Genomics of Bacterial Pathogens Across Evolutionary Scales

By

Abigail C. Shockey

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Microbiology)

at the

UNIVERSITY OF WISCONSIN-MADISON

2019

Date of final oral examination: 07/01/2019

The dissertation is approved by the following members of the Final Oral Committee:

Caitlin Pepperell, Associate Professor, Medical Microbiology & Immunology, Medicine

Joseph Dillard, Professor, Medical Microbiology & Immunology

Garret Suen, Associate Professor, Bacteriology

Colin Dewey, Professor, Biostatistics & Medical Informatics, Computer Sciences

Laurence Loewe, Assistant Professor, Genetics

Table of Contents

Abstract iii
Acknowledgements v
List of Figures vii
List of Tables ix
Chapter 1: Introduction 1
Genomic plasticity and its role in intragenomic interactions 2
Within-host evolution of bacterial pathogens 6
Migratory histories of bacterial pathogens. 9
Chapter 2: Adaptation of the Gonococcal Genetic Island 25
Abstract 26
Introduction 26
Methods 27
Results 30
Discussion 36
Acknowledgements 41
References 41
Figures 46
Tables 56
Supplementary Figures 57
Supplementary Tables 62
Chapter 3: Effects of Host, Sample, and *in vitro* Culture on Genomic Diversity of Pathogenic Mycobacteria 103
Abstract 104
Introduction 104
Methods 106
Results 111
Discussion 115
Acknowledgements 122
References 122
Figures 128
Tables 135
Supplementary Figures 136

Supplementary Tables 142
Chapter 4: Bayesian and Site Frequency Based Inference of *Mycobacterium tuberculosis'* Migratory Histories 178
Abstract 179
Introduction 179
Methods 181
Results 183
Discussion 184
Acknowledgements 186
References 186
Figures 190
Tables 195
Supplementary Tables 195
Chapter 5: Conclusions and Future Directions 202
The gonococcal genetic island and identifying intragenomic interactions 203
Within-host evolution of pathogenic Mycobacteria 205
Migratory history of *Mycobacterium tuberculosis* 207
References 208
Appendix 1 211
Sample description of the Old World Collection 212
Reference guided assembly of the Old World Collection 212
SNP alignment of the Old World Collection 213
Genomic data and sample description of L4.4 213
SNP alignment of L4.4 214
Bayesian phylogenetic analysis of L4.4.1.1 214
Assessment of temporal signal for tip-based calibration for L4.4.1.1 215
Molecular dating of L4.4.1.1 215
References 216

Abstract

Advances in sequencing technologies and increasingly sophisticated computational tools for analyzing sequence data have facilitated in-depth characterization of genomic variation in bacterial pathogens. In this thesis, I examine the evolution of pathogenic bacteria across three scales, intragenomic, within-host and across the globe, using whole genome sequence data (WGS) from natural populations.

The plasticity of bacterial genomes facilitates intragenomic conflict, particularly when DNA is acquired from lateral gene transfer (LGT), but intragenomic conflict can be resolved through co-adaptation. Using WGS from natural populations of *Neisseria gonorrhoeae*, I characterized the evolutionary history of the gonococcal genetic island (GGI), which encodes a Type IV secretion system that facilitates LGT, and investigated co-adaptation between the GGI and *N. gonorrhoeae's* core genome. I identified patterns of core genome recombination that differentiate GGI+ and GGI- populations. Additionally, I found evidence of interactions and co-adaptation between the GGI and *N. gonorrhoeae's* core genome.

Most genomic studies of bacterial pathogens are based on single isolates cultured *in vitro*. However, in their natural hosts, bacteria comprise complex populations that diversify over the course of infection. In order to elucidate differences between evolutionary pressures encountered within the host and *in vitro*, I compared patterns of genetic diversity of pathogenic mycobacteria in sputum to bacteria grown *in vitro*. I identified effects of sampling, patient, and sample type on bacterial genetic diversity. Genetic diversity was more variable and higher in sputum than in culture samples, which suggests *in vitro* manipulation reshapes bacterial populations.

As bacterial pathogens diversify and spread to new hosts, they acquire mutations that can be used to reconstruct their migratory history. Using distinct methods of demographic inference, I reconstructed the migratory history of two *M. tb* lineages. Inferring the migratory

history of *M. tb* Lineage 1, I found SFS based methods performed poorly. Using Bayesian methods, I reconstructed *M. tb* L4.4.1.1's migratory history in the South Pacific and found it was likely dispersed from Europe via colonial migrations.

This work provides new insights into the evolution of bacterial pathogens and demonstrates the power of WGS and associated analytical methods to characterize the ecology and evolution of infectious disease.

Acknowledgements

First and foremost, I thank my advisor, Caitlin Pepperell, for your mentorship. Thank you for giving me the opportunity to work in your lab and numerous opportunities to learn and grow as a scientist. Your guidance has been invaluable. Thank you to the past and present members of the Pepperell Lab. Your insight and contributions are deeply appreciated. Thank you Tatum Mortimer and Mary O'Neill for the foundation you built for present and future graduate students. Lindsey Bohr, I'm so glad you joined the lab. Working alongside you has been wonderful. Tracy Smith, I am grateful for your encouragement, and your humor has always been delightful. Erin Zwick, thank you for your wit and expertise. Madison Youngblom, I'm so glad you joined the lab and to have met you. I would also like to thank Donny Xiong, Jesse Dabney, Linda Xiong, and the many system administrators who kept our servers running strong.

Thank you to my committee members, Joe Dillard, Garret Suen, Colin Dewey and Laurence Loewe. I have also been able to work with wonderful collaborators, Joe Dillard, Melanie Callaghan, Drew Kitchen and Claire Mulholland. Thank you to Laura Knoll, who pointed me to the Pepperell Lab and gave me the opportunity to improve as an educator. I would like to acknowledge my funding sources, the National Institutes of Health and the University of Wisconsin School of Medicine and Public Health from the Wisconsin Partnership Program. I would also like to thank the Microbiology Doctoral Training Program and the Department of Medical Microbiology and Immunology for their financial and administrative support.

The thanks I wish to give my family and friends for their love and support throughout graduate school could go on forever. To my mother, Diane Eastman-Shockey, thank you for the years spent watching Public Broadcasting and Nova that sparked my love of science, and thank you for being my biggest fan. To my father, Orville Shockey, thank you for being a pillar of strength and source of wisdom. I love both of you so much. Thank you to my grandparents, Trudy and Richard Eastman, who taught me every change in life was an opportunity for adventure and to march to the beat of my own drum. To all my friends in the Microbiology

Doctoral Training Program, Kelsey Florek, Gloria Larson, Kayleigh Nyffeler, Jenny Bratburd, Melanie Callaghan and Tyler Jacobson, I'm grateful for your friendship and I can't wait to see what all of you accomplish next. Thank you Sam Manaktola for your encouragement, compassion, commiseration and queer sense of humor. You have made every week I've known you bright. Ross Llewallyn, thank you for being thoughtful and always cheering me on. Thank you Kelly Jones and James Luksich for your friendship and support, even from afar. Kati Niekerk, I could not have made it through graduate school without you. You have been a font of love, compassion, joy, inspiration and support, and I am so grateful to have you in my life. And to Boss, my somber but loyal hound, thank you for snoozing by my desk every day I've worked through graduate school. All of you have made Madison such a wonderful home. Thank you!

