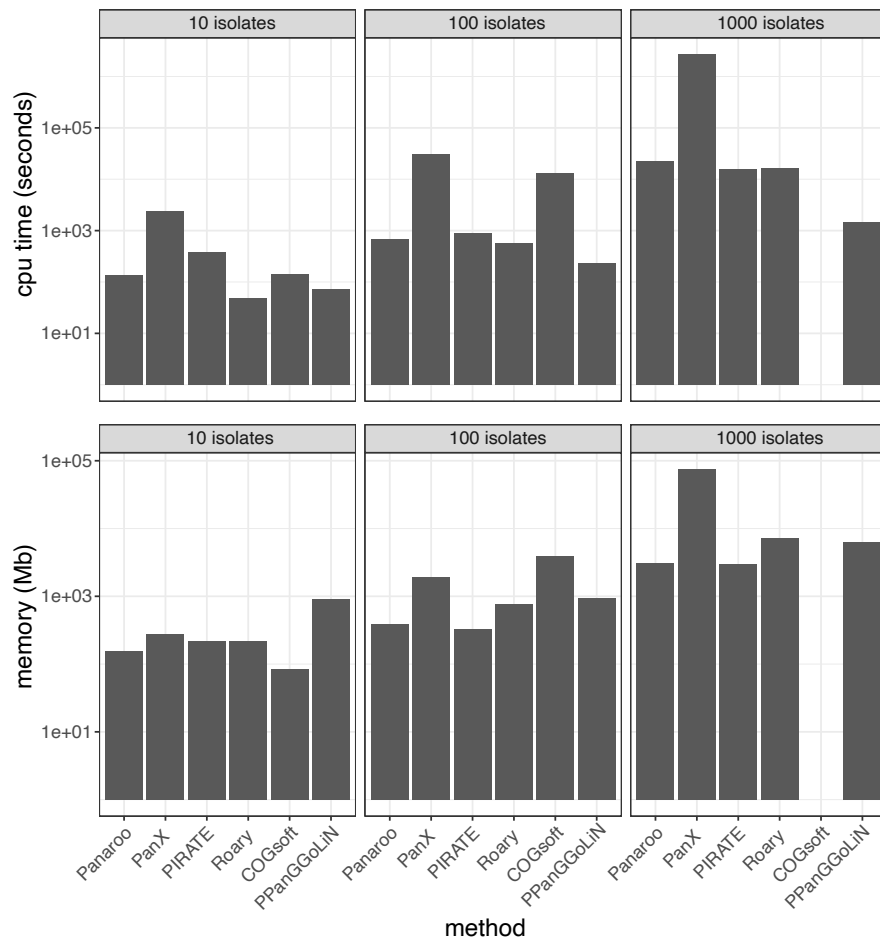


Figure 18: the cpu time and memory required for each of the algorithms for 10, 100 and 1000 *N. gonorrhoeae* isolates. Each tool was run with 5 cpus

approach, we identified a deletion in the genome of *N. gonorrhoeae* in a large European collection that confers resistance to tetracycline. We demonstrated the utility of Panaroo to disentangle highly similar genetic structures through identification of a plasmid family in *N. gonorrhoeae* (Figure 16c). By combining this high resolution picture with structural variant pan-GWAS we identified that some members of this plasmid family carry tetracycline resistance and were able to accurately determine the *tetM* gene as the cause of resistance.

As part of the Panaroo package, we include implementations of recently proposed pangenome evolution models, which are more appropriate than the more frequently used gene accumulation curves. We demonstrated the effectiveness of such methods through the analysis of the 51 major GPSCs where we observed an association between recombination rate and pangenome size (Figure 17). We also identified an association between pneumococcal clade invasiveness and gene gain rate.

Panaroo is written in python (versions 3.6+) and is available under the open source MIT licence from <https://github.com/gtonkinhill/panaroo>. Taking gene annotation errors into account is vital to recover an accurate pangenome, something previous methods have struggled to do in a systematic manner. Panaroo uses gene adjacency in a population-graph to provide a fast method for pangenome analysis, which is robust to a wide range of error sources. In the future, we plan to further improve the computational performance of Panaroo to allow it to scale to datasets involving hundreds of thousands or millions of genomes. We will also extend the post-processing tools available to analyse the resulting pangenome graph.

TRANSMISSION INFERENCE

Declaration of contributions

Jukka Corander, Simon Frost and Stephen Bentley supervised this work. The adaptation of the TransPhylo algorithm to endemic disease was implemented with suggestions from Caroline Colijn. Caroline Colijn, Yuanwei Xu and Xavier Didelot have since implemented a version of the TransPhylo algorithm that allows parameters to be shared between different phylogenies.

4.1 BACKGROUND

The tracking of infectious disease through a population has historically been dominated by epidemiological methods such as contact tracing and case counting. One of the best-known examples of the tracking of transmission is from the nineteenth century when John Snow mapped the location of cholera cases to identify the source of an outbreak as contaminated water from the Broad Street pump in the Soho district of London. More recently, by leveraging pathogen genetic sequence data, it is possible to build a more detailed picture of transmission and to disentangle transmission patterns that are obscured when only epidemiological data are available [209]. Previously, transmission networks were inferred from detailed contact tracing information. While this approach is still highly informative, person to person contact information is not always readily available. Additionally, contact information alone cannot always differentiate between transmission patterns such as those seen in person to person transmission versus zoonotic transmission. Previous genotyping methods including multilocus sequence typing report genetic variation at a small number of sites resulting in limited temporal resolution to distinguish transmission chains. As whole-genome sequencing has become more affordable, near-real-time analysis of pathogen transmission is starting to become routine [64]. One of the first examples of the use of whole-genome sequencing in studying bacterial transmission at a clinically relevant time-scale involved the analysis of methicillin-resistant *Staphylococcus aureus* (MRSA) in a neonatal intensive care unit [274]. Here, WGS allowed for the identification of direct transmission, including one previously unidentified transmission event. Additionally, the WGS clearly distinguished those cases involved in the outbreak from similar infections occurring through alter-

native transmission routes. To date, the increased resolution provided by genomic data has often been more valuable in ruling out potential transmission scenarios rather than in identifying complete transmission networks. Although there undoubtedly exist new opportunities for unravelling past and future transmission networks, there are still many methodological challenges plaguing who-infected-whom studies. Biased sequence sampling, population structure, genetic recombination, within-host evolution and multiple infection can all affect the accuracy of inferred transmission networks [182].

Unlike the previous chapters, this chapter is split into three sections that describe three general strategies for transmission inference using genomic data. In each section we detail novel extensions to existing methods including a new approach to the analysis of deep sequencing data. We start by concentrating on pairwise genetic distance based methods that rely on WGS of single genomes. Next, we consider methods that attempt to infer the full transmission network and describe adaptations to an existing algorithm which leads to improved results when applied to transmission in endemic disease. Finally, we describe the limitations of relying only on a single genome per host and suggest a novel approach for inferring the direction of transmission using deep WGS to account for the genetic diversity of pathogens within the host.

4.2 PAIRWISE TRANSMISSION INFERENCE

One of the simplest and most commonly used methods for identifying transmission links using genomic data is to count the number of single nucleotide polymorphisms (SNPs) that separate two genomes. Depending on the pathogen species being considered a SNP distance threshold is often used to either include or exclude the possibility of a direct transmission between the respective hosts. Due to the high variability in evolutionary rates and genome sizes of different pathogens, SNP threshold approaches provide vastly different temporal resolutions depending upon the pathogen [56]. The very low evolutionary rate in *M. tuberculosis* often leads to many identical genomes even when an outbreak spans a number of years [57]. In this case, distinguishing direct from indirect transmission can be impossible without the aid of additional high quality epidemiological information. Instead, in these scenarios, genome sequencing is used to exclude the possibility of direct transmission. This can help to identify the number of independent introductions of a pathogen into a population, which can be highly informative for the public health response. A recent example of this is the use of genomics to identify the number of independent introductions of SARS-COV-2 into the United Kingdom, Australia and New Zealand in 2020 [186, 403, 444]. Here, using genomics, public health officials were able to ex-

clude the possibility of ongoing transmissions within the Australian and New Zealand populations and instead identified failures in hotel quarantine as the source of novel outbreaks of COVID-19 [186, 444]. In contrast to the relatively slow evolutionary rate of *M. tuberculosis*, some bacterial pathogens evolve at rates that can be more informative for identifying recent transmissions. A notable example is MRSA. With an evolutionary rate of approximately 1 core genome SNP every 6 weeks, WGS has been effectively used to track transmission of the clinical important bacterial pathogen within hospitals [209, 274].

4.2.1 Calculating SNP distance matrices

Multiple sequence alignments (MSA) of genomes from a single pathogen species typically consist of many alignment columns which are highly conserved. This is particularly the case when analysing outbreaks of highly similar genomes. The low sequence divergence in many bacterial core genes also leads to conserved columns in core genome MSAs. The majority of existing tools for calculating pairwise SNP distance compare each position in each genome of an MSA sequentially [41, 442]. Thus, a naive pairwise distance calculation of 100 genomes with an MSA length of 1000 positions would lead to $\sim 9.9 \times 10^6$ nucleotide comparisons. However, due to the high level of sequence conservation the vast majority of these comparisons are often redundant.

The need to perform sparse matrix algebra is a common theme in scientific computing and has led to the development of efficient algorithms and libraries for the computation of sparse matrix addition, subtraction and multiplication. By only considering sites in an MSA that differ from the consensus we can represent an MSA as a sparse matrix. This allows for the efficient calculation of pairwise SNP distances.

Let M be a matrix indicating the alleles that differ from the consensus in a MSA. Thus, an MSA with 100 genomes and an alignment 1000 nucleotides in length will lead to a sparse matrix M with 100 rows and 1000 columns. If genome i differs from the consensus at position j and has the allele a then $M_{ij} = a$. Let I_k be a binary matrix such that

$$I_{k,i,j} = \begin{cases} 1 & \text{if } M_{ij} = k \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Using this notation, a symmetric matrix indicating the number of shared non-consensus SNPs between every two genomes can be calculated as

$$S = I_a I_a^T + I_c I_c^T + I_g I_g^T + I_t I_t^T \quad (2)$$

To account for the total number of non-consensus alleles in each genome we construct the matrix P

$$P_{ij} = \sum_{\forall l} \sum_{k \in \{a,c,g,t\}} I_{k,i,l} + \sum_{\forall l} \sum_{k \in \{a,c,g,t\}} I_{k,j,l} \quad (3)$$

To avoid double counting differences in alleles at the same location in the alignment we construct the matrix Q as

$$Q = \left(\sum_{k \in \{a,c,g,t\}} I_k \right) \left(\sum_{k \in \{a,c,g,t\}} I_k \right)^T \quad (4)$$

The pairwise distance matrix can then be given by

$$D = P - Q - S \quad (5)$$

This approach was implemented in the package ‘pairsnp’ in c++, python and R for the convenience of users and is available at <https://github.com/gtonkinhill/pairsnp/>. A small extension to this approach allows for the consideration of missing or unknown sites which we include by default in the implementation. Figure 19 indicates that the sparse approach implemented in pairsnp substantially improves the time it takes to find all pairwise SNP distances in a large alignment of 3156 pneumococcal genomes [71]. The implementation is able to take advantage of additional cpu cores further increasing the speed at which results can be achieved.

4.2.2 Transcluster

While SNP distance methods are simple to use and can provide important insights when investigating pathogen transmission, the choice of an appropriate distance threshold is often difficult. By only considering genetic distance these methods ignore other epidemiological information including the times at which the samples were taken. Transcluster, a recent extension to the common SNP distance approach helps to address this issue [468]. Instead of reporting genetic distance, the algorithm attempts to infer the distribution of the number of intermediate hosts separating two infections. This is achieved by

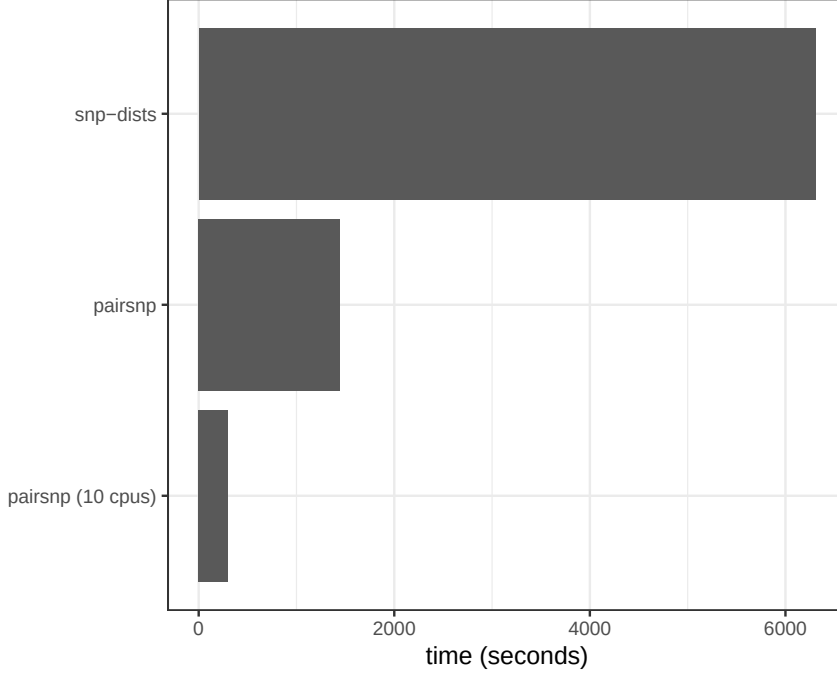


Figure 19: A barchart indicating the time required to calculate all 9.96 million pairwise distances of 3156 pneumococcal genomes from Chewapreecha *et al.*, 2014.

considering the time between samples, the evolutionary rate of the pathogen and the average time between successive infections (generation time). We provide a brief description of this approach as well as a small extension that improves the scalability of method allowing to be applied to datasets comprising over 100,000 sequences.

Let N be the SNP distance separating two genomes and δ the time difference between when the samples were taken. We would like to estimate $h + \delta$, the total time along both lines of descent between the MRCA and the two sample times. The number of SNPs per unit time can be modelled as a Poisson process with evolutionary rate λ . We first decompose N such that $N = N_h + N_\delta$ where N_h represents the number of SNPs that accumulated before sampling the first case and N_δ accounts for any SNPs that accumulated in the subsequently sampled case after the initial sample has been taken (see Figure 20).

Without this restriction N_δ would be Poisson distributed with parameter $\lambda\delta$. After conditioning on the fact that the number of SNPs N_δ can not be greater than the total number of SNPs observed ($0 \leq N_\delta \leq N$) we get the distribution of N_δ as

$$P(N_\delta|\delta) = \left(\frac{e^{-\lambda\delta}(\lambda\delta)^{N_\delta}}{F(N_\delta)} \right) \quad (6)$$

where F is the CDF, $F(N_\delta) = \sum_{i=0}^{N_\delta} e^{-\lambda\delta}(\lambda\delta)^i/i!$

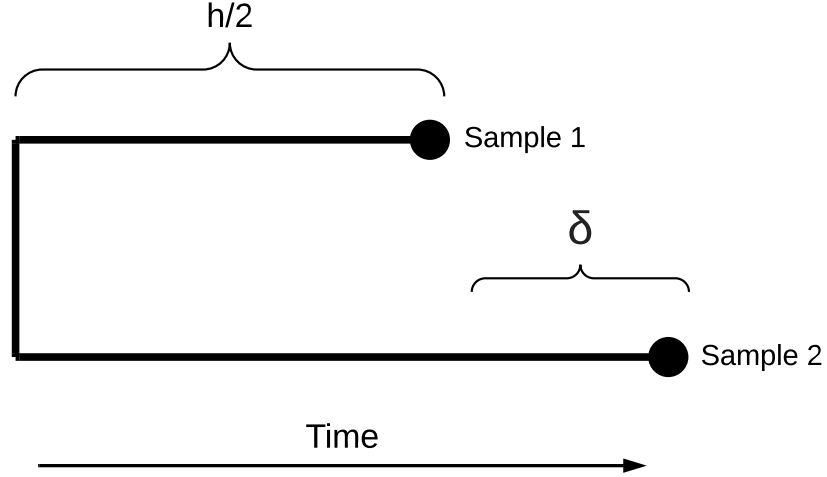


Figure 20: A diagram describing the parameters used in the Transcluster algorithm. δ indicates the time between samples while h indicates the total time along both lines of descent from the MRCA to the time the first sample was taken. Thus, the time between the MRCA and the first sample is $h/2$.

We now consider N_h , the number of SNPs accumulated prior to sampling. This can be modelled as a Poisson process with rate λ and thus the waiting time until the N_h th SNP as described in Stimson *et al.*, [468], can be given by

$$\frac{\lambda^{N_h} h^{N_h-1} e^{-\lambda h}}{(N_h - 1)!} \quad (7)$$

As we know that N_h SNPs have already accumulated we are interested in the waiting time for the $(N_h + 1)$ th SNP. Thus we can write the likelihood of h given N_h as

$$\mathcal{L}(h|N_h) = \frac{\lambda^{N_h+1} h^{N_h} e^{-\lambda h}}{N_h!} \quad (8)$$

To find the likelihood h given N and δ we sum over all possible values of N_δ as

$$\mathcal{L}(h|N, \delta) = \sum_{i=0}^N \mathcal{L}(h|i, \delta) P((N - i)|\delta) \quad (9)$$

which can be expanded as

$$\mathcal{L}(h|N, \delta) = \frac{e^{-\lambda(h+\delta)} \lambda^{N+1}}{F(N)} \sum_{i=0}^N \frac{h^i \delta^{N-i}}{i!(N-i)!} \quad (10)$$

Given h and δ , we now assume that the generation time (the time between transmission events) is parameterised by β , the rate at which the pathogen jumps to a new host. We assume β is constant resulting in another Poisson process for the number of intermediate hosts between the two samples. This can be expressed as

$$P(k|h, \delta) = \frac{\beta^k (h + \delta)^k e^{-\beta(h+\delta)}}{k!} \quad (11)$$

integrating over h gives

$$P(k|h, \delta) = \int_{h=0}^{\infty} \mathcal{L}(h|N, \delta) P(k|h) dh \quad (12)$$

this expression can be expanded as described in Stimson *et al.*, [468], to give

$$P(k|h, \delta) = \frac{e^{-\delta(\lambda+\beta)} \lambda^{N+1} \beta^k}{k! F(N)} \int_{h=0}^{\infty} e^{-h(\lambda+\beta)} (h + \delta)^k \sum_{i=0}^N h^i \left(\frac{\delta^{N-i}}{i!(N-i)!} \right) dh \quad (13)$$

In the original implementation of Transcluster, Equation 13 is solved via numerical integration. Whilst this is suitable for a small number of comparisons when dealing with a very large number of genomes it can become a bottleneck in analysis. Instead, using integration by parts we can express Equation 13 as

$$\frac{\lambda^{N+1} \beta^k (n+k)!}{e^{\delta\beta} n! k! \sum_{i=0}^N \frac{(\lambda\delta)^i}{i!}} \sum_{i=0}^{N+k} \frac{\delta^{N+k-i}}{(N+k-i)!(\lambda+\beta)^{i+1}} \quad (14)$$

As this is a simple sum and as many of the components of this sum will be identical when comparing multiple genome pairs we can make use of efficient caching (storage) of the intermediate results. We have implemented this approach using python and coupled it with pairsnp to enable analysis of very large datasets comprising over 100 thousand genomes. This implementation is available at <https://github.com/gtonkinhill/fasttranscluster>. We describe the application of this implementation in our analysis of SARS-CoV-2 data in Part iii.

4.3 TRANSMISSION NETWORK INFERENCE

Pairwise transmission inference methods are simple to interpret and highly scalable allowing them to be applied to many thousands of samples. However, by only considering pairwise relationships they

do not leverage the polytree structure of transmission chains. A polytree is a Directed Acyclic Graph (DAG) whose underlying undirected graph is a tree in that any two vertices are connected by exactly one path. This structure places constraints on the possible transmission scenarios which is not present when we consider each pairwise interaction independently. As an example, assuming only single infections per host, if host A infects host B which goes on to infect host C, then host C can not have infected host A. In addition to these constraints, reconstructing a phylogeny from sampled sequences distinguishes between those mutations that are shared between multiple isolates and those that are unique to a single branch of the tree. Unique mutations are more likely to be the result of sequencing errors and thus a phylogenetic reconstruction can help to weight the importance of different mutations in transmission inference.

Initial approaches to modelling the full transmission network using genomic data often started with a model of transmission on top of which a genetic model was added [528]. These approaches frequently make significant simplifications, including assuming no within-host evolution and that the branching times of the underlying phylogeny coincide with transmission events.

Notable early examples of models that combined genetic and epidemiological data include the modelling of viral transmission in livestock such as avian influenza (H7N7) between poultry farms in the Netherlands [528] and Foot-and-Mouth Disease Virus (FMDV) in the UK [369]. In both these cases only a single sequence was modelled per host with the former assuming that mutation only occurred at the start of an infection. Simulation studies have shown that ignoring within-host diversity can significantly reduce the accuracy of transmission inference estimates [515, 517]. Other limitations of these approaches include their assumption of a single introduction event and that all cases within the outbreak were observed.

The Outbreaker transmission inference R package addressed a number of these limitations including accounting for both missing cases and multiple independent introductions of a pathogen into a population [249]. However, this approach still assumes that evolution only occurs during transmission and thus ignores the impact of within-host variation. The Outbreaker2 package which superseded this approach provides a modular platform allowing users to easily specify custom models including those that account for within-host evolution [55].

Methods that explicitly account for within-host diversity include those that rely on an initial timed phylogeny [128, 380] described in more detail below, and those that jointly infer both the transmission and phylogenetic tree [115, 129, 206, 265]. While accounting for within-host diversity, a number of these methods including the Phybreak method of Klinkenberg *et al.*, 2017 and the approach of Hall *et al.*,

2015 again assume that all hosts have been sampled. While this may be true in very controlled scenarios such as within hospital wards, it is rarely the case. The fully Bayesian method SCOTTI (Structured COalescent Transmission Tree Inference) accounts for within-host diversity by modelling transmission as the migration of the pathogen between hosts. Using this framework the authors are able to account for both missing hosts and multiple infections of the same host [115]. A limitation of the migration assumption is that SCOTTI does not account for the transmission bottleneck and thus may be more applicable to pathogens that typically have large transmission inocula.

These methods represent important steps forward in using genomic data to inform transmission. However, a number of unresolved challenges remain. One limitation that we have yet to mention, is that the majority of methods tend to focus on outbreaks of acute infections. A notable exception to this is the extension of the approach of Morelli *et al.*, 2012 to endemic disease [365]. However, no code was provided with this publication and similar to the original method it assumes that each host can be represented by a single sequence.

In ‘outbreak’ scenarios all infections are generally assumed to originate from a single introduction, sampling is dense, and the number of new cases over time is highly informative due to the short duration of infection. By contrast, in endemic disease the pathogen is already entrenched often with many origins of infection. High prevalence can also lead to sparse sampling.

In this section we focus on describing the full network inference algorithm TransPhylo [128]. In particular, we describe an extension of the original TransPhylo algorithm of Didelot *et al.*, 2017 to endemic disease. We compare this adapted version to both simulated and real data of *S. pneumoniae* from a refugee camp and demonstrate that accounting for the different features of endemic disease can lead to improvements in transmission inference.

4.3.1 Transphylo

The TransPhylo algorithm of Didelot *et al.*, 2017 attempts to overlay a transmission network onto a precalculated dated phylogeny. In contrast to other methods, the algorithm allows for both missing samples and crucially in the case of endemic disease, is able to account for scenarios in which an ‘outbreak’ is ongoing. The complexity and computational efficiency of the algorithm is improved by separating the problem of transmission inference into the two stages of inferring a timed phylogeny and overlaying a transmission network. This allows for the approach to be used on larger datasets that are infeasible for other ‘one-step’ Bayesian methods that jointly infer both the phylogeny and transmission network. However, unlike these ‘one-step’ methods TransPhylo does not incorporate uncertainty in the initial

phylogeny. Although TransPhylo is able to account for an ongoing and sparsely sampled outbreak the algorithm assumes that a single source seeded the outbreak and that sampling occurs evenly over the outbreak time period. In endemic disease, there is usually considerable evolutionary history within the population prior to sampling. Thus large portions of infections near the root of phylogeny will be unsampled, breaking the assumptions of TransPhylo (Figure 21). For many pathogen outbreaks the generation time (the time between primary and secondary infections) is well understood such as in SARS-CoV-2 where it is thought to be approximately one week [303]. In contrast, many endemic pathogens such as *S. pneumoniae* and *M. tuberculosis* have poorly defined generation intervals. As these pathogens are endemic, their prevalence within a population is known more accurately and thus the probability of sampling an infection can be estimated. Due to the dependence between the probability of sampling an infected host and the generation time of transmission, the TransPhylo algorithm is only able to infer one parameter. The initial implementation provided in Didelot *et al.*, 2017 only allowed for the inference of the sampling probability. To extend the TransPhylo algorithm to endemic disease we adapted the algorithm to estimate the mean generation time as well as allowing for unsampled hosts at the start of an outbreak. In the original publication of Didelot *et al.*, 2017 the authors describe results for both a ‘finished’ and ‘ongoing’ outbreak scenario. Here, we outline the main results of the ‘ongoing’ scenario and leave a more detailed description to the original publication.

Given a timed phylogenetic tree $\mathcal{P}$ the TransPhylo algorithm attempts to infer the transmission tree $\mathcal{T}$ with underlying transmission parameters θ and within-host population size N_e and generation time g modelled as one parameter $N_e g$. The probability of these parameters and the transmission tree can be separated as

$$P(\mathcal{T}, \theta, N_e g) \propto P(\mathcal{P} | N_e g, \mathcal{T}) P(\mathcal{T} | \theta) P(\theta) P(N_e g) \quad (15)$$

The algorithm assumes a stochastic branching process to model the number of secondary infections $\alpha(k)$ for any infected host. This is assumed to follow a negative binomial distribution with parameters (r, p) such that

$$\alpha(k) = \binom{k+r-1}{k} p^k (1-p)^r \quad (16)$$

The generation time $\gamma(\tau)$ is assumed to follow a Gamma distribution with parameters (α, β) such that

$$\gamma(\tau) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \tau^{\alpha-1} e^{-\frac{\tau}{\beta}} \quad (17)$$

Similarly, the time between infection and sampling $\sigma(\tau)$ is assumed to follow a Gamma distribution. A host is said to be ‘included’ if either they or a host downstream from them in the transmission chain is sampled otherwise they are said to be ‘excluded’. Given a uniform sampling probability π we must account for those hosts that became infected just prior to the cessation of sampling. Thus the probability of sampling a host at time t given sampling stopped at time T can be stated as

$$\pi_t = \int_0^{T-1} \sigma(\tau) d\tau \quad (18)$$

The stopping time T also impacts on the probability a host is excluded. Let w be the probability of a host being excluded. We can thus write the probability that an infected individual at time t is excluded as

$$w_t = (1 - \pi_t) \sum_{k=0}^{\infty} \alpha(k) \left(\int_0^{\infty} \gamma(\tau) w_{t+\tau} d\tau \right)^k \quad (19)$$

As we have assumed the offspring distribution to be negative binomial, we can use the generating function of the negative binomial distribution and let $\bar{w}_t = \int_0^{\infty} \gamma(\tau) w_{t+\tau} d\tau$ to give $w_t = (1 - \pi_t) G(\bar{w}_t)$. The TransPhylo algorithm approximates $\bar{w}_t$ using numerical integration. The probability that a host has d offspring that are included can then be written as

$$P(d \text{ included}) = \sum_{k=d}^{\infty} \binom{k}{d} \alpha(k) \bar{w}_t^{k-d} (1 - \bar{w}_t)^d \quad (20)$$

which is approximated in the TransPhylo algorithm by taking the sum to a large k .

The algorithm assumes a transmission tree $\mathcal{T}$ of n nodes representing infected hosts with $s_i = 1$ if host i is sampled and 0 otherwise. The sampling times, infection times and number of included offspring of host i are denoted by t_i^{sam} , t_i^{inf} and d_i respectively. Given the parameters θ the probability of $\mathcal{T}$ can be derived using induction by considering the root ρ of the tree and the resulting subtrees $\{\mathcal{T}_j\}_{j=1..d_p}$ to give

$$P(\mathcal{T}|\theta) = \prod_{i=1}^n \left[(1 - \pi_{t_i^{\text{inf}}})^{1-s_i} (\pi_{t_i^{\text{inf}}} \sigma_t(t_i^{\text{sam}} - t_i^{\text{inf}}))^{s_i} \alpha_{t_i^{\text{inf}}}(d_i) \prod_{j=1}^{d_i} \gamma_t(t_j^{\text{inf}} - t_i^{\text{inf}}) \right] \quad (21)$$

where $\alpha_t(d) = \sum_{k=d}^{\infty} \binom{k}{d} \alpha(k) \bar{w}_t^{k-d}$

To model the probability of the phylogeny $P(\mathcal{P}|\mathcal{T}, N_{eg})$, the TransPhylo algorithm assumes a complete bottleneck at transmission. Consequently, the probability of the phylogeny $\mathcal{P}$ is equal to the product of the probabilities of each subtree which given $\mathcal{T}$ and N_{eg} are calculated under the coalescent with constant population size [263]. Using these definitions, the TransPhylo algorithm performs Bayesian inference of the transmission tree $\mathcal{T}$ given a precalculated phylogeny. To account for the fact that the number of missing hosts is unknown the algorithm explores the parameter space using reversible jump Monte-Carlo Markov Chain (rjMCMC).

4.3.1.1 Adaptations to the TransPhylo algorithm

The original implementation of the Transphylo algorithm treats the parameters of the Gamma distribution defining the generation time as known. In many endemic pathogens such as *S. pneumoniae* the generation time is poorly defined. However, due to the stable rate of infection within communities in endemic disease it is possible to estimate the probability of an infected individual is sampled π . It is not possible to jointly infer both the parameters of the generation time distribution (α, β) and the sampling probability π . As a result, we adapted the algorithm to treat π as known and to instead estimate the parameters of the Gamma distribution. To reduce the complexity of the problem we fixed the shape parameter $\alpha = 1$, which results in an exponential distribution for the generation time. Similar to the treatment of the N_{eg} parameter we used standard Metropolis-Hastings moves to estimate the scale parameter β with an exponential prior.

As mentioned previously the TransPhylo algorithm assumes that sampling occurs evenly over the course of an outbreak. This is inconsistent with the large evolutionary history prior to sampling that is usually present in endemic disease (Figure 21). To account for this, we took advantage of the independence between subtrees in the phylogeny made possible by the assumption of a complete bottleneck. After building a dated phylogeny of a set of samples we cut the phylogeny at the date sampling started. This produces a number of subtrees for which the original assumptions of the TransPhylo algorithm hold. By enforcing that the parameters θ are the same across the subtrees we can improve our statistical power to infer them accurately. Thus, we adapted the TransPhylo algorithm to accept a list of dated phylogenies as input. The same MCMC moves are used as in the original version of the algorithm with the exception of the proposal for moving events on the transmission tree $\mathcal{T}$. Here, during each MCMC move a random subtree is chosen uniformly and the original proposal is then used to suggest a move on the corresponding sub-transmission tree.

4.3.1.2 *Simulation*

To generate realistic simulated data we created a forward simulation algorithm by coupling both a simple stochastic susceptible-infectious-susceptible (SIS) model to simulate a transmission network using the Gillespie algorithm with the efficient forward genetic evolution package FFPopSim [533]. FFPopSim is written in c++ and collapses identical genotypes into clones, improving the speed and scalability of the method which allowed us to simulate transmission networks involving hundreds of hosts whilst accounting for within-host diversity.

To simulate the underlying transmission network we assumed an SIS model with recovery rate λ , infection rate α , total simulation time t_{sim} and host adjacency matrix A where $0 \leq A_{ij} \leq 1$, is an indication of the strength of the contact between hosts i and j . $A_{ij} = 0$ indicates that there is no contact between the two hosts and direct transmission can not occur. The status of each host is given by

$$S_i = \begin{cases} 1, & \text{if host } i \text{ is infected} \\ 0, & \text{if host } i \text{ is susceptible} \end{cases} \quad (22)$$

The algorithm proceeds as outlined in the following pseudocode where H_e , I_e and T_e are vectors indicating the host, infector and time of event e in the transmission chain respectively.

Algorithm 1: Stochastic SIS model

```

input :  $A, S, \lambda, \beta$ 
output:  $H, I, T$ 
 $e \leftarrow 1$ 
while  $t < t_{\text{sim}}$  do
   $r \leftarrow \sum_{i=0}^N \alpha \mathbb{1}_{S_i=0} + \sum_{i=0}^N \lambda \mathbb{1}_{S_i=1}$ 
  sample a single host  $i$ , w.p  $(\alpha \mathbb{1}_{S_i=0} + \lambda \mathbb{1}_{S_i=1}) / r$ 
   $H_e \leftarrow i$ 
   $T_e \leftarrow t$ 
  if  $S_i = 0$  then
     $S_i \leftarrow 1$ 
     $I_e \leftarrow$  sample a host  $j$  uniformly, where  $S_j = 1$  and  $j \neq i$ 
  else
     $S_i \leftarrow 0$ 
     $I_e \leftarrow \text{NA}$ 
  end
   $t \leftarrow t + \tau$  where  $\tau \sim \text{Exp}(r)$ 
   $e \leftarrow e + 1$ 
end

```

Given a network generated using the stochastic SIS model, we simulate the genetic evolution of the underlying pathogen population using FFPopSim. This is a discrete time forward simulation and thus the time steps between successive generations of the pathogen is given as

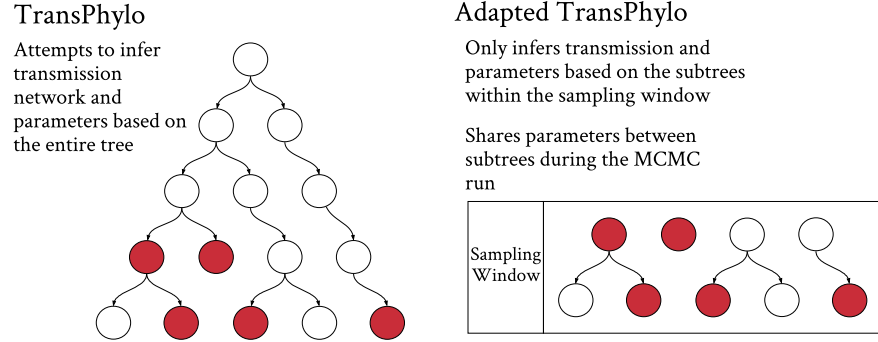


Figure 21: A diagram indicating the difference between the original TransPhylo approach and the adapted method. After cutting at the date sampling started, transmission chains are simultaneously inferred on the resulting sub-trees by sharing the inferred model parameters.

input. Similar to the coalescent simulation the mean bottleneck size, sample size, evolutionary rate, recombination rate and within-host population size are also given to the algorithm. Due to the differences between coalescent and forward simulation the population size parameters are not directly comparable.

4.3.1.3 The Maela *S. pneumoniae* dataset

To investigate the algorithm's ability to infer transmission in real data we clustered 3,085 genomes from a major genomic carriage study of *S. pneumoniae* in the Maela refugee camp in Thailand [71]. Clustering was performed using the hierBAPS algorithm resulting in 36 major clusters [70]. For each cluster a phylogeny was constructed using fasttree2 after first removing recombination with Gubbins [104, 406]. The phylogenies were dated using the 'node.dating' function in the R package 'ape' which implements a fast maximum likelihood method similar to the approach described by Felsenstein (1981) [165, 251]. These phylogenies were then cut at the date sampling started in 2007. The resulting parameter estimates could be compared to previous studies of *S. pneumoniae* carriage and transmission. As household location data was available for the majority of the sampled genomes within the refugee camp the inferred transmission links could also be compared with household geographic distance.

4.3.2 Results

4.3.2.1 Simulations

Initially, we compared the results of TransPhylo, SCOTTI, Outbreaker2 and the adapted TransPhylo algorithm with simulated data. We simulated a simplified situation similar to the Maela camp based on an

SIS model, using FFPopSim to model the within-host diversity. The mean carriage duration was drawn from an exponential distribution with a mean of eight weeks. A genome size of 1,000nt was simulated with each host having a carrying capacity of 1,000 genomes. Recombination was not simulated and the mutation rate was set separately to 5.3, 20 and 100 mutations per genome per year, with a reproduction generation time of one hour. A transmission rate of 0.2 transmissions per week was used. The simulation was run without sampling for the equivalent of one year in order for the system to reach equilibrium. Samples were then taken at one week time steps in the simulation over a six month interval. At each time step an infected host was sampled with probability 0.05.

To compare the results of running the methods on simulated data we calculated the transmission distance between each simulated sample. Transmission distance is the number of intermediate hosts between two samples. Thus, in a direct transmission between two hosts the transmission distance will be zero. As indicated in Figure 22, Outbreaker2 performed worst, which can be attributed to its inability to model within-host diversity. The adapted TransPhylo algorithm performed the best out of all four methods tested. It should be noted that SCOTTI performed well without customisation. However, its running time was an order of magnitude longer than the TransPhylo methods. Interestingly, SCOTTI struggled with very low mutation rates, which could be due to issues with convergence of the MCMC.

4.3.2.2 *Transmission of S. pneumoniae in a refugee camp*

The adapted TransPhylo algorithm was then applied to each major cluster from the WGS Maela data set of [71], after first clustering using hierBAPS [70]. R_e was set to 1 by specifying appropriate parameters for the offspring distribution. The mean time between infection and sampling was set to 2 weeks as samples were collected monthly. The mean epidemiological generation was inferred from the data after fixing the probability of sampling an infected individual to the proportion of children under 5 that were sampled in the camp. Figure 23 indicates the distribution of close transmission (< 5 inferred intermediate hosts) with respect to geographic distance between the homes of the sampled patients. The high concentration of close transmissions at short geographic distances suggests the algorithm is working. A transmission distance of < 5 was chosen as it corresponded to a peak in the distribution of transmission distances. This threshold corresponds with a SNP distance of approximately 10 SNPs which is similar to previous studies [56]. Finally, the inferred mean epidemiological generation time for each major BAPS cluster is shown in Figure 24. The majority of clusters have a mean generation time of approximately 2 months which is consistent with a uniform transmission rate over previous estimates of carriage duration [287]. Clusters

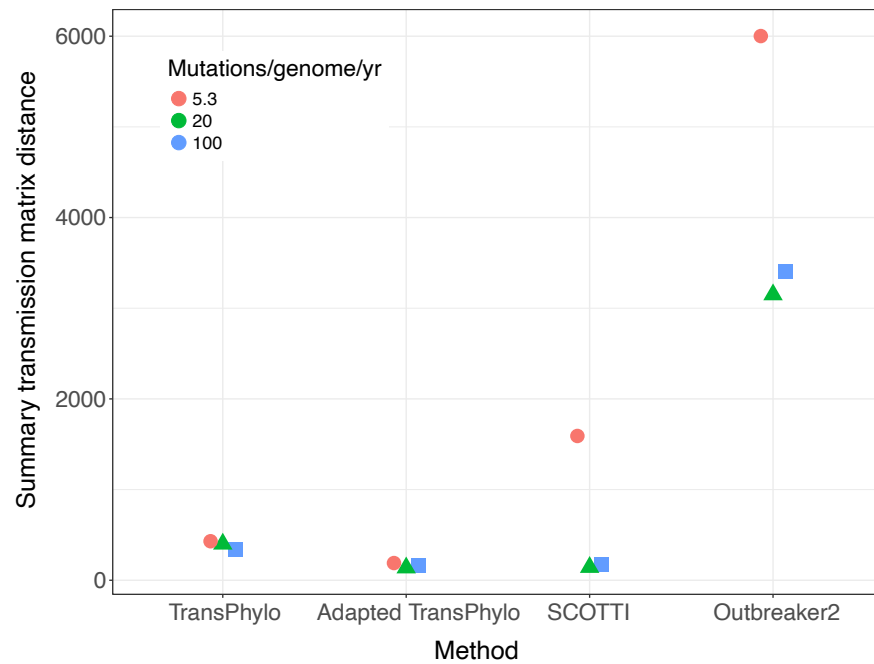


Figure 22: The results of running SCOTTI, Outbreaker 2, TransPhylo and the adapted TransPhylo algorithm on simulated data. The inferred transmission distance between samples (number of intermediate infections) is compared to the simulated results. The adapted algorithm performs best. However, SCOTTI performs well without modification except for lower mutation rates. This could be an issue with convergence of the MCMC. The TransPhylo based methods were significantly faster than SCOTTI.

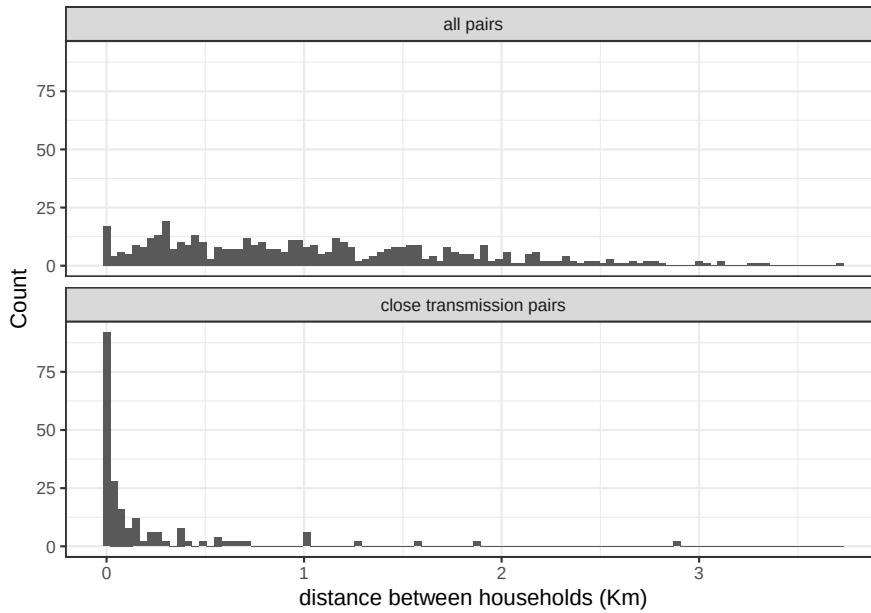


Figure 23: A comparison between the distribution of all pairwise geographic distances in the Maela data set and those found to be within a transmission distance of 5 intermediate hosts. The high concentration of inferred transmission at short geographic distances suggests the method is working.

15, 23 and 31 have very high estimates for the mean generation time. These clusters were found to have poor temporal signal which could explain the discrepancy. The results suggest the adapted TransPhylo algorithm was a suitable choice for identifying transmissions. The resulting close transmission links were used to aid in the selection of samples for the large deep sequencing pneumococcal study described in [Part ii](#).

4.4 TRANSMISSION INFERENCE WITH DEEP SEQUENCING

To this point, we have concentrated on methods that rely on a single pathogen genomes per infection for modelling transmission. [Figure 25](#) indicates the valuable additional resolution that information on within-host genetic diversity can provide for the inference of transmission. Given a phylogeny built from single genomes from three infected hosts, there are five possible transmission scenarios that are concordant with the tree. Given sufficient within-host diversity we can determine the true transmission chain by obtaining additional pathogen genomes from each of the hosts. Obtaining multiple genomes per infection can be prohibitively expensive. Instead, many studies attempt to gain information on within-host diversity by pooling pathogen genetic data from a sample and using deep sequencing to obtain a high level of coverage across the genome. This allows

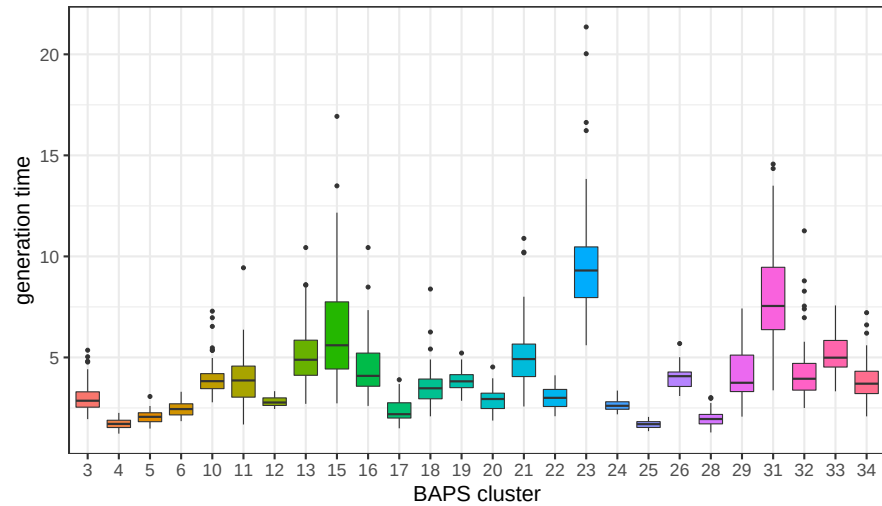


Figure 24: Boxplots indicating the distribution of the inferred mean epidemiological generation time of each BAPS cluster. Clusters 15, 23, and 31 had poor temporal signal which could have impacted the results.

for the identification of heterozygous nucleotide base calls at positions within the genome which provide an indication of the diversity present within the host. Generally, short read sequencing is used in these studies which represents an additional challenge as the genetic linkage between loci in the genome is obscured.

One of the simplest methods for inferring putative transmission links using deep sequencing data is to consider the number of shared variants between two samples. By taking the consensus at each site along the genome, a consensus SNP distance can also be calculated to supplement the information provided by minority variants [516]. This approach is sometimes referred to as shared variant clustering (SVC) and has been used to investigate outbreaks of both Ebola and *M. tuberculosis* [330, 516]. In the analysis of an *M. tuberculosis* outbreak amongst the Canadian inuit population, deep sequencing was able to identify a novel ‘super-spreader’ thought to have infected 17 others [286]. While such within-host distance based methods have been effective they retain many of the limitations of SNP distance based approaches. The lack of a realistic model of pathogen evolution limits the resolution of the method and as distance based approaches are usually symmetric they are unable to identify the direction of transmission.

As an alternative, De Maio and colleagues proposed a Bayesian model of transmission (BadTriP) [114]. The approach relies on a polymorphism aware phylogenetic model (PoMo) that describes the evolution of the pathogen population within each host and assumes that sites within the genome are unlinked [112]. The model based approach incorporates both the signal provided by within-host variants

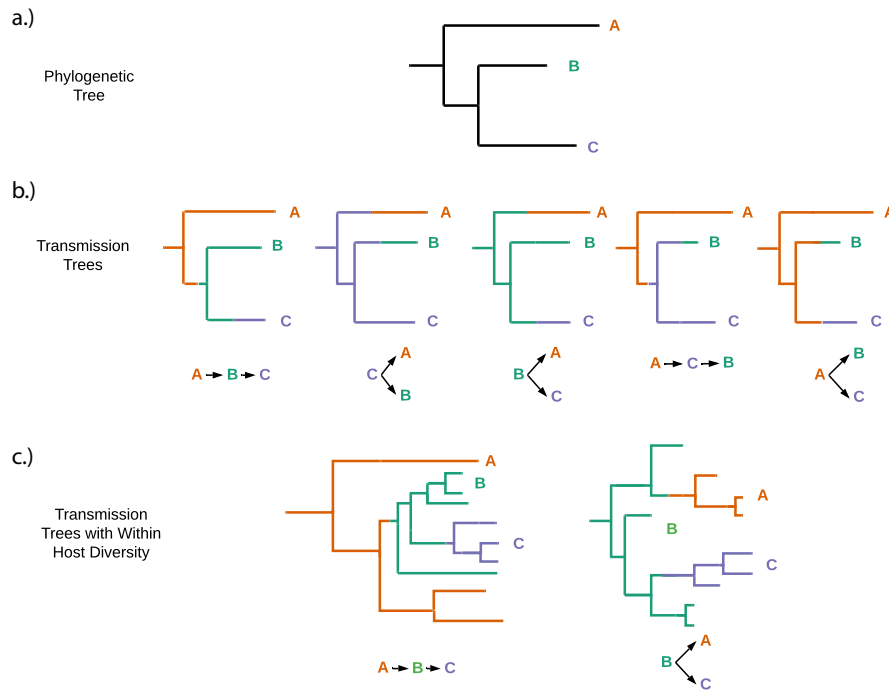


Figure 25: (a) The figure indicates a phylogeny built from three genomes sampled from separate hosts. (b) The five possible transmission scenarios concordant with (a) are indicated by colouring the phylogeny by the host in which that stage of the branch occurred. (c) Additional genomes from each host indicate the nested relationship between hosts from different genomes allowing for the direction of transmission to be determined.

and allows for multiple samples per host. It is able to incorporate additional epidemiological data such as exposure and sampling time information. Consequently, the method is able to estimate both the direction of transmission as well as the timing of transmission events. It also provides estimates of the within-host evolutionary rate and the transmission bottleneck size. A limitation of the approach is its assumption that all cases within the transmission chain have been sampled or are at least known. The algorithm is also computationally expensive as it attempts to infer the full transmission chain which limits its application to datasets with less than 100 samples [114]. As such, it is not well suited to the analysis of large sparsely sampled datasets from an endemic pathogen.

In contrast to the BadTriP algorithm which assumes alleles are unlinked, PHYLOSCANNER is an alternative algorithm, which investigates windows within the genomes that can be entirely spanned by sequencing reads [523]. By building phylogenies within these windows the algorithm is able to combine the phylogenetic signal between windows to give an estimate of the underlying transmission network. PHYLOSCANNER is able to account for multiple infections and contamination in pre-filtering steps. Unlike BadTriP, PHYLOSCANNER does not make use of sample times and does not provide an estimate of the transmission bottleneck size. The window based approach has proven to be effective in the analysis of HIV transmission chains where the long carriage duration and high within-host evolutionary rate lead to many phylogenetically informative windows, even within the length of a single read [417]. However, in order to build a phylogeny within a window multiple variant sites must be present. When considering large bacterial genomes with lower evolutionary rates this is unlikely to be the case given short read data [56].

4.4.1 *A simulation based alternative*

Many of the simplifying assumptions made by the methods described so far are motivated by the complexity of modelling transmission using short read deep sequencing data. This makes defining an accurate statistical model and calculating the corresponding likelihood difficult. In many cases, even when the likelihood function can be defined its calculation is intractable. In contrast, it is generally much easier to create realistic simulations of transmission that account for within-host diversity. As a result we explored the application of recent advances in simulation based likelihood-free inference techniques to the problem of inferring transmission using deep sequencing data.

4.4.2 Likelihood-free inference

Problems in a diverse range of fields including astronomy, ecology, epidemiology and genetics are often best modelled via simulation. Despite this, the probability density that they implicitly define is frequently intractable. This has led to the development of techniques that rely on simulation to develop a surrogate for the likelihood function. The most common strategy for likelihood-free inference is approximate Bayesian computation (ABC) [475]. While many variations of ABC exist, they typically rely on comparing simulations to data via a distance function. The ABC rejection algorithm is one of the simplest versions. It starts by sampling simulation parameters from a prior distribution. A simulation is then run and the summary statistics of the resulting simulated data are compared to the real data via a distance function d . If this distance is less than a predefined threshold $\epsilon \geq 0$ the set of parameters is accepted [475]. While this approach can alleviate difficulties with intractable likelihoods, it is computationally expensive. This has motivated the development of more efficient techniques such as those based on regression [25], Markov Chain Monte Carlo (MCMC) [327] and Sequential Monte Carlo (SMC) [453]. These methods can lead to large improvements in the efficiency of ABC algorithms. However, they normally rely on the simulation being run many thousands to millions of times. This can present a challenge when the simulation itself is computationally expensive and has led to the development of methods that make use of Bayesian optimisation to dramatically reduce the number of simulations required [203].

ABC likelihood-free methods usually require both the choice of a distance function as well a tolerance ϵ although methods have been developed to adapt the tolerance as the algorithm progresses [418]. In parallel to the development of ABC based methods and to resolve some of these issues, synthetic likelihood methods have been proposed [514]. These assume that some well chosen summary statistics are approximately multivariate Gaussian for each value of the parameters. Using this assumption a Gaussian likelihood is constructed using simulation to estimate the mean and covariance matrix of the summary statistics. Similar to ABC, synthetic likelihood methods have since been extended to incorporate more advanced Bayesian techniques such as the Metropolis-Hasting algorithm and Sequential Monte Carlo (SMC) [405]. Approaches have also been developed to reduce the number of evaluations of the synthetic likelihood when the simulation itself is computationally expensive [203, 246].

The loss of genetic linkage information as the result of short read sequencing presents a significant challenge to obtaining a realistic model with a tractable likelihood for pathogen transmission. This has led to approximations such as the PoMo model used in BadTrIP which assumes no genetic linkage and approximates the transmission

bottleneck by additional branches that evolve over zero time, making interpretation of the resulting parameter difficult [114]. Instead, we investigate the use of a simulation based strategy to infer pathogen transmission parameters. This allows for a more realistic and interpretable model of transmission. Furthermore, as we do not rely on a derived likelihood, extensions and adjustments to the model can be made easily.

4.4.3 Synthetic likelihood

Unlike ABC methods which use a distance function and a threshold ϵ to compare summary statistics between real and simulated datasets, synthetic likelihood approaches assume that the summary statistics are approximately Gaussian conditional on the set of parameters θ [514]. Given data y , an intractable likelihood $p(y|\theta)$ and summary statistics $S = S(y)$ the likelihood for the summary statistics is assumed to follow a multivariate normal distribution with mean $\mu(\theta) = E(S|\theta)$ and covariance matrix $\Sigma(\theta) = \text{Cov}(S|\theta)$. As $\mu(\theta)$ and $\Sigma(\theta)$ are generally unknown we can estimate them via simulation as

$$\hat{\mu}(\theta) = \frac{\sum_{i=1}^N S_i}{N} \quad (23)$$

$$\hat{\Sigma}(\theta) = \frac{1}{N-1} \sum_{i=1}^N (S_i - \hat{\mu}(\theta))(S_i - \hat{\mu}(\theta))^T \quad (24)$$

Thus, in contrast to ABC methods, calculating the synthetic likelihood depends only on the number of replicate simulations of the model to evaluate the likelihood for each set of parameters θ . We then approximate the likelihood as

$$p(y|\theta) \approx \phi(s|\mu(\theta), \Sigma(\theta)) \approx \phi(s|\hat{\mu}(\theta), \hat{\Sigma}(\theta)) \quad (25)$$

4.4.4 Bayesian optimisation

Bayesian optimisation attempts to maximise (or minimise) the output of a blackbox objective function by modelling the inputs and outputs using a Gaussian Process [362]. The posterior distribution of the Gaussian Process is used to iteratively select new points within the parameter space at which to evaluate the objective function. Bayesian optimisation is most useful in applications where the blackbox function is difficult or expensive to evaluate. The method's efficiency in exploring parameter space has made it an appealing strategy for likelihood-free inference approaches where the simulations themselves are difficult to evaluate [203, 246].

4.4.4.1 Gaussian Processes (GPs)

A Gaussian Process is a collection of a possibly infinite number of random variables such that the joint distribution over any finite subset is multivariate Gaussian defined by a mean $\mu(x)$ and positive definite covariance matrix $\Sigma(x, x')$ [465]. Equivalently, a prior is placed over the possibly infinite set of finite dimensional functions defined by a mean and covariance matrix. The ‘smoothness’ of this prior can be controlled by the choice of covariance function. A common choice is the ‘squared-exponential’ $K_{SE}(x, x') = \exp\left(\frac{-|x-x'|^2}{2\ell^2}\right)$ where ℓ is the length scale of the process, controlling how close two points need to be to influence each other. As is done in the laGP R package we add a ‘nugget’ to the kernel such that $K_{SE}(x, x') = \exp\left(\frac{-|x-x'|^2}{2\ell^2}\right) + \eta \mathbb{1}_{x=x'}$ which has been shown to improve robustness and smooth noisy data [197, 198].

4.4.4.2 Expected improvement

In Bayesian Optimisation, the acquisition function determines which point in parameter space will be evaluated next. There is a trade off between exploitation (sampling near the maximum of the objective function) and exploration (sampling in regions of high uncertainty). A number of strategies exist including entropy search, upper confidence bound and probability of improvement [121, 218, 279]. In our approach we make use of the ‘Expected Improvement’ acquisition function [363] which attempts to identify the parameters θ to maximise the objective function f such that

$$EI(x) = \mathbb{E}[\max(f(f) - f(x_{\max}))] \quad (26)$$

A simple outline of a Bayesian Optimisation strategy using expected improvement as the acquisition strategy is outlined in [algorithm 2](#). Initial training points $\mathbf{x}_t$ can be chosen by sampling uniformly from the set of possible parameters, sampling from the prior or using grid based strategies such as generating a Latin hypercube.

These are then evaluated using the objective function to generate possibly noisy samples y .

Algorithm 2: Bayesian Optimisation

```

N ← total number of iterations
Ninit ← number of initial training points
for t ∈ 1, ..., N do
  if t ≤ Ninit then
    |  $\mathbf{x}_t \leftarrow$  select from latin hypercube
  else
    |  $\mathbf{x}_t \leftarrow \operatorname{argmax}_x \operatorname{EI}(x|D_{1:t-1})$ 
  end
   $y_t \leftarrow f(\mathbf{x}_t) + \epsilon_t$ 
   $\mathcal{D}_{1:t} \leftarrow \mathcal{D}_{1:t-1}, (\mathbf{x}_t, y_t)$ 
end

```

4.4.5 Importance resampling

So far we have discussed methods for approximating the likelihood function using simulations. As with traditional likelihood functions we would often like to incorporate information about our prior beliefs of the likely parameter distributions. The posterior distribution in terms of the parameters θ , data y , likelihood $p(y|\theta)$ and prior $p(\theta)$ can be written as

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_{\theta' \in \Theta} p(y|\theta')p(\theta')d\theta'} \quad (27)$$

This is often intractable and thus given the approximate likelihood function we can use traditional Bayesian inference techniques such as rejection sampling, the Metropolis-Hastings algorithm or Gibbs sampling to provide estimates of the posterior distribution. In our case we implement importance resampling to account for the prior distribution of each parameter. We first describe importance sampling before detailing an extension that allows for sampling from the posterior distribution.

Suppose we are interested in the expected value of $h(\theta)$ but can not directly sample from $p(\theta|y)$ but can instead sample from a distribution $g(\theta)$. Then,

$$E(h(\theta)) = \frac{\int h(\theta)p(\theta|y)d\theta}{\int p(\theta|y)d\theta} = \frac{\int h(\theta) \frac{p(\theta|y)}{g(\theta)} g(\theta)d\theta}{\int \frac{p(\theta|y)}{g(\theta)} g(\theta)d\theta} \quad (28)$$

This can be estimated by drawing N times from $g(\theta)$ as

$$\frac{\frac{1}{N} \sum_{i=1}^N h(\theta^i) w(\theta^i)}{\frac{1}{N} \sum_{i=1}^N w(\theta^i)} \quad (29)$$

where

$$w(\theta^i) = \frac{p(\theta^i|y)}{g(\theta^i)} \quad (30)$$

are the importance weights.

This idea can be extended to approximate drawing from the posterior distribution by drawing N samples from the prior distribution $g(\theta)$. A subset of the N samples are then resampled with probability proportional to its importance weight which reduces to the likelihood as

$$w(\theta^i) = \frac{p(\theta^i|y)}{p(\theta^i)} = \frac{p(y|\theta^i)p(\theta^i)}{p(\theta^i)} = p(y|\theta^i) \quad (31)$$

4.4.6 Simulating transmission

To use likelihood free techniques we require a realistic simulation model of pathogen transmission. The number of possible transmission networks equals the number of possible trees which given n nodes is given by Cayley's formula as n^{n-2} [61]. This complexity makes inference of all but the smallest transmission networks intractable to infer using likelihood free techniques. Instead we focus our attention on pairwise relationships and only simulate a large transmission network to test the accuracy of our algorithm. Given samples from two infected hosts, there are three overriding transmission scenarios as shown in Figure 26. Either one host can infect the other possible via a number of intermediate hosts, the inverse can occur or alternatively an unsampled host can have infected both sampled hosts separately, possibly via intermediate hosts. To simulate the evolution of genetic sequences in each of these scenarios we make use of the approximated coalescent with recombination (scrm) [463]. This approximation allows us to simulate each scenario within a reasonable amount of time for use in a likelihood-free inference framework whilst allowing for recombination between genomes within each host.

4.4.6.1 Pairwise simulation

The parameters required for the simulation can be split into two broad categories (Figure 26). Local parameters are specific to each pairwise combination of hosts while global parameters are common to all pairs. Given pair j , local parameters include the times when hosts 1 and 2 are infected (t_{1j} and t_{2j}). The time the pathogen population could last be considered to be a single population (t_{mj}) and the number of unsampled intermediate hosts separating the two samples

(d_j). The number of samples of host 1 and 2 are given as $n_{s_{1j}}$ and $n_{s_{2j}}$ respectively with the times of sample k in host i given by s_i^k . Global parameters common to every pairwise relationship include the average transmission bottleneck size β , the evolutionary rate μ , the recombination rate r , the mean sample size v , and the effective population in a host N_e . Given these parameters, a pairwise transmission scenario is simulated using *scrm* by simulating d bottlenecks of size β at times chosen uniformly in the interval $\{t : t < t_1, t < t_2, t > t_m\}$. Bottlenecks are also simulated at t_1 and t_2 to account for the bottleneck of the transmission to the sampled host. A population split is simulated at time t_m .

While it is possible to use the same simulation algorithm to perform both the inference and assess the accuracy of a likelihood-free inference approach, in this case the pairwise simulation is an oversimplification of the underlying transmission problem. To generate simulated datasets for testing we used the full network simulation approach described in [Section 4.3.2.1](#).

4.4.7 A synthetic likelihood approach to transmission inference

Although the pairwise coalescent simulation is significantly faster than simulating a full transmission network using forward simulation, the time taken per simulation is still not trivial. This leads us to consider approaches similar to the likelihood-free Bayesian optimisation methods of Gutmann and Corander, 2016. The effects of mutation, recombination, transmission and sampling bottlenecks leads our simulator to produce highly stochastic results even when given the same set of input parameters. To overcome this difficulty we make use of a synthetic likelihood based approach using a large number of simulations for each set of parameters. We then explore this parameter space using Bayesian optimisation. As our approach is based on simulation and was initially designed for the inference of transmission in endemic disease contexts we dubbed it ‘endsim’.

Given the global parameters, each pairwise model is treated independently. Consequently, to allow much of the inference algorithm to be run in parallel and to reduce the dimension of the fitted GPs we separate the problem by iteratively solving for the global and local parameters. Given a set of initial values for the local parameters, the posterior distribution of the global parameters is found using the Bayesian optimisation synthetic likelihood approach. The posterior distribution of the local parameters is then inferred using the same method where the simulations are run using global parameters sampled from the previously inferred global posterior distribution. These two steps can then be repeated to improve the estimation of the respective global and local posterior distributions. In practice, we have yet to thoroughly explore the impact of different numbers of itera-

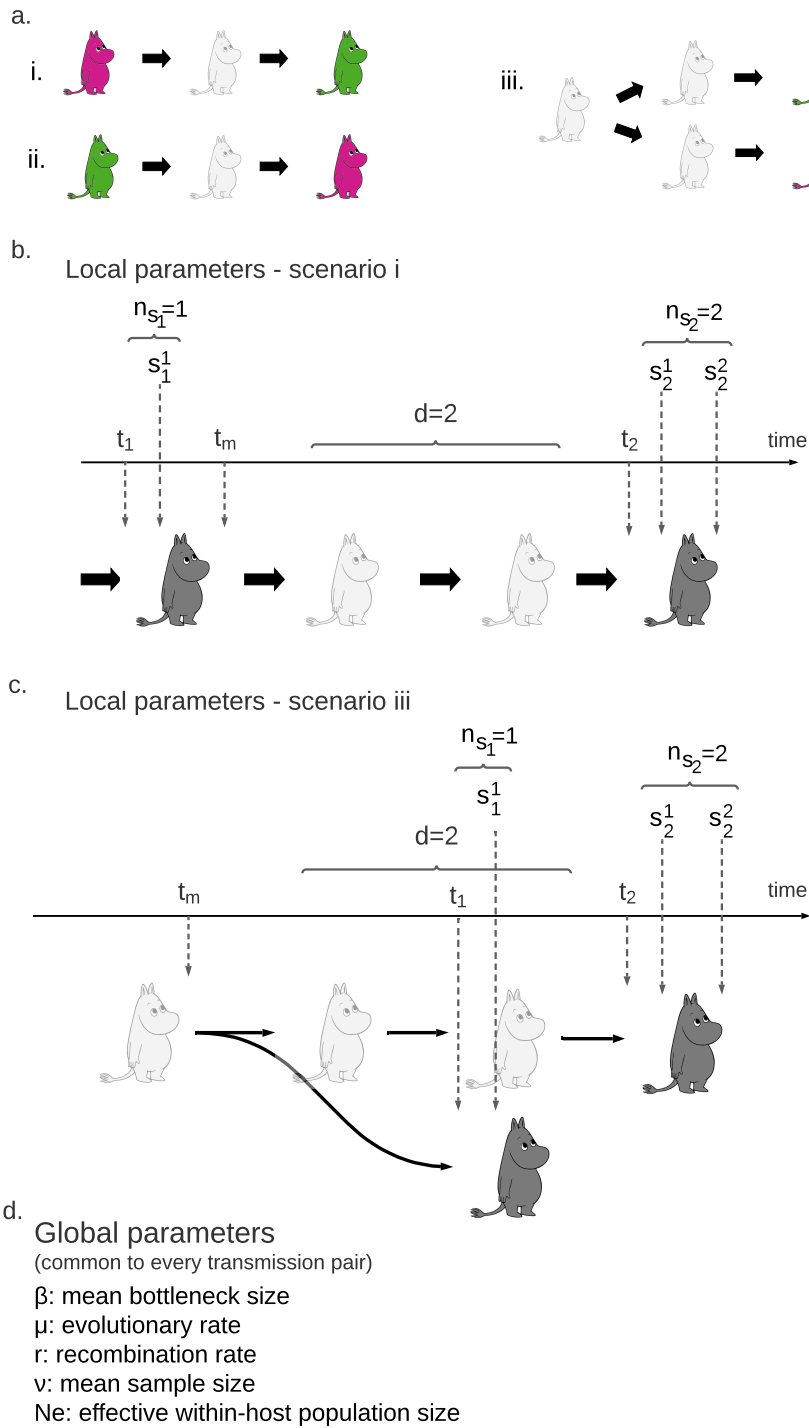


Figure 26: a.) The three possible transmission scenarios when considering two hosts. i.) Either one host can infect the other, possibly through intermediate hosts. ii.) the reverse can occur or iii.) an unsampled host can infect both sampled hosts possible via additional unsampled hosts. b. & c.) Diagrams indicating the local parameters involved in the simulation of scenarios i and iii. d.) A list of the global parameters assumed to be the same in each pairwise interaction.

tions yet have found that even one iteration is sufficient to provide local parameter estimates that improve over previous transmission inference algorithms.

Let N be the number of transmission pairs and let T_{ij} and S_{ij} be vectors of sample times and summary statistics for host i in pair j . Fixed parameters for the algorithm include the number of iterations to evaluate the synthetic likelihood I_s as well as the initial number of points I_p and number of iterations I_r used in the Bayesian optimisation stage of the algorithm. To improve the computational efficiency of the algorithm we restrict the range of the inferred parameters such that: $T_{ij} - t_{ij} \leq t_{\max}$ and $0 \leq d_j \leq d_{\max}$. The global parameters are also restricted to reasonable ranges and can be fixed if known.

The parameters are initialised by sampling from their respective prior distributions. An initial grid of I_p points for evaluation is chosen by generating a Latin hypercube from draws from a set of uniform distributions over the valid ranges for each parameter using the ‘maximin’ function from the lhs R package [466].

The log likelihood, local and global parameters for the k th evaluation of the synthetic log likelihood are given by ℓ_k , G_k and L_k

respectively and the summary statistics of the dataset is denoted D . The algorithm proceeds as outlined in [algorithm 3](#).

Algorithm 3: Likelihood free transmission inference

```

 $L \leftarrow$  sample from prior
 $G_{1..I_p} \leftarrow$  select  $I_p$  points from Latin hypercube
for  $k$  in  $1, \dots, I_p$  do
  |  $\ell_k \leftarrow$  SyntheticLLK( $G_k, L, D$ )
end
Build GP with  $G_{1..I_p}, \ell_{1..I_p}$ 
for  $k$  in  $I_p + 1, \dots, I_p + I_r$  do
  |  $G_k \leftarrow$  MaxExpectedImprovement( $\ell_{1..,k-1}, G_{1..,k-1}$ )
  |  $\ell_k \leftarrow$  SyntheticLLK( $G_k, L, D$ )
  | Update GP with  $G_k, \ell_k$ 
end
Posterior distribution of  $G \leftarrow$ 
  ImportanceReSampling(GP, prior on  $L$ )
 $L_{1..,I_p} \leftarrow$  select  $I_p$  points from latin hypercube
 $G \leftarrow$  sample from posterior inferred in Stage 1
for  $k$  in  $1, \dots, I_p$  do
  |  $\ell_k \leftarrow$  SyntheticLLK( $G, L_k, D$ )
end
Build GP with  $L_{1..I_p}, \ell_{1..I_p}$ 
for  $k$  in  $I_p + 1, \dots, I_p + I_r$  do
  |  $G_k \leftarrow$  MaxExpectedImprovement( $\ell_{1..,k-1}, L_{1..,k-1}$ )
  |  $\ell_k \leftarrow$  SyntheticLLK( $G, L_k, D$ )
  | Update GP with  $L_k, \ell_k$ 
end
Posterior distribution of  $L \leftarrow$ 
  ImportanceReSampling(GP, prior on  $L$ )

```

4.4.8 Results

The work on this method is ongoing but initial results indicate that the approach is able to improve over alternative methods. To investigate the ability of the method to accurately infer transmission parameters we simulated a full endemic transmission scenario using forward simulation as described in [Section 4.3.2.1](#). The simulation was run from a single introduction to the equivalent of 10 years of evolutionary time with an evolutionary rate equivalent to 1 SNPs/genome/year. This is a relatively low mutation rate but comparable to many important pathogens such as *M. tuberculosis* (0.04–2.2 SNPs/genome/year) [351] and *S. pneumoniae* (2.90–9.62 SNPs/genome/year) [71]. To mimic sampling in an endemic scenario we only sampled from the simula-

tion during the final year of the ‘outbreak’. As a result, many of our samples are separated by tens of intermediate transmissions events.

While such a low mutation rate might preclude the use of traditional consensus based transmission inference techniques, by accounting for the evolution of the population within the host we are able to distinguish different routes of transmission. For example, if a within-host population consisted of 10,000 bacteria then there would be 10,000 times the genetic material from which mutations could occur. Thus, if we ignored the impact of selection and drift, a population of this size could generate a number of *de novo* mutations four orders of magnitude greater than that seen in a single genome. In reality, many of these mutations are quickly lost. Under a tight transmission bottleneck the amount of diversity present within a sample is then an indicator of the length of time since the transmission. Our simulation based inference approach allows us to use this to help determine the likely direction of transmission as described in [Figure 26](#).

Only a limited number of existing tools are suitable for the analysis of short read deep sequencing pathogen data. As we are simulating a slowly evolving pathogen population the distance between SNPs on the genome is nearly always greater than the common length of a NGS sequencing read (150bp). In fact, in the majority of our simulations, zero SNPs were within this range. Consequently, the approach adopted by PHYLOSCANNER, which relies on linkage information between SNPs is not applicable here and the tool simply rejects the data from our simulations [523]. While BadTrIP and the shared variant (SVC) method of Worby *et al.*, 2017, assume that all individuals have been sampled, these methods do accept unlinked variant data and can be used to infer the direction of transmission. Our simulation based inference approach only considers pairwise interactions. To compare with the full network inference results of BadTrIP and SVC we only consider the subset of pairwise relationships whose genetic distance are in the top 5% of all pairwise relationships. As well as allowing for a direct comparison this reduces the impact of ignoring missing samples in the BadTrIP and SVC approaches, as any discrepancies found in more distant pairs are ignored and thus this should lead to an improvement in their accuracy. The computational complexity of the BadTrIP algorithm prevents us from evaluating it using our full simulated sample sets. As a result we compare the full simulated sample set only against the SVC approach. To compare against BadTrIP we randomly sampled 10 transmission pairs from each simulated sample set in the full comparison ([Figure 27b](#)).

[Figure 27a](#) indicates the resulting accuracy of the simulation based inference approach (endsim) and SVC across nine different scenarios where three independent transmission networks have been simulated for each scenario. The endsim approach improves over SVC in 5/9 scenarios indicating that more accurately modelling the underlying

transmission scenario can provide substantial improvements in some scenarios. The scenarios for which the SVC approach provides more accurate results are those where the transmission bottleneck is large. In this case, the transmission event has little impact on the underlying within-host population. Our method relies on accurately determining the timing of transmission and thus if the bottleneck signal is weak it struggles to determine transmission direction. The SVC method simply compares the number of shared within-host variants and thus by modelling the full transmission network is able to take advantage of the large number of shared variants that occur under a loose transmission bottleneck. A similar result was observed in the comparison of BadTrIP on the subsampled datasets shown in [Figure 27b](#). Here, endsim outperformed BadTrIP and SVC in the tight bottleneck scenarios but was outperformed consistently by BadTrIP when the bottleneck size was large. In reality for many notable pathogens, the transmission bottleneck is thought to be much tighter than that simulated in the majority of these scenarios. Only a single bacterium was thought to transmit in lab experiments of *S. pneumoniae* transmission [271] whilst the bottleneck in Influenza A was found to be between 1-13 virions [188]. Consequently, many transmission inference algorithms developed for use with whole genome data assume a complete bottleneck including TransPhylo and Phybreak [128, 265]. Of the scenarios where endsim outperformed SVC and BadTrIP, it achieved the highest accuracy of $\sim 75\%$ when the bottleneck was reduced to a single particle or when recombination was present.

It is important to consider the robustness of the endsim algorithm to differences in the global transmission parameters. In these simulations, we have fixed all the global parameters to their simulated values with the exception of the within-host population size. [Figure 28](#) indicates the resulting histograms of the inferred within-host population size under the difference simulation scenarios. As the endsim inference is based on coalescent simulation while the simulated scenarios are generated using a forward simulation strategy the population size estimates can not be directly compared. Instead, we can compare the relative estimates between the different scenarios. The bottom-left panel of [Figure 28](#) indicates that when the underlying scenario simulated a large within-host population size the endsim algorithm estimated a shift to larger population sizes when compared to the histograms of the other scenarios. Despite this, there appears to be a bias towards very low population sizes in the inferred distributions from endsim. Additional work is needed to properly diagnose this issue and to determine whether the exponential prior on the population size could be having an overly restrictive impact on the estimates.

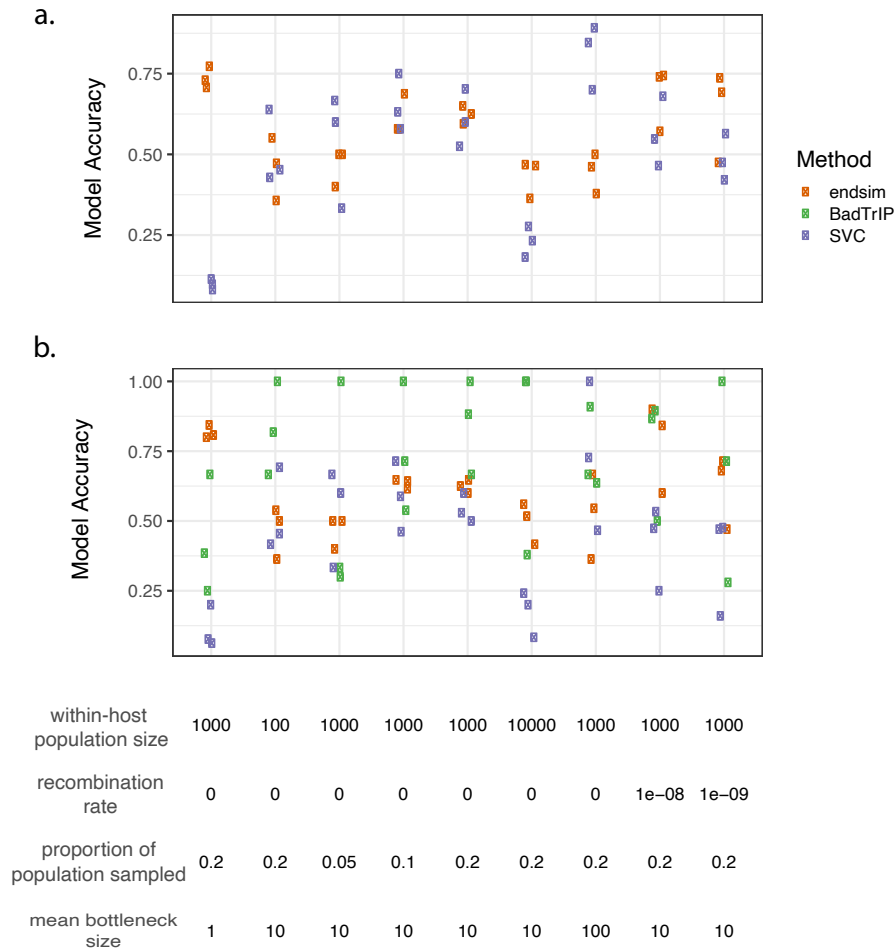


Figure 27: a.) A dotplot indicating the accuracy of endsim and SVC across nine different scenarios where three independent transmission networks have been simulated for each scenario. b.) The accuracy of the endsim, SVC and BadTriP algorithms on 10 random pairs, downsampled from those in a.