List of Figures

Figure 2.1. Distribution of recombinant fragments across *N. gonorrhoeae* core genome. 46
Figure 2.2. Proportion of recombinant sites per genome and proportion of alignment affected by recombination in GGI+ and GGI- isolates. 47
Figure 2.3. Proportion of recombinant sites per genome in individual core genomes of GGI+ and GGI- isolates. 48
Figure 2.4. Core genome network of isolates with and without the GGI. 49
Figure 2.5. Intensity of recombination across *N. gonorrhoeae* genomes. 50
Figure 2.6. Co-ancestry matrix of GGI+ and GGI- populations. 51
Figure 2.7. Network of the GGI. 52
Figure 2.8. Maximum-likelihood phylogeny of *N. gonorrhoeae* isolates with presence/absence of the GGI indicated. 53
Figure 2.9. Principal component analysis of diversity within the core genome of the GGI. 54
Figure 2.10. Reconstruction of the ancestral state of the GGI locus. 55
Figure 2.S1. Distribution of gene content within the bounds of the GGI. 57
Figure 2.S2. Distribution of gene content in the pan-genome of *N. gonorrhoeae*. 58
Figure 2.S3. Distribution of recombinant fragment lengths. 59
Figure 2.S4. Reconstruction of the ancestral state of the phage tail protein. 60
Figure 2.S5. Distribution of genome-wide F_{ST} values. 61
Figure 2.S6. Distribution of mobile genetic elements in the pan-genome. 62
Figure 3.1. Smear score.. 128
Figure 3.2. Bacterial diversity across genome windows in sputum and culture samples. 129
Figure 3.3. Differences in bacterial diversity among inter-patient and intra-patient pairs of samples. 130
Figure 3.4. Genome-wide diversity in sputum and culture. 131
Figure 3.5. Differences in gene diversity of sputum and culture samples. 132
Figure 3.6. Linear regression of gene diversity in culture vs sputum. 133
Figure 3.7. Distribution of fold change in nucleotide diversity per gene. 134
Figure 3.S1. Sliding window diversity of Patient 14 in sputum, culture and culture spiked with negative sputum. 136

Figure 3.S2. Patterns of nucleotide diversity in composite patient sample. 137
Figure 3.S3. Patterns of Watterson's theta in sputum and culture. 138
Figure 3.S4. Patterns of Tajima's D in sputum and culture. 139
Figure 3.S5. Patterns of fold change in nucleotide diversity across the genome. 140
Figure 3.S6. Histogram of F_{ST} outliers in *M. tuberculosis* samples. 141
Figure 4.1. Observed and inferred two-dimensional sSFS of India and RoW. 190
Figure 4.2. Poisson residuals for best-fit migration models..... 191
Figure 4.3. Demography of *M. tb* L4.4.1.1. 192
Figure 4.4. MCC phylogeny of *M. tb* L4.4.1.1. 193
Figure 4.5. Migration matrix of *M. tb* L4.4.1.1. 194

List of Tables

Table 2.1. Core genome analysis and global diversity estimates. 56
Table 2.2. BaTS association statistics. 56
Table 2.3. Accessory gene content associated with the GGI. 56
Table 2.S1. List of isolate accessions and GGI presence/absence. 62
Table 2.S2. GGI genes in GGI+ isolates. 66
Table 2.S3. Genes with extreme H_i (≥ 99th percentile) in GGI+ and GGI- Isolates. 69
Table 2.S4. Genes with extreme fold changes in GGI+ and GGI- Isolates. 69
Table 2.S5. Homoplastic F_{ST} outliers. 70
Table 2.S6. Homoplastic F_{ST} outliers gained with the GGI on the core genome phylogeny. 88
Table 2.S7. Mobile genetic element genes identified in N. gonorrhoeae pan-genome. 98
Table 2.S8. Restriction modification genes identified in pan-genome of *N. gonorrhoeae*. 101
Table 2.S9. Toxin-antitoxin genes identified in pan-genome of *N. gonorrhoeae*. 102
Table 3.1. *M. tb* genes with extreme patterns of variation across multiple patients and multiple measures. 135
Table 3.S1. Regions removed. 142
Table 3.S2. Variation at lineage defining loci. 164
Table 3.S3.1 Coverage statistics. 165
Table 3.S3.2 Lineage typing of *M. tb* samples. 167
Table 3.S4. Smear Score ANOVA of *M. tb* samples. 168
Table 3.S5.1. *M. tb* windows of diversity (π) overlap. 168
Table 3.S5.2. *Mb* windows of diversity (π) overlap. 168
Table 3.S6.1 Outlier genes: lm and fold change. 170
Table 3.S6.2 Outlier genes: Absolute change and F_{ST}. 173
Table 4.1. Summary of ∂a∂i migration analyses for *M. tb* L1 195
Table 4.S1. *M. tb* L1 accessions and metadata. 195
Table 4.S2. *M. tb* L4.4.1.1 accessions and metadata. 199
Table 1. Appendix 1 Old World Collection accessions and metadata 221
Table 2. Appendix 1 *M. tb* L4.4 accessions and metadata 256

Chapter 1: Introduction

Advances in sequencing technologies, as well as decreasing costs of sequencing, have increased the accessibility and affordability of whole genome sequencing as a tool for generating bacterial sequence data (1). Hundreds of thousands of whole-genome sequences (WGS) are now publicly available for a wide range of bacterial pathogens. The National Center for Biotechnology Information's Pathogen Detection Isolates Browser alone consists of ~380,000 genomes from over 25 pathogenic bacterial species (2). Additionally, method development and analysis tools for WGS are advancing alongside sequencing (3).

This wealth of genomic data and analytical tools allows for in-depth characterization of genomic variation in natural populations of bacterial pathogens (4). Additionally, WGS is being used increasingly in clinical settings to identify outbreaks (5–8), detect antibiotic resistance (9–17), perform pathogen surveillance (7, 13, 18, 19), and enhance diagnostics (20–24).

My research has been focused on understanding the evolution of bacterial pathogens using WGS from natural populations. Each of my projects has examined bacterial evolution at a different scale: intragenomic, within hosts, and across the globe. In this chapter, I review bacterial genomic plasticity and its role in intragenomic interactions. Additionally, I review the selective pressures that shape the genomes of bacterial pathogens during infection and insights into within-host evolution gained from WGS. Finally, I review the role of WGS in analyzing pathogen migratory histories and notable examples of pathogen dispersal facilitated by human migrations.

Genomic plasticity and its role in intragenomic interactions

Bacterial genomes exhibit a wide range of plasticity (25, 26), which facilitates adaptation to selective pressures bacteria encounter in their environment. Broadly, sources of bacterial genomic plasticity include point mutations, genomic rearrangements and lateral gene transfer.