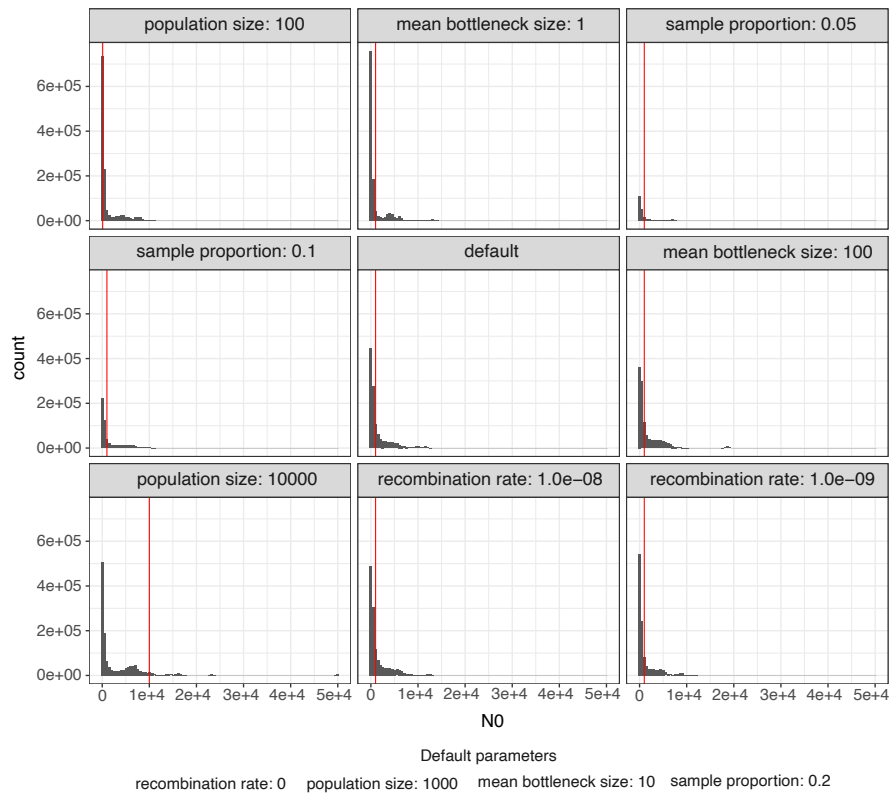


Figure 28: Histograms of the inferred within-host population size under the difference simulation scenarios given in [Figure 27](#)

4.4.9 *Limitations and further work*

While we have shown that the simulation based inference approach of endsim improves over competing methods in some scenarios it is important that the method is tested more thoroughly. This includes exploring how well the approach is able to recover other ‘global’ parameters such as the transmission bottleneck, recombination rate and sample size. In addition, the observed bias in the estimation of N_0 will need to be more carefully considered. Provided these tests are successful it is important that the approach be tested on real data for which high quality epidemiological information is available. This approach was originally designed to be applied to the deep sequencing study of pneumococcal transmission in [Part ii](#). We do not make use of this approach on real data in this thesis. However, this pneumococcal dataset along with the explosion in sequencing and tracking of SARS-COV-2 as a result of the recent pandemic present important applications for which this approach could potentially improve over the limited set of current methods. The endsim method currently assumes that every sample is the result of a single infection and that each site along the genome evolves at the same rate. Our analysis of pneumococcal and SARS-COV-2 infection in parts [ii](#) and [iii](#) of this thesis respectively suggests that these assumptions may be overly restrictive and that further work is needed to account for mixed infection and hypermutating sites. A further extension to the endsim method includes adapting it to infer parameters from purely longitudinal data of repeated samples of the same host. By ignoring transmission it may be possible to obtain cleaner estimates of within-host evolutionary rates and possibly allow for the investigation of selection within the host.

4.5 DISCUSSION

In this chapter, we have outlined three distinct strategies for the analysis of pathogen transmission using whole genome sequencing data. When choosing an analysis method the limitations of each approach must be considered. The phylogenetic and phylodynamic characteristics of each pathogen and outbreak will impact on the choice of the appropriate transmission inference approach.

We have shown that phylodynamic factors such as the prevalence of a pathogen within a population as well as whether a pathogen is in an endemic or outbreak state can influence the choice of approach, with endemic pathogens breaking the assumptions of many transmission inference algorithms. Sparse sampling, mixed infections, biased sampling, the pathogen’s evolutionary rate and within-host evolution are all important factors when considering which inference method to use.

In addition to ensuring the modelling assumptions of a chosen algorithm are suitable for the situation at hand, it is important to consider the feasibility of implementing such an inference approach. While in an academic setting more complex and accurate methods may be preferable, the added difficulty of implementing these approaches can make them impractical to use in a clinical setting.

When the evolutionary rate of a pathogen is low genomics is often used to exclude potential transmission scenarios rather than to identify complete transmission networks. In these scenarios, simple and fast pairwise comparisons are often best used in combination with traditional epidemiological techniques. In many cases the simplicity of these approaches makes them attractive for use in clinical and public health scenarios.

For faster evolving pathogens, or when the timescales of interest are larger, methods based on building phylogenetic trees and inferring the complete transmission network can bring valuable insights into transmission dynamics. While these approaches generally require more experience to implement, they can provide additional estimates of transmission directionality as well as estimates of the generation time, mutation rate, bottleneck size and within-host population size of the pathogen.

Finally, we have shown the limitations of relying solely on single genome samples and described methods that take advantage of information on within-host diversity to infer the direction of transmission. The added complexity of obtaining and processing deep sequencing data makes the routine implementation of these approaches in clinical settings difficult. However, this could change as our understanding of the evolution of pathogens within the host improves and as methods become more automated.

The utility of using genome sequencing to help track and respond to disease outbreaks has been highlighted by the SARS-CoV-2 pandemic. In the UK, the establishment of the Coronavirus Disease 2019 (COVID-19) Genomics UK Consortium (COG-UK) has led to the most rapid sequencing effort of a single pathogen in history with over 100,000 genomes sequenced by the consortium by the 10th of October 2020 [53]. Much of the motivation for this massive effort was to improve the tracking and surveillance of SARS-CoV-2 transmission within the UK and early analyses were able to identify Spain, France and Italy as the source of the majority of SARS-CoV-2 lineages during the first wave of the pandemic [410].

Despite the advances in both the routine sequencing of pathogens as well as transmission inference algorithms there remains many challenges to the effective use of genomics in epidemiology. Many of the algorithms we have described either make restrictive assumptions, are overly complicated or can be slow to run. The recent pandemic has highlighted the need for simple, robust and easily interpretable

tools with fast turnaround times. While the advanced sequencing and analytics conducted as part of the COG-UK study is highly valuable, it is out of reach for many communities and countries which are most affected by the pandemic. Consequently, future work will need to focus on improving current techniques whilst attempting to make them more broadly accessible.

Part II

THE PNEUMOCOCCUS

Part II describes the analysis of a large sequencing study investigating the within-host genetic diversity of *Streptococcus pneumoniae*

INTRODUCTION

Declaration of contributions

Jukka Corander, Simon Frost, Nick Croucher, Paul Turner and Stephen Bentley supervised this part of the thesis. Sample preparation, culture and DNA extraction was performed by Claire Ling and Zannah Salter.

Streptococcus pneumoniae (a.k.a. the pneumococcus) is a human nasopharyngeal commensal and a major human pathogen, which was responsible for approximately 294,000 deaths in children during 2015 [498]. The high burden of disease caused by *S. pneumoniae*, coupled with complexities in its recent evolution in response to antibiotics and vaccines, have led it to be classified as a priority pathogen by the World Health Organisation (WHO) [519]. The evolution of *S. pneumoniae* at the host population level has received substantial attention, most recently in the collection and whole genome sequencing of over 13,000 *S. pneumoniae* genomes collected from across the world [190]. In comparison, as is true for many bacterial species, relatively little is known about the genomic evolution of *S. pneumoniae* within its human host as well as the process of transmission between hosts. To explore the dynamics of pneumococcal evolution within and between hosts we considered a set of nasopharyngeal swabs sampled from 999 infants as well as a subset of mothers over 24 months, from the Maela refugee camp in western Thailand between 2007-2010.

This chapter starts by providing a background of the history and biology of the pneumococcus. The Maela deep pneumococcal study is then described in detail along with an explanation of how the samples were chosen and an outline of the analysis performed in the following chapters.

5.1 PNEUMOCOCCAL HISTORY AND BIOLOGY

Discovered in 1881, *S. pneumoniae* is a gram-positive bacterium from the Firmicutes phylum that can asymptotically colonise the human upper respiratory tract. Both Louis Pasteur and George Sternberg, a US army surgeon are credited with first isolating the pneumococcus after inoculating rodents and rabbits respectively with infected saliva taken from an infected child or in the case of Sternberg, himself [392, 467]. The pathogen was first found to cause pneumonia in 1882



Figure 29: A cartogram of the global mortality rate of *S. pneumoniae* in 2015. The high burden in Africa is clear, whilst India appears small despite its high number of deaths due to the country's large population size.

by Carl Friedländer and within the decade, had been found to cause meningitis and otitis media [181, 445, 535].

S. pneumoniae is a major global pathogen with invasive disease being most common in young children, the elderly and people with underlying health issues. In 2015, pneumonia accounted for the majority of pneumococcal deaths in HIV-uninfected children (81%), followed by meningitis (12%) [498]. The burden of disease is heterogeneously distributed across the globe with the highest mortality rates found on the African continent (Figure 29).

5.1.1 The pneumococcal capsule

Studies of the rough and smooth colony variants of pneumococci were first described by Pasteur in 1881 as 'le aureole' (the halo). Subsequent, investigations of convalescent-phase sera taken from patients indicated diverse serologic groups of pneumococci, which were later confirmed in animal protection experiments [33, 157]. Studies in the early 20th century confirmed that immunity to *S. pneumoniae* is primarily serotype specific [135, 309, 375].

The pneumococcus is an encapsulated bacterium with a polysaccharide capsule that is strongly related to the pathogen's virulence, as well as its ability to evade the host immune system and cause systemic infections [14]. With the recent discovery of serotype 10D, there are now 100 known capsular types [184].

Historically, the serotype of a pneumococcal colony has been determined through Quellung serotyping. This involves testing a cell sus-

pension with antisera specific for the capsular polysaccharide. The resulting reactions are then observed using a microscope. Whilst this approach is often referred to as the 'gold standard', it is labour intensive and is not cost effective for determining the serotypes of hundreds or thousands of potentially mixed samples. To improve the scalability of these approaches researchers developed a latex agglutination technique in which cultured colonies from a sample are suspended in saline and serotyped by latex agglutination [222]. Latex agglutination works by coating latex microbeads with serotype specific antibodies. After mixing these beads with the saline, any agglutination (clumping) of a serotype specific bead indicates the presence of that serotype. A follow on study confirmed that this accurately identified multiple serotypes in mixed carriage but that a microarray based approach were more sensitive [490].

As whole genome sequencing of *S. pneumoniae* isolates has become increasingly common in the surveillance of the pathogen a number of algorithms have been developed to call serotypes directly from WGS data. These typically rely on aligning sequence reads against a database of known capsular genes. Initial approaches focused on calling serotypes from WGS of single isolates. However, recently efforts have been made to expand this to mixed samples as described in [Section 6.2.2](#).

5.1.2 Vaccines and immunity

Initial pneumococcal vaccines consisted of dead pneumococci, with the first recorded mass inoculations of humans conducted in 1911 in the gold mining industry in South Africa by Almroth Wright [12]. The identification of serotype specific immunity led to early tests investigating the immune response when the polysaccharide was injected intradermally [480]. While initial attempts at a capsule-based vaccine were inconclusive, a tetravalent vaccine targeting capsule type 1, 2, 5 and 7 was shown to reduce pneumonia by 85% amongst US air force personnel during the second world war [13, 321].

The introduction of penicillin as a treatment of pneumonia led to waning interest in pneumococcal vaccines over the following two decades. However, the emergence of antibiotic resistant pneumococci led to renewed enthusiasm, and a 13 valent vaccine formulation was tested in the 1970's again in South African miners [12]. Whilst both a 14 and 23 valent version were later released in the USA, they proved successful mainly in immunocompetent adults and had little effect on paediatric disease [95, 323, 448].

Although the pneumococcal capsule polysaccharide was first conjugated to a protein in the early 19th century it was not until the late 1990's that a 7 valent pneumococcal conjugate polysaccharide vaccine (PCV7) was released in the USA [14, 420]. By conjugating with a pro-

tein the immune response in children was improved and following the introduction of PCV7, incidence of invasive pneumococcal disease decreased in the USA by 60-70% [306]. Despite this, the overall carriage rates of *S. pneumoniae* did not decrease due to the replacement of vaccine serotypes with less virulent non-vaccine types [367]. Following the introduction of PCV7; 10 valent and 13 valent forms of the vaccine have been introduced [84, 268].

5.1.3 *Other virulence factors*

Aside from the polysaccharide capsule there are a number of other prominent pneumococcal virulence factors. Whilst a number of enzymes and proteins are known to interact with the host, their roles in colonisation and disease are often less well understood. The toxin pneumolysin (ply), is well studied and known to be essential for transmission and colonisation [429, 530]. Immunogenicity has been demonstrated in several other pneumococcal surface proteins including pspA and pspC also known as CbpA which have both been associated with increased virulence [386]. These proteins have been shown to elicit cross reactive immune responses despite sequence diversity in their immunodominant regions and thus have been suggested as possible alternative vaccine targets [364]. Other virulence factors that have been associated with immune evasion include proteases zmpA, zmpB, zmpC and zmpD, which have been shown to aid in the evasion of human immunoglobulin [27]. The main pneumococcal autolysis (LytA) was also associated with increased virulence. LytA is a member of a group of cell wall degrading enzymes which are involved in the release of ply. Whilst LytA has not been studied extensively, the relatively low sequence diversity of the protein has led to increased interest in its potential as a vaccine target [253, 506].

5.1.4 *Antibiotic resistance*

Antimicrobial resistance was first observed in *S. pneumoniae* in the early 1900's. Early antibiotics included optochin and sulphanilamide. Resistance to optochin developed quickly and this coupled with the drug's poor effectiveness and significant side effects including blindness prevented it from becoming a common treatment [366].

The development of penicillin by Alexander Fleming in 1929 revolutionised the treatment of many bacterial infections including those caused by *S. pneumoniae*. The drug became widely used in the 1940's and coupled with the development of cephalosporins, pneumococcal disease decreased significantly over the following decades [95, 258]. Resistance to penicillin was first reported in Australia in the 1960's in strains with serogroups 23 and 6 [264]. Over the following two

decades resistance was found in North America, Africa and Europe [167, 243, 333].

Following the discovery and widespread use of penicillin a number of other antibiotics were developed in the 1940's and 1950's. These included; tetracycline aureomycin, chloramphenicol and macrolide erythromycin [142, 145, 345]. Following their widespread use resistant pneumococci soon emerged. Tetracycline resistance was found in 1963 in Australia, erythromycin in 1966 in the USA and chloramphenicol resistance in 1973 in France [109, 156, 264]. The following decade saw the combination treatment trimethoprim with sulfamethoxazole (co-trimoxazole) to treat pneumococcal infections, followed quickly by the emergence of resistant strains [230, 235].

The discovery of nalidixic acid in 1962 marked the start of decades of development and use of quinolones with fluoroquinolones coming to prominence in the 1970's and 1980's [149]. Although fluoroquinolone resistance was observed as early as the 1990's the drug is often used to treat respiratory infections and rates of resistance in pneumococci remain fairly low [65, 262].

The emergence of multidrug resistant strains in the 1970's motivated the development of the Pneumococcal Molecular Epidemiology Network (PMEN) to track the development and emergence of multidrug resistant clones [266]. Vancomycin remains as one of the only antibiotics that is effective against all pneumococci although tolerance to the drug has been observed [340, 379]. More recently the development of Oxazolidinone and Ketolide antibiotics has expanded the repertoire of drugs available to treat multidrug resistant *S. pneumoniae* clones [262].

5.1.5 *Pneumococcal recombination*

In 1928, Frederick Griffith, a British medical officer demonstrated that by injecting mice with both dead pneumococci from a virulent strain as well as live non-virulent pneumococci, the mice could develop a fatal infection. He identified that the bacteria were able to 'transform' to the virulent, encapsulated type [199]. Following on from these experiments, Avery and colleagues showed that DNA rather than protein as had previously been believed was the material that had been transferred between the live and dead cells [15]. Pneumococci have since been shown to recombine both homologously, which amounts to the replacement of DNA sequence with a highly similar sequence as well as illegitimate or non-homologously which involves the integration of a dissimilar DNA sequence often comprising a bacteriophage or transposon.

5.1.5.1 Competence

S. pneumoniae has a competence system dedicated to the uptake of exogenous DNA. Competence in streptococci is a temporary state which can last up to 40 minutes [252]. In pneumococci, the state is regulated by extracellular signalling, generated by the comAB transporter [212, 482]. The extracellular pheromone Competence Stimulating Peptide (CSP) that it generates is detected by the comDE two component system [398]. This leads to the activation of stress response systems as well as the competence genes that ultimately results in a physiological state known as the 'X state' [76]. Two stages are associated with the X state. The first involves the comABCDE genes whilst the later involves activation of genes involved in promoting DNA processing and genetic recombination including genes for the competence pseudopilus, DNA uptake pore and ssDNA binding [108]. Genes involved in the later stage also include those related to the lysis of cells to release DNA, including *lytA*, *cbpD* and the bacteriocin *cibAB*. The genes *cibC* and *comM* help to prevent lysis of the host cell [200, 213]. This enables fratricide whereby pneumococci are able to kill related cells that exhibit a different response to CSP. This is one mechanism by which pneumococci in mixed infections of multiple lineages are able to compete.

In carriage, competence is likely to be induced by stress to enhance the pneumococcal's ability to adapt to external stressors within the host. Competency has been shown to be stimulated by antibiotic treatment including aminoglycosides, fluoroquinolones and mitocin [408].

5.1.5.2 Transduction

Distinct from the complement system, recombination can occur in *S. pneumoniae* through transduction, where the foreign DNA is introduced via a viral vector. A significant proportion of *S. pneumoniae* isolates are lysogenic with genetic elements that resemble prophages observed in up to 70% of clinical isolates [416]. While prophage infected pneumococci were first described in 1975, the impact of these prophages on the pneumococcal host is still being understood [116, 343]. The predator-prey interaction between phages and bacteria has been shown to both promote and impede bacterial fitness: increasing persistence within the host; improving the bacteria's ability to adapt to changes in the human host environment such as antibiotic treatment; promote the bacteria's ability to colonize new hosts; and impact the bacteria's ability to cause disease [45, 68, 116, 492]. Several pneumophage genomes have been sequenced including the temperate phages Φ MM1 and Φ MM1-1998, the lytic phages Φ Cp-1, Φ Dp-1 and SV-1 as well as the recently characterised phage like element SpnCI [312, 339, 383, 416, 430].

5.1.5.3 Conjugation

Conjugation describes the process of genetic transfer of material through direct contact between bacterial cells. The evolution of *S. pneumoniae* involves two types of conjugative elements - plasmids and integrative and conjugative elements (ICEs). While it has been shown that *S. pneumoniae* is able to accept and transfer plasmids, they very rarely occur naturally [19, 31]. Two distinct *S. pneumoniae* plasmids that have been identified include pDP1 and pSpnP1. However, their impact on the pneumococcal host cell is unknown leading them to be labeled as cryptic [426, 458].

Unlike plasmids, ICE elements are common in *S. pneumoniae* and have been strongly associated with antimicrobial resistance [52, 105]. Evidence for tetracycline resistance conferring ICE elements present as early as the late 1960's has recently been found in Australia [524]. This placed the emergence of the resistance ICE element within a decade of the introduction of the antibiotic. Of the common antibiotic classes a study has linked resistance to macrolides, lincosamides and aminoglycosides to ICE elements but not penicillin, trimethoprim or sulphonamides [52].

5.1.5.4 Pneumococcal strains

Similar to many other pathogens, *S. pneumoniae* genomes cluster into genetically similar groups or strains despite recombination. Both the conservation of core and accessory genome content, coupled with the fact that different typing methods group genomes into similar clusters indicates that pneumococcal strains represent significant biological entities [98, 457]. This is highlighted by the development of PMEN to track the distribution of multidrug resistant strains [266, 344]. The terms 'lineage', 'strain' and 'Global Pneumococcal Sequence Cluster (GPSC)' will often be used interchangeably throughout this thesis.

While the existence of distinct *S. pneumoniae* strains is well accepted, the underlying evolutionary mechanism for the development and maintenance of this structure is less well understood [459]. Prior to the development and widespread use of whole genome sequencing a number of neutral models were developed to help explain the co-existence of multiple strains. New models have subsequently been suggested that better align with the recent large scale sequencing efforts such as the Global Pneumococcal Sequencing Project [86, 190].

Initial mechanisms for maintaining strain structure in bacterial populations suggested that they could result from many lineages existing over very short durations, possibly as the result of micro-epidemics [177]. However, pneumococcal strains have since been shown to remain stable over long periods of time [97, 100]. Isolation by distance is another mechanism for maintaining distinct strains where each strain is geographically separated, preventing between strain recombination

and thus maintaining structure. Again, WGS has shown that this is unlikely given the global distribution of many *S. pneumoniae* strains [190]. A similar but much smaller scale version of this hypothesis is that strains have adapted to specific ecological niches. However, frequent observations of coinfection of multiple strains within the same host niche and little private gene content suggests that this is unlikely [98, 471]. Distinct from this model, immunological niches have been suggested as an alternative explanation for bacterial strains whereby negative frequency-dependent selection (NFDS) acts on variable antigens, reducing the fitness of common strains due to increased host immunity and maintaining intermediate strain frequencies [201, 296]. The lack of discordance of variable antigens between populations implies this model is unlikely [96]. NFDS has also been suggested as a mechanism for maintaining the coexistence of both antibiotic resistant and susceptible strains via within-host competition enabled by the ability of multiple strains to co-colonise the same host [77, 111]. Genetic linkage between resistance loci and those involved in determining the carriage duration of a strain represents an alternative mechanism for the coexistence of resistant and susceptible strains [294]. Strains with longer carriage durations would benefit more from the acquisition of resistance, even if this resistance came with a fitness cost. In this scenario, recombination events that upset this epistatic interaction would be selected against, maintaining multiple distinct strains [373]. However, genome wide epistasis analysis of large pneumococcal WGS datasets has suggested that strong epistatic interactions are relatively rare and focused on a small number of loci [396, 454].

An alternative NFDS model, assumed that accessory loci found at intermediate frequency were the result of NFDS acting via multiple phenotypes [86]. In addition to antibiotic resistance, it was suggested that NFDS may act via competition for different nutrient resources as well as direct competition between bacteria via differing bacteriocin biosynthesis gene clusters. Similar to Gupta *et al.* 1996 it was suggested that NFDS was acting on genes encoding immunogenic structures. However, this was not suggested as a mechanism of maintaining overall strain structure [86]. Instead, a recent extension to this work has suggested that asymmetric recombination whereby deletion of accessory loci is more likely than insertion can result in distinct strains and has a better fit to large whole genome pneumococcal datasets [210].

5.1.6 Transmission

S. pneumoniae is reliant on being able to effectively transmit between human hosts as it does not persist in the environment. As such the pneumococcus stimulates inflammation and shedding from mucosal

surfaces in the nasopharynx which facilitates the transmission of pneumococci between hosts. The subsequent colonisation of a new host represents a major component of the overall transmission process. Common laboratory models of pneumococcal transmission include infant mice and ferrets [427, 449]. Due to the increased difficulty in studying transmission in natural human carriage, these models have proved critical in the identification of genes involved in both transmission and colonisation [505].

Transmission involves a number of key steps including: exiting from the current host; survival in the environment; acquisition by a new host and finally colonisation. The pneumococcus is able to transmit from the current host via nasal secretion. Pathogen induced inflammation can stimulate both the expression of mucus glycoproteins in addition to the flow of mucus. The virulence factor pneumolysin, encoded by the *ply* gene, has been shown to be both necessary and sufficient in causing the required inflammation in an infant mouse model [530]. Expression of *ply* has also been shown to increase the survivability of pneumococcal in nutrient poor conditions outside the host, which can be related to the increased inflammation leading to larger and more nutrient rich secretions [530]. More recently, the *dlt* locus has been shown to promote transmission via inflammation in addition to enhancing resistance to antimicrobial peptides [529].

Acquisition in the new host is inhibited by pre-existing immunity, mainly in response to the pneumococcal polysaccharide capsule as described previously. In addition to variation in its capsule sequence, *S. pneumoniae* is able evade clearance mediated by immunoglobulin A1 (IgA1) via expression of the zinc metalloprotease ZmpA which prevents the agglutinating activity of IgA1 [423].

Finally, successful colonisation of the human nasopharynx involves several factors including: adherence to the nasopharynx; competition with resident nasopharyngeal flora; and maintenance within the host for a period sufficient to allow for onwards transmission [505]. Pneumococci can adhere to the host mucosa [163]. However, by avoiding entrapment in the mucus, pneumococci can adhere to host epithelial cells via adhesins including pneumococcal adherence and virulence protein A and B (PavA/B), enolase (Eno), phosphorylcholine (ChoP), and choline-binding protein A (CbpA) [30, 106, 226, 505, 536].

5.2 THE MAELA DEEP SEQUENCING STUDY

5.2.1 Background

For many pathogens including *S. pneumoniae*, success depends on their ability to transmit between, and survive within, hosts. The vast majority of genomic studies into *S. pneumoniae* have so far consisted of sequencing single colonies from cultured samples. While this has

provided many insights into the evolution of the pathogen, relatively little is known about the genomic diversity of *S. pneumoniae* within the host and during transmission. The Maela deep sequencing study attempts to fill this gap by sequencing a large number of whole plate sweeps of cultured pneumococcal samples to a high depth. By deep sequencing the whole plate rather than a single colony we are able to efficiently investigate the genomic diversity of *S. pneumoniae* within each sample.

5.2.1.1 *Within-host bacterial evolution*

Initial estimates of long-term bacterial evolutionary rates were of the order of 10^{-10} to 10^{-9} substitutions per year, indicating that within-host bacterial populations would have been monomorphic [130]. As whole genome sequencing has become more readily available, short-term evolutionary rates in bacteria are now understood to be considerably higher (10^{-7} to 10^{-5}). Coupled with the larger genomes of bacteria, the per genome mutation rate of some bacteria is comparable to or exceeds that of many viruses [130]. A higher mutation rate has important ramifications for the adaptation of bacteria to antimicrobials, as well as host immunity and competition between strains. Faster mutations also enable the use of whole genome sequencing in the epidemiological surveillance of bacterial pathogens, as mutations can be tracked allowing for the inference of transmission chains [209].

Within the host, a pathogen must evade the host's immune system, compete with other species or strains and resist the impact of any medications that have been taken. The spread of antibiotic resistant strains of many bacteria is a direct result of their ability to adapt to selection pressures within the host. A notable example of such adaptation is the evolution of multidrug resistant *M. tuberculosis* within a single patient [148]. Bacteria must also evolve to outcompete other species or strains, whilst avoiding the host immune system. These often competing goals can result in complex within-host evolutionary dynamics.

Competing evolutionary pressures favour strains and species that are able to adapt to changes in selective pressures quickly [130]. A higher mutation rate can increase the adaptability of a strain. Hypermutating strains can occur when the mismatch repair system becomes disrupted, thereby allowing for a strain to cross valleys in fitness; a feat that would be less likely at lower mutation rates [195]. Phase variation, where frequent reversible genetic changes alter specific gene expression, allows a strain to respond to fast changes within its environment. By harbouring resistant genotypes with possible deleterious mutations at a low frequency, a bacterial population can respond quickly to the introduction of antimicrobials. This is known as heteroresistance [377]. The presence of low frequency genotypes that are suited to specific environmental pressures has also been as-

sociated with virulence. For example, serotype 1 pneumococci are rarely found in carriage but are often observed in invasive disease [62]. Whilst these adaptive abilities are important for a pathogen's longevity within a host, the survival of the species or strain also relies on effective transmission between hosts. By deep sequencing a large collection of whole plate sweeps of pneumococcal from a refugee camp in Maela, Thailand, we are able to gain a better understanding of the evolutionary dynamics of *S. pneumoniae* within the host and during transmission.

5.2.1.2 The Maela collection

Maela is the largest of nine long term refugee camps along the Thai Myanmar border, established in 1984 to accommodate displaced persons from Myanmar (Figure 30). The camp's population is mostly Karen, one of Myanmar's largest ethnic minorities [478]. Between 2007 and 2010 approximately 40,000 refugees lived within the camp's 4 square kilometer area. Movement in and out of the camp is severely restricted. However, some residents risk being fined by Thai authorities to take up agricultural work in the surrounding region [164]. Whilst camp residents receive the WHO expanded program on immunization vaccines, the *Haemophilus influenzae* type B (HiB) and pneumococcal vaccines were not available during the period over which this study was conducted [491]. The Shoklo Malaria Research Unit (SMRU) was established within the camp in 1986 to provide medical care to residents. To help understand the dynamics of *S. pneumoniae* carriage in a rural South East Asian population Turner *et al.*, [491], investigated pneumococcal carriage in mother-infant pairs from birth for 24 months. Pregnant women attending the SMRU clinic at 28-30 weeks gestation were invited to participate in the study between October 2007 and November 2008 resulting in 999 women being selected for the study [491]. Following delivery, children and a subset of mothers had nasopharyngeal swabs taken monthly from 1-24 months of age. These samples have subsequently been used in a large number of follow-on studies including a large WGS study of a subset of 3,085 randomly selected samples [71]. In this study, swabs were plated onto selective agar before the selection of single colonies of *S. pneumoniae* for sequencing. This resulted in a single pneumococcal genome for each swab.

With the advent of affordable next-generation sequencing, it is now possible to perform WGS at a high depth of coverage across a large population of samples [313]. This provides a greater level of detail for the study of within-host diversity and transmission. To obtain a comprehensive overview of the genome dynamics of *S. pneumoniae* within the host, during carriage, and transmission we selected a subset of 4,000 nasopharyngeal swabs from the Maela collection. These swabs were plated onto selective agar and cultured before the whole

a.



b.



Figure 30: a.) Location of the Maela refugee camp along the Thailand-Myanmar border. b.) A photo overlooking the Maela refugee camp.

plate was scrapped, DNA extracted, and sequenced to a mean coverage of 250 fold across the *S. pneumoniae* genome. By sequencing the whole plate to a high depth of coverage rather than a single colony as was done previously, we are able to gain insight into the diversity of *S. pneumoniae* within the host.

5.2.2 Sample selection

A subset of 4,000 out of ~ 11,000 swab samples from a study of pneumococcal carriage in Maela, Thailand was chosen for deep sequencing in this study [491]. In order to choose the most informative samples for our aims we made use of 3,085 previous single colony WGS samples from the same data set [71]. To investigate the phylodynamics of transmission, samples involved in direct or near-direct transmissions were of particular interest.

The adapted TransPhylo algorithm was applied to each major cluster from the WGS Maela data set of [71], after first clustering using hierBAPS [70, 484] as described in Chapter 4. Those sample pairs with an inferred transmission distance of < 5 intermediate hosts were selected for culture and sequencing. This threshold corresponds with a SNP distance of approximately 10 SNPs which is similar to previous studies [56]. We also selected any remaining transmission pairs that were within this 10 SNP threshold.

In addition to selecting samples plausibly linked by transmission we identified those associated with antimicrobial treatment and resistance to help investigate the impact of treatment on within-host diversity. In order to choose samples to best achieve our aims, we chose samples approximately split into four categories of resistant, susceptible, treated with antimicrobials and untreated. Samples were classified as treated if the patient had been given antimicrobials within the previous two months. As the overall number of treated samples within the data set was low, all treated samples were selected for deep sequencing. The previous untreated samples for the same patients were also chosen to allow for the comparison of within-host diversity before and after treatment. Finally, patients with more than three samples already selected were treated as possible ‘index’ cases. Additional samples were chosen from these index cases to a maximum of twelve samples per patient to gain a longitudinal view of transmission and within-host diversity. The set of index cases includes a number of mother/child pairs, which are of interest to compare with more geographically distant transmission events.

5.2.3 *Analysis outline*

In the following chapters we discuss the analysis of the resulting deep sequencing data to investigate pneumococcal evolution within the host, during carriage, after treatment and in transmission.

[Chapter 6](#) describes the methods used to deconvolve the possibly mixed pneumococcal samples into lineages. The observed distribution of within host frequencies of each lineage are discussed before concentrating on the major virulence factors of the pneumococcus including its serotype. Finally, association between instances of multiple carriage and virulence factors are considered.

Antibiotic resistance determinants are focused on in [Chapter 7](#). We demonstrate that it is possible to accurately identify antibiotic resistance elements from the deep sequencing data and compare the sensitivity of this approach to that of single colony picks. Similar to the analysis of virulence factors we then examine the association between multiple carriage and antibiotic resistance determinants. To explore the impact of treatment on pneumococcal diversity within the host, samples taken pre- and post-antibiotic treatment are compared. The impacts of treatment are considered both at the lineage level, using the lineages assigned in [Chapter 6](#), as well as the locus level through the use of a custom bacterial genome wide association study (GWAS) pipeline.

Finally in [Chapter 8](#), we investigate fine scale variation generated at the sub-lineage level within the host. Using the inferred minority variants we consider the dynamics of transmission within the camp concentrating specifically on the relationship between mother and child. We go on to explore the transmission bottleneck before using the inferred minority variants to investigate the mutational processes and selective forces driving the evolution of *S. pneumoniae* within the host.

LINEAGES, SEROTYPES AND OTHER VIRULENCE FACTORS

6.1 BACKGROUND

The accurate identification of *S. pneumoniae* within the nasopharynx is vital to understanding which strains and serotypes are circulating within a population. This is important in determining the effectiveness of vaccines, the prevalence of drug resistant or virulent lineages and to gain an understanding of the mechanisms driving pneumococcal population dynamics.

Initial efforts to distinguish between different pneumococcal lineages focused on identifying the pneumococcal capsular serotypes within carriage samples. This was initially achieved using Quellung serotyping which was later followed by latex sweeps and DNA microarrays [490]. Most recently, whole genome sequencing has allowed the determination of strain structure at a much higher resolution enabling the characterisation of the global population structure of *S. pneumoniae* in the Global Pneumococcal Sequencing Project [190]. In addition to *in-silico* serotyping, WGS data provides information on sequence variation outside the pneumococcal capsular genes allowing for the development of bespoke lineage identifications techniques such as PopPunk [290]. A wide range of other tools for the analysis of WGS data have also been developed including those that identify virulence and resistance genes.

Unfortunately, similar to Quellung serotyping, typical pneumococcal sequencing studies rely on the WGS of a single pneumococcal colony. Identifying multiple strains within a sample can thus be both time consuming and expensive. Deep WGS of whole plate sweeps offers an alternative, allowing for both the increased resolution afforded by WGS and enabling the identification of multiple strains within a sample.

6.1.1 Strategies in analysing deep sequencing data

Short read deep sequencing data of whole plate sweeps presents an analysis challenge as the sequencing reads may be made up of a mixture of multiple strains, can include contaminating species and may be subject to plate effects induced by the culturing stage of the experiment. Initial quality control steps such as identifying highly contaminated samples or those with poor coverage of the *S. pneumoniae* genome are important. Following the removal of poor quality sam-

ples there are two main strategies that can be used to analyse deep sequencing data. The ‘deconvolution’ based strategy involves first attempting to separate sequencing reads into the individual strains that make up the mixed sample. This allows for traditional analysis pipelines to be run on the deconvoluted data and thus presents an attractive solution. However, the deconvolution process itself is not trivial and poorly deconvoluted samples can have a profound impact on downstream analyses potentially leading to erroneous conclusions. An alternative strategy is to perform analyses that do not require single strains. Many algorithms designed for the identification of strain types, virulence and resistance elements do not require deconvoluted samples thus avoiding the potentially difficult deconvolution step. To generate robust results we make use of both strategies. This allows us to take advantage of the many algorithms designed to work with single bacterial genomes whilst ensuring our results are robust.

6.1.2 *Multiple carriage*

The co-colonisation of a single host with multiple pneumococcal strains represents a critical component to understanding the evolutionary dynamics of *S. pneumoniae*. Multiple carriage allows for the exchange of genetic material between strains enabling them to adapt to selective pressures far faster than would be possible solely via clonal evolution. The competition between strains within the host has also been suggested to contribute to the maintenance of strain structure within the wider pneumococcal population allowing antibiotic resistant and susceptible strains to stably coexist over long time periods [80, 111, 210, 308]. Models of the maintenance of gene distributions within the pneumococcal population also rely on the coexistence and competition of pneumococcal strains within the host [86]. Furthermore, within-host competition has been postulated to lead to a ‘founder takes all’ model where an occupying pneumococcal strain is able to outcompete potential intruders leading to the selection for transmissibility under a tight transmission bottleneck [244]. It is only by gaining a strong understanding of the dynamics of multiple carriage in *S. pneumoniae* that we will be able to tease apart which of these often competing models is most likely and whether there are additional details that have yet to be considered in current modeling efforts.

In this chapter, we summarise the methods used to perform quality control, call features and deconvolute the deep sequenced whole plate sweeps from the Maela refugee camp. By comparing our results with those obtained from previous latex sweeps [490] and WGS of single colony picks [71] we are able to demonstrate the accuracy and increased sensitivity of the approach in detecting pneumococcal lineages, serotypes and virulence genes. Using the resulting calls we go on to consider the implications for pneumococcal surveillance by

considering the detection sensitivity of disease causing serotypes as well as other known virulence factors. Finally, we borrow techniques commonly used in ecology to investigate multiple carriage in *S. pneumoniae* and discuss the implications our results have for models of pneumococcal strain structure.

6.2 METHODS

6.2.1 Lineage deconvolution

Whilst there are a large number of tools for identifying the species composition of mixed samples from sequencing data there are relatively few that go further and deconvolute the sequencing reads into species specific bins. Of these tools only a minority are able to deconvolute reads at a high enough resolution to separate strains within a species.

Two of the most popular composition based tools include Kraken and Centrifuge [260, 512, 513]. Whilst both classify reads at the species level, they do not allow for reads to be allocated to multiple species and instead classify such multi-mapping reads to a higher order such as the genus. As many bacteria share homology this can complicate downstream analyses. This problem becomes particularly acute in the case of highly similar strains. A number of alternative tools have been created to resolve the deconvolution problem including those that rely on a reference database [319, 357, 382, 540] and those that attempt to construct deconvoluted assemblies directly from metagenomic data [298, 381]. Whilst *de novo* assembly based methods have proven highly useful in metagenome studies they have shown to perform poorly in generating accurate assemblies at the strain level [440]. Recently, the mGEMS pipeline has been developed that enables both quantification and deconvolution at strain resolution [324]. The mGEMS pipeline builds off recent work on quantification provided in the mSWEEP program [325].

6.2.1.1 mSWEEP

The mSWEEP algorithm estimates the relative abundance of different predefined bacterial lineages or groups in a mixed sample [325]. It requires a curated dataset where each genome has been pre-classified into biologically meaningful groups. In our case we use a manually curated subset of 20,047 assemblies from the Global Pneumococcal Sequencing (GPS) project where potentially contaminated samples have been excluded. For the grouping we use the pre-defined lineages generated by the PopPunk algorithm which was originally designed for use on the GPS project. The mSWEEP algorithm relies on the pseudoalignment of reads to references using Themisto [324]. Pseudoalignment makes use of a fast approximate kmer based algorithm

to assign reads to references allowing each read to be assigned multiple times. Following alignment, a vector $r_n = \{r_{n,k}\}$ is generated which indicates the number of alignments for read n in each group k . Using this, the mSWEEP algorithm estimates the probability that each read originated from each predefined group. It assumes that if groups have the same number of reference sequences, a read is more likely to originate from the group that has a higher number of alignments. The mSWEEP algorithm further assumes that the likelihood of a read originating from a group should not depend on the number of references in that group.

Given the vector r_n the probability that a read belongs to group k is given by

$$p(r_n | I_n = k) = \begin{cases} 0.01 & \text{if } r_{n,k} = 0 \\ 0.99 & r_{n,k} > 0 \text{ and } M_k = 1 \\ 0.99 \frac{f(r_{n,k}, k)}{Z(r_n)} & \text{when } r_{n,k} > 0 \text{ and } M_k > 1 \end{cases} \quad (32)$$

where

$$f(r_{n,k}, k) \propto \binom{M_k}{r_{n,k}} \frac{B(\alpha_k + r_{n,k}, M_k - r_{n,k} + \beta_k)}{B(\alpha_k + M_k, \beta_k)} \quad (33)$$

$$Z(n_k) = \prod_{k=1}^K \prod_{j=1}^{M_k} \frac{\alpha_k + r_{n,k} + (j-1)}{\alpha_k + \beta_k + 2M_k + (j-1)} \quad (34)$$

Here, M_k is the number of references in group k , $B(a, b)$ is the Beta function and $f(r_{n,k}, k)$ is the compound distribution after assuming that the probability of a read originating from a group follows a binomial distribution with a Beta prior on the probability of success. $Z(n_k)$ is a normalising term that accounts for all $f(r_n)$. The mSWEEP algorithm uses a variational inference strategy to fit these parameters efficiently.

The mSWEEP algorithm has been shown to work well when the groups of interest are well represented in the database but has been found to struggle when an unknown group is present (communication with the mSWEEP authors). In this case the algorithm generally assigns reads from the unknown group to multiple different lineages as there does not exist a strong match. To counter this we make use of the Mash screen algorithm and take the intersection between the two approaches to ensure that any mSWEEP calls are likely to be present in the mixture [387].

6.2.1.2 Mash screen

Mash screen is an extension to the fast kmer based distance method Mash for measuring the containment of one genome in another [387].

Unlike Kraken and mSWEEP, Mash screen does not estimate the proportion of a sample assigned to a particular reference sequence. However, its increased efficiency enables a large database of similar references to be searched. For each reference genome, Mash screen identifies the presence of a subset of reference kmers that are also present in the mixed sample of interest. By ensuring that any groups identified by mSWEEP also have a high number of shared kmers with at least one representative reference sequence we are able to exclude spurious mSWEEP calls driven by strains or species not being accurately represented in the database. While this approach is heavily dependent upon the reference database used, as we include over 2000 whole genome assemblies obtained via WGS of single colony picks from the Maela collection, our database should accurately reflect the vast majority of pneumococcal diversity in the Maela camp.

6.2.1.3 mGEMS

The mGEMs pipeline is an extension of mSWEEP that makes use of the inferred read assignment probabilities to deconvolute the mixed sequence data into individual fastq files for each strain [324]. By coupling this pipeline with the assembly tool Shovill (a SPAdes optimising pipeline), we are able to run many of the conventional analysis tools designed for single strains [17, 441]. mGEMs assumes that each group or lineage is only represented once in a mixed sample. While this may not always hold, disentangling samples below this level of resolution presents a significant challenge as it can become hard to distinguish multiple independent carriage events from diversity generated by within-host evolution. Instead, in this chapter, we restrict much of our analysis to reporting at the pneumococcal lineage level. The mGEMs algorithm assigns each read to a lineage if the mSWEEP assignment probability $\gamma_{n,k}$ is greater than a lineage specific threshold q_k . In practice, q_k is chosen using the estimated proportion of the lineage found in the mixture by mSWEEP. Thus, the set of reads for each group identified by mSWEEP is $G_k = \{r_n : \gamma_{n,k} \geq q_k\}$. This allows for reads to be allocated to multiple groups, an important feature given the high level of similarity between lineages. In practice, to allow for subsequent assembly of the reads, we restrict our analysis to those samples inferred to make up at least 1% of the reads in a sample.

6.2.2 Serotype calling

To identify serotypes from the potentially mixed samples obtained from the whole plate sweeps, we make use of both a deconvolution based strategy and recent methods developed to infer serotypes directly from mixed samples. This involved running SeroBA against

the deconvoluted strains generated by the mGEMS pipeline as well as running Serocall directly on the deep sequenced samples [152, 267].

6.2.2.1 *SeroBA and Pneumocat*

The first algorithm to call serotypes from WGS data was Pneumocat [254]. This algorithm runs in two stages. In the first stage, sequencing reads are aligned to a database of *S. pneumoniae* capsular genes. If the genes for a particular serotype have a coverage of at least 90% and uniquely specify that serotype the algorithm returns the serotype call and terminates. If multiple serotypes have the same gene set then the algorithm proceeds to the next stage where finer scale features such as SNPs and frameshifts are considered. If no serotype has at least 90% coverage over its capsular genes an inconclusive result is returned.

The SeroBA algorithm uses a fast kmer based approach in replacement of the full alignment used by pneumocat in its initial stage [152]. This improves the speed of the algorithm making it simpler to scale analyses to 1000's of samples. In place of the second stage of the Pneumocat algorithm, SeroBA uses the Ariba pipeline to identify finer scale features in order to discriminate between serotypes with similar capsular sequences.

6.2.2.2 *Serocall*

Pneumocat and SeroBA are both designed for use with single isolates. Instead, the Serocall algorithm is able to identify serotypes in mixed samples and reports the proportion of each serotype found within a sample [267]. The algorithm initially aligns the sequencing reads against a database of capsular genes to produce binned counts of unique and multi-mapped reads aligned to different locations in the capsular sequences. Each of these bins is associated with one or more serotypes and compared to the expected number of alignments generated by randomly simulating reads from each serotype and aligning them back to the capsular sequences. An optimisation strategy is then used to identify the most likely combination of serotypes that would generate the observed bin counts.

6.2.3 *Virulence calls*

6.2.3.1 *Ariba*

The Ariba resistance calling pipeline uses a local assembly based approach to identify genes and mutations in sequencing data from a database of known features of interest. Although the pipeline was originally designed for identifying resistance elements it can also be used to call other genomic features as exemplified in its use in the

final stage of the SeroBA pipeline. As Ariba reports the proportion of reads that support a given feature it is suitable for use with mixed deep sequencing data. However, it can not determine which lineage the observed feature originated from within a mixture.

The Ariba algorithm proceeds by first clustering the reference dataset into non-redundant groups using `cd-hit-est` [304]. Reads are then aligned to the clustered sequences using `minimap` and the local clusters of reads assembled using `fermi-lite` [300]. The resulting assemblies are compared to the reference sequences using `MUMmer` and `nucmer` allowing for the identification of resistance causing SNPs as well as more complex variants such as indels and deletions [120]. Finally `Bowtie2` and `samtools` are used to align the original reads back to the assembly which enables the identification of cases of both mixed and susceptible strains [281, 301].

To identify virulence features beyond the capsular genes we ran Ariba on all the deep sequencing samples using the full Virulence Factor Database (VFDB) [66]. We restricted our analysis to those features found in *S. pneumoniae* genomes.

6.2.4 Association with multiple infection

The association of lineages with multiple carriage, as well as virulence and resistance determinants, can provide important clues as to the ability of different features to coexist with other strains. Lineages and features that are more often found in mixed infections are also prone to being missed in genomic surveys that focus only on single colony picks. Thus identifying the relationship between different lineages and instances of multiple carriage is important for both the surveillance of the pathogen and in understanding the competition between lineages within the host.

We consider two main strategies for investigating the association of lineages, and other features of interest, with multiple carriage. In both cases we make use of a Mixed Model, allowing us to efficiently control for variation introduced by each host using a random effect. To account for the potential impact of antimicrobial treatment on multiple carriage we include a binary covariate with a fixed effect. A sample is considered to be impacted by treatment if the host was treated within 4 weeks prior of the sample being taken. This threshold was chosen after reviewing the decline in the proportion of resistant isolates tested via disk diffusion and Etest testing of all swabs positive for *S. pneumoniae* (Figure 31). It is also consistent with that described in a previous review of the literature [307].

In the simpler regression the proportion estimates from the mGEMs/Mash analysis were used as the dependent variable and each lineage or serotype of interest was tested independently using a Linear Mixed Model (LMM). The Benjamini-Hochberg (BH) method was

used to control for multiple testing [28]. This approach assumes that the lineage proportions estimated by the mGEMs pipeline are representative of what would be found in the host that the sample originated from. As these proportions are subject to the effects of both the swabbing procedure and being cultured on a plate, it is uncertain whether they provide an accurate representation of the true proportion within the host.

As an alternative we considered whether certain lineages and features of interest were more likely to be found in multiple carriage, treating multiple carriage as a binary dependent variable. We used a modified link function to account for the expected increase in prevalence of a feature that would occur simply because a larger number of lineages are sampled in multiple infections. As an extreme example, if we were to sample a mixed infection with one million lineages there would be a very high likelihood we would observe resistance within the mixture. However, this would not necessarily be due to an association with multiple carriage but rather due to the very large number of lineages present. This is a similar problem to that seen in ecology in the study of nest success where the overall age of the nest must be accounted for [446]. In this case the standard link function used in logistic regression is augmented to account for the exposure or in our case the number of lineages.

The standard logistic link function can be written as

$$g(\theta) = \log \left(\frac{\theta}{1 - \theta} \right) \quad (35)$$

We augment this, replacing θ with $\theta^{\frac{1}{N}}$ where N is the number of lineages present in a mixture to give

$$g(\theta) = \log \left(\frac{\theta^{\frac{1}{N}}}{1 - \theta^{\frac{1}{N}}} \right) \quad (36)$$

Here, we are assuming that ‘success’ or a 1 in the regression represents the absence of the feature of interest. The addition of the power term accounts for the expected reduction in the probability of absence when multiple lineages are present. With this link function we used a Generalised Linear Mixed Model to identify associations between multiple carriage and specific lineages or serotypes and corrected for multiple testing with the BH method.

6.3 RESULTS

6.3.1 Quality control

Of the original 4000 samples selected for sequencing, 3961 were found to have sufficient quantities of DNA after culturing to attempt sequencing. Although a selective agar designed to target *S. pneumoniae*

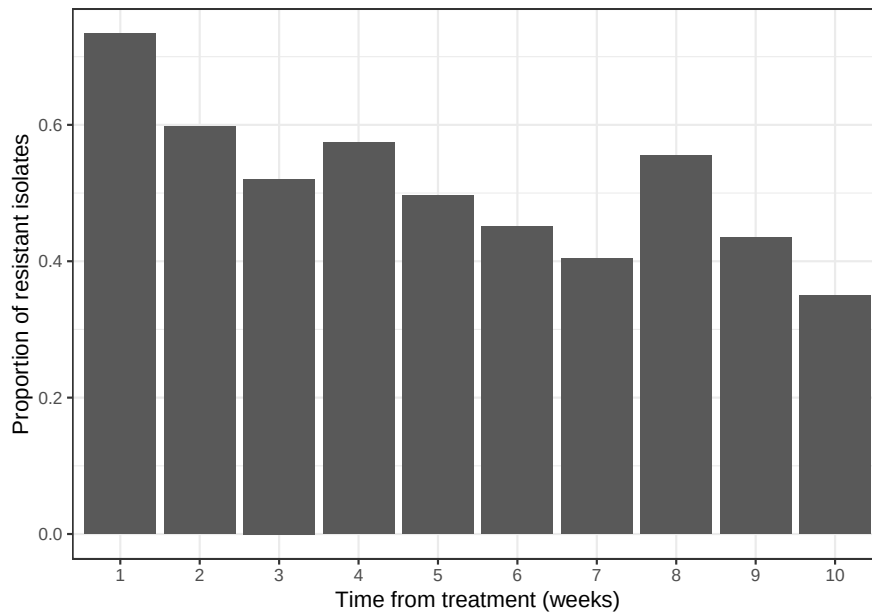


Figure 31: A barplot indicating the proportion of resistant isolates found in children following antibiotic treatment.

was used in the culturing of the whole plate sweeps, it does not eliminate the possibility of related species contaminating the sequence data. The low biomass in some samples can also lead to relatively little bacterial growth in the culture step resulting in samples with poor sequence coverage over the *S. pneumoniae* genome. These issues can hamper efforts to accurately characterise strains and genomic features within deep sequencing data. To exclude those samples that were badly affected by these issues we introduced thresholds on the minimum fraction of both the average depth of sequencing across the *S. pneumoniae* reference as well as the maximum fraction of contaminating reads from different bacterial species.

6.3.1.1 Coverage

Given the chosen plex per sequencing lane, previous sequencing efforts in the GPS project indicated that we should expect a mean sequencing coverage of approximately 300 fold across the pneumococcal reference genome. A histogram of the mean coverage across the ATCC-700669 reference genome for all 3961 non-control samples sequenced is given in [Figure 32a](#). This indicates that the median coverage was 247 fold, slightly below what was expected. In contrast to the GPS project, dual indexing was used as it reduces the probability of ‘index hopping’ [92]. It has been observed that Illumina HiSeqX, HiSeq4000 and NovaSeq sequencing machines are subject to index hopping whereby reads from samples that are multiplexed on the same lane can ‘hop’ between indexes leading to cross-contamination.

As we are interested in low prevalence strains and with observed rates of index hopping of between 0.2-6% it was essential that dual indexes were used to correct for this issue [92]. Whilst dual indexing helps to prevent cross-contamination it has been found to reduce the overall yield of the sequencing run (internal communication with the Sanger sequencing team). In order to concentrate our efforts on those samples where there was sufficient data to retrieve reliable results we excluded samples with a mean coverage below 50 fold representing 20% of the median coverage observed across all samples. Whilst it is hard to choose an optimal coverage threshold, 50x has been shown to be a reasonable coverage for the assembly of bacterial genomes [123].

6.3.1.2 Contamination

The Wellcome Sanger Institute bacterial sequencing pipelines automatically run the Kraken metagenomic classification algorithm on all samples [513]. A histogram of the proportion of each sample assigned by the Kraken algorithm to *S. pneumoniae* is given in Figure 32b, indicating the vast majority of samples consisted predominantly of *S. pneumoniae*. Choosing an exact threshold for excluding contaminated samples is challenging as a very strict threshold can lead to a loss of power in downstream analyses whilst a relaxed threshold may introduce errors into the final results. We attempted to choose a practical threshold of requiring at least 75% of sequencing reads to be assigned to *S. pneumoniae* which led to the exclusion of 20% of the sequenced samples.

To investigate what made up the contamination of the remaining 3188 samples we plotted the estimated proportion as given by Kraken for all species observed in at least 5 samples above 1% in frequency (Figure 33). This indicates that the majority of contamination is assigned to unknown species which could be the result of some *S. pneumoniae* strains not being accurately represented in the Kraken database. *Streptococcus pseudopneumoniae* was the second most prevalent species with the following top 5 species all being of the *Staphylococcus* genus. As *S. pseudopneumoniae* is a close relative of *S. pneumoniae* it is also possible that some of this contamination is driven by divergent strains of *S. pneumoniae* being incorrectly assigned to *S. pseudopneumoniae*. To ensure downstream results are not driven by low level contamination of these species we include a number of further checks at each stage of the analysis to account for the potential impacts of contamination. This includes the use of the Mash filter as described in Section 6.2.1.2 as well as a novel variant filtering strategy detailed in Chapter 8. In addition, throughout the analysis we scrutinize the results to determine whether they could be explained by low level contamination.

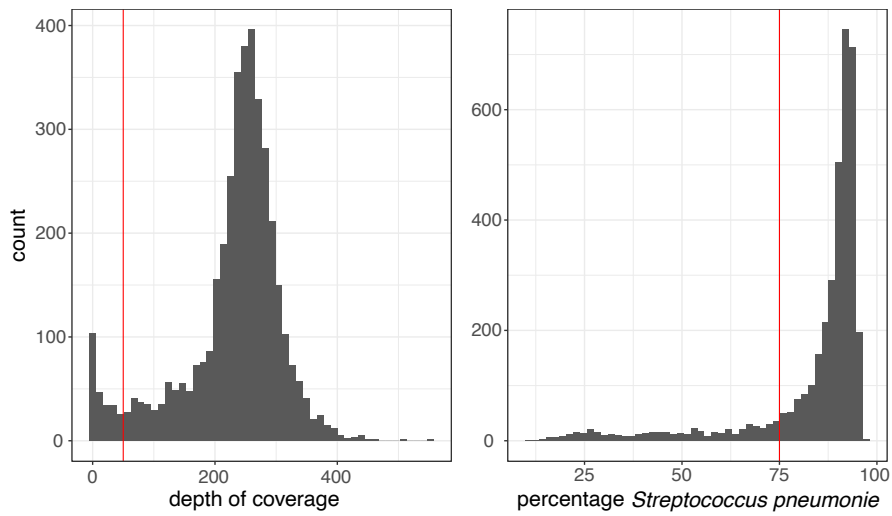


Figure 32: a.) A histogram indicating the distribution of the mean depth of coverage across the ATCC-700669 reference genome found in each of the 3961 sequenced samples. b.) A histogram indicating the proportion of reads assigned to *S. pneumoniae* in each sample. Red lines indicate the quality control thresholds chosen.

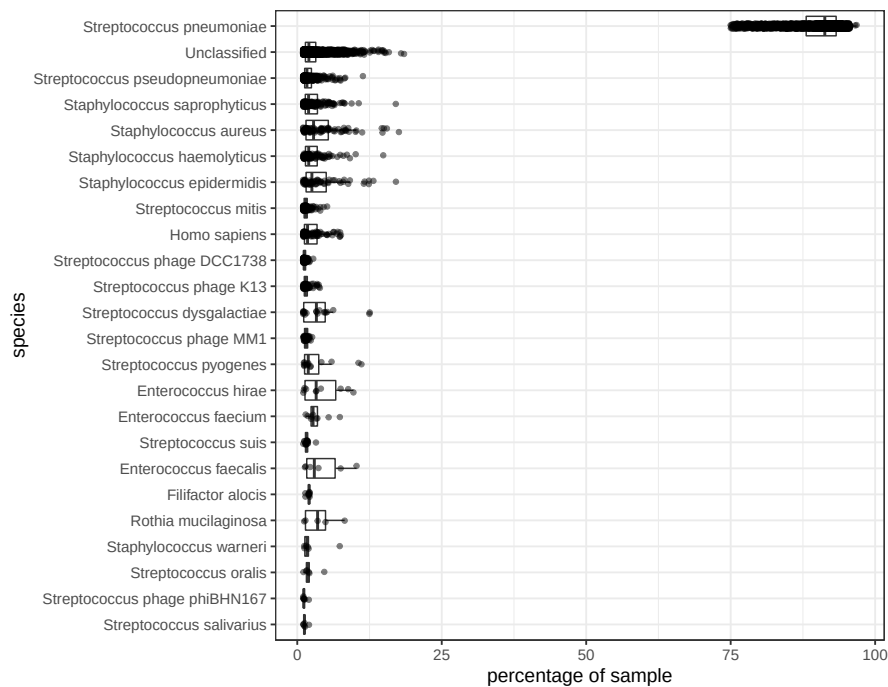


Figure 33: Boxplots of the proportion of reads allocated by the Kraken algorithm to different species in the 3188 samples that passed the initial quality control thresholds.

6.3.2 Verification of mGEMS/Mash pipeline

To verify that the mGEMS pipeline is able to accurately determine the lineages present in mixed samples we ran the pipeline on 44 laboratory generated mixtures of between 1 and 10 lineages [267]. Figure 34a indicates that the default mGEMS pipeline was able to achieve a precision of 0.92 with a recall rate of 0.93. As mGEMS, and the mSWEEP algorithm it relies on, are known to suffer when lineages are present that are not found in the database, we also considered adding the additional condition that any mGEMS call must be supported by a Mash screen call with at least 990/1000 hashes. This threshold was found empirically to exclude lineages that were not found within a mixture. Figure 34b indicates that including this additional requirement effectively removed the small number of false positives without any loss in recall. The remaining false negative calls were found to occur at low frequencies in more complicated mixtures of 10 lineages. As we do not see this level of multiplicity of infection in practice, this is unlikely to have a large impact on our results.

To verify the approach on a larger and more diverse dataset we considered the 1158 samples which have been sequenced previously using single colony picks [71]. Whilst it is not possible to determine the false positive rate using these samples 1149/1183 (97.1%) of serotypes that were found originally using single colony picks were refound using the deep sequencing data and the mGEMS/Mash pipeline Figure 34c. This suggests that the sensitivity observed in the known laboratory mixtures extends to the much larger and more diverse dataset.

6.3.3 Verification of serotype calls

To verify that both the combined SeroBA/mGEMS pipeline and the Serocall algorithm could accurately identify serotypes within mixed samples we compared the results using the same test datasets as used in the verification of the lineage level calls. In addition we were able to take advantage of previous serotype calls obtained via latex sweep for all the samples from the Maela collection.

Using the 44 laboratory mixtures generated described in Knight *et al.*, [267], we ran the seroBA algorithm on the output of the mGEMS pipeline after removing those that were not supported by Mash calls. This approach achieved a precision of 1 with a recall of 0.93. This corresponds to seroBA correctly identifying the serotype in all lineages that were correctly deconvoluted by the mGEMS/Mash pipeline indicating that the pipeline is able to achieve a high level of accuracy.

To test both approaches on a much larger collection, we ran the mGEMS/Mash approach as well as the Serocall algorithm on all 3188 samples that passed the initial quality control filters. Figure 35a indicates that 65.4% of serotype calls were consistent across all three

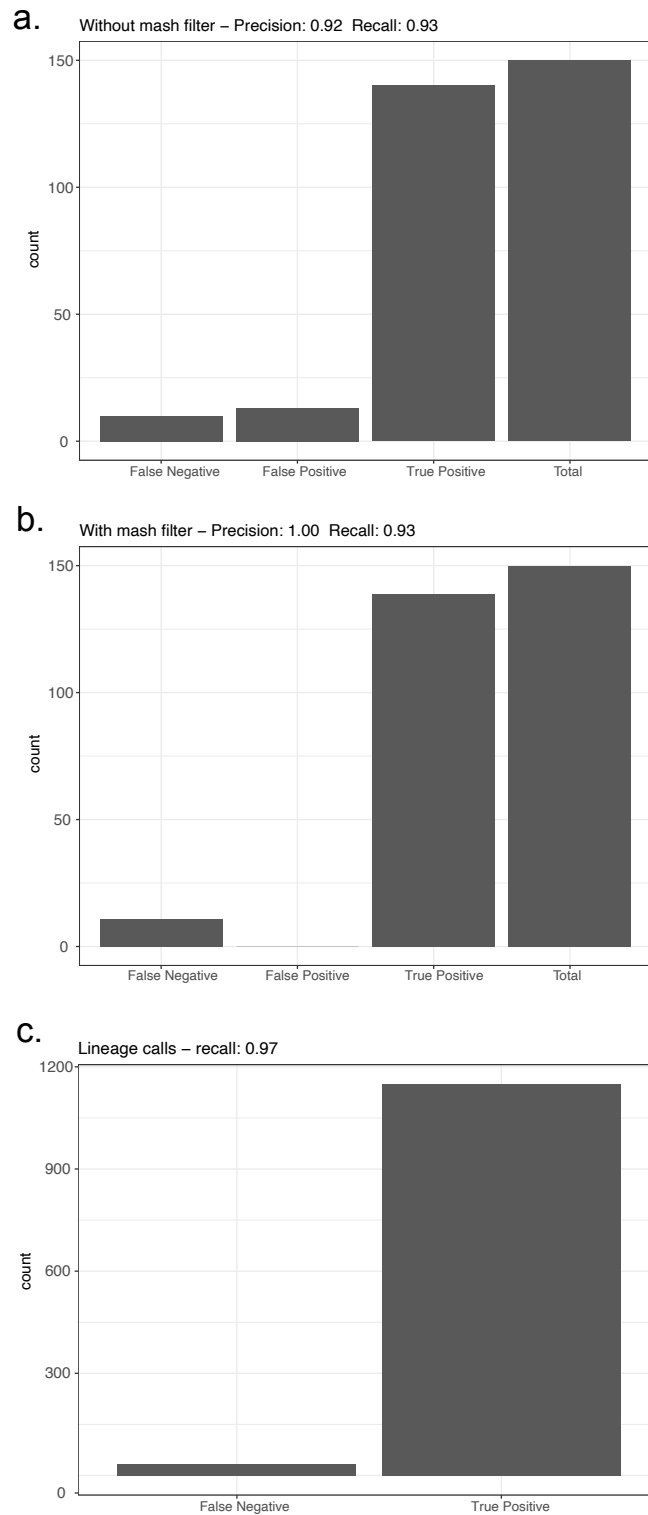


Figure 34: a.) A barplot of the comparing the number of errors and correct lineage calls in the analysis of 44 laboratory mixtures of pneumococcal strains. b.) similar to a. but with the additional requirement that lineages be called using the Mash algorithm. c.) False Negative and True Positive calls for 1158 samples from the Maela dataset which were sequenced using both single colony picks and whole plates

methods. The majority of the discrepancy between the approaches was caused by the increased sensitivity of the deep WGS based approaches with an additional 12.9% of calls being found by both Serocall and the mGEMs/seroBA pipeline.

In contrast to the high concordance between the latex sweep and deep sequencing based methods in identifying serotypes there was a higher level of discrepancy when calling non-typable (NT) strains (Figure 35b). Both the Serocall and SeroBA algorithms use two NT reference sequences (NT_alib and Swiss_NT) to identify potential NT strains. However, these are not universally present in all NT isolates. The mGEMs pipeline is able to identify potential NT strains when there is no clear capsular sequence present but this may also be the result of poor coverage over the capsular genes following the deconvolution step. Previous comparisons have suggested that microarray based approaches are more sensitive in identifying NTs than latex sweeps [490]. As both microarrays and our approach are sequence based this could potentially explain some of the additional NT calls. Overall, we find that despite the high level of concordance at the lineage and serotype level, identifying NTs accurately is challenging. For this reason, we focus most of our analysis on either serotype or lineage calls.

As the Serocall algorithm is able to identify serotypes below the 1% abundance threshold required for the mGEMs pipeline we can improve the sensitivity of the calls by taking the union of the two approaches. The two algorithms generally agree at the serogroup level. However, they can provide conflicting calls at the type level. To resolve conflicting type calls we chose to take the mGEMs/seroBA call as this pipeline was found to better agree with the latex sweep results (Figure 36b). After taking the union we correctly recover 93.6% of serotypes originally identified by latex sweeps. We identify an additional 874 (25.3%) serotypes through the use of the deep sequencing based approach highlighting the higher sensitivity of the method. This is consistent with the higher sensitivity observed through the use of microarrays [490]. Figure 36a compares the estimated abundances of serotypes called using both the mGEMs/seroBA pipeline and Serocall and suggests that both methods provide broadly consistent estimates. A number of points in Figure 36a are concentrated at 100% in only one of the two algorithms. This is mostly driven by discrepancies in the calling of non-typables as Serocall was not designed to recover strains without capsular sequences.

6.3.4 Deep sequencing increases sensitivity in detecting multiple carriage

Accurately identifying cases of multiple carriage is important in gaining a thorough understanding of the epidemiology and disease dynamics of *S. pneumoniae*. Multiple carriage allows for the exchange of

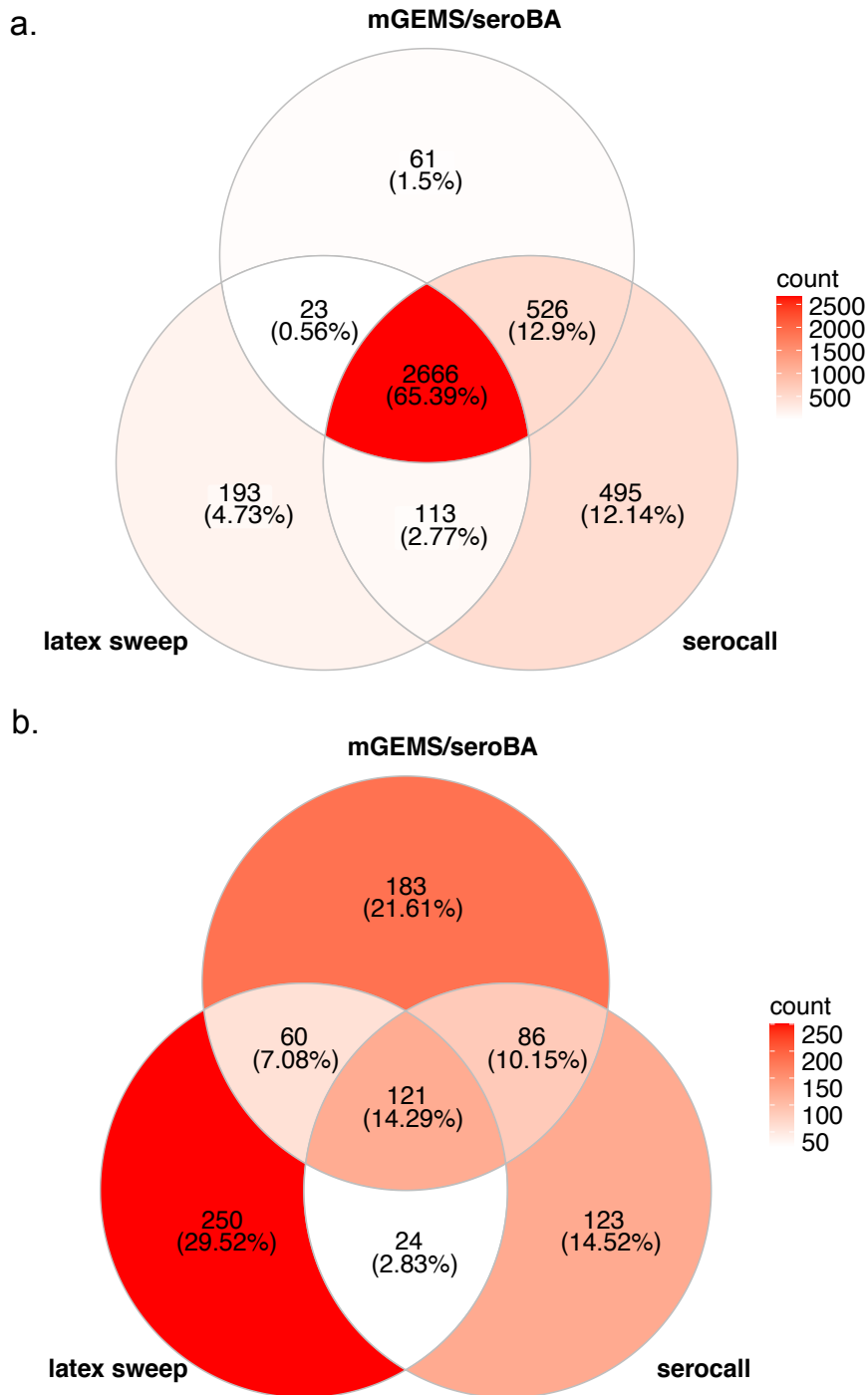


Figure 35: a.) A Venn diagram of the intersection of the serotype calls obtained using latex sweep as well as the in-silico methods Serocall and the mGEMS/seroBA pipeline. b.) similar to a. but comparing the non-typable calls of each of the three methods.

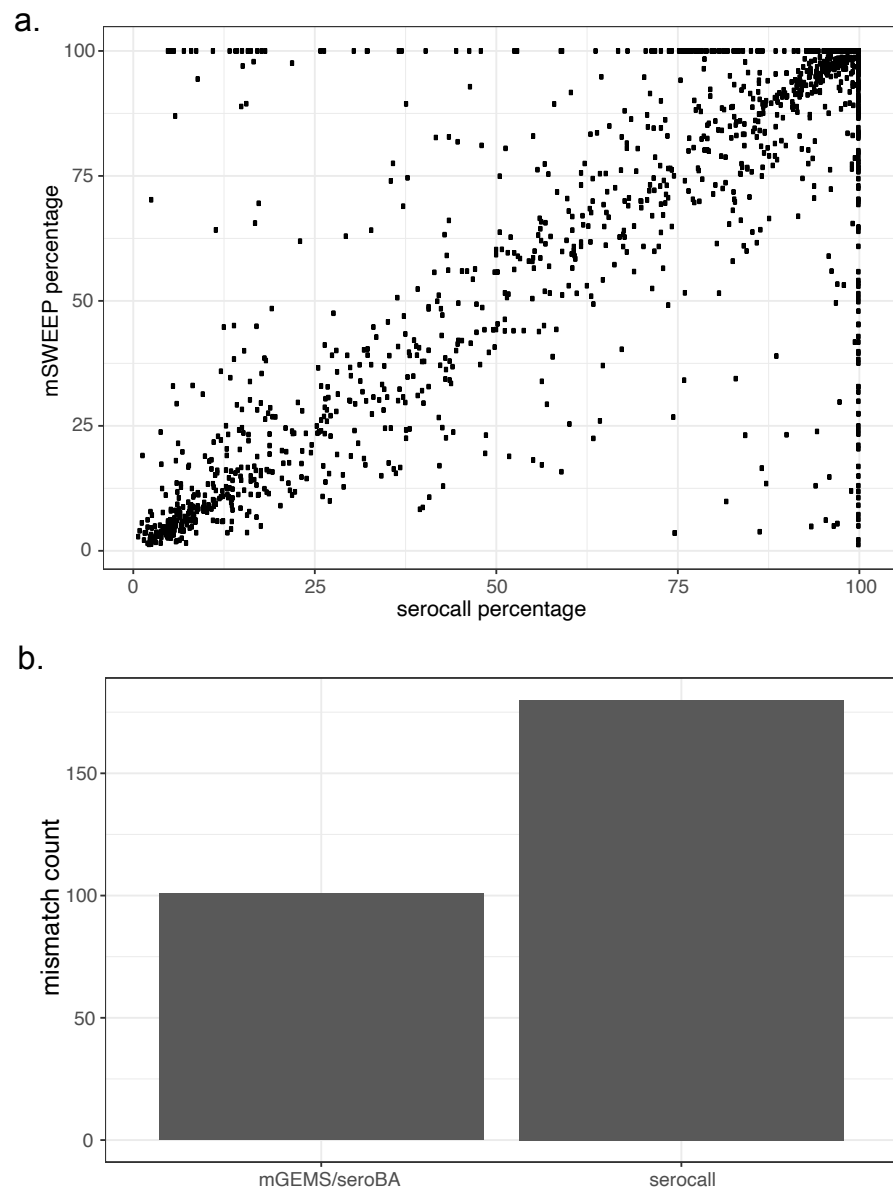


Figure 36: a.) a dotplot indicating the concordance between the estimated serotype proportions obtained using Serocall and mSWEEP after identifying the serotype using the mGEMS/seroBA pipeline. b.) The number of mismatches between the serotypes observed using latex sweeps and those found using Serocall and the mGEMS/seroBA pipeline.

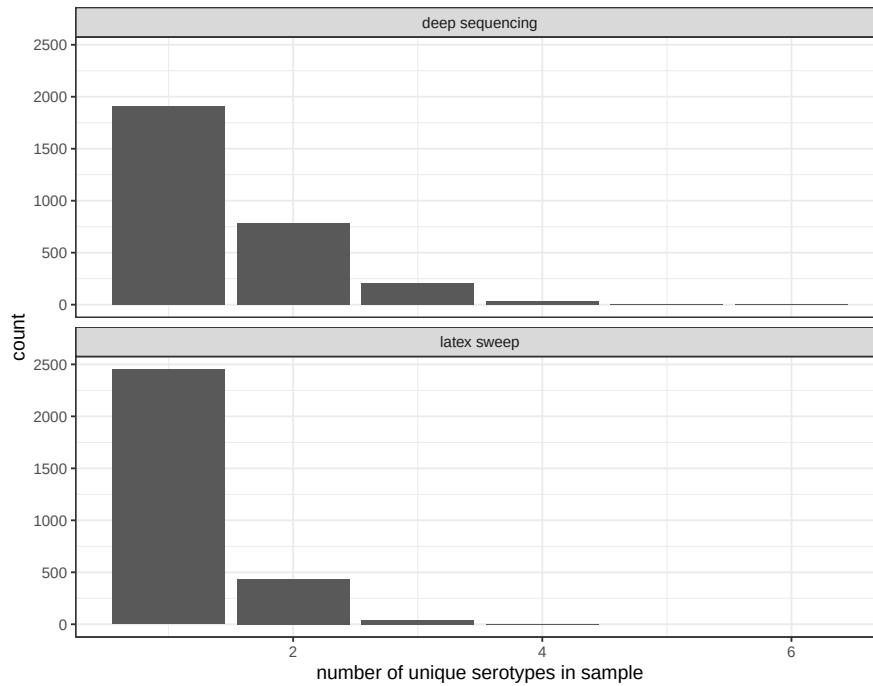


Figure 37: The distribution of the number of unique serotypes found in samples separated by those found using whole plate sweeps with deep sequencing and latex sweeps.

genetic material including resistance determinants and virulence factors including capsular genes. As the most common pneumococcal vaccines target a subset of the known *S. pneumoniae* serotypes, switching of capsular sequences can have a major impact on our ability to control the pathogen using vaccines [81, 102]. Similarly, the sharing of resistance elements within cases of multiple carriage allows susceptible strains that are not targeted by current vaccines to acquire resistance, thus threatening the effectiveness of vaccination programs for reducing resistance.

Across the 3188 samples that passed our quality controls filters, deep sequencing identified an additional 874 lineages, representing an increase of 17.7% from that found using latex sweeps. The distribution of the multiplicity of infection is shown in Figure 37 and suggests that latex sweeps consistently underestimates the rates of multiple infection ($p < 0.001$, 2-sided t-test). Overall the estimated rate of multiple infection increased from 16.2% using latex sweeps to 35.0% with the deep sequencing approach. The maximum number of distinct strains found in a single sample was 6, found in one sample from a child who had recently been treated with antibiotics.

The substantial increase in the sensitivity of a whole plate sweep, deep sequencing based approach suggests it could provide substantial benefits in the genomic surveillance of *S. pneumoniae*.