Point mutations provide the raw material for bacterial evolution, and the accumulation of point mutations varies across bacterial species due to differences in per site mutation rates.

Additionally, some bacteria are hypermutators: they have higher rates of point mutation due to the disruption of genes encoding mismatch repair (MMR) systems or components of other systems that ensure replication fidelity and prevent DNA damage (25, 27). For example, mutation accumulation analysis has shown MMR-deficiency in *Escherichia coli* can result in a 138X increase in the number of base-pair substitutions (28). Hypermutators have been observed in several species of bacterial pathogens (29–34) and appear commonly associated with infectious disease (35, 36); despite the potential fitness effects of rapidly accumulating (potentially deleterious) mutations, hypermutation may accelerate the evolution of bacterial pathogens by increasing the variation of surface antigens recognized by the host immune system and facilitating the acquisition of pathoadaptive and/or antibiotic resistance mutations (29, 30, 37–40).

Genome rearrangements, which affect long segments of the chromosome and can occur during DNA recombination, replication and repair, produce structural variants (SVs) in bacterial chromosomes (41). SVs can be divided into five major classes: deletions, insertions, duplications, inversions and translocations. Genome rearrangements can alter gene function and expression, but the consequences vary depending on their length and the region of the chromosome they affect (42). For example, large inversions may attenuate virulence and alter growth kinetics as in *Bacillus anthracis* (43) or affect cell viability as in *E. coli* (44). Additionally, the frequency and significance of genome rearrangements differs between and within bacterial species. Genome rearrangements played a role in the divergence of *Burkholderia mallei* from *Burkholderia pseudomallei,* (45, 46), and pathogenic *Yersinia*, *Bordatella pertussis* and *Gardnerella vaginalis* differ widely in their genomic synteny (47–49).

Arguably the largest contributor to bacterial genomic plasticity is lateral gene transfer (LGT), the process by which bacteria exchange DNA. LGT facilitates bacterial adaptation by shuffling beneficial and deleterious alleles and is the dominant source of novel gene content in

bacterial populations. LGT occurs within and between bacterial species (28, 50), and many species differ in their propensity to engage in LGT (51, 52).

The three classical mechanisms of LGT are conjugation, transduction and transformation. Conjugation involves the transfer of plasmids, self-replicating extrachromosomal molecules of DNA, from donor to recipient cell via a Type IV secretion system [T4SS; (53)]. Transduction is mediated by the bacteria specific viruses known as bacteriophage. Bacterial DNA can be mistakenly packaged within the phage particle during phage replication and assembly; this DNA can be integrated into the chromosome of a new host upon phage infection (54). Transformation involves the uptake of environmental DNA by a cell followed by recombination into the cell's chromosome. Homologous recombination of the environmentally acquired DNA may replace the DNA encoded at that locus, while illegitimate recombination introduces new gene content into the recipient chromosome (55).

Mobile genetic elements (MGEs) are vectors of LGT; they encode machinery to facilitate their transfer within and between genomes. MGEs comprise a wide set of elements, including the previously described plasmids and phage. Other MGEs include insertion sequences, transposons, genomic islands and many other elements that are genetic mosaics of plasmid and phage (56). Many MGEs carry beneficial traits such as virulence factors and antibiotic resistance (57–61), but they may also carry addiction modules than ensure their maintenance at the expense of their host (62, 63). Additionally, the integration of MGEs can contribute to genomic plasticity by inducing rearrangements in their host chromosomes (64–68).

The acquisition of novel gene content via LGT creates a wide variability in the gene content present within a bacterial species (69, 70). The genes shared by all members of a bacterial species are known as the core genome, while genes shared by fewer than all members are known as the accessory genome. The pan-genome refers to the collection of all genes present within a bacterial species.

While genomic plasticity allows bacteria to adapt to different selective pressures, it can result in genomes that are mosaics of interacting loci with aligned and/or competing interests. This is frequently the result of LGT events. When novel gene content enters a new genetic background, it often functions inefficiently because it has evolved under different selective pressures. This can impose a fitness cost for the cell and creates intragenomic conflict (71). Fitness costs imposed by LGT include chromosome and/or regulatory network disruption [described above under genomic rearrangements; (72, 73)], cytotoxicity (74), increased metabolic burden (75) and sequestration of limited resources (76, 77). Additionally, as previously described, many MGEs are “selfish” elements that ensure their own replication and transmission by imposing the direct fitness cost of cell death (62, 63, 78, 79).

Fitness costs imposed by LGT may be mitigated by compensatory mutations in the chromosome, laterally acquired elements or both (80–84). For example, experimental evolution of *E. coli* containing a tetracycline resistance plasmid has shown that, over the course of antibiotic treatment, mutations facilitating co-adaptation occur on both the chromosome and the plasmid. Mutations that conferred antibiotic resistance through reduced membrane permeability and increased chromosomally encoded drug efflux emerged on the host chromosome, while mutations that impaired plasmid encoded drug efflux arose on the tetracycline resistance plasmid. Together, these co-adaptations allowed for the maintenance of the resistance plasmid while reducing the associated fitness cost (85). Laterally acquired genes and/or elements can also interact with one another, and their fitness costs can be reduced by co-adaptation. Experimental evolution and RNA-seq experiments have shown costly changes in gene expression induced by the acquisition of a small plasmid in *Pseudomonas aeruginosa* can be mitigated by compensatory mutations that inactivate two genes encoded on recently acquired, chromosomally integrated MGEs (86, 87).

In chapter 2 of this thesis, I use WGS from natural populations of *Neisseria gonorrhoeae* to characterize the evolutionary history of the gonococcal genetic island (GGI), a genetic island

encoding a T4SS that secretes DNA, and investigate co-adaptation between the GGI and *N. gonorrhoeae's* core genome.

Within-host evolution of bacterial pathogens

The plasticity of bacterial genomes allows bacteria to respond to selective pressures they encounter and adapt to changes in their environment. The host environment is extremely hostile and exerts numerous selective pressures on bacterial pathogens, which necessitates adaptation during the course of infection. Selective pressures bacteria face in the host environment include the host immune system, competition with the native microbiota and medical interventions like antibiotic treatment (88).

Innate immunity includes defenses such as physical barriers (i.e. mucosal epithelium), the complement system and phagocytic cells, while adaptive immunity includes antibody production via B cells and subsequent cytotoxicity via T cells. Bacterial pathogens can evolve many different mechanisms to evade and survive their hosts' immune system, both innate and adaptive (89). The mucosal layer can be broken down by mucinases such as the Pic enzyme of *Shigella spp.* (90), which is encoded on a pathogenicity island, or the sialidases of *Gardnerella vaginalis* (91). S*taphylococcus aureus* possesses a complement inhibiting protein, the staphylococcal complement inhibitor (SCIN), which is also encoded on a pathogenicity island (92). To evade intracellular killing by phagocytes, *N. gonorrhoeae* has evolved the capacity to delay phagosomal maturation by inhibiting granule fusion with phagosomes (93). Finally, pathogenic *Neisseria spp.* can avoid recognition by human antibodies through recombination of loci associated with their pilin protein, a process known as antigenic variation (94).

Invading pathogens must also compete with the host's resident microbiota for space and nutrients (95). Bacteria can eliminate their competition by producing bacteriocins, antibacterial peptides that are primarily encoded on plasmids and produced by a wide range of bacteria (95,

96). For example, the enteric pathogen *Salmonella enterica* upregulates bacteriocin production in the presence of inflammation, giving it a fitness advantage over commensal *E. coli* (97). Additionally, many bacterial pathogens utilize Type VI secretion systems, which inject effectors directly into other bacterial cells, to kill their hosts' microbiota (98–100). Alternatively, bacterial pathogens can sequester nutrients from competing resident microbes. Iron is limited in the host environment, and bacterial pathogens have evolved several mechanisms to scavenge iron (101). One notable adaptation is the iron-chelating siderophores, some of which are species-specific and require specific re-uptake machinery to bind and import the iron-siderophore complex (102).

Antibiotics are the primary host medical intervention used against bacterial pathogens. As antibiotic resistance has risen to dangerous levels across the globe (103), antibiotic resistance has become one of the most widely studied examples of bacterial adaptation to a within-host selective pressure. Bacterial pathogens have evolved many mechanisms of resistance to avoid killing by antibiotics, which can be broken down into four categories: modification/destruction of the antimicrobial, decreased antimicrobial penetrance and drug efflux, changes to the target site, and resistance due to global cell adaptation (104). Destruction of the amide bond in β-lactam antibiotics is a classic example of antibiotic resistance facilitated by modification/destruction of the antibiotic. β-lactamases are encoded on both chromosomes and MGEs, and over 2000 β-lactamases from > 500 bacterial species have been described (105), highlighting the numerous ways bacteria have evolved to combat this class of antibiotics. Additionally, resistance to β-lactam antibiotics epitomizes the rapid adaptation of bacterial populations to host medical interventions: Widespread resistance to the β-lactam penicillin, and eventually its synthetic successors, emerged in hospital settings within a decade after their introduction as therapeutics (106–108).

Until the advent of WGS, little was known about within-host adaptation of bacterial pathogens. Prior to the routine use of WGS in the analysis of bacterial evolution, low resolution

techniques based on small segments of the chromosome, such as pulse field gel electrophoresis (PFGE), variable-number tandem repeats (VNTR) and multi-locus sequence typing (MLST), were used to characterize within-host populations (109, 110). These techniques are sensitive enough to detect the presence of different bacterial lineages, particularly in the case of mixed infections, but cannot delineate the evolution of individual bacterial lineages over the course of infection.

WGS from multiple bacterial isolates sampled simultaneously and/or longitudinally from the same host have provided researchers a higher resolution picture of within-host adaptation. For example, measurements of within-host point mutation rates have revealed rates as high as ~30/year/genome for *Helicobacter pylori* (111, 112) and as low ~0.5/year/genome in *Mycobacterium tuberculosis* and *Mycobacterium abscessus* (6, 113, 114). Strikingly, a recent study of *Klebsiella pneumoniae* isolated from a single patient across six body sites revealed no mutation accumulation during a recurrent urinary tract infection (115).

WGS has confirmed the presence of LGT during infection and highlighted its importance as a driver of within-host adaptation (116–122). A recent longitudinal analysis of *Burkholderia multivorans* isolates from a single cystic fibrosis patient identified recombination events, and variants associated with pathoadaptation and β-lactam resistance were over-represented in those recombinogenic regions (123). Additionally, analysis of within-host adaptation has also identified convergent evolution in mutations within intergenic regions of *Pseudomonas aeruginosa's* genome, suggesting a role for intergenic regions in host adaptation (124).

Within-host studies can also identify adaptations that allow for niche expansion during infection. Paired throat and blood samples from patients with invasive meningococcal disease revealed few (≤ 3) mutations facilitate *Neisseria meningitidis*' transition from asymptomatic carriage to invasive disease (119), but paired stool and blood samples from *Enterococcus faecium* patients failed to reveal within-host mutations associated with bloodstream infection,

which suggests the route to invasive disease may be more complex in this species (118). Conversely, a longitudinal analysis of *Burkholderia pseudomallei* from a single patient showed substantial genomic changes occur during asymptomatic carriage, including genome reduction, loss of pathogenicity genes and attenuation of virulence (125).

As described above, antibiotic resistance is one of the most widely studied bacterial adaptations, and antibiotic resistance has been the focus of many within-host evolution studies (126–130). Many within-host analyses of *M. tuberculosis* in particular have examined antibiotic resistance (131). For example, a longitudinal study of *M. tuberculosis* isolated from a single patient over 3 years showed the emergence of extensive drug resistance from a drug-sensitive ancestor (132–134). Resistance mutations emerged for each of the seven drugs the patient was exposed to and a single multidrug resistant lineage emerged by the end of the sampling period after displacing its competitor lineages.

While many studies of within-host evolution of *M. tuberculosis* have focused on antibiotic resistance, a recent study that characterized genome-wide patterns of diversity in five patients' *M. tuberculosis* populations over the course of infection found *M. tuberculosis* patients harbor diverse populations of *M. tuberculosis* during infection. Notable changes in diversity occurred in genes involved in the regulation, synthesis, and transportation of lipids and glycolipids of the *M. tuberculosis'* cell envelope (135).

In chapter 3 of this thesis, I compare patterns of genetic diversity of pathogenic mycobacteria in sputum to bacteria grown *in vitro* in order to elucidate differences between evolutionary pressures encountered within the host and those imposed by *ex vivo* manipulation of bacterial populations.

Migratory histories of bacterial pathogens

Bacterial pathogens that adapt, survive and replicate within their hosts can subsequently transmit to a new host. As bacterial pathogens diversify during infection and

spread to new hosts, they can acquire informative mutations within their genomes. This process of diversification and dispersal can leave a quantifiable footprint within their genomes. As such, phylogenetic analyses of WGS can be used to reconstruct these transmission events (109, 136).

Incorporating geographic and temporal information into a phylogenetic framework can be used map bacterial populations back to their origin (4). These analyses are referred phylogenomic analyses (137). Inferring the origin of a bacterial pathogen can refer to a historically recent outbreak, but also includes the geographic origin of a bacterial pathogen and its global patterns of dispersal over time. When combined, the evolutionary history of a bacterial pathogen and its patterns of dispersal can elucidate the key factors and dynamics underlying epidemics, which can be used to inform control strategies.