6.3.5 *Differences in carriage rates between mothers and children*

Immunity to different pneumococcal strains tends to increase through childhood with adults usually developing immunity to a wider range of pneumococcal serotypes. Historically, this was the result of the accumulation of exposure through multiple colonisation events. However, more recently the introduction of pneumococcal conjugate vaccines have changed the dynamics with which immunity is acquired. As the samples considered here were collected in the Maela refugee camp prior to the introduction of a pneumococcal vaccine we were interested to see whether the prevalence of multiple carriage, serotypes and lineages differed between mothers and children. The mean number of serotypes found in each host split into mothers and children is shown in [Figure 38a](#). It indicates the significant difference between the prevalence of multiple carriage in mothers and children ($p < 0.001$, two sided t-test). The majority of carriage events in mothers consisted of only a single strain which is consistent with the higher rates of immunity to *S. pneumoniae* serotypes in adult populations.

[Figure 38b](#) and [38c](#) indicates those serotypes and lineages which have a significantly different prevalence between mothers and children. The most common serotypes 19F and 23F differ most with a smaller proportion of mothers found to be infected with these strains. This is consistent with the high prevalence of these strains leading to a higher level of immunity in older individuals. Similarly, GPSCs 1 and 20 are the most prevalent in this dataset and are also found at lower rates in mothers.

Mothers tended to be colonised with rarer serotypes and lineages. Interestingly, serotype 3 and the corresponding GPSC 128 were found at higher rates in mothers. Serotype 3 pneumococci exhibit less genetic variation than other strains and have also been implicated in higher rates of invasive disease. In particular, the risk of invasive serotype 3 disease has been shown to increase with age and as carriage is a precursor to disease is consistent with our results [240, 439]. It has been suggested that the association with age may correspond to the high immunogenicity of the capsule in children [103, 138]. Despite its inclusion in the PCV13 vaccine serotype 3 has continued to be a major cause of invasive disease.

6.3.6 *Prevalence in multiple carriage varies by lineage and serotype which has implications for surveillance and models of pneumococcal strain structure*

The majority of current pneumococcal surveillance strategies rely on the sequencing of single colony picks. As up to one third of our samples consist of multiple pneumococcal lineages, surveillance strategies based on the analysis of a single genome per sample could po-

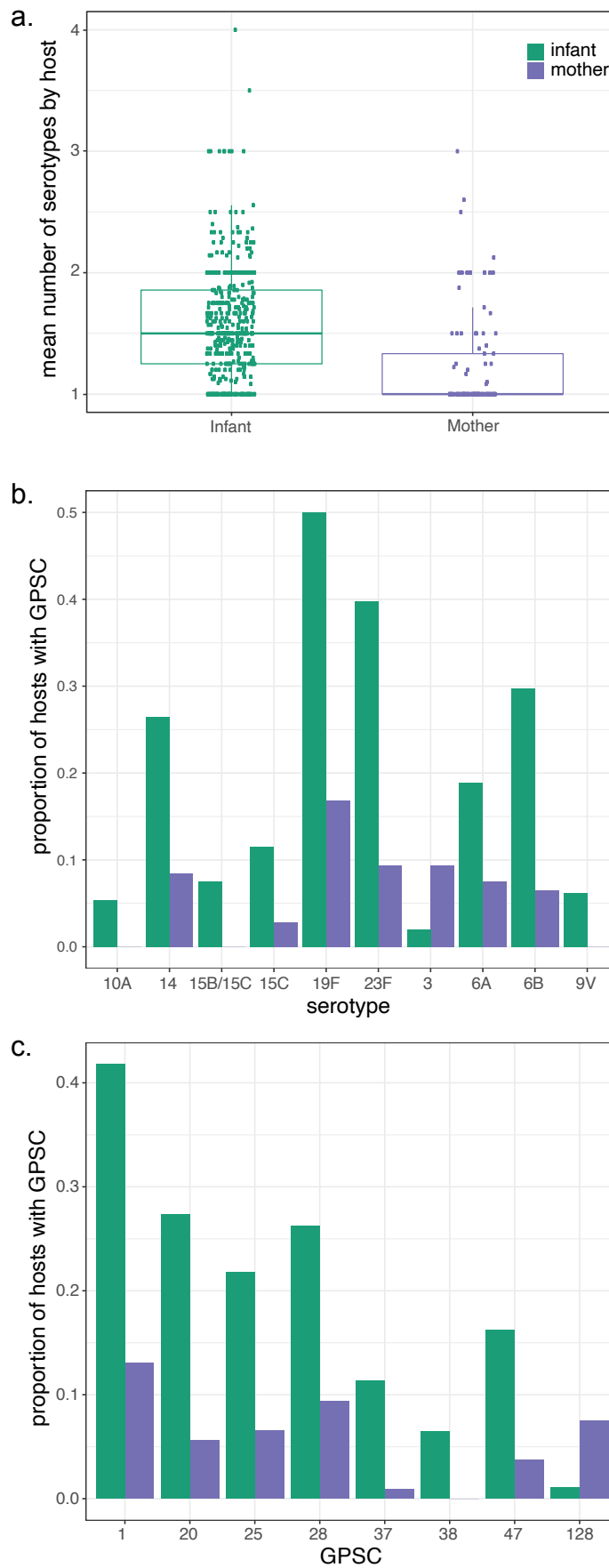


Figure 38: a.) a boxplot of the distribution in the multiplicity of infection of mothers versus children. b.) the proportion of hosts found to carry each serotype found to have significantly different carriage rates in mothers and children. c.) Similar to b. but summarised at the lineage (GPSC) level.

tentially lead to biased estimates of the prevalence of different lineages. Multiple carriage events also have implications for the evolutionary dynamics of the pneumococcus as co-colonisation allows for the transfer of genetic material between lineages.

To investigate the potential impacts of frequent multiple carriage on surveillance, we considered both whether certain lineages and serotypes preferentially occurred in mixed infections as well as whether lineages were more likely to be found at lower frequencies indicating they may often be missed by strategies focusing on single colony picks. To compare our whole plate sweep strategy with the more common single colony pick approach we examined 1158 samples which were sequenced both using whole plate sweeps in this study as well as single colony picks in the previous study of Chewapreecha *et al.*, [71].

The estimated proportions of each lineage in each sample are given in Figure 39. As expected the majority of proportions are found at 100%, consistent with the majority of our samples consisting of a single lineage. However, a number of lineages appear to occur frequently at lower proportions when compared to the overall distribution. This pattern was reflected in the distribution of serotype proportions as is expected given the high level of correspondence between serotypes and lineages (Figure 40). To identify those lineages with unusual abundance patterns we compared the distribution of each lineage and serotype with the combined dataset using a one sided Kolmogorov-Smirnov test. To account for multiple samples from the same host we only considered the highest proportion found within a single host for each serotype and lineage. Those lineages and serotypes with significantly outlying proportions are given in Table 6 and Table 7 respectively.

6.3.6.1 *Significant associations between lineage virulence and multiple carriage*

Strikingly, lineages that were found to occur more often than expected at lower proportions were mostly composed of those known to be associated with invasive pneumococcal disease and non-typables. Lineages 2 and 28 have both been associated with high rates of invasive disease with GPSC2 consisting of serotype 1 isolates [190]. Similarly, performing the same analysis at the serotype level also identified serotype 1 as well as serotype 8 which is also often found in invasive disease [190, 311]. In both analyses, non-typables were found to have significantly different abundance patterns although this is likely to be impacted by the lower accuracy our approach has to identify non-typable isolates.

Whilst our use of the Kolmogorov-Smirnov test allows for the identification of outlying lineages and serotypes in terms of the distribution of abundance, this does not necessarily correspond with those

lineages being overrepresented in multiple carriage. The abundance estimates are also dependent upon an initial culture step and thus any biases introduced during culture would impact the results. As an alternative, we fit a logistic mixed model using a modified link function as described in [Section 6.2.4](#). Instead of relying on abundance estimates we tested whether some lineages or serotypes occurred in multiple carriage at higher or lower rates than expected. We included the host as a random effect and treatment as a fixed effect. The top results at the lineage and serotype level are given in [Table 8](#) and [Table 9](#) respectively. Overall the regression based approach confirmed the results based on the KS test. Both lineage 28 and the non-typable lineages were found to be overrepresented in cases of multiple carriage. Whilst neither serotype 1 or 8 reached significance in this analysis the estimated effect was in the same direction (more likely in mixed infection) and thus this is likely to be the result of the relatively fewer examples of these serotypes leading to insufficient power to detect an association.

Collectively, these results suggest that the common practice of sequencing single colony picks may lead to an underestimate of the prevalence of virulent pneumococcal strains. This could have important implications for both the treatment and the surveillance of *S. pneumoniae* as carriage based surveillance studies may systematically underestimate the prevalence of virulent serotypes 1, 8 and 14.

6.3.6.2 *Lineage competition within the host*

The logistic mixed model based approach also identified a number of lineages that were associated with being less likely to occur in mixed carriage. These included GPSC₁ and GPSC₁₀ which are both known to have high rates of antibiotic resistance [190]. Association between resistance and reduced rates of multiple carriage could result from antibiotic resistance conferring a fitness cost, leading to resistant strains being outcompeted within the host and reducing their frequency in cases of multiple infection. Alternatively, it is possible that these lineages outcompete other GPSCs resulting in them being rarely observed in multiple carriage. An increased ability to outcompete other lineages and persist in carriage could then result in an increased selective pressure to develop resistance determinants due to the increased likelihood of exposure to antibiotics brought on by longer carriage duration [294].

The observation that different pneumococcal lineages appear in multiple infections at different rates has important implications for models of the maintenance of pneumococcal strain structure. Recent efforts to model the coexistence of resistant and susceptible pneumococcal strains in populations have relied on within-host competition of pneumococcal strains to drive negative frequency dependent selection (NFDS) of resistant strains [111]. This model is consistent with

GPSC	statistic	p-value	adj p-value	serotypes
42	0.413	1.11E-05	5.65E-04	NT
60	0.365	1.06E-04	2.70E-03	NT
28	0.163	1.73E-03	2.93E-02	14
2	0.446	2.67E-03	3.40E-02	1

Table 6: The GPSCs found to be significantly different from the combined frequency distribution of all GPSCs using the Kolmogorov-Smirnov test.

serotype	statistic	p-value	adj p-value
NT	0.516	7.70E-28	3.90E-26
1	0.472	1.50E-04	2.80E-03
8	0.853	1.70E-04	2.80E-03

Table 7: The serotypes found to be significantly different from the combined frequency distribution of all serotypes using the Kolmogorov-Smirnov test.

our data. However, an alternative model suggests that the maintenance of intermediate levels of antibiotic resistance in the population is driven by genetic linkage between resistance and carriage duration determinants [294]. As already mentioned, if the phenotypes driving extended carriage duration also lead to reduced rates of multiple carriage then our data could equivalently be explained by this model. Finally, NFDS models of the distribution of accessory genes in the pneumococcal population do not currently incorporate structured sharing of accessory loci between different pneumococcal lineages [86, 210].

6.3.7 Little evidence for founder advantage in pneumococcal transmission

Distinct from the previous models of pneumococcal strain structure, recent experimental and modeling studies have proposed an occupant-intruder competitive strategy whereby the first variant to colonise a host is able to prevent later arriving variants from contributing to infection [244, 449]. To investigate whether there was support for this model in our dataset we considered all instances of consecutive monthly samples within a single host where a GPSC initially appeared in isolation and was then found to co-colonise with one additional GPSC. We restricted our analysis to those samples unaffected by antimicrobial treatment. We identified 59 such instances. Of these, in 33 cases the initial GPSC appeared as the minority at a frequency below 50% in the subsequent co-colonisation sample. This was not significantly different ($p=0.435$, binomial test) from what we might

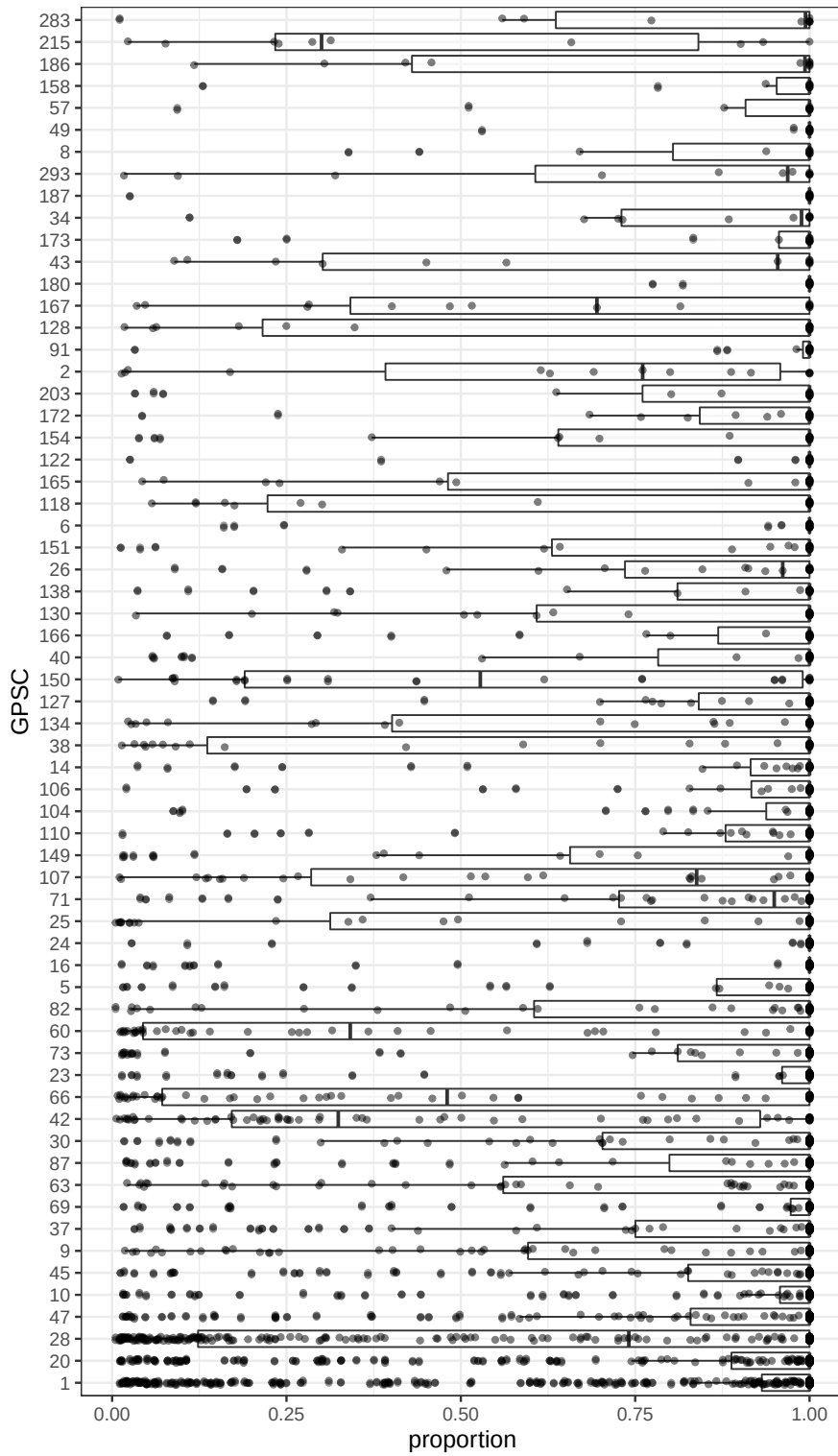


Figure 39: Boxplots indicating the distribution of the inferred within-host frequencies for each GPSC.

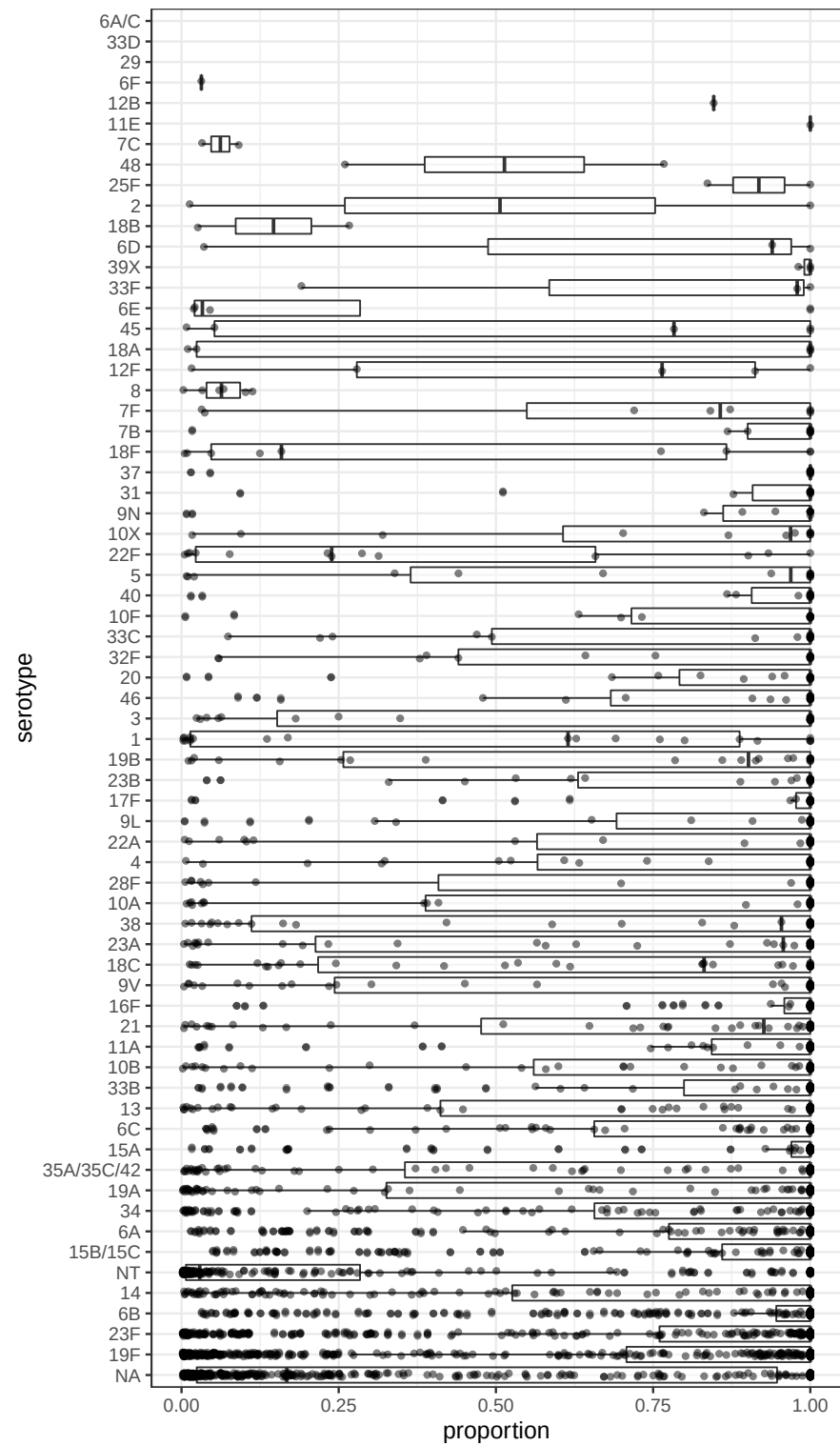


Figure 40: Boxplots indicating the distribution of the inferred within-host frequencies for each serotype. Here, the final merged results of both the mSWEEP and Serocall algorithms are given. Deconvoluted samples that could not be accurately assigned to either a serotype or one of the known non-capsular sequences are designated 'NA'.

GPSC	estimate	std.error	statistic	p-value	adj p-value	serotypes
1	0.718	0.120	5.963	0.00E+00	1.00E-07	19F,19A
28	-0.569	0.153	-3.711	2.06E-04	4.02E-03	14
60	-1.382	0.401	-3.452	5.57E-04	7.24E-03	NT
42	-1.251	0.376	-3.329	8.71E-04	8.49E-03	NT
24	1.174	0.395	2.971	2.97E-03	2.31E-02	23F
66	-1.078	0.379	-2.842	4.49E-03	2.92E-02	
10	0.515	0.206	2.504	1.23E-02	6.84E-02	19A,17F,7B
20	0.362	0.149	2.423	1.54E-02	7.50E-02	23F
6	1.290	0.577	2.235	2.54E-02	1.10E-01	9V
47	0.399	0.194	2.056	3.98E-02	1.44E-01	6B,6A
23	0.703	0.349	2.016	4.38E-02	1.44E-01	6B,6A
69	0.531	0.264	2.011	4.43E-02	1.44E-01	15A
149	0.684	0.382	1.792	7.32E-02	2.20E-01	28F,32F

Table 8: The GPSCs found to be present in multiple carriage at either significantly higher (estimate<0) or lower (estimate>0) rates than expected given their prevalence in the population.

serotype	estimate	std.error	statistic	p-value	adj p-value
unknown	-1.607	0.137	-11.691	0.00E+00	0.00E+00
15B/15C	0.621	0.165	3.765	1.66E-04	4.24E-03
6B	0.473	0.141	3.353	7.99E-04	1.36E-02
19F	0.324	0.109	2.970	2.98E-03	3.80E-02
23F	0.297	0.123	2.416	1.57E-02	1.60E-01
1	-1.300	0.558	-2.329	1.98E-02	1.69E-01
22F	-2.352	1.129	-2.084	3.72E-02	2.71E-01
15A	0.457	0.262	1.745	8.10E-02	4.97E-01

Table 9: The serotypes's found to be present in multiple carriage at either significantly higher (estimate<0) or lower (estimate>0) rates than expected given their prevalence in the population.

expect if there was no competitive advantage of colonising first assuming a null distribution of there being an even chance either GPSC appears as the dominant lineage in the co-colonised sample. However, this result is reliant on the estimated lineage frequencies within each sample which may be subject to variation introduced by the culturing step prior to sequencing. To further investigate the occupant-intruder hypothesis we considered the subset of these cases where additional samples following the co-colonisation were available indicating which lineage prevailed in the longer term. Of the 59 cases we had additional longitudinal resolution on 15 cases. Similar to the previous finding 7/15 of the cases consisted of the founding GPSC maintaining its position and returning to a single lineage infection following the co-colonisation event. A higher resolution description of two of the 15 cases is given in [Figure 41](#).

Although this result does not support the occupant-intruder model of pneumococcal colonisation it does not conclusively demonstrate that there is no advantage to colonising first. The limited number of examples coupled with the low temporal resolution of the samples makes it hard to accurately determine whether such competition exists. For instance, it is plausible that competitive-exclusion generally occurs on a much shorter time scale than our monthly sampling schedule, preventing us from observing the majority of interactions.

To investigate whether colonisation with a serotype reduced the probability of a subsequent colonisation event consisting of the same serotype we considered the inferred carriage episodes found in Lees *et al.*, [287]. Here, the authors made use of a Hidden Markov Model to accurately characterise distinct carriage episodes in the Maela dataset from the serotype presence/absence data provided by latex sweeps. To generate a null model of the expected distribution of carriage in the absence of immunity we permuted the serotype labels for each sample. Using 10,000 permutations, we found that the expected number of repeated carriage episodes of the same serotype was significantly less than would be expected in the absence of immunity ($p < 1e-3$). This is consistent with the long established theory of serotype specific immunity in *S. pneumoniae*. Thus, although we do not find evidence of a competitive advantage to the strain which first colonises a host, this may not be true when considering strains with the same serotype.

6.3.8 Evidence for preferential co-colonisation

Preferential co-colonisation between different pneumococcal strains can impact the evolution of the pathogen, surveillance and treatment decisions. In the absence of an intermediate species, *S. pneumoniae* strains must co-colonise in order to exchange genetic material. Thus gaining an understanding of whether certain strains preferentially

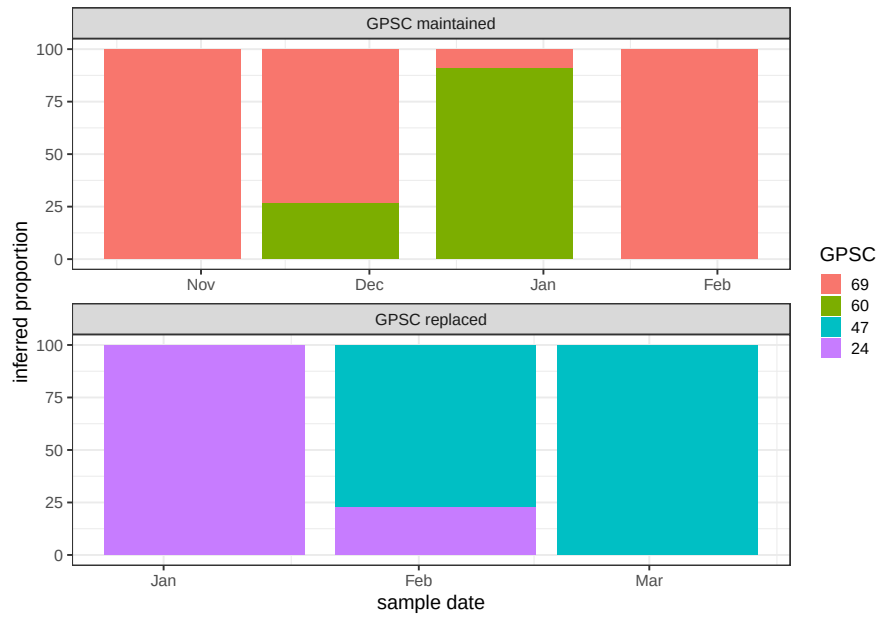


Figure 41: An example of competing lineages within two independent hosts. In the top facet, GPSC 69 colonises first and ultimately prevails. Thus, GPSC 69 would be considered to excluded the intruding GPSC 60 lineage. In the lower facet GPSC 24 is ultimately overtaken by the intruding lineage GPSC 47.

co-colonise can help identify how genomic features and their corresponding phenotypes move through the broader *S. pneumoniae* population. Understanding co-colonisation dynamics can also inform treatment decisions. If a more virulent lineage is found more often with another benign lineage, it may motivate more careful observation of patients found to have the benign strain.

To date, most studies investigating co-colonisation have lacked the power to identify significant associations due to the large number of potential pairwise interactions [43, 44]. Whilst this still remains a challenge, the large sample size of our dataset coupled with the increased sensitivity of the deep sequencing based approach enables us to identify signals of co-colonisation.

To investigate co-colonisation we used a logistic mixed model to account for host effects. For each pairwise interaction we compared the presence of one lineage as the dependent variable with the presence of the other lineage. We performed the same analysis both at the GPSC and serotype levels. After adjusting for multiple testing (BH) only two significant associations were observed, both involving non-typables with serotypes 19F and 6B. Serotypes 19F and 6B are the 1st and 3rd most prevalent serotypes in our dataset indicating that even in this relatively large collection we still have limited power to observe co-colonisation associations between lineages. Although our ability to accurately identify non-typables could have im-

pacted this result, the preferential co-colonisation of non-typeables with these two dominant serotypes supports past findings that non-typeables play a significant role in the evolution and genetic exchange of *S. pneumoniae* [71].

6.3.9 *Virulence factors are negatively associated with multiple carriage*

The pneumococcal polysaccharide capsule (CPS) is the major virulence factor of the pathogen which induces type specific antibodies and is the target of most pneumococcal vaccines. Unfortunately, poor immunogenicity of CPS antigens has been observed in children. While this has been overcome to some extent by the introduction of multi-serotype protein conjugate vaccines PCV7, PCV10, PCV13 and most recently PCV23, the observed replacement of vaccine types with non-vaccine types has reignited interest in identifying alternative vaccine targets that are common to all serotypes [100, 253].

To investigate the association between non-CPS virulence factors and multiple carriage we downloaded a database of known *Streptococcus* virulence factors from the Virulence Factor Database (VFDB) [66]. We used Ariba to identify the presence or absence of each feature in the deep sequenced samples and employed the same logistic mixed model as used for the analysis of serotypes to identify associations between the determinants and multiple carriage. To account for lineage effects we performed principal component analysis on the binary presence absence matrix of lineages per sample. We included the first 10 principal components as additional covariates in the regression. Whilst this will not adjust for fine scale population structure it should account for lineage level signals and is a common approach used in bacterial genome wide association studies [289]. It should be noted that Ariba groups highly similar gene sequences into clusters and that variants of the same ortholog can be represented by different clusters.

Unlike that seen in the analysis of serotypes, the majority of virulence factors included in this database were found to be negatively associated with multiple carriage (Table 10). These included variants of the choline binding protein A (CbpA), which is involved in upper-airway colonisation and enhances the pneumococci's ability to evade complement-mediated innate immunity [253]. Other genes negatively associated with multiple carriage included 3 variants of the hyper-variable *pspA* gene, shown to inhibit complement-dependent host defense mechanisms and variants of zinc metalloproteases *IgA* and *zmpB* found in 3 and 4 distinct variant clusters respectively. Variants of these genes have been implicated as having a role in adhesion and inflammation [37, 488]. Variants of both *lytA* and *lytB* genes were also found to be less likely to occur in multiple carriage. These genes are known to be involved in colonisation with *lytA* also linked to

fratricidal lysis [147]. Whilst it is feasible that many of these gene variants could enhance a lineage's ability to exclude or outcompete other strains, particularly in the case of *lytA*, these variants could also confer a fitness cost which would reduce their frequency in multiple carriage.

Interestingly, the most significant genes associated with an increase in their presence in multiple carriage included those involved with type 2 pili (PI-2). These included all known genes in the PI-2 region including *pitA*, *pitB*, *srtG-1*, *srtG-2* and *sipA*. PI-2 has been shown to mediate adhesion to host cells and is associated with the virulent serotypes 1 as well as serotype 2 which has also been found to be the dominant serotype in invasive disease in some instances [16, 432]. Supporting this link, the capsule polysaccharide biosynthesis protein *wzh* found in serotype 2 was also associated with being more common in multiple carriage. Whilst serotype 2 is very rare in our dataset, having been found in only one mother child pair, this result is consistent with the observed association with PI-2. Other capsular proteins found at increased rates in multiple carriage included the variants of the *wze* gene commonly found in serotype 4. In contrast to *zmpB* variants, a *ZmpC* variant cluster was found to be positively associated with multiple carriage. *ZmpC* has been shown to be involved with migration across the blood-brain barrier [353] as well as mediating shedding of syndecan-1 in response to inflammation [67].

Intriguingly, there appears to be a distinction between virulence factors associated with the pneumococcal capsule including serotypes 1, 2 and 8 which appear more often in cases of multiple carriage and other known virulence associated genes outside the capsular sequence which are negatively associated with multiple carriage. Although it is not clear what exactly causes this distinction, it is likely related to the different mechanisms that drive the associations between these genetic factors and virulence.

6.4 DISCUSSION

The increased sensitivity to detect multiple lineages and serotypes provided by deep sequencing whole plate sweeps over single colony picks or latex sweeps presents a strong argument for the routine use of deep sequencing in the surveillance of *S. pneumoniae*. By processing a large number of samples using a number of different technologies, we have demonstrated that it is possible to robustly identify both pneumococcal lineages and serotypes from sequenced whole plate sweeps. Critically, this revealed that well known virulent lineages and serotypes, including serotype 1, are often present in multiple carriage events at low frequencies. As tracking the effectiveness of pneumococcal vaccines often relies on either latex sweeps or WGS of single colonies, these technologies could potentially be system-

virulence factor	estimate	std.error	statistic	p-value	adj p-value
pspC_cbpA_1	1.364	0.082	16.598	7.20E-62	5.90E-60
cbpA	1.174	0.086	13.659	1.80E-42	7.30E-41
zmpB_3	1.882	0.140	13.451	3.00E-41	8.30E-40
pspA_2	1.244	0.097	12.809	1.50E-37	3.00E-36
lytA	2.043	0.162	12.634	1.40E-36	2.30E-35
lytB	1.909	0.154	12.427	1.90E-35	2.60E-34
pspA_1	0.864	0.083	10.408	2.30E-25	2.70E-24
pspA	0.956	0.094	10.160	3.00E-24	3.10E-23
pspA_3	1.108	0.113	9.824	8.90E-23	8.10E-22
iga	0.901	0.110	8.217	2.10E-16	1.70E-15
pspC_cbpA	0.886	0.120	7.387	1.50E-13	1.10E-12
iga_1	0.826	0.112	7.363	1.80E-13	1.20E-12
eno+	4.355	0.612	7.121	1.10E-12	6.80E-12
zmpB	0.819	0.116	7.053	1.80E-12	1.00E-11
pspC_cbpA_2	0.571	0.085	6.730	1.70E-11	9.30E-11
zmpB_2	0.751	0.127	5.918	3.30E-09	1.70E-08
pitB	-1.793	0.318	-5.631	1.80E-08	8.70E-08
srtG1	-1.472	0.293	-5.014	5.30E-07	2.40E-06
lytB_1	0.636	0.136	4.682	2.80E-06	1.20E-05
iga_3	1.723	0.372	4.637	3.50E-06	1.50E-05
srtG2	-1.466	0.319	-4.595	4.30E-06	1.70E-05
wzd_1	-0.562	0.124	-4.533	5.80E-06	2.20E-05
rrgB_1	0.756	0.176	4.296	1.70E-05	6.20E-05
pitA	-1.239	0.301	-4.112	3.90E-05	1.30E-04
pclA	-0.491	0.125	-3.915	9.00E-05	3.00E-04
zmpC_1	-0.705	0.189	-3.724	2.00E-04	6.20E-04
wchA	-0.438	0.125	-3.495	4.70E-04	1.40E-03
cps4D+	-0.392	0.113	-3.475	5.10E-04	1.50E-03
sipA	-1.046	0.308	-3.391	7.00E-04	2.00E-03
zmpB_1	0.388	0.118	3.295	9.80E-04	2.70E-03
wzx	0.685	0.213	3.218	1.30E-03	3.30E-03
SPH_0459+	-0.358	0.111	-3.215	1.30E-03	3.30E-03
SPH_0456+	-0.370	0.121	-3.065	2.20E-03	5.40E-03
nanA_1	0.413	0.138	2.988	2.80E-03	6.80E-03
SPH_0455+	-1.348	0.466	-2.895	3.80E-03	8.90E-03
rrgB	-0.610	0.213	-2.866	4.20E-03	9.50E-03
cps4C	-0.692	0.254	-2.728	6.40E-03	1.40E-02
cps2T	-0.596	0.240	-2.485	1.30E-02	2.80E-02
nanB	0.885	0.368	2.404	1.60E-02	3.40E-02
SPH_0458+	-0.259	0.115	-2.251	2.40E-02	5.00E-02

Table 10: Those virulence factors from VFDB that were found to occur significantly less often (estimate>0) or more often (estimate<0) in multiple carriage.

atically underestimating the prevalence of virulent lineages in the community. The reduced sensitivity could also help to explain why serotype 1 is very rarely found in carriage isolates [62]. Consistent with previous studies we observed a higher prevalence of pneumococcal carriage in children with a corresponding increase in the rate of multiple carriage [169]. Interestingly, common lineages were more likely to be found in children with rarer lineages disproportionately found in adults. In particular, serotype 3 was found to have a higher prevalence in mothers consistent with the increased risk of serotype 3 related disease found in adults [240, 439]. In contrast to the association observed between virulent lineages and multiple carriage, after accounting for lineage level effects we found the many known virulence factors were less likely to appear in multiple carriage events. These included the well known virulence genes *CbpA*, *pspA*, *ZmpB*, *lytA* and *lytB*. While these genes have been linked to the virulence of pneumococcal infections many have also been shown to play a role in the competition between pneumococci within the host [147]. Hence, lineages with these genes may be better able to outcompete other lineages leading to lower rates of multiple carriage. Genes in the PI-2 regions were found to be more common in multiple carriage. As they have been associated with virulent serotypes 1 and 2 this could suggest that similar mechanisms are acting to increase their prevalence in multiple carriage events.

The heterogeneity observed in the association of different lineages and serotypes within multiple carriage provides new insights into the competition dynamics between pneumococcal strains. Although some models of pneumococcal strain structure incorporate differences in fitness in within-host competition [111], many assume a uniform probability that any two strains interact or do not incorporate within-host competition at all [80, 210, 294]. Similarly, recent advances in the modeling of gene frequencies in pneumococcal populations do not currently consider heterogeneity in the ability of different strains to interact and thus exchange genetic material. While current models of pneumococcal strain structure often account for the coexistence of antimicrobial resistant and susceptible strains, we do not currently know of any that also incorporate differences in the carriage rates and prevalence of virulent strains. As our results indicate that virulent lineages behave differently to resistant lineages this represents a promising avenue of further research to better characterise pneumococcal population dynamics. While we do not find evidence for a fitness advantage to resident pneumococci over challengers as described in Jacobs and Weiser [244], a plausible model similar in nature to that posed in Jacobs and Weiser, could be that more virulent pneumococci survive by transmitting quickly, facilitated by the elicitation of a strong immune response. As such they would not require the genes necessary for long term carriage including resistance and the

ability to fend off competing pneumococci. While this model fits with the observation of the correlation between carriage and resistance in Lehtinen *et al.*, [294], additional modeling would be needed to determine whether it provides a better fit to the available data.

ANTIBIOTIC RESISTANCE AND TREATMENT

7.1 BACKGROUND

The high transformability of the pneumococcus has allowed the species to rapidly adapt to the introduction of antibiotics with the emergence of multidrug resistance in the 1970's leading to revived interest in pneumococcal vaccines as an alternative method for controlling the pathogen. Although the genetic determinants of antibiotic resistance in *S. pneumoniae* have been well studied, the impact of antibiotic resistance and treatment on within-host genetic diversity has received relatively little attention outside modeling studies. Deep sequencing of whole plate sweeps offers not only the possibility of increasing our ability to detect resistance in cases of multiple carriage but allows for the characterisation of the impacts of antibiotic resistance and treatment within the host.

7.1.1 *Mechanisms of resistance*

7.1.1.1 *Beta-lactams*

Resistance to beta-lactams in *S. pneumoniae* is driven by combinations of altered penicillin binding proteins (PBPs). These proteins function to remodel the peptidoglycan component of the cell wall which serves essential roles in cell expansion, division, surface anchoring, diffusion and in the maintenance of cell integrity. The majority of such mutations are found in three of the six PBPs found in *S. pneumoniae*: PBP2b, PBP2x and PBP1a [20] (Table 11). The remaining three PBP genes include PBP1b, PBP2a and PBP3 [205]. Whilst mutations in the PBP genes are required for high level resistance to beta-lactams, alterations to the MurM protein, involved with synthesising cross-links between peptidoglycan chains also appears to be associated with high level resistance to beta-lactams [456].

7.1.1.2 *Macrolides and Lincosamides*

Ribosomal modification by 23S rRNA methylation caused by modifications to the encoding erm(B) gene is the primary form of macrolide resistance in *S. pneumoniae* [504]. erm(B) is often found in the Tn916 family of transposons which frequently carries the tet(M) gene associated with tetracycline resistance leading to such strains being multidrug resistant [494]. Rare modifications to erm(A) which encodes a methylase has also been shown to confer resistance to macrolides

and lincosamides [504]. Active drug efflux has been shown to confer lower level resistance to erythromycin but not lincosamides. These are mediated by *med(A)*, *mef(E)* and in rare cases *mef(I)* genes which have been found on a number of different but related mobile genetic elements [78, 262].

7.1.1.3 Fluoroquinolones

Resistance to fluoroquinolones can be driven by modifications in the quinolone resistance determining region (QRDR) primarily in genes encoding the DNA gyrase (*gyrA*) and DNA topoisomerase IV (*parC*) subunits. An overexpression of the ABC transporter genes *patA* and *patB* have also been implicated in efflux mechanisms of fluoroquinolone resistance but are not well understood [23, 262].

7.1.1.4 Other antibiotics

Tetracycline resistance is frequently observed in *S. pneumoniae* and occurs via protection of the 30S ribosome subunit against antibiotic binding via the acquisition of *tet* genes. In pneumococci the *tet(M)* gene is most common with the *tet(O)* gene occurring in rare instances [50, 507]. Co-Trimoxazole resistance is often found in combination with resistance to penicillin and is the result of mutation in the *folA* and *folP* genes encoding dihydrofolate reductase (DHFR) protein and the dihydropteroate synthase (DHPS) proteins respectively [2, 335].

7.1.2 Maintenance of resistance in *S. pneumoniae*

Whilst the mechanisms of resistance to many antibiotics are well known in *S. pneumoniae*, the evolution and maintenance of resistance at the population level remains poorly understood. Drug resistance in the malaria parasite *Plasmodium falciparum* has steadily accumulated and swept to fixation in many populations [11]. Similarly, penicillin resistance in *Staphylococcus aureus* became nearly universal following the introduction of the drug [317]. In stark contrast, despite the increase in the use of antibiotics, resistance in *S. pneumoniae* has remained at intermediate frequencies [154]. Similar patterns have been observed in the resistance to aminopenicillins and fluoroquinolones in *E. coli* and vancomycin resistance in *Enterococcus* [111, 150, 342].

The proportion of resistant *S. pneumoniae* strains within a population correlates strongly with antibiotic use. This has been observed both at the country level in Europe and North America as well as within countries with periods of high antibiotic usage correlating with high levels of resistance in southern Israel [42, 107, 220]. The decline of resistance in the absence of selection pressures from antibiotic treatment suggests that resistance comes at a fitness cost. This fitness cost has also been observed experimentally in rats [486].

Abbreviation	Antibiotic	Genes
PEN	Penicillin	pbp1a, pbp2b, pbp2x
AMO	Amoxicillin	pbp1a, pbp2b, pbp2x
MER	Meropenem	pbp1a, pbp2b, pbp2x
TAX	Cefotaxime	pbp1a, pbp2b, pbp2x
CFT	Ceftriaxone	pbp1a, pbp2b, pbp2x
CFX	Cefuroxime	pbp1a, pbp2b, pbp2x
ERY	Erythromycin	mefA and ermB
CLI	Clindamycin	ermB
COT	Cotrimoxazole	folA and folP
TET	Tetracycline	tet(M), tet(O), tet(S/M)
CHL	Chloramphenicol	cat
RIF	Rifampin	rpoB
DOX	Doxycycline	tet(M)

Table 11: Antibiotic abbreviations and the genes linked to resistance in *S. pneumoniae*.

As described in [Chapter 5](#), a number of models have been suggested to explain the co-existence of resistant and susceptible strains over long time periods. These include those that consider within-host competition, negative frequency dependent selection, carriage duration and competitive exclusion [[77](#), [80](#), [86](#), [111](#), [210](#), [244](#), [295](#)].

7.1.3 Surveillance of resistance

Historically antibiotic resistance surveillance in *S. pneumoniae* as with most bacteria have focused on culture based phenotypic susceptibility methods. This remains the gold standard for clinical laboratories and forms the basis of many countries' public health surveillance systems. Following the identification of resistance elements within the *S. pneumoniae* genome, PCR based methods were introduced [[191](#), [207](#)]. This improved the speed with which results could be obtained. With the identification of multi-drug resistant strains the Pneumococcal Molecular Epidemiology Network (PMEN) was set up to help track drug resistant lineages [[266](#)]. Most recently, approaches based on whole genome sequencing have become increasingly common. Large WGS surveillance studies such as the Global Pneumococcal Sequencing (GPS) project coupled with detailed laboratory based phenotypic studies have allowed for detailed investigations and tracking of resistance elements in *S. pneumoniae* globally [[190](#), [311](#)].

7.1.4 Impacts of antimicrobial treatment

Much of the *S. pneumoniae* genomics literature has focused on investigating genomic signatures of antibiotic resistance. This has both increased our understanding of the mechanisms of resistance as well as aiding in the surveillance of resistant *S. pneumoniae* lineages [72, 190, 262]. Despite this, less attention has been paid to the impact antibiotic treatment itself has on the within-host genetic diversity of *S. pneumoniae*. This is in part due to the difficulty in obtaining relevant samples as surveillance studies typically consider a single sample per host and the treatment history of the host is usually unknown. A number of studies have looked at the impact of antimicrobial treatment on the prevalence of resistance in treated individuals. One multi-centre study in French children that considered nasopharyngeal swabs before and after antimicrobial treatment found that treatment led to a decrease in pneumococcal carriage rates but did not increase the risk a child would become colonised with a resistant lineage. However, the study did observe an increase in the frequency of resistant lineages post treatment driven by the elimination of susceptible lineages [495].

As antibiotics are often prescribed when an individual is unwell, by investigating the genomic features of *S. pneumoniae* associated with treatment we can potentially identify novel virulence and resistance elements. The consistent monthly sampling in the Maela dataset coupled with the accurate recording of the prescription of antimicrobials makes it an ideal dataset for investigating the impacts of treatment on the within-host genetic diversity of *S. pneumoniae*.

In this chapter, we describe methods for the identifying resistance elements from sequenced whole plate sweeps. We verify these approaches using both the artificial laboratory mixtures of Knight *et al.*, [267], as well as the previous single colony pick data from Maela described in Chewapreecha *et al.*, [71]. Using the resulting resistance calls, we consider the increased sensitivity to detected resistance provided by whole plate sweeps and the association between resistance and multiple carriage. Finally, we examine the impact of antimicrobial treatment on the within-host genetic diversity of *S. pneumoniae*.

7.2 METHODS

7.2.1 Antimicrobial resistance typing

A number of bioinformatic pipelines have been developed to identify resistance elements from WGS data. Initial efforts such as ResFinder and ARG-ANNOT focused on calling resistance elements directly from *de novo* assemblies of WGS data [202, 534]. These usually make use of local alignment tools such as BLAST which search a database of known resistance sequences against each assembly [6]. Al-

ternative approaches including SRST2 and KmerResistance align sequencing reads directly to the database of known elements using predefined coverage and identity thresholds to call resistance elements [75, 241]. As described previously in [Section 6.2.3.1](#), Ariba takes a hybrid approach, using local assemblies of reads at the gene level to more accurately identify possible resistance elements and mutations [236]. To determine resistance elements with Ariba we made use of the comprehensive antibiotic resistance database (CARD) [337]. We supplemented this database with a custom set of known *S. pneumoniae* resistance elements which have previously been used with Ariba to call resistance in *S. pneumoniae* [311]. Species specific databases and methods have also been developed including the CDC's resistance typing pipeline developed for *S. pneumoniae* [355]. The CDC resistance calling pipeline for *S. pneumoniae* relies on a database of PBP genes with known resistance profiles [305]. A random forest is trained on these profiles using the mutations in the corresponding PBP genes as features for the model. Given a novel set of PBP genes the pipeline is able to predict the likely resistance profile of a strain of *S. pneumoniae*. However, the pipeline was designed for the analysis of single genomes and thus relies on lineage deconvolution to be applicable to whole plate sweep sequencing data. Similar to our analysis of serotypes in the previous chapter we make use of both the deconvolution independent Ariba method as well as running the CDC pipeline on the strains reconstructed using mGEMs.

7.3 THE IMPACTS OF ANTIMICROBIAL TREATMENT

7.3.1 Bacterial GWAS

Genome wide association studies (GWAS) have been used extensively in human genetics. The technique identifies genetic variants that are associated with a phenotype of interest, for example heart disease, predisposition to cancer, eye colour or even one's preference for coriander [91, 153, 489, 493]. Unlike in humans, bacterial populations are often highly related as they reproduce clonally. This can result in spurious associations or a considerable loss in statistical power if common human genetics pipelines are used to analyse bacterial phenotypes. Many bacterial species also consist of large pangenomes and thus the common assumption in human genetics that most genes are shared between individuals is often invalid.

To resolve these issues a number of bacteria specific GWAS methods have been proposed. The bugwas pipeline distinguishes lineage level associations from locus-specific associations [144]. When bacterial population structure prevents fine mapping of associations between phenotypes and specific genetic features such as single nucleotide polymorphisms (SNPs), the method will instead report as-

sociations at the lineage level. In *S. pneumoniae* this could refer to GP-SCs as an example. An alternative solution uses a pre-generated phylogeny of the bacterial genomes to create a simulated null distribution for the features of interest which has been shown to increase specificity and sensitivity over common dimension-reduction techniques used in human GWAS [82].

Both the bugwas pipeline, and pyseer, a similar bacterial GWAS method, make use of Linear Mixed Models (LMMs) to control for population structure which have been found to be more efficient in human genetics studies and thus are preferable to methods that rely solely on fixed effects [144, 289]. In human genetics, it is common to use SNPs generated by alignment to a reference genome as the genetic features of interest. This is problematic in bacterial genetics where many genes are rare and only common to a subset of isolates. To address this, a number of pipelines have been developed that instead rely on alignment free approaches making use of alternative features including k-mers and unitigs [245, 293].

7.3.1.1 *K-mers and Unitigs*

K-mers are short sub-sequences of k nucleotides. Unlike SNPs, the presence or absence of k-mers can be generated directly from raw sequence data without the need for alignment to a reference. Given a sequence of length L there are up to $L - k + 1$ unique k-mers out of a possible n^k where n is usually equal to 4 (the number of nucleotides). While the use of k-mers allows for better characterisation of the pan-genome, the large number of k-mers that are identified in bacterial genomes can lead to a very large number of association tests when compared to more common SNP based analyses. This can be remedied by using burden tests as well as by considering the number of unique kmer presence/absence patterns rather than the total number of unique k-mers [289]. An alternative solution is to use a reduced set of variable length k-mers. Unitigs which are variable length k-mers, can be found by generating a compressed De Bruijn graph from either genome assemblies or raw sequencing data. De Bruijn graphs connect overlapping k-mers. Paths through this graph that do not result in forks can then be collapsed into unitigs which can be used as an alternative to k-mers [245].

To generate frequency counts of each unitig present in our samples we used the Bifrost algorithm to define the set of unitigs [225]. We then found the count of each unitig in each sample using a custom python script that iterated through the original sequencing reads. To avoid testing very rare features, we only considered those unitigs present in at least 1% of the samples of interest in our presence/absence based analysis and in at least 2% of our paired analysis discussed below.

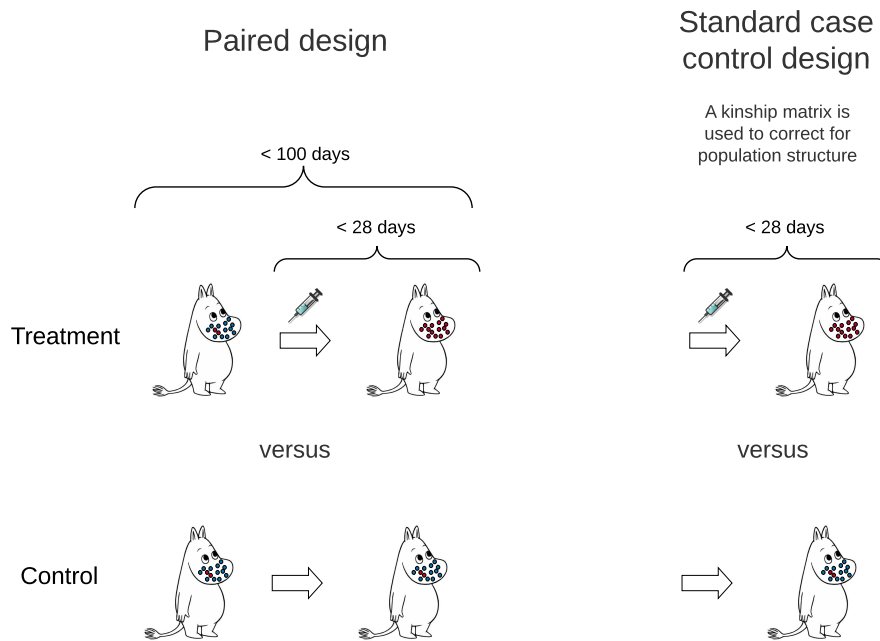


Figure 42: A diagram describing the two unitig association studies performed. In the paired designed, population structure is accounted for by requiring that the same lineage is sampled prior to treatment. In this analysis the change in frequency between the first and second sample in a pair is treated as the dependent variable. In the standard design, the presence and absence of unitigs taken from samples within 28 days of antibiotic treatment are compared to those where no treatment has occurred. Here, population structure is accounted by including a kinship matrix in the LMM.

7.3.2 Treatment association study design

To investigate the impact of antibiotic treatment on *S. pneumoniae* carriage we performed two main analyses (Figure 42). The first consisted of a typical case control design and compared samples that were within a recent antimicrobial treatment event to those where no treatment had occurred. This allowed us to investigate features associated with recent antibiotic treatment but does not consider the changes that occur within an individual that is already colonised with *S. pneumoniae* prior to treatment. In order to shed light on this scenario our second analysis investigated the impact of treatment on pairs of consecutive samples from the same patient where some patients had taken antibiotics inbetween samples (Figure 42).

7.3.2.1 Treated vs untreated

Samples were classified as treated if they were within 28 days of an antimicrobial treatment event. This is the same definition as was used in the analysis of virulence and serotypes and was chosen after re-

viewing the decline in the proportion of resistant isolates tested via disk diffusion and Etest testing of all swabs positive for *S. pneumoniae* (Figure 31). The python implementation of the seer algorithm was then used to identify unitigs significantly associated with treatment. Here, rather than using counts unitigs were called as either present or absent. To control for population structure pyseer was run using a Linear Mixed Model with a kinship matrix generated by taking the cross product of the binary unitig presence/absence matrix.

7.3.2.2 Paired analysis

Our unique sampling allows us to compare samples from the same individual before and after treatment. We first identified sample pairs where there were at most 100 days separating pneumococcal positive nasopharyngeal swabs from the same individual. We restricted our analysis to infants as treatment information for mothers was not available. To ensure previous treatments prior to the first sample of an individual were not confounding our results we excluded pairs with any treatment event within 28 days of the first swab. This resulted in 615 sets of paired samples. We classified these pairs into treated and untreated groups based on whether or not the individual had received antibiotic treatment in the time between swabs.

We only considered paired samples where the infants were positive for *S. pneumoniae* in both samples. As a result we are not considering the impact of antibiotic treatment on overall carriage rates but rather the differences in *S. pneumoniae* genomes pre and post antibiotic treatment. Using this paired design we considered the impact of treatment both at the lineage (GPSC) level as well as the locus level. This distinction between lineage and locus is similar to that defined in [144]. Unlike many previous bacterial GWAS studies which typically focus on the presence or absence of a feature we considered the frequency of both lineages and loci within each sample. This improves our ability to identify more subtle changes that can be obscured by ignoring within-host diversity.

7.3.2.3 Lineage model

At the lineage level we considered the estimated frequencies of each lineage obtained using the mSWEEP algorithm. We used a simple linear model to test whether treatment impacted the frequency of the second sample of a pair after controlling for the observed frequency in the first sample as well as the difference in time between the two samples.

$$f_i^B = \beta_0 + \beta_1 f_i^A + \beta_2 (t_i^B - t_i^A) + \beta_3 \tau_i \quad (37)$$

Where f_i^A and f_i^B are the frequencies of the lineage in sample i in the first and second swab of the pair respectively; t_i is the date of the samples, and τ_i is a binary variable indicating whether or not pair i has a treatment event between the samples.

7.3.2.4 *Locus model*

To investigate locus level effects we considered the frequency of each unitig in each sample. In order to control for lineage level effects we concentrated on pairs where the same lineage was present in both samples. This reduced the analysis to 445 pairs.

Unlike the lineage level analysis where we used estimated frequencies, unitigs are represented by the number of times they were observed in the raw reads from each sample. This is a similar problem to that found in the analysis of RNAseq datasets where the number of RNA reads aligned to a gene is used as a proxy for the expression of that gene. Here, methods have been developed that make use of negative binomial regression [338, 422]. Unlike the Poisson distribution, which assumes that variation is equal to the mean, a negative binomial distribution allows for overdispersion of the count data. In the case of the RNAseq specific methods, information is shared between gene counts to estimate the dispersion parameters. This increases the statistical power of these tests. Whilst such methods are applicable to our analysis they were originally designed for the analysis of tens of samples and thousands of genes and thus do not scale to our dataset of over 10 million unitigs and thousands of samples.

This inability to scale was also observed when analysing k-mer count data [414]. Here, the authors took a similar approach to that used in seer but used Poisson regression in place of the linear or logistic regression options available in seer [293]. Whilst Poisson regression resolves some of the scaling issues we found it to be unsuitable for the analysis of our unitig data as the strong assumption of the mean variance relationship led to an unrealistic number of significant unitigs.

Instead we make use of a quasi-poisson generalised linear model. Like a negative binomial model the quasi-poisson GLM allows for overdispersed count data and has the same number of parameters. The quasi-poisson model does not have a distributional form and assumes that the variance is a linear function of the mean as opposed to the negative binomial model where this relationship is quadratic. The negative binomial model is generally preferred in genomics due to its distributional form allowing for model comparison methods such as the Akaike information criteria (AIC) to be used. However, we found that fitting a negative binomial to all unitigs could not be achieved within a reasonable time frame and thus the quasi-poisson approach strikes a balance between computational efficiency and accuracy.

Using the quasi-poisson GLM we tested the counts obtained for each unitig for an association with treatment. We included an offset term to control for the total number of unitigs present in each sample as is done in RNAseq analysis. We also included covariates for each pair as well as the time between samples in a pair. The final model including the quasi-poisson link function is can be written as

$$\log(E(U_{ij})) = \beta_0 + \beta_1 p_i + \beta_2 t_i + \beta_3 \tau_i + \log\left(\sum_{\forall j} U_{ij}\right) \quad (38)$$

Here, U_j is the count of unitig j in sample i , and p_i is a factor indicating which pair sample i belongs to. Unlike the lineage analysis, t_i refers to the time since the first sample. Thus this is set to zero for the first sample of each pair. Similar to the lineage analysis τ_i indicates whether this pair involved an antimicrobial treatment event.

7.4 RESULTS

7.4.1 Verification of the mGEMS/CDC pipeline

To test whether combining the mGEMS lineage deconvolution algorithm with the CDC resistance calling pipeline could successfully identify resistance in mixed samples we considered the artificial laboratory mixtures from Knight *et al.*, [267]. As this is a small dataset we also compared the calls identified in 1158 samples that were processed both in the previous study of single colony picks in Maela [71] as well as in the current study.

Figure 43a indicates that the combined mGEMS + CDC pipeline approach achieves a precision and recall of 0.750 and 0.825 respectively in identifying resistant lineages. The majority of these errors are found in the artificial mixture containing 10 lineages. This would correspond to an extremely high multiplicity of infection which we do not observe in our dataset. Figure 43b indicates the precision and recall of the mGEMS + CDC pipeline method excluding this sample. In this case we are able to achieve a recall of 100% and a precision of 83%. If we only consider resistance calls at the sample level rather than at the lineage level the number of false positives drops to zero indicating that the remaining errors are the result of the failure of the mGEMS algorithm to separate some resistance elements into the correct lineages (Figure 43c). For this reason we focus the majority of our analysis on sample level resistance calls.

To ensure the mGEMS + CDC resistance pipeline remained effective across a larger and more representative dataset we compared the results of the pipeline on 1158 samples for which corresponding assemblies existed from single colony picks taken from the same swabs [71]. As we are comparing single colony picks with the deep sequenc-

ing data it was not possible to determine the number of false positives. However, this approach allows us to obtain accurate estimates of the recall of the combined pipeline. [Figure 43d](#) indicates that the mGEMs + CDC resistance pipeline is able to achieve a recall rate of 96.9%. This indicates that the combined pipeline is able to accurately identify resistance from deep sequenced plate sweeps.

7.4.2 *Whole plate sweeps identify 14.6% more antibiotic resistance calls*

Genomic surveillance in *S. pneumoniae* is predominantly achieved through the sequencing of single colonies [190, 311]. To investigate the increase in sensitivity provided by the deep sequencing of whole culture plates we compared the number of resistance calls across 1158 samples for which we had both single colony picks and whole plate sweeps. Over the nine resistance classes considered the whole plate deep sequencing based strategy identified an additional 520 resistance calls corresponding to an a 14.6% increase over the single colony pick method ([Figure 44](#)). This increase varied by antibiotic with the number of chloramphenicol resistance calls almost doubling from 44 to 82. These results suggest that sequencing whole plate sweeps not only identifies additional lineages and serotypes but offers profound improvements to current antimicrobial resistance surveillance strategies. Furthermore, treatment decisions based on resistance calls obtained from sequencing a single genome could systematically underestimate the presence of resistance leading to the suboptimal prescription of antibiotics.

7.4.3 *Resistance elements are less likely to be observed in mixed infections*

In [Chapter 6](#), we found that common multidrug resistant pneumococcal lineages were less likely to be observed in multiple carriage. This has important implications for models of the maintenance of resistance in the broader pneumococcal population [77, 80, 111, 295]. However, as we only considered resistance at the lineage level it is possible that the association observed was driven by other genetic features of the lineages unrelated to resistance. To investigate the association between genetic markers of resistance and multiple carriage independent of the background population structure, we performed a similar analysis as in [Section 6.2.4](#). To control for population structure we constructed a pairwise distance matrix between every sample using the Mash algorithm [388]. We performed multidimensional scaling (MDS) on this matrix and retained 10 dimensions to use as covariates to control for population structure as is standard in bacterial GWAS studies [293]. To investigate whether resistance elements appear at different proportions we compared the inferred mSWEEP strain frequencies of strains that were found to have resistant ele-

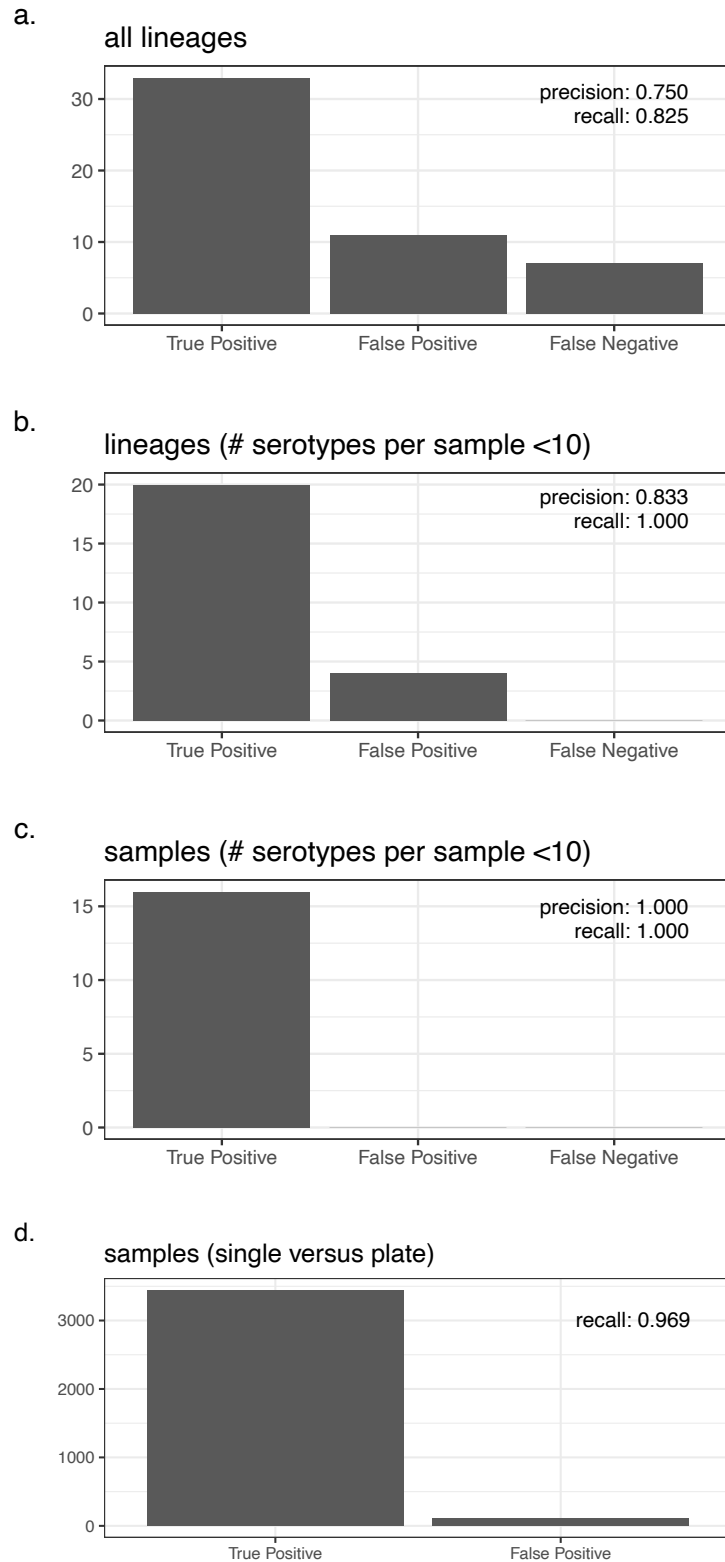


Figure 43: a.) the number of true positive, false positive and false negative resistance calls obtained using the mGEMs + CDC pipeline applied to the artificial laboratory mixtures of Knight *et al.*, [267]. b.) the same as a. but excluding samples containing 10 or more lineages. c.) the same as a. and b. but reporting the resistance calls by sample rather than lineage. d.) The true positive and false positive resistance calls obtained by comparing the 1158 whole genomes from Chewapreecha *et al.*, [71] with the calls obtained using the mGEMs + CDC pipeline.

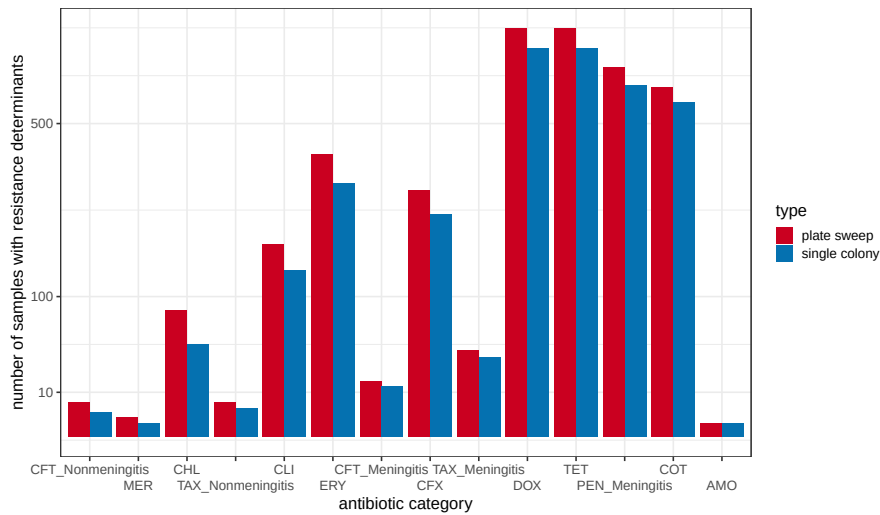


Figure 44: The number of samples identified to be resistant split by antibiotic for the 1158 samples where both single colony picks and whole plate sweeps have been processed.

ments according to the CDC pipeline using a linear mixed model with fixed effects for treatment and population structure as well as a random effect for the host a sample originated from. Figure 45 indicates the distribution of the mSWEEP frequencies separated by antibiotic. Overall, resistant elements were found at higher frequencies despite controlling for the background population structure with 8/14 CDC antibiotic categories found at higher frequencies (Table 12). This suggests that the negative association with multiple carriage is not driven by the lineage but by the antibiotic resistance features themselves. As most laboratory studies have demonstrated a fitness cost to resistance it is likely that these lineages are being outcompeted by susceptible strains within the host [428, 487].

To ensure that this result was not caused by biases in the estimated frequencies induced by the culturing step of the experiment or in the deconvolution of mixed samples, we ran a logistic regression with the same modified link function as was used in the lineage level analysis in Section 6.2.4. The use of logistic rather than linear regression reduced the statistical power of the test. However, consistent with the frequency based analysis we found that resistant features were less likely to be observed in multiple carriage with 5/14 resistance categories found to be significantly more likely to appear in samples comprising a single lineage (Table 13). As a final comparison, to conduct an analysis entirely independent of the mSWEEP/mGEMs pipelines we performed the same logistic regression based test using the resistance calls obtained by running Ariba with the extended CARD database. Although the resistance categories are not identical between the Ariba and CDC pipelines, we identified that all 4 of the resistance categories identified by Ariba were associated with a decreased pres-

antibiotic	estimate	Std. Error	t value	p-value	adj p-value
PEN_Meningitis	20.07	1.94	10.36	3.58E-24	5.01E-23
CFX	18.97	2.52	7.52	1.04E-13	7.30E-13
ERY	9.93	2.05	4.85	1.33E-06	6.21E-06
TAX_Meningitis	19.25	4.52	4.25	2.19E-05	7.68E-05
CFT_Meningitis	20.04	5.13	3.9	9.64E-05	2.70E-04
TET	4.82	1.65	2.91	3.65E-03	7.08E-03
DOX	4.82	1.65	2.91	3.65E-03	7.08E-03
COT	4.36	1.52	2.88	4.04E-03	7.08E-03
CLI	1.71	2.14	0.8	4.25E-01	6.62E-01
AMO	8.72	13.35	0.65	5.14E-01	7.12E-01
MER	-6.03	10.34	-0.58	5.60E-01	7.12E-01
TAX_Nonmeningitis	-2.48	9.29	-0.27	7.90E-01	9.22E-01
CHL	0.36	2.4	0.15	8.82E-01	9.29E-01
CFT_Nonmeningitis	0.87	9.69	0.09	9.29E-01	9.29E-01

Table 12: The results testing the association between resistance and the inferred mSWEEP frequencies using a linear mixed model to account for treatment, host and lineage (GPSC) effects.

ence in multiple carriage (Table 14). These results suggest that it is the genetic features of resistance themselves that are associated with a reduced prevalence in multiple carriage rather than the association being driven by other lineage specific genetic features.

7.4.4 Resistance levels differ between mothers and children

A number of models of resistance in *S. pneumoniae* suggest that structure within the population could explain the maintenance of intermediate frequencies of resistance [77, 80]. A common motivation for such structure is that children may experience different transmission dynamics and rates of antibiotic treatment than the adult population. This is often referred to as the ‘day care’ model as it is plausible that transmission dynamics within childcare and schools could be significantly different to that seen more broadly. Similar arguments can also be made for population structure induced by care homes in elderly individuals [541]. To investigate whether there was evidence for such dynamics within our dataset we compared the incidence of resistance within samples taken from children to that found in mothers. Figure 46 indicates the number of mothers and children that were found to have resistant lineages to each of the antibiotics considered. Logistic regression was used to identify those antibiotics that had significantly different profiles in mothers than children after controlling for the number of lineages found in each individual (Table 15). Penicillin, cotrimoxazole, erythromycin and cefuroxime were all found

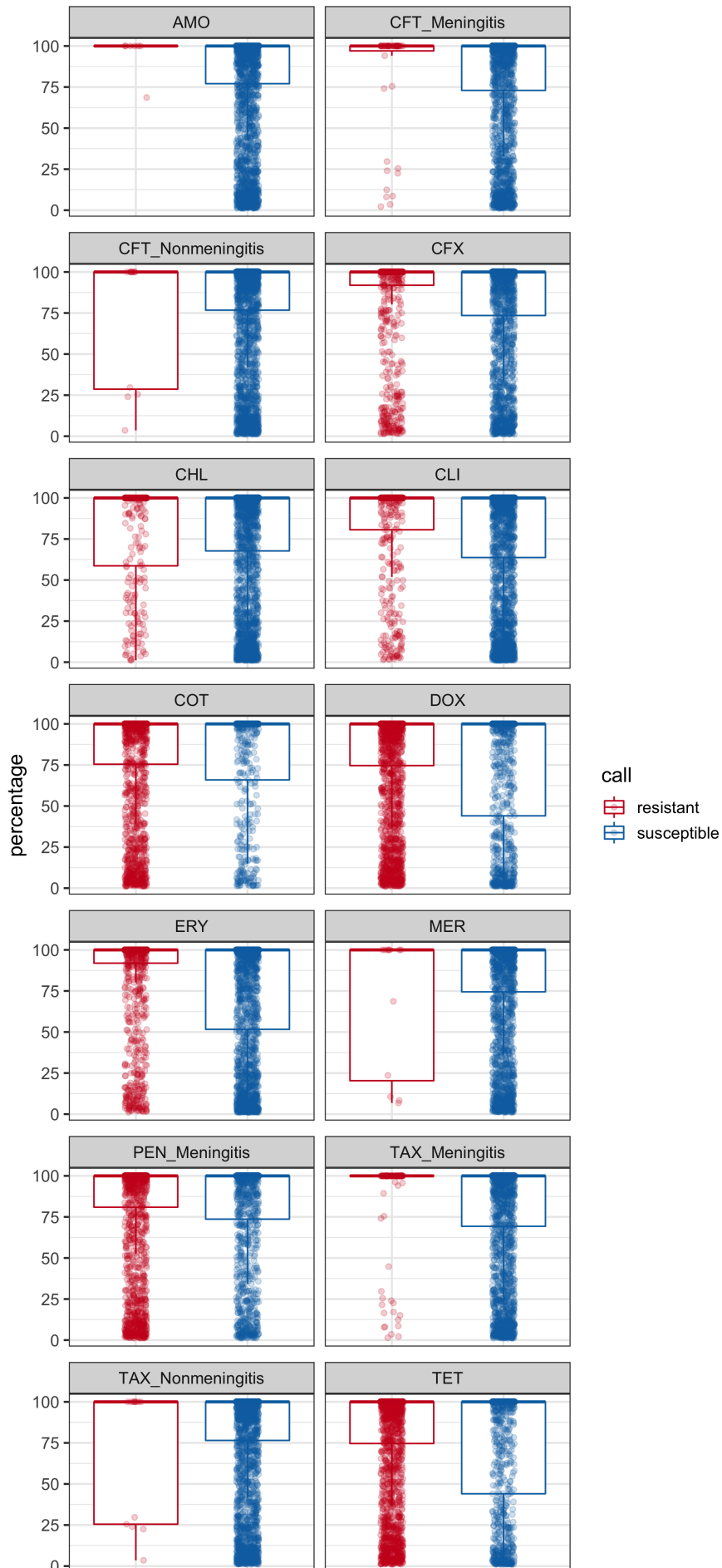


Figure 45: Boxplots indicating the corresponding lineage frequencies inferred using mSWEEP for those lineages identified as being resistant and susceptible by the mGEMs + CDC pipeline.

antibiotic	estimate	std.error	statistic	p-value	adj p-value
PEN_Meningitis	0.5	0.1	5.09	3.59E-07	4.31E-06
COT	0.4	0.1	3.91	9.37E-05	5.62E-04
CLI	0.5	0.15	3.37	7.47E-04	2.99E-03
TET	0.3	0.12	2.72	6.53E-03	1.57E-02
DOX	0.3	0.12	2.72	6.53E-03	1.57E-02
TAX_Meningitis	0.6	0.3	1.93	5.42E-02	1.08E-01
CFX	-0.2	0.15	-1.7	8.95E-02	1.46E-01
ERY	0.2	0.13	1.66	9.72E-02	1.46E-01
CHL	0.2	0.16	0.96	3.38E-01	4.51E-01
CFT_Meningitis	-0.4	0.57	-0.64	5.20E-01	6.25E-01
TAX_Nonmeningitis	-0.5	1.14	-0.43	6.65E-01	7.26E-01
CFT_Nonmeningitis	-0.2	1.25	-0.19	8.47E-01	8.47E-01

Table 13: The results of association tests comparing co-occurrence of resistance and multiple carriage using a logistic generalised linear mixed model with a modified link function to account for the number of lineages present in each sample. Terms are included to account for treatment, host and lineage effects.

antibiotic	estimate	std.error	statistic	p-value	adj p-value
COT	3.6	0.53	6.78	1.24E-11	4.96E-11
RIF	1	0.23	4.27	1.99E-05	3.97E-05
TET	0.6	0.16	3.67	2.47E-04	3.29E-04
ERY	0.4	0.13	2.82	4.74E-03	4.74E-03

Table 14: The results of association tests comparing co-occurrence of resistance calls obtained using the Ariba pipeline and multiple carriage. Test were performed using a logistic generalised linear mixed model with a modified link function to account for the number of lineages present in each sample. Terms are included to account for treatment, host and lineage effects.

antibiotic	estimate	std.error	statistic	p-value	adj p-value
PEN_Meningitis	-1.25	0.27	-4.58	4.74E-06	3.35E-05
COT	-1.18	0.26	-4.57	4.79E-06	3.35E-05
CFX	-1.13	0.27	-4.21	2.51E-05	1.17E-04
ERY	-1.01	0.25	-4.04	5.24E-05	1.83E-04
CLI	-0.54	0.28	-1.90	5.71E-02	1.60E-01
CFT_Meningitis	-1.59	1.04	-1.53	1.27E-01	2.97E-01
TAX_Meningitis	-0.86	0.62	-1.37	1.69E-01	3.39E-01
CHL	-0.29	0.31	-0.94	3.49E-01	6.11E-01
DOX	-0.09	0.34	-0.28	7.81E-01	9.95E-01
TET	-0.09	0.34	-0.28	7.81E-01	9.95E-01
TAX_Nonmeningitis	-0.18	1.12	-0.16	8.71E-01	9.95E-01
CFT_Nonmeningitis	0.12	1.14	0.10	9.17E-01	9.95E-01
MER	-15.95	1712.32	-0.01	9.93E-01	9.95E-01
AMO	-16.68	2824.41	-0.01	9.95E-01	9.95E-01

Table 15: The results of association tests comparing the number of resistant samples in mothers versus children using logistic regression to account for the number of lineages found in each individual

more commonly in children than mothers. This is likely a reflection of the increased levels of antibiotic treatment in young children in the Maela refugee camp. Whilst there was a significant difference in the levels of resistance between mothers and children, resistance was still very much prevalent in both categories. In addition, we did not find a significant association between the age of a child and resistance indicating there was no additional age driven sub-structure within the infant population. Models that explain the maintenance of intermediate levels of resistance using population structure have been shown to be very sensitive to the level of mixing between the sub-populations. Consequently, the relatively small but significant difference in resistance rates between mothers and children is unlikely to drive the co-existence of resistant and susceptible strains at the population level.

7.4.5 Treatment

7.4.5.1 Lineage associations

In order to examine the changes in both serotypes and GPSC frequencies after treatment we compared pairs of samples taken from the same individual taken within 100 days. Frequencies were estimated using the mSWEEP algorithm and sample pairs were classified by whether or not a treatment event had occurred in the time between samples. After correcting for multiple testing, GPSC 1 was the only lineage to vary significantly in frequency after antibiotic treatment

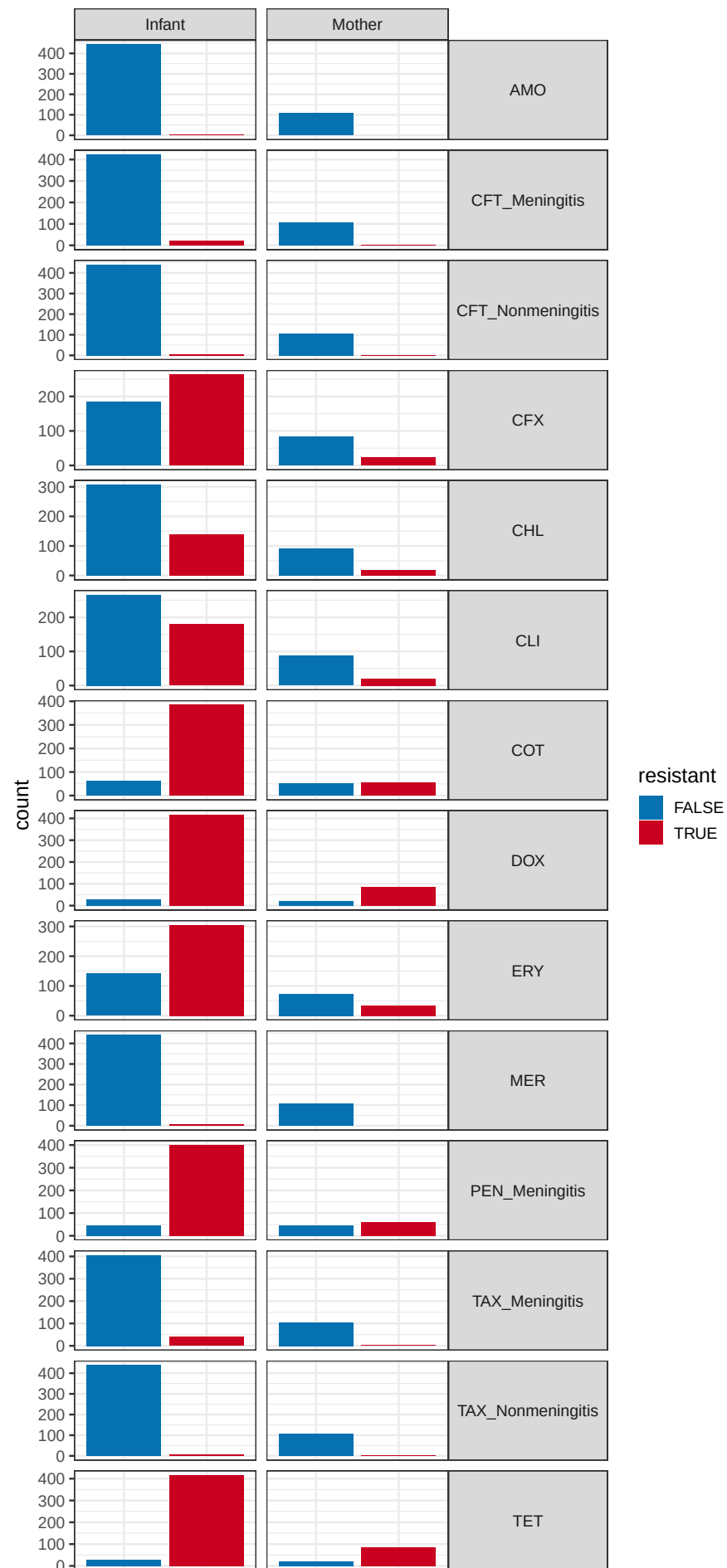


Figure 46: The number of samples identified as potentially resistant by the mGEMs + CDC pipeline in mothers and children, split by antibiotic.

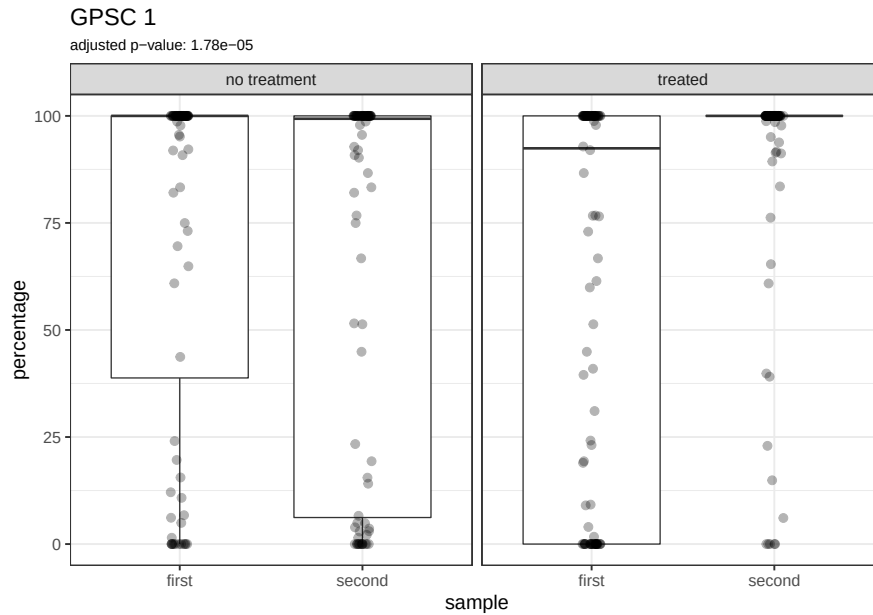


Figure 47: The inferred mSWEEP lineage frequencies for GPSC1 lineages taken in pairs of samples with 100 days from the same host where at least one sample was found to have a GPSC1 lineage. The left panel indicates those pairs where no antibiotic treatment occurred while the right includes pairs where a treatment event occurred between the sampling times. The increase in the frequency of GPSC1 lineages post-treatment was found to be the result of both the elimination of susceptible strains as well as the transmission of GPSC1 to an individual following treatment.

(Figure 47). GPSC 1 is the most prevalent lineage in the Maela dataset and is also common globally [190]. In addition to being highly prevalent, GPSC 1 was found to be the most common multi-drug resistant lineage in the global pneumococcal sequencing project.

In previous studies it was found that an increase in the frequency of resistant isolates post-treatment was due to the elimination of susceptible lineages [495]. However, after only considering the binary presence and absence of each lineage and controlling for presence in the first sample of a pair the association between GPSC 1 colonisation and treatment remained. This suggests that not only does treatment increase the frequency of GPSC 1 by eliminating competing lineages, it also increases the chance that a child will be colonised with the resistant lineage following treatment. This could have important ramifications for the prescription of antimicrobials as it may motivate the prescription of preventative broad spectrum antibiotics in patients that have higher risks of complications if they were to become colonized with a resistant strain.

7.4.6 Locus associations

7.4.6.1 Treated versus untreated

To identify sub-lineage genomic features of *S. pneumoniae* associated with recent antimicrobial treatment we compared the presence and absence of unitigs in samples taken within 28 days of treatment to those that had not been treated. We made use of a linear mixed model as implemented in pyseer to control for population structure [289]. This is a conservative approach as we are not considering the frequency of each unitig. This approach will also ignore resistant features that persist beyond the 28 day cut-off. Consequently, we are likely to only identify unitigs that are associated with the short time period following treatment. Using this approach we identified 8 unitigs significantly associated with antimicrobial treatment (Table 16). Of these, 6 were the result of low level contamination as their closest BLAST match against the NCBI nt database was to *Staphylococcus haemolyticus* [248]. *S. haemolyticus* is a commensal bacterium of the human skin and is thought to have the highest level of antimicrobial resistance amongst the Coagulase-negative staphylococci [179]. It is thus plausible that contamination from an infant's skin flora was more likely to contain *S. haemolyticus* following antimicrobial treatment.

The remaining two unitigs were consistent with changes in response to antimicrobial treatment in *S. pneumoniae* with one matching an efflux ABC transporter, permease protein. Active efflux has been associated with resistance to fluoroquinolones although the exact mechanisms are not well understood [23, 262]. Consistent with our analysis, the development of resistance to fluoroquinolones during treatment has been observed previously, although this was primarily a finding from observations in elderly patients with lung disease [110].

The final unitig was found in penicillin-binding protein 2 (PBP2b) and was negatively associated with treatment. This protein has been extensively studied and mutations in the penicillin binding proteins are known to confer resistance to a variety of beta-lactam antibiotics [262]. Thus this unitig may represent a susceptible form of the PBP2b which is less likely to be present when treated with beta-lactams such as ampicillin.

7.4.6.2 Paired analysis

To investigate genomic alterations within single infections in response to antibiotic treatment we focused on pairs of samples from the same infant. Unlike in the previous pairwise analysis of GPSC's we now only consider sample pairs where the same GPSC was present in both samples. This controls for lineage effects and allows us to identify finer scale genomic changes in response to antibiotic treatment.

seq	BLAST matches	beta	beta std-err	lrt p-value	notes
>tgacatatcatcttgaatc ttaaggggagataaaca atgactgaatttgattat...	<i>S. haemolyticus</i>	0.5	0.080	1.98E-10	contamination
>tgaagaaacaaatccat cagtgtattttcattctat cctaaatttgct	<i>S. haemolyticus</i>	0.4	0.073	5.55E-09	contamination
>ataaaaccgatacgctc cagtaagattttagctcc ttt	<i>S. pseudopneumoniae</i> <i>S. pneumoniae</i>	0.4	0.066	8.27E-09	efflux ABC transporter permease protein
>tttacatctgctgtgctac atttgacttgaaaatgga aagattgcagtattaag...	<i>S. haemolyticus</i> <i>S. hominis</i>	0.4	0.077	7.35E-09	contamination
>aatggtaaaacaacgac tatttcagggtgctaact atgg	<i>S. hominis</i>	0.4	0.074	2.27E-09	contamination
>atcatcagacaatctatt gtacacatactcgactgc ttcaaccattttctt	<i>S. haemolyticus</i>	0.4	0.077	8.60E-09	contamination
>aagtattctgaagggtg ctatgcagtcgccc	<i>S. pneumoniae</i>	-0.2	0.026	1.02E-08	penicillin binding protein 2
>tgtattacttttcaactt agtcgtggtgtggaaaa tcgaaataataaacga...	<i>S. haemolyticus</i>	0.5	0.084	1.20E-08	contamination

Table 16: The results of the unitig presence/absence association study comparing unitigs observed in samples within 28 days of antibiotic to those that were not treated. Notes obtained after search the NCBI BLAST datanase are given in the final column along with associated species in the second column.

Unlike in the previous unitig analysis which focused on features specific to samples after recent antibiotic treatment we are now concentrating on changes within the same carriage episode. As such rather than relying on the presence or absence of a unitig we now consider the overall frequency of a unitig within each sample. By considering frequency, we are able to identify changes to the pneumococcal population within the host that would be hidden by looking at presence/absence data alone. This allows us to identify features that would otherwise remain hidden such as the existence of heteroresistance. Heteroresistance is the existence of a resistant subpopulation within a population of susceptible cells. It has been demonstrated in a number of bacterial species that heteroresistance can be caused by gene amplification [377]. In such a scenario, the resistant gene would be present at different frequencies in both samples of a pair. Thus any associated unitig would also be present in both samples and it is only by considering the frequency of the unitig that any association with antibiotic treatment could be determined. In addition to antibiotic resistant mechanisms, if a treatment event has occurred then it is likely that an infant is unwell. Consequently, they are likely to mount an immune response to either the colonising and possible invasive pneumococci or other pathogens. As such we may expect to see genomic features associated with virulence and immune evasion as well as resistance when looking at the impact of treatment within the host.

Using the pairwise model described in [Section 7.3.2.4](#), we investigated the changes in unitig frequency in response to treatment. [Figure 48](#) indicates the resulting gene level summaries for those unitigs found to be significantly associated with treatment. Each gene is represented by a point with the y-axis indicating the p-value after adjusting for multiple testing and the x-axis indicating the average effect size for that gene. The unitigs were annotated by aligning them to the whole genome assemblies from the previous study of *S. pneumoniae* genomics in Maela using single colony picks [71]. By only considering those with strong matches to these assemblies we can exclude any unitigs that are likely to be the result of contaminating species. In [Figure 48](#), a positive effect indicates that the unitigs corresponding to that gene are more prevalent after treatment whilst a negative effect indicates the unitig(s) are less prevalent after treatment. [Table 17](#) reproduces that data displayed in [Figure 48](#) with additional notes for each annotation.

The unitigs with the largest average effect size corresponded to genes associated with both virulence and the cell wall. Four unitigs corresponding to TrpD2 were found to be negatively associated with treatment with a higher frequency of these unitigs prior to treatment with antibiotics ([Figure 49](#)). TrpD2 is involved in the synthesis of the aromatic amino acid tryptophan and has been shown to be essential for invasion of the lung [211]. The higher frequency prior to treatment

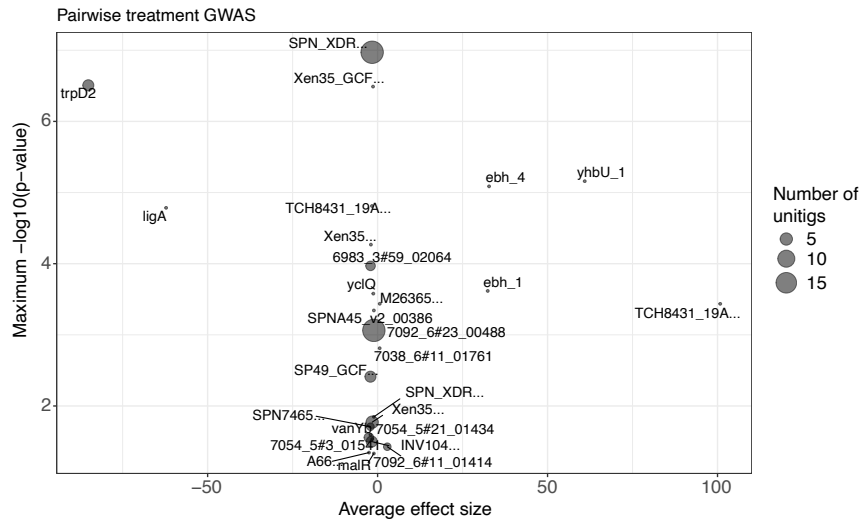


Figure 48: A dotplot indicating the average effect size and minimum p-value of the presence or absence of unitigs associated with antibiotic treatment. Points are annotated by the gene they were found in and their indicates the number of significant unitigs that were found within that gene.

is consistent with a more virulent form of *S. pneumoniae*. The lower frequency after treatment suggests that pneumococci with these unitigs are susceptible to antibiotic treatment. YhbU has also been associated with colonization of the lung in *Klebsiella pneumoniae* and is involved in peptidase activity [389]. Unlike TrpD2, the unitig corresponding to YhbU increases in frequency following treatment and decreases in those pairs without treatment (Figure 49). This could be driven by selection against this unitig in the absence of antimicrobial pressure.

A number of cell wall associated unitigs were also identified. These included unitigs corresponding to the giant Hyperosmolarity resistance protein Ebh. Ebh has been associated with cell size in *Staphylococcus aureus* and variants have been associated with increased susceptibility to the antibiotic oxacillin as well as complement-mediated killing [69]. Associations were also identified with known cell wall surface anchor family proteins, which are commonly associated with virulence [136]. Unitigs corresponding to the inner membrane protein tolA were also identified. TolA proteins are involved in the translocation of group A colicins which can act as toxins against other bacteria.

Finally, a number of phage associated unitigs were identified. Phages are known to be an important factor driving the evolution of *S. pneumoniae* and can introduce genes through horizontal gene transfer that encode for toxins or other virulence factors that can help the pneumococcus to persist during infection [117, 492]. The stress induced by antimicrobial treatment could also lead to an increased susceptibility to invasion by phages [134].

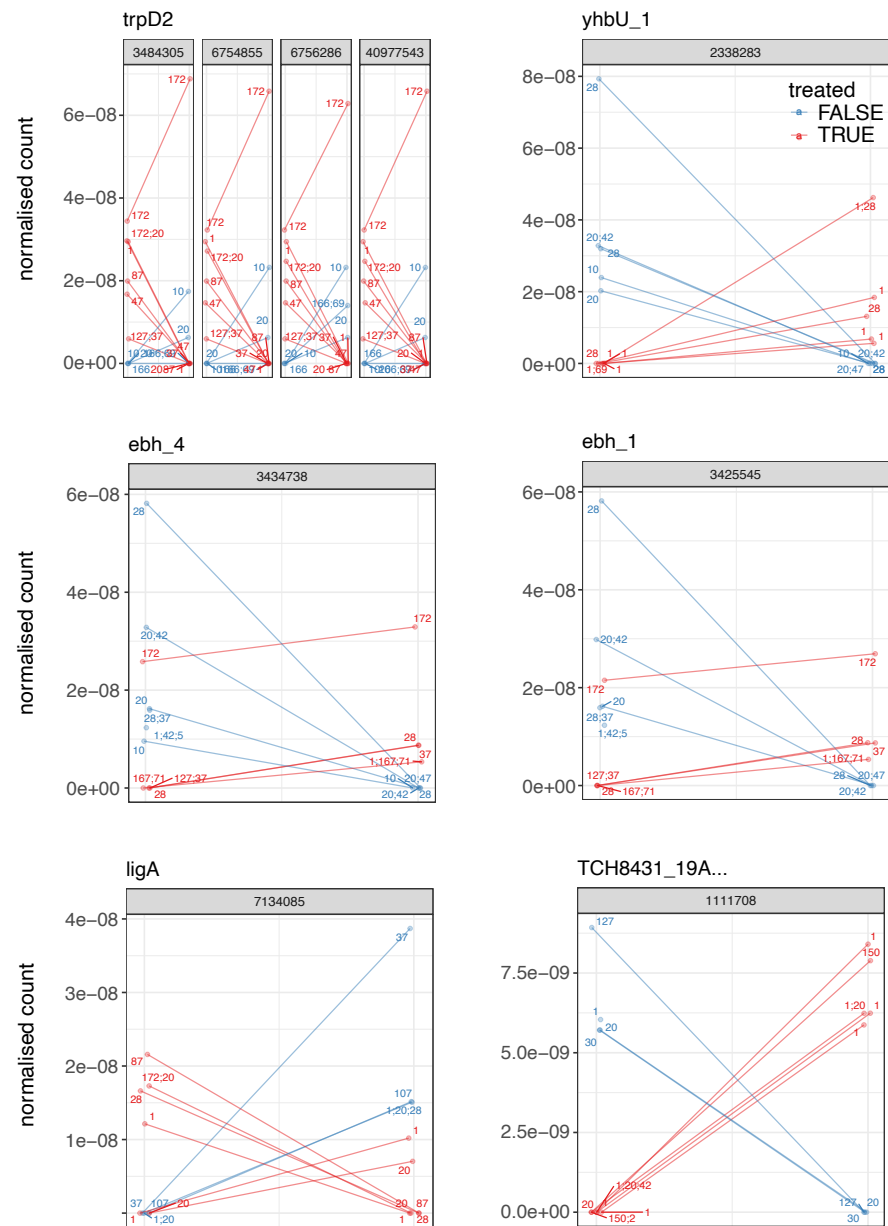


Figure 49: Lines indicating the normalised count of unitigs before and after antibiotic treatment that were found to be significantly associated with treatment. The colour indicates whether or not the pair of samples included an antibiotic treatment event inbetween sampling times.

annotated gene	avg beta	min adjusted p-value	number of unitigs	note
SPN_XDR...	-1.6	1.07E-07	18	chlorohydrolase (trzA)
trpD2	-85.1	3.11E-07	4	lung invasion
Xen35..	-1.34	3.24E-07	1	chlorohydrolase (trzA)
yhbU_1	60.94	6.93E-06	1	virulence in <i>K. pneumoniae</i>
ebh_4	32.8	8.18E-06	1	Ebh
TCH8431...	-1.66	1.55E-05	1	satellite prophage
ligA	-62.21	1.64E-05	1	ligase
Xen35...	-1.97	5.39E-05	1	cell wall surface anchor family protein
6983_3#59_02064	-2.07	1.07E-04	3	YSIRK signal domain, cell wall surface anchor family protein
ebh_1	32.4	2.42E-04	1	Hyperosmolarity resistance protein Ebh, cell wall
yclQ	-1.25	2.64E-04	1	
M26365...	0.64	3.66E-04	1	chlorohydrolase
TCH8431...	100.79	3.67E-04	1	phage
SPNA45_v2_00386	-1.1	4.54E-04	1	phage hyaluronidase
7092_6#23_00488	-1.1	8.66E-04	18	phage
7038_6#11_01761	0.61	1.54E-03	1	metallo-beta-lactamase superfamily protein 1
SP49...	-2.1	3.88E-03	4	TolA
SPN_XDR...	-1.1	1.42E-02	1	PblB, phage-related protein
Xen35_...	-1.59	1.69E-02	5	TolA
SPN7465...	-2.1	1.96E-02	2	TolA
vanYb	-2.08	2.54E-02	1	Programed cell death
7054_5#3_01541	-2.57	2.77E-02	3	
7054_5#21_01434	-1.56	2.92E-02	1	TolA
INV104...	-1.73	3.16E-02	4	TolA
7092_6#11_01414	2.85	3.74E-02	2	
A66_GCF...	-2.53	4.52E-02	1	transposase and inactivated derivatives
malR	-1.1	4.66E-02	1	transcriptional repressor of the maltosaccharide utilization operons malxCD and malMP

Table 17: Those unitigs found to significantly change in frequency following an antibiotic treatment event using the paired sample design. Only unitigs that could be annotated to the original collection of ~3100 genome assemblies from Chewapreecha *et al.*, [71] are included.

7.5 DISCUSSION

Deep sequencing of whole plate sweeps allows for the identification of resistance elements in samples containing multiple pneumococcal lineages. The resulting increased sensitivity led to the identification of 14.6% more resistance calls than using single colony picks. This increase in sensitivity has important implications for the surveillance and treatment of pneumococcal disease. It suggests that relying on resistance calls obtained from a single genome may not be sufficient to determine whether or not treatment with a particular antibiotic is likely to be effective. While we found that resistance elements were less likely to be found in multiple carriage than expected, the large increase in the number of resistance calls suggests that surveillance programs such as the Pneumococcal Molecular Epidemiology Network (PMEN) may underestimate the prevalence of drug resistance clones when relying on single genome data.

In [Chapter 6](#), we found that multidrug resistant lineages are less likely to be found in instances of multiple carriage. In this chapter, we have demonstrated that a similar signal can be seen in resistance elements after accounting for the background population structure. This suggests that it is the resistance elements themselves that are driving the reduced presence in multiple carriage. Given many laboratory studies have demonstrated that resistance often comes at a fitness cost [[428](#), [487](#)], the reduction in multiple carriage could be explained by resistant lineages being outcompeted by susceptible strains within the host. This is consistent with the recent modelling study of Davies *et al.*, [[111](#)], which demonstrated that the coexistence of resistant and susceptible strains within a population could be driven by the interaction between susceptible lineages and less fit resistant lineages with the host. However, it is unclear how this reduction in fitness would allow for the increased carriage duration observed in many multidrug resistant lineages [[295](#)].

The additional insights provided by deep sequencing whole plate sweeps enabled us to consider the impact of antibiotic treatment on the genetic diversity of *S. pneumoniae* within the host. We identified that children were more likely to be colonised with the multidrug resistant lineage GPSC 1 following antimicrobial treatment. This was in contrast to a previous study which suggested that the increase in frequency post treatment was due to the elimination of susceptible strains [[495](#)]. When treating patients it is thus important to consider the implications of a resistant lineage taking the place of the current infection. If this represents a significant risk, it could motivate the prescription of broader spectrum antibiotics in the initial treatment.

The impact of treatment at the locus level revealed a number of hits to genes related to both antimicrobial resistance and virulence. These included hits to both an efflux ABC transporter protein and the well

known penicillin-binding protein 2. The ability to identify these well known resistance features without the need to perform laboratory based experiments such as disk diffusions, demonstrates the added value high resolution epidemiological data can provide to pathogen genetic association studies. Furthermore, by comparing pairs of samples from the same host infected with a single lineage pre- and post-treatment we identified a number of hits to phages as well as genes associated with virulence and the pneumococcal cell wall. Many virulence genes are associated with the pathogen's ability to persist in carriage in response to stressors including antimicrobials. Thus, these hits may be related to the pathogen's stress response [253]. As phages are known to facilitate horizontal gene transfer, the identification of phage unitigs suggests that the pneumococcus may use phages as a way to quickly adapt to changes in its environment. Equally, the link to phages could be driven by the increased susceptibility of pneumococci to invasion by phages when subjected to antimicrobials [134].

Overall, we have shown that sequencing whole plate sweeps increases our ability to detect antibiotic resistance mechanisms in pneumococcal carriage. This coupled with an improved understanding of the response of pneumococci to antimicrobial treatment could enhance future antibiotic surveillance, intervention and treatment strategies.

8.1 BACKGROUND

Historically, relatively little attention has been paid to the genetic diversity present within the host at the sub-lineage level in pneumococcal carriage. Similar to much of our analysis in Chapters 6 and 7, the majority of studies have only considered within-host diversity in terms of multiple carriage events of distinct pneumococcal lineages. This was in part due to the expectation that the relatively low evolutionary rate in most bacteria would lead to within-host populations of a single lineage being nearly entirely monomorphic [130]. Recently, studies have demonstrated that the within-host evolutionary rates of *S. pneumoniae* are up to an order of magnitude larger than estimates based on diverse datasets representing many years of evolution [63].

In addition to allowing the pneumococcus to adapt to antimicrobial treatment as described in the previous chapter, this increased evolutionary rate has important implications for the pathogen's ability to evade the host immune system and compete with other bacteria including other pneumococci. The increased rate of the *de novo* generation of diversity within pneumococcal carriage also has implications for genomic epidemiology including the identification of transmission chains and the characterisation of the transmission bottleneck. As described in Chapter 4, within-host diversity can complicate transmission inference efforts based on the analysis of single genomes. By concentrating on fine scale variation at the sub-lineage level we are able to better understand the mutational and selective forces acting within pneumococcal carriage in addition to improving our understanding of the transmission *S. pneumoniae*.

8.1.1 Transmission

A significant benefit of the additional information on within-host diversity provided by whole plate sweeps is the increased resolution it provides for the inference of transmission. Previous analyses of pneumococcal transmission has for the most part relied solely on epidemiological information or genomic data from single colony picks [169, 380]. Whilst this can provide valuable insights into the dynamics of transmission, the high rate of co-infection of multiple lineages limits the resolution of these approaches. In this chapter, we describe an extension to the Transcluster algorithm discussed in Chapter 4, allowing it to be applied to mixed infections in order to provide an upper

bound on the distribution of the number of intermediate hosts that separate two infections.

The increased sensitivity of this approach allows us to further investigate the transmission dynamics of *S. pneumoniae*. There is currently limited information on which age groups are responsible for seeding the majority of pneumococcal transmission events [532]. A recent modeling study based on pneumococcal carriage prevalence data found that infants under the age of one contribute significantly to the number of secondary infections per capita but that children between 1-5 years of age play a much larger role in the population level infection pressure [169]. Supporting this finding, an analysis of the single colony pick genome data from the Maela refugee camp identified a strong geographic structure in likely transmission events and found that people were 4.5 times more likely to be infected by a member of their household than from the wider community [380]. By identifying which age groups contribute the most to transmission as well as the interaction between parents and children more targeted intervention strategies can be designed. Our previous analysis of the impact of antibiotic treatment on the likelihood of being colonised with a resistant strain (Chapter 7) adds to the importance of understanding the dynamics of transmission between parents and children. If a vulnerable member of a family is treated with antibiotics, understanding whether they are likely to be susceptible to infection from a family member carrying a resistance strain could motivate the use of preventative antimicrobials. Identifying pairs of infections linked closely by transmission also allows us to examine the transmission bottleneck in pneumococcal infection. Previous studies focused on mouse models of pneumococcal transmission observed a tight transmission bottleneck of a single bacterium [271]. Such tight bottlenecks are possible targets for interventions as the pathogen's diversity is at a minimum and thus vulnerable to treatment and competition.

8.1.2 Within-host mutation

The additional information on within-host minority variants provided by sequencing whole plate sweeps allows us to examine the mutational and selective forces driving the evolution of *S. pneumoniae* within the host. An analysis of single genomes obtained from longitudinal samples of pneumococcal carriage episodes in the Gambia identified that the within-host evolutionary rate was up to 40 fold higher than that observed between hosts over longer time scales [63]. By analysing the mutational spectrum that corresponds to this elevated evolutionary rate we can gain insights into the mutational processes behind pneumococcal evolution.

It has been suggested that an increased mutation rate can provide the genetic variation in bacterial cells on which selection can act to

improve the fitness of populations under stress [278]. Modelling and laboratory studies have gone on to show that depending upon the scenario, faster mutating strains can either outcompete or be outcompeted by slower mutating strains [93]. The vast majority of mutations with phenotypic effects are likely to be deleterious and thus in general selection will act to reduce the mutational load in the absence of external stressors. Thus while at short timescales an increased mutational rate may be beneficial to respond to factors such as the immune response or antimicrobial treatment, selection is likely to reduce the evolutionary rate over longer timescales. The ability of bacterial populations to switch between faster and more slowly mutating states was demonstrated in the well known *E. coli* Long-Term Evolution Experiment (LTEE) in which 12 cell populations have been serially propagated for more than 60,000 generations. Here, one population was observed to revert to a reduced mutational rate after being dominated by a high rate of mutations for over 10,000 generations [195].

When considering mutational spectra we are often interested not only in single nucleotide base changes, but also in the broader sequence context of such mutations. Biases based on mutational sequence context have been observed in a number of settings driven by both different mutagens and different DNA repair mechanisms. Analysis of the broader context of mutational spectra has proved useful in a diverse range of organisms including providing remarkable insights into the origins of different human cancers [173]. While there have been a number of studies into the mutational spectrum of a select few bacterial species, most notably *E. coli* the majority of these have been in laboratory isolates. Relatively little is known about the mutational processes behind the evolution of *S. pneumoniae* within the host. By comparing the mutational spectrum observed between hosts to that seen in minority variants identified from whole plate sweeps we can gain a better understanding about the different mutational processes driving pneumococcal evolution at different time scales.

8.1.3 Within-host selection

Analysing within-host minority variants also allows us to consider the role of selection on pneumococcal evolution during carriage. An excess of nonsynonymous mutations compared to silent synonymous mutations, usually expressed as the d_N/d_S ratio can indicate whether there is evidence of adaptive change occurring within the host. Conversely, a lower d_N/d_S ratio can indicate purifying selection [360, 374]. Although the interpretability of the d_N/d_S has seen it used widely in population genetics, the approach has some well-known limitations. Notably, the d_N/d_S ratio is expected to vary even in a constant environment as younger variants with lower exposure to the forces of selection may have ratios closer to one [341]. When considering pos-

itive selection, the approach is also most sensitive to diversifying selection, where any amino acid change is favoured over the current state. Fortunately, the scenarios under which diversifying selection is likely to occur overlap well with those we might expect to see in short term within-host selection. Such scenarios include arms races between competing strains or the host immune system. We might also expect to see diversifying selection in response to antibiotic pressure due to the different antibiotic exposures from host to host [130]. The utility of using d_N/d_S ratios to investigate selection acting over short time scales has motivated the development of bespoke techniques such as dNdScv; a method originally designed to detect cancer driver genes from somatic mutation data [331]. The simplicity of the d_N/d_S approach has also motivated the development of methods that can scale to the very large bacterial genomics datasets that are currently being generated [509].

8.1.4 *Characterising minority variants presents additional challenges*

Typically in the analysis of sequencing data from single bacterial colony picks, sequencing reads are aligned to a reference genome and the consensus allele at each position in the reference is taken as being representative of the genome for the colony [51]. The consensus method is robust to low levels of contamination and sequencing errors. However, whole plate sweeps allow us to examine not only the consensus allele but also minority variants that appear at lower frequencies. This presents additional challenges as sequencing errors, contamination and multi-copy gene families impact on our ability to identify real within-host minority variants. Despite aligning reads to a pneumococcal reference genome, contamination of the culture plate with closely related bacterial species represents a challenge as regions of homology between the respective genomes of the two species can lead to contamination being incorrectly called as within-host pneumococcal variation. Similarly, multi-copy gene families can lead to misleading signals. If a reference genome contains a single copy of a gene when the sampled genome contains multiple copies the additional versions of the gene can appear as minority variants within the final alignment. To negotiate these challenges we have implemented a conservative variant calling pipeline that is able to filter out sequencing errors as well as masking regions that are likely to be impacted by low level contamination or multi-copy gene families.

In this chapter, we start by examining within-host variants in the context of transmission focusing on the relationship between mother and child where our prior expectation of a plausible transmission link is strong. Concentrating on samples consisting of a single lineage we next consider how within-host diversity changes over the course of a carriage episode. Using the same variant calls we go on to look at

the mutational spectrum within pneumococcal carriage and compare it to that found over longer timescales. Finally, using methods originally designed for investigating somatic mutations in cancer, we explore the d_N/d_S ratio at both the lineage level and gene level within pneumococcal carriage.

8.2 METHODS

8.2.1 *Calling within-host variants*

The accurate identification of single-nucleotide polymorphisms (SNPs) from bacterial sequencing data is important for a wide array of applications from identifying antibiotic resistance features to tracking transmission. While the range of methods developed to identify SNPs from sequencing data of single colony picks is well established, relatively few studies have focused on identifying minority variants which are critical to the analysis of within-host diversity using whole plate sweeps [51]. The identification of minority variants is hampered by a number of confounding factors including low level contamination from non-target bacterial species, multiple infections of different strains of the same species and the impacts of multi-copy gene families. Although algorithms such as mGEMS allow for deconvolution of mixed samples of the same species, these approaches are usually designed to separate mixtures with clearly distinct lineages such as GPSCs. Their ability to identify sub-lineage variation generated within a carriage episode is largely unexplored and the added difficulty of this problem is likely to lead to an elevated error rate. Instead we focus on samples where only a single GPSC was identified which avoids the problem of first deconvoluting separate lineages.

Despite this, genomic regions of homology between the target and a contaminating species can still lead to minority variants being incorrectly called. Similarly, the high level of homology between multiple copies of the same gene within a single genome can give a misleading signal of within-host variation. Fortunately, in both of these scenarios it is highly likely that the regions of overlapping homology will lead to an elevated number of polymorphisms when compared to the overall background rate of within-host diversity in a sample. By using a sliding window approach inspired by the Gubbins algorithm we are able to identify these potentially confounding regions allowing us to filter out erroneous within-host variant calls [104]. Coupled with a conservative variant calling pipeline we can mitigate the impact of sequencing errors, low level contamination and multi-copy gene families.

8.2.2 *Variant calling pipeline*

To identify within-host minority variants we used the LoFreq variant caller as it has been shown to generate robust minority variant calls and accounts for base call qualities and alignment uncertainty to reduce the impact of sequencing errors [51, 508]. As it has been shown that the choice of reference genome can dramatically impact the quality of variant calls, we selected a representative genome assembly for each GPSC cluster which was used as a reference for those samples found to contain that GPSC. Pairwise mash distances were calculated for each GPSC cluster in the reference database described in [Chapter 6](#). The medoid of each cluster of genomes G was then identified such that

$$g_{\text{medoid}} = \arg \min_{y \in G} \sum_{i=1}^N d(y, x_i) \quad (39)$$

The pipeline suggested in the LoFreq manual was used to call variants after first aligning reads for each sample to the respective GPSC assembly. Reads were aligned to the chosen reference genomes using BWA v0.7.17-r1188. The Picard tools v2.23.8 ‘CleanSam’ function was then used to soft clip reads aligned to the end of contigs and to set the alignment qualities of unaligned reads to zero. Pysamstats v1.1.2 was run to provide allele counts for each location of the aligned reference for use in the transmission analysis. The LoFreq pipeline v2.1.5 was run initially with stricter filters requiring a coverage of at least 10 reads to identify a variant. The resulting variant calls were used along with the read alignment as input to the GATK BaseRecalibrator tool v4.1.9 as suggested in the LoFreq manual to improve the estimated base quality scores. Finally the LoFreq pipeline was run for a second time with a reduced coverage requirement of 3 reads. The resulting variant calls were only considered if there was support for the variant on at least one read in both the positive and minus strand.

8.2.3 *Identifying problematic regions*

To identify problematic variants that were likely the result of low level contamination or multi-copy gene families we implemented a similar approach to that used to identify recombination in the tool Gubbins [104]. A scan statistic was used to identify regions of the alignment with an elevated number of polymorphisms. Assuming that within host variants are relatively rare and should be distributed fairly evenly across the genome, regions with a high number of polymorphisms are likely to be the result of confounding factors and can thus be filtered out. We assumed a null hypothesis H_0 , that the number of polymorphisms occurring in a window s_w follows a binomial

distribution based upon the number of bases within the window w and the mean density of polymorphisms across the whole alignment. We chose w for each sample such that $E(s_w) = 1$. A window centred at each polymorphism was then considered and a one-tailed Binomial test was performed to determine if that window contained an elevated number of polymorphisms. After adjusting for multiple testing using the Benjamini-Hochberg method those windows with a p-value < 0.05 were selected and combined if they overlapped with another window [28]. To more accurately define the edges of each region, we assumed each combined window conformed to an alternative hypothesis H_{1r} where the number of polymorphisms s_r also followed a binomial distribution with a rate based on the length of the window l_r and the number of polymorphisms within the window s_r . Each end of the window is then progressively moved inward to the location of the next polymorphism until the likelihood of H_{1r} relative to H_0 no longer increases. The resulting final windows were then called as potential problematic regions if they satisfied the inequality

$$\frac{0.05}{g/l_f} > 1 - \sum_{i=0}^{s_f-1} \binom{l_f}{i} d_0^i (1 - d_0)^{l_f-i} \quad (40)$$

Where l_f is the length of the final window, g is the length of the reference genome and d_0 is the expected rate of polymorphisms under the null hypothesis. The left hand side of Equation 40 accounts for the possible number of similarly-sized non-overlapping windows in the reference.

An example of the output of this approach is given in Figure 50, which indicates that the algorithm accurately identifies regions that are likely to be problematic for the analysis of within-host diversity.

8.2.4 Transmission

A number of algorithms have been developed to infer putative transmission links by taking advantage of information on within-host diversity provided by deep sequencing data including BADTRIP and SVA [114, 516]. These algorithms have been shown to improve on estimates that rely on single representative genomes from a sample. However, they make a number of assumptions that make them unsuitable for the analysis in this chapter. Importantly, both approaches assume that all infections have been sampled in an outbreak which is highly unlikely in the analysis of the endemic transmission of *S. pneumoniae* in the Maela camp. Whilst our novel simulation based approach described in Chapter 4 was originally developed with this study in mind, we have yet to show concretely that it is able to provide consistent results over a wide range of scenarios. Instead in this chapter we take a more conservative approach and adapt the Tran-

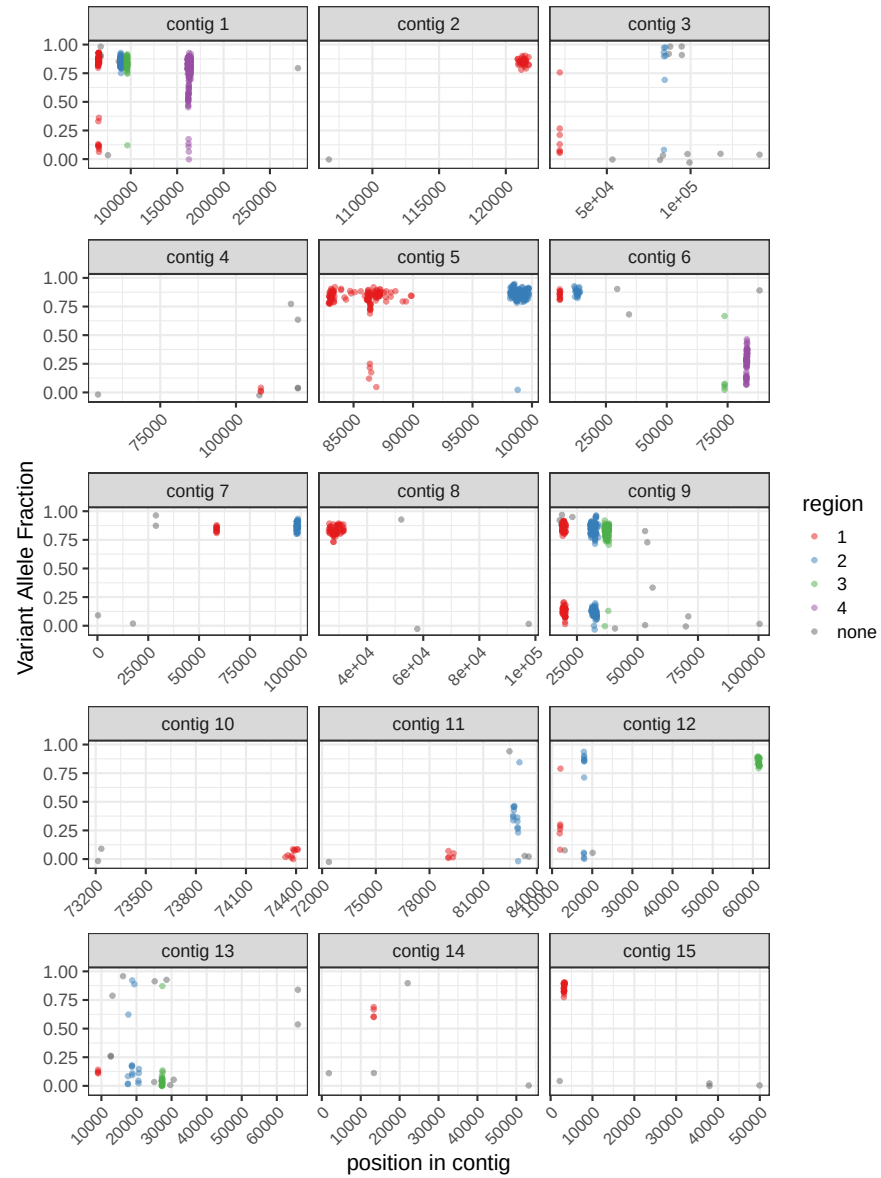


Figure 50: An example of the regions inferred to be spurious in a sample aligned to the 15 contigs of the reference assembly of GPSC 1. Each region identified in each contig by the binomial scan statistic is given a different colour to allow them to be distinguished.

sCluster algorithm described in [Chapter 4](#) to provide an upper bound estimate on the distribution of the number of intermediate hosts separating two samples [468].

In place of the core genome pairwise SNP distance used in the original TransCluster algorithm, for each pair of samples that share a GPSC, we calculate the number of sites along the reference genome that share no matching alleles between each pair of samples. The allele counts at each site along the reference are taken from the output of pysamstats as described in [Section 8.2.2](#). Regions of low coverage (less than 3 reads) in one sample are not counted towards the overall distance. If we assume that each sample represents the diversity within each host then this approach identifies the minimum theoretical SNP distance between any two pneumococcal genomes in each host. As such it provides an upper bound for the probability distribution of the number of intermediate hosts. We assume a mean generation time of 2 months and an evolutionary rate of 5.3 SNPs/genome/year as calculated from the analysis of the pneumococcal genomes from the original single colony pick study into Maela described in [Chapter 5](#).

8.2.5 *Mutational spectrum*

In the mutational spectrum analysis of human cancers, normal samples are usually taken along with samples of the cancer to allow for somatic mutations to be distinguished from germline mutations. As we can not be sure which alleles were present at the start of a pneumococcal carriage episode we can not be certain of the direction a mutation occurred in. For example, it is difficult to distinguish between a A>C and a C>A mutation. Instead, we consider the difference between the consensus and minority variants at each site in the reference genome. If we assume that the colonising variant typically dominates the diversity within an infection then this approach corresponds with the direction of mutation. To account for the context of each mutation we considered the consensus nucleotide bases either side of the mutation. These were then normalised to account for the overall composition of the reference genome for each GPSC. The normalised mutation rates (r) for each of the 192 possible changes (j) in a trinucleotide context were calculated as:

$$r_j = \frac{n_j}{L_j \sum_j \frac{n_j}{L_j}} \quad (41)$$

Where n_j is the total number of mutations observed for a trinucleotide change j , and L_j is the total number of times that the corresponding trinucleotide is present in the reference genome. To avoid double counting the same mutation, each variant was only counted once per host.

To compare with the mutational spectrum observed across a longer timescale we considered the recombination filtered alignments of 7 major sequence clusters generated in the original publication of the single colony pick analysis of the Maela dataset [71]. We used Iqtree v2.1.2 to build a maximum likelihood phylogeny for each alignment using a General Time Reversible (GTR) model with 4 rate categories and enabled the ‘ancestral’ option to reconstruct the sequences at the internal nodes of the resulting phylogeny [358]. Mutations were called by considering changes in alleles between consecutive nodes of the phylogeny and the mutational spectrum was normalised using the trinucleotide frequencies in the reconstructed ancestral sequence of the root node.

8.2.6 Selection

Traditional d_N/d_S methods were developed to investigate selection acting across large timescales with many studies considering datasets comprising multiple species. These approaches typically rely on Markov-chain codon substitution models [193]. More recently, the sequencing of large human cancer datasets has necessitated the development of methods such as dNdScv which have been designed for more closely related genomes where observed changes often represent individual mutation events [331]. Here, mutations are modelled using a Poisson framework allowing for more complex and realistic substitution models to be considered. Using the filtered variant calls, we make use of the default substitution model in dNdScv which considers 192 rate parameters to model all possible mutations in both strands separated by their trinucleotide context. The default model also considers two additional parameters to estimate the d_N/d_S ratios for missense and nonsense mutations separately. We fit this model to each GPSC separately where there were at least 10 hosts infected with the single GPSC. We considered both the global d_N/d_S estimates as well as estimating d_N/d_S parameters at the gene level. In the case of the gene level estimates, dNdScv incorporates global estimates for the variation rate across genes as well as accounting for the sequence composition of the gene.

8.3 RESULTS

8.3.1 Transmission links suggest strong geographic structure

To explore the structure of putative pneumococcal transmission links within the Maela camp we compared the pairwise geographic distance of the household from which each sample was taken with the mean probability the two samples were the result of a direct transmission. Figure 51, indicates that samples taken from people

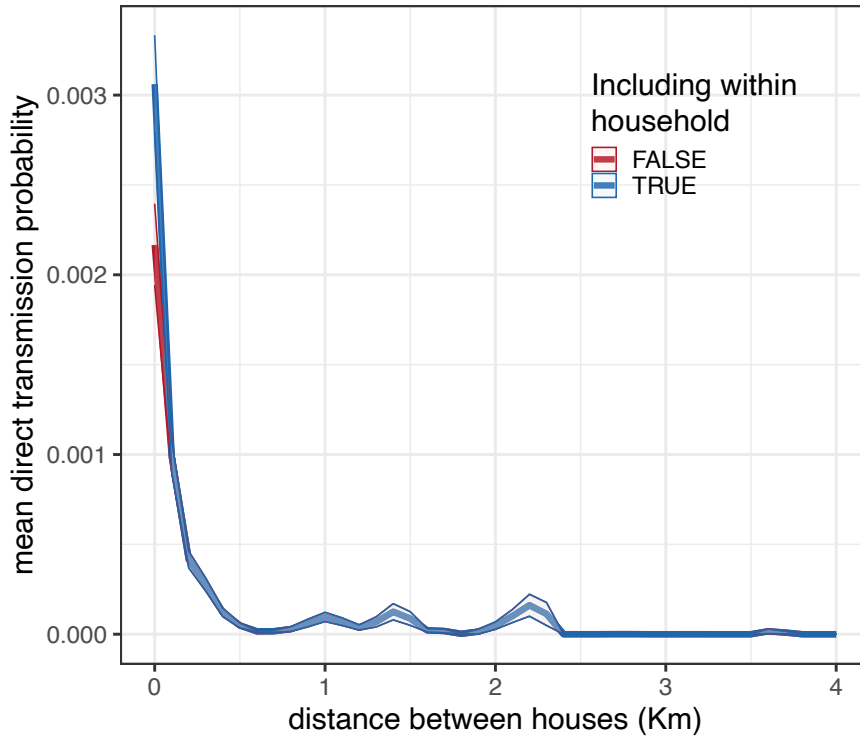


Figure 51: The mean probability of a direct transmission inferred using the TransCluster algorithm adapted for within-host variants compared to geographic distance determined as the straight line distance between homes. The mean has been calculated over a rolling window of 0.1km. Mother/child pairs have been removed from the red line indicating that the geographic is not driven solely by within-household relationships.

living in close proximity within the camp have an increased probability of being related by a direct transmission event. Although some of this geographic structure is driven by transmission events between samples obtained from both mothers and children within the same household, after removing these pairs the strong geographic structure remained (Figure 51). This result corresponds well with both our previous analysis of single colony picks using TransPhylo in Chapter 5 as well as with the approach of Numminen *et al.*, [380]. Given the small size of the Maela camp and as people are able to freely move throughout the camp it is somewhat surprising that we observe such tight geographic structure. This has important implications for local interventions targeting both pneumococcal disease as well as other respiratory pathogens that transmit and colonise in a similar fashion. For example, if a virulent strain such as serotype 1 is observed then neighbouring households may be at elevated risk. Knowing this, measures could be taken to minimise the risk to vulnerable people in close proximity to an infected household.

8.3.2 *Asymmetric transmission between mother and child*

Arguably the closest contact an infant will have in the early stages of their life is with their mother. Accordingly, developing a strong understanding of the transmission dynamics between mothers and their children could inform both the treatment and prevention of pneumococcal infections. To examine the transmission relationship between mother and child we considered all pairs where we identified the same GPSC in both mother child. This resulted in 70 such pairs being identified. Using the TransCluster approach described in [Section 8.2.4](#), we identified 48/70 pairs where the probability of less than 5 intermediate hosts separating two samples exceeded 1%. This corresponds to a maximum pairwise SNP distance of 12 which is a similar threshold to that used in other studies [56]. We complemented the transmission information provided by the sequencing data with the extended set of serotype calls obtained via latex sweeps available for both mother and child with a representative example given in [Figure 52](#). We assigned a putative direction to each of the sequence supported transmission links if the serotype involved was observed in the donor host prior to the recipient and within 2 months of the recipient becoming infected with the same serotype. If this condition was not met the link was designated an ambiguous direction. Within the 48 mother/child pairs we identified 66 transmission links. Of these 16 were suggestive of mother to child transmission, 22 were ambiguous and 28 were consistent with transmission from the child to their mother as shown in [Figure 53](#). The utility of the additional information on within-host diversity provided by whole plate sweeps is highlighted by the fact that 23/66 transmission links involved a sample that contained more than one GPSC. This indicates that in this dataset, up to a third of the identified transmission links through sequencing could have been obscured if only genomes from single colonies were considered.

These results suggest that children are significantly more likely to transmit to their mothers than vice-versa (p-value of 0.048 using an exact binomial test with a null model of an even chance of transmission in either direction). This is consistent with a recent modeling study conducted across a wide range of countries investigating pneumococcal transmission dynamics from carriage prevalence data [169]. In this study the authors observed that <1 year-old children, while still very frequently causing secondary infections, had a lower contribution to the force of infection than children between 1 and 5 years of age. To investigate this finding in our dataset we repeated the analysis only considering those links that occurred prior to the child turning one. This resulted in 16 mother to child events, 16 ambiguous events and 17 putative child to mother transmission links. Strikingly, this is consistent with a much higher rate of transmission from mother to child in the first year of life (16 - 0). In addition to being consistent

with the previous modelling study, this result could have considerable importance for the design of preventative treatment and vaccination strategies to prevent disease in very young children.

8.3.3 *Close transmission pairs are more likely to share minority variants*

Transmission bottlenecks are important points in the evolution of the pneumococcus where the population size and genetic diversity is usually at its lowest, making the pathogen particularly vulnerable to intervention strategies. To date, most studies investigating the size of the bottleneck in pneumococcal transmission have made use of animal models, usually infecting mice with multiple strains of genetically marked pneumococci to estimate how many bacteria are transmitted to a new host [271, 310]. These studies have suggested that pneumococcal transmission involves a very tight bottleneck with Kono et al, [271] suggesting that transmission consisted of only a single cell. While these studies have provided valuable insights into the dynamics of pneumococcal transmission, the extent that they can be generalised to the transmission of *S. pneumoniae* in humans remains unclear.

To investigate the transmission bottleneck using our data, we counted the number of shared minority variants between every pair of samples that contained a single pneumococcal lineage from the same GPSC. We then plotted the resulting distribution of counts separating those pairs that had a probability of being within 5 intermediate hosts of at least 1%, the same threshold used in our analysis of mother/child transmission pairs as shown in Figure 54. Unlike many common methods for investigating the transmission bottleneck this approach does not rely on well identified direct transmission pairs. However, it is more limited in that it does not provide an estimate for the actual size of the transmission bottleneck and rather indicates whether minority variants are likely to be transmitted or not. As transmission is a highly stochastic process, exact estimates of the mean bottleneck size may be of limited use and thus we focus on the simpler and possibly more important question of whether within-host diversity can be transmitted.

Figure 54, indicates that the median number of shared within-host variants is significantly higher amongst those pairs of samples where the probability of close transmission is high. The distributions were compared using a Poisson generalised linear model with the number of shared variants found to be significantly higher in both the close transmission pairs and the close transmission of mother/child pairs ($p < 0.001$). While there are still a substantial number of close transmission pairs that share zero minority variants, the elevated number of shared variants in this category is consistent with the transmission bottleneck not being constrained to a single bacterium. This would

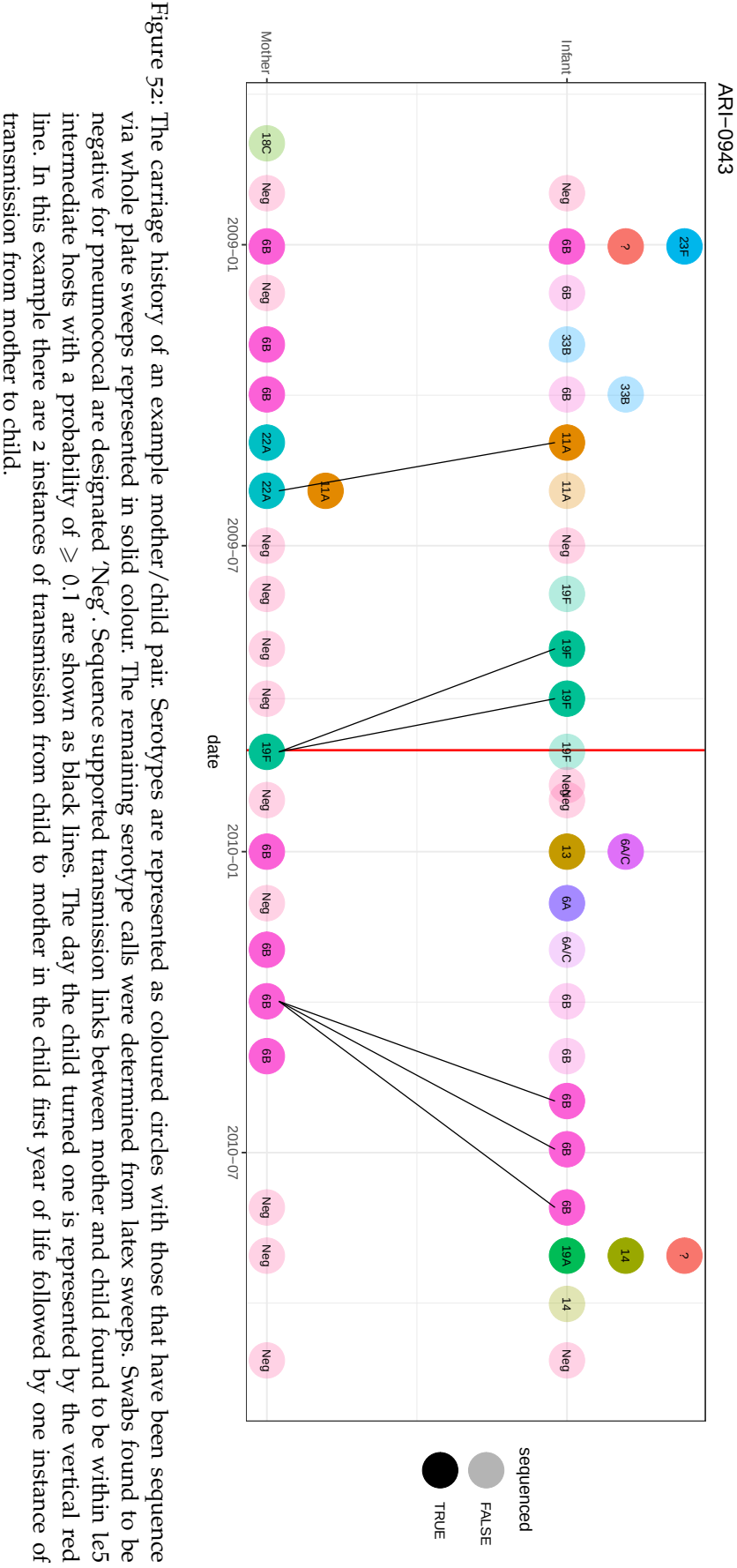


Figure 52: The carriage history of an example mother/child pair. Serotypes are represented as coloured circles with those that have been sequence via whole plate sweeps represented in solid colour. The remaining serotype calls were determined from latex sweeps. Swabs found to be negative for pneumococcal are designated 'Neg'. Sequence supported transmission links between mother and child found to be within 1e5 intermediate hosts with a probability of ≥ 0.1 are shown as black lines. The day the child turned one is represented by the vertical red line. In this example there are 2 instances of transmission from child to mother in the child first year of life followed by one instance of transmission from mother to child.

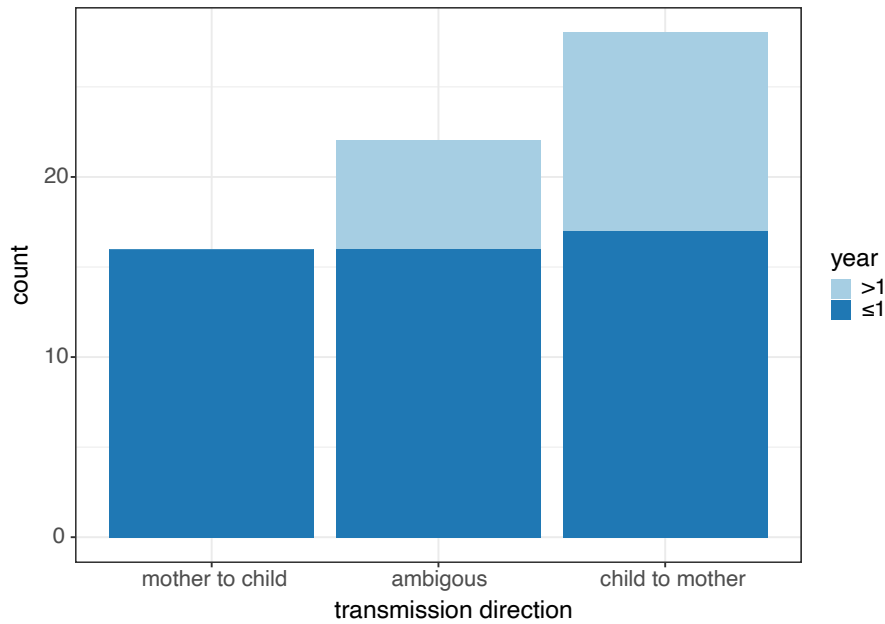


Figure 53: A stacked barplot of the number of transmissions from mother to child over all pairs split by direction and whether the transmission occurred prior to the child turning one. Those transmission links where the direction was unclear are represented as 'ambiguous'.

allow for the maintenance of within-host diversity through multiple carriage events allowing the pneumococci to more easily adapt to environmental changes in a new host. It should be noted that the increased number of shared variants could also be the result of multiple transmission events between the same two hosts. However, recent experimental studies have suggested that resident strains may outcompete newcomers which could suggest that this increased diversity is transmitted in a single event [449]. While the number of shared variants in sample pairs that are unlikely to be linked closely by transmission is significantly lower there are still a sizable number of sample pairs that share a large number of minority variants. This is consistent with some mutations being more likely to occur than others which could be the result of certain sites having higher mutation rates or be driven by the effects of selection discussed in more detail in [Section 8.3.6](#).

8.3.4 *Despite significant heterogeneity in the diversity of minority variants in carriage there is evidence for parallel within-host evolution*

Obtaining an accurate picture of the variability of within-host diversity in pneumococcal carriage is important in understanding how representative a single sample from an infection is likely to be of the broader population dynamics. Consistent changes in the diversity of

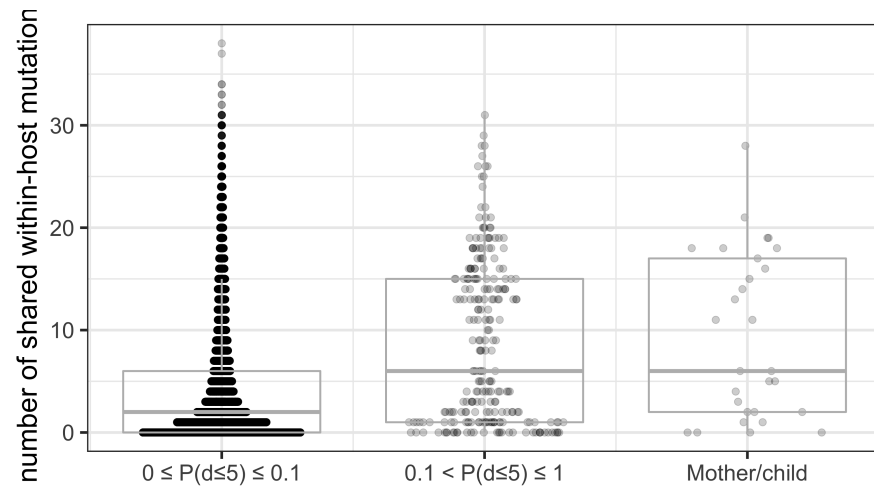


Figure 54: Boxplots indicating the distribution of the number of shared minority variants for: all pairwise relationships; those found to be within a transmission distance of ≤ 5 intermediate hosts with a probability of ≥ 0.1 ; and the subset of these that consisted of mother/child relationships. The increase in the number of shared minority variants at higher transmission probabilities suggests that the mean transmission bottleneck is greater than one.

variation within pneumococcal carriage could be informative for estimating the time at which a person was infected. Conversely, high variation between consecutive samples could result from repeated bottlenecks induced by within-host competition, the immune response of the host or from the sampling process itself. To explore the diversity of within-host variation during carriage we considered those samples where a single GPSC was found in the same host over multiple consecutive months and where no antimicrobial treatment had occurred during this period or within a month of the first sample. We found a high level of variability in the number of within-host minority variants found between consecutive samples. To test whether there was any consistent change in the number of minority variants between the first time a GPSC was sampled in a host and later samples we used a Poisson generalised linear mixed model with a random effect to account for host level differences such as the timing of the first sample within a carriage episode. No significant trend was identified. The highly stochastic nature of the number of within-host minority variants is exemplified in [Figure 55](#), which indicates the number of minority variants over time for the 6 carriage events where at least 4 timepoints were sampled.

Despite this high level of variability, a number of mutations were observed as minority variants repeatedly in different hosts. [Figure 56](#), indicates that while the majority of mutations are only observed in a single sample, consistent with neutral evolution some mutations ap-

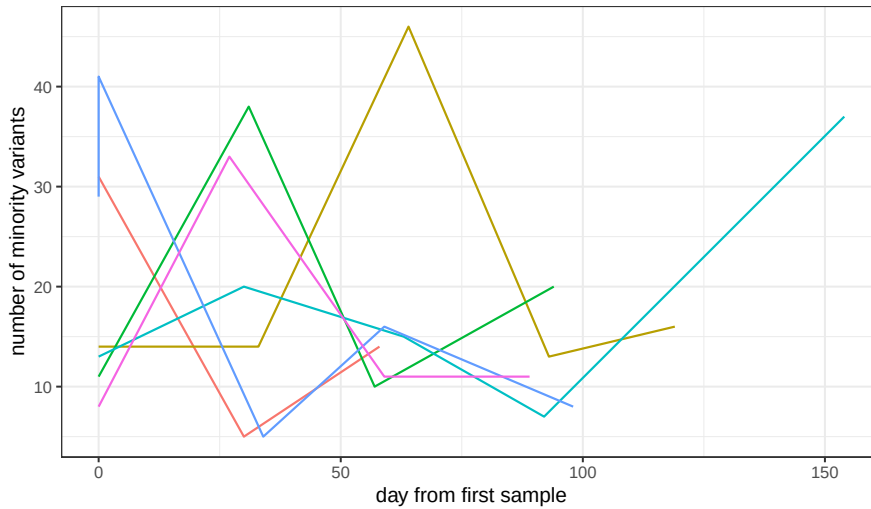


Figure 55: The number of minority variants found for each sample taken from a host where the same single lineage was found in at least 4 consecutive samples. Lines for the 6 separate hosts are represented as different colours. The high variability indicates there is not a consistent increase or decrease in the diversity of sampled variants through carriage.

pear in up to 122 hosts. Repeated examples of the same mutation were found in both genic and intergenic regions. While biases introduced by the culture, sequencing or bioinformatics processing of these samples could lead to similar results, the fact that shared minority variants were found to correlate with the probability of transmission suggests that experimental artefacts are unlikely to be the main source of this signal. Moreover, similar results have been observed in an independent study of pneumococcal carriage in the Gambia where a different experimental setup consisting of the sequencing of consecutive single colony picks also identified a number of recurring within-host mutations [63]. These results could be consistent with parallel evolution within the host as a result of selection which we consider further in a subsequent section.

8.3.5 *The mutational spectrum of *S. pneumoniae* is heterogeneous across different timescales*

Understanding the rate and profile of spontaneous mutations within *S. pneumoniae* is fundamental to accurately characterising the molecular processes driving its evolution. To investigate the within-host mutational spectrum of the pneumococcus, we made the simplifying assumption that a minority variant was likely to have mutated from the consensus variant at that location in the genome. This assumes that the bottleneck in pneumococcal transmission is tight and that the founding variants are likely to dominate due to the additional

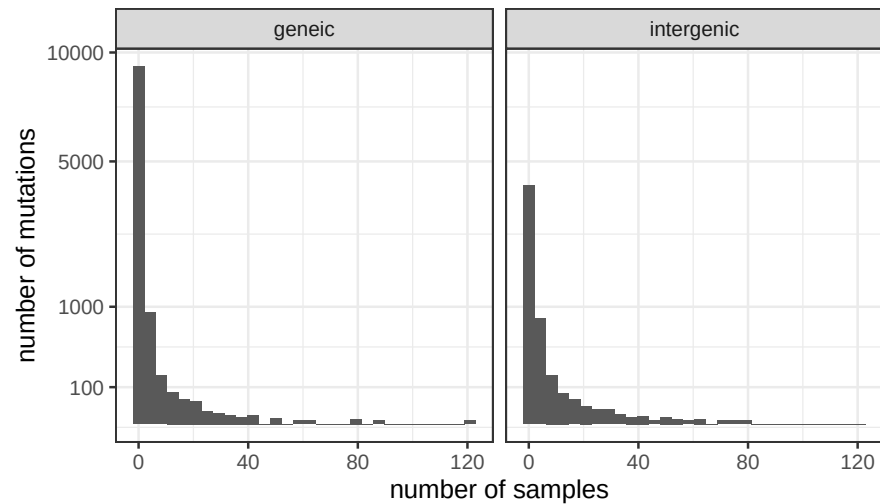


Figure 56: Histograms indicating the distribution of the number of times a minority variant is seen in different hosts, split by whether the variant was found in a genic or intergenic region. Although most variants are found only once, a number are found many times suggestive of parallel evolution.

time in which they have to proliferate. Across all single infections we counted the number of unique substitution events, allowing a single variant to be counted only once per host. Similar to the study of mutational spectra in cancer genomics we also examined the genomic context of these mutations by considering the two neighbouring bases either side of the mutation. Figure 57 indicates the resulting mutation spectra with and without this contextual information after normalising for the sequence composition of the reference genome. To compare this spectrum with that observed between hosts over a longer timescale we considered the number of independent mutations in each of the major sequence clusters of the genomes obtained from the single colony picks in the original genomics study of *S. pneumoniae* in the Maela camp [71].

Figure 58 reveals that both the within-host and between host mutational spectrum is dominated by transitions mainly consisting of C>T mutations. This is similar to that observed in mutational accumulation studies of other bacterial species including *E. coli*, *V. fischeri* and *V. cholerae* involving the serial passaging of bacteria between culture plates in a laboratory [132, 219, 285]. The consistency with which C>T mutations occur as the predominant mutation across a varied array of bacterial species has been speculated to be simply the result of cytosine deamination, which may dominant other mutational sources such as replication and repair mechanisms [219]. The similarity of our derived mutational spectrum with previous studies coupled with the similarity between the mutational profiles of the between and within-

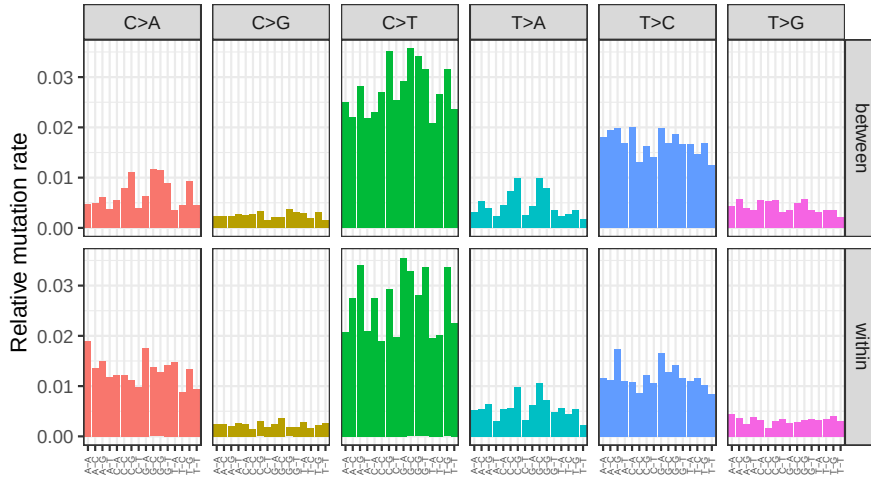


Figure 57: The mutational spectrum of between-host and within-host mutations split by their trinucleotide sequence context. Frequencies have been normalised to account for the background distribution in the reference (within-host) and root ancestral state (between-host).

host spectra suggests that our results are unlikely to be the result of experimental biases.

Strikingly, in addition to C>T mutations, the within-host mutational profile consisted of an excess of C>A transversions with weak sequence context when compared to that observed between hosts. This is consistent with a recent laboratory experiment into the mutational spectrum of *E. coli* which utilised a highly accurate sequencing approach (Maximum-depth sequencing) to detect extremely rare mutations in populations of bacteria [247]. In this study the authors identified an excess of short lived C>A mutations consistent with the differences we observe in the spectrum at the within and between host timescales. They observed that these mutations were unlikely to be the result of polymerase or sequencing errors and that increasing the treatment of in vivo DNA with fpg, a protein that excises purines reduced the rate of C>A substitutions. This is consistent with other studies investigating the impacts of oxidative stress on the mutation rates of bacteria [424]. Instead, the authors suggest that the excess of C>A substitutions may be driven by the misincorporation of As into C sites. These misincorporations are then reversed prior to genome replication, explaining the excess of C>A mutations at the short timescales observed within the host.

8.3.6 Evidence for selection acting within the host

The identification of multiple identical within-host minority variants suggests that signals of selection acting within the host may be evi-

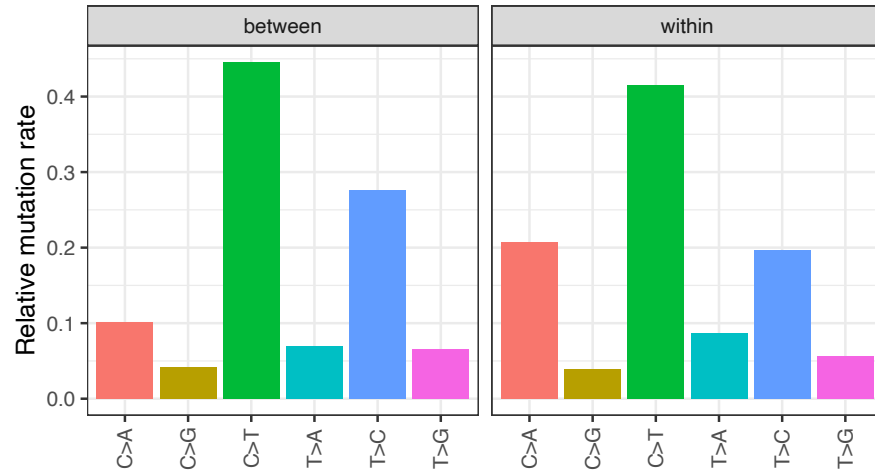


Figure 58: The mutational spectrum of between-host and within-host mutations. Frequencies have been normalised to account for the background distribution in the reference (within-host) and root ancestral state (between-host).

dent in our dataset. As the lifecycle of *S. pneumoniae* occurs nearly exclusively within its human host, characterising selection at this resolution can be informative for understanding the processes driving pneumococcal evolution across longer timescales. To investigate within-host selection we utilised the Poisson mutation model included in the dNdScv package originally developed for identifying driver mutations in cancer genetics studies. This approach is able to account for the fact that within-host mutations often represent individual mutation events [331]. We concentrated our analysis on the 10 most common GPSCs in our dataset. This provided sufficient statistical power while allowing us to make use of the more accurate within-host variant calls obtained by aligning samples to closely related reference genomes. It also enabled us to consider a wider range of genes over that found in the limited core set of genes common to all pneumococci.

Figure 59a, indicates the global d_N/d_S estimates for each of the GPSC's considered with separate estimates for both missense and nonsense mutations. In both cases the estimated d_N/d_S ratios were below 1 for all GPSCs indicating that purifying selection is the predominant force driving pneumococcal evolution within the host. As expected, these ratios are higher than those reported from the analysis of diverse pneumococcal datasets representing years of evolution [96, 288]. This can be explained by the fact that deleterious mutations are not purged instantaneously, allowing them to be observed in samples representing the broader population diversity present within the host [60]. With the exception of GPSC 25, the estimated d_N/d_S ratio for nonsense mutations was lower than for missense mutations which

is consistent with nonsense mutations being more likely to lead to the loss of essential functions. Despite the lower number of nonsense mutations observed, there was not a significant association between the variant allele frequencies of nonsense mutations when compared to missense and synonymous mutations (Figure 59b). Interestingly, the observed d_N/d_S ratios for both nonsense and missense mutations are very similar to that observed in SARS-CoV-2, another pathogen present in the nasopharynx which we describe in more detail in Chapter 9.

In addition to estimates of the global d_N/d_S ratio, the large number of samples in our dataset allows for the investigation of within-host d_N/d_S ratios at the gene level. Similar to the global estimates we made use of dNdScv to estimate ratios for each gene in each of the top 10 GPSCs and conducted a likelihood ratio test with the corresponding neutral model to identify genes that had a significant signal of selection. The resulting significant d_N/d_S ratios for missense mutations after correcting for multiple tests are given in Figure 60. Interestingly this identified a number of genes that have been linked to the virulence of *S. pneumoniae* including *pcpA* and *lytA* [253]. This is consistent with a previous study of selection in the broader *Streptococcus* genus which similarly identified an association between selection and virulence [10]. *LytA* is conserved across a diverse range of strains suggesting that similar selective forces may act on this gene across different timescales [506]. *PspA* is known to diversify frequently through recombination and it has been suggested that this may reflect positive selection for evading adaptive immune responses which may then be subject to negative selection if the alteration reduces the functionality of the protein [96].

We also identified a large number of transposable elements. Short-term pneumococcal evolution is characterised by the movement of phage and intragenomic rearrangements [98]. The identification of these elements is thus consistent with selection acting over a short-time scale. However, the high mobility of these elements could also result in recombination events leading to biased estimates of the d_N/d_S ratios [60]. Consequently, further confirmatory analyses possibly making use of long read data is needed to ensure the identified mutations were unbiased by recombination.

Finally, we identified significant diversifying selection in the *cglC* gene in 4 of the 10 GPSCs. *CglC* is known to play a role in pneumococcal competence with laboratory experiments demonstrating that the absence of the *cgl* locus led to the complete loss of transformability [399].