Many phylogenomic methods were first developed for the analysis of viral phylogeography (138). Due to viruses' rapid rate of evolution, the epidemiological processes that shape their diversity operate on a similar timescale as mutation fixation within their populations, making them an ideal organism to study phylogenomics (139). However, with the exponential increase of sequenced bacterial genomes, phylogenomic analyses are increasingly used to study bacterial pathogens. In particular, the phylogeographic history of *H. pylori* has been extensively characterized (140). Human acquisition of *H. pylori* is predicted to have occurred at least 100,000 years ago (141, 142), and ancient *H. pylori* populations appear to have accompanied humans out of Africa twice. Reconstructions of *Mycobacterium leprae's* migration patterns suggest *M. leprae* was transmitted to European populations from Central Asian populations, likely via trade routes (143). Additionally, a recent phylogeographic analysis suggests *M. leprae* originated in East Asia (144). *N. gonorrhoeae* may have originated in Europe or Africa as late as the 16th century; older populations of *N. gonorrhoeae* were initially geographically separated and more recently disseminated globally (145). Analyses of *Vibrio cholerae* suggest the seventh cholera pandemic spread from the Bay of Bengal in at least three

independent, overlapping waves of global transmission, and the origin of this pandemic was estimated to be ~60 years ago (146).

Similar to *H. pylori*, out of Africa migrations in association with humans has been proposed as a hypothesis for the dispersal of ancient *M. tuberculosis* populations. Under this hypothesis the most recent common ancestor (MRCA) of *M. tuberculosis* emerged in Africa ~73,000 years ago and was dispersed via subsequent waves of human migration (147). However, recently sequenced ancient DNA from members of the *M. tb* complex (MTBC) infer a more recent time to the MRCA for the MTBC: < 6,000 years ago (148, 149). Recent analyses of extant *M. tuberculosis* populations in the Americas suggest that dispersal of *M. tuberculosis* lineage 4 has been influenced by European colonialism and recent immigration (150, 151); however, the role of other historical phenomena in driving the global dispersal of *M. tuberculosis* is not well understood.

In chapter 4 of this thesis, I reconstruct the migratory history of two lineages of *M. tuberculosis*: *M. tuberculosis* lineage 1 (L1) and *M. tuberculosis* sub-lineage 4.4.1.1. For each lineage, I use a different method of migration inference, highlight the utility of each method and where applicable discuss phenomena that may have contributed to the present global distribution of these lineages.

References

1. NHGRI. DNA Sequencing Costs. genome.gov.

2. NCBI. Pathogen Detection. https://www.ncbi.nlm.nih.gov/pathogens/

3. Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW. 2015. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics 15:141–161.

4. Klemm E, Dougan G. 2016. Advances in Understanding Bacterial Pathogenesis Gained from Whole-Genome Sequencing and Phylogenetics. Cell Host Microbe 19:599–610.

5. Quainoo S, Coolen JPM, van Hijum SAFT, Huynen MA, Melchers WJG, van Schaik W, Wertheim HFL. 2017. Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. Clin Microbiol Rev 30:1015–1063.

6. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW, Parkhill J, Harris D, Walker AS, Bowden R, Monk P, Smith EG, Peto TEA. 2013. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis 13:137–146.

7. Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W, McLaughlin SF, Henkhaus JK, Leopold B, Bielaszewska M, Prager R, Brzoska PM, Moore RL, Guenther S, Rothberg JM, Karch H. 2011. Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. PloS One 6:e22751–e22751.

8. Whaley MJ, Joseph SJ, Retchless AC, Kretz CB, Blain A, Hu F, Chang H-Y, Mbaeyi SA, MacNeil JR, Read TD, Wang X. 2018. Whole genome sequencing for investigations of meningococcal outbreaks in the United States: a retrospective analysis. Sci Rep 8:15803–15803.

9. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 67:2640–2644.

10. Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, Keane JA, Harris SR. 2017. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. Microb Genomics 3:e000131–e000131.

11. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, Lago BA, Dave BM, Pereira S, Sharma AN, Doshi S, Courtot M, Lo R, Williams LE, Frye JG, Elsayegh T, Sardar D, Westman EL, Pawlowski AC, Johnson TA, Brinkman FSL, Wright GD, McArthur AG. 2017.

CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res 45:D566–D573.

12. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, Tyson GH, Zhao S, Hsu C-H, McDermott PF, Tadesse DA, Morales C, Simmons M, Tillman G, Wasilenko J, Folster JP, Klimke W. 2019. Using the NCBI AMRFinder Tool to Determine Antimicrobial Resistance Genotype-Phenotype Correlations Within a Collection of NARMS Isolates. bioRxiv 550707.

13. Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, Tomita T, Zobel J, Holt KE. 2014. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Med 6:90–90.

14. Rowe W, Baker KS, Verner-Jeffreys D, Baker-Austin C, Ryan JJ, Maskell D, Pearce G. 2015. Search Engine for Antimicrobial Resistance: A Cloud Compatible Pipeline and Web Interface for Rapidly Detecting Antimicrobial Resistance Genes Directly from Sequence Data. PloS One 10:e0133492–e0133492.

15. de Man TJB, Limbago BM. 2016. SSTAR, a Stand-Alone Easy-To-Use Antimicrobial Resistance Gene Predictor. mSphere 1:e00050-15.

16. Clausen PTLC, Zankari E, Aarestrup FM, Lund O. 2016. Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. J Antimicrob Chemother 71:2484–2488.

17. Lemon JK, Khil PP, Frank KM, Dekker JP. 2017. Rapid Nanopore Sequencing of Plasmids and Resistance Gene Detection in Clinical Isolates. J Clin Microbiol 55:3530.

18. Votintseva AA, Bradley P, Pankhurst L, del Ojo Elias C, Loose M, Nilgiriwala K, Chatterjee A, Smith EG, Sanderson N, Walker TM, Morgan MR, Wyllie DH, Walker AS, Peto TEA, Crook DW, Iqbal Z. 2017. Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. J Clin Microbiol 55:1285.

19. Kwong JC, Mercoulia K, Tomita T, Easton M, Li HY, Bulach DM, Stinear TP, Seemann T, Howden BP. 2016. Prospective Whole-Genome Sequencing Enhances National Surveillance of Listeria monocytogenes. J Clin Microbiol 54:333–342.

20. Hasman H, Saputra D, Sicheritz-Ponten T, Lund O, Svendsen CA, Frimodt-Møller N, Aarestrup FM. 2014. Rapid Whole-Genome Sequencing for Detection and Characterization of Microorganisms Directly from Clinical Samples. J Clin Microbiol 52:139.

21. Kuroda M, Sekizuka T, Shinya F, Takeuchi F, Kanno T, Sata T, Asano S. 2012. Detection of a possible bioterrorism agent, Francisella sp., in a clinical specimen by use of next-generation direct DNA sequencing. J Clin Microbiol 50:1810–1812.

22. Thomsen MCF, Ahrenfeldt J, Cisneros JLB, Jurtz V, Larsen MV, Hasman H, Aarestrup FM, Lund O. 2016. A Bacterial Analysis Platform: An Integrated System for Analysing Bacterial Whole Genome Sequencing Data for Clinical Diagnostics and Surveillance. PLOS ONE 11:e0157718.