Transformation is known to result in a tradeoff between fitness benefits and costs. Benefits include the ability to acquire drug resistance genes, evade the immune system through allelic replacement and repair damaged DNA [408]. However, transformation has also been

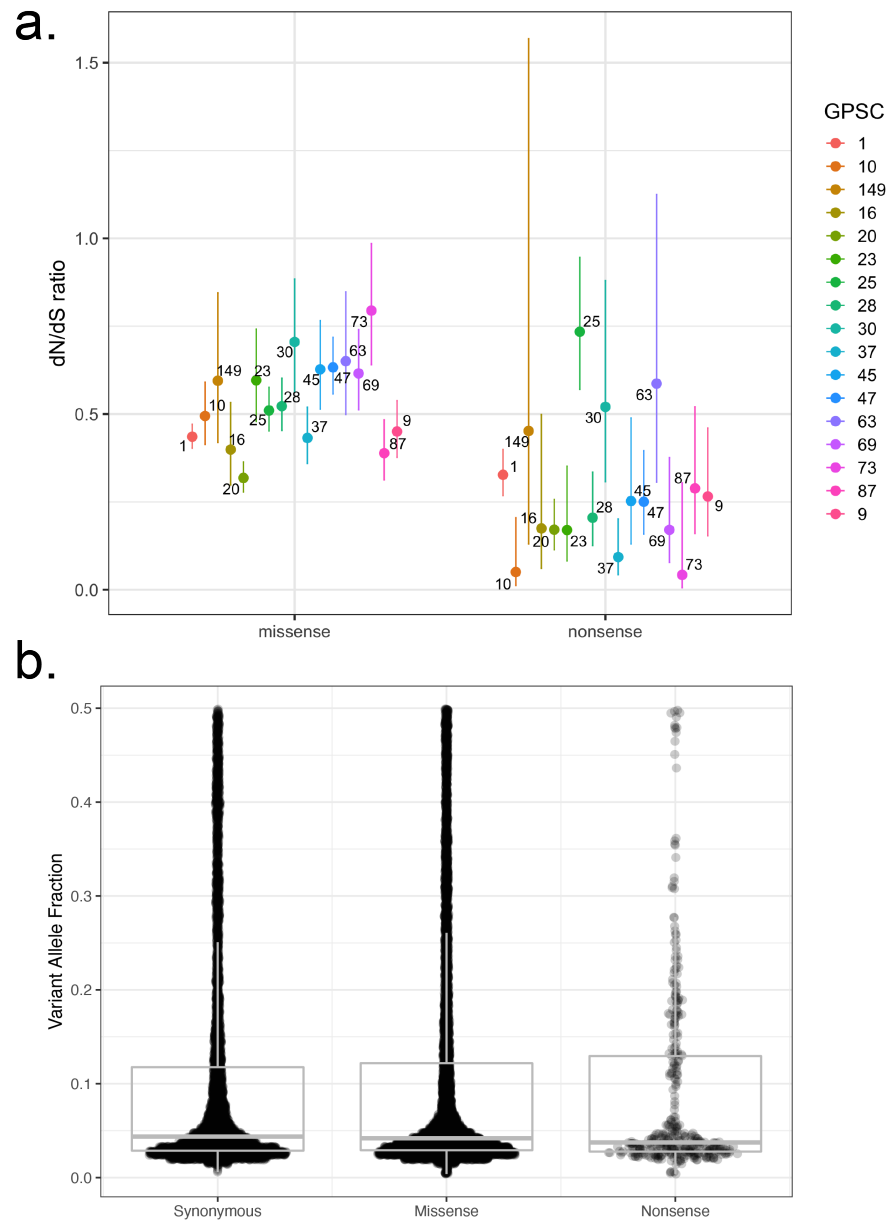


Figure 59: a.) The estimate d_N/d_S ratios found in minority variants using dNdScv split by missense and nonsense mutations. The algorithm was run separately for each GPSC. b.) The inferred variant allele frequencies of minority variants split by whether they caused synonymous, missense or nonsense mutations. Each variant is plotted as a transparent point with the far fewer number of nonsense mutations leading to the lighter colour in the final boxplot.

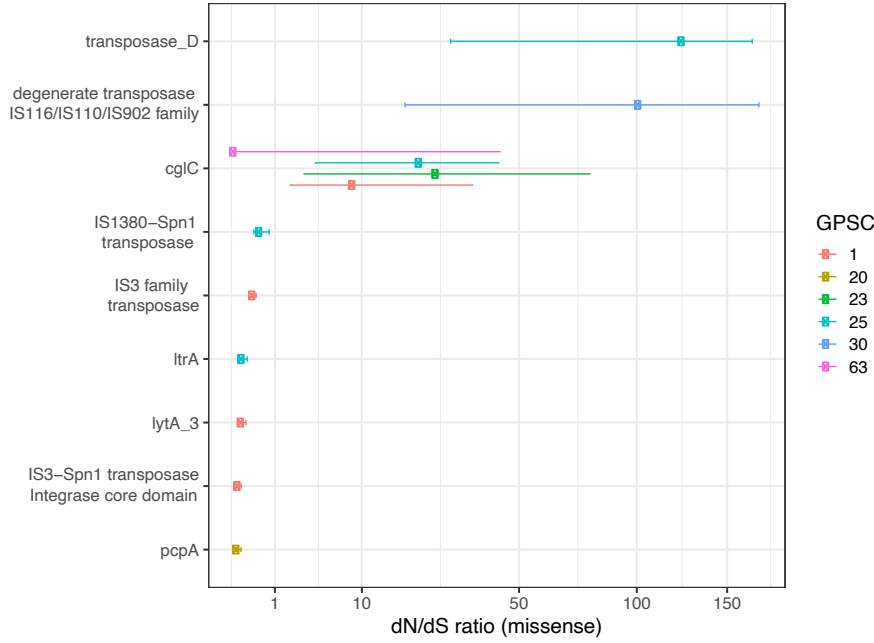


Figure 60: Estimated missense d_N/d_S ratios for those genes found to be under significant selection using the default model in dNdScv. The algorithm was run for separately for each GPSC.

shown to lead to transient growth arrest, the disruption of beneficial alleles and fratricidal killing of non-competent cells [213]. As a result, maintaining variable levels of transformation within the population of pneumococcal cells in an infection may allow the pathogen to better adapt to the presence of antimicrobial, antibodies or competing infections with the fitness costs of this adaptability preventing these cells from sweeping to fixation [399].

8.4 DISCUSSION

The additional information provided by deep sequencing such a large collection of whole plate sweeps from pneumococcal positive nasopharyngeal swabs has provided one of the clearest pictures of the dynamics of pneumococcal evolution both within the host and in transmission.

We have demonstrated that the approach can be used to improve our ability to distinguish different transmission scenarios using genomic data. This allowed us to confirm the previous finding that close transmission links have a strong geographic structure. The added resolution of the method also allowed for the characterisation of the typical transmission direction between mothers and their children. Our finding that mothers are more likely to infect infants in their first year of life could have important implications for the design of intervention strategies such as when and to whom vaccines should be

administered to as well as whether antibiotics could be used to preemptively reduce the likelihood of invasive disease. Looking at the number of shared minority variants between hosts identified as close transmission links indicates that consistent with previous studies, the transmission bottleneck is likely to be small. However, it is unlikely to be restricted to only a single bacterium as has been previously postulated [271].

By focusing on minority variants within pneumococcal infections consisting of a single lineage we were able to shed light on the selection and mutational processes driving pneumococcal evolution within the host. We found a distinct mutational spectrum of within-host minority variants when compared to that seen over longer timescales which was consistent with previous laboratory studies of minority variants in *E. coli* populations. This coupled with identification of diversifying selection acting on genes linked to pneumococcal competence suggests that consistent variation generated by mutational processes within the host may allow the pneumococcus to maintain the ability to quickly react to external stressors including the host immune response and antimicrobial treatment. This is supported by our previous GWAS analysis of the impact of antimicrobial treatment on within-host variation in [Chapter 7](#).

In this part of the thesis we have shown that the deep sequencing of whole plate sweeps of cultured pneumococcal samples can improve our ability to accurately identify cases of multiple carriage as well as pneumococcal resistance and virulence factors. We have further demonstrated that accurately characterising within-host variation can improve transmission inference efforts and can provide insights into the mutational and selective forces driving the evolution of *S. pneumoniae* within the host. With the ever decreasing cost of sequencing and as many of the laboratory steps required to culture and sequence whole plate sweeps of pneumococcal samples are identical to those required for single colony picks, this part of the thesis presents a compelling case for the routine use of deep sequencing whole plate sweeps in the surveillance of *S. pneumoniae*.

Part III

SARS-COV-2

Part III describes the analysis of the genetic diversity of SARS-CoV-2 within the host, conducted in response to the global SARS-CoV-2 pandemic.

TRANSMISSION AND WITHIN-HOST EVOLUTION OF SARS-COV-2

Declaration of contributions

This chapter involved the combined efforts of many people from the SARS-CoV-2 genomics teams based at the Sanger Institute and the University of Cambridge. In particular, Dominic Kwiatkowski supervised the analysis of within-host diversity in SARS-CoV-2 infections. Estée Török and Will Hamilton supervised the analysis of SARS-CoV-2 transmission in care homes. The analysis of transmission in care homes was performed in collaboration with Will Hamilton. Inigo Martincorena performed the analysis of the mutational spectrum and selection within SARS-CoV-2 infections which is described briefly in this chapter for completeness. Andrew Lawson and Roberto Amato called the within-host minority variants.

9.1 BACKGROUND

The SARS-CoV-2 pandemic has caused global disruption and more than one million deaths in 2020 [520]. Originating in Wuhan, Hubei province, China in late 2019, the virus is the seventh coronavirus known to infect people after having jumped from an animal reservoir with similar coronaviruses isolated from both bats and pangolins [7].

The majority of SARS-CoV-2 infections are mild, with only a fraction of people developing more severe symptoms including life threatening pneumonia [525]. Severe symptoms of coronavirus disease 19 (COVID-19) are disproportionately seen in the elderly and people with pre-existing conditions such as cardiovascular disease and diabetes [74, 137]. This has led to considerable heterogeneity in the infection fatality ratio between regions, ranging from 0.24% in Slovenia to 1.88% in New York, depending on the region-specific demographics of those infected [137].

Similar to other beta-coronaviruses, SARS-CoV-2 is an enveloped virus with a large positive-sense single-stranded RNA genome, nearly 30kb in length [522]. After entering a host cell, the viral genomic RNA is translated to produce nonstructural proteins from the two open reading frames, ORF1a and ORF1b [261]. The viral genome is used as a template to generate negative-sense RNA intermediates templates which are used to synthesise positive-sense genomic RNA (gRNA) and subgenomic RNAs (sgRNAs). While gRNA are packaged to as-

semble progeny virions, sgRNAs encode conserved proteins including: the spike protein (S), envelope protein (E), membrane protein (M) and nucleocapsid protein (N) as well as a number of accessory proteins including 3a, 6, 7a, 7b, 8, and 10 [261]. The spike protein (S), mediates entry into host cells and interacts with host receptors such as angiotensin-converting enzyme 2 (ACE2). As a result, this protein has been a major target for many vaccine efforts [402].

Aside from vaccine design, the initial sequencing of the SARS-CoV-2 genome in early 2020, has yielded important insights into the origins and spread of the pandemic [522]. Routine sequencing has become part of the public health response in a number of countries including Australia, New Zealand and the UK where over 100,000 viral genomes were sequenced by the 10th October 2020 by the COVID-19 Genomics UK Consortium (COG-UK) [53, 186, 444]. Genome sequencing has also provided insights into the adaptation of the virus to its human host with recent studies indicating the D614G mutation which occurred in the early stages of the pandemic may have increased the transmissibility of the pathogen [402, 497].

Tracking the transmission and spread of the virus is the major motivation behind the high level of sequencing being performed by multiple public health bodies. In this chapter we focus on two distinct but related questions on the use of genomics to track SARS-CoV-2. Firstly, can we use genomics to gain a better understanding of transmission into and within high risk groups such as elderly care homes? Secondly, how does the virus evolve within the human host and can within-host diversity be used to improve transmission inference efforts?

9.1.1 SARS-CoV-2 transmission in care homes

The disproportionately high mortality seen in the elderly and people with preexisting conditions means that elderly care homes are particularly vulnerable to outbreaks of COVID-19. As of the end of June 2020, the United Kingdom (UK) Office for National Statistics (ONS) estimated that 30.2% (13,417 deaths) of all COVID-19 deaths in England occurred in care homes [384]. According to the ONS, there were an estimated 29,393 excess deaths in care homes between 28th December 2019 and the 12th June 2020 indicating that the true number of care home COVID-19 deaths may be significantly higher. This discrepancy is likely driven by COVID-19 testing limitations during the early stages of the outbreak in the UK [385]. Consequently, obtaining a strong understanding of the transmission of SARS-CoV-2 into and around care homes is of particular importance. In particular, identifying whether SARS-CoV-2 typically spreads between care home residents within a home or is the result of multiple introductions can be crucial in deciding appropriate mitigation strategies. Similarly, under-

standing the role of healthcare workers in transmission could lead to improved transmission control procedures.

9.1.2 *Within-host diversity of SARS-CoV-2*

Genomic epidemiology typically relies on obtaining a single representative genome per infected individual. Whilst this can provide valuable information in the tracking of infectious disease, inconsistencies between the phylogeny of the sampled genomes and the underlying transmission chain can complicate transmission inference efforts. By obtaining additional information on the within-host diversity of a pathogen we can generate improved estimates of transmission. A more thorough description of the difficulties and benefits of within-host data is given in [Chapter 4](#). In addition to potentially improving estimates of transmission chains, analysis of the within-host mutational spectrum of the virus can improve our understanding of the origins of mutations and the dominant mutational mechanisms driving the evolution of the virus.

9.1.3 *Data*

To answer both of these questions we collected two distinct datasets from across the East of England. We identified 1,167 positive COVID-19 samples between 26th February and 10th May 2020 from the residents of 337 care homes of which 700 (60%) had been sent for genome sequencing and returned a reliable consensus sequence. These data were coupled with genomes obtained from tests conducted on health care workers within Cambridge University Hospitals (CUH) to allow for the investigation of transmission relationships between health care workers and care home residents.

To investigate the within-host genomic diversity of SARS-CoV-2 we obtained genomic data from 1,181 SARS-CoV-2 positive samples collected in March and April 2020, comprising 1,123 infected individuals of whom 49 had multiple samples collected at different timepoints. To gain insights into within-host diversity and to allow for the investigation of the reproducibility of within-host variant calls, two libraries were sequenced for each sample with separate reverse transcription (RT), PCR amplification and library preparation steps. To develop reliable methods for analysing within-host variation in the context of ongoing genomic epidemiological studies, we used the ARTIC protocol which is a common method used for SARS-CoV-2 genome sequencing by many labs around the world [400].

Using these data, the following sections investigate the major questions surrounding the transmission of SARS-CoV-2 within care homes and whether information on within-host diversity can provide addi-

tional resolution for transmission inference as well as insights into the mechanisms of SARS-CoV-2 evolution.

9.2 METHODS

9.2.1 *Variant calling*

To identify both within-host and consensus variants we made use of the ShearwaterML algorithm [187]. Originally developed for the analysis of deep sequencing data in cancer studies, the algorithm uses a Beta Binomial model to account for sequencing errors. For each site, a null model of the probability of sequencing errors is learned from a set of control samples. The calling of variants is then formulated as a model selection problem through the use of a likelihood ratio test. In cancer studies, the null model is usually fit to a collection of alternative non-cancer (normal) samples under the assumption that each variant is likely to be rare in the population.

As we do not have a similar set of reference samples from which to learn the null model, we instead fit the null model to a subset of 50 replicates and set sites with a non-reference allele frequency of $\geq 1\%$ as uninformative. The resulting ShearwaterML calls were merged if consecutive calls differed in VAFs ≤ 0.05 and were filtered to ensure: a total depth of $\geq 100\times$; a BH adjusted p-value of ≤ 0.05 in one sample and an unadjusted p-value ≤ 0.01 in the replicate sample. This is a conservative approach which allowed us to be more confident that the resulting variants were not the result of sequencing artefacts. However, the ShearwaterML pipeline provided similar overall results to more standard and less conservative methods such as bcftools [299]. In the analysis of care home data, the variants were obtained directly from the COG-UK assemblies [53].

9.2.2 *Mutational spectrum and selection*

To investigate selection and the mutational spectrum within the host we made use of similar approaches to that used in the analysis of within-host diversity in *S. pneumoniae* in Chapter 8. We adopted the Wuhan-Hu-1 reference genome as the ancestral state with which to call mutations. As the occurrence of the same mutation across different samples may not be independent we counted mutations only once when considering the consensus genomes and up to five times when considering within-host variants. This avoided a small number of highly-recurrent hotspots from dominating the spectrum. Similar to our analysis of the pneumococcus we calculated the d_N/d_S ratios using the dNdScv R package with the default substitution model comprising 192 rate parameters.

9.2.3 *Consensus phylogeny construction*

Consensus genomes for each replicate were generated using the ARTIC SARS-CoV-2 bioinformatics pipeline. A multiple sequence alignment was generated using MAFFT v7.464 and any sites that were discordant between replicates were set to be ambiguous [257]. Sites that have previously been identified to create difficulties in generating phylogenies were masked using bedtools v2.29.2 using the VCF described in De Maio et al, 2020 [113, 412]. Fasttree v2.1.11 was used to generate a maximum likelihood phylogeny [406].

9.2.4 *Transmission model*

To investigate transmission, samples were only considered if both replicates produced high quality consensus genomes. When multiple samples from the same host were available the earliest sample was used. Pairwise SNP distances were generated between the consensus genomes using pairsnp v0.2.0 [483]. The distribution of the underlying number of intermediate transmission events between each pair of samples was then inferred using an implementation of the TransCluster algorithm as described in Chapter 4. The serial interval and evolutionary rate were set to 5 days and $1e-3$ substitutions/site/year [162, 538].

9.2.5 *Identification of potential mixed infections*

Potential mixed infections were identified using a linear model by testing whether the allele frequencies in a sample could be better explained by the inclusion of an additional consensus genome from the COG-UK dataset of the 29th May 2020. Additional samples mixtures were considered if the addition of a COG-UK consensus genome could explain at least 2 iSNVs and have at most 1 variant that was not found in the alleles of the sample. This identified 54 putative mixtures which were then screened manually to obtain 36 potentially mixed samples.

9.3 RESULTS

9.3.1 *Transmission in care homes*

The genomic data considered here represents a subset of a larger collection of samples taken from 1,167 care home residents from 337 care homes. This collection included 193/337 residential homes and 144/337 nursing homes representing approximately half of the care homes in the East of England that reported confirmed COVID-19 outbreaks to PHE by the 11th of May 2020 [409].

Across this collection, there was a median of two COVID-19 cases per care home (IQR: 1-5, range: 1-22), with the top 10 care homes containing 164/1167 of the cases. This dataset was collected prior to the implementation of routine asymptomatic testing in care homes. As a result, the number of positive cases likely represents a significant underestimate due to the high rate of asymptomatic cases of SARS-CoV-2 infection [361]. The seriousness of the situation in care homes is highlighted by the difference in mortality between care home residents and non-care home residents at the Cambridge University Hospitals. Of the 72/464 care home patients admitted to CUH, 34/72 (47.2%) died within 30 days of a positive COVID-19 test. This compared with 78/392 (19.9%) of non-care home residents. The difference in mortality between care home and non-care home patients was significant even after controlling for age and whether or not patients were admitted to the intensive care unit. This highlights the gravity of the situation in many care homes and demonstrates the need for a thorough understanding of the transmission of the virus into and within care homes. To investigate the patterns of transmission within care homes we considered a subset of 700 samples representing 292 care homes, collected from residents between 26th February and 10th May 2020 for which high quality genomic and epidemiological data was available. In order to determine cases that could be plausibly linked by recent transmission events we made use of the implementation of the TransCluster algorithm described in Chapter 4. This enabled us to account for the genetic distance between viral genomes as well as the time difference between samples. To concentrate on the most likely transmission links we used a pragmatic cut-off of at least a 15% probability that any two samples were connected by at most 2 intermediate hosts. This corresponded with a maximum of 4 SNP differences between any pair of genomes and a maximum of 22 days separating the sample dates (Figure 61). As we are not considering any community samples outside of care homes, we did not attempt to infer possible transmission events between care homes. Without the strong prior assumption that people living within the same care home are more likely to transmit to one and other, the low viral diversity within the UK during this period makes inferring close transmission links difficult [403].

Within the 292 care homes, we identified 409 distinct transmission clusters with a median of one cluster per care home and a maximum of four. This can be compared with the 316 clusters that would occur if we ignored the genomic data, equivalently treating all the genomes as identical. The increase in the number of identified clusters highlights the value of genomics in distinguishing separate introduction events. The distribution of genetic and temporal distances between each pairwise comparison within a cluster is shown in Figure 61 indicating 673/775 (86.8%) of pairwise links had zero or 1 pairwise

SNP differences, and 756/775 (97.5%) were sampled < 14 days apart. Transmission links for the top ten care homes with the most sampled genomes are shown in [Figure 62](#), indicating that infections in some care homes could be explained by a single transmission cluster (e.g. CARE0314), while others comprised multiple introduction events (e.g. CARE0061). This highlights the importance of both preventing the introduction of the virus into care homes as well as implementing effective infection control procedures within care homes to prevent the onwards transmission of the disease.

9.3.2 *Healthcare workers*

While it was not possible to consider transmission links between care homes, detailed epidemiological information on the location and movement of healthcare workers within CUH allowed us to consider possible transmission links between care home residents and healthcare staff. Similar to the previous analysis we considered each care home independently. However, we allowed healthcare workers to belong to clusters from multiple care homes to account for the possibility of multiple interactions between a healthcare worker and different care homes. Using the same threshold of at least a 15% probability of a transmission event within 2 intermediate hosts we identified 38/54 care home residents with possible links to healthcare workers ([Figure 63](#)). Of these, there was strong epidemiological supporting evidence for 14/38 or 26% of the total care home residents for which healthcare worker data was available. These involved the care homes CARE0063 and CARE0114 ([Figure 63](#)). The cluster found in CARE0063 was described in a previous publication and included a carer at the home, a carer from a different care home as well as paramedics and a number of family members of the healthcare workers [354]. The cluster at CARE0114 had not been described previously and included acute medical staff from CUH, including one staff member who cared for at least one of the residents of the care home. As described in [Chapter 4](#), the TransCluster pairwise transmission algorithm does not assign a direction to transmission events and can not exclude the possibility of both infections being the result of separate transmissions from a third unsampled host ([Figure 26c](#)). As such, the number of links between healthcare workers and care home residents may be inflated. However, the strong epidemiological evidence for a number of interactions suggests that some transmission between healthcare workers and care homes is likely to have occurred. Interestingly, an additional link was identified between a third care home (CARE0273) and the same paramedics involved in the CARE0063 cluster. Consecutive pairwise transmission links could be made between these care homes within zero SNPs and three days. These care homes are located within 1km of each other suggesting that there

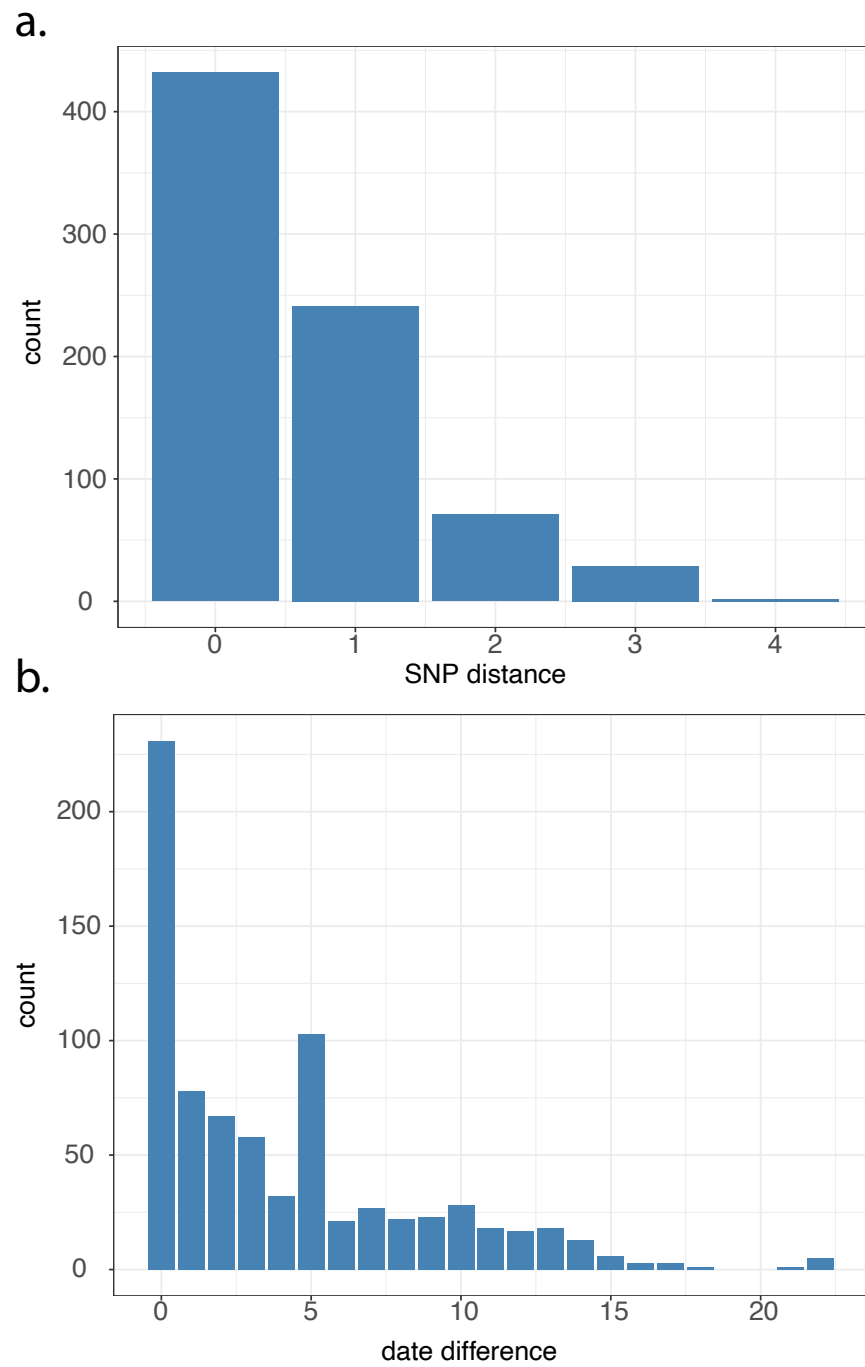


Figure 61: The distribution of pairwise SNP (a) and date (b) distances between those transmission links that were found to have at least a 15% probability of being within two intermediate hosts using the TransCluster algorithm

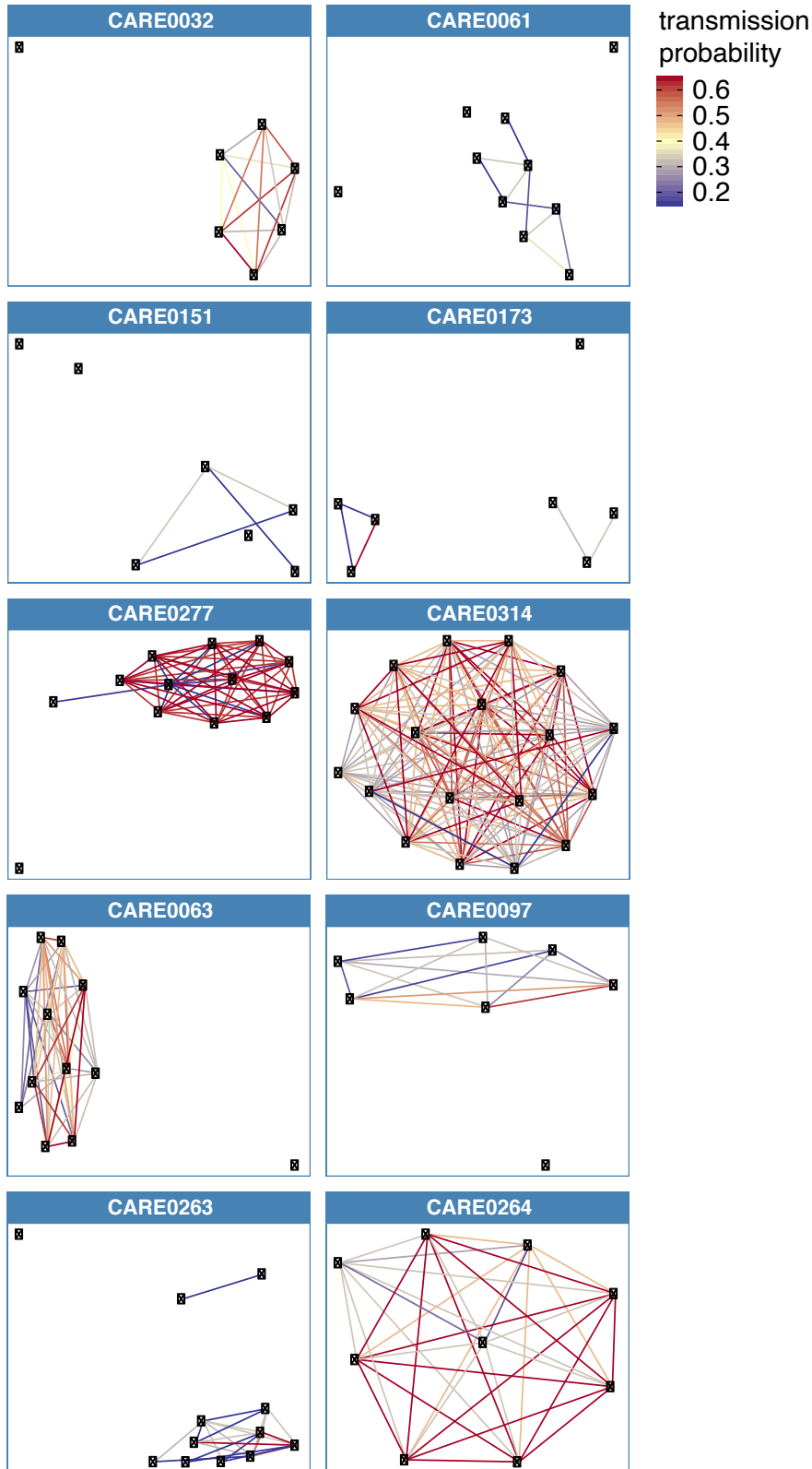


Figure 62: Transmission networks generated using the TransCluster algorithm including links that were found to have at least a 15% probability of being within two intermediate hosts. These networks may not necessarily reflect the true transmission network, and hospital-acquired infections for some of the clusters is a possibility alongside multiple introductions into the same care homes.

could plausibly have been transmission between homes, possibly facilitated by healthcare staff. While this is speculative, it highlights the struggle healthcare workers face in preventing transmission and the difficulties posed by a highly connected healthcare system.

9.3.3 *Within-host diversity*

Our analysis of care home data has highlighted the value of WGS in distinguishing different transmission scenarios, notably by enabling us to identify multiple introductions into care homes in addition to links between healthcare workers and care home outbreaks. As outlined in [Chapter 4](#), although single genomes can be informative for transmission inference, a phylogeny built from single genome samples from each host can often be explained by multiple transmission scenarios ([Figure 25](#)). To determine whether additional information on the within-host diversity of SARS-CoV-2 could improve transmission inference efforts and to gain a better understanding of the evolution of SARS-CoV-2 within the host we generated high depth sequencing data for 1,181 samples from the East of England. These samples were distinct from those used in the analysis of care home transmission and involved 1,123 infected individuals of whom 49 had multiple samples collected.

To identify reliable within-host mutations and to determine the reproducibility of using amplicon sequencing to investigate within-host diversity we sequenced all 1,181 samples in duplicate. Two libraries were sequenced for each sample with separate reverse transcription (RT), PCR amplification and library preparation steps. To develop reliable methods for analysing within-host variation in the context of ongoing genomic epidemiological studies, we used the ARTIC protocol which is a common method used for SARS-CoV-2 genome sequencing by many labs around the world [[400](#), [411](#)].

Variants were called using the ShearWaterML algorithm using a site-specific error model learned from a collection of reference samples as described above. To investigate the concordance between replicates we compared all sites with a variant allele frequency (VAF) $> 0.5\%$ supported by at least 5 reads. While the majority of samples showed high concordance between replicates others were discordant ([Figure 64](#)). This could be driven by uneven amplification of a small number of RNA molecules with either rare within-host mutations or RT/PCR errors. Using a beta-binomial, we compared the diagnostic Ct values with the discordance between replicates which identified that samples with $Ct \geq 24$ showed high levels of discordance but that in general samples with lower Ct values had good concordance between replicates. This suggests that lower viral loads can compound amplification biases and other sequencing artefacts which should be thought through when using RT-PCR protocols to study within-host

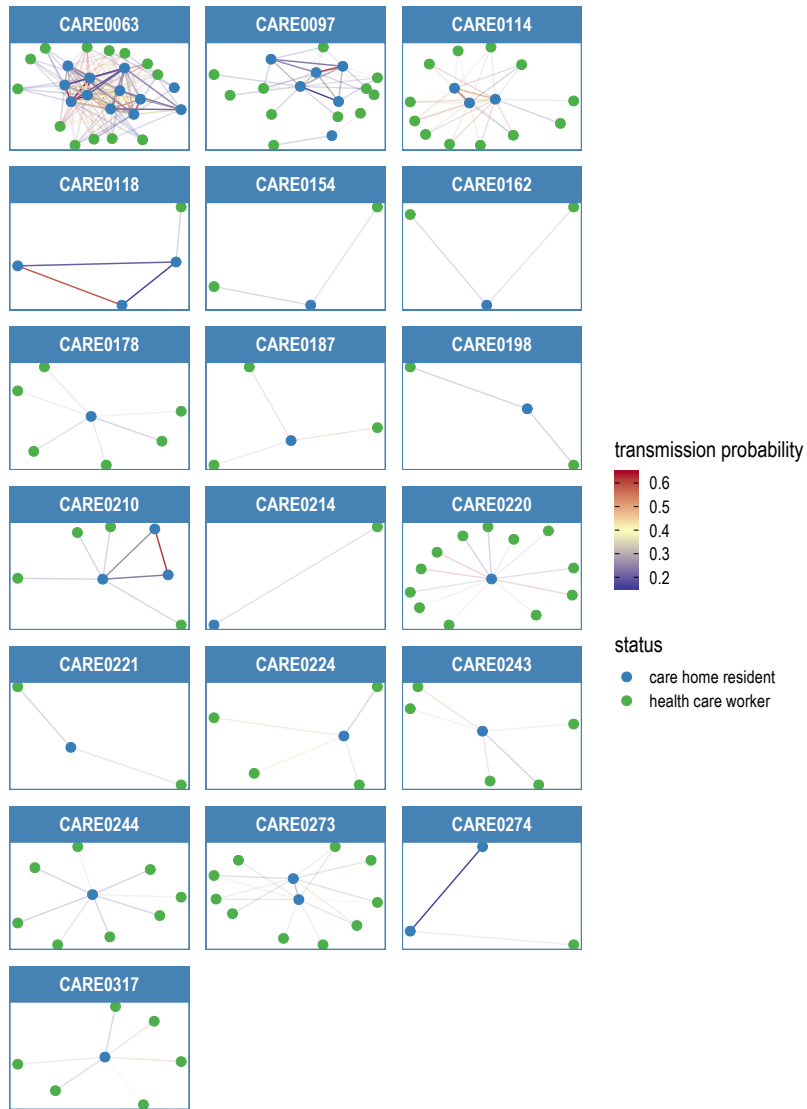


Figure 63: Transmission networks including healthcare workers generated using the TransCluster algorithm including links that were found to have at least a 15% probability of being within two intermediate hosts. Healthcare workers may be present in the networks of multiple care homes.

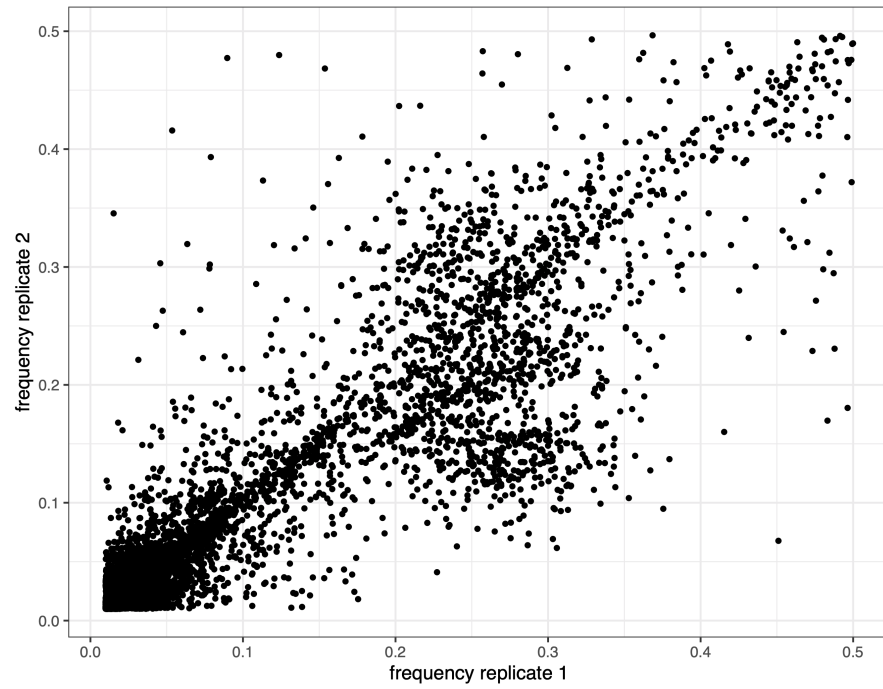


Figure 64: Dot plot indicating the concordance between variant allele frequency estimates across sequencing replicates. Each point is a single variant called in an individual sample.

diversity. To alleviate many of these issues we only consider variants found in both replicates for the remaining analyses.

In total, we identified 18,888 putative variants, including 7,190 consensus variants ($\text{VAF} \geq 95\%$, affecting 948 sites) and 11,698 within-host variants (affecting 5,625 sites). Within-host variants included 7,561 single-nucleotide substitutions, 3,235 small deletions, 548 small insertions and 354 putative multi-nucleotide variants. Minority variants were detected in the majority of samples (95.2%), with a median of 8 within-host variants per sample at a mean allele frequency of 7.2% (median 1.9%, minimum 0.05%). This is likely to be an underestimate as in some cases the absence of within-host variants will be due to technical issues driven by low sequencing depth in one or both of the replicates. However, the results were inline with a similar study on the within-host diversity of SARS-CoV-2 [320].

9.3.4 *Mutational spectrum*

To gain a better understanding of the dominant mutational mechanisms driving SARS-CoV-2 evolution we examined the mutational spectrum of both within-host mutations and ‘between’ host mutations which included those variants above 95% frequency which differed from the Wuhan-1 reference genome (Figure 65). The similarity between within-host mutations and ‘between’ host mutations indi-

cates the similar processes may be driving the spectrum in both cases and provides extra confidence to our within-host mutation calls. The frequencies given in [Figure 65](#) indicate the observed changes after mapping to the plus strand reference genome. Consequently, the excess of C>U mutations over G>A indicates an asymmetry in the mutational profile between the plus and minus strands of the virus. In addition to this asymmetry we see an excess of C>U and G>U mutations (shown as C>A in [Figure 65](#)).

We would not expect to see these asymmetries as a result of replication as each RNA molecule has been through the same number of replication steps between plus and minus strands using the same RNA-dependent RNA-polymerase. An alternative explanation is that RNA editing or damage occurs at a similar rate at all times in molecules of both strands. The higher molarity of the plus strand would thus provide more time over which mutations could occur leading to the asymmetry in the mutational spectrum [437]. Mutational processes that could explain the dominance of C>U over G>A and G>U over C>A include spontaneous cytosine deamination and guanine oxidation respectively which have both been observed as common mutagenic processes in human cells [216, 277]. Previous efforts to look at the mutational mechanisms in SARS-CoV-2 have suggested that two families of RNA-editing enzymes may contribute significantly: APOBEC cytosine deaminase enzymes (C>U); and ADAR enzymes causing adenosine-to-inosine changes (A>G) [124]. However, while we see a high rate of C>U changes we do not see a high frequency of A>G mutations as previously indicated [450]. In human cancers, mutations driven by APOBEC enzymes are characterised by strong sequence context leading to changes occurring predominantly at TpC sites, which we do not observe ([Figure 65](#)) [216]. Thus while we can not rule out a role for these enzymes in the mutational mechanisms behind SARS-CoV-2 evolution, our data is more consistent with spontaneous cytosine deamination and guanine oxidation.

9.3.5 *Within-host mutational load and selection*

As the within-host variant allele frequency provides an estimate of the number of viral RNA molecules that carry a mutation within a sample, we can estimate the mean number of within-host mutation per viral genome. This can be achieved by summing the allele frequencies of all within-host variants within a sample [332]. We estimated a mean mutation load over all samples of 0.72 (median 0.37) ([Figure 65d](#)). This is an approximation as it assumes that each RNA molecule consists of the entire viral genome which is not always the case given the presence of RNA fragments including sgRNAs [397, 511]. It also fails to account for mutations that fall below our quality control thresholds and thus likely represents an underestimate of

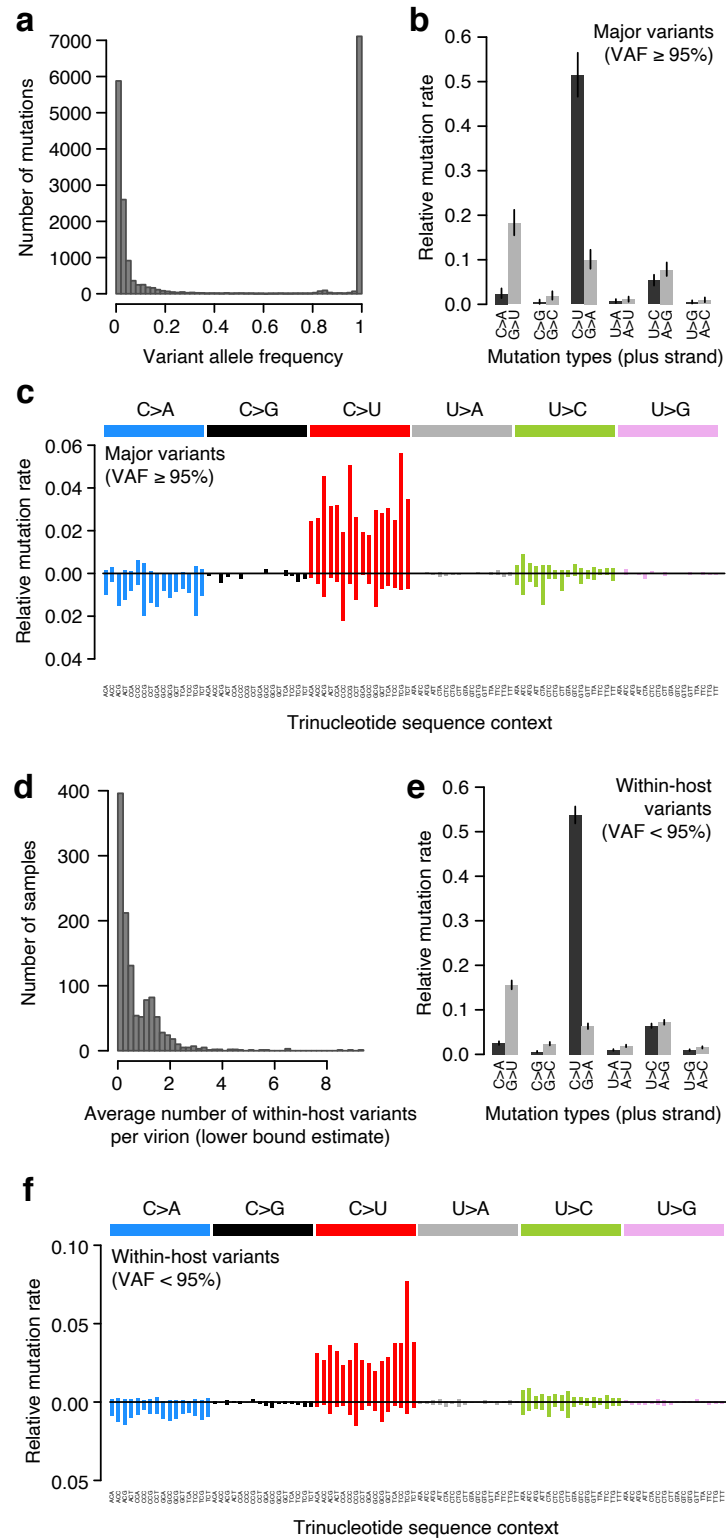


Figure 65: (a) Histogram of the VAFs of all variant calls. (b, e) Mutational spectra of major (b) and within-host (e) variants, as mapped to the reference strand indicating mutational strand asymmetries. Error bars depict Poisson 95% confidence intervals. (d) Histogram of the lower bound estimates of the mean number of within-host variants per viral genome per sample. (c, f) Mutational spectra of major (c) and within-host (f) variants, as in [5]. Mutations are represented as mapped to the pyrimidine base, depicted above the horizontal line if the pyrimidine base is in the reference (plus) strand and below it if the pyrimidine base is in the minus strand.

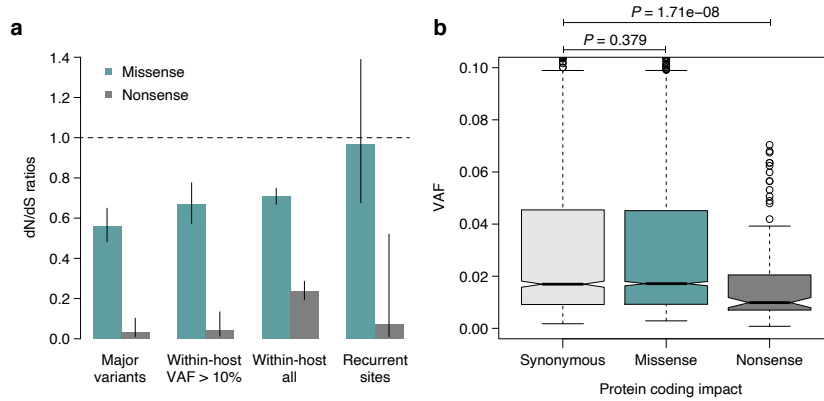


Figure 66: (a) Genome-wide d_N/d_S ratios for missense and nonsense mutations (Methods). Error bars depict 95% confidence intervals from the Poisson maximum-likelihood model. (b) VAFs of within-host variants as a function of their predicted coding impact. P-values were calculated with Wilcoxon tests.

the true value. Given an estimated evolutionary rate in SARS-CoV-2 of ~ 0.001 mutations/bp/year [162], or ~ 0.082 mutations/genome/day, this would suggest that the estimated mutation load could be acquired relatively quickly over a few days.

To investigate potential differences in selection acting at the within and between host level we used dNdScv, a tool designed to estimate d_N/d_S ratios for populations that undergo little or no recombination [331]. The d_N/d_S ratios for dominant variants (VAF > 95%) and moderate variants (VAF > 10%) were under strong purifying selection, while within-host variants overall exhibited weaker purifying selection (Figure 66). This is consistent with the selection observed at the consensus sequence level being driven by within-host selection rather than during transmission.

9.3.6 Longitudinal samples of the same host

To investigate the possible accumulation of *de novo* mutations during the course of an infection, we studied 43 individuals for whom we had multiple samples collected at different timepoints (Figure 67a). Overall the number of within-host variants tended to increase over time and this trend was significant using a Poisson mixed model to control for host specific effects ($p=0.007$; Figure 67b). To put this finding in context, there was considerable variation in the observed number of within-host variants among samples from the same individual, even if taken on the same day, possibly as a result of artefactual bottlenecks caused by the different sampling methods which included sputum, swabs and bronchoalveolar lavage. High variability between longitudinal samples has also been observed in six out of nine hospitalised patients in Austria [404]. Here, the authors found using data

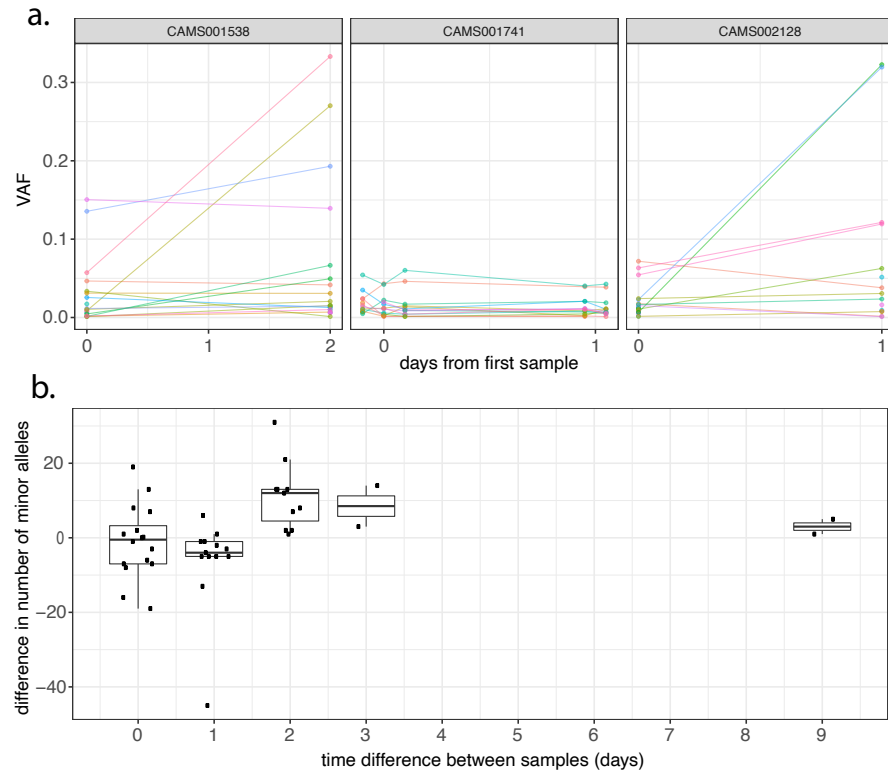


Figure 67: (a) Frequencies of within-host variants for three selected hosts where multiple samples were taken over consecutive days. Samples taken on the same day have been offset by a small distance. (b) The difference in the number of within-host variants between pairwise combinations of samples taken from the same host. The order for samples taken on the same day was randomised

from nine patients with longitudinal samples that the divergence in variant frequencies between serial samples from the same patient was on average greater than that between donor recipient transmission pairs. The authors observed multiple instances of the fixation of a variant over the course of an infection. In contrast, we observed no change in the consensus genome sequence in any of the individuals from whom multiple samples were collected. It is possible that the accumulation of within-host mutations could in future be used to estimate time since infection, but validation of this approach would require analysis of time-series data in a larger number of individuals with high quality epidemiological data.

9.3.7 Shared within-host variants

If the population bottleneck in SARS-CoV-2 transmission is large, identifying shared within-host variants between samples could potentially inform transmission inference efforts. This was shown in [Chapter 4](#), where the SVC algorithm was found to more accurately

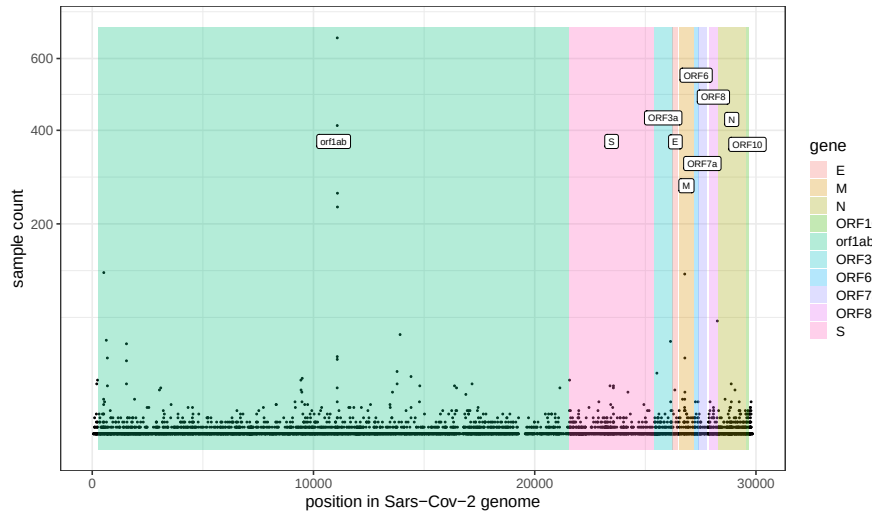


Figure 68: dot plot indicating the number of samples that have a within-host variant for each site in the SARS-CoV-2 genome. Colours represent the different open reading frames within the genome.

identify transmission chains under a loose transmission bottleneck. Recurrent mutations have been identified in both the global consensus phylogeny of SARS-CoV-2 genomes as well as in previous studies of within-host variation [113, 320]. However, while shared variants could inform estimates of transmission there are a number of potentially confounding factors including: sequencing errors, contamination between samples, coinfection of the same sample with multiple lineages, recombination, mutational hotspots, and positive selection. It is vital that the source of shared within-host variants are well understood prior to their use in transmission inference to avoid potentially erroneous conclusions which could potentially impact on public health decision making.

Across our dataset 196 mutations are observed as within-host variants in samples from at least 5 hosts (0.45%) (Figure 68). Due to the use of replicates and the ShearWaterML variant calling algorithm which models the error profile of each site, these recurrent mutations are unlikely to be the result of sequencing or PCR errors.

In order to investigate the genetic background of our samples we generated a maximum likelihood phylogeny of all consensus genomes produced by the COG-UK consortium by the end of May 2020, including those for the samples on our dataset (Figure 69a). This showed that our samples represent a broad range of the SARS-CoV-2 genetic diversity found in the UK at that time, and that the diversity observed among our samples was not primarily driven by geographical location. To explore the relationship between within-host variants and the consensus phylogeny, we identified within-host variants that are shared between samples, as illustrated by links between the tips of the phylogenetic tree in Figure 69b. This confirmed that within-host

variants are often shared between samples that are distant on the consensus phylogeny. Both a high level of recombination and a large transmission bottleneck would be required for within-host variants to be maintained across long transmission chains in order to explain the simultaneous preservation of some within-host variants with the fixation of others. The sharing of within-host variants between consensus genomes as divergent as 10 SNPs, indicating multiple months of separation between the samples suggests that a more likely explanation is that these shared within-host variants are the result of recurrent mutation or co-infection.

To investigate the correspondence between shared within-host variants and transmission, we used the TransCluster algorithm to infer the probability distribution of the number of intermediate hosts that separate each pair of samples[468]. This pairwise approach accounts for the serial interval of the virus as well as its evolutionary rate using the difference in the number of SNPs between each pair of consensus genomes. Figure 70a illustrates the relationship between the mean number of shared within-host variants and the probability that the respective infections were a result of direct transmission. A weak correlation with transmission probability was found for shared within-host variants of high allele frequency, but not for those of low allele frequency.

Similar to our use of the TransCluster algorithm we can investigate the background genotypes of each within-host variant in more detail. If shared within-host variant were the result of transmission we would expect them to occur within similar genomic backgrounds. In the absence of significant recombination a shared within-host variant is incompatible with the simultaneous fixation of additional variants as the within-host allele would either be purged or hitchhike and become fixed. Figure 71, indicates a selection of three notable within-host variants along with the VAF of all other non-reference calls. Similar to our initial transmission analysis, the heatmaps suggest a complex genomic background for these variants that is more consistent with independent recurrent mutation rather than the transmission of minority variants.

Recurrent within-host mutations could be driven by both mutational hotspots or convergent positive selection. As we observed that 25% (51/205) of recurrent ($n \geq 5$) within-host variants are predicted to cause synonymous mutations and as our estimate of the d_N/d_S ratio for within-host variants indicated that they are subject to purifying selection these recurrent variants are more likely to be the result of hypermutable sites. However, this does not exclude the possibility that some recurrent within-host variants are under positive selection. The spike glycoprotein mutation D614G, which has been shown to increase the transmissibility of the virus [402, 497] appears as a within-host variant in 12 of our samples with a higher VAF than

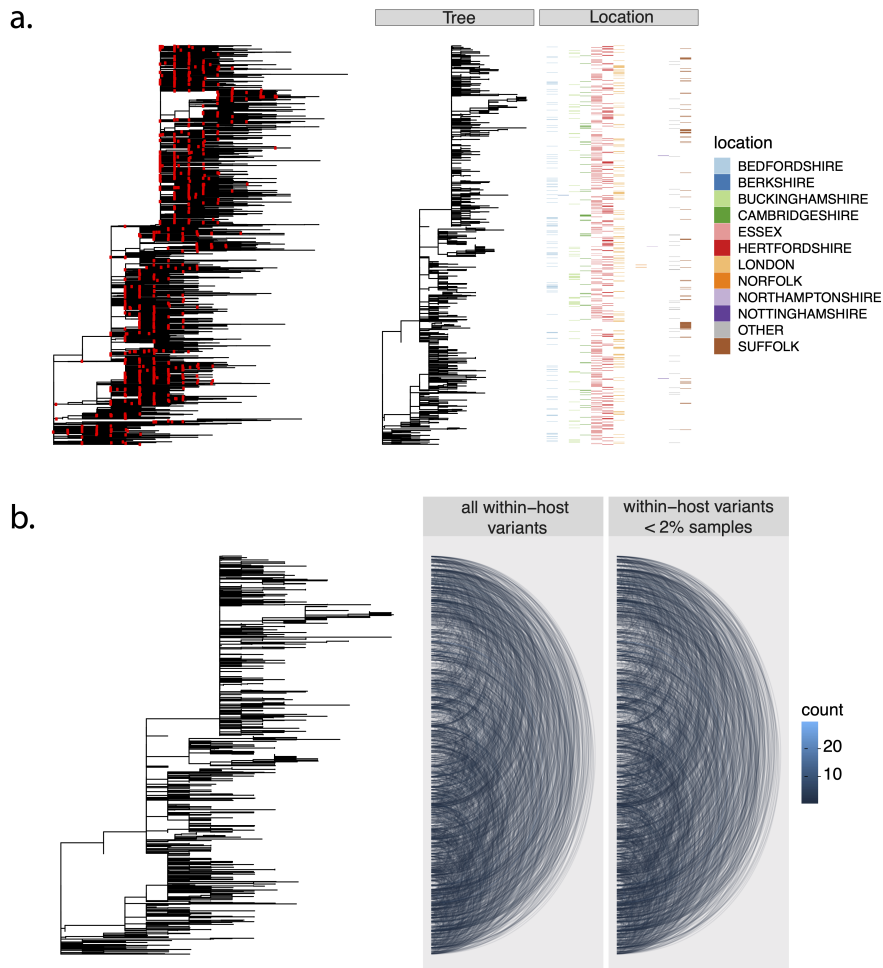


Figure 69: (a) Left panel - a maximum likelihood phylogeny of all COG-UK consensus genomes available on the 29th May 2020. Red dots indicate the location of those samples that were deep sequenced in replicate. Middle panel - the same phylogeny restricted to those samples taken for deep sequencing. Right panel - the region each patient's home address was located. (b) Left panel - maximum likelihood phylogeny of samples deep sequenced to investigate minority variants. Middle panel - links are drawn between tips of the phylogeny that share within-host minority. Right panel - links restricted to those variants seen in less than 2% of individuals.

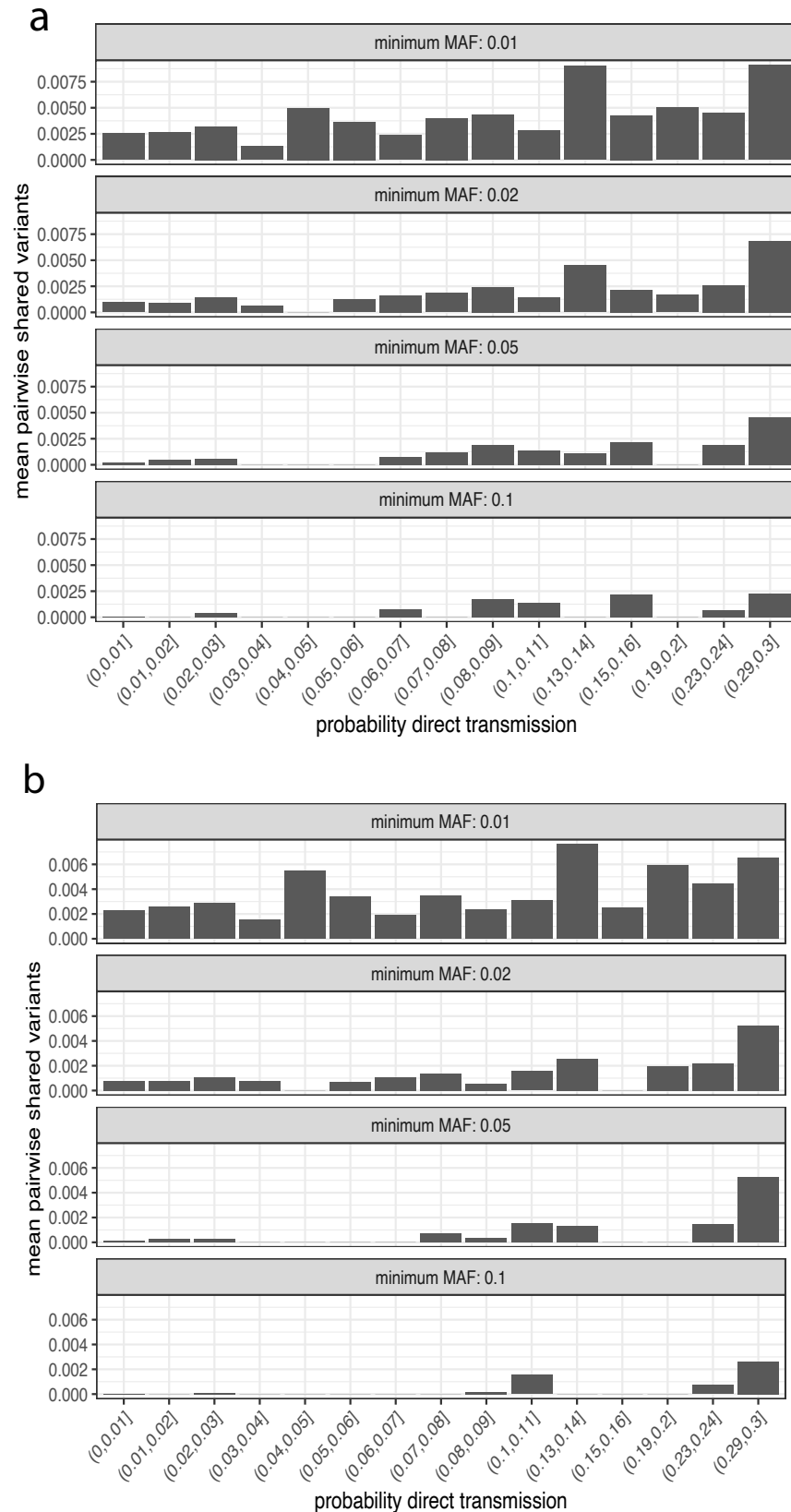


Figure 70: (a) The median number of shared iSNVs shared by each pair of samples binned by the probability they were the result of a direct transmission according to the model of Stimson *et al.*, 2019. Results, with a minimum minor allele frequency of 0.01, 0.02, 0.05 and 0.1 are shown in each of the facets. (b) The same plot as B but having removed all samples that were inferred to be mixed infections.

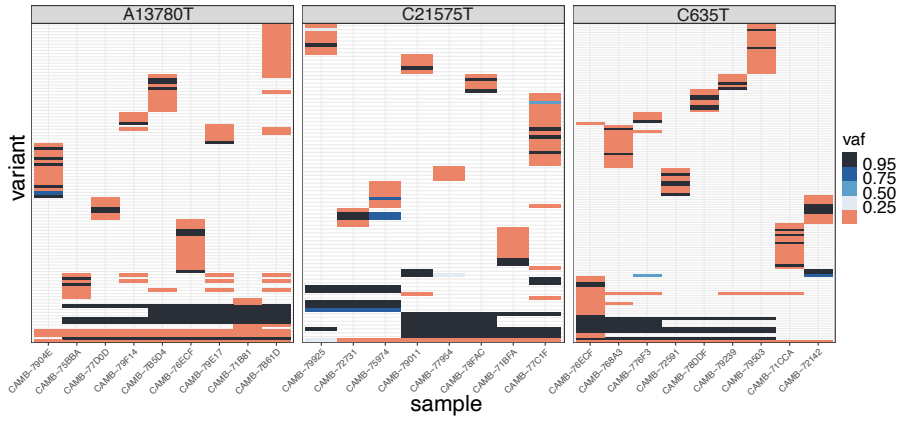


Figure 71: Heatmaps of variant allele frequencies in samples containing three common within-host variants are shown. The diversity of major variants (black tiles) across samples is better explained by independent acquisitions of the minority variant rather than transmission.

other within-host variants (median VAF 0.56 vs 0.036, Wilcoxon test $P=3.98e-4$).

9.3.8 Mixed infections

While most within-host variants are more concordant with recurrent mutation it is important to consider the phylogenetic transmission signal we observed in within-host variants at higher allele frequencies. It has been proposed that co-infection by different lineages is a significant source of within-host variation in SARS-CoV-2 (Lythgoe *et al.* 2020). Co-infection by multiple lineages would result in the allele frequencies of divergent bases to appear as within-host variants and reflect the proportions of either haplotype. To identify possible co-infected samples we compared the site frequency spectrum for each of our samples with all of the consensus genomes in the COG-UK dataset at the end of May 2020. A linear model was used to identify mixtures of two consensus genomes that could explain the allele frequencies of variants within a sample better than a single consensus genome. Additional mixtures were considered if two consensus genomes could explain at least two additional within-host variants over that of a single consensus genome. In this case the additional consensus genome was restricted to have at most one variant that was not found in the sample. Samples identified by the model as being a possible mixture of consensus genomes were then visually inspected to determine whether the putative co-infections were convincing. This resulted in 36 putative co-infected samples, with a representative example shown in [Figure 72](#). To determine the potential impact of co-infections on transmission inference we excluded all 36 samples and re-ran the pairwise transmission analysis using the Tran-

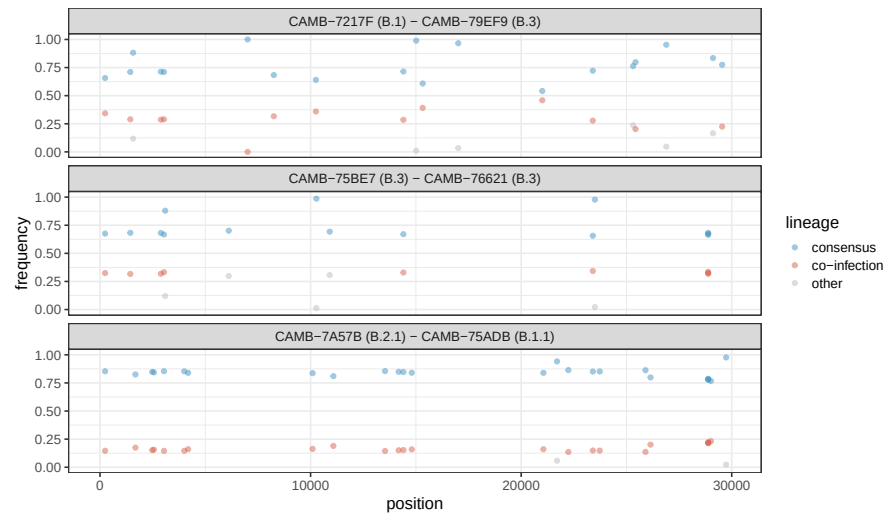


Figure 72: An example of three samples identified as potential mixtures. The consensus lineage is given first and coloured blue while the potentially co-infecting lineage is given second and coloured red. Minority variants that do not match the co-infecting lineage are coloured grey.

sCluster algorithm on the remaining samples. Figure 70b indicates that the remaining within-host variants no longer correlate as well with the probability of direct transmission, suggesting that much of the correlation observed previously was driven by co-infections.

It has been suggested that correlations between co-infections and the consensus phylogeny could be driven by the co-transmission of multiple strains [320]. However, as the transmission signal in the minority variants in the remaining samples has reduced; a more likely explanation is that co-infection of certain strains is driven by multiple episodes of superinfection (co-infection from two different infection sources). As these samples and those from the Lythgoe study were acquired from hospitalised patients, the correspondence with the consensus phylogeny could then be driven by structured superinfection within hospital COVID-19 wards. The prevalence of infection within hospitals, particularly during the first ‘wave’ of the COVID-19 pandemic was often considerably higher than in the community with up to 50% of the consultant A&E workforce at one Welsh hospital testing positive in April, 2020 [34]. This contrasts with the finding that approximately 6% of the UK population had been infected with SARS-CoV-2 by the end of June, 2020 [501]. Within hospital transmission was also common in the early stages of the outbreak in Wuhan with 41% of 138 patients were found to have contracted SARS-CoV-2 in hospital [499]. While it is not possible to rule out cross-contamination between samples, this contamination would have to be structured to explain the correlations between the inferred co-infections and the consensus phylogeny. Batch effects were carefully

controlled for by Lythgoe *et al.*, [320], and as we identified a similar signal using an independent dataset with samples sequenced in replicate, structured superinfection provides a better explanation. In cases of higher prevalence, repeated episodes of superinfection rather than co-transmission could complicate the use of within-host minority variants in transmission inference.

9.4 DISCUSSION

We have demonstrated the utility of using genomics to inform estimates of transmission and to gain a better understanding of the mechanisms of evolution of SARS-CoV-2 within the host.

By incorporating genomic data in the analysis of SARS-CoV-2 transmission in care homes in the East of England, we were able to identify an additional 93 potential introductions of the virus into care homes that were indistinguishable using symptom onset time data alone. Our transmission analysis suggested that infections in a number of care homes could have been the result of outbreaks with all or most infections plausibly connected by recent transmission events. These transmission clusters could also plausibly be driven by multiple introductions owing to the low viral diversity of SARS-CoV-2 within the broader community at the time these samples were taken [403]. These results suggest that preventing new introductions into care homes as well as limiting the possibility for onwards transmission is vital in preventing outbreaks within this vulnerable subpopulation.

Our findings of plausible transmission links between care home residents and healthcare workers including paramedics, care homes workers and hospital staff demonstrates the difficulties of an interconnected healthcare system. This was highlighted by a plausible transmission link between two care homes that involved healthcare workers. The higher prevalence of SARS-CoV-2 infection within hospitals during the first wave of the pandemic makes the interaction between healthcare workers and care home residents particularly problematic. Consequently, ensuring healthcare workers have access to adequate personal protection equipment coupled with other infection prevention efforts such as asymptomatic testing screens are critical in reducing the burden of COVID-19 within care homes.

Limitations to our analysis of SARS-CoV-2 transmission in care homes includes the limited number of positive samples for which genomic data was available as well as the low viral diversity within the broader population making it difficult to definitely describe chains of transmission. We also did not have data on healthcare worker status for those samples from tests carried out outside CUH, which limits our ability to accurately describe the origin of some care home infections as they could plausibly have been acquired within hospitals or from carers.

While the diversity between hosts in the care home dataset was limited, we identified considerable genetic diversity within individual SARS-CoV-2 infections in a separate dataset from the East of England that cannot be explained by technical artefacts. We estimated that each viral genome within a host, on average deviated from the consensus genome for that host by 0.72 variants. As most samples were collected within hospitals and thus were likely in the later stages of infection, this is consistent with the accumulation of most within-host genetic diversity during infection. This was supported by our observation in longitudinal samples from the same host that the number of within-host mutations tends to increase over time.

Our analysis of the mutational spectrum of within-host variants suggests that they are characterised by purifying selection, similar to variants found in the consensus phylogeny. This implies that selection may be primarily occurring within the host rather than at transmission. Unlike previous studies we observed a mutational spectrum with a clear asymmetry between plus and minus strands dominated by C>U and G>U mutations indicating the direct demand of cytosine and guanine bases could be a common mechanism of mutation in SARS-CoV-2.

Similar to previous studies we identify a large number of within-host variants that are shared between different hosts [320]. Many of these variants are shared between hosts that are distant on the corresponding consensus phylogeny including a number of variants that also appear frequently as homoplasies at the consensus level [113]. This coupled with the observation that in the absence of significant recombination, the transmission of within-host variants is incompatible with the simultaneous fixation of additional variants suggests that recurrent mutation rather than transmission is likely to be responsible for most shared variants although co-infection between lineages could also be a relevant factor in high transmission settings.

In our dataset, we identified approximately 3-4% of samples which were plausibly the result of co-infection with multiple SARS-CoV-2 lineages. This is similar to a separate analysis of within-host infection conducted in England where it was reported that the sharing of within-host variants correlated with the hospitals the samples were taken from. Here, Lythgoe and colleagues suggested these variants could be explained by localised episodes of superinfection (where an individual acquires infection from multiple sources) and subsequent co-transmission of multiple lineages [320]. Using the TransCluster algorithm, we find some correlation between shared within-host variants at higher allele frequencies and the probability of direct transmission. However, we find this signal is partially driven by co-infection. The higher prevalence of COVID-19 in many hospital wards increases the chances of superinfection. Thus, in samples taken from hospitals treating multiple patients with COVID-19 it may be hard to separate

shared within-host variants that are driven by superinfection from those caused by transmission complicating the use of within-host variants for transmission inference.

Overall we find substantial genetic diversity of SARS-CoV-2 within the host which appears to be primarily driven by the accumulation of *de novo* mutations during the course of an infection often at highly mutable sites. We identify a small number of plausible instances of co-infection that could further complicate transmission inference efforts. We thus recommend caution in the interpretation of minority variants for transmission inference and suggest that epidemiological data is essential in accurately determining transmission chains in SARS-CoV-2.

CONCLUSION

This thesis has described methodological advances in three key areas of bacterial genome analysis. The analysis of two new large sequencing datasets has provided unique insights into the transmission and within-host evolution of *S. pneumoniae* and SARS-CoV-2.

10.1 BACTERIAL GENOME ANALYSIS

Bacteria have been studied in detail since the discovery of the first bacterial cells in the 17th century. However, it was only in the 21st century that high throughput genomics enabled the study of bacterial evolution at scale and over different timescales. The increasing size and complexity of bacterial genomic datasets has required the development of new tools and techniques. This thesis has addressed three significant problems in the analysis of large WGS datasets, namely the identification of population structure, characterisation of gene content and tracking of transmission.

By approximating a Dirichlet process mixture (DPM) model, our genetic clustering method ‘fastbaps’ is able to cluster pathogen genetic datasets 10-100 times larger than previous approaches. We demonstrated the accuracy and speed of this method by clustering over 110,000 HIV pol genes as well as over 3000 pneumococcal genomes. We also provided an extension of the algorithm that is able to partition a predefined phylogeny to maximise the DPM model marginal likelihood in linear time. This allowed for the full phylogeny of HIV-1 pol genes to be partitioned in 45.8 seconds using a single cpu. The speed improvements provided by the ‘fastbaps’ algorithm have allowed for model based clustering of previously intractably large datasets. The approach also has considerable potential for use in surveillance settings where models are refitted regularly as new data becomes available.

The automated annotation of prokaryotic genomes is imperfect with errors due to fragmented assemblies, contamination, diverse gene-families and mis-assemblies leading to substantial problems when attempting to determine a species’ pangenome. In large genome collections these errors can accumulate. We have demonstrated that failing to account for errors in gene annotation in an extremely clonal *M. tuberculosis* dataset can lead to over 30% of the inferred gene clusters in the resulting pangenome being erroneous. We developed Panaroo, a tool that constructs a full graphical representation of the pangenome and leverages the resulting contextual information sur-

rounding erroneous clusters to correct for annotation errors. Panaroo was able to achieve more than a 10 fold reduction in the number of erroneous clusters. In addition to developing Panaroo to cluster genes into orthologous and paralogous groups we produced a set of downstream analysis tools to aid in the interpretation of the results. This allowed for the identification of gene gain and loss rates across 51 of the major Global Pneumococcal Sequencing Clusters and identified an association between a reduced rate of gene gain correlated with an increased odds ratio for that GPSC to cause invasive disease. A separate tool allowed for the interrogation of the graphical structure of the inferred pangenome, which we demonstrated on the EuroGASP collection of 1054 *Neisseria gonorrhoeae* isolates. Using this tool, we identified associations between antibiotic resistance and structural rearrangements near phage-associated transposase and pilin genes.

The SARS-CoV-2 pandemic has dramatically highlighted the importance of being able to accurately track the spread of infectious diseases through a population. We present extensions to existing algorithms including increasing the speed of TransCluster, an algorithm based upon pairwise SNP distance, by more than an order of magnitude allowing the algorithm to scale to datasets comprising over 100,000 genomes. The differences in tracking the outbreak of a disease versus transmission of an endemic pathogen were discussed and we show that adapting the TransPhylo transmission algorithm achieves improved estimates of transmission in endemic scenarios. A further adaptation of the TransPhylo algorithm enables a pathogen's generation time (time between infections) to be inferred if the sampling probability is known. This allowed for the inference of the generation times of the major clades of *S. pneumoniae* in the Maela refugee camp in Thailand. Finally, we describe initial efforts to implement a simulation based transmission inference algorithm. By simulating the within-host genetic diversity within a host and during transmission we showed that using recent advances in likelihood-free inference techniques, accurate transmission inferences could be made. However, as we only tested this approach on a limited set of simulated scenarios, additional work is needed to determine how robust the technique is and to determine how successfully it can be applied to real pathogen data.

10.2 GENETIC DIVERSITY OF THE PNEUMOCOCCUS WITHIN THE HOST AND IN TRANSMISSION

Deep sequencing whole plate sweeps of cultured pneumococci significantly increases our ability to detect virulence factors, resistance elements and instances of multiple carriage. In the Maela pneumococcal study, deep sequencing identified 17.7% more lineages than found using single colony picks which is the current standard in the genomic

surveillance of *S. pneumoniae* [190]. Crucially, the increased sensitivity of deep sequencing showed that well known virulent pneumococcal serotypes, including serotype 1, were more likely to be found at lower frequencies in cases of multiple carriage suggesting current techniques may systematically underestimate their prevalence within carriage. The 14.6% increase in the number of resistance calls, characterised using whole plates sequencing over single colony picks, highlights how this strategy adds value by informing both the surveillance of resistance within the pneumococcal population and the treatment of individual infections. As the cost and complexity of sequencing whole plate sweeps is comparable to that of single colony picks, this approach presents an attractive alternative for improving the surveillance of *S. pneumoniae*.

Deep sequencing whole plate sweeps has provided unique insights into the dynamics of multiple carriage in *S. pneumoniae*. Virulent serotypes were found more often in multiple carriage than would be expected given their overall prevalence. In contrast, resistance elements and other well known virulence factors were found at lower than expected rates in multiple carriage. These findings have important implications for models of the maintenance of strain structure within the *S. pneumoniae* population. In particular, the reduced rates of resistance elements in multiple carriage is consistent with these elements conferring a fitness cost leading them to be outcompeted within the host. This fits with the model of the maintenance of resistance at intermediate frequencies based on negative frequency dependent selection proposed by Davies et al. [111]. Alternatively, this signal could be driven by the correlation between certain resistance and virulence factors and the duration of carriage of different pneumococcal strains [294]. In the latter case, either lineages with longer carriage durations could outcompete other lineages within the host leading to a reduced presence in multiple carriage or, the increased carriage duration could lead to longer periods of single carriage interspersed with shorter periods of dual carriage driving the association. These results thus provide important additional data to improve future modelling studies.

The ability to observe within-host minority variants in pneumococcal carriage samples allowed us to better characterise transmission within the Maela refugee camp. Strikingly, we identified that mothers were more likely to transmit to their children in the first year of the child's life. This, along with the observation that the majority of transmission occurred from child to mother, is consistent with previous modelling studies fit to age stratified pneumococcal prevalence data [169]. Similarly, and consistent with previous studies, we identified strong spatial structure in transmission even within such a small geographic area as the Maela refugee camp [380]. These results have important implications for the design and implementation of inter-

ventions to try and prevent pneumococcal disease. For example, it may be beneficial to vaccinate mothers prior to them giving birth in order to better protect a child from pneumococcal infection during their first year of life. Using the additional information provided by within-host minority variants, we also found evidence that the transmission bottleneck in *S. pneumoniae* can be greater than one. This is in contrast to laboratory studies, which supported a pneumococcal bottleneck involving only a single bacterium [271]. This may allow for the transmission of heteroresistant populations of pneumococci allowing resistance to persist over longer timescales at low frequencies.

Within-host minority variants observed in samples containing only a single lineage allowed us to gain one of the most detailed pictures of the mutational and selective forces driving pneumococcal evolution over short timescales. We identified a mutational spectrum found uniquely within the host. This consisted of an elevated C>A mutation rate similar to that seen in previous laboratory studies of low frequency variants in *E. coli* which were found to likely be the result of misincorporation of As into C sites. These mutations are thought to be reversed prior to genome replication. Similar to other respiratory pathogens, variation within the host was dominated by purifying selection, with nonsense mutation being far rarer than both missense and synonymous variants. As well as identifying variants that occurred many times in different hosts we identified a number of genes under selection including those linked to the virulence of the pneumococcus and its competence system. These could be related to the competition between pneumococci within the host.

Overall, the deep sequencing of whole plate sweeps of pneumococcal samples has provided a fascinating level of detail into how selection, mutation and transmission shape the evolution of *S. pneumoniae*. This, coupled with the increased sensitivity to detect virulence and resistance factors and equivalent cost, suggests that routine use of whole plate sweeps could be preferable to current genomic surveillance strategies.

10.3 WITHIN-HOST DIVERSITY AND TRANSMISSION OF SARS-COV-

2

The characterisation of within-host diversity across 1,181 samples taken from the East of England and sequenced in replicate provided a detailed picture of the mutational and selective processes driving the evolution of SARS-CoV-2 within and between hosts. Critically, we found that many of the same within-host minority variants occurred independently in different samples. This suggests that considerable caution is needed for the interpretation of within-host variants in the context of transmission and the characterisation of the transmission

bottleneck. The detection of multiple putative cases of co-infection with different strains, possibly driven by transmission in COVID19 hospital wards, further complicates the use of minority variants in the inference of SARS-CoV-2 transmission chains when the prevalence of the virus is high. As with *S. pneumoniae*, within-host variants appeared to be under strong purifying selection. The mutational spectrum was characterised by strong strand asymmetries consistent with damage or RNA editing of the plus strand rather than replication errors. These results can help to inform future studies into the evolution of SARS-CoV-2 and provide vital context for efforts to track the transmission of the pathogen using within-host genome sequencing data.

10.4 FUTURE WORK

A number of chapters in this thesis describe methods and analyses that are ongoing. The development and improvement of the simulation based, likelihood-free transmission inference algorithm requires further refinement and testing to ensure it can be accurately applied to real datasets. The impacts of the sampling and culture steps in the Maela pneumococcal deep sequencing study are also being studied in more detail. This will involve the analysis of replicates where the same culture and sequencing steps have been repeated for a subset of samples. As the pneumococcal single colony picks used in Chewapreecha et al., [71] all undergo a subsequent culture step after being selected, the analysis of minority variants present in this data will allow for a better understanding of the variation that can be introduced by culture. Finally, sequencing a subset of the pneumococcal samples using long read technologies will improve our ability to verify the results of the mGEMs deconvolution pipeline and allow for investigation of recombination within the host.

BIBLIOGRAPHY

- [1] David M Aanensen, Edward J Feil, et al. "Whole-Genome Sequencing for Routine Pathogen Surveillance in Public Health: a Population Snapshot of Invasive *Staphylococcus aureus* in Europe." en. In: *MBio* 7.3 (May 2016). ISSN: 2150-7511. DOI: [10.1128/mBio.00444-16](https://doi.org/10.1128/mBio.00444-16). URL: <http://dx.doi.org/10.1128/mBio.00444-16>.
- [2] P V Adrian and K P Klugman. "Mutations in the dihydrofolate reductase gene of trimethoprim-resistant isolates of *Streptococcus pneumoniae*." en. In: *Antimicrob. Agents Chemother.* 41.11 (Nov. 1997), pp. 2406–2413. ISSN: 0066-4804. URL: <https://www.ncbi.nlm.nih.gov/pubmed/9371341>.
- [3] H Akaike. "A new look at the statistical model identification." In: *IEEE Trans. Automat. Contr.* 19.6 (Dec. 1974), pp. 716–723. ISSN: 0018-9286. DOI: [10.1109/TAC.1974.1100705](https://doi.org/10.1109/TAC.1974.1100705). URL: <http://dx.doi.org/10.1109/TAC.1974.1100705>.
- [4] David H Alexander, John Novembre, and Kenneth Lange. "Fast model-based estimation of ancestry in unrelated individuals." en. In: *Genome Res.* 19.9 (Sept. 2009), pp. 1655–1664. ISSN: 1088-9051, 1549-5469. DOI: [10.1101/gr.094052.109](https://doi.org/10.1101/gr.094052.109). URL: <http://dx.doi.org/10.1101/gr.094052.109>.
- [5] Ludmil B Alexandrov, Serena Nik-Zainal, et al. "Signatures of mutational processes in human cancer." en. In: *Nature* 500.7463 (Aug. 2013), pp. 415–421. ISSN: 0028-0836, 1476-4687. DOI: [10.1038/nature12477](https://doi.org/10.1038/nature12477). URL: <http://dx.doi.org/10.1038/nature12477>.
- [6] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." en. In: *Nucleic Acids Res.* 25.17 (Sept. 1997), pp. 3389–3402. ISSN: 0305-1048, 1362-4962. DOI: [10.1093/nar/25.17.3389](https://doi.org/10.1093/nar/25.17.3389). URL: <http://nar.oxfordjournals.org/content/25/17/3389>.
- [7] Kristian G Andersen, Andrew Rambaut, W Ian Lipkin, Edward C Holmes, and Robert F Garry. "The proximal origin of SARS-CoV-2." en. In: *Nat. Med.* 26.4 (Apr. 2020), pp. 450–452. ISSN: 1078-8956, 1546-170X. DOI: [10.1038/s41591-020-0820-9](https://doi.org/10.1038/s41591-020-0820-9). URL: <http://dx.doi.org/10.1038/s41591-020-0820-9>.

- [8] E C Anderson and E A Thompson. "A model-based method for identifying species hybrids using multilocus genetic data." en. In: *Genetics* 160.3 (Mar. 2002), pp. 1217–1229. ISSN: 0016-6731. URL: <https://www.ncbi.nlm.nih.gov/pubmed/11901135>.
- [9] S G Andersson, A Zomorodipour, J O Andersson, T Sicheritz-Pontén, U C Alsmark, R M Podowski, A K Näslund, A S Eriksson, H H Winkler, and C G Kurland. "The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria." en. In: *Nature* 396.6707 (Nov. 1998), pp. 133–140. ISSN: 0028-0836. DOI: [10.1038/24094](https://doi.org/10.1038/24094). URL: <http://dx.doi.org/10.1038/24094>.
- [10] Maria Anisimova, Joseph Bielawski, Katherine Dunn, and Ziheng Yang. "Phylogenomic analysis of natural selection pressure in *Streptococcus* genomes." en. In: *BMC Evol. Biol.* 7 (Aug. 2007), p. 154. ISSN: 1471-2148. DOI: [10.1186/1471-2148-7-154](https://doi.org/10.1186/1471-2148-7-154). URL: <http://dx.doi.org/10.1186/1471-2148-7-154>.
- [11] Hiasindh Ashmi Antony and Subhash Chandra Parija. "Antimalarial drug resistance: an overview." In: *Trop. Parasitol.* 6.1 (2016), p. 30. ISSN: 2229-5070. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/pmc4778180/>.
- [12] R Austrian. "The Jeremiah Metzger Lecture: Of gold and pneumococci: a history of pneumococcal vaccines in South Africa." en. In: *Trans. Am. Clin. Climatol. Assoc.* 89 (1978), pp. 141–161. ISSN: 0065-7778. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30204>.
- [13] R Austrian. "Pneumococcal otitis media and pneumococcal vaccines, a historical perspective." en. In: *Vaccine* 19 Suppl 1 (Dec. 2000), S71–7. ISSN: 0264-410X. DOI: [10.1016/S0264-410X\(00\)00282-6](https://doi.org/10.1016/S0264-410X(00)00282-6). URL: [http://dx.doi.org/10.1016/S0264-410X\(00\)00282-6](http://dx.doi.org/10.1016/S0264-410X(00)00282-6).
- [14] O T Avery and W F Goebel. "Chemo-immunological studies on conjugated carbohydrate-proteins: V. The immunological specificity of an antigen prepared by combining the capsular polysaccharide of type III pneumococcus with foreign protein." en. In: *J. Exp. Med.* 54.3 (July 1931), pp. 437–447. ISSN: 0022-1007. DOI: [10.1084/jem.54.3.437](https://doi.org/10.1084/jem.54.3.437). URL: <http://dx.doi.org/10.1084/jem.54.3.437>.
- [15] Oswald T Avery, Colin M MacLeod, and Maclyn McCarty. "Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III." In: *J. Exp. Med.* 79.2 (1944), pp. 137–158. ISSN: 0022-1007. URL: <https://rupress.org/jem/article-abstract/79/2/137/4753>.

- [16] Fabio Bagnoli, Monica Moschioni, et al. "A second pilus type in *Streptococcus pneumoniae* is prevalent in emerging serotypes and mediates adhesion to host cells." en. In: *J. Bacteriol.* 190.15 (Aug. 2008), pp. 5480–5492. ISSN: 0021-9193, 1098-5530. DOI: [10.1128/JB.00384-08](https://doi.org/10.1128/JB.00384-08). URL: <http://dx.doi.org/10.1128/JB.00384-08>.
- [17] Anton Bankevich, Sergey Nurk, et al. "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing." en. In: *J. Comput. Biol.* 19.5 (May 2012), pp. 455–477. ISSN: 1066-5277, 1557-8666. DOI: [10.1089/cmb.2012.0021](https://doi.org/10.1089/cmb.2012.0021). URL: <http://dx.doi.org/10.1089/cmb.2012.0021>.
- [18] Yinon M Bar-On, Rob Phillips, and Ron Milo. "The biomass distribution on Earth." en. In: *Proc. Natl. Acad. Sci. U. S. A.* 115.25 (June 2018), pp. 6506–6511. ISSN: 0027-8424, 1091-6490. DOI: [10.1073/pnas.1711842115](https://doi.org/10.1073/pnas.1711842115). URL: <http://dx.doi.org/10.1073/pnas.1711842115>.
- [19] F Barany and A Tomasz. "Genetic transformation of *Streptococcus pneumoniae* by heterologous plasmid deoxyribonucleic acid." en. In: *J. Bacteriol.* 144.2 (Nov. 1980), pp. 698–709. ISSN: 0021-9193. DOI: [10.1128/JB.144.2.698-709.1980](https://doi.org/10.1128/JB.144.2.698-709.1980). URL: <http://dx.doi.org/10.1128/JB.144.2.698-709.1980>.
- [20] V A Barcus, K Ghanekar, M Yeo, T J Coffey, and C G Dowson. "Genetics of high level penicillin resistance in clinical isolates of *Streptococcus pneumoniae*." en. In: *FEMS Microbiol. Lett.* 126.3 (Mar. 1995), pp. 299–303. ISSN: 0378-1097. DOI: [10.1111/j.1574-6968.1995.tb07433.x](https://doi.org/10.1111/j.1574-6968.1995.tb07433.x). URL: <http://dx.doi.org/10.1111/j.1574-6968.1995.tb07433.x>.
- [21] Franz Baumdicker, Wolfgang R Hess, and Peter Pfaffelhuber. "The infinitely many genes model for the distributed genome of bacteria." en. In: *Genome Biol. Evol.* 4.4 (Feb. 2012), pp. 443–456. ISSN: 1759-6653. DOI: [10.1093/gbe/evs016](https://doi.org/10.1093/gbe/evs016). URL: <http://dx.doi.org/10.1093/gbe/evs016>.
- [22] Franz Baumdicker and Peter Pfaffelhuber. "The infinitely many genes model with horizontal gene transfer." en. In: *Electron. J. Probab.* 19 (2014). ISSN: 1083-6489. DOI: [10.1214/EJP.v19-2642](https://doi.org/10.1214/EJP.v19-2642). URL: <https://projecteuclid.org/euclid.ejp/1465065757>.
- [23] Alison J Baylay, Alasdair Ivens, and Laura J V Piddock. "A novel gene amplification causes upregulation of the PatAB ABC transporter and fluoroquinolone resistance in *Streptococcus pneumoniae*." en. In: *Antimicrob. Agents Chemother.* 59.6 (Mar. 2015), pp. 3098–3108. ISSN: 0066-4804, 1098-6596. DOI: [10.1128/AAC.04858-14](https://doi.org/10.1128/AAC.04858-14). URL: <http://dx.doi.org/10.1128/AAC.04858-14>.

- [24] Sion C Bayliss, Harry A Thorpe, Nicola M Coyle, Samuel K Sheppard, and Edward J Feil. "PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria." en. Apr. 2019. URL: <https://www.biorxiv.org/content/10.1101/598391v1.abstract>.
- [25] Mark A Beaumont, Wenyang Zhang, and David J Balding. "Approximate Bayesian computation in population genetics." en. In: *Genetics* 162.4 (Dec. 2002), pp. 2025–2035. ISSN: 0016-6731. URL: <https://www.ncbi.nlm.nih.gov/pubmed/12524368>.
- [26] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel W H Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. "Dimensionality reduction for visualizing single-cell data using UMAP." In: *Nat. Biotechnol.* 37 (Dec. 2018), p. 38. ISSN: 1087-0156. DOI: [10.1038/nbt.4314](https://doi.org/10.1038/nbt.4314). URL: <https://doi.org/10.1038/nbt.4314>.
- [27] Malene Bek-Thomsen, Knud Poulsen, and Mogens Kilian. "Occurrence and evolution of the paralogous zinc metalloproteases IgA1 protease, ZmpB, ZmpC, and ZmpD in *Streptococcus pneumoniae* and related commensal species." en. In: *MBio* 3.5 (Oct. 2012). ISSN: 2150-7511. DOI: [10.1128/mBio.00303-12](https://doi.org/10.1128/mBio.00303-12). URL: <http://dx.doi.org/10.1128/mBio.00303-12>.
- [28] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." In: *J. R. Stat. Soc. Series B Stat. Methodol.* 57.1 (Jan. 1995), pp. 289–300. ISSN: 1369-7412, 0035-9246. URL: <http://www.jstor.org/stable/2346101>.
- [29] S D Bentley, K F Chater, et al. "Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2)." en. In: *Nature* 417.6885 (May 2002), pp. 141–147. ISSN: 0028-0836. DOI: [10.1038/417141a](https://doi.org/10.1038/417141a). URL: <http://dx.doi.org/10.1038/417141a>.
- [30] S Bergmann, M Rohde, G S Chhatwal, and S Hammerschmidt. "alpha-Enolase of *Streptococcus pneumoniae* is a plasmin(ogen)-binding protein displayed on the bacterial cell surface." en. In: *Mol. Microbiol.* 40.6 (June 2001), pp. 1273–1287. ISSN: 0950-382X. DOI: [10.1046/j.1365-2958.2001.02448.x](https://doi.org/10.1046/j.1365-2958.2001.02448.x). URL: <http://dx.doi.org/10.1046/j.1365-2958.2001.02448.x>.
- [31] A M Berry, E M Glare, D Hansman, and J C Paton. "Presence of a small plasmid in clinical isolates of *Streptococcus pneumoniae*." en. In: *FEMS Microbiol. Lett.* 53.3 (Dec. 1989), pp. 275–278. ISSN: 0378-1097. DOI: [10.1016/0378-1097\(89\)90230-9](https://doi.org/10.1016/0378-1097(89)90230-9). URL: [http://dx.doi.org/10.1016/0378-1097\(89\)90230-9](http://dx.doi.org/10.1016/0378-1097(89)90230-9).
- [32] Marie-Pauline Beugin, Thibault Gayet, Dominique Pontier, Sébastien Devillard, and Thibaut Jombart. "A fast likelihood solution to the genetic clustering problem." In: *Methods Ecol. Evol.* 9.4 (Apr.

- 2018). Ed. by Thomas Hansen, pp. 1006–1016. ISSN: 2041-210X. DOI: [10.1111/2041-210X.12968](https://doi.org/10.1111/2041-210X.12968). URL: <http://doi.wiley.com/10.1111/2041-210X.12968>.
- [33] F Bezançon and V Griffon. “Pouvoir agglutinatif du sérum dans les infections expérimentales et humaines à pneumocoques.” In: *ICR Soc Biol* 49 (1897), p. 551.
- [34] James R M Black, Chris Bailey, Joanna Przewrocka, Krijn K Dijkstra, and Charles Swanton. “COVID-19: the case for health-care worker screening to prevent hospital transmission.” en. In: *Lancet* 395.10234 (May 2020), pp. 1418–1420. ISSN: 0140-6736, 1474-547X. DOI: [10.1016/S0140-6736\(20\)30917-X](https://doi.org/10.1016/S0140-6736(20)30917-X). URL: [http://dx.doi.org/10.1016/S0140-6736\(20\)30917-X](http://dx.doi.org/10.1016/S0140-6736(20)30917-X).
- [35] David M Blei and Michael I Jordan. “Variational inference for Dirichlet process mixtures.” en. In: *Bayesian Anal.* 1.1 (Mar. 2006), pp. 121–143. ISSN: 1936-0975, 1931-6690. DOI: [10.1214/06-BA104](https://doi.org/10.1214/06-BA104). URL: <https://projecteuclid.org/euclid.ba/1340371077>.
- [36] M Blot. “Transposable elements and adaptation of host bacteria.” en. In: *Genetica* 93.1-3 (1994), pp. 5–12. ISSN: 0016-6707. DOI: [10.1007/BF01435235](https://doi.org/10.1007/BF01435235). URL: <http://dx.doi.org/10.1007/BF01435235>.
- [37] C E Blue, G K Paterson, A R Kerr, M Bergé, J P Claverys, and T J Mitchell. “ZmpB, a novel virulence factor of *Streptococcus pneumoniae* that induces tumor necrosis factor alpha production in the respiratory tract.” en. In: *Infect. Immun.* 71.9 (Sept. 2003), pp. 4925–4935. ISSN: 0019-9567. DOI: [10.1128/iai.71.9.4925-4935.2003](https://doi.org/10.1128/iai.71.9.4925-4935.2003). URL: <http://dx.doi.org/10.1128/iai.71.9.4925-4935.2003>.
- [38] Kirsten I Bos, Verena J Schuenemann, et al. “A draft genome of *Yersinia pestis* from victims of the Black Death.” en. In: *Nature* 478.7370 (Oct. 2011), pp. 506–510. ISSN: 0028-0836, 1476-4687. DOI: [10.1038/nature10549](https://doi.org/10.1038/nature10549). URL: <http://dx.doi.org/10.1038/nature10549>.
- [39] Luis Boto. “Horizontal gene transfer in evolution: facts and challenges.” en. In: *Proc. Biol. Sci.* 277.1683 (Mar. 2010), pp. 819–827. ISSN: 0962-8452, 1471-2954. DOI: [10.1098/rspb.2009.1679](https://doi.org/10.1098/rspb.2009.1679). URL: <http://dx.doi.org/10.1098/rspb.2009.1679>.
- [40] Phelim Bradley, N Claire Gordon, et al. “Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*.” en. In: *Nat. Commun.* 6 (Dec. 2015), p. 10063. ISSN: 2041-1723. DOI: [10.1038/ncomms10063](https://doi.org/10.1038/ncomms10063). URL: <http://dx.doi.org/10.1038/ncomms10063>.
- [41] Karel Břinda. *disty*. URL: <https://github.com/c2-d2/disty>.

- [42] Stef L A M Bronzwaer, Stef L A, et al. *The Relationship between Antimicrobial Use and Antimicrobial Resistance in Europe*. 2002. DOI: [10.3201/eid0803.010192](https://doi.org/10.3201/eid0803.010192). URL: <http://dx.doi.org/10.3201/eid0803.010192>.
- [43] Silvio D Brugger, Pascal Frey, Suzanne Aepli, Jason Hinds, and Kathrin Mühlemann. "Multiple colonization with *S. pneumoniae* before and after introduction of the seven-valent conjugated pneumococcal polysaccharide vaccine." en. In: *PLoS One* 5.7 (July 2010), e11638. ISSN: 1932-6203. DOI: [10.1371/journal.pone.0011638](https://doi.org/10.1371/journal.pone.0011638). URL: <http://dx.doi.org/10.1371/journal.pone.0011638>.
- [44] Silvio D Brugger, Lucy J Hathaway, and Kathrin Mühlemann. "Detection of *Streptococcus pneumoniae* strain cocolonization in the nasopharynx." en. In: *J. Clin. Microbiol.* 47.6 (June 2009), pp. 1750–1756. ISSN: 0095-1137, 1098-660X. DOI: [10.1128/JCM.01877-08](https://doi.org/10.1128/JCM.01877-08). URL: <http://dx.doi.org/10.1128/JCM.01877-08>.
- [45] Harald Brüssow, Carlos Canchaya, and Wolf-Dietrich Hardt. "Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion." en. In: *Microbiol. Mol. Biol. Rev.* 68.3 (Sept. 2004), 560–602, table of contents. ISSN: 1092-2172. DOI: [10.1128/MMBR.68.3.560-602.2004](https://doi.org/10.1128/MMBR.68.3.560-602.2004). URL: <http://dx.doi.org/10.1128/MMBR.68.3.560-602.2004>.
- [46] Josephine M Bryant, Dorothy M Grogono, et al. "Emergence and spread of a human-transmissible multidrug-resistant non-tuberculous mycobacterium." en. In: *Science* 354.6313 (Nov. 2016), pp. 751–757. ISSN: 0036-8075, 1095-9203. DOI: [10.1126/science.aaf8156](https://doi.org/10.1126/science.aaf8156). URL: <http://dx.doi.org/10.1126/science.aaf8156>.
- [47] Josephine M Bryant, Simon R Harris, et al. "Whole-genome sequencing to establish relapse or re-infection with *Mycobacterium tuberculosis*: a retrospective observational study." en. In: *Lancet Respir Med* 1.10 (Dec. 2013), pp. 786–792. ISSN: 2213-2619, 2213-2600. DOI: [10.1016/S2213-2600\(13\)70231-5](https://doi.org/10.1016/S2213-2600(13)70231-5). URL: [http://dx.doi.org/10.1016/S2213-2600\(13\)70231-5](http://dx.doi.org/10.1016/S2213-2600(13)70231-5).
- [48] Ola Brynildsrud, Jon Bohlin, Lonneke Scheffer, and Vegard Eldholm. "Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary." en. In: *Genome Biol.* 17.1 (Nov. 2016), p. 238. ISSN: 1465-6906. DOI: [10.1186/s13059-016-1108-8](https://doi.org/10.1186/s13059-016-1108-8). URL: <http://dx.doi.org/10.1186/s13059-016-1108-8>.
- [49] Benjamin Buchfink, Chao Xie, and Daniel H Huson. "Fast and sensitive protein alignment using DIAMOND." en. In: *Nat. Methods* 12.1 (Jan. 2015), pp. 59–60. ISSN: 1548-7091, 1548-7105. DOI: [10.1038/nmeth.3176](https://doi.org/10.1038/nmeth.3176). URL: <http://dx.doi.org/10.1038/nmeth.3176>.

- [50] V Burdett, J Inamine, and S Rajagopalan. "Heterogeneity of tetracycline resistance determinants in *Streptococcus*." en. In: *J. Bacteriol.* 149.3 (Mar. 1982), pp. 995–1004. ISSN: 0021-9193. URL: <https://www.ncbi.nlm.nih.gov/pubmed/6801018>.
- [51] Stephen J Bush, Dona Foster, David W Eyre, Emily L Clark, Nicola De Maio, Liam P Shaw, Nicole Stoesser, Tim E A Peto, Derrick W Crook, and A Sarah Walker. "Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism-calling pipelines." en. In: *Gigascience* 9.2 (Feb. 2020). ISSN: 2047-217X. DOI: [10.1093/gigascience/giaa007](https://doi.org/10.1093/gigascience/giaa007). URL: <http://dx.doi.org/10.1093/gigascience/giaa007>.
- [52] A Buu-Hoï and T Horodniceanu. "Conjugative transfer of multiple antibiotic resistance markers in *Streptococcus pneumoniae*." en. In: *J. Bacteriol.* 143.1 (July 1980), pp. 313–320. ISSN: 0021-9193. DOI: [10.1128/JB.143.1.313-320.1980](https://doi.org/10.1128/JB.143.1.313-320.1980). URL: <http://dx.doi.org/10.1128/JB.143.1.313-320.1980>.
- [53] COG-UK. "An integrated national scale SARS-CoV-2 genomic surveillance network." In: *The Lancet Microbe* (June 2020). ISSN: 2666-5247. DOI: [10.1016/S2666-5247\(20\)30054-9](https://doi.org/10.1016/S2666-5247(20)30054-9). URL: [https://doi.org/10.1016/S2666-5247\(20\)30054-9](https://doi.org/10.1016/S2666-5247(20)30054-9).
- [54] Joseph H Camin and Robert R Sokal. "A method for deducing branching sequences in phylogeny." In: *Evolution* 19.3 (Sept. 1965), pp. 311–326. ISSN: 0014-3820, 1558-5646. DOI: [10.1111/j.1558-5646.1965.tb01722.x](https://doi.org/10.1111/j.1558-5646.1965.tb01722.x). URL: <http://dx.doi.org/10.1111/j.1558-5646.1965.tb01722.x>.
- [55] Finlay Campbell, Xavier Didelot, Rich Fitzjohn, Neil Ferguson, Anne Cori, and Thibaut Jombart. "outbreaker2: a modular platform for outbreak reconstruction." en. In: *BMC Bioinformatics* 19.Suppl 11 (Oct. 2018), p. 363. ISSN: 1471-2105. DOI: [10.1186/s12859-018-2330-z](https://doi.org/10.1186/s12859-018-2330-z). URL: <http://dx.doi.org/10.1186/s12859-018-2330-z>.
- [56] Finlay Campbell, Camilla Strang, Neil Ferguson, Anne Cori, and Thibaut Jombart. "When are pathogen genome sequences informative of transmission events?" en. In: *PLoS Pathog.* 14.2 (Feb. 2018), e1006885. ISSN: 1553-7366, 1553-7374. DOI: [10.1371/journal.ppat.1006885](https://doi.org/10.1371/journal.ppat.1006885). URL: <http://dx.doi.org/10.1371/journal.ppat.1006885>.
- [57] Nicola Casali, Agnieszka Broda, Simon R Harris, Julian Parkhill, Timothy Brown, and Francis Drobniewski. "Whole Genome Sequence Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective Observational Study." en. In: *PLoS Med.* 13.10 (Oct. 2016), e1002137. ISSN: 1549-1277, 1549-1676. DOI: [10.1371/journal.pmed.1002137](https://doi.org/10.1371/journal.pmed.1002137). URL: <http://dx.doi.org/10.1371/journal.pmed.1002137>.

- [58] Eric Cascales and Peter J Christie. "The versatile bacterial type IV secretion systems." en. In: *Nat. Rev. Microbiol.* 1.2 (Nov. 2003), pp. 137–149. ISSN: 1740-1526. DOI: [10.1038/nrmicro753](https://doi.org/10.1038/nrmicro753). URL: <http://dx.doi.org/10.1038/nrmicro753>.
- [59] Sherwood R Casjens, Emmanuel F Mongodin, Wei-Gang Qiu, John J Dunn, Benjamin J Luft, Claire M Fraser-Liggett, and Steve E Schutzer. "Whole-genome sequences of two *Borrelia afzelii* and two *Borrelia garinii* Lyme disease agent isolates." en. In: *J. Bacteriol.* 193.24 (Dec. 2011), pp. 6995–6996. ISSN: 0021-9193, 1098-5530. DOI: [10.1128/JB.05951-11](https://doi.org/10.1128/JB.05951-11). URL: <http://dx.doi.org/10.1128/JB.05951-11>.
- [60] Santiago Castillo-Ramírez, Simon R Harris, Matthew T G Holden, Miao He, Julian Parkhill, Stephen D Bentley, and Edward J Feil. "The impact of recombination on dN/dS within recently emerged bacterial clones." en. In: *PLoS Pathog.* 7.7 (July 2011), e1002129. ISSN: 1553-7366, 1553-7374. DOI: [10.1371/journal.ppat.1002129](https://doi.org/10.1371/journal.ppat.1002129). URL: <http://dx.doi.org/10.1371/journal.ppat.1002129>.
- [61] A Cayley. "A theorem on trees." In: *Quart. J. Pure Appl. Math.* 23 (1889), pp. 376–378.
- [62] Chrispin Chaguza, Jennifer E Cornick, et al. "Understanding pneumococcal serotype 1 biology through population genomic analysis." en. In: *BMC Infect. Dis.* 16.1 (Nov. 2016), p. 649. ISSN: 1471-2334. DOI: [10.1186/s12879-016-1987-z](https://doi.org/10.1186/s12879-016-1987-z). URL: <http://dx.doi.org/10.1186/s12879-016-1987-z>.
- [63] Chrispin Chaguza, Madikay Senghore, et al. "Within-host microevolution of *Streptococcus pneumoniae* is rapid and adaptive during natural colonisation." en. In: *Nat. Commun.* 11.1 (July 2020), pp. 1–14. ISSN: 2041-1723, 2041-1723. DOI: [10.1038/s41467-020-17327-w](https://doi.org/10.1038/s41467-020-17327-w). URL: <https://www.nature.com/articles/s41467-020-17327-w>.
- [64] Marie Anne Chattaway, Timothy J Dallman, et al. "The Transformation of Reference Microbiology Methods and Surveillance for *Salmonella* With the Use of Whole Genome Sequencing in England and Wales." en. In: *Front Public Health* 7 (Nov. 2019), p. 317. ISSN: 2296-2565. DOI: [10.3389/fpubh.2019.00317](https://doi.org/10.3389/fpubh.2019.00317). URL: <http://dx.doi.org/10.3389/fpubh.2019.00317>.
- [65] D K Chen, A McGeer, J C de Azavedo, and D E Low. "Decreased susceptibility of *Streptococcus pneumoniae* to fluoroquinolones in Canada. Canadian Bacterial Surveillance Network." en. In: *N. Engl. J. Med.* 341.4 (July 1999), pp. 233–239. ISSN: 0028-4793. DOI: [10.1056/NEJM199907223410403](https://doi.org/10.1056/NEJM199907223410403). URL: <http://dx.doi.org/10.1056/NEJM199907223410403>.

- [66] Lihong Chen, Dandan Zheng, Bo Liu, Jian Yang, and Qi Jin. "VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on." en. In: *Nucleic Acids Res.* 44.D1 (Jan. 2016), pp. D694–7. ISSN: 0305-1048, 1362-4962. DOI: [10.1093/nar/gkv1239](https://doi.org/10.1093/nar/gkv1239). URL: <http://dx.doi.org/10.1093/nar/gkv1239>.
- [67] Ye Chen, Atsuko Hayashida, Allison E Bennett, Susan K Hollingshead, and Pyong Woo Park. "Streptococcus pneumoniae sheds syndecan-1 ectodomains through ZmpC, a metalloproteinase virulence factor." en. In: *J. Biol. Chem.* 282.1 (Jan. 2007), pp. 159–167. ISSN: 0021-9258. DOI: [10.1074/jbc.M608542200](https://doi.org/10.1074/jbc.M608542200). URL: <http://dx.doi.org/10.1074/jbc.M608542200>.
- [68] Yi-Yin Chen, Jin-Town Wang, Tzu-Lung Lin, Yu-Nong Gong, Ting-Hsuan Li, Ya-Yu Huang, and Yu-Chia Hsieh. "Prophage Excision in *Streptococcus pneumoniae* Serotype 19A ST320 Promote Colonization: Insight Into Its Evolution From the Ancestral Clone Taiwan 19F-14 (ST236)." In: *Front. Microbiol.* 10 (2019), p. 205. ISSN: 1664-302X. DOI: [10.3389/fmicb.2019.00205](https://doi.org/10.3389/fmicb.2019.00205). URL: <https://www.frontiersin.org/article/10.3389/fmicb.2019.00205>.
- [69] Alice G Cheng, Dominique Missiakas, and Olaf Schneewind. "The giant protein Ebh is a determinant of *Staphylococcus aureus* cell size and complement resistance." en. In: *J. Bacteriol.* 196.5 (Mar. 2014), pp. 971–981. ISSN: 0021-9193, 1098-5530. DOI: [10.1128/JB.01366-13](https://doi.org/10.1128/JB.01366-13). URL: <http://dx.doi.org/10.1128/JB.01366-13>.
- [70] Lu Cheng, Thomas R Connor, Jukka Sirén, David M Aanensen, and Jukka Corander. "Hierarchical and spatially explicit clustering of DNA sequences with BAPS software." en. In: *Mol. Biol. Evol.* 30.5 (May 2013), pp. 1224–1228. ISSN: 0737-4038, 1537-1719. DOI: [10.1093/molbev/mst028](https://doi.org/10.1093/molbev/mst028). URL: <http://dx.doi.org/10.1093/molbev/mst028>.
- [71] Claire Chewapreecha, Simon R Harris, et al. "Dense genomic sampling identifies highways of pneumococcal recombination." en. In: *Nat. Genet.* 46.3 (Mar. 2014), pp. 305–309. ISSN: 1061-4036, 1546-1718. DOI: [10.1038/ng.2895](https://doi.org/10.1038/ng.2895). URL: <http://dx.doi.org/10.1038/ng.2895>.
- [72] Claire Chewapreecha, Pekka Marttinen, et al. "Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes." en. In: *PLoS Genet.* 10.8 (Aug. 2014), e1004547. ISSN: 1553-7390, 1553-7404. DOI: [10.1371/journal.pgen.1004547](https://doi.org/10.1371/journal.pgen.1004547). URL: <http://dx.doi.org/10.1371/journal.pgen.1004547>.

- [73] Chen-Shan Chin, Jon Sorenson, et al. "The origin of the Haitian cholera outbreak strain." en. In: *N. Engl. J. Med.* 364.1 (Jan. 2011), pp. 33–42. ISSN: 0028-4793, 1533-4406. DOI: [10.1056/NEJMoa1012928](https://doi.org/10.1056/NEJMoa1012928). URL: <http://dx.doi.org/10.1056/NEJMoa1012928>.
- [74] Andrew Clark, Mark Jit, et al. "Global, regional, and national estimates of the population at increased risk of severe COVID-19 due to underlying health conditions in 2020: a modelling study." In: *The Lancet Global Health* 8.8 (Aug. 2020), e1003–e1017. ISSN: 2214-109X. DOI: [10.1016/S2214-109X\(20\)30264-3](https://doi.org/10.1016/S2214-109X(20)30264-3). URL: [https://doi.org/10.1016/S2214-109X\(20\)30264-3](https://doi.org/10.1016/S2214-109X(20)30264-3).
- [75] Philip T L C Clausen, Frank M Aarestrup, and Ole Lund. "Rapid and precise alignment of raw reads against redundant databases with KMA." en. In: *BMC Bioinformatics* 19.1 (Aug. 2018), p. 307. ISSN: 1471-2105. DOI: [10.1186/s12859-018-2336-6](https://doi.org/10.1186/s12859-018-2336-6). URL: <http://dx.doi.org/10.1186/s12859-018-2336-6>.
- [76] Jean-Pierre Claverys, Marc Prudhomme, and Bernard Martin. "Induction of competence regulons as a general response to stress in gram-positive bacteria." en. In: *Annu. Rev. Microbiol.* 60 (2006), pp. 451–475. ISSN: 0066-4227. DOI: [10.1146/annurev.micro.60.080805.142139](https://doi.org/10.1146/annurev.micro.60.080805.142139). URL: <http://dx.doi.org/10.1146/annurev.micro.60.080805.142139>.
- [77] 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." en. In: *J. R. Soc. Interface* 14.133 (Aug. 2017). ISSN: 1742-5689, 1742-5662. DOI: [10.1098/rsif.2017.0295](https://doi.org/10.1098/rsif.2017.0295). URL: <http://dx.doi.org/10.1098/rsif.2017.0295>.
- [78] Ileana Cochetti, Manuela Vecchi, Marina Mingoia, Emily Tili, Maria R Catania, Aldo Manzin, Pietro E Varaldo, and Maria Pia Montanari. "Molecular characterization of pneumococci with efflux-mediated erythromycin resistance and identification of a novel *mef* gene subclass, *mef* (I)." In: *Antimicrob. Agents Chemother.* 49.12 (2005), pp. 4999–5006. ISSN: 0066-4804. URL: <https://aac.asm.org/content/49/12/4999.short>.
- [79] S T Cole, R Brosch, et al. "Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence." en. In: *Nature* 393.6685 (June 1998), pp. 537–544. ISSN: 0028-0836. DOI: [10.1038/31159](https://doi.org/10.1038/31159). URL: <http://dx.doi.org/10.1038/31159>.
- [80] 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*?" en. In: *J. R. Soc. Interface* 7.47 (June 2010), pp. 905–919. ISSN: 1742-5689, 1742-5662. DOI: [10.1098/rsif.2010.0295](https://doi.org/10.1098/rsif.2010.0295).

- rsif.2009.0400. URL: <http://dx.doi.org/10.1098/rsif.2009.0400>.
- [81] Caroline Colijn, Jukka Corander, and Nicholas J Croucher. “Designing ecologically optimized pneumococcal vaccines using population genomics.” en. In: *Nat Microbiol* 5.3 (Mar. 2020), pp. 473–485. ISSN: 2058-5276. DOI: [10.1038/s41564-019-0651-y](https://doi.org/10.1038/s41564-019-0651-y). URL: <http://dx.doi.org/10.1038/s41564-019-0651-y>.
- [82] Caitlin Collins and Xavier Didelot. “A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination.” en. In: *PLoS Comput. Biol.* 14.2 (Feb. 2018), e1005958. ISSN: 1553-734X, 1553-7358. DOI: [10.1371/journal.pcbi.1005958](https://doi.org/10.1371/journal.pcbi.1005958). URL: <http://dx.doi.org/10.1371/journal.pcbi.1005958>.
- [83] R Eric Collins and Paul G Higgs. “Testing the Infinitely Many Genes Model for the Evolution of the Bacterial Core Genome and Pangenome.” In: *Mol. Biol. Evol.* 29.11 (Nov. 2012), pp. 3413–3425. ISSN: 0737-4038. DOI: [10.1093/molbev/mss163](https://doi.org/10.1093/molbev/mss163). URL: <https://academic.oup.com/mbe/article-pdf/29/11/3413/13648372/mss163.pdf>.
- [84] David Cooper, Xinhong Yu, Mohinder Sidhu, Moon H Nahm, Philip Fernsten, and Kathrin U Jansen. “The 13-valent pneumococcal conjugate vaccine (PCV13) elicits cross-functional opsonophagocytic killing responses in humans to *Streptococcus pneumoniae* serotypes 6C and 7A.” en. In: *Vaccine* 29.41 (Sept. 2011), pp. 7207–7211. ISSN: 0264-410X, 1873-2518. DOI: [10.1016/j.vaccine.2011.06.056](https://doi.org/10.1016/j.vaccine.2011.06.056). URL: <http://dx.doi.org/10.1016/j.vaccine.2011.06.056>.
- [85] Jukka Corander, Nicholas Croucher, Simon Harris, John Lees, and Gerry Tonkin-Hill. *Handbook of Statistical Genomics*. Ed. by David Balding, Ida Moltke, and John Marioni. 1st ed. Vol. 162. Wiley, July 2019, pp. 997–1020. ISBN: 9781119429142, 9781119487845. DOI: [10.1002/9781119487845.ch36](https://doi.org/10.1002/9781119487845.ch36). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119487845.ch36>.
- [86] Jukka Corander, Christophe Fraser, Michael U Gutmann, Brian Arnold, William P Hanage, Stephen D Bentley, Marc Lipsitch, and Nicholas J Croucher. “Frequency-dependent selection in vaccine-associated pneumococcal population dynamics.” en. In: *Nature Ecology & Evolution* (Oct. 2017), p. 1. ISSN: 2397-334X, 2397-334X. DOI: [10.1038/s41559-017-0337-x](https://doi.org/10.1038/s41559-017-0337-x). URL: <https://www.nature.com/articles/s41559-017-0337-x>.
- [87] Jukka Corander and Pekka Marttinen. “Bayesian identification of admixture events using multilocus molecular markers.” en. In: *Mol. Ecol.* 15.10 (Sept. 2006), pp. 2833–2843. ISSN: 0962-1083. DOI: [10.1111/j.1365-294X.2006.02994.x](https://doi.org/10.1111/j.1365-294X.2006.02994.x). URL: <http://dx.doi.org/10.1111/j.1365-294X.2006.02994.x>.

- [88] Jukka Corander, Pekka Marttinen, Jukka Sirén, and Jing Tang. “Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations.” en. In: *BMC Bioinformatics* 9 (Dec. 2008), p. 539. ISSN: 1471-2105. DOI: [10.1186/1471-2105-9-539](https://doi.org/10.1186/1471-2105-9-539). URL: <http://dx.doi.org/10.1186/1471-2105-9-539>.
- [89] Jukka Corander, Patrik Waldmann, Pekka Marttinen, and Mikko J Sillanpää. “BAPS 2: enhanced possibilities for the analysis of genetic population structure.” en. In: *Bioinformatics* 20.15 (Oct. 2004), pp. 2363–2369. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/bth250](https://doi.org/10.1093/bioinformatics/bth250). URL: <http://dx.doi.org/10.1093/bioinformatics/bth250>.
- [90] Jukka Corander, Patrik Waldmann, and Mikko J Sillanpää. “Bayesian analysis of genetic differentiation between populations.” en. In: *Genetics* 163.1 (Jan. 2003), pp. 367–374. ISSN: 0016-6731. URL: <https://www.ncbi.nlm.nih.gov/pubmed/12586722>.
- [91] Heather J Cordell, Jamie Benthall, et al. “Genome-wide association study of multiple congenital heart disease phenotypes identifies a susceptibility locus for atrial septal defect at chromosome 4p16.” en. In: *Nat. Genet.* 45.7 (July 2013), pp. 822–824. ISSN: 1061-4036, 1546-1718. DOI: [10.1038/ng.2637](https://doi.org/10.1038/ng.2637). URL: <http://dx.doi.org/10.1038/ng.2637>.
- [92] Maura Costello, Mark Fleharty, et al. “Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms.” en. In: *BMC Genomics* 19.1 (May 2018), p. 332. ISSN: 1471-2164. DOI: [10.1186/s12864-018-4703-0](https://doi.org/10.1186/s12864-018-4703-0). URL: <http://dx.doi.org/10.1186/s12864-018-4703-0>.
- [93] Alejandro Couce and Olivier Tenaillon. “Mutation bias and GC content shape antimutator invasions.” en. In: *Nat. Commun.* 10.1 (July 2019), p. 3114. ISSN: 2041-1723. DOI: [10.1038/s41467-019-11217-6](https://doi.org/10.1038/s41467-019-11217-6). URL: <http://dx.doi.org/10.1038/s41467-019-11217-6>.
- [94] Oliver M Crook, Laurent Gatto, and Paul D W Kirk. “Fast approximate inference for variable selection in Dirichlet process mixtures, with an application to pan-cancer proteomics.” In: (Oct. 2018). arXiv: [1810.05450](https://arxiv.org/abs/1810.05450) [stat.ME]. URL: <http://arxiv.org/abs/1810.05450>.
- [95] Nicholas J Croucher. “Functional and evolutionary analyses of pneumococcal genome variation.” PhD thesis. University of Cambridge, Aug. 2011. URL: <ftp://ftp.sanger.ac.uk/pub/resources/theses/nc3/abstract.pdf>.
- [96] Nicholas J Croucher, Joseph J Campo, Timothy Q Le, Xiaowu Liang, Stephen D Bentley, William P Hanage, and Marc Lipsitch. “Diverse evolutionary patterns of pneumococcal anti-

- gens identified by pangenome-wide immunological screening." en. In: *Proc. Natl. Acad. Sci. U. S. A.* 114.3 (Jan. 2017), E357–E366. ISSN: 0027-8424, 1091-6490. DOI: [10.1073/pnas.1613937114](https://doi.org/10.1073/pnas.1613937114). URL: <http://dx.doi.org/10.1073/pnas.1613937114>.
- [97] Nicholas J Croucher, Claire Chewapreecha, et al. "Evidence for soft selective sweeps in the evolution of pneumococcal multidrug resistance and vaccine escape." en. In: *Genome Biol. Evol.* 6.7 (June 2014), pp. 1589–1602. ISSN: 1759-6653. DOI: [10.1093/gbe/evu120](https://doi.org/10.1093/gbe/evu120). URL: <http://dx.doi.org/10.1093/gbe/evu120>.
- [98] 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." en. In: *Nat. Commun.* 5 (Nov. 2014), p. 5471. ISSN: 2041-1723. DOI: [10.1038/ncomms6471](https://doi.org/10.1038/ncomms6471). URL: <http://dx.doi.org/10.1038/ncomms6471>.
- [99] Nicholas J Croucher and Xavier Didelot. "The application of genomics to tracing bacterial pathogen transmission." en. In: *Curr. Opin. Microbiol.* 23 (Feb. 2015), pp. 62–67. ISSN: 1369-5274, 1879-0364. DOI: [10.1016/j.mib.2014.11.004](https://doi.org/10.1016/j.mib.2014.11.004). URL: <http://dx.doi.org/10.1016/j.mib.2014.11.004>.
- [100] 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." en. In: *Nat. Genet.* 45.6 (June 2013), pp. 656–663. ISSN: 1061-4036, 1546-1718. DOI: [10.1038/ng.2625](https://doi.org/10.1038/ng.2625). URL: <http://dx.doi.org/10.1038/ng.2625>.
- [101] Nicholas J Croucher, Simon R Harris, Yonatan H Grad, and William P Hanage. "Bacterial genomes in epidemiology—present and future." In: *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 368.1614 (Mar. 2013), p. 20120202. ISSN: 0962-8436. DOI: [10.1098/rstb.2012.0202](https://doi.org/10.1098/rstb.2012.0202). URL: <https://doi.org/10.1098/rstb.2012.0202>.
- [102] 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." en. In: *PLoS Genet.* 11.3 (Mar. 2015), e1005095. ISSN: 1553-7390, 1553-7404. DOI: [10.1371/journal.pgen.1005095](https://doi.org/10.1371/journal.pgen.1005095). URL: <http://dx.doi.org/10.1371/journal.pgen.1005095>.
- [103] Nicholas J Croucher, Andrea M Mitchell, et al. *Dominant Role of Nucleotide Substitution in the Diversification of Serotype 3 Pneumococci over Decades and during a Single Infection*. 2013. DOI: [10.1371/journal.pgen.1003868](https://doi.org/10.1371/journal.pgen.1003868). URL: <http://dx.doi.org/10.1371/journal.pgen.1003868>.

- [104] 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." en. In: *Nucleic Acids Res.* (Nov. 2014), gku1196. ISSN: 0305-1048, 1362-4962. DOI: [10.1093/nar/gku1196](https://doi.org/10.1093/nar/gku1196). URL: <http://nar.oxfordjournals.org/content/early/2014/11/20/nar.gku1196>.
- [105] Nicholas J Croucher, Danielle Walker, et al. "Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone *Streptococcus pneumoniae* Spain23F ST81." en. In: *J. Bacteriol.* 191.5 (Mar. 2009), pp. 1480–1489. ISSN: 0021-9193, 1098-5530. DOI: [10.1128/JB.01343-08](https://doi.org/10.1128/JB.01343-08). URL: <http://dx.doi.org/10.1128/JB.01343-08>.
- [106] D R Cundell, N P Gerard, C Gerard, I Idanpaan-Heikkila, and E I Tuomanen. "*Streptococcus pneumoniae* anchor to activated human cells by the receptor for platelet-activating factor." en. In: *Nature* 377.6548 (Oct. 1995), pp. 435–438. ISSN: 0028-0836. DOI: [10.1038/377435a0](https://doi.org/10.1038/377435a0). URL: <http://dx.doi.org/10.1038/377435a0>.
- [107] Ron Dagan, Galia Barkai, Noga Givon-Lavi, Amir Z Sharf, Daniel Vardy, Ted Cohen, Marc Lipsitch, and David Greenberg. "Seasonality of antibiotic-resistant *Streptococcus pneumoniae* that causes acute otitis media: a clue for an antibiotic-restriction policy?" en. In: *J. Infect. Dis.* 197.8 (Apr. 2008), pp. 1094–1102. ISSN: 0022-1899. DOI: [10.1086/528995](https://doi.org/10.1086/528995). URL: <http://dx.doi.org/10.1086/528995>.
- [108] Adilia Dagkessamanskaia, Miriam Moscoso, Vincent Hénard, Sébastien Guiral, Karin Overweg, Mark Reuter, Bernard Martin, Jerry Wells, and Jean-Pierre Claverys. "Interconnection of competence, stress and CiaR regulons in *Streptococcus pneumoniae*: competence triggers stationary phase autolysis of ciaR mutant cells." en. In: *Mol. Microbiol.* 51.4 (Feb. 2004), pp. 1071–1086. ISSN: 0950-382X. DOI: [10.1111/j.1365-2958.2003.03892.x](https://doi.org/10.1111/j.1365-2958.2003.03892.x). URL: <http://dx.doi.org/10.1111/j.1365-2958.2003.03892.x>.
- [109] A Dang-Van, G Tiraby, J F Acar, W V Shaw, and D H Bouanchaud. "Chloramphenicol resistance in *Streptococcus pneumoniae*: enzymatic acetylation and possible plasmid linkage." en. In: *Antimicrob. Agents Chemother.* 13.4 (Apr. 1978), pp. 577–583. ISSN: 0066-4804. URL: <https://www.ncbi.nlm.nih.gov/pubmed/27138>.
- [110] Ross Davidson, Rodrigo Cavalcanti, James L Brunton, Darrin J Bast, Joyce C S de Azavedo, Pamela Kibsey, Christine Fleming, and Donald E Low. "Resistance to levofloxacin and failure

- of treatment of pneumococcal pneumonia." en. In: *N. Engl. J. Med.* 346.10 (Mar. 2002), pp. 747–750. ISSN: 0028-4793, 1533-4406. DOI: [10.1056/NEJMoa012122](https://doi.org/10.1056/NEJMoa012122). URL: <http://dx.doi.org/10.1056/NEJMoa012122>.
- [111] Nicholas G Davies, Stefan Flasche, Mark Jit, and Katherine E Atkins. "Within-host dynamics shape antibiotic resistance in commensal bacteria." en. In: *Nature Ecology & Evolution* (Feb. 2019), p. 1. ISSN: 2397-334X, 2397-334X. DOI: [10.1038/s41559-018-0786-x](https://doi.org/10.1038/s41559-018-0786-x). URL: <https://www.nature.com/articles/s41559-018-0786-x>.
- [112] Nicola De Maio, Dominik Schrempf, and Carolin Kosiol. "PoMo: An Allele Frequency-Based Approach for Species Tree Estimation." en. In: *Syst. Biol.* 64.6 (Nov. 2015), pp. 1018–1031. ISSN: 1063-5157, 1076-836X. DOI: [10.1093/sysbio/syv048](https://doi.org/10.1093/sysbio/syv048). URL: <http://dx.doi.org/10.1093/sysbio/syv048>.
- [113] Nicola De Maio, Conor Walker, and Nick Goldman. *Issues with SARS-CoV-2 sequencing data*. May 2020. URL: https://github.com/W-L/ProblematicSites_SARS-CoV2.
- [114] Nicola De Maio, Colin J Worby, Daniel J Wilson, and Nicole Stoesser. "Bayesian reconstruction of transmission within outbreaks using genomic variants." en. In: *PLoS Comput. Biol.* 14.4 (Apr. 2018), e1006117. ISSN: 1553-734X, 1553-7358. DOI: [10.1371/journal.pcbi.1006117](https://doi.org/10.1371/journal.pcbi.1006117). URL: <http://dx.doi.org/10.1371/journal.pcbi.1006117>.
- [115] Nicola De Maio, Chieh-Hsi Wu, and Daniel J Wilson. "SCOTTI: Efficient Reconstruction of Transmission within Outbreaks with the Structured Coalescent." en. In: *PLoS Comput. Biol.* 12.9 (Sept. 2016), e1005130. ISSN: 1553-734X, 1553-7358. DOI: [10.1371/journal.pcbi.1005130](https://doi.org/10.1371/journal.pcbi.1005130). URL: <http://dx.doi.org/10.1371/journal.pcbi.1005130>.
- [116] Hilary K DeBardeleben, Elena S Lysenko, Ankur B Dalia, and Jeffrey N Weiser. "Tolerance of a phage element by *Streptococcus pneumoniae* leads to a fitness defect during colonization." en. In: *J. Bacteriol.* 196.14 (July 2014), pp. 2670–2680. ISSN: 0021-9193, 1098-5530. DOI: [10.1128/JB.01556-14](https://doi.org/10.1128/JB.01556-14). URL: <http://dx.doi.org/10.1128/JB.01556-14>.
- [117] Hilary K DeBardeleben, Elena S Lysenko, Ankur B Dalia, and Jeffrey N Weiser. "Tolerance of a phage element by *Streptococcus pneumoniae* leads to a fitness defect during colonization." en. In: *J. Bacteriol.* 196.14 (July 2014), pp. 2670–2680. ISSN: 0021-9193, 1098-5530. DOI: [10.1128/JB.01556-14](https://doi.org/10.1128/JB.01556-14). URL: <http://dx.doi.org/10.1128/JB.01556-14>.

- [118] A L Delcher, D Harmon, S Kasif, O White, and S L Salzberg. "Improved microbial gene identification with GLIMMER." en. In: *Nucleic Acids Res.* 27.23 (Dec. 1999), pp. 4636–4641. ISSN: 0305-1048. DOI: [10.1093/nar/27.23.4636](https://doi.org/10.1093/nar/27.23.4636). URL: <http://dx.doi.org/10.1093/nar/27.23.4636>.
- [119] Arthur L Delcher, Kirsten A Bratke, Edwin C Powers, and Steven L Salzberg. "Identifying bacterial genes and endosymbiont DNA with Glimmer." en. In: *Bioinformatics* 23.6 (Mar. 2007), pp. 673–679. ISSN: 1367-4803, 1367-4811. DOI: [10.1093/bioinformatics/btm009](https://doi.org/10.1093/bioinformatics/btm009). URL: <http://dx.doi.org/10.1093/bioinformatics/btm009>.
- [120] Arthur L Delcher, Adam Phillippy, Jane Carlton, and Steven L Salzberg. "Fast algorithms for large-scale genome alignment and comparison." en. In: *Nucleic Acids Res.* 30.11 (June 2002), pp. 2478–2483. ISSN: 0305-1048, 1362-4962. DOI: [10.1093/nar/30.11.2478](https://doi.org/10.1093/nar/30.11.2478). URL: <http://nar.oxfordjournals.org/content/30/11/2478>.
- [121] Susan John Dennis D. Cox. "SDO: A Statistical Method for Global Optimization." In: *Multidisciplinary Design Optimization: State-of-the-Art*. SIAM, 1997, pp. 315–329. URL: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.52.4169>.
- [122] James F Denton, Jose Lugo-Martinez, Abraham E Tucker, Daniel R Schrider, Wesley C Warren, and Matthew W Hahn. "Extensive error in the number of genes inferred from draft genome assemblies." en. In: *PLoS Comput. Biol.* 10.12 (Dec. 2014), e1003998. ISSN: 1553-734X, 1553-7358. DOI: [10.1371/journal.pcbi.1003998](https://doi.org/10.1371/journal.pcbi.1003998). URL: <http://dx.doi.org/10.1371/journal.pcbi.1003998>.
- [123] Aarti Desai, Veer Singh Marwah, Akshay Yadav, Vineet Jha, Kishor Dhaygude, Ujwala Bangar, Vivek Kulkarni, and Abhay Jere. "Identification of optimum sequencing depth especially for de novo genome assembly of small genomes using next generation sequencing data." en. In: *PLoS One* 8.4 (Apr. 2013), e60204. ISSN: 1932-6203. DOI: [10.1371/journal.pone.0060204](https://doi.org/10.1371/journal.pone.0060204). URL: <http://dx.doi.org/10.1371/journal.pone.0060204>.
- [124] Salvatore Di Giorgio, Filippo Martignano, Maria Gabriella Torcia, Giorgio Mattiuz, and Silvestro G Conticello. "Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2." en. In: *Science Advances* 6.25 (June 2020), eabb5813. ISSN: 2375-2548. DOI: [10.1126/sciadv.abb5813](https://doi.org/10.1126/sciadv.abb5813). URL: <https://advances.sciencemag.org/content/6/25/eabb5813>.
- [125] Alex Diaz-Papkovich, Luke Anderson-Trocme, and Simon Gravel. "Revealing multi-scale population structure in large cohorts." en. Sept. 2018. URL: <https://www.biorxiv.org/content/early/2018/09/23/423632>.

- [126] Xavier Didelot, Nicholas J Croucher, Stephen D Bentley, Simon R Harris, and Daniel J Wilson. "Bayesian inference of ancestral dates on bacterial phylogenetic trees." en. In: *Nucleic Acids Res.* 46.22 (Dec. 2018), e134. ISSN: 0305-1048, 1362-4962. DOI: [10.1093/nar/gky783](https://doi.org/10.1093/nar/gky783). URL: <http://dx.doi.org/10.1093/nar/gky783>.
- [127] Xavier Didelot and Daniel Falush. "Inference of bacterial microevolution using multilocus sequence data." en. In: *Genetics* 175.3 (Mar. 2007), pp. 1251–1266. ISSN: 0016-6731. DOI: [10.1534/genetics.106.063305](https://doi.org/10.1534/genetics.106.063305). URL: <http://dx.doi.org/10.1534/genetics.106.063305>.
- [128] Xavier Didelot, Christophe Fraser, Jennifer Gardy, and Caroline Colijn. "Genomic Infectious Disease Epidemiology in Partially Sampled and Ongoing Outbreaks." en. In: *Mol. Biol. Evol.* 34.4 (Apr. 2017), pp. 997–1007. ISSN: 0737-4038, 1537-1719. DOI: [10.1093/molbev/msw275](https://doi.org/10.1093/molbev/msw275). URL: <http://dx.doi.org/10.1093/molbev/msw275>.
- [129] Xavier Didelot, Jennifer Gardy, and Caroline Colijn. "Bayesian inference of infectious disease transmission from whole-genome sequence data." en. In: *Mol. Biol. Evol.* 31.7 (July 2014), pp. 1869–1879. ISSN: 0737-4038, 1537-1719. DOI: [10.1093/molbev/msu121](https://doi.org/10.1093/molbev/msu121). URL: <http://dx.doi.org/10.1093/molbev/msu121>.
- [130] Xavier Didelot, A Sarah Walker, Tim E Peto, Derrick W Crook, and Daniel J Wilson. "Within-host evolution of bacterial pathogens." en. In: *Nat. Rev. Microbiol.* 14.3 (Mar. 2016), pp. 150–162. ISSN: 1740-1526, 1740-1534. DOI: [10.1038/nrmicro.2015.13](https://doi.org/10.1038/nrmicro.2015.13). URL: <http://dx.doi.org/10.1038/nrmicro.2015.13>.
- [131] Xavier Didelot and Daniel J Wilson. "ClonalFrameML: efficient inference of recombination in whole bacterial genomes." en. In: *PLoS Comput. Biol.* 11.2 (Feb. 2015), e1004041. ISSN: 1553-734X, 1553-7358. DOI: [10.1371/journal.pcbi.1004041](https://doi.org/10.1371/journal.pcbi.1004041). URL: <http://dx.doi.org/10.1371/journal.pcbi.1004041>.
- [132] Marcus M Dillon, Way Sung, Michael Lynch, and Vaughn S Cooper. "Periodic Variation of Mutation Rates in Bacterial Genomes Associated with Replication Timing." en. In: *MBio* 9.4 (Aug. 2018). ISSN: 2150-7511. DOI: [10.1128/mBio.01371-18](https://doi.org/10.1128/mBio.01371-18). URL: <http://dx.doi.org/10.1128/mBio.01371-18>.
- [133] Wei Ding, Franz Baumdicker, and Richard A Neher. "panX: pan-genome analysis and exploration." en. In: *Nucleic Acids Res.* 46.1 (Jan. 2018), e5. ISSN: 0305-1048, 1362-4962. DOI: [10.1093/nar/gkx977](https://doi.org/10.1093/nar/gkx977). URL: <http://dx.doi.org/10.1093/nar/gkx977>.

- [134] Svetolik Djurkovic, Jutta M Loeffler, and Vincent A Fischetti. "Synergistic killing of *Streptococcus pneumoniae* with the bacteriophage lytic enzyme Cpl-1 and penicillin or gentamicin depends on the level of penicillin resistance." en. In: *Antimicrob. Agents Chemother.* 49.3 (Mar. 2005), pp. 1225–1228. ISSN: 0066-4804. DOI: [10.1128/AAC.49.3.1225-1228.2005](https://doi.org/10.1128/AAC.49.3.1225-1228.2005). URL: <http://dx.doi.org/10.1128/AAC.49.3.1225-1228.2005>.
- [135] A R Dochez and L J Gillespie. "A biologic classification of pneumococci by means of immunity reactions." In: *JAMA* 61.10 (Sept. 1913), pp. 727–732. ISSN: 0098-7484, 0002-9955. DOI: [10.1001/jama.1913.04350100005003](https://doi.org/10.1001/jama.1913.04350100005003). URL: <https://jamanetwork.com/journals/jama/article-abstract/215899>.
- [136] Claudio Donati, N Luisa Hiller, et al. "Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species." en. In: *Genome Biol.* 11.10 (Oct. 2010), R107. ISSN: 1465-6906. DOI: [10.1186/gb-2010-11-10-r107](https://doi.org/10.1186/gb-2010-11-10-r107). URL: <http://dx.doi.org/10.1186/gb-2010-11-10-r107>.
- [137] Gabriel Ribeiro Dos Santos, Lin Wang, Derek A T Cummings, Andrew S Azman, Juliette Paireau, Arnaud Fontanet, Simon Cauchemez, and Henrik Salje. "Age-specific mortality and immunity patterns of SARS-CoV-2 infection in 45 countries." en. In: *medRxiv* (Aug. 2020), p. 2020.08.24.20180851. DOI: [10.1101/2020.08.24.20180851](https://doi.org/10.1101/2020.08.24.20180851). URL: <https://www.medrxiv.org/content/10.1101/2020.08.24.20180851v1.abstract>.
- [138] R M Douglas, J C Paton, S J Duncan, and D J Hansman. "Antibody response to pneumococcal vaccination in children younger than five years of age." en. In: *J. Infect. Dis.* 148.1 (July 1983), pp. 131–137. ISSN: 0022-1899. DOI: [10.1093/infdis/148.1.131](https://doi.org/10.1093/infdis/148.1.131). URL: <http://dx.doi.org/10.1093/infdis/148.1.131>.
- [139] A J Drummond and A Rambaut. "BEAST: Bayesian evolutionary analysis by sampling trees." en. In: *BMC Evol. Biol.* 7 (Nov. 2007), p. 214. ISSN: 1471-2148. DOI: [10.1186/1471-2148-7-214](https://doi.org/10.1186/1471-2148-7-214). URL: <http://dx.doi.org/10.1186/1471-2148-7-214>.
- [140] A J Drummond, A Rambaut, B Shapiro, and O G Pybus. "Bayesian coalescent inference of past population dynamics from molecular sequences." en. In: *Mol. Biol. Evol.* 22.5 (May 2005), pp. 1185–1192. ISSN: 0737-4038. DOI: [10.1093/molbev/msi103](https://doi.org/10.1093/molbev/msi103). URL: <http://dx.doi.org/10.1093/molbev/msi103>.
- [141] Gytis Dudas, Luiz Max Carvalho, et al. "Virus genomes reveal factors that spread and sustained the Ebola epidemic." en. In: *Nature* 544.7650 (Apr. 2017), pp. 309–315. ISSN: 0028-0836, 1476-4687. DOI: [10.1038/nature22040](https://doi.org/10.1038/nature22040). URL: <http://dx.doi.org/10.1038/nature22040>.

- [142] B M Duggar. "Aureomycin; a product of the continuing search for new antibiotics." en. In: *Ann. N. Y. Acad. Sci.* 51.Art. 2 (Nov. 1948), pp. 177–181. ISSN: 0077-8923. DOI: [10.1111/j.1749-6632.1948.tb27262.x](https://doi.org/10.1111/j.1749-6632.1948.tb27262.x). URL: <http://dx.doi.org/10.1111/j.1749-6632.1948.tb27262.x>.
- [143] "E pluribus unum." en. In: *Nat. Methods* 7.5 (May 2010), p. 331. ISSN: 1548-7091, 1548-7105. DOI: [10.1038/nmeth0510-331](https://doi.org/10.1038/nmeth0510-331). URL: <http://dx.doi.org/10.1038/nmeth0510-331>.
- [144] Sarah G Earle, Chieh-Hsi Wu, et al. "Identifying lineage effects when controlling for population structure improves power in bacterial association studies." en. In: *Nat Microbiol* 1 (Apr. 2016), p. 16041. ISSN: 2058-5276. DOI: [10.1038/nmicrobiol.2016.41](https://doi.org/10.1038/nmicrobiol.2016.41). URL: <http://dx.doi.org/10.1038/nmicrobiol.2016.41>.
- [145] John Ehrlich, Quentin R Bartz, Robert M Smith, Dwight A Joslyn, and Others. "Chloromyetin, a New Antibiotic from a Soil Actinomycete." In: *Science* (1947). ISSN: 0193-4511. URL: <https://www.cabdirect.org/cabdirect/abstract/19482901354>.
- [146] K Eiglmeier, J Parkhill, et al. "The decaying genome of *Mycobacterium leprae*." en. In: *Lepr. Rev.* 72.4 (Dec. 2001), pp. 387–398. ISSN: 0305-7518. URL: <https://www.ncbi.nlm.nih.gov/pubmed/11826475>.
- [147] Vegard Eldholm, Ola Johnsborg, Kristine Haugen, Hilde Solheim Ohnstad, and Leiv Sigve Håvarstein. "Fratricide in *Streptococcus pneumoniae*: contributions and role of the cell wall hydrolases CbpD, LytA and LytC." en. In: *Microbiology* 155.Pt 7 (July 2009), pp. 2223–2234. ISSN: 0026-2617, 1350-0872. DOI: [10.1099/mic.0.026328-0](https://doi.org/10.1099/mic.0.026328-0). URL: <http://dx.doi.org/10.1099/mic.0.026328-0>.
- [148] Vegard Eldholm, Gunnstein Norheim, Bent von der Lippe, Wibeke Kinander, Ulf R Dahle, Dominique A Caugant, Turid Mannsåker, Anne Torunn Mengshoel, Anne Ma Dyrhol-Riise, and Francois Balloux. "Evolution of extensively drug-resistant *Mycobacterium tuberculosis* from a susceptible ancestor in a single patient." en. In: *Genome Biol.* 15.11 (Nov. 2014), p. 490. ISSN: 1465-6906. DOI: [10.1186/s13059-014-0490-3](https://doi.org/10.1186/s13059-014-0490-3). URL: <http://dx.doi.org/10.1186/s13059-014-0490-3>.
- [149] A M Emmerson and A M Jones. "The quinolones: decades of development and use." In: *J. Antimicrob. Chemother.* 51.suppl_1 (May 2003), pp. 13–20. ISSN: 0305-7453. DOI: [10.1093/jac/dkg208](https://doi.org/10.1093/jac/dkg208). URL: https://academic.oup.com/jac/article-abstract/51/suppl_1/13/2473462.
- [150] V I Enne, D M Livermore, P Stephens, and L M Hall. "Persistence of sulphonamide resistance in *Escherichia coli* in the UK despite national prescribing restriction." en. In: *Lancet* 357.9265

- (Apr. 2001), pp. 1325–1328. ISSN: 0140-6736. DOI: [10.1016/S0140-6736\(00\)04519-0](https://doi.org/10.1016/S0140-6736(00)04519-0). URL: [http://dx.doi.org/10.1016/S0140-6736\(00\)04519-0](http://dx.doi.org/10.1016/S0140-6736(00)04519-0).
- [151] A J Enright, S Van Dongen, and C A Ouzounis. “An efficient algorithm for large-scale detection of protein families.” en. In: *Nucleic Acids Res.* 30.7 (Apr. 2002), pp. 1575–1584. ISSN: 0305-1048, 1362-4962. DOI: [10.1093/nar/30.7.1575](https://doi.org/10.1093/nar/30.7.1575). URL: <http://www.ncbi.nlm.nih.gov/pubmed/11917018>.
- [152] Lennard Epping, Andries J van Tonder, Rebecca A Gladstone, The Global Pneumococcal Sequencing Consortium, Stephen D Bentley, Andrew J Page, and Jacqueline A Keane. “SeroBA: rapid high-throughput serotyping of *Streptococcus pneumoniae* from whole genome sequence data.” en. In: *Microb Genom* 4.7 (July 2018). ISSN: 2057-5858. DOI: [10.1099/mgen.0.000186](https://doi.org/10.1099/mgen.0.000186). URL: <http://dx.doi.org/10.1099/mgen.0.000186>.
- [153] Nicholas Eriksson, Shirley Wu, Chuong B Do, Amy K Kiefer, Joyce Y Tung, Joanna L Mountain, David A Hinds, and Uta Francke. “A genetic variant near olfactory receptor genes influences cilantro preference.” In: *Flavour* 1.1 (Nov. 2012), p. 22. ISSN: 2044-7248. DOI: [10.1186/2044-7248-1-22](https://doi.org/10.1186/2044-7248-1-22). URL: <https://doi.org/10.1186/2044-7248-1-22>.
- [154] European Centre for Disease Prevention and Control. *European Centre for Disease Prevention and Control (2015) Antimicrobial Resistance Surveillance in Europe 2015*. Tech. rep. European Centre for Disease Prevention and Control, 2015. URL: <https://www.ecdc.europa.eu/sites/default/files/media/en/publications/Publications/antimicrobial-resistance-europe-2015.pdf>.
- [155] G Evanno, S Regnaut, and J Goudet. “Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study.” In: *Mol. Ecol.* 14.8 (July 2005), pp. 2611–2620. ISSN: 0962-1083. DOI: [10.1111/j.1365-294X.2005.02553.x](https://doi.org/10.1111/j.1365-294X.2005.02553.x). URL: <http://dx.doi.org/10.1111/j.1365-294X.2005.02553.x>.
- [156] W Evans and D Hansman. “Tetracycline-resistant pneumococcus.” In: *Lancet* 281.7278 (Feb. 1963), p. 451. ISSN: 0140-6736. DOI: [10.1016/S0140-6736\(63\)92350-X](https://doi.org/10.1016/S0140-6736(63)92350-X). URL: [http://dx.doi.org/10.1016/S0140-6736\(63\)92350-X](http://dx.doi.org/10.1016/S0140-6736(63)92350-X).
- [157] J W H Eyre and J W Washbourn. “Further Experiments With Pane’s Antipneumococcus Serum.” In: *Br. Med. J.* 2.2027 (1899), pp. 1247–1249. ISSN: 0007-1447. URL: <http://www.jstor.org/stable/20262354>.

- [158] Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. "Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies." en. In: *Genetics* 164.4 (Aug. 2003), pp. 1567–1587. ISSN: 0016-6731. URL: <https://www.ncbi.nlm.nih.gov/pubmed/12930761>.
- [159] Daniel Falush, Thierry Wirth, et al. "Traces of human migrations in *Helicobacter pylori* populations." en. In: *Science* 299.5612 (Mar. 2003), pp. 1582–1585. ISSN: 0036-8075, 1095-9203. DOI: [10.1126/science.1080857](https://doi.org/10.1126/science.1080857). URL: <http://dx.doi.org/10.1126/science.1080857>.
- [160] N R Faria, J Quick, et al. "Establishment and cryptic transmission of Zika virus in Brazil and the Americas." en. In: *Nature* 546.7658 (June 2017), pp. 406–410. ISSN: 0028-0836, 1476-4687. DOI: [10.1038/nature22401](https://doi.org/10.1038/nature22401). URL: <http://dx.doi.org/10.1038/nature22401>.
- [161] James S Farris. "Methods for Computing Wagner Trees." en. In: *Syst. Biol.* 19.1 (Mar. 1970), pp. 83–92. ISSN: 1063-5157. DOI: [10.1093/sysbio/19.1.83](https://doi.org/10.1093/sysbio/19.1.83). URL: <https://academic.oup.com/sysbio/article-abstract/19/1/83/1647484>.
- [162] Joseph R Fauver, Mary E Petrone, et al. "Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States." en. In: *Cell* 181.5 (May 2020), 990–996.e5. ISSN: 0092-8674, 1097-4172. DOI: [10.1016/j.cell.2020.04.021](https://doi.org/10.1016/j.cell.2020.04.021). URL: <http://dx.doi.org/10.1016/j.cell.2020.04.021>.
- [163] C Feldman, R Read, A Rutman, P K Jeffery, A Brain, V Lund, T J Mitchell, P W Andrew, G J Boulnois, and H C Todd. "The interaction of *Streptococcus pneumoniae* with intact human respiratory mucosa in vitro." en. In: *Eur. Respir. J.* 5.5 (May 1992), pp. 576–583. ISSN: 0903-1936. URL: <https://www.ncbi.nlm.nih.gov/pubmed/1612157>.
- [164] Gracia Fellmeth, May May Oo, Billion Lay, and Rose McGready. "Paired suicide in a young refugee couple on the Thai-Myanmar border." en. In: *BMJ Case Rep.* 2016 (Sept. 2016). ISSN: 1757-790X. DOI: [10.1136/bcr-2016-215527](https://doi.org/10.1136/bcr-2016-215527). URL: <http://dx.doi.org/10.1136/bcr-2016-215527>.
- [165] Joseph Felsenstein. "Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates." en. In: *Evolution* 35.6 (Nov. 1981), pp. 1229–1242. ISSN: 0014-3820, 1558-5646. DOI: [10.1111/j.1558-5646.1981.tb04991.x](https://doi.org/10.1111/j.1558-5646.1981.tb04991.x). URL: <http://dx.doi.org/10.1111/j.1558-5646.1981.tb04991.x>.
- [166] Pamela Ferretti, Edoardo Pasolli, et al. "Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome." en. In: *Cell Host Microbe*

- 24.1 (July 2018), 133–145.e5. ISSN: 1931-3128, 1934-6069. DOI: [10.1016/j.chom.2018.06.005](https://doi.org/10.1016/j.chom.2018.06.005). URL: <http://dx.doi.org/10.1016/j.chom.2018.06.005>.
- [167] M Finland, C Garner, C Wilcox, and L D Sabath. “Susceptibility of pneumococci and *Haemophilus influenzae* to antibacterial agents.” en. In: *Antimicrob. Agents Chemother.* 9.2 (Feb. 1976), pp. 274–287. ISSN: 0066-4804. DOI: [10.1128/aac.9.2.274](https://doi.org/10.1128/aac.9.2.274). URL: <http://dx.doi.org/10.1128/aac.9.2.274>.
- [168] Walter M Fitch. “Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology.” en. In: *Syst. Biol.* 20.4 (Dec. 1971), pp. 406–416. ISSN: 1063-5157. DOI: [10.1093/sysbio/20.4.406](https://doi.org/10.1093/sysbio/20.4.406). URL: <https://academic.oup.com/sysbio/article-abstract/20/4/406/1673276>.
- [169] Stefan Flasche, Marc Lipsitch, John Ojal, and Amy Pinsent. “Estimating the contribution of different age strata to vaccine serotype pneumococcal transmission in the pre vaccine era: a modelling study.” en. In: *BMC Med.* 18.1 (June 2020), p. 129. ISSN: 1741-7015. DOI: [10.1186/s12916-020-01601-1](https://doi.org/10.1186/s12916-020-01601-1). URL: <http://dx.doi.org/10.1186/s12916-020-01601-1>.
- [170] R D Fleischmann, M D Adams, O White, R A Clayton, E F Kirkness, A R Kerlavage, C J Bult, J F Tomb, B A Dougherty, and J M Merrick. “Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd.” en. In: *Science* 269.5223 (July 1995), pp. 496–512. ISSN: 0036-8075. DOI: [10.1126/science.7542800](https://doi.org/10.1126/science.7542800). URL: <http://dx.doi.org/10.1126/science.7542800>.
- [171] R D Fleischmann, D Alland, et al. “Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains.” en. In: *J. Bacteriol.* 184.19 (Oct. 2002), pp. 5479–5490. ISSN: 0021-9193. DOI: [10.1128/jb.184.19.5479-5490.2002](https://doi.org/10.1128/jb.184.19.5479-5490.2002). URL: <http://dx.doi.org/10.1128/jb.184.19.5479-5490.2002>.
- [172] Anthony R Florita, Larry J Brackney, Todd P Otanicar, and Jeffrey Robertson. “Classification of Commercial Building Electrical Demand Profiles for Energy Storage Applications.” In: *J. Sol. Energy Eng.* 135.3 (Aug. 2013), p. 031020. ISSN: 0199-6231. DOI: [10.1115/1.4024029](https://doi.org/10.1115/1.4024029). URL: <http://solarenergyengineering.asmedigitalcollection.asme.org/article.aspx?articleid=1697039>.
- [173] Simon A Forbes, David Beare, et al. “COSMIC: exploring the world’s knowledge of somatic mutations in human cancer.” en. In: *Nucleic Acids Res.* 43.Database issue (Jan. 2015), pp. D805–11. ISSN: 0305-1048, 1362-4962. DOI: [10.1093/nar/gku1075](https://doi.org/10.1093/nar/gku1075). URL: <http://dx.doi.org/10.1093/nar/gku1075>.