23. Reuter S, Ellington MJ, Cartwright EJP, Köser CU, Török ME, Gouliouris T, Harris SR, Brown NM, Holden MTG, Quail M, Parkhill J, Smith GP, Bentley SD, Peacock SJ. 2013. Rapid bacterial whole-genome sequencing to enhance diagnostic and public health microbiology. JAMA Intern Med 173:1397–1404.

24. Yu Y, Hu F, Zhu C, Chen E, Lu L, Gao Y. 2018. Use of Next Generation Sequencing and Synergy Susceptibility Testing in Diagnosis and Treatment of Carbapenem-Resistant Klebsiella pneumoniae Blood Stream Infection. Case Rep Infect Dis 2018:3295605–3295605.

25. Williams AB. 2016. Chapter 5 - Genome Instability in Bacteria: Causes and Consequences, p. 69–85. *In* Kovalchuk, I, Kovalchuk, O (eds.), Genome Stability. Academic Press, Boston.

26. Patel S. 2016. Drivers of bacterial genomes plasticity and roles they play in pathogen virulence, persistence and drug resistance. Infect Genet Evol 45:151–164.

27. Denamur E, Matic I. 2006. Evolution of mutation rates in bacteria. Mol Microbiol 60:820–827.

28. Lee H, Popodi E, Tang H, Foster PL. 2012. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. Proc Natl Acad Sci U S A 109:E2774–E2783.

29. Waine DJ, Honeybourne D, Smith EG, Whitehouse JL, Dowson CG. 2008. Association between Hypermutator Phenotype, Clinical Variables, Mucoid Phenotype, and Antimicrobial Resistance in Pseudomonas aeruginosa. J Clin Microbiol 46:3491.

30. Bucci C, Lavitola A, Salvatore P, Del Giudice L, Massardo DR, Bruni CB, Alifano P. 1999. Hypermutation in Pathogenic Bacteria: Frequent Phase Variation in Meningococci Is a Phenotypic Trait of a Specialized Mutator Biotype. Mol Cell 3:435–445.

31. Björkholm B, Sjölund M, Falk PG, Berg OG, Engstrand L, Andersson DI. 2001. Mutation frequency and biological cost of antibiotic resistance in Helicobacter pylori. Proc Natl Acad Sci 98:14607.

32. del Campo R, Morosini M-I, de la Pedrosa EG-G, Fenoll A, Muñoz-Almagro C, Máiz L, Baquero F, Cantón R. 2005. Population Structure, Antimicrobial Resistance, and Mutation Frequencies of Streptococcus pneumoniae Isolates from Cystic Fibrosis Patients. J Clin Microbiol 43:2207.

33. Richardson AR, Yu Z, Popovic T, Stojiljkovic I. 2002. Mutator clones of Neisseria meningitidis in epidemic serogroup A disease. Proc Natl Acad Sci 99:6103.
34. LeClerc JE, Li B, Payne WL, Cebula TA. 1996. High Mutation Frequencies Among Escherichia coli and Salmonella Pathogens. Science 274:1208.
35. Denamur E, Bonacorsi S, Giraud A, Duriez P, Hilali F, Amorin C, Bingen E, Andremont A, Picard B, Taddei F, Matic I. 2002. High frequency of mutator strains among human uropathogenic Escherichia coli isolates. J Bacteriol 184:605–609.
36. Ciofu O, Riis B, Pressler T, Poulsen HE, Høiby N. 2005. Occurrence of hypermutable Pseudomonas aeruginosa in cystic fibrosis patients is associated with the oxidative stress caused by chronic lung inflammation. Antimicrob Agents Chemother 49:2276–2282.
37. Smith EE, Buckley DG, Wu Z, Saenphimmachak C, Hoffman LR, D'Argenio DA, Miller SI, Ramsey BW, Speert DP, Moskowitz SM, Burns JL, Kaul R, Olson MV. 2006. Genetic adaptation by Pseudomonas aeruginosa to the airways of cystic fibrosis patients. Proc Natl Acad Sci U S A 103:8487–8492.
38. Canfield GS, Schwingel JM, Foley MH, Vore KL, Boonanantanasarn K, Gill AL, Sutton MD, Gill SR. 2013. Evolution in Fast Forward: a Potential Role for Mutators in Accelerating Staphylococcus aureus Pathoadaptation. J Bacteriol 195:615.
39. Hammerstrom TG, Beabout K, Clements TP, Saxer G, Shamoo Y. 2015. Acinetobacter baumannii Repeatedly Evolves a Hypermutator Phenotype in Response to Tigecycline That Effectively Surveys Evolutionary Trajectories to Resistance. PLOS ONE 10:e0140489.
40. Mehta HH, Prater AG, Beabout K, Elworth RAL, Karavis M, Gibbons HS, Shamoo Y. 2019. The essential role of hypermutation in rapid adaptation to antibiotic stress. Antimicrob Agents Chemother AAC.00744-19.
41. Periwal V, Scaria V. 2014. Insights into structural variations and genome rearrangements in prokaryotic genomes. Bioinformatics 31:1–9.
42. Morrow JD, Cooper VS. 2012. Evolutionary effects of translocations in bacterial genomes. Genome Biol Evol 4:1256–1262.
43. Okinaka RT, Price EP, Wolken SR, Gruendike JM, Chung WK, Pearson T, Xie G, Munk C, Hill KK, Challacombe J, Ivins BE, Schupp JM, Beckstrom-Sternberg SM, Friedlander A, Keim P. 2011. An attenuated strain of Bacillus anthracis (CDC 684) has a large chromosomal inversion and altered growth kinetics. BMC Genomics 12:477–477.
44. Esnault E, Valens M, Espéli O, Boccard F. 2007. Chromosome structuring limits genome plasticity in Escherichia coli. PLoS Genet 3:e226–e226.