- [174] L R Foulds and R L Graham. "The steiner problem in phylogeny is NP-complete." In: *Adv. Appl. Math.* 3.1 (Mar. 1982), pp. 43–49. ISSN: 0196-8858. DOI: [10.1016/S0196-8858\(82\)80004-3](https://doi.org/10.1016/S0196-8858(82)80004-3). URL: <http://www.sciencedirect.com/science/article/pii/S0196885882800043>.
- [175] Derrick E Fouts, Lauren Brinkac, Erin Beck, Jason Inman, and Granger Sutton. "PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species." en. In: *Nucleic Acids Res.* 40.22 (Dec. 2012), e172–e172. ISSN: 0305-1048. DOI: [10.1093/nar/gks757](https://doi.org/10.1093/nar/gks757). URL: <https://academic.oup.com/nar/article-pdf/40/22/e172/25347099/gks757.pdf>.
- [176] E B Fowlkes and C L Mallows. "A Method for Comparing Two Hierarchical Clusterings." In: *J. Am. Stat. Assoc.* 78.383 (Sept. 1983), pp. 553–569. ISSN: 0162-1459. DOI: [10.1080/01621459.1983.10478008](https://doi.org/10.1080/01621459.1983.10478008). URL: <https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1983.10478008>.
- [177] Christophe Fraser, William P Hanage, and Brian G Spratt. "Neutral microepidemic evolution of bacterial pathogens." en. In: *Proc. Natl. Acad. Sci. U. S. A.* 102.6 (Feb. 2005), pp. 1968–1973. ISSN: 0027-8424. DOI: [10.1073/pnas.0406993102](https://doi.org/10.1073/pnas.0406993102). URL: <http://dx.doi.org/10.1073/pnas.0406993102>.
- [178] Julius Fredens, Kaihang Wang, et al. "Total synthesis of *Escherichia coli* with a recoded genome." en. In: *Nature* 569.7757 (May 2019), pp. 514–518. ISSN: 0028-0836, 1476-4687. DOI: [10.1038/s41586-019-1192-5](https://doi.org/10.1038/s41586-019-1192-5). URL: <http://dx.doi.org/10.1038/s41586-019-1192-5>.
- [179] Elizabeth Gladys Aarag Fredheim, Claus Klingenberg, Holger Rohde, Stephanie Frankenberger, Peter Gaustad, Trond Flaegstad, and Johanna Ericson Sollid. "Biofilm formation by *Staphylococcus haemolyticus*." en. In: *J. Clin. Microbiol.* 47.4 (Apr. 2009), pp. 1172–1180. ISSN: 0095-1137, 1098-660X. DOI: [10.1128/JCM.01891-08](https://doi.org/10.1128/JCM.01891-08). URL: <http://dx.doi.org/10.1128/JCM.01891-08>.
- [180] Eric Frichot, François Mathieu, Théo Trouillon, Guillaume Bouchard, and Olivier François. "Fast and efficient estimation of individual ancestry coefficients." en. In: *Genetics* 196.4 (Apr. 2014), pp. 973–983. ISSN: 0016-6731, 1943-2631. DOI: [10.1534/genetics.113.160572](https://doi.org/10.1534/genetics.113.160572). URL: <http://dx.doi.org/10.1534/genetics.113.160572>.
- [181] C Friedlaender. *Ueber die Schizomyceten bei der acuten fibrösen Pneumonie*. 1882. DOI: [10.1007/bf01880516](https://doi.org/10.1007/bf01880516). URL: <http://dx.doi.org/10.1007/bf01880516>.

- [182] Simon D W Frost, Oliver G Pybus, Julia R Gog, Cecile Viboud, Sebastian Bonhoeffer, and Trevor Bedford. "Eight challenges in phylodynamic inference." en. In: *Epidemics* 10 (Mar. 2015), pp. 88–92. ISSN: 1755-4365, 1878-0067. DOI: [10.1016/j.epidem.2014.09.001](https://doi.org/10.1016/j.epidem.2014.09.001). URL: <http://dx.doi.org/10.1016/j.epidem.2014.09.001>.
- [183] Marek Gagolewski, Maciej Bartoszek, and Anna Cena. "Genie: A new, fast, and outlier-resistant hierarchical clustering algorithm." In: *Inf. Sci.* 363 (Oct. 2016), pp. 8–23. ISSN: 0020-0255. DOI: [10.1016/j.ins.2016.05.003](https://doi.org/10.1016/j.ins.2016.05.003). URL: <http://www.sciencedirect.com/science/article/pii/S002002551630319X>.
- [184] Feroze Ganaie, Jamil S Saad, et al. "A New Pneumococcal Capsule Type, 10D, is the 100th Serotype and Has a Large cps Fragment from an Oral Streptococcus." en. In: *MBio* 11.3 (May 2020). ISSN: 2150-7511. DOI: [10.1128/mBio.00937-20](https://doi.org/10.1128/mBio.00937-20). URL: <http://dx.doi.org/10.1128/mBio.00937-20>.
- [185] Guillaume Gautreau, Adelme Bazin, et al. "PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph." en. In: *PLoS Comput. Biol.* 16.3 (Mar. 2020), e1007732. ISSN: 1553-734X, 1553-7358. DOI: [10.1371/journal.pcbi.1007732](https://doi.org/10.1371/journal.pcbi.1007732). URL: <http://dx.doi.org/10.1371/journal.pcbi.1007732>.
- [186] Jemma L Geoghegan, Xiaoyun Ren, et al. "Genomic epidemiology reveals transmission patterns and dynamics of SARS-CoV-2 in Aotearoa New Zealand." en. In: *medRxiv* (Aug. 2020), p. 2020.08.05.20168930. DOI: [10.1101/2020.08.05.20168930](https://doi.org/10.1101/2020.08.05.20168930). URL: <https://www.medrxiv.org/content/10.1101/2020.08.05.20168930v3.abstract>.
- [187] Moritz Gerstung, Elli Papaemmanuil, and Peter J Campbell. "Subclonal variant calling with multiple samples and prior knowledge." en. In: *Bioinformatics* 30.9 (May 2014), pp. 1198–1204. ISSN: 1367-4803, 1367-4811. DOI: [10.1093/bioinformatics/btt750](https://doi.org/10.1093/bioinformatics/btt750). URL: <http://dx.doi.org/10.1093/bioinformatics/btt750>.
- [188] Mahan Ghafari, Casper K Lumby, Daniel B Weissman, and Christopher J R Illingworth. "Inferring transmission bottleneck size from viral sequence data using a novel haplotype reconstruction method." en. In: *J. Virol.* (Apr. 2020). ISSN: 0022-538X, 1098-5514. DOI: [10.1128/JVI.00014-20](https://doi.org/10.1128/JVI.00014-20). URL: <http://dx.doi.org/10.1128/JVI.00014-20>.
- [189] Stephen K Gire, Augustine Goba, et al. "Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak." en. In: *Science* 345.6202 (Sept. 2014), pp. 1369–1372. ISSN: 0036-8075, 1095-9203. DOI: [10.1126/science.1259657](https://doi.org/10.1126/science.1259657). URL: <http://dx.doi.org/10.1126/science.1259657>.

- [190] Rebecca A Gladstone, Stephanie W Lo, et al. "International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact." en. In: *EBioMedicine* 43 (May 2019), pp. 338–346. ISSN: 2352-3964. DOI: [10.1016/j.ebiom.2019.04.021](https://doi.org/10.1016/j.ebiom.2019.04.021). URL: <http://www.ebiomedicine.com/article/S2352396419302592/abstract>.
- [191] S Carvalho Maria da Gloria, Maria Lucia Tondella, et al. "Evaluation and improvement of real-time PCR assays targeting *lytA*, *ply*, and *psaA* genes for detection of pneumococcal DNA." In: *J. Clin. Microbiol.* 45.8 (2007), pp. 2460–2466. ISSN: 0095-1137. URL: <https://jcm.asm.org/content/45/8/2460.short>.
- [192] Maxime Godfroid, Tal Dagan, and Anne Kupczok. "Recombination Signal in Mycobacterium tuberculosis Stems from Reference-guided Assemblies and Alignment Artefacts." en. In: *Genome Biol. Evol.* 10.8 (Aug. 2018), pp. 1920–1926. ISSN: 1759-6653. DOI: [10.1093/gbe/evy143](https://doi.org/10.1093/gbe/evy143). URL: <http://dx.doi.org/10.1093/gbe/evy143>.
- [193] N Goldman and Z Yang. "A codon-based model of nucleotide substitution for protein-coding DNA sequences." en. In: *Mol. Biol. Evol.* 11.5 (Sept. 1994), pp. 725–736. ISSN: 0737-4038. DOI: [10.1093/oxfordjournals.molbev.a040153](https://doi.org/10.1093/oxfordjournals.molbev.a040153). URL: <http://dx.doi.org/10.1093/oxfordjournals.molbev.a040153>.
- [194] Zijian Gong, Wei Lai, Min Liu, Zhengshuang Hua, Yayin Sun, Qingfang Xu, Yue Xia, Yue Zhao, and Xiaoyuan Xie. "Novel Genes Related to Ceftriaxone Resistance Found among Ceftriaxone-Resistant *Neisseria gonorrhoeae* Strains Selected In Vitro." en. In: *Antimicrob. Agents Chemother.* 60.4 (Apr. 2016), pp. 2043–2051. ISSN: 0066-4804, 1098-6596. DOI: [10.1128/AAC.00149-15](https://doi.org/10.1128/AAC.00149-15). URL: <http://dx.doi.org/10.1128/AAC.00149-15>.
- [195] Benjamin H Good, Michael J McDonald, Jeffrey E Barrick, Richard E Lenski, and Michael M Desai. "The dynamics of molecular evolution over 60,000 generations." en. In: *Nature* (Oct. 2017). ISSN: 0028-0836, 1476-4687. DOI: [10.1038/nature24287](https://doi.org/10.1038/nature24287). URL: <http://dx.doi.org/10.1038/nature24287>.
- [196] Yonatan H Grad, Marc Lipsitch, et al. "Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011." en. In: *Proc. Natl. Acad. Sci. U. S. A.* 109.8 (Feb. 2012), pp. 3065–3070. ISSN: 0027-8424, 1091-6490. DOI: [10.1073/pnas.1121491109](https://doi.org/10.1073/pnas.1121491109). URL: <http://dx.doi.org/10.1073/pnas.1121491109>.
- [197] Robert B Gramacy and Herbert K H Lee. "Cases for the nugget in modeling computer experiments." In: *Stat. Comput.* 22.3 (May 2012), pp. 713–722. ISSN: 0960-3174, 1573-1375. DOI: [10.1007/s11222-010-9224-x](https://doi.org/10.1007/s11222-010-9224-x). URL: <https://doi.org/10.1007/s11222-010-9224-x>.

- [198] Robert Gramacy. "laGP: Large-Scale Spatial Modeling via Local Approximate Gaussian Processes in R." In: *Journal of Statistical Software, Articles* 72.1 (2016), pp. 1–46. ISSN: 1548-7660. DOI: [10.18637/jss.v072.i01](https://doi.org/10.18637/jss.v072.i01). URL: <https://www.jstatsoft.org/v072/i01>.
- [199] F Griffith. "The Significance of Pneumococcal Types." en. In: *J. Hyg.* 27.2 (Jan. 1928), pp. 113–159. ISSN: 0022-1724. DOI: [10.1017/s0022172400031879](https://doi.org/10.1017/s0022172400031879). URL: <http://dx.doi.org/10.1017/s0022172400031879>.
- [200] Sébastien Guiral, Tim J Mitchell, Bernard Martin, and Jean-Pierre Claverys. "Competence-programmed predation of non-competent cells in the human pathogen *Streptococcus pneumoniae*: genetic requirements." en. In: *Proc. Natl. Acad. Sci. U. S. A.* 102.24 (June 2005), pp. 8710–8715. ISSN: 0027-8424. DOI: [10.1073/pnas.0500879102](https://doi.org/10.1073/pnas.0500879102). URL: <http://dx.doi.org/10.1073/pnas.0500879102>.
- [201] S Gupta, M C Maiden, I M Feavers, S Nee, R M May, and R M Anderson. "The maintenance of strain structure in populations of recombining infectious agents." en. In: *Nat. Med.* 2.4 (Apr. 1996), pp. 437–442. ISSN: 1078-8956. DOI: [10.1038/nm0496-437](https://doi.org/10.1038/nm0496-437). URL: <http://dx.doi.org/10.1038/nm0496-437>.
- [202] Sushim Kumar Gupta, Babu Roshan Padmanabhan, Seydina M Diene, Rafael Lopez-Rojas, Marie Kempf, Luce Landraud, and Jean-Marc Rolain. "ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes." en. In: *Antimicrob. Agents Chemother.* 58.1 (2014), pp. 212–220. ISSN: 0066-4804, 1098-6596. DOI: [10.1128/AAC.01310-13](https://doi.org/10.1128/AAC.01310-13). URL: <http://dx.doi.org/10.1128/AAC.01310-13>.
- [203] Michael U Gutmann and Jukka Corander. "Bayesian Optimization for Likelihood-free Inference of Simulator-based Statistical Models." In: *J. Mach. Learn. Res.* 17.1 (Jan. 2016), pp. 4256–4302. ISSN: 1532-4435. URL: <http://dl.acm.org/citation.cfm?id=2946645.3007078>.
- [204] D S Guttman and D E Dykhuizen. "Clonal divergence in *Escherichia coli* as a result of recombination, not mutation." en. In: *Science* 266.5189 (Nov. 1994), pp. 1380–1383. ISSN: 0036-8075. DOI: [10.1126/science.7973728](https://doi.org/10.1126/science.7973728). URL: <http://dx.doi.org/10.1126/science.7973728>.
- [205] R Hakenbeck, H Ellerbrok, T Briesse, S Handwerger, and A Tomasz. "Penicillin-binding proteins of penicillin-susceptible and -resistant pneumococci: immunological relatedness of altered proteins and changes in peptides carrying the beta-lactam binding site." en. In: *Antimicrob. Agents Chemother.* 30.4 (Oct. 1986), pp. 553–558. ISSN: 0066-4804. DOI: [10.1128/aac.30.4.553](https://doi.org/10.1128/aac.30.4.553). URL: <http://dx.doi.org/10.1128/aac.30.4.553>.

- [206] Matthew Hall, Mark Woolhouse, and Andrew Rambaut. "Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set." en. In: *PLoS Comput. Biol.* 11.12 (Dec. 2015), e1004613. ISSN: 1553-734X, 1553-7358. DOI: [10.1371/journal.pcbi.1004613](https://doi.org/10.1371/journal.pcbi.1004613). URL: <http://dx.doi.org/10.1371/journal.pcbi.1004613>.
- [207] Kathryn A Harris, Paul Turner, Elaine A Green, and John C Hartley. "Duplex real-time PCR assay for detection of *Streptococcus pneumoniae* in clinical samples and determination of penicillin susceptibility." en. In: *J. Clin. Microbiol.* 46.8 (Aug. 2008), pp. 2751–2758. ISSN: 0095-1137, 1098-660X. DOI: [10.1128/JCM.02462-07](https://doi.org/10.1128/JCM.02462-07). URL: <http://dx.doi.org/10.1128/JCM.02462-07>.
- [208] Simon R Harris, Michelle J Cole, et al. "Public health surveillance of multidrug-resistant clones of *Neisseria gonorrhoeae* in Europe: a genomic survey." en. In: *Lancet Infect. Dis.* 18.7 (July 2018), pp. 758–768. ISSN: 1473-3099, 1474-4457. DOI: [10.1016/S1473-3099\(18\)30225-1](https://doi.org/10.1016/S1473-3099(18)30225-1). URL: [http://dx.doi.org/10.1016/S1473-3099\(18\)30225-1](http://dx.doi.org/10.1016/S1473-3099(18)30225-1).
- [209] Simon R Harris, Edward J Feil, et al. "Evolution of MRSA during hospital transmission and intercontinental spread." en. In: *Science* 327.5964 (Jan. 2010), pp. 469–474. ISSN: 0036-8075, 1095-9203. DOI: [10.1126/science.1182395](https://doi.org/10.1126/science.1182395). URL: <http://dx.doi.org/10.1126/science.1182395>.
- [210] 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." en. June 2020. URL: <https://www.biorxiv.org/content/10.1101/2020.06.05.118836v2>.
- [211] David L Hava and Andrew Camilli. "Large-scale identification of serotype 4 *Streptococcus pneumoniae* virulence factors." In: *Mol. Microbiol.* 45.5 (2002), pp. 1389–1406. ISSN: 0950-382X. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2958.2002.03106.x>.
- [212] L S Håvarstein, G Coomaraswamy, and D A Morrison. "An unmodified heptadecapeptide pheromone induces competence for genetic transformation in *Streptococcus pneumoniae*." en. In: *Proc. Natl. Acad. Sci. U. S. A.* 92.24 (Nov. 1995), pp. 11140–11144. ISSN: 0027-8424. DOI: [10.1073/pnas.92.24.11140](https://doi.org/10.1073/pnas.92.24.11140). URL: <http://dx.doi.org/10.1073/pnas.92.24.11140>.
- [213] Leiv Sigve Håvarstein, Bernard Martin, Ola Johnsborg, Chantal Granadel, and Jean-Pierre Claverys. "New insights into the pneumococcal fratricide: relationship to clumping and identification of a novel immunity factor." en. In: *Mol. Microbiol.* 59.4

- (Feb. 2006), pp. 1297–1307. ISSN: 0950-382X. DOI: [10.1111/j.1365-2958.2005.05021.x](https://doi.org/10.1111/j.1365-2958.2005.05021.x). URL: <http://dx.doi.org/10.1111/j.1365-2958.2005.05021.x>.
- [214] T Hayashi, K Makino, et al. “Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12.” en. In: *DNA Res.* 8.1 (Feb. 2001), pp. 11–22. ISSN: 1340-2838. DOI: [10.1093/dnares/8.1.11](https://doi.org/10.1093/dnares/8.1.11). URL: <http://dx.doi.org/10.1093/dnares/8.1.11>.
- [215] J F Heidelberg, J A Eisen, et al. “DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*.” en. In: *Nature* 406.6795 (Aug. 2000), pp. 477–483. ISSN: 0028-0836. DOI: [10.1038/35020000](https://doi.org/10.1038/35020000). URL: <http://dx.doi.org/10.1038/35020000>.
- [216] Thomas Helleday, Saeed Eshtad, and Serena Nik-Zainal. “Mechanisms underlying mutational signatures in human cancers.” en. In: *Nat. Rev. Genet.* 15.9 (Sept. 2014), pp. 585–598. ISSN: 1471-0056, 1471-0064. DOI: [10.1038/nrg3729](https://doi.org/10.1038/nrg3729). URL: <http://dx.doi.org/10.1038/nrg3729>.
- [217] Katherine A Heller and Zoubin Ghahramani. “Bayesian Hierarchical Clustering.” In: *Proceedings of the 22Nd International Conference on Machine Learning. ICML '05*. Bonn, Germany: ACM, 2005, pp. 297–304. ISBN: 9781595931801. DOI: [10.1145/1102351.1102389](https://doi.org/10.1145/1102351.1102389). URL: <http://doi.acm.org/10.1145/1102351.1102389>.
- [218] Philipp Hennig and Christian J Schuler. “Entropy search for information-efficient global optimization.” In: *J. Mach. Learn. Res.* 13.null (June 2012), pp. 1809–1837. ISSN: 1532-4435. URL: <https://dl.acm.org/doi/abs/10.5555/2188385.2343701>.
- [219] Ruth Hershberg and Dmitri A Petrov. “Evidence that mutation is universally biased towards AT in bacteria.” en. In: *PLoS Genet.* 6.9 (Sept. 2010), e1001115. ISSN: 1553-7390, 1553-7404. DOI: [10.1371/journal.pgen.1001115](https://doi.org/10.1371/journal.pgen.1001115). URL: <http://dx.doi.org/10.1371/journal.pgen.1001115>.
- [220] L A Hicks, Y-W Chien, T H Taylor, M Haber, K P Klugman, and on behalf of the Active Bacterial Core Surveillance (ABCs) Team. *Outpatient Antibiotic Prescribing and Nonsusceptible *Streptococcus pneumoniae* in the United States, 1996-2003*. 2011. DOI: [10.1093/cid/cir443](https://doi.org/10.1093/cid/cir443). URL: <http://dx.doi.org/10.1093/cid/cir443>.
- [221] Elena Hilario, Johann Peter Gogarten, and Others. “Horizontal transfer of ATPase genes—the tree of life becomes a net of life.” In: *Biosystems*. 31.2-3 (1993), pp. 111–119. ISSN: 0303-2647. URL: http://www.academia.edu/download/43639300/Horizontal_transfer_of_ATPase_genes_-_Th20160311-3241-eh16yg.pdf.

- [222] Philip C Hill, Yin Bun Cheung, Abiodun Akisanya, Kawsu Sankareh, George Lahai, Brian M Greenwood, and Richard A Adegbola. "Nasopharyngeal carriage of *Streptococcus pneumoniae* in Gambian infants: a longitudinal study." en. In: *Clin. Infect. Dis.* 46.6 (Mar. 2008), pp. 807–814. ISSN: 1058-4838, 1537-6591. DOI: [10 . 1086 / 528688](https://doi.org/10.1086/528688). URL: [http://dx.doi.org/10 . 1086/528688](http://dx.doi.org/10.1086/528688).
- [223] R Himmelreich, H Hilbert, H Plagens, E Pirkl, B C Li, and R Herrmann. "Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*." en. In: *Nucleic Acids Res.* 24.22 (Nov. 1996), pp. 4420–4449. ISSN: 0305-1048. DOI: [10 . 1093 / nar / 24 . 22 . 4420](https://doi.org/10.1093/nar/24.22.4420). URL: [http://dx.doi.org/10 . 1093 / nar / 24 . 22 . 4420](http://dx.doi.org/10.1093/nar/24.22.4420).
- [224] Matthew T G Holden, Helena M B Seth-Smith, et al. "The genome of *Burkholderia cenocepacia* J2315, an epidemic pathogen of cystic fibrosis patients." en. In: *J. Bacteriol.* 191.1 (Jan. 2009), pp. 261–277. ISSN: 0021-9193, 1098-5530. DOI: [10 . 1128 / JB . 01230 - 08](https://doi.org/10.1128/JB.01230-08). URL: <http://dx.doi.org/10.1128/JB.01230-08>.
- [225] Guillaume Holley and Páll Melsted. *Bifrost – Highly parallel construction and indexing of colored and compacted de Bruijn graphs*. DOI: [10 . 1101 / 695338](https://doi.org/10.1101/695338). URL: [http://dx.doi.org/10 . 1101 / 695338](http://dx.doi.org/10.1101/695338).
- [226] A R Holmes, R McNab, K W Millsap, M Rohde, S Hammer-schmidt, J L Mawdsley, and H F Jenkinson. "The *pavA* gene of *Streptococcus pneumoniae* encodes a fibronectin-binding protein that is essential for virulence." en. In: *Mol. Microbiol.* 41.6 (Sept. 2001), pp. 1395–1408. ISSN: 0950-382X. DOI: [10 . 1046 / j . 1365 - 2958 . 2001 . 02610 . x](https://doi.org/10.1046/j.1365-2958.2001.02610.x). URL: [http://dx.doi.org/10.1046 / j . 1365 - 2958 . 2001 . 02610 . x](http://dx.doi.org/10.1046/j.1365-2958.2001.02610.x).
- [227] Kathryn E Holt, Julian Parkhill, et al. "High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi." en. In: *Nat. Genet.* 40.8 (Aug. 2008), pp. 987–993. ISSN: 1061-4036, 1546-1718. DOI: [10 . 1038 / ng . 195](https://doi.org/10.1038/ng.195). URL: <http://dx.doi.org/10.1038/ng.195>.
- [228] Kathryn E Holt, Heiman Wertheim, et al. "Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health." en. In: *Proc. Natl. Acad. Sci. U. S. A.* 112.27 (July 2015), E3574–81. ISSN: 0027-8424, 1091-6490. DOI: [10 . 1073 / pnas . 1501049112](https://doi.org/10.1073/pnas.1501049112). URL: [http://dx.doi.org/10 . 1073 / pnas . 1501049112](http://dx.doi.org/10.1073/pnas.1501049112).
- [229] Manuel Holtgrewe. *Mason: a read simulator for second generation sequencing data*. Tech. rep. TR-B-10-06. Institut für Mathematik und Informatik, Freie Universität Berlin, 2010. URL: <http://publications.imp.fu-berlin.de/962/2/mason201009.pdf>.

- [230] J G Howe and T S Wilson. "Co-trimoxazole-resistant pneumococci." en. In: *Lancet* 2.7769 (July 1972), pp. 184–185. ISSN: 0140-6736. DOI: [10.1016/s0140-6736\(72\)91354-2](https://doi.org/10.1016/s0140-6736(72)91354-2). URL: [http://dx.doi.org/10.1016/s0140-6736\(72\)91354-2](http://dx.doi.org/10.1016/s0140-6736(72)91354-2).
- [231] Weichun Huang, Leping Li, Jason R Myers, and Gabor T Marth. "ART: a next-generation sequencing read simulator." en. In: *Bioinformatics* 28.4 (Feb. 2012), pp. 593–594. ISSN: 1367-4803, 1367-4811. DOI: [10.1093/bioinformatics/btr708](https://doi.org/10.1093/bioinformatics/btr708). URL: <http://dx.doi.org/10.1093/bioinformatics/btr708>.
- [232] J P Huelsenbeck and F Ronquist. "MRBAYES: Bayesian inference of phylogenetic trees." en. In: *Bioinformatics* 17.8 (Aug. 2001), pp. 754–755. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/17.8.754](https://doi.org/10.1093/bioinformatics/17.8.754). URL: <http://dx.doi.org/10.1093/bioinformatics/17.8.754>.
- [233] John P Huelsenbeck and Peter Andolfatto. "Inference of population structure under a Dirichlet process model." en. In: *Genetics* 175.4 (Apr. 2007), pp. 1787–1802. ISSN: 0016-6731. DOI: [10.1534/genetics.106.061317](https://doi.org/10.1534/genetics.106.061317). URL: <http://dx.doi.org/10.1534/genetics.106.061317>.
- [234] John P Huelsenbeck, Peter Andolfatto, and Edna T Huelsenbeck. "Structurama: bayesian inference of population structure." en. In: *Evol. Bioinform. Online* 7 (June 2011), pp. 55–59. ISSN: 1176-9343. DOI: [10.4137/EB0.S6761](https://doi.org/10.4137/EB0.S6761). URL: <http://dx.doi.org/10.4137/EB0.S6761>.
- [235] D T D Hughes. "Single-blind comparative trial of trimethoprim-sulphamethoxazole and ampicillin in the treatment of exacerbations of chronic bronchitis." In: *BMJ* 4.5681 (1969), pp. 470–473. DOI: [10.1136/bmj.4.5681.470](https://doi.org/10.1136/bmj.4.5681.470). URL: <http://dx.doi.org/10.1136/bmj.4.5681.470>.
- [236] Martin Hunt, Alison E Mather, Leonor Sánchez-Busó, Andrew J Page, Julian Parkhill, Jacqueline A Keane, and Simon R Harris. "ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads." en. In: *Microb Genom* 3.10 (Oct. 2017), e000131. ISSN: 2057-5858. DOI: [10.1099/mgen.0.000131](https://doi.org/10.1099/mgen.0.000131). URL: <http://dx.doi.org/10.1099/mgen.0.000131>.
- [237] Doug Hyatt, Gwo-Liang Chen, Philip F Locascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. "Prodigal: prokaryotic gene recognition and translation initiation site identification." en. In: *BMC Bioinformatics* 11 (Mar. 2010), p. 119. ISSN: 1471-2105. DOI: [10.1186/1471-2105-11-119](https://doi.org/10.1186/1471-2105-11-119). URL: <http://dx.doi.org/10.1186/1471-2105-11-119>.

- [238] Hal Daume Iii. "Fast search for Dirichlet process mixture models." In: *Proceedings of Machine Learning Research* 2 (2007). Ed. by Marina Meila and Xiaotong Shen, pp. 83–90. URL: <http://proceedings.mlr.press/v2/daume07a.html>.
- [239] Piotr Indyk and Rajeev Motwani. "Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality." In: *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*. STOC '98. Dallas, Texas, USA: ACM, 1998, pp. 604–613. ISBN: 9780897919623. DOI: [10.1145/276698.276876](https://doi.org/10.1145/276698.276876). URL: <http://doi.acm.org/10.1145/276698.276876>.
- [240] J Inostroza, A M Vinet, G Retamal, P Lorca, G Ossa, R R Facklam, and R U Sorensen. "Influence of patient age on Streptococcus pneumoniae serotypes causing invasive disease." en. In: *Clin. Diagn. Lab. Immunol.* 8.3 (May 2001), pp. 556–559. ISSN: 1071-412X. DOI: [10.1128/CDLI.8.3.556-559.2001](https://doi.org/10.1128/CDLI.8.3.556-559.2001). URL: <http://dx.doi.org/10.1128/CDLI.8.3.556-559.2001>.
- [241] 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." en. In: *Genome Med.* 6.11 (Nov. 2014), p. 90. ISSN: 1756-994X. DOI: [10.1186/s13073-014-0090-6](https://doi.org/10.1186/s13073-014-0090-6). URL: <http://dx.doi.org/10.1186/s13073-014-0090-6>.
- [242] Zamin Iqbal, Mario Caccamo, Isaac Turner, Paul Flicek, and Gil McVean. "De novo assembly and genotyping of variants using colored de Bruijn graphs." en. In: *Nat. Genet.* 44.2 (Jan. 2012), pp. 226–232. ISSN: 1061-4036, 1546-1718. DOI: [10.1038/ng.1028](https://doi.org/10.1038/ng.1028). URL: <http://dx.doi.org/10.1038/ng.1028>.
- [243] M R Jacobs, H J Koornhof, et al. "Emergence of multiply resistant pneumococci." en. In: *N. Engl. J. Med.* 299.14 (Oct. 1978), pp. 735–740. ISSN: 0028-4793. DOI: [10.1056/NEJM197810052991402](https://doi.org/10.1056/NEJM197810052991402). URL: <http://dx.doi.org/10.1056/NEJM197810052991402>.
- [244] Nathan T Jacobs and Jeffrey N Weiser. "Competitive exclusion strengthens selection for transmissibility and increases the benefit of recombination for within-host adaptation." en. June 2020. URL: <https://www.biorxiv.org/content/10.1101/2020.06.18.158956v2>.
- [245] Magali Jaillard, Leandro Lima, Maud Tournoud, Pierre Mahé, Alex van Belkum, Vincent Lacroix, and Laurent Jacob. "A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events." en. In: *PLoS Genet.* 14.11 (Nov. 2018), e1007758. ISSN: 1553-7390, 1553-7404. DOI: [10.1371/journal.pgen.1007758](https://doi.org/10.1371/journal.pgen.1007758). URL: <http://dx.doi.org/10.1371/journal.pgen.1007758>.

- [246] Marko Järvenpää, Michael U Gutmann, Aki Vehtari, and Pekka Marttinen. "Parallel Gaussian Process Surrogate Bayesian Inference with Noisy Likelihood Evaluations." en. In: *Bayesian Anal.* (2020). ISSN: 1936-0975, 1931-6690. DOI: [10 . 1214 / 20 - BA1200](https://doi.org/10.1214/20-BA1200). URL: <https://projecteuclid.org/euclid.ba/1584583229>.
- [247] Justin Jee, Aviram Rasouly, Ilya Shamovsky, Yonatan Akivis, Susan R Steinman, Bud Mishra, and Evgeny Nudler. "Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing." en. In: *Nature* 534:7609 (June 2016), pp. 693–696. ISSN: 0028-0836, 1476-4687. DOI: [10 . 1038 / nature18313](https://doi.org/10.1038/nature18313). URL: <http://dx.doi.org/10.1038/nature18313>.
- [248] Mark Johnson, Irena Zaretskaya, Yan Raytselis, Yuri Merezhuik, Scott McGinnis, and Thomas L Madden. "NCBI BLAST: a better web interface." en. In: *Nucleic Acids Res.* 36.Web Server issue (July 2008), W5–9. ISSN: 0305-1048, 1362-4962. DOI: [10 . 1093 / nar / gkn201](https://doi.org/10.1093/nar/gkn201). URL: <http://dx.doi.org/10.1093/nar/gkn201>.
- [249] Thibaut Jombart, Anne Cori, Xavier Didelot, Simon Cauchemez, Christophe Fraser, and Neil Ferguson. "Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data." en. In: *PLoS Comput. Biol.* 10.1 (Jan. 2014), e1003457. ISSN: 1553-734X, 1553-7358. DOI: [10 . 1371 / journal . pcbi . 1003457](https://doi.org/10.1371/journal.pcbi.1003457). URL: <http://dx.doi.org/10.1371/journal.pcbi.1003457>.
- [250] Thibaut Jombart, Sébastien Devillard, and François Balloux. "Discriminant analysis of principal components: a new method for the analysis of genetically structured populations." en. In: *BMC Genet.* 11 (Oct. 2010), p. 94. ISSN: 1471-2156. DOI: [10 . 1186 / 1471 - 2156 - 11 - 94](https://doi.org/10.1186/1471-2156-11-94). URL: <http://dx.doi.org/10.1186/1471-2156-11-94>.
- [251] Bradley R Jones and Art F Y Poon. "node.dating: dating ancestors in phylogenetic trees in R." en. In: *Bioinformatics* 33.6 (Mar. 2017), pp. 932–934. ISSN: 1367-4803, 1367-4811. DOI: [10 . 1093 / bioinformatics / btw744](https://doi.org/10.1093/bioinformatics/btw744). URL: <http://dx.doi.org/10.1093/bioinformatics/btw744>.
- [252] Roger Junges, Gabriela Salvadori, Tsute Chen, Donald A Morrison, and Fernanda C Petersen. "Hidden Gems in the Transcriptome Maps of Competent Streptococci." en. In: *Front Mol Biosci* 5 (2018), p. 116. ISSN: 2296-889X. DOI: [10 . 3389 / fmolb . 2018 . 00116](https://doi.org/10.3389/fmolb.2018.00116). URL: <http://dx.doi.org/10.3389/fmolb.2018.00116>.
- [253] Aras Kadioglu, Jeffrey N Weiser, James C Paton, and Peter W Andrew. "The role of Streptococcus pneumoniae virulence factors in host respiratory colonization and disease." en. In: *Nat. Rev. Microbiol.* 6.4 (Apr. 2008), pp. 288–301. ISSN: 1740-1526,

- 1740-1534. DOI: [10.1038/nrmicro1871](https://doi.org/10.1038/nrmicro1871). URL: <http://dx.doi.org/10.1038/nrmicro1871>.
- [254] 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." en. In: *PeerJ* 4 (Sept. 2016), e2477. ISSN: 2167-8359. DOI: [10.7717/peerj.2477](https://doi.org/10.7717/peerj.2477). URL: <http://dx.doi.org/10.7717/peerj.2477>.
- [255] Alboukadel Kassambara. *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*. en. STHDA, Aug. 2017. ISBN: 9781542462709. URL: <https://market.android.com/details?id=book-plEyDwAAQBAJ>.
- [256] Zoubin Ghahramani Katherine A. Heller. "Randomized Algorithms for Fast Bayesian Hierarchical Clustering." In: *cite-seerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.60.298* (2005). URL: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.60.298>.
- [257] Kazutaka Katoh, Kazuharu Misawa, Kei-Ichi Kuma, and Takashi Miyata. "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform." en. In: *Nucleic Acids Res.* 30.14 (July 2002), pp. 3059-3066. ISSN: 0305-1048, 1362-4962. DOI: [10.1093/nar/gkf436](https://doi.org/10.1093/nar/gkf436). URL: <http://nar.oxfordjournals.org/content/30/14/3059>.
- [258] Powel Kazanjian. "Changing interest among physicians toward pneumococcal vaccination throughout the twentieth century." en. In: *J. Hist. Med. Allied Sci.* 59.4 (Oct. 2004), pp. 555-587. ISSN: 0022-5045. DOI: [10.1093/jhmas/jrh110](https://doi.org/10.1093/jhmas/jrh110). URL: <http://dx.doi.org/10.1093/jhmas/jrh110>.
- [259] David G Kendall and Others. "On the generalized" birth-and-death" process." In: *Ann. Math. Stat.* 19.1 (1948), pp. 1-15. URL: <https://projecteuclid.org/euclid.aoms/1177730285>.
- [260] Daehwan Kim, Li Song, Florian P Breitwieser, and Steven L Salzberg. "Centrifuge: rapid and sensitive classification of metagenomic sequences." en. In: *Genome Res.* 26.12 (Dec. 2016), pp. 1721-1729. ISSN: 1088-9051, 1549-5469. DOI: [10.1101/gr.210641.116](https://doi.org/10.1101/gr.210641.116). URL: <http://dx.doi.org/10.1101/gr.210641.116>.
- [261] Dongwan Kim, Joo-Yeon Lee, Jeong-Sun Yang, Jun Won Kim, V Narry Kim, and Hyeshik Chang. "The Architecture of SARS-CoV-2 Transcriptome." en. In: *Cell* 181.4 (May 2020), 914-921.e10. ISSN: 0092-8674, 1097-4172. DOI: [10.1016/j.cell.2020.04.011](https://doi.org/10.1016/j.cell.2020.04.011). URL: <http://dx.doi.org/10.1016/j.cell.2020.04.011>.

- [262] 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." en. In: *Clin. Microbiol. Rev.* 29.3 (July 2016), pp. 525–552. ISSN: 0893-8512, 1098-6618. DOI: [10.1128/CMR.00058-15](https://doi.org/10.1128/CMR.00058-15). URL: <http://dx.doi.org/10.1128/CMR.00058-15>.
- [263] J F C Kingman. "The coalescent." In: *Stochastic Process. Appl.* 13.3 (Sept. 1982), pp. 235–248. ISSN: 0304-4149. DOI: [10.1016/0304-4149\(82\)90011-4](https://doi.org/10.1016/0304-4149(82)90011-4). URL: <http://www.sciencedirect.com/science/article/pii/0304414982900114>.
- [264] J W Kislak. "Type 6 pneumococcus resistant to erythromycin and lincomycin." en. In: *N. Engl. J. Med.* 276.15 (Apr. 1967), p. 852. ISSN: 0028-4793. DOI: [10.1056/NEJM196704132761508](https://doi.org/10.1056/NEJM196704132761508). URL: <http://dx.doi.org/10.1056/NEJM196704132761508>.
- [265] Don Klinkenberg, Jantien A Backer, Xavier Didelot, Caroline Colijn, and Jacco Wallinga. "Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks." en. In: *PLoS Comput. Biol.* 13.5 (May 2017), e1005495. ISSN: 1553-734X, 1553-7358. DOI: [10.1371/journal.pcbi.1005495](https://doi.org/10.1371/journal.pcbi.1005495). URL: <http://dx.doi.org/10.1371/journal.pcbi.1005495>.
- [266] K Klugman. "Pneumococcal molecular epidemiology network." In: *ASM News* 64.7 (1998), pp. 371–371. ISSN: 0044-7897.
- [267] James R Knight, Eileen M Dunne, E Kim Mulholland, Sudipta Saha, Catherine Satzke, Adrienn Tothpal, and Daniel M Weinberger. "Determining the serotype composition of mixed samples of pneumococcus using whole genome sequencing." en. Aug. 2019. URL: <https://www.biorxiv.org/content/10.1101/741603v1>.
- [268] Mirjam J Knol, Gertjan H J Wagenvoort, Elisabeth A M Sanders, Karin Elberse, Bart J Vlamincx, Hester E de Melker, and Arie van der Ende. "Invasive Pneumococcal Disease 3 Years after Introduction of 10-Valent Pneumococcal Conjugate Vaccine, the Netherlands." en. In: *Emerg. Infect. Dis.* 21.11 (Nov. 2015), pp. 2040–2044. ISSN: 1080-6040, 1080-6059. DOI: [10.3201/eid2111.140780](https://doi.org/10.3201/eid2111.140780). URL: <http://dx.doi.org/10.3201/eid2111.140780>.
- [269] A H Knoll. *Life on a young planet: the first three billion years of evolution on earth*. Princeton Univ. Press Princeton, NJ, 2003.
- [270] D Koelman, P Kremer, J Lees, M Brouwer, S Bentley, and D van de Beek. "Bacterial hypervirulence in *Haemophilus influenzae* meningitis identified by whole genome sequencing." In: *J. Neurol. Sci.* 381 (Oct. 2017), pp. 181–182. ISSN: 0022-510X. DOI: [10.1016/j.jns.2017.08.523](https://doi.org/10.1016/j.jns.2017.08.523). URL: <http://www.sciencedirect.com/science/article/pii/S0022510X17310201>.

- [271] Masamitsu Kono, M Ammar Zafar, Marisol Zuniga, Aoife M Roche, Shigeto Hamaguchi, and Jeffrey N Weiser. "Single Cell Bottlenecks in the Pathogenesis of *Streptococcus pneumoniae*." en. In: *PLoS Pathog.* 12.10 (Oct. 2016), e1005887. ISSN: 1553-7366, 1553-7374. DOI: [10.1371/journal.ppat.1005887](https://doi.org/10.1371/journal.ppat.1005887). URL: <http://dx.doi.org/10.1371/journal.ppat.1005887>.
- [272] Sergei L Kosakovsky Pond and Simon D W Frost. "Not so different after all: a comparison of methods for detecting amino acid sites under selection." en. In: *Mol. Biol. Evol.* 22.5 (May 2005), pp. 1208–1222. ISSN: 0737-4038. DOI: [10.1093/molbev/msi105](https://doi.org/10.1093/molbev/msi105). URL: <http://dx.doi.org/10.1093/molbev/msi105>.
- [273] Sergei L Kosakovsky Pond, David Posada, Michael B Gravenor, Christopher H Woelk, and Simon D W Frost. "GARD: a genetic algorithm for recombination detection." en. In: *Bioinformatics* 22.24 (Dec. 2006), pp. 3096–3098. ISSN: 1367-4803, 1367-4811. DOI: [10.1093/bioinformatics/btl474](https://doi.org/10.1093/bioinformatics/btl474). URL: <http://dx.doi.org/10.1093/bioinformatics/btl474>.
- [274] Claudio U Köser, Matthew T G Holden, et al. "Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak." en. In: *N. Engl. J. Med.* 366.24 (June 2012), pp. 2267–2275. ISSN: 0028-4793, 1533-4406. DOI: [10.1056/NEJMoa1109910](https://doi.org/10.1056/NEJMoa1109910). URL: <http://dx.doi.org/10.1056/NEJMoa1109910>.
- [275] P H C Kremer, J A Lees, et al. "Benzalkonium tolerance genes and outcome in *Listeria monocytogenes* meningitis." en. In: *Clin. Microbiol. Infect.* 23.4 (Apr. 2017), 265.e1–265.e7. ISSN: 1198-743X, 1469-0691. DOI: [10.1016/j.cmi.2016.12.008](https://doi.org/10.1016/j.cmi.2016.12.008). URL: <http://dx.doi.org/10.1016/j.cmi.2016.12.008>.
- [276] David M Kristensen, Lavanya Kannan, Michael K Coleman, Yuri I Wolf, Alexander Sorokin, Eugene V Koonin, and Arcady Mushegian. "A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches." en. In: *Bioinformatics* 26.12 (June 2010), pp. 1481–1487. ISSN: 1367-4803, 1367-4811. DOI: [10.1093/bioinformatics/btq229](https://doi.org/10.1093/bioinformatics/btq229). URL: <http://dx.doi.org/10.1093/bioinformatics/btq229>.
- [277] Hans E Krokan, Finn Drabløs, and Geir Slupphaug. "Uracil in DNA—occurrence, consequences and repair." en. In: *Oncogene* 21.58 (Dec. 2002), pp. 8935–8948. ISSN: 0950-9232. DOI: [10.1038/sj.onc.1205996](https://doi.org/10.1038/sj.onc.1205996). URL: <http://dx.doi.org/10.1038/sj.onc.1205996>.
- [278] Thomas A Kunkel and Dorothy A Erie. "DNA mismatch repair." en. In: *Annu. Rev. Biochem.* 74 (2005), pp. 681–710. ISSN: 0066-4154. DOI: [10.1146/annurev.biochem.74.082803.133243](https://doi.org/10.1146/annurev.biochem.74.082803.133243). URL: <http://dx.doi.org/10.1146/annurev.biochem.74.082803.133243>.

- [279] H J Kushner. "A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise." In: *J. Basic Eng* 86.1 (Mar. 1964), pp. 97–106. ISSN: 0021-9223. DOI: [10.1115/1.3653121](https://doi.org/10.1115/1.3653121). URL: <https://asmedigitalcollection.asme.org/fluidsengineering/article-abstract/86/1/97/392213/A-New-Method-of-Locating-the-Maximum-Point-of-an?redirectedFrom=fulltext>.
- [280] P A Lachenbruch and M Goldstein. "Discriminant Analysis." In: *Biometrics* 139.2 (Mar. 1979). ISSN: 0006-341X. DOI: [10.2307/2529937](https://doi.org/10.2307/2529937). URL: https://www.researchgate.net/publication/267057025_Discriminant_Analysis.
- [281] Ben Langmead and Steven L Salzberg. "Fast gapped-read alignment with Bowtie 2." en. In: *Nat. Methods* 9.4 (Mar. 2012), pp. 357–359. ISSN: 1548-7091, 1548-7105. DOI: [10.1038/nmeth.1923](https://doi.org/10.1038/nmeth.1923). URL: <http://dx.doi.org/10.1038/nmeth.1923>.
- [282] Emily K Latch, Guha Dharmarajan, Jeffrey C Glaubitz, and Olin E Rhodes. "Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation." In: *Conserv. Genet.* 7.2 (Apr. 2006), pp. 295–302. ISSN: 1566-0621, 1572-9737. DOI: [10.1007/s10592-005-9098-1](https://doi.org/10.1007/s10592-005-9098-1). URL: <https://doi.org/10.1007/s10592-005-9098-1>.
- [283] Daniel John Lawson, Garrett Hellenthal, Simon Myers, and Daniel Falush. "Inference of population structure using dense haplotype data." In: *PLoS Genet.* 8.1 (Jan. 2012), e1002453. ISSN: 1553-7390, 1553-7404. DOI: [10.1371/journal.pgen.1002453](https://doi.org/10.1371/journal.pgen.1002453). URL: <http://dx.doi.org/10.1371/journal.pgen.1002453>.
- [284] Si Quang Le and Olivier Gascuel. "An improved general amino acid replacement matrix." en. In: *Mol. Biol. Evol.* 25.7 (July 2008), pp. 1307–1320. ISSN: 0737-4038, 1537-1719. DOI: [10.1093/molbev/msn067](https://doi.org/10.1093/molbev/msn067). URL: <http://dx.doi.org/10.1093/molbev/msn067>.
- [285] Heewook Lee, Ellen Popodi, Haixu Tang, and Patricia L Foster. "Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing." en. In: *Proc. Natl. Acad. Sci. U. S. A.* 109.41 (Oct. 2012), E2774–83. ISSN: 0027-8424, 1091-6490. DOI: [10.1073/pnas.1210309109](https://doi.org/10.1073/pnas.1210309109). URL: <http://dx.doi.org/10.1073/pnas.1210309109>.
- [286] Robyn S Lee, Jean-François Proulx, Fiona McIntosh, Marcel A Behr, and William P Hanage. "Previously undetected super-spreading of *Mycobacterium tuberculosis* revealed by deep sequencing." en. In: *Elife* 9 (Feb. 2020). ISSN: 2050-084X. DOI: [10.7554/eLife.53245](https://doi.org/10.7554/eLife.53245). URL: <http://dx.doi.org/10.7554/eLife.53245>.

- [287] John A Lees, Nicholas J Croucher, David Goldblatt, François Nosten, Julian Parkhill, Claudia Turner, Paul Turner, and Stephen D Bentley. "Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration." en. In: *Elife* 6 (July 2017). ISSN: 2050-084X. DOI: [10.7554/eLife.26255](https://doi.org/10.7554/eLife.26255). URL: <http://dx.doi.org/10.7554/eLife.26255>.
- [288] John A Lees, Bart Ferwerda, et al. "Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis." en. In: *Nat. Commun.* 10.1 (May 2019), p. 2176. ISSN: 2041-1723. DOI: [10.1038/s41467-019-09976-3](https://doi.org/10.1038/s41467-019-09976-3). URL: <http://dx.doi.org/10.1038/s41467-019-09976-3>.
- [289] John A Lees, Marco Galardini, Stephen D Bentley, Jeffrey N Weiser, and Jukka Corander. "pyseer: a comprehensive tool for microbial pangenome-wide association studies." en. In: *Bioinformatics* 34.24 (Dec. 2018), pp. 4310–4312. ISSN: 1367-4803, 1367-4811. DOI: [10.1093/bioinformatics/bty539](https://doi.org/10.1093/bioinformatics/bty539). URL: <http://dx.doi.org/10.1093/bioinformatics/bty539>.
- [290] 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." en. In: *Genome Res.* 29.2 (Feb. 2019), pp. 304–316. ISSN: 1088-9051, 1549-5469. DOI: [10.1101/gr.241455.118](https://doi.org/10.1101/gr.241455.118). URL: <http://dx.doi.org/10.1101/gr.241455.118>.
- [291] John A Lees, Philip H C Kremer, et al. "Large scale genomic analysis shows no evidence for pathogen adaptation between the blood and cerebrospinal fluid niches during bacterial meningitis." en. In: *Microb Genom* 3.1 (Jan. 2017), e000103. ISSN: 2057-5858. DOI: [10.1099/mgen.0.000103](https://doi.org/10.1099/mgen.0.000103). URL: <http://dx.doi.org/10.1099/mgen.0.000103>.
- [292] John A Lees, T Tien Mai, Marco Galardini, Nicole E Wheeler, Samuel T Horsfield, Julian Parkhill, and Jukka Corander. "Improved Prediction of Bacterial Genotype-Phenotype Associations Using Interpretable Pangenome-Spanning Regressions." en. In: *MBio* 11.4 (July 2020). ISSN: 2150-7511. DOI: [10.1128/mBio.01344-20](https://doi.org/10.1128/mBio.01344-20). URL: <http://dx.doi.org/10.1128/mBio.01344-20>.
- [293] John A Lees, Minna Vehkala, et al. "Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes." en. In: *Nat. Commun.* 7 (Sept. 2016), p. 12797. ISSN: 2041-1723. DOI: [10.1038/ncomms12797](https://doi.org/10.1038/ncomms12797). URL: <http://dx.doi.org/10.1038/ncomms12797>.

- [294] 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*. 2017. DOI: [10.1073/pnas.1617849114](https://doi.org/10.1073/pnas.1617849114). URL: <http://dx.doi.org/10.1073/pnas.1617849114>.
- [295] 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." en. In: *Proc. Natl. Acad. Sci. U. S. A.* 114.5 (Jan. 2017), pp. 1075–1080. ISSN: 0027-8424, 1091-6490. DOI: [10.1073/pnas.1617849114](https://doi.org/10.1073/pnas.1617849114). URL: <http://dx.doi.org/10.1073/pnas.1617849114>.
- [296] B R Levin, J Antonovics, and H Sharma. "Frequency-Dependent Selection in Bacterial Populations [and Discussion]." In: *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 319.1196 (1988), pp. 459–472. ISSN: 0962-8436, 0080-4622. URL: <http://www.jstor.org/stable/2396631>.
- [297] P O Lewis. "A likelihood approach to estimating phylogeny from discrete morphological character data." en. In: *Syst. Biol.* 50.6 (Nov. 2001), pp. 913–925. ISSN: 1063-5157. DOI: [10.1080/106351501753462876](https://doi.org/10.1080/106351501753462876). URL: <http://dx.doi.org/10.1080/106351501753462876>.
- [298] Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. "MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph." en. In: *Bioinformatics* 31.10 (May 2015), pp. 1674–1676. ISSN: 1367-4803, 1367-4811. DOI: [10.1093/bioinformatics/btv033](https://doi.org/10.1093/bioinformatics/btv033). URL: <http://dx.doi.org/10.1093/bioinformatics/btv033>.
- [299] Heng Li. "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data." en. In: *Bioinformatics* 27.21 (Nov. 2011), pp. 2987–2993. ISSN: 1367-4803, 1367-4811. DOI: [10.1093/bioinformatics/btr509](https://doi.org/10.1093/bioinformatics/btr509). URL: <http://dx.doi.org/10.1093/bioinformatics/btr509>.
- [300] Heng Li. "Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences." en. In: *Bioinformatics* 32.14 (July 2016), pp. 2103–2110. ISSN: 1367-4803, 1367-4811. DOI: [10.1093/bioinformatics/btw152](https://doi.org/10.1093/bioinformatics/btw152). URL: <http://dx.doi.org/10.1093/bioinformatics/btw152>.
- [301] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. "The Sequence Alignment/Map format and SAMtools." en. In: *Bioinformatics* 25.16 (Aug. 2009), pp. 2078–2079. ISSN: 1367-4803, 1460-2059. DOI: [10.1093/bioinformatics/btp352](https://doi.org/10.1093/bioinformatics/btp352).

- URL: <http://bioinformatics.oxfordjournals.org/content/25/16/2078>.
- [302] Li Li, Christian J Stoeckert Jr, and David S Roos. "OrthoMCL: identification of ortholog groups for eukaryotic genomes." en. In: *Genome Res.* 13.9 (Sept. 2003), pp. 2178–2189. ISSN: 1088-9051. DOI: [10.1101/gr.1224503](https://doi.org/10.1101/gr.1224503). URL: <http://dx.doi.org/10.1101/gr.1224503>.
 - [303] Qun Li, Xuhua Guan, et al. "Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia." en. In: *N. Engl. J. Med.* 382.13 (Mar. 2020), pp. 1199–1207. ISSN: 0028-4793, 1533-4406. DOI: [10.1056/NEJMoa2001316](https://doi.org/10.1056/NEJMoa2001316). URL: <http://dx.doi.org/10.1056/NEJMoa2001316>.
 - [304] Weizhong Li and Adam Godzik. "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." en. In: *Bioinformatics* 22.13 (July 2006), pp. 1658–1659. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btl158](https://doi.org/10.1093/bioinformatics/btl158). URL: <http://dx.doi.org/10.1093/bioinformatics/btl158>.
 - [305] Yuan Li, Benjamin J Metcalf, et al. "Penicillin-Binding Protein Transpeptidase Signatures for Tracking and Predicting β -Lactam Resistance Levels in *Streptococcus pneumoniae*." en. In: *MBio* 7.3 (June 2016). ISSN: 2150-7511. DOI: [10.1128/mBio.00756-16](https://doi.org/10.1128/mBio.00756-16). URL: <http://dx.doi.org/10.1128/mBio.00756-16>.
 - [306] Philana Ling Lin, Marian G Michaels, Janine Janosky, Michelle Ortenzo, Ellen R Wald, and Edward O Mason Jr. "Incidence of invasive pneumococcal disease in children 3 to 36 months of age at a tertiary care pediatric center 2 years after licensure of the pneumococcal conjugate vaccine." en. In: *Pediatrics* 111.4 Pt 1 (Apr. 2003), pp. 896–899. ISSN: 0031-4005, 1098-4275. DOI: [10.1542/peds.111.4.896](https://doi.org/10.1542/peds.111.4.896). URL: <http://dx.doi.org/10.1542/peds.111.4.896>.
 - [307] M Lipsitch. "Measuring and interpreting associations between antibiotic use and penicillin resistance in *Streptococcus pneumoniae*." en. In: *Clin. Infect. Dis.* 32.7 (Apr. 2001), pp. 1044–1054. ISSN: 1058-4838. DOI: [10.1086/319604](https://doi.org/10.1086/319604). URL: <http://dx.doi.org/10.1086/319604>.
 - [308] Marc Lipsitch, Caroline Colijn, Ted Cohen, William P Hanage, and Christophe Fraser. "No coexistence for free: neutral null models for multistrain pathogens." en. In: *Epidemics* 1.1 (Mar. 2009), pp. 2–13. ISSN: 1755-4365, 1878-0067. DOI: [10.1016/j.epidem.2008.07.001](https://doi.org/10.1016/j.epidem.2008.07.001). URL: <http://dx.doi.org/10.1016/j.epidem.2008.07.001>.
 - [309] Frederick Spencer Lister. *Specific serological reactions with pneumococci from different sources*. South African Institute for Medical Research, 1913.

- [310] Xue Liu, Jacqueline M Kimmey, Vincent de Bakker, Victor Nizet, and Jan-Willem Veening. "Exploration of bacterial bottlenecks and *Streptococcus pneumoniae* pathogenesis by CRISPRi-seq." en. Apr. 2020. URL: <https://www.biorxiv.org/content/10.1101/2020.04.22.055319v1.full>.
- [311] Stephanie W Lo, Rebecca A Gladstone, et al. "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." en. In: *Lancet Infect. Dis.* 19.7 (July 2019), pp. 759–769. ISSN: 1473-3099, 1474-4457. DOI: [10.1016/S1473-3099\(19\)30297-X](https://doi.org/10.1016/S1473-3099(19)30297-X). URL: [http://dx.doi.org/10.1016/S1473-3099\(19\)30297-X](http://dx.doi.org/10.1016/S1473-3099(19)30297-X).
- [312] Jutta M Loeffler and Vincent A Fischetti. "Lysogeny of *Streptococcus pneumoniae* with MM1 phage: improved adherence and other phenotypic changes." en. In: *Infect. Immun.* 74.8 (Aug. 2006), pp. 4486–4495. ISSN: 0019-9567. DOI: [10.1128/IAI.00020-06](https://doi.org/10.1128/IAI.00020-06). URL: <http://dx.doi.org/10.1128/IAI.00020-06>.
- [313] Nicholas J Loman, Chrystala Constantinidou, Jacqueline Z M Chan, Mihail Halachev, Martin Sergeant, Charles W Penn, Esther R Robinson, and Mark J Pallen. "High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity." en. In: *Nat. Rev. Microbiol.* 10.9 (Sept. 2012), pp. 599–606. ISSN: 1740-1526, 1740-1534. DOI: [10.1038/nrmicro2850](https://doi.org/10.1038/nrmicro2850). URL: <http://dx.doi.org/10.1038/nrmicro2850>.
- [314] Nicholas J Loman and Mark J Pallen. "Twenty years of bacterial genome sequencing." en. In: *Nat. Rev. Microbiol.* 13.12 (Dec. 2015), pp. 787–794. ISSN: 1740-1526, 1740-1534. DOI: [10.1038/nrmicro3565](https://doi.org/10.1038/nrmicro3565). URL: <http://dx.doi.org/10.1038/nrmicro3565>.
- [315] M G Lorenz and W Wackernagel. "Bacterial gene transfer by natural genetic transformation in the environment." In: *Microbiological Reviews* 58.3 (1994), pp. 563–602. DOI: [10.1128/mmbr.58.3.563-602.1994](https://doi.org/10.1128/mmbr.58.3.563-602.1994). URL: <http://dx.doi.org/10.1128/mmbr.58.3.563-602.1994>.
- [316] Los Alamos National Laboratory. *HIV Databases*. <https://www.hiv.lanl.gov/content/index>. Accessed: 2018-10-25. URL: <https://www.hiv.lanl.gov/content/index>.
- [317] Franklin D Lowy. "Antimicrobial resistance: the example of *Staphylococcus aureus*." en. In: *J. Clin. Invest.* 111.9 (May 2003), pp. 1265–1273. ISSN: 0021-9738. DOI: [10.1172/JCI18535](https://doi.org/10.1172/JCI18535). URL: <http://dx.doi.org/10.1172/JCI18535>.

- [318] Ari Löytynoja. "Phylogeny-aware alignment with PRANK." en. In: *Methods Mol. Biol.* 1079 (2014), pp. 155–170. ISSN: 1064-3745, 1940-6029. DOI: [10.1007/978-1-62703-646-7_10](https://doi.org/10.1007/978-1-62703-646-7_10). URL: http://dx.doi.org/10.1007/978-1-62703-646-7_10.
- [319] Chengwei Luo, Rob Knight, Heli Siljander, Mikael Knip, Ramnik J Xavier, and Dirk Gevers. "ConStrains identifies microbial strains in metagenomic datasets." en. In: *Nat. Biotechnol.* 33.10 (Oct. 2015), pp. 1045–1052. ISSN: 1087-0156, 1546-1696. DOI: [10.1038/nbt.3319](https://doi.org/10.1038/nbt.3319). URL: <http://dx.doi.org/10.1038/nbt.3319>.
- [320] Katrina A Lythgoe, Matthew David Hall, et al. "Shared SARS-CoV-2 diversity suggests localised transmission of minority variants." en. May 2020. URL: <https://www.biorxiv.org/content/10.1101/2020.05.28.118992v2>.
- [321] C M Macleod, R G Hodges, M Heidelberger, and W G Bernhard. "Prevention of pneumococcal pneumonia by immunization with specific capsula polysaccharides." en. In: *J. Exp. Med.* 82.6 (Nov. 1945), pp. 445–465. ISSN: 0022-1007. URL: <https://www.ncbi.nlm.nih.gov/pubmed/19871511>.
- [322] M C J Maiden, J A Bygraves, et al. "Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms." In: *Proceedings of the National Academy of Sciences* 95.6 (1998), pp. 3140–3145. DOI: [10.1073/pnas.95.6.3140](https://doi.org/10.1073/pnas.95.6.3140). URL: <http://dx.doi.org/10.1073/pnas.95.6.3140>.
- [323] P Helena Mäkelä, Maija Leinonen, Juhani Pukander, and Pekka Karma. "A Study of the Pneumococcal Vaccine in Prevention of Clinically Acute Attacks of Recurrent Otitis Media." en. In: *Rev. Infect. Dis.* 3.Supplement_1 (Mar. 1981), S124–S132. ISSN: 0162-0886. DOI: [10.1093/clinids/3.Supplement_1.S124](https://doi.org/10.1093/clinids/3.Supplement_1.S124). URL: https://academic.oup.com/cid/article-abstract/3/Supplement_1/S124/314321.
- [324] Tommi Mäklin, Teemu Kallonen, Jarno Alanko, Veli Mäkinen, Jukka Corander, and Antti Honkela. "Genomic Epidemiology with Mixed Samples." en. Apr. 2020. URL: <https://www.biorxiv.org/content/10.1101/2020.04.03.021501v1>.
- [325] Tommi Mäklin, Teemu Kallonen, et al. "High-resolution sweep metagenomics using fast probabilistic inference." In: *Wellcome Open Research* 5.14 (2020), p. 14. URL: <https://wellcomeopenresearch.org/articles/5-14>.
- [326] Marcel Margulies, Michael Egholm, et al. "Genome sequencing in microfabricated high-density picolitre reactors." en. In: *Nature* 437.7057 (Sept. 2005), pp. 376–380. ISSN: 0028-0836, 1476-4687. DOI: [10.1038/nature03959](https://doi.org/10.1038/nature03959). URL: <http://dx.doi.org/10.1038/nature03959>.

- [327] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. "Markov chain Monte Carlo without likelihoods." en. In: *Proc. Natl. Acad. Sci. U. S. A.* 100.26 (Dec. 2003), pp. 15324–15328. ISSN: 0027-8424. DOI: [10.1073/pnas.0306899100](https://doi.org/10.1073/pnas.0306899100). URL: <http://dx.doi.org/10.1073/pnas.0306899100>.
- [328] B Marrs. "Genetic recombination in *Rhodopseudomonas capsulata*." en. In: *Proc. Natl. Acad. Sci. U. S. A.* 71.3 (Mar. 1974), pp. 971–973. ISSN: 0027-8424. DOI: [10.1073/pnas.71.3.971](https://doi.org/10.1073/pnas.71.3.971). URL: <http://dx.doi.org/10.1073/pnas.71.3.971>.
- [329] Darren P Martin, Philippe Lemey, Martin Lott, Vincent Moulton, David Posada, and Pierre Lefevre. "RDP3: a flexible and fast computer program for analyzing recombination." en. In: *Bioinformatics* 26.19 (Oct. 2010), pp. 2462–2463. ISSN: 1367-4803, 1460-2059. DOI: [10.1093/bioinformatics/btq467](https://doi.org/10.1093/bioinformatics/btq467). URL: <http://bioinformatics.oxfordjournals.org/content/26/19/2462>.
- [330] Michael A Martin, Robyn S Lee, Lauren A Cowley, Jennifer L Gardy, and William P Hanage. "Within-host *Mycobacterium tuberculosis* diversity and its utility for inferences of transmission." en. In: *Microb Genom* 4.10 (Oct. 2018). ISSN: 2057-5858. DOI: [10.1099/mgen.0.000217](https://doi.org/10.1099/mgen.0.000217). URL: <http://dx.doi.org/10.1099/mgen.0.000217>.
- [331] Iñigo Martincorena, Keiran M Raine, Moritz Gerstung, Kevin J Dawson, Kerstin Haase, Peter Van Loo, Helen Davies, Michael R Stratton, and Peter J Campbell. "Universal Patterns of Selection in Cancer and Somatic Tissues." en. In: *Cell* 171.5 (Nov. 2017), 1029–1041.e21. ISSN: 0092-8674, 1097-4172. DOI: [10.1016/j.cell.2017.09.042](https://doi.org/10.1016/j.cell.2017.09.042). URL: <http://dx.doi.org/10.1016/j.cell.2017.09.042>.
- [332] Iñigo Martincorena, Amit Roshan, et al. "Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin." en. In: *Science* 348.6237 (May 2015), pp. 880–886. ISSN: 0036-8075, 1095-9203. DOI: [10.1126/science.aaa6806](https://doi.org/10.1126/science.aaa6806). URL: <http://dx.doi.org/10.1126/science.aaa6806>.
- [333] A Marton, M Gulyas, R Munoz, and A Tomasz. "Extremely high incidence of antibiotic resistance in clinical isolates of *Streptococcus pneumoniae* in Hungary." en. In: *J. Infect. Dis.* 163.3 (Mar. 1991), pp. 542–548. ISSN: 0022-1899. DOI: [10.1093/infdis/163.3.542](https://doi.org/10.1093/infdis/163.3.542). URL: <http://dx.doi.org/10.1093/infdis/163.3.542>.
- [334] 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." en. In: *Nucleic Acids Res.* 40.1

- (Jan. 2012), e6. ISSN: 0305-1048, 1362-4962. DOI: [10.1093/nar/gkr928](https://doi.org/10.1093/nar/gkr928). URL: <http://dx.doi.org/10.1093/nar/gkr928>.
- [335] Jeffrey P Maskell, Armine M Sefton, and Lucinda M C Hall. "Multiple Mutations Modulate the Function of Dihydrofolate Reductase in Trimethoprim-Resistant *Streptococcus pneumoniae*." en. In: *Antimicrob. Agents Chemother.* 45.4 (Apr. 2001), pp. 1104–1108. ISSN: 0066-4804, 1098-6596. DOI: [10.1128/AAC.45.4.1104-1108.2001](https://doi.org/10.1128/AAC.45.4.1104-1108.2001). URL: <https://aac.asm.org/content/45/4/1104.short>.
- [336] J Maynard Smith and N H Smith. "Detecting recombination from gene trees." en. In: *Mol. Biol. Evol.* 15.5 (May 1998), pp. 590–599. ISSN: 0737-4038. DOI: [10.1093/oxfordjournals.molbev.a025960](https://doi.org/10.1093/oxfordjournals.molbev.a025960). URL: <http://dx.doi.org/10.1093/oxfordjournals.molbev.a025960>.
- [337] Andrew G McArthur, Nicholas Waglechner, et al. "The comprehensive antibiotic resistance database." en. In: *Antimicrob. Agents Chemother.* 57.7 (July 2013), pp. 3348–3357. ISSN: 0066-4804, 1098-6596. DOI: [10.1128/AAC.00419-13](https://doi.org/10.1128/AAC.00419-13). URL: <http://dx.doi.org/10.1128/AAC.00419-13>.
- [338] Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation." en. In: *Nucleic Acids Res.* 40.10 (May 2012), pp. 4288–4297. ISSN: 0305-1048, 1362-4962. DOI: [10.1093/nar/gks042](https://doi.org/10.1093/nar/gks042). URL: <http://dx.doi.org/10.1093/nar/gks042>.
- [339] Tera L McCool, Thomas R Cate, Gregory Moy, and Jeffrey N Weiser. "The immune response to pneumococcal proteins during experimental human carriage." en. In: *J. Exp. Med.* 195.3 (Feb. 2002), pp. 359–365. ISSN: 0022-1007. DOI: [10.1084/jem.20011576](https://doi.org/10.1084/jem.20011576). URL: <http://dx.doi.org/10.1084/jem.20011576>.
- [340] Jonathan A McCullers, B Keith English, and Rodger Novak. "Isolation and characterization of vancomycin-tolerant *Streptococcus pneumoniae* from the cerebrospinal fluid of a patient who developed recrudescence meningitis." In: *J. Infect. Dis.* 181.1 (2000), pp. 369–373. ISSN: 0022-1899. URL: <https://academic.oup.com/jid/article-abstract/181/1/369/895773>.
- [341] John H McDonald and Martin Kreitman. *Adaptive protein evolution at the Adh locus in Drosophila*. 1991. DOI: [10.1038/351652a0](https://doi.org/10.1038/351652a0). URL: <http://dx.doi.org/10.1038/351652a0>.
- [342] L Clifford McDonald. "Trends in Antimicrobial Resistance in Health Care–Associated Pathogens and Effect on Treatment." In: *Clin. Infect. Dis.* 42.Supplement_2 (Jan. 2006), S65–S71. ISSN: 1058-4838. DOI: [10.1086/499404](https://doi.org/10.1086/499404). URL: https://academic.oup.com/cid/article-abstract/42/Supplement_2/S65/377160.

- [343] Maureen McDonnell, Concepcion Ronda-Lain, and Alexander Tomasz. ““Diplophage”: A bacteriophage of *Diplococcus pneumoniae*.” In: *Virology* 63.2 (Feb. 1975), pp. 577–582. ISSN: 0042-6822. DOI: [10.1016/0042-6822\(75\)90329-3](https://doi.org/10.1016/0042-6822(75)90329-3). URL: <http://www.sciencedirect.com/science/article/pii/0042682275903293>.
- [344] L McGee, L McDougal, et al. “Nomenclature of major antimicrobial-resistant clones of *Streptococcus pneumoniae* defined by the pneumococcal molecular epidemiology network.” en. In: *J. Clin. Microbiol.* 39.7 (July 2001), pp. 2565–2571. ISSN: 0095-1137. DOI: [10.1128/JCM.39.7.2565-2571.2001](https://doi.org/10.1128/JCM.39.7.2565-2571.2001). URL: <http://dx.doi.org/10.1128/JCM.39.7.2565-2571.2001>.
- [345] J H McGuire, R L Bunch, R C Anderson, H E Boaz, E H Flynn, H W Powell, and J W Smith. “Ilotycin (Erythromycin-Lilly).” In: *Antibiotics & Chemother* 2 (1952), p. 281.
- [346] Leland McInnes and John Healy. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.” In: (Feb. 2018). arXiv: [1802.03426](https://arxiv.org/abs/1802.03426) [stat.ML]. URL: <http://arxiv.org/abs/1802.03426>.
- [347] Alan McNally, Teemu Kallonen, Christopher Connor, Khalil Abudahab, David M Aanensen, Carlyne 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.” en. In: *MBio* 10.2 (Apr. 2019). ISSN: 2150-7511. DOI: [10.1128/mBio.00644-19](https://doi.org/10.1128/mBio.00644-19). URL: <http://dx.doi.org/10.1128/mBio.00644-19>.
- [348] Duccio Medini, Claudio Donati, Hervé Tettelin, Vega Masignani, and Rino Rappuoli. “The microbial pan-genome.” In: *Curr. Opin. Genet. Dev.* 15.6 (Dec. 2005), pp. 589–594. ISSN: 0959-437X. DOI: [10.1016/j.gde.2005.09.006](https://doi.org/10.1016/j.gde.2005.09.006). URL: <http://www.sciencedirect.com/science/article/pii/S0959437X05001759>.
- [349] Duccio Medini, Claudio Donati, Hervé Tettelin, Vega Masignani, and Rino Rappuoli. “The microbial pan-genome.” en. In: *Curr. Opin. Genet. Dev.* 15.6 (Dec. 2005), pp. 589–594. ISSN: 0959-437X. DOI: [10.1016/j.gde.2005.09.006](https://doi.org/10.1016/j.gde.2005.09.006). URL: <http://dx.doi.org/10.1016/j.gde.2005.09.006>.
- [350] Alexander Mellmann, Stefan Bletz, Thomas Böking, Frank Kipp, Karsten Becker, Anja Schultes, Karola Prior, and Dag Harsen. “Real-Time Genome Sequencing of Resistant Bacteria Provides Precision Infection Control in an Institutional Setting.” en. In: *J. Clin. Microbiol.* 54.12 (Dec. 2016), pp. 2874–2881. ISSN: 0095-1137, 1098-660X. DOI: [10.1128/JCM.00790-16](https://doi.org/10.1128/JCM.00790-16). URL: <http://dx.doi.org/10.1128/JCM.00790-16>.

- [351] Fabrizio Menardo, Sebastian Duchêne, Daniela Brites, and Sebastien Gagneux. "The molecular clock of *Mycobacterium tuberculosis*." en. In: *PLoS Pathog.* 15.9 (Sept. 2019), e1008067. ISSN: 1553-7366, 1553-7374. DOI: [10.1371/journal.ppat.1008067](https://doi.org/10.1371/journal.ppat.1008067). URL: <http://dx.doi.org/10.1371/journal.ppat.1008067>.
- [352] Jennifer E Mendell, Kendall D Clements, J Howard Choat, and Esther R Angert. "Extreme polyploidy in a large bacterium." en. In: *Proc. Natl. Acad. Sci. U. S. A.* 105.18 (May 2008), pp. 6730–6734. ISSN: 0027-8424, 1091-6490. DOI: [10.1073/pnas.0707522105](https://doi.org/10.1073/pnas.0707522105). URL: <http://dx.doi.org/10.1073/pnas.0707522105>.
- [353] Balaraj B Menon and Bharathi Govindarajan. "Identification of an atypical zinc metalloproteinase, ZmpC, from an epidemic conjunctivitis-causing strain of *Streptococcus pneumoniae*." en. In: *Microb. Pathog.* 56 (Mar. 2013), pp. 40–46. ISSN: 0882-4010, 1096-1208. DOI: [10.1016/j.micpath.2012.11.006](https://doi.org/10.1016/j.micpath.2012.11.006). URL: <http://dx.doi.org/10.1016/j.micpath.2012.11.006>.
- [354] Luke W Meredith, William L Hamilton, et al. "Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study." en. In: *Lancet Infect. Dis.* 20.11 (Nov. 2020), pp. 1263–1272. ISSN: 1473-3099, 1474-4457. DOI: [10.1016/S1473-3099\(20\)30562-4](https://doi.org/10.1016/S1473-3099(20)30562-4). URL: [http://dx.doi.org/10.1016/S1473-3099\(20\)30562-4](http://dx.doi.org/10.1016/S1473-3099(20)30562-4).
- [355] B J Metcalf, R E Gertz Jr, et al. "Strain features and distributions in pneumococci from children with invasive disease before and after 13-valent conjugate vaccine implementation in the USA." en. In: *Clin. Microbiol. Infect.* 22.1 (Jan. 2016), pp. 60.e9–60.e29. ISSN: 1198-743X, 1469-0691. DOI: [10.1016/j.cmi.2015.08.027](https://doi.org/10.1016/j.cmi.2015.08.027). URL: <http://dx.doi.org/10.1016/j.cmi.2015.08.027>.
- [356] R Milkman and M M Bridges. "Molecular evolution of the *Escherichia coli* chromosome. IV. Sequence comparisons." en. In: *Genetics* 133.3 (Mar. 1993), pp. 455–468. ISSN: 0016-6731. URL: <https://www.ncbi.nlm.nih.gov/pubmed/8095913>.
- [357] Ian J Miller, Evan R Rees, Jennifer Ross, Izaak Miller, Jared Baxa, Juan Lopera, Robert L Kerby, Federico E Rey, and Jason C Kwan. "Autometa: automated extraction of microbial genomes from individual shotgun metagenomes." en. In: *Nucleic Acids Res.* 47.10 (June 2019), e57. ISSN: 0305-1048, 1362-4962. DOI: [10.1093/nar/gkz148](https://doi.org/10.1093/nar/gkz148). URL: <http://dx.doi.org/10.1093/nar/gkz148>.
- [358] Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear. "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era." en.