45. Losada L, Ronning CM, DeShazer D, Woods D, Fedorova N, Kim HS, Shabalina SA, Pearson TR, Brinkac L, Tan P, Nandi T, Crabtree J, Badger J, Beckstrom-Sternberg S, Saqib M, Schutzer SE, Keim P, Nierman WC. 2010. Continuing evolution of Burkholderia mallei through genome reduction and large-scale rearrangements. Genome Biol Evol 2:102–116.
46. Bochkareva OO, Moroz EV, Davydov II, Gelfand MS. 2018. Genome rearrangements and selection in multi-chromosome bacteria Burkholderia spp. BMC Genomics 19:965.
47. Darling AE, Miklós I, Ragan MA. 2008. Dynamics of Genome Rearrangement in Bacterial Populations. PLOS Genet 4:e1000128.
48. Bowden KE, Weigand MR, Peng Y, Cassiday PK, Sammons S, Knipe K, Rowe LA, Loparev V, Sheth M, Weening K, Tondella ML, Williams MM. 2016. Genome Structural Diversity among 31 Bordetella pertussis Isolates from Two Recent U.S. Whooping Cough Statewide Epidemics. mSphere 1:e00036-16.
49. Yeoman CJ, Yildirim S, Thomas SM, Durkin AS, Torralba M, Sutton G, Buhay CJ, Ding Y, Dugan-Rocha SP, Muzny DM, Qin X, Gibbs RA, Leigh SR, Stumpf R, White BA, Highlander SK, Nelson KE, Wilson BA. 2010. Comparative genomics of Gardnerella vaginalis strains reveals substantial differences in metabolic and virulence potential. PloS One 5:e12411–e12411.
50. Lawrence JG, Ochman H. 1998. Molecular archaeology of the Escherichia coli genome. Proc Natl Acad Sci U S A 95:9413–9417.
51. Didelot X, Maiden MCJ. 2010. Impact of recombination on bacterial evolution. Trends Microbiol 18:315–322.
52. Eldholm V, Balloux F. 2016. Antimicrobial Resistance in Mycobacterium tuberculosis: The Odd One Out. Trends Microbiol 24:637–648.
53. Llosa M, Gomis-Rüth FX, Coll M, Cruz F de la. 2002. Bacterial conjugation: a two-step mechanism for DNA transport. Mol Microbiol 45:1–8.
54. ZINDER ND, LEDERBERG J. 1952. Genetic exchange in Salmonella. J Bacteriol 64:679–699.
55. Thomas CM, Nielsen KM. 2005. Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. Nat Rev Microbiol 3:711–721.
56. Mark Osborn A, Böltner D. 2002. When phage, plasmids, and transposons collide: genomic islands, and conjugative- and mobilizable-transposons as a mosaic continuum. Plasmid 48:202–212.
57. Cheetham BF, Katz ME. 1995. A role for bacteriophages in the evolution and transfer of bacterial virulence determinants. Mol Microbiol 18:201–208.

58. Palmer KL, Kos VN, Gilmore MS. 2010. Horizontal gene transfer and the genomics of enterococcal antibiotic resistance. Curr Opin Microbiol 13:632–639.
59. Balcazar JL. 2014. Bacteriophages as vehicles for antibiotic resistance genes in the environment. PLoS Pathog 10:e1004219–e1004219.
60. Brown-Jaque M, Calero-Cáceres W, Muniesa M. 2015. Transfer of antibiotic-resistance genes via phage-related mobile elements. Plasmid 79:1–7.
61. Carattoli A. 2013. Plasmids and the spread of resistance. Spec Issue Antibiot Resist 303:298–304.
62. Tsang J. 2017. Bacterial plasmid addiction systems and their implications for antibiotic drug development. Postdoc J J Postdr Res Postdr Aff 5:3–9.
63. Engelberg-Kulka H, Glaser G. 1999. Addiction Modules and Programmed Cell Death and Antideath in Bacterial Cultures. Annu Rev Microbiol 53:43–70.
64. Haack KR, Roth JR. 1995. Recombination between chromosomal IS200 elements supports frequent duplication formation in Salmonella typhimurium. Genetics 141:1245–1252.
65. Daveran-Mingot M-L, Campo N, Ritzenthaler P, Le Bourgeois P. 1998. A Natural Large Chromosomal Inversion inLactococcus lactis Is Mediated by Homologous Recombination between Two Insertion Sequences. J Bacteriol 180:4834.
66. Desmet L, Faelen M, Lefèbvre N, Résibois A, Toussaint A, van Gijsegem F. 1981. Genetic Study of Mu Transposition and Mu-mediated Chromosomal Rearrangements. Cold Spring Harb Symp Quant Biol 45:355–363.
67. Kusumoto M, Ooka T, Nishiya Y, Ogura Y, Saito T, Sekine Y, Iwata T, Akiba M, Hayashi T. 2011. Insertion sequence-excision enhancer removes transposable elements from bacterial genomes and induces various genomic deletions. Nat Commun 2:152.
68. Rice LB, Carias LL, Marshall S, Rudin SD, Hutton-Thomas R. 2005. Tn5386, a novel Tn916-like mobile element in Enterococcus faecium D344R that interacts with Tn916 to yield a large genomic deletion. J Bacteriol 187:6668–6677.
69. McInerney JO, McNally A, O'Connell MJ. 2017. Why prokaryotes have pangenomes. Nat Microbiol 2:17040.
70. Rouli L, Merhej V, Fournier P-E, Raoult D. 2015. The bacterial pangenome as a new tool for analysing pathogenic bacteria. New Microbes New Infect 7:72–85.
71. Baltrus DA. 2013. Exploring the costs of horizontal gene transfer. Trends Ecol Evol 28:489–495.
72. Park C, Zhang J. 2012. High Expression Hampers Horizontal Gene Transfer. Genome Biol Evol 4:523–532.

73. Pál C, Papp B, Lercher MJ. 2005. Horizontal gene transfer depends on gene content of the host. Bioinformatics 21:ii222–ii223.
74. Allan Drummond D, Wilke CO. 2009. The evolutionary consequences of erroneous protein synthesis. Nat Rev Genet 10:715.
75. Diaz Ricci JC, Hernández ME. 2000. Plasmid Effects on Escherichia coli Metabolism. Crit Rev Biotechnol 20:79–108.
76. Bragg JG, Wagner A. 2009. Protein material costs: single atoms can make an evolutionary difference. Trends Genet 25:5–8.
77. Shachrai I, Zaslaver A, Alon U, Dekel E. 2010. Cost of Unneeded Proteins in E. coli Is Reduced after Several Generations in Exponential Growth. Mol Cell 38:758–767.
78. Rankin DJ, Turner LA, Heinemann JA, Brown SP. 2012. The coevolution of toxin and antitoxin genes drives the dynamics of bacterial addiction complexes and intragenomic conflict. Proc Biol Sci 279:3706–3715.
79. Cooper TF, Paixão T, Heinemann JA. 2010. Within-host competition selects for plasmid-encoded toxin-antitoxin systems. Proc Biol Sci 277:3149–3155.
80. Bouma JE, Lenski RE. 1988. Evolution of a bacteria/plasmid association. Nature 335:351–352.
81. De Gelder L, Williams JJ, Ponciano JM, Sota M, Top EM. 2008. Adaptive plasmid evolution results in host-range expansion of a broad-host-range plasmid. Genetics 178:2179–2190.
82. Zwanzig M, Harrison E, Brockhurst MA, Hall JPJ, Berendonk TU, Berger U. 2019. Mobile Compensatory Mutations Promote Plasmid Survival. mSystems 4:e00186-18.
83. Stalder T, Rogers LM, Renfrow C, Yano H, Smith Z, Top EM. 2017. Emerging patterns of plasmid-host coevolution that stabilize antibiotic resistance. Sci Rep 7:4853–4853.
84. Dahlberg C, Chao L. 2003. Amelioration of the cost of conjugative plasmid carriage in Eschericha coli K12. Genetics 165:1641–1649.
85. Bottery MJ, Wood AJ, Brockhurst MA. 2017. Adaptive modulation of antibiotic resistance through intragenomic coevolution. Nat Ecol Evol 1:1364–1369.
86. San Millan A, Toll-Riera M, Qi Q, MacLean RC. 2015. Interactions between horizontally acquired genes create a fitness cost in Pseudomonas aeruginosa. Nat Commun 6:6845–6845.
87. San Millan A, Peña-Miller R, Toll-Riera M, Halbert ZV, McLean AR, Cooper BS, MacLean RC. 2014. Positive selection and compensatory adaptation interact to stabilize non-transmissible plasmids. Nat Commun 5:5208–5208.