- In: *Mol. Biol. Evol.* 37.5 (May 2020), pp. 1530–1534. ISSN: 0737-4038, 1537-1719. DOI: [10.1093/molbev/msaa015](https://doi.org/10.1093/molbev/msaa015). URL: <http://dx.doi.org/10.1093/molbev/msaa015>.
- [359] Thomas Minka and Zoubin Ghahramani. “Expectation propagation for infinite mixtures.” In: *NIPS Workshop on Nonparametric Bayesian Methods and Infinite Models*. Vol. 19. 2003. URL: <https://pdfs.semanticscholar.org/43dc/d3855e7990076976952476ef9d0f2aadb.pdf>.
- [360] T Miyata and T Yasunaga. “Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application.” en. In: *J. Mol. Evol.* 16.1 (Sept. 1980), pp. 23–36. ISSN: 0022-2844. DOI: [10.1007/BF01732067](https://doi.org/10.1007/BF01732067). URL: <http://dx.doi.org/10.1007/BF01732067>.
- [361] Kenji Mizumoto, Katsushi Kagaya, Alexander Zarebski, and Gerardo Chowell. “Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020.” en. In: *Euro Surveill.* 25.10 (Mar. 2020). ISSN: 1025-496X, 1560-7917. DOI: [10.2807/1560-7917.ES.2020.25.10.2000180](https://doi.org/10.2807/1560-7917.ES.2020.25.10.2000180). URL: <http://dx.doi.org/10.2807/1560-7917.ES.2020.25.10.2000180>.
- [362] J Mockus. “On Bayesian Methods for Seeking the Extremum and their Application.” In: *IFIP Congress* (1977). URL: <https://www.semanticscholar.org/paper/dc33f2fa9da784788e71ef6261bdb37e36bfe1c8>.
- [363] J Mockus, Vytautas Tiesis, and Antanas Zilinskas. “The application of Bayesian methods for seeking the extremum.” In: *Towards Global Optimisation 2*. Ed. by L C W Dixon and G P Szego. Vol. 2. North-Holland, Sept. 2014, pp. 117–129. URL: <http://dx.doi.org/>.
- [364] Kristin L Moffitt and Richard Malley. “Next generation pneumococcal vaccines.” en. In: *Curr. Opin. Immunol.* 23.3 (June 2011), pp. 407–413. ISSN: 0952-7915, 1879-0372. DOI: [10.1016/j.coi.2011.04.002](https://doi.org/10.1016/j.coi.2011.04.002). URL: <http://dx.doi.org/10.1016/j.coi.2011.04.002>.
- [365] Nardus Mollentze, Louis H Nel, Sunny Townsend, Kevin le Roux, Katie Hampson, Daniel T Haydon, and Samuel Soubeyrand. “A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data.” en. In: *Proc. Biol. Sci.* 281.1782 (May 2014), p. 20133251. ISSN: 0962-8452, 1471-2954. DOI: [10.1098/rspb.2013.3251](https://doi.org/10.1098/rspb.2013.3251). URL: <http://dx.doi.org/10.1098/rspb.2013.3251>.
- [366] Henry F Moore and Alan M Chesney. “A study of ethylhydrocuprein optochin in the treatment of acute lobar pneumonia.” In: *Arch. Intern. Med.* XIX.4 (Apr. 1917), pp. 611–682. ISSN:

- 0730-188X. DOI: [10.1001/archinte.1917.00080230114010](https://doi.org/10.1001/archinte.1917.00080230114010). URL: <https://jamanetwork.com/journals/jamainternalmedicine/article-abstract/653823>.
- [367] Matthew R Moore, Terri B Hyde, et al. "Impact of a conjugate vaccine on community-wide carriage of nonsusceptible *Streptococcus pneumoniae* in Alaska." en. In: *J. Infect. Dis.* 190.11 (Dec. 2004), pp. 2031–2038. ISSN: 0022-1899. DOI: [10.1086/425422](https://doi.org/10.1086/425422). URL: <http://dx.doi.org/10.1086/425422>.
- [368] Nancy A Moran. "Microbial minimalism: genome reduction in bacterial pathogens." en. In: *Cell* 108.5 (Mar. 2002), pp. 583–586. ISSN: 0092-8674. DOI: [10.1016/S0092-8674\(02\)00665-7](https://doi.org/10.1016/S0092-8674(02)00665-7). URL: [http://dx.doi.org/10.1016/S0092-8674\(02\)00665-7](http://dx.doi.org/10.1016/S0092-8674(02)00665-7).
- [369] Marco J Morelli, Gaël Thébaud, Joël Chadœuf, Donald P King, Daniel T Haydon, and Samuel Soubeyrand. "A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data." en. In: *PLoS Comput. Biol.* 8.11 (Nov. 2012), e1002768. ISSN: 1553-734X, 1553-7358. DOI: [10.1371/journal.pcbi.1002768](https://doi.org/10.1371/journal.pcbi.1002768). URL: <http://dx.doi.org/10.1371/journal.pcbi.1002768>.
- [370] Daniel Müllner and Others. "fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python." In: *J. Stat. Softw.* 53.9 (2013), pp. 1–18. URL: <https://www.jstatsoft.org/article/view/v053i09/v53i09.pdf>.
- [371] Fionn Murtagh and Pierre Legendre. "Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?" In: *J. Classification* 31.3 (Oct. 2014), pp. 274–295. ISSN: 0176-4268, 1432-1343. DOI: [10.1007/s00357-014-9161-z](https://doi.org/10.1007/s00357-014-9161-z). URL: <https://doi.org/10.1007/s00357-014-9161-z>.
- [372] Atsushi Nakabachi, Atsushi Yamashita, Hidehiro Toh, Hajime Ishikawa, Helen E Dunbar, Nancy A Moran, and Masahira Hattori. "The 160-kilobase genome of the bacterial endosymbiont *Carsonella*." en. In: *Science* 314.5797 (Oct. 2006), p. 267. ISSN: 0036-8075, 1095-9203. DOI: [10.1126/science.1134196](https://doi.org/10.1126/science.1134196). URL: <http://dx.doi.org/10.1126/science.1134196>.
- [373] Richard A Neher and Boris I Shraiman. "Competition between recombination and epistasis can cause a transition from allele to genotype selection." en. In: *Proc. Natl. Acad. Sci. U. S. A.* 106.16 (Apr. 2009), pp. 6866–6871. ISSN: 0027-8424, 1091-6490. DOI: [10.1073/pnas.0812560106](https://doi.org/10.1073/pnas.0812560106). URL: <http://dx.doi.org/10.1073/pnas.0812560106>.
- [374] M Nei and T Gojobori. "Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions." en. In: *Mol. Biol. Evol.* 3.5 (Sept. 1986), pp. 418–426. ISSN: 0737-4038. DOI: [10.1093/oxfordjournals.molbev](https://doi.org/10.1093/oxfordjournals.molbev).

- a040410. URL: <http://dx.doi.org/10.1093/oxfordjournals.molbev.a040410>.
- [375] F Neufeld and L Haendel. "Weitere untersuchungen über pneumokokkenheilsere." In: *III. Mitteilung Arbeiten aus dem Kaiserlichen Gesundheitsamte* 34 (1910), pp. 293–304.
- [376] 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." en. In: *Mol. Biol. Evol.* 32.1 (Jan. 2015), pp. 268–274. ISSN: 0737-4038, 1537-1719. DOI: [10.1093/molbev/msu300](https://doi.org/10.1093/molbev/msu300). URL: <http://dx.doi.org/10.1093/molbev/msu300>.
- [377] Hervé Nicoloff, Karin Hjort, Bruce R Levin, and Dan I Andersson. "The high prevalence of antibiotic heteroresistance in pathogenic bacteria is mainly caused by gene amplification." en. In: *Nat Microbiol* (Feb. 2019). ISSN: 2058-5276. DOI: [10.1038/s41564-018-0342-0](https://doi.org/10.1038/s41564-018-0342-0). URL: <http://dx.doi.org/10.1038/s41564-018-0342-0>.
- [378] B Noonan, B C Guild, B L deJonge, and Others. "Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen." In: *Helicobacter pylori. Nature* 397 (1999), pp. 176–180.
- [379] R Novak, B Henriques, E Charpentier, S Normark, and E Tuomanen. "Emergence of vancomycin tolerance in *Streptococcus pneumoniae*." en. In: *Nature* 399.6736 (June 1999), pp. 590–593. ISSN: 0028-0836. DOI: [10.1038/21202](https://doi.org/10.1038/21202). URL: <http://dx.doi.org/10.1038/21202>.
- [380] Elina Numminen, Claire Chewapreecha, Jukka Sirén, Claudia Turner, Paul Turner, Stephen D Bentley, and Jukka Corander. "Two-phase importance sampling for inference about transmission trees." en. In: *Proc. Biol. Sci.* 281.1794 (Nov. 2014), p. 20141324. ISSN: 0962-8452, 1471-2954. DOI: [10.1098/rspb.2014.1324](https://doi.org/10.1098/rspb.2014.1324). URL: <http://dx.doi.org/10.1098/rspb.2014.1324>.
- [381] Sergey Nurk, Dmitry Meleshko, Anton Korobeynikov, and Pavel A Pevzner. "metaSPAdes: a new versatile metagenomic assembler." en. In: *Genome Res.* 27.5 (May 2017), pp. 824–834. ISSN: 1088-9051, 1549-5469. DOI: [10.1101/gr.213959.116](https://doi.org/10.1101/gr.213959.116). URL: <http://dx.doi.org/10.1101/gr.213959.116>.
- [382] John D O'Brien, Xavier Didelot, Zamin Iqbal, Lucas Amenga-Etego, Bartu Ahiska, and Daniel Falush. "A Bayesian approach to inferring the phylogenetic structure of communities from metagenomic data." en. In: *Genetics* 197.3 (July 2014), pp. 925–937. ISSN: 0016-6731, 1943-2631. DOI: [10.1534/genetics.114.161299](https://doi.org/10.1534/genetics.114.161299). URL: <http://dx.doi.org/10.1534/genetics.114.161299>.

- [383] Virginia Obregón, José L García, Ernesto García, Rubens López, and Pedro García. "Genome organization and molecular analysis of the temperate bacteriophage MM1 of *Streptococcus pneumoniae*." en. In: *J. Bacteriol.* 185.7 (Apr. 2003), pp. 2362–2368. ISSN: 0021-9193. DOI: [10.1128/jb.185.7.2362-2368.2003](https://doi.org/10.1128/jb.185.7.2362-2368.2003). URL: <http://dx.doi.org/10.1128/jb.185.7.2362-2368.2003>.
- [384] Office for National Statistics. *Deaths involving COVID-19, England and Wales*. <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsinvolvingcovid19englandandwales/deathsoccurringinjune2020>. Accessed: 2020-7-30. 2020. URL: <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsinvolvingcovid19englandandwales/deathsoccurringinjune2020>.
- [385] Office for National Statistics. *Deaths involving COVID-19 in the care sector, England and Wales*. <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/deathsinvolvingcovid19inthecaresectorenglandandwales/deathsoccurringupto12june2020andregisteredupto20june2020provisional>. Accessed: 2020-7-30. 2020. URL: <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/deathsinvolvingcovid19inthecaresectorenglandandwales/deathsoccurringupto12june2020andregisteredupto20june2020provisional>.
- [386] Abiodun D Ogunniyi, Marcin Grabowicz, David E Briles, Jan Cook, and James C Paton. "Development of a vaccine against invasive pneumococcal disease based on combinations of virulence proteins of *Streptococcus pneumoniae*." en. In: *Infect. Immun.* 75.1 (Jan. 2007), pp. 350–357. ISSN: 0019-9567. DOI: [10.1128/IAI.01103-06](https://doi.org/10.1128/IAI.01103-06). URL: <http://dx.doi.org/10.1128/IAI.01103-06>.
- [387] Brian D Ondov, Gabriel J Starrett, Anna Sappington, Aleksandra Kostic, Sergey Koren, Christopher B Buck, and Adam M Phillippy. "Mash Screen: high-throughput sequence containment estimation for genome discovery." en. In: *Genome Biol.* 20.1 (Nov. 2019), p. 232. ISSN: 1465-6906. DOI: [10.1186/s13059-019-1841-x](https://doi.org/10.1186/s13059-019-1841-x). URL: <http://dx.doi.org/10.1186/s13059-019-1841-x>.
- [388] Brian D Ondov, Todd J Treangen, Páll Melsted, Adam B Mallonee, Nicholas H Bergman, Sergey Koren, and Adam M Phillippy. "Mash: fast genome and metagenome distance estimation using MinHash." en. In: *Genome Biol.* 17.1 (June 2016), p. 132. ISSN: 1465-6906. DOI: [10.1186/s13059-016-0997-x](https://doi.org/10.1186/s13059-016-0997-x). URL: <http://dx.doi.org/10.1186/s13059-016-0997-x>.

- [389] Michelle K Paczosa, Rebecca J Silver, Anne L McCabe, Albert K Tai, Colin H McLeish, David W Lazinski, and Joan Mecsas. "Transposon Mutagenesis Screen of *Klebsiella pneumoniae* Identifies Multiple Genes Important for Resisting Antimicrobial Activities of Neutrophils in Mice." en. In: *Infect. Immun.* 88.4 (Mar. 2020). ISSN: 0019-9567, 1098-5522. DOI: [10.1128/IAI.00034-20](https://doi.org/10.1128/IAI.00034-20). URL: <http://dx.doi.org/10.1128/IAI.00034-20>.
- [390] 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." en. In: *Bioinformatics* 31.22 (Nov. 2015), pp. 3691–3693. ISSN: 1367-4803, 1367-4811. DOI: [10.1093/bioinformatics/btv421](https://doi.org/10.1093/bioinformatics/btv421). URL: <http://dx.doi.org/10.1093/bioinformatics/btv421>.
- [391] Donovan H Parks, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. "CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes." en. In: *Genome Res.* 25.7 (July 2015), pp. 1043–1055. ISSN: 1088-9051, 1549-5469. DOI: [10.1101/gr.186072.114](https://doi.org/10.1101/gr.186072.114). URL: <http://dx.doi.org/10.1101/gr.186072.114>.
- [392] Louis Pasteur. "Note sur la maladie nouvelle provoquee par la salive d'un enfant mort de la rage." In: *Bull. Acad. Med.* 10 (1881), pp. 94–103.
- [393] Jerome Pella and Michele Masuda. "The Gibbs and split-merge sampler for population mixture analysis from genetic data with incomplete baselines." In: *Can. J. Fish. Aquat. Sci.* 63.3 (Apr. 2011), pp. 576–596. ISSN: 0706-652X. DOI: [10.1139/f05-224](https://doi.org/10.1139/f05-224). URL: https://www.researchgate.net/publication/237175801_The_Gibbs_and_split-merge_sampler_for_population_mixture_analysis_from_genetic_data_with_incomplete_baselines.
- [394] Ye Peng, Shanmei Tang, et al. "MetaPGN: a pipeline for construction and graphical visualization of annotated pangenome networks." en. In: *Gigascience* 7.11 (Nov. 2018). ISSN: 2047-217X. DOI: [10.1093/gigascience/giy121](https://doi.org/10.1093/gigascience/giy121). URL: <http://dx.doi.org/10.1093/gigascience/giy121>.
- [395] Johan Pensar, Santeri Puranen, et al. "Genome-wide epistasis and co-selection study using mutual information." en. In: *Nucleic Acids Res.* (July 2019). ISSN: 0305-1048, 1362-4962. DOI: [10.1093/nar/gkz656](https://doi.org/10.1093/nar/gkz656). URL: <http://dx.doi.org/10.1093/nar/gkz656>.
- [396] Johan Pensar, Santeri Puranen, et al. *Genome-wide epistasis and co-selection study using mutual information*. 2019. DOI: [10.1093/nar/gkz656](https://doi.org/10.1093/nar/gkz656). URL: <http://dx.doi.org/10.1093/nar/gkz656>.

- [397] Ranawaka A P M Perera, Eugene Tso, et al. "SARS-CoV-2 Virus Culture and Subgenomic RNA for Respiratory Specimens from Patients with Mild Coronavirus Disease." en. In: *Emerg. Infect. Dis.* 26.11 (Nov. 2020), pp. 2701–2704. ISSN: 1080-6040, 1080-6059. DOI: [10.3201/eid2611.203219](https://doi.org/10.3201/eid2611.203219). URL: <http://dx.doi.org/10.3201/eid2611.203219>.
- [398] E V Pestova, L S Håvarstein, and D A Morrison. "Regulation of competence for genetic transformation in *Streptococcus pneumoniae* by an auto-induced peptide pheromone and a two-component regulatory system." In: *Mol. Microbiol.* 21.4 (1996), pp. 853–862. ISSN: 0950-382X. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2958.1996.501417.x>.
- [399] E V Pestova and D A Morrison. "Isolation and characterization of three *Streptococcus pneumoniae* transformation-specific loci by use of a lacZ reporter insertion vector." en. In: *J. Bacteriol.* 180.10 (May 1998), pp. 2701–2710. ISSN: 0021-9193. DOI: [10.1128/JB.180.10.2701-2710.1998](https://doi.org/10.1128/JB.180.10.2701-2710.1998). URL: <http://dx.doi.org/10.1128/JB.180.10.2701-2710.1998>.
- [400] D N A Pipelines, Benjamin Farr, et al. *COVID-19 ARTIC v3 Illumina library construction and sequencing protocol v4*. Tech. rep. May 2020. DOI: [10.17504/protocols.io.bgxjjxkn](https://doi.org/10.17504/protocols.io.bgxjjxkn). URL: <https://www.protocols.io/view/covid-19-artic-v3-illumina-library-construction-an-bgxjjxkn>.
- [401] M Pizza, V Scarlato, et al. "Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing." en. In: *Science* 287.5459 (Mar. 2000), pp. 1816–1820. ISSN: 0036-8075. DOI: [10.1126/science.287.5459.1816](https://doi.org/10.1126/science.287.5459.1816). URL: <http://dx.doi.org/10.1126/science.287.5459.1816>.
- [402] Jessica A Plante, Yang Liu, et al. "Spike mutation D614G alters SARS-CoV-2 fitness." In: *Nature* (Oct. 2020). ISSN: 0028-0836, 1476-4687. DOI: [10.1038/s41586-020-2895-3](https://doi.org/10.1038/s41586-020-2895-3). URL: <https://doi.org/10.1038/s41586-020-2895-3>.
- [403] Louis du Plessis, John T McCrone, et al. "Establishment & lineage dynamics of the SARS-CoV-2 epidemic in the UK." en. In: *medRxiv* (Oct. 2020), p. 2020.10.23.20218446. DOI: [10.1101/2020.10.23.20218446](https://doi.org/10.1101/2020.10.23.20218446). URL: <https://www.medrxiv.org/content/10.1101/2020.10.23.20218446v1.abstract>.
- [404] Alexandra Popa, Jakob-Wendelin Genger, et al. "Mutational dynamics and transmission properties of SARS-CoV-2 super-spreading events in Austria." en. July 2020. URL: <https://www.biorxiv.org/content/10.1101/2020.07.15.204339v1.abstract>.

- [405] L F Price, C C Drovandi, A Lee, and D J Nott. "Bayesian Synthetic Likelihood." In: *J. Comput. Graph. Stat.* 27.1 (Jan. 2018), pp. 1–11. ISSN: 1061-8600. DOI: [10.1080/10618600.2017.1302882](https://doi.org/10.1080/10618600.2017.1302882). URL: <https://doi.org/10.1080/10618600.2017.1302882>.
- [406] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. "Fast-Tree 2 – Approximately Maximum-Likelihood Trees for Large Alignments." In: *PLoS One* 5.3 (Mar. 2010), e9490. DOI: [10.1371/journal.pone.0009490](https://doi.org/10.1371/journal.pone.0009490). URL: <http://dx.doi.org/10.1371/journal.pone.0009490>.
- [407] J K Pritchard, M Stephens, and P Donnelly. "Inference of population structure using multilocus genotype data." en. In: *Genetics* 155.2 (June 2000), pp. 945–959. ISSN: 0016-6731. URL: <http://www.ncbi.nlm.nih.gov/pubmed/10835412>.
- [408] Marc Prudhomme, Laetitia Attaiech, Guillaume Sanchez, Bernard Martin, and Jean-Pierre Claverys. "Antibiotic stress induces genetic transformability in the human pathogen *Streptococcus pneumoniae*." en. In: *Science* 313.5783 (July 2006), pp. 89–92. ISSN: 0036-8075, 1095-9203. DOI: [10.1126/science.1127912](https://doi.org/10.1126/science.1127912). URL: <http://dx.doi.org/10.1126/science.1127912>.
- [409] Public Health England. COVID-19: number of outbreaks in care homes - management information. <https://www.gov.uk/government/statistical-data-sets/covid-19-number-of-outbreaks-in-care-homes-management-information>. Accessed: 2020-7-30. 2020. URL: <https://www.gov.uk/government/statistical-data-sets/covid-19-number-of-outbreaks-in-care-homes-management-information>.
- [410] Oliver Pybus, Andrew Rambaut, et al. Preliminary analysis of SARS-CoV-2 importation & establishment of UK transmission lineages. <https://virological.org/t/preliminary-analysis-of-sars-cov-2-importation-establishment-of-uk-transmission-lineages/507>. Accessed: 2020-10-23. June 2020. URL: <https://virological.org/t/preliminary-analysis-of-sars-cov-2-importation-establishment-of-uk-transmission-lineages/507>.
- [411] Josh Quick. "nCoV-2019 sequencing protocol v2 (GunIt)." In: *Protocols.io* (Apr. 2020). DOI: [10.17504/protocols.io.bdp7i5rn](https://doi.org/10.17504/protocols.io.bdp7i5rn). URL: <https://www.protocols.io/view/ncov-2019-sequencing-protocol-v2-bdp7i5rn.pdf>.
- [412] Aaron R Quinlan and Ira M Hall. "BEDTools: a flexible suite of utilities for comparing genomic features." en. In: *Bioinformatics* 26.6 (Mar. 2010), pp. 841–842. ISSN: 1367-4803, 1460-2059. DOI: [10.1093/bioinformatics/btq033](https://doi.org/10.1093/bioinformatics/btq033). URL: <http://bioinformatics.oxfordjournals.org/content/26/6/841>.

- [413] Manon Ragonnet-Cronin, Emma Hodcroft, Stéphane Hué, Esther Fearnhill, Valerie Delpech, Andrew J Leigh Brown, Samantha Lycett, and UK HIV Drug Resistance Database. "Automated analysis of phylogenetic clusters." en. In: *BMC Bioinformatics* 14 (Nov. 2013), p. 317. ISSN: 1471-2105. DOI: [10.1186/1471-2105-14-317](https://doi.org/10.1186/1471-2105-14-317). URL: <http://dx.doi.org/10.1186/1471-2105-14-317>.
- [414] Atif Rahman, Ingileif Hallgrímsdóttir, Michael Eisen, and Lior Pachter. "Association mapping from sequencing reads using k-mers." en. In: *Elife* 7 (June 2018). ISSN: 2050-084X. DOI: [10.7554/eLife.32920](https://doi.org/10.7554/eLife.32920). URL: <http://dx.doi.org/10.7554/eLife.32920>.
- [415] Anil Raj, Matthew Stephens, and Jonathan K Pritchard. "fast-STRUCTURE: variational inference of population structure in large SNP data sets." en. In: *Genetics* 197.2 (June 2014), pp. 573–589. ISSN: 0016-6731, 1943-2631. DOI: [10.1534/genetics.114.164350](https://doi.org/10.1534/genetics.114.164350). URL: <http://dx.doi.org/10.1534/genetics.114.164350>.
- [416] M Ramirez, E Severina, and A Tomasz. "A high incidence of prophage carriage among natural isolates of *Streptococcus pneumoniae*." en. In: *J. Bacteriol.* 181.12 (June 1999), pp. 3618–3625. ISSN: 0021-9193. DOI: [10.1128/JB.181.12.3618-3625.1999](https://doi.org/10.1128/JB.181.12.3618-3625.1999). URL: <http://dx.doi.org/10.1128/JB.181.12.3618-3625.1999>.
- [417] Oliver Ratmann, M Kate Grabowski, et al. "Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis." en. In: *Nat. Commun.* 10.1 (Mar. 2019), p. 1411. ISSN: 2041-1723. DOI: [10.1038/s41467-019-09139-4](https://doi.org/10.1038/s41467-019-09139-4). URL: <http://dx.doi.org/10.1038/s41467-019-09139-4>.
- [418] Oliver Ratmann, Ole Jørgensen, Trevor Hinkley, Michael Stumpf, Sylvia Richardson, and Carsten Wiuf. "Using likelihood-free inference to compare evolutionary dynamics of the protein networks of *H. pylori* and *P. falciparum*." en. In: *PLoS Comput. Biol.* 3.11 (Nov. 2007), e230. ISSN: 1553-734X, 1553-7358. DOI: [10.1371/journal.pcbi.0030230](https://doi.org/10.1371/journal.pcbi.0030230). URL: <http://dx.doi.org/10.1371/journal.pcbi.0030230>.
- [419] M Remm, C E Storm, and E L Sonnhammer. "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons." en. In: *J. Mol. Biol.* 314.5 (Dec. 2001), pp. 1041–1052. ISSN: 0022-2836. DOI: [10.1006/jmbi.2000.5197](https://doi.org/10.1006/jmbi.2000.5197). URL: <http://dx.doi.org/10.1006/jmbi.2000.5197>.

- [420] M B Rennels, K M Edwards, H L Keyserling, K S Reisinger, D A Hogerman, D V Madore, I Chang, P R Paradiso, F J Malinoski, and A Kimura. "Safety and immunogenicity of heptavalent pneumococcal vaccine conjugated to CRM197 in United States infants." en. In: *Pediatrics* 101.4 Pt 1 (Apr. 1998), pp. 604–611. ISSN: 0031-4005. URL: <https://www.ncbi.nlm.nih.gov/pubmed/9521941>.
- [421] Sylvia Richardson and Peter J Green. "On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion)." In: *J. R. Stat. Soc. Series B Stat. Methodol.* 59.4 (Nov. 1997), pp. 731–792. ISSN: 1369-7412, 1467-9868. DOI: [10.1111/1467-9868.00095](https://doi.org/10.1111/1467-9868.00095). URL: <http://doi.wiley.com/10.1111/1467-9868.00095>.
- [422] Mark D Robinson and Gordon K Smyth. "Moderated statistical tests for assessing differences in tag abundance." eng. In: *Bioinformatics* 23.21 (Nov. 2007), pp. 2881–2887. ISSN: 1367-4803, 1367-4811. DOI: [10.1093/bioinformatics/btm453](https://doi.org/10.1093/bioinformatics/btm453). URL: <http://dx.doi.org/10.1093/bioinformatics/btm453>.
- [423] A M Roche, A L Richard, J T Rahkola, E N Janoff, and others. "Antibody blocks acquisition of bacterial colonization through agglutination." In: *Mucosal Immunol.* (2015). ISSN: 1933-0219. URL: <https://www.nature.com/articles/mi201455>.
- [424] Alexandro Rodríguez-Rojas, Joshua Jay Kim, Paul R Johnston, Olga Makarova, Murat Eravci, Christoph Weise, Regine Hengge, and Jens Rolff. "Non-lethal exposure to H₂O₂ boosts bacterial survival and evolvability against oxidative stress." en. In: *PLoS Genet.* 16.3 (Mar. 2020), e1008649. ISSN: 1553-7390, 1553-7404. DOI: [10.1371/journal.pgen.1008649](https://doi.org/10.1371/journal.pgen.1008649). URL: <http://dx.doi.org/10.1371/journal.pgen.1008649>.
- [425] Claudia M Romero, David DeShazer, Tamara Feldblyum, Jacques Ravel, Donald Woods, H Stanley Kim, Yan Yu, Catherine M Ronning, and William C Nierman. "Genome sequence alterations detected upon passage of *Burkholderia mallei* ATCC 23344 in culture and in mammalian hosts." en. In: *BMC Genomics* 7 (Sept. 2006), p. 228. ISSN: 1471-2164. DOI: [10.1186/1471-2164-7-228](https://doi.org/10.1186/1471-2164-7-228). URL: <http://dx.doi.org/10.1186/1471-2164-7-228>.
- [426] 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*." en. In: *Plasmid* 58.1 (July 2007), pp. 51–60. ISSN: 0147-619X. DOI: [10.1016/j.plasmid.2006.12.006](https://doi.org/10.1016/j.plasmid.2006.12.006). URL: <http://dx.doi.org/10.1016/j.plasmid.2006.12.006>.

- [427] Hannah M Rowe, Erik Karlsson, Haley Echlin, Ti-Cheng Chang, Lei Wang, Tim van Opijnen, Stanley B Pounds, Stacey Schultz-Cherry, and Jason W Rosch. "Bacterial Factors Required for Transmission of *Streptococcus pneumoniae* in Mammalian Hosts." en. In: *Cell Host Microbe* 25.6 (June 2019), 884–891.e6. ISSN: 1931-3128, 1934-6069. DOI: [10.1016/j.chom.2019.04.012](https://doi.org/10.1016/j.chom.2019.04.012). URL: <http://dx.doi.org/10.1016/j.chom.2019.04.012>.
- [428] Daniel E Rozen, Lesley McGee, Bruce R Levin, and Keith P Klugman. "Fitness costs of fluoroquinolone resistance in *Streptococcus pneumoniae*." en. In: *Antimicrob. Agents Chemother.* 51.2 (Feb. 2007), pp. 412–416. ISSN: 0066-4804. DOI: [10.1128/AAC.01161-06](https://doi.org/10.1128/AAC.01161-06). URL: <http://dx.doi.org/10.1128/AAC.01161-06>.
- [429] J B Rubins, A H Paddock, D Charboneau, A M Berry, J C Paton, and E N Janoff. "Pneumolysin in pneumococcal adherence and colonization." en. In: *Microb. Pathog.* 25.6 (Dec. 1998), pp. 337–342. ISSN: 0882-4010. DOI: [10.1006/mpat.1998.0239](https://doi.org/10.1006/mpat.1998.0239). URL: <http://dx.doi.org/10.1006/mpat.1998.0239>.
- [430] Mourad Sabri, Roman Häuser, Marc Ouellette, Jing Liu, Mohammed Dehbi, Greg Moeck, Ernesto García, Björn Titz, Peter Uetz, and Sylvain Moineau. "Genome annotation and intraviral interactome for the *Streptococcus pneumoniae* virulent phage Dp-1." en. In: *J. Bacteriol.* 193.2 (Jan. 2011), pp. 551–562. ISSN: 0021-9193, 1098-5530. DOI: [10.1128/JB.01117-10](https://doi.org/10.1128/JB.01117-10). URL: <http://dx.doi.org/10.1128/JB.01117-10>.
- [431] Pavel Sagulenko, Vadim Puller, and Richard A Neher. "TreeTime: Maximum-likelihood phylodynamic analysis." en. In: *Virus Evol.* 4.1 (Jan. 2018), vxo042. ISSN: 2057-1577. DOI: [10.1093/ve/vex042](https://doi.org/10.1093/ve/vex042). URL: <http://dx.doi.org/10.1093/ve/vex042>.
- [432] Samir K Saha, Hassan M Al Emran, et al. "*Streptococcus pneumoniae* serotype-2 childhood meningitis in Bangladesh: a newly recognized pneumococcal infection threat." en. In: *PLoS One* 7.3 (Mar. 2012), e32134. ISSN: 1932-6203. DOI: [10.1371/journal.pone.0032134](https://doi.org/10.1371/journal.pone.0032134). URL: <http://dx.doi.org/10.1371/journal.pone.0032134>.
- [433] N Saitou and M Nei. "The neighbor-joining method: a new method for reconstructing phylogenetic trees." en. In: *Mol. Biol. Evol.* 4.4 (July 1987), pp. 406–425. ISSN: 0737-4038, 1537-1719. URL: <http://mbe.oxfordjournals.org/content/4/4/406>.
- [434] Steven L Salzberg. "Next-generation genome annotation: we still struggle to get it right." In: *Genome Biol.* 20.1 (May 2019), p. 92. ISSN: 1465-6906, 1474-760X. DOI: [10.1186/s13059-019-1715-2](https://doi.org/10.1186/s13059-019-1715-2). URL: <https://doi.org/10.1186/s13059-019-1715-2>.

- [435] F Sanger, S Nicklen, and A R Coulson. "DNA sequencing with chain-terminating inhibitors." en. In: *Proc. Natl. Acad. Sci. U. S. A.* 74.12 (Dec. 1977), pp. 5463–5467. ISSN: 0027-8424. DOI: [10.1073/pnas.74.12.5463](https://doi.org/10.1073/pnas.74.12.5463). URL: <http://dx.doi.org/10.1073/pnas.74.12.5463>.
- [436] Richard S Savage, Katherine Heller, Yang Xu, Zoubin Ghahramani, William M Truman, Murray Grant, Katherine J Denby, and David L Wild. "R/BHC: fast Bayesian hierarchical clustering for microarray data." en. In: *BMC Bioinformatics* 10 (Aug. 2009), p. 242. ISSN: 1471-2105. DOI: [10.1186/1471-2105-10-242](https://doi.org/10.1186/1471-2105-10-242). URL: <http://dx.doi.org/10.1186/1471-2105-10-242>.
- [437] Stanley G Sawicki, Dorothea L Sawicki, and Stuart G Siddell. "A contemporary view of coronavirus transcription." en. In: *J. Virol.* 81.1 (Jan. 2007), pp. 20–29. ISSN: 0022-538X. DOI: [10.1128/JVI.01358-06](https://doi.org/10.1128/JVI.01358-06). URL: <http://dx.doi.org/10.1128/JVI.01358-06>.
- [438] Susanne Schneiker, Olena Perlova, et al. "Complete genome sequence of the myxobacterium *Sorangium cellulosum*." en. In: *Nat. Biotechnol.* 25.11 (Nov. 2007), pp. 1281–1289. ISSN: 1087-0156. DOI: [10.1038/nbt1354](https://doi.org/10.1038/nbt1354). URL: <http://dx.doi.org/10.1038/nbt1354>.
- [439] J A Scott, A J Hall, et al. "Serogroup-specific epidemiology of *Streptococcus pneumoniae*: associations with age, sex, and geography in 7,000 episodes of invasive disease." en. In: *Clin. Infect. Dis.* 22.6 (June 1996), pp. 973–981. ISSN: 1058-4838. DOI: [10.1093/clinids/22.6.973](https://doi.org/10.1093/clinids/22.6.973). URL: <http://dx.doi.org/10.1093/clinids/22.6.973>.
- [440] Alexander Sczyrba, Peter Hofmann, et al. "Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software." en. In: *Nat. Methods* 14.11 (Nov. 2017), pp. 1063–1071. ISSN: 1548-7091, 1548-7105. DOI: [10.1038/nmeth.4458](https://doi.org/10.1038/nmeth.4458). URL: <http://dx.doi.org/10.1038/nmeth.4458>.
- [441] T Seeman. *Shovill: faster SPAdes assembly of Illumina reads (v0.9.0)*. 2018.
- [442] Torsten Seemann. *snp-dists*. URL: <https://github.com/tseemann/snp-dists>.
- [443] Torsten Seemann. "Prokka: rapid prokaryotic genome annotation." en. In: *Bioinformatics* 30.14 (July 2014), pp. 2068–2069. ISSN: 1367-4803, 1367-4811. DOI: [10.1093/bioinformatics/btu153](https://doi.org/10.1093/bioinformatics/btu153). URL: <http://dx.doi.org/10.1093/bioinformatics/btu153>.

- [444] Torsten Seemann, Courtney R Lane, et al. "Tracking the COVID-19 pandemic in Australia using genomics." en. In: *Nat. Commun.* 11.1 (Sept. 2020), p. 4376. ISSN: 2041-1723. DOI: [10.1038/s41467-020-18314-x](https://doi.org/10.1038/s41467-020-18314-x). URL: <http://dx.doi.org/10.1038/s41467-020-18314-x>.
- [445] Emil Senger. "Bacteriologische Untersuchungen über die Pneumonie und pneumonische Metastasen." de. In: *Archiv für experimentelle Pathologie und Pharmakologie* 20.5 (Jan. 1886), pp. 389–410. ISSN: 1432-1912, 1432-1912. DOI: [10.1007/BF01831345](https://doi.org/10.1007/BF01831345). URL: <https://link.springer.com/article/10.1007/BF01831345>.
- [446] Terry L Shaffer. "A unified approach to analyzing nest success." In: *Auk* 121.2 (2004), p. 526. ISSN: 0004-8038, 0006-3657. DOI: [10.1642/0004-8038\(2004\)121\[0526:AUATAN\]2.0.CO;2](https://doi.org/10.1642/0004-8038(2004)121[0526:AUATAN]2.0.CO;2). URL: <https://academic.oup.com/auk/article/121/2/526-540/5147518>.
- [447] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. "Cytoscape: a software environment for integrated models of biomolecular interaction networks." en. In: *Genome Res.* 13.11 (Nov. 2003), pp. 2498–2504. ISSN: 1088-9051. DOI: [10.1101/gr.1239303](https://doi.org/10.1101/gr.1239303). URL: <http://dx.doi.org/10.1101/gr.1239303>.
- [448] E D Shapiro and J D Clemens. "A controlled evaluation of the protective efficacy of pneumococcal vaccine for patients at high risk of serious pneumococcal infections." en. In: *Ann. Intern. Med.* 101.3 (Sept. 1984), pp. 325–330. ISSN: 0003-4819. DOI: [10.7326/0003-4819-101-3-325](https://doi.org/10.7326/0003-4819-101-3-325). URL: <http://dx.doi.org/10.7326/0003-4819-101-3-325>.
- [449] Pamela Shen, John A Lees, Gavyn Chern Wei Bee, Sam P Brown, and Jeffrey N Weiser. "Pneumococcal quorum sensing drives an asymmetric owner-intruder competitive strategy during carriage via the competence regulon." en. In: *Nat Microbiol* 4.1 (Jan. 2019), pp. 198–208. ISSN: 2058-5276. DOI: [10.1038/s41564-018-0314-4](https://doi.org/10.1038/s41564-018-0314-4). URL: <http://dx.doi.org/10.1038/s41564-018-0314-4>.
- [450] Zijie Shen, Yan Xiao, et al. *Genomic Diversity of Severe Acute Respiratory Syndrome–Coronavirus 2 in Patients With Coronavirus Disease 2019*. 2020. DOI: [10.1093/cid/ciaa203](https://doi.org/10.1093/cid/ciaa203). URL: <http://dx.doi.org/10.1093/cid/ciaa203>.
- [451] Fabian Sievers, Andreas Wilm, et al. "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega." en. In: *Mol. Syst. Biol.* 7.1 (Oct. 2011), p. 539. ISSN: 1744-4292. DOI: [10.1038/msb.2011.75](https://doi.org/10.1038/msb.2011.75). URL: <http://dx.doi.org/10.1038/msb.2011.75>.

- [452] Martin Simonsen, Thomas Mailund, and Christian N S Pedersen. "Rapid Neighbour-Joining." In: *Algorithms in Bioinformatics*. Springer Berlin Heidelberg, 2008, pp. 113–122. DOI: [10.1007/978-3-540-87361-7_10](https://doi.org/10.1007/978-3-540-87361-7_10). URL: http://dx.doi.org/10.1007/978-3-540-87361-7_10.
- [453] S A Sisson, Y Fan, and Mark M Tanaka. "Sequential Monte Carlo without likelihoods." en. In: *Proc. Natl. Acad. Sci. U. S. A.* 104.6 (Feb. 2007), pp. 1760–1765. ISSN: 0027-8424. DOI: [10.1073/pnas.0607208104](https://doi.org/10.1073/pnas.0607208104). URL: <http://dx.doi.org/10.1073/pnas.0607208104>.
- [454] Marcin J Skwark, Nicholas J Croucher, et al. "Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis." en. In: *PLoS Genet.* 13.2 (Feb. 2017), e1006508. ISSN: 1553-7390, 1553-7404. DOI: [10.1371/journal.pgen.1006508](https://doi.org/10.1371/journal.pgen.1006508). URL: <http://dx.doi.org/10.1371/journal.pgen.1006508>.
- [455] Chris Smillie, M Pilar Garcillán-Barcia, M Victoria Francia, Eduardo P C Rocha, and Fernando de la Cruz. "Mobility of plasmids." en. In: *Microbiol. Mol. Biol. Rev.* 74.3 (Sept. 2010), pp. 434–452. ISSN: 1092-2172, 1098-5557. DOI: [10.1128/MMBR.00020-10](https://doi.org/10.1128/MMBR.00020-10). URL: <http://dx.doi.org/10.1128/MMBR.00020-10>.
- [456] 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*." en. In: *Antimicrob. Agents Chemother.* 45.8 (Aug. 2001), pp. 2393–2396. ISSN: 0066-4804. DOI: [10.1128/AAC.45.8.2393-2396.2001](https://doi.org/10.1128/AAC.45.8.2393-2396.2001). URL: <http://dx.doi.org/10.1128/AAC.45.8.2393-2396.2001>.
- [457] J M Smith, E J Feil, and N H Smith. "Population structure and evolutionary dynamics of pathogenic bacteria." en. In: *Bioessays* 22.12 (Dec. 2000), pp. 1115–1122. ISSN: 0265-9247. DOI: [10.1002/1521-1878\(200012\)22:12<1115::AID-BIES9>3.0.CO;2-R](https://doi.org/10.1002/1521-1878(200012)22:12<1115::AID-BIES9>3.0.CO;2-R). URL: [http://dx.doi.org/10.1002/1521-1878\(200012\)22:12<1115::AID-BIES9>3.0.CO;2-R](http://dx.doi.org/10.1002/1521-1878(200012)22:12<1115::AID-BIES9>3.0.CO;2-R).
- [458] M D Smith and W R Guild. *A plasmid in Streptococcus pneumoniae*. 1979. DOI: [10.1128/jb.137.2.735-739.1979](https://doi.org/10.1128/jb.137.2.735-739.1979). URL: <http://dx.doi.org/10.1128/jb.137.2.735-739.1979>.
- [459] Lars Snipen, Trygve Almøy, and David W Ussery. "Microbial comparative pan-genomics using binomial mixture models." en. In: *BMC Genomics* 10 (Aug. 2009), p. 385. ISSN: 1471-2164. DOI: [10.1186/1471-2164-10-385](https://doi.org/10.1186/1471-2164-10-385). URL: <http://dx.doi.org/10.1186/1471-2164-10-385>.

- [460] Robert Reuven Sokal and Charles Duncan Michener. *A Statistical Method for Evaluating Systematic Relationships*. en. University of Kansas, 1958. URL: <https://market.android.com/details?id=book-o1B1HAAACAAJ>.
- [461] Martin Šošić and Mile Šikić. "Edlib: a C/C++ library for fast, exact sequence alignment using edit distance." In: *Bioinformatics* 33.9 (2017), pp. 1394–1395. ISSN: 1367-4803. URL: <https://academic.oup.com/bioinformatics/article-abstract/33/9/1394/2964763>.
- [462] Paul R Staab and Dirk Metzler. "Coala: an R framework for coalescent simulation." en. In: *Bioinformatics* 32.12 (June 2016), pp. 1903–1904. ISSN: 1367-4803, 1367-4811. DOI: [10.1093/bioinformatics/btw098](https://doi.org/10.1093/bioinformatics/btw098). URL: <http://dx.doi.org/10.1093/bioinformatics/btw098>.
- [463] Paul R Staab, Sha Zhu, Dirk Metzler, and Gerton Lunter. "scrm: efficiently simulating long sequences using the approximated coalescent with recombination." en. In: *Bioinformatics* 31.10 (May 2015), pp. 1680–1682. ISSN: 1367-4803, 1367-4811. DOI: [10.1093/bioinformatics/btu861](https://doi.org/10.1093/bioinformatics/btu861). URL: <http://dx.doi.org/10.1093/bioinformatics/btu861>.
- [464] Alexandros Stamatakis. "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies." en. In: *Bioinformatics* 30.9 (May 2014), pp. 1312–1313. ISSN: 1367-4803, 1367-4811. DOI: [10.1093/bioinformatics/btu033](https://doi.org/10.1093/bioinformatics/btu033). URL: <http://dx.doi.org/10.1093/bioinformatics/btu033>.
- [465] Michael L Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York, NY, 1999. ISBN: 9781461271666. DOI: [10.1007/978-1-4612-1494-6](https://doi.org/10.1007/978-1-4612-1494-6). URL: <https://link.springer.com/book/10.1007/978-1-4612-1494-6>.
- [466] Michael Stein. "Large Sample Properties of Simulations Using Latin Hypercube Sampling." In: *Technometrics* 29.2 (1987), pp. 143–151. ISSN: 0040-1706. DOI: [10.2307/1269769](https://doi.org/10.2307/1269769). URL: <http://www.jstor.org/stable/1269769>.
- [467] G M Sternberg. "A fatal form of septicemia in the rabbit produced by subcutaneous injection of human saliva." In: *Natl Board of Health Bull* 3.87 (1881), p. 108.
- [468] James Stimson, Jennifer Gardy, Barun Mathema, Valeriu Crudu, Ted Cohen, and Caroline Colijn. "Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions." en. In: *Mol. Biol. Evol.* 36.3 (Mar. 2019), pp. 587–603. ISSN: 0737-4038, 1537-1719. DOI: [10.1093/molbev/msy242](https://doi.org/10.1093/molbev/msy242). URL: <http://dx.doi.org/10.1093/molbev/msy242>.

- [469] Alexander Strehl and Joydeep Ghosh. "Cluster Ensembles — A Knowledge Reuse Framework for Combining Multiple Partitions." In: *J. Mach. Learn. Res.* 3, Dec (2002), pp. 583–617. ISSN: 1532-4435, 1533-7928. URL: <http://www.jmlr.org/papers/volume3/strehl02a/strehl02a.pdf>.
- [470] Jeet Sukumaran and Mark T Holder. "DendroPy: a Python library for phylogenetic computing." en. In: *Bioinformatics* 26.12 (June 2010), pp. 1569–1571. ISSN: 1367-4803, 1367-4811. DOI: [10.1093/bioinformatics/btq228](https://doi.org/10.1093/bioinformatics/btq228). URL: <http://dx.doi.org/10.1093/bioinformatics/btq228>.
- [471] R K Syrjänen, T M Kilpi, T H Kaijalainen, E E Herva, and A K Takala. "Nasopharyngeal carriage of *Streptococcus pneumoniae* in Finnish children younger than 2 years old." en. In: *J. Infect. Dis.* 184.4 (Aug. 2001), pp. 451–459. ISSN: 0022-1899. DOI: [10.1086/322048](https://doi.org/10.1086/322048). URL: <http://dx.doi.org/10.1086/322048>.
- [472] K Takahashi and M Nei. "Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used." en. In: *Mol. Biol. Evol.* 17.8 (Aug. 2000), pp. 1251–1258. ISSN: 0737-4038. DOI: [10.1093/oxfordjournals.molbev.a026408](https://doi.org/10.1093/oxfordjournals.molbev.a026408). URL: <http://dx.doi.org/10.1093/oxfordjournals.molbev.a026408>.
- [473] Hua Tang, Marc Coram, Pei Wang, Xiaofeng Zhu, and Neil Risch. "Reconstructing genetic ancestry blocks in admixed individuals." en. In: *Am. J. Hum. Genet.* 79.1 (July 2006), pp. 1–12. ISSN: 0002-9297. DOI: [10.1086/504302](https://doi.org/10.1086/504302). URL: <http://dx.doi.org/10.1086/504302>.
- [474] R L Tatusov, E V Koonin, and D J Lipman. "A genomic perspective on protein families." en. In: *Science* 278.5338 (Oct. 1997), pp. 631–637. ISSN: 0036-8075. DOI: [10.1126/science.278.5338.631](https://doi.org/10.1126/science.278.5338.631). URL: <http://dx.doi.org/10.1126/science.278.5338.631>.
- [475] S Tavaré, D J Balding, R C Griffiths, and P Donnelly. "Inferring coalescence times from DNA sequence data." en. In: *Genetics* 145.2 (Feb. 1997), pp. 505–518. ISSN: 0016-6731. URL: <https://www.ncbi.nlm.nih.gov/pubmed/9071603>.
- [476] Simon Tavaré. "Some probabilistic and statistical problems in the analysis of DNA sequences." In: *Lectures on mathematics in the life sciences* 17.2 (1986), pp. 57–86. URL: https://books.google.com/books?hl=en&lr=&id=8aI1phhOKhgC&oi=fnd&pg=PA57&dq=tavare+1986&ots=roKy4NEdPf&sig=_tkqlyDTKly8askP1qScJ95a4NU.
- [477] Hervé Tettelin, Vega Massignani, et al. "Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome"." en. In: *Proc. Natl.*

- Acad. Sci. U. S. A.* 102.39 (Sept. 2005), pp. 13950–13955. ISSN: 0027-8424. DOI: [10.1073/pnas.0506758102](https://doi.org/10.1073/pnas.0506758102). URL: <http://dx.doi.org/10.1073/pnas.0506758102>.
- [478] The Border Consortium. *Refugee and IDP camp populations: January 2016*. Mae Sot. <http://www.theborderconsortium.org/media/66743/2016-01-jan-map-tbc-unhcr.pdf>. 2016. URL: <http://www.theborderconsortium.org/media/66743/2016-01-jan-map-tbc-unhcr.pdf>.
- [479] Christopher M Thomas and Kaare M Nielsen. “Mechanisms of, and barriers to, horizontal gene transfer between bacteria.” en. In: *Nat. Rev. Microbiol.* 3.9 (Sept. 2005), pp. 711–721. ISSN: 1740-1526. DOI: [10.1038/nrmicro1234](https://doi.org/10.1038/nrmicro1234). URL: <http://dx.doi.org/10.1038/nrmicro1234>.
- [480] W S Tillet and T Francis. “Reacciones serológicas en neumonía con una fracción somática no-proteica del neumococo.” In: *J. Exp. Med.* (1930). ISSN: 0022-1007.
- [481] Thu-Hien To, Matthieu Jung, Samantha Lycett, and Olivier Gascuel. “Fast Dating Using Least-Squares Criteria and Algorithms.” en. In: *Syst. Biol.* 65.1 (Jan. 2016), pp. 82–97. ISSN: 1063-5157, 1076-836X. DOI: [10.1093/sysbio/syv068](https://doi.org/10.1093/sysbio/syv068). URL: <http://dx.doi.org/10.1093/sysbio/syv068>.
- [482] A Tomasz and J L Mosser. “On the nature of the pneumococcal activator substance.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 55.1 (Jan. 1966), pp. 58–66. ISSN: 0027-8424. DOI: [10.1073/pnas.55.1.58](https://doi.org/10.1073/pnas.55.1.58). URL: <http://dx.doi.org/10.1073/pnas.55.1.58>.
- [483] Gerry Tonkin-Hill. *pairsnp*. Aug. 2018. URL: <https://github.com/gtonkinhill/pairsnp>.
- [484] 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.” In: *Wellcome Open Res* 3 (July 2018), p. 93. ISSN: 2398-502X. DOI: [10.12688/wellcomeopenres.14694.1](https://doi.org/10.12688/wellcomeopenres.14694.1). URL: <https://wellcomeopenresearch.org/articles/3-93/v1>.
- [485] Gerry Tonkin-Hill, Neil MacAlasdair, et al. “Producing polished prokaryotic pangenomes with the Panaroo pipeline.” In: *Genome Biol.* 21.1 (July 2020), p. 180. ISSN: 1465-6906, 1474-760X. DOI: [10.1186/s13059-020-02090-4](https://doi.org/10.1186/s13059-020-02090-4). URL: <https://doi.org/10.1186/s13059-020-02090-4>.
- [486] Krzysztof Trzciński, Claudette M Thompson, Andrea M Gilbey, Christopher G Dowson, and Marc Lipsitch. “Incremental Increase in Fitness Cost with Increased β -Lactam Resistance in Pneumococci Evaluated by Competition in an Infant Rat Nasal Colonization Model.” In: *J. Infect. Dis.* 193.9 (May 2006), pp. 1296–1303. ISSN: 0022-1899. DOI: [10.1086/501367](https://doi.org/10.1086/501367). URL: <https://doi.org/10.1086/501367>.

- academic.oup.com/jid/article-abstract/193/9/1296/1013899.
- [487] Krzysztof Trzciński, Claudette M Thompson, Andrea M Gilbey, Christopher G Dowson, and Marc Lipsitch. "Incremental increase in fitness cost with increased β -lactam resistance in pneumococci evaluated by competition in an infant rat nasal colonization model." In: *J. Infect. Dis.* 193.9 (2006), pp. 1296–1303. ISSN: 0022-1899. URL: <https://academic.oup.com/jid/article-abstract/193/9/1296/1013899>.
- [488] A H Tu, R L Fulgham, M A McCrory, D E Briles, and A J Szalai. "Pneumococcal surface protein A inhibits complement activation by *Streptococcus pneumoniae*." en. In: *Infect. Immun.* 67.9 (Sept. 1999), pp. 4720–4724. ISSN: 0019-9567. DOI: [10.1128/IAI.67.9.4720-4724.1999](https://doi.org/10.1128/IAI.67.9.4720-4724.1999). URL: <http://dx.doi.org/10.1128/IAI.67.9.4720-4724.1999>.
- [489] Clare Turnbull, Shahana Ahmed, et al. "Genome-wide association study identifies five new breast cancer susceptibility loci." en. In: *Nat. Genet.* 42.6 (June 2010), pp. 504–507. ISSN: 1061-4036, 1546-1718. DOI: [10.1038/ng.586](https://doi.org/10.1038/ng.586). URL: <http://dx.doi.org/10.1038/ng.586>.
- [490] Paul Turner, Jason Hinds, Claudia Turner, Auscharee Jankhot, Katherine Gould, Stephen D Bentley, François Nosten, and David Goldblatt. "Improved detection of nasopharyngeal colonization by multiple pneumococcal serotypes by use of latex agglutination or molecular serotyping by microarray." en. In: *J. Clin. Microbiol.* 49.5 (May 2011), pp. 1784–1789. ISSN: 0095-1137, 1098-660X. DOI: [10.1128/JCM.00157-11](https://doi.org/10.1128/JCM.00157-11). URL: <http://dx.doi.org/10.1128/JCM.00157-11>.
- [491] 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." en. In: *PLoS One* 7.5 (May 2012), e38271. ISSN: 1932-6203. DOI: [10.1371/journal.pone.0038271](https://doi.org/10.1371/journal.pone.0038271). URL: <http://dx.doi.org/10.1371/journal.pone.0038271>.
- [492] K Ubukata, M Konno, and R Fujii. "Transduction of drug resistance to tetracycline, chloramphenicol, macrolides, lincomycin and clindamycin with phages induced from *Streptococcus pyogenes*." en. In: *J. Antibiot.* 28.9 (Sept. 1975), pp. 681–688. ISSN: 0021-8820. DOI: [10.7164/antibiotics.28.681](https://doi.org/10.7164/antibiotics.28.681). URL: <http://dx.doi.org/10.7164/antibiotics.28.681>.

- [493] Sheila Ulivi, Massimo Mezzavilla, and Paolo Gasparini. "Genetics of eye colours in different rural populations on the Silk Road." en. In: *Eur. J. Hum. Genet.* 21.11 (Nov. 2013), pp. 1320–1323. ISSN: 1018-4813, 1476-5438. DOI: [10.1038/ejhg.2013.41](https://doi.org/10.1038/ejhg.2013.41). URL: <http://dx.doi.org/10.1038/ejhg.2013.41>.
- [494] Pietro E Varaldo, Maria Pia Montanari, and Eleonora Giovanetti. "Genetic elements responsible for erythromycin resistance in streptococci." en. In: *Antimicrob. Agents Chemother.* 53.2 (Feb. 2009), pp. 343–353. ISSN: 0066-4804, 1098-6596. DOI: [10.1128/AAC.00781-08](https://doi.org/10.1128/AAC.00781-08). URL: <http://dx.doi.org/10.1128/AAC.00781-08>.
- [495] E Varon, C Levy, F De La Rocque, M Boucherat, D Deforche, I Podglajen, M Navel, and R Cohen. "Impact of antimicrobial therapy on nasopharyngeal carriage of *Streptococcus pneumoniae*, *Haemophilus influenzae*, and *Branhamella catarrhalis* in children with respiratory tract infections." en. In: *Clin. Infect. Dis.* 31.2 (Aug. 2000), pp. 477–481. ISSN: 1058-4838. DOI: [10.1086/313981](https://doi.org/10.1086/313981). URL: <http://dx.doi.org/10.1086/313981>.
- [496] E M Volz and S D W Frost. "Scalable relaxed clock phylogenetic dating." In: *Virus Evol* 3.2 (July 2017). DOI: [10.1093/ve/vex025](https://doi.org/10.1093/ve/vex025). URL: <https://academic.oup.com/ve/article/4100592>.
- [497] Erik M Volz, Verity Hill, et al. "Evaluating the effects of SARS-CoV-2 Spike mutation D614G on transmissibility and pathogenicity." en. In: *medRxiv* (Sept. 2020), p. 2020.07.31.20166082. DOI: [10.1101/2020.07.31.20166082](https://doi.org/10.1101/2020.07.31.20166082). URL: <https://www.medrxiv.org/content/10.1101/2020.07.31.20166082v2.abstract>.
- [498] Brian Wahl, Katherine L O'Brien, et al. "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." In: *The Lancet Global Health* 6.7 (July 2018), e744–e757. ISSN: 2214-109X. DOI: [10.1016/S2214-109X\(18\)30247-X](https://doi.org/10.1016/S2214-109X(18)30247-X). URL: [http://dx.doi.org/10.1016/S2214-109X\(18\)30247-X](http://dx.doi.org/10.1016/S2214-109X(18)30247-X).
- [499] Dawei Wang, Bo Hu, et al. "Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China." In: *JAMA* 323.11 (Mar. 2020), pp. 1061–1069. ISSN: 0098-7484. DOI: [10.1001/jama.2020.1585](https://doi.org/10.1001/jama.2020.1585). URL: <https://jamanetwork.com/journals/jama/fullarticle/2761044>.
- [500] Lianming Wang and David B Dunson. "Fast Bayesian Inference in Dirichlet Process Mixture Models." en. In: *J. Comput. Graph. Stat.* 20.1 (Jan. 2011). ISSN: 1061-8600. DOI: [10.1198/jcgs.2010.07081](https://doi.org/10.1198/jcgs.2010.07081). URL: <http://dx.doi.org/10.1198/jcgs.2010.07081>.

- [501] Helen Ward, Christina J Atchison, et al. *Antibody prevalence for SARS-CoV-2 in England following first peak of the pandemic: REACT2 study in 100,000 adults*. DOI: [10.1101/2020.08.12.20173690](https://doi.org/10.1101/2020.08.12.20173690). URL: <http://dx.doi.org/10.1101/2020.08.12.20173690>.
- [502] Hua Wei and Leiv Sigve Håvarstein. "Fratricide is essential for efficient gene transfer between pneumococci in biofilms." en. In: *Appl. Environ. Microbiol.* 78.16 (Aug. 2012), pp. 5897–5905. ISSN: 0099-2240, 1098-5336. DOI: [10.1128/AEM.01343-12](https://doi.org/10.1128/AEM.01343-12). URL: <http://dx.doi.org/10.1128/AEM.01343-12>.
- [503] Lucy A Weinert, Roy R Chaudhuri, et al. "Genomic signatures of human and animal disease in the zoonotic pathogen *Streptococcus suis*." en. In: *Nat. Commun.* 6 (Mar. 2015), p. 6740. ISSN: 2041-1723. DOI: [10.1038/ncomms7740](https://doi.org/10.1038/ncomms7740). URL: <http://dx.doi.org/10.1038/ncomms7740>.
- [504] B Weisblum. "Erythromycin resistance by ribosome modification." en. In: *Antimicrob. Agents Chemother.* 39.3 (Mar. 1995), pp. 577–585. ISSN: 0066-4804. DOI: [10.1128/aac.39.3.577](https://doi.org/10.1128/aac.39.3.577). URL: <http://dx.doi.org/10.1128/aac.39.3.577>.
- [505] Jeffrey N Weiser, Daniela M Ferreira, and James C Paton. "Streptococcus pneumoniae: transmission, colonization and invasion." en. In: *Nat. Rev. Microbiol.* 16.6 (June 2018), pp. 355–367. ISSN: 1740-1526, 1740-1534. DOI: [10.1038/s41579-018-0001-8](https://doi.org/10.1038/s41579-018-0001-8). URL: <http://dx.doi.org/10.1038/s41579-018-0001-8>.
- [506] A M Whatmore and C G Dowson. "The autolysin-encoding gene (*lytA*) of *Streptococcus pneumoniae* displays restricted allelic variation despite localized recombination events with genes of pneumococcal bacteriophage encoding cell wall lytic enzymes." en. In: *Infect. Immun.* 67.9 (Sept. 1999), pp. 4551–4556. ISSN: 0019-9567. DOI: [10.1128/IAI.67.9.4551-4556.1999](https://doi.org/10.1128/IAI.67.9.4551-4556.1999). URL: <http://dx.doi.org/10.1128/IAI.67.9.4551-4556.1999>.
- [507] C A Widdowson, K P Klugman, and others. "Identification of the tetracycline resistance gene, *tet* (O), in *Streptococcus pneumoniae*." In: *Antimicrob. Agents Chemother.* (1996). ISSN: 0066-4804. URL: <https://aac.asm.org/content/40/12/2891.short>.
- [508] Andreas Wilm, Pauline Poh Kim Aw, Denis Bertrand, Grace Hui Ting Yeo, Swee Hoe Ong, Chang Hua Wong, Chiea Chuen Khor, Rosemary Petric, Martin Lloyd Hibberd, and Niranjana Nagarajan. "LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets." en. In: *Nucleic Acids Res.* 40.22 (Dec. 2012), pp. 11189–11201. ISSN: 0305-1048, 1362-4962. DOI: [10.1093/nar/gks918](https://doi.org/10.1093/nar/gks918). URL: <http://dx.doi.org/10.1093/nar/gks918>.

- [509] Daniel J Wilson and CRyPTIC Consortium. "GenomeMap: Within-Species Genome-Wide dN/dS Estimation from over 10,000 Genomes." en. In: *Mol. Biol. Evol.* 37.8 (Aug. 2020), pp. 2450–2460. ISSN: 0737-4038, 1537-1719. DOI: [10.1093/molbev/msaa069](https://doi.org/10.1093/molbev/msaa069). URL: <http://dx.doi.org/10.1093/molbev/msaa069>.
- [510] Daniel N Wilson. "The ABC of Ribosome-Related Antibiotic Resistance." en. In: *MBio* 7.3 (May 2016). ISSN: 2150-7511. DOI: [10.1128/mBio.00598-16](https://doi.org/10.1128/mBio.00598-16). URL: <http://dx.doi.org/10.1128/mBio.00598-16>.
- [511] Roman Wölfel, Victor M Corman, et al. "Virological assessment of hospitalized patients with COVID-2019." en. In: *Nature* 581.7809 (May 2020), pp. 465–469. ISSN: 0028-0836, 1476-4687. DOI: [10.1038/s41586-020-2196-x](https://doi.org/10.1038/s41586-020-2196-x). URL: <http://dx.doi.org/10.1038/s41586-020-2196-x>.
- [512] Derrick E Wood, Jennifer Lu, and Ben Langmead. "Improved metagenomic analysis with Kraken 2." en. In: *Genome Biol.* 20.1 (Nov. 2019), p. 257. ISSN: 1465-6906. DOI: [10.1186/s13059-019-1891-0](https://doi.org/10.1186/s13059-019-1891-0). URL: <http://dx.doi.org/10.1186/s13059-019-1891-0>.
- [513] Derrick E Wood and Steven L Salzberg. "Kraken: ultrafast metagenomic sequence classification using exact alignments." en. In: *Genome Biol.* 15.3 (Mar. 2014), R46. ISSN: 1465-6906. DOI: [10.1186/gb-2014-15-3-r46](https://doi.org/10.1186/gb-2014-15-3-r46). URL: <http://dx.doi.org/10.1186/gb-2014-15-3-r46>.
- [514] Simon N Wood. "Statistical inference for noisy nonlinear ecological dynamic systems." en. In: *Nature* 466.7310 (Aug. 2010), pp. 1102–1104. ISSN: 0028-0836, 1476-4687. DOI: [10.1038/nature09319](https://doi.org/10.1038/nature09319). URL: <http://dx.doi.org/10.1038/nature09319>.
- [515] Colin J Worby, Marc Lipsitch, and William P Hanage. "Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data." en. In: *PLoS Comput. Biol.* 10.3 (Mar. 2014), e1003549. ISSN: 1553-734X, 1553-7358. DOI: [10.1371/journal.pcbi.1003549](https://doi.org/10.1371/journal.pcbi.1003549). URL: <http://dx.doi.org/10.1371/journal.pcbi.1003549>.
- [516] Colin J Worby, Marc Lipsitch, and William P Hanage. "Shared Genomic Variants: Identification of Transmission Routes Using Pathogen Deep-Sequence Data." en. In: *Am. J. Epidemiol.* 186.10 (Nov. 2017), pp. 1209–1216. ISSN: 0002-9262, 1476-6256. DOI: [10.1093/aje/kwx182](https://doi.org/10.1093/aje/kwx182). URL: <http://dx.doi.org/10.1093/aje/kwx182>.
- [517] Colin J Worby and Timothy D Read. "'SEEDY' (Simulation of Evolutionary and Epidemiological Dynamics): An R Package to Follow Accumulation of Within-Host Mutation in Pathogens." en. In: *PLoS One* 10.6 (June 2015), e0129745. ISSN: 1932-6203.

- DOI: [10.1371/journal.pone.0129745](https://doi.org/10.1371/journal.pone.0129745). URL: <http://dx.doi.org/10.1371/journal.pone.0129745>.
- [518] World Health Organisation. *Global tuberculosis report*. Tech. rep. WHO, Oct. 2019. URL: <https://apps.who.int/iris/bitstream/handle/10665/329368/9789241565714-eng.pdf>.
- [519] World Health Organization. *Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics*. Tech. rep. World Health Organization, 2017. URL: https://www.who.int/medicines/publications/WHO-PPL-Short-Summary_25Feb-ET_NM_WHO.pdf.
- [520] World Health Organization. *Weekly update on COVID-19*. Tech. rep. World Health Organization, Oct. 2020. URL: <https://www.who.int/publications/m/item/weekly-update-on-covid-19-9-october-2020>.
- [521] S Wright. "Genetical structure of populations." en. In: *Nature* 166.4215 (Aug. 1950), pp. 247–249. ISSN: 0028-0836. DOI: [10.1038/166247a0](https://doi.org/10.1038/166247a0). URL: <http://dx.doi.org/10.1038/166247a0>.
- [522] Fan Wu, Su Zhao, et al. "A new coronavirus associated with human respiratory disease in China." en. In: *Nature* 579.7798 (Mar. 2020), pp. 265–269. ISSN: 0028-0836, 1476-4687. DOI: [10.1038/s41586-020-2008-3](https://doi.org/10.1038/s41586-020-2008-3). URL: <http://dx.doi.org/10.1038/s41586-020-2008-3>.
- [523] Chris Wymant, Matthew Hall, Oliver Ratmann, David Bonsall, Tanya Golubchik, Mariateresa de Cesare, Astrid Gall, Marion Cornelissen, Christophe Fraser, and STOP-HCV Consortium, The Maela Pneumococcal Collaboration, and The BEE-HIVE Collaboration. "PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity." en. In: *Mol. Biol. Evol.* (Nov. 2017). ISSN: 0737-4038, 1537-1719. DOI: [10.1093/molbev/msx304](https://doi.org/10.1093/molbev/msx304). URL: <http://dx.doi.org/10.1093/molbev/msx304>.
- [524] Kelly L Wyres, Andries van Tonder, Lotte M Lambertsen, Regine Hakenbeck, Julian Parkhill, Stephen D Bentley, and Angela B Brueggemann. "Evidence of antimicrobial resistance-conferring genetic elements among pneumococci isolated prior to 1974." en. In: *BMC Genomics* 14 (July 2013), p. 500. ISSN: 1471-2164. DOI: [10.1186/1471-2164-14-500](https://doi.org/10.1186/1471-2164-14-500). URL: <http://dx.doi.org/10.1186/1471-2164-14-500>.
- [525] Juan Yang, Xinhua Chen, et al. "Disease burden and clinical severity of the first pandemic wave of COVID-19 in Wuhan, China." In: *Nat. Commun.* 11.1 (Oct. 2020), p. 5411. ISSN: 2041-1723. DOI: [10.1038/s41467-020-19238-2](https://doi.org/10.1038/s41467-020-19238-2). URL: <https://doi.org/10.1038/s41467-020-19238-2>.

- [526] Z Yang. "Among-site rate variation and its impact on phylogenetic analyses." en. In: *Trends Ecol. Evol.* 11.9 (Sept. 1996), pp. 367–372. ISSN: 0169-5347. DOI: [10.1016/0169-5347\(96\)10041-0](https://doi.org/10.1016/0169-5347(96)10041-0). URL: [http://dx.doi.org/10.1016/0169-5347\(96\)10041-0](http://dx.doi.org/10.1016/0169-5347(96)10041-0).
- [527] Zhirong Yang, Jukka Corander, and Erkki Oja. "Low-Rank Doubly Stochastic Matrix Decomposition for Cluster Analysis." In: *J. Mach. Learn. Res.* 17.187 (2016), pp. 1–25. ISSN: 1532-4435. URL: <http://jmlr.org/papers/volume17/15-549/15-549.pdf>.
- [528] R J F Ypma, A M A Bataille, A Stegeman, G Koch, J Wallinga, and W M van Ballegooijen. "Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data." en. In: *Proc. Biol. Sci.* 279.1728 (Feb. 2012), pp. 444–450. ISSN: 0962-8452, 1471-2954. DOI: [10.1098/rspb.2011.0913](https://doi.org/10.1098/rspb.2011.0913). URL: <http://dx.doi.org/10.1098/rspb.2011.0913>.
- [529] M Ammar Zafar, Alexandria J Hammond, Shigeto Hamaguchi, Weisheng Wu, Masamitsu Kono, Lili Zhao, and Jeffrey N Weiser. "Identification of Pneumococcal Factors Affecting Pneumococcal Shedding Shows that the dlt Locus Promotes Inflammation and Transmission." en. In: *MBio* 10.3 (June 2019). ISSN: 2150-7511. DOI: [10.1128/mBio.01032-19](https://doi.org/10.1128/mBio.01032-19). URL: <http://dx.doi.org/10.1128/mBio.01032-19>.
- [530] M Ammar Zafar, Yang Wang, Shigeto Hamaguchi, and Jeffrey N Weiser. "Host-to-Host Transmission of Streptococcus pneumoniae Is Driven by Its Inflammatory Toxin, Pneumolysin." en. In: *Cell Host Microbe* 21.1 (Jan. 2017), pp. 73–83. ISSN: 1931-3128, 1934-6069. DOI: [10.1016/j.chom.2016.12.005](https://doi.org/10.1016/j.chom.2016.12.005). URL: <http://dx.doi.org/10.1016/j.chom.2016.12.005>.
- [531] Seyed Alireza Zamani-Dahaj, Mohamed Okasha, Jakub Kosakowski, and Paul G Higgs. "Estimating the Frequency of Horizontal Gene Transfer Using Phylogenetic Models of Gene Gain and Loss." en. In: *Mol. Biol. Evol.* 33.7 (July 2016), pp. 1843–1857. ISSN: 0737-4038, 1537-1719. DOI: [10.1093/molbev/msw062](https://doi.org/10.1093/molbev/msw062). URL: <http://dx.doi.org/10.1093/molbev/msw062>.
- [532] Kevin van Zandvoort, Francesco Checchi, et al. "Pneumococcal conjugate vaccine use during humanitarian crises." en. In: *Vaccine* 37.45 (Oct. 2019), pp. 6787–6792. ISSN: 0264-410X, 1873-2518. DOI: [10.1016/j.vaccine.2019.09.038](https://doi.org/10.1016/j.vaccine.2019.09.038). URL: <http://dx.doi.org/10.1016/j.vaccine.2019.09.038>.
- [533] Fabio Zanini and Richard A Neher. "FFPopSim: an efficient forward simulation package for the evolution of large populations." en. In: *Bioinformatics* 28.24 (Dec. 2012), pp. 3332–3333. ISSN: 1367-4803, 1367-4811. DOI: [10.1093/bioinformatics/](https://doi.org/10.1093/bioinformatics/bts611)

- bts633. URL: <http://dx.doi.org/10.1093/bioinformatics/bts633>.
- [534] Ea Zankari, Henrik Hasman, Salvatore Cosentino, Martin Vester-gaard, Simon Rasmussen, Ole Lund, Frank M Aarestrup, and Mette Voldby Larsen. "Identification of acquired antimicrobial resistance genes." en. In: *J. Antimicrob. Chemother.* 67.11 (Nov. 2012), pp. 2640–2644. ISSN: 0305-7453, 1460-2091. DOI: [10.1093/jac/dks261](https://doi.org/10.1093/jac/dks261). URL: <http://dx.doi.org/10.1093/jac/dks261>.
- [535] Emanuel Zaufal. "Mikroorganismen im Secrete der Otitis media acuta." In: *Prager Medicinische Wochenschrift* 12 (1887), pp. 225–227.
- [536] J R Zhang, K E Mostov, M E Lamm, M Nanno, S Shimida, M Ohwaki, and E Tuomanen. "The polymeric immunoglobulin receptor translocates pneumococci across human nasopharyngeal epithelial cells." en. In: *Cell* 102.6 (Sept. 2000), pp. 827–837. ISSN: 0092-8674. DOI: [10.1016/S0092-8674\(00\)00071-4](https://doi.org/10.1016/S0092-8674(00)00071-4). URL: [http://dx.doi.org/10.1016/S0092-8674\(00\)00071-4](http://dx.doi.org/10.1016/S0092-8674(00)00071-4).
- [537] Ji Zhang, Jani Halkilahti, Marja-Liisa Hänninen, and Mirko Rossi. "Refinement of whole-genome multilocus sequence typing analysis by addressing gene paralogy." en. In: *J. Clin. Microbiol.* 53.5 (May 2015), pp. 1765–1767. ISSN: 0095-1137, 1098-660X. DOI: [10.1128/JCM.00051-15](https://doi.org/10.1128/JCM.00051-15). URL: <http://dx.doi.org/10.1128/JCM.00051-15>.
- [538] Juanjuan Zhang, Maria Litvinova, et al. "Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study." en. In: *Lancet Infect. Dis.* (Apr. 2020). ISSN: 1473-3099, 1474-4457. DOI: [10.1016/S1473-3099\(20\)30230-9](https://doi.org/10.1016/S1473-3099(20)30230-9). URL: [http://dx.doi.org/10.1016/S1473-3099\(20\)30230-9](http://dx.doi.org/10.1016/S1473-3099(20)30230-9).
- [539] Yongbing Zhao, Jiayan Wu, Junhui Yang, Shixiang Sun, Jingfa Xiao, and Jun Yu. "PGAP: pan-genomes analysis pipeline." en. In: *Bioinformatics* 28.3 (Feb. 2012), pp. 416–418. ISSN: 1367-4803, 1367-4811. DOI: [10.1093/bioinformatics/btr655](https://doi.org/10.1093/bioinformatics/btr655). URL: <http://dx.doi.org/10.1093/bioinformatics/btr655>.
- [540] Sha Joe Zhu, Jacob Almagro-Garcia, and Gil McVean. "Deconvolution of multiple infections in *Plasmodium falciparum* from high throughput sequencing data." In: *Bioinformatics* (Aug. 2017). ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btx530](https://doi.org/10.1093/bioinformatics/btx530). URL: <https://academic.oup.com/bioinformatics/article/doi/10.1093/bioinformatics/btx530/4091117/Deconvolution-of-multiple-infections-in-Plasmodium>.
- [541] Paul N Zivich, John D Grabenstein, Sylvia I Becker-Dreps, and David J Weber. "Streptococcus pneumoniae outbreaks and implications for transmission and control: a systematic review."

- en. In: *Pneumonia (Nathan)* 10 (Nov. 2018), p. 11. ISSN: 2200-6133. DOI: [10.1186/s41479-018-0055-4](https://doi.org/10.1186/s41479-018-0055-4). URL: <http://dx.doi.org/10.1186/s41479-018-0055-4>.
- [542] E Zuckerkandl and L Pauling. "Molecules as documents of evolutionary history." en. In: *J. Theor. Biol.* 8.2 (Mar. 1965), pp. 357–366. ISSN: 0022-5193. DOI: [10.1016/0022-5193\(65\)90083-4](https://doi.org/10.1016/0022-5193(65)90083-4). URL: [http://dx.doi.org/10.1016/0022-5193\(65\)90083-4](http://dx.doi.org/10.1016/0022-5193(65)90083-4).

COLOPHON

This document was typeset using the typographical look-and-feel classicthesis developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*".

<https://bitbucket.org/amiede/classicthesis/>

Final Version as of January 28, 2021 (classicthesis).