88. Bliven KA, Maurelli AT. 2016. Evolution of Bacterial Pathogens Within the Human Host. Microbiol Spectr 4:10.1128/microbiolspec.VMBF-0017–2015.
89. Finlay BB, McFadden G. 2006. Anti-Immunology: Evasion of the Host Immune System by Bacterial and Viral Pathogens. Cell 124:767–782.
90. Henderson IR, Czeczulin J, Eslava C, Noriega F, Nataro JP. 1999. Characterization of pic, a secreted protease of Shigella flexneri and enteroaggregative Escherichia coli. Infect Immun 67:5587–5596.
91. Hardy L, Jespers V, Van den Bulck M, Buyze J, Mwambarangwe L, Musengamana V, Vaneechoutte M, Crucitti T. 2017. The presence of the putative Gardnerella vaginalis sialidase A gene in vaginal specimens is associated with bacterial vaginosis biofilm. PLOS ONE 12:e0172522.
92. Rooijakkers SHM, van Wamel WJB, Ruyken M, van Kessel KPM, van Strijp JAG. 2005. Anti-opsonic properties of staphylokinase. Microbes Infect 7:476–484.
93. Johnson MB, Criss AK. 2013. Neisseria gonorrhoeae phagosomes delay fusion with primary granules to enhance bacterial survival inside human neutrophils. Cell Microbiol 15:1323–1340.
94. Vink C, Rudenko G, Seifert HS. 2012. Microbial antigenic variation mediated by homologous DNA recombination. FEMS Microbiol Rev 36:917–948.
95. Vonaesch P, Anderson M, Sansonetti PJ. 2018. Pathogens, microbiome and the host: emergence of the ecological Koch's postulates. FEMS Microbiol Rev 42:273–292.
96. Inglis RF, Bayramoglu B, Gillor O, Ackermann M. The role of bacteriocins as selfish genetic elements. Biol Lett 9:20121173–20121173.
97. Nedialkova LP, Denzler R, Koeppel MB, Diehl M, Ring D, Wille T, Gerlach RG, Stecher B. 2014. Inflammation Fuels Colicin Ib-Dependent Competition of Salmonella Serovar Typhimurium and E. coli in Enterobacterial Blooms. PLOS Pathog 10:e1003844.
98. Sana TG, Flaugnatti N, Lugo KA, Lam LH, Jacobson A, Baylot V, Durand E, Journet L, Cascales E, Monack DM. 2016. Salmonella Typhimurium utilizes a T6SS-mediated antibacterial weapon to establish in the host gut. Proc Natl Acad Sci U S A 113:E5044–E5051.
99. Unterweger D, Miyata ST, Bachmann V, Brooks TM, Mullins T, Kostiuk B, Provenzano D, Pukatzki S. 2014. The Vibrio cholerae type VI secretion system employs diverse effector modules for intraspecific competition. Nat Commun 5:3549.
100. Anderson MC, Vonaesch P, Saffarian A, Marteyn BS, Sansonetti PJ. 2017. Shigella sonnei Encodes a Functional T6SS Used for Interbacterial Competition and Niche Occupancy. Cell Host Microbe 21:769-776.e3.

101. Wandersman C, Delepelaire P. 2004. Bacterial Iron Sources: From Siderophores to Hemophores. Annu Rev Microbiol 58:611–647.
102. Miethke M, Marahiel MA. 2007. Siderophore-Based Iron Acquisition and Pathogen Control. Microbiol Mol Biol Rev 71:413.
103. WHO. Antibiotic resistance.
104. Munita JM, Arias CA. 2016. Mechanisms of Antibiotic Resistance. Microbiol Spectr 4:10.1128/microbiolspec.VMBF-0016–2015.
105. NCBI. Bacterial Antimicrobial Resistance Reference Gene Database.
106. Kong K-F, Schneper L, Mathee K. 2010. Beta-lactam antibiotics: from antibiosis to resistance and bacteriology. APMIS Acta Pathol Microbiol Immunol Scand 118:1–36.
107. Chambers HF. 2001. The changing epidemiology of Staphylococcus aureus? Emerg Infect Dis 7:178–182.
108. Paterson DL, Bonomo RA. 2005. Extended-spectrum beta-lactamases: a clinical update. Clin Microbiol Rev 18:657–686.
109. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. 2012. Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet 13:601–612.
110. Köser CU, Ellington MJ, Cartwright EJP, Gillespie SH, Brown NM, Farrington M, Holden MTG, Dougan G, Bentley SD, Parkhill J, Peacock SJ. 2012. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog 8:e1002824–e1002824.
111. Didelot X, Nell S, Yang I, Woltemate S, van der Merwe S, Suerbaum S. 2013. Genomic evolution and transmission of Helicobacter pylori in two South African families. Proc Natl Acad Sci U S A 110:13880–13885.
112. Kennemann L, Didelot X, Aebischer T, Kuhn S, Drescher B, Droege M, Reinhardt R, Correa P, Meyer TF, Josenhans C, Falush D, Suerbaum S. 2011. Helicobacter pylori genome evolution during human infection. Proc Natl Acad Sci U S A 108:5033–5038.
113. Ford CB, Lin PL, Chase MR, Shah RR, Iartchouk O, Galagan J, Mohaideen N, Ioerger TR, Sacchettini JC, Lipsitch M, Flynn JL, Fortune SM. 2011. Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. Nat Genet 43:482–486.
114. Bryant JM, Grogono DM, Greaves D, Foweraker J, Roddick I, Inns T, Reacher M, Haworth CS, Curran MD, Harris SR, Peacock SJ, Parkhill J, Floto RA. 2013. Whole-genome sequencing to identify transmission of Mycobacterium abscessus between patients with cystic fibrosis: a retrospective cohort study. Lancet Lond Engl 381:1551–1560.

115. Wylie KM, Wylie TN, Minx PJ, Rosen DA. 2019. Whole-Genome Sequencing of Klebsiella pneumoniae Isolates to Track Strain Progression in a Single Patient With Recurrent Urinary Tract Infection. Front Cell Infect Microbiol 9:14.
116. Sheppard SK, Didelot X, Jolley KA, Darling AE, Pascoe B, Meric G, Kelly DJ, Cody A, Colles FM, Strachan NJC, Ogden ID, Forbes K, French NP, Carter P, Miller WG, McCarthy ND, Owen R, Litrup E, Egholm M, Affourtit JP, Bentley SD, Parkhill J, Maiden MCJ, Falush D. 2013. Progressive genome-wide introgression in agricultural Campylobacter coli. Mol Ecol 22:1051–1064.
117. Pandey A, Cleary DW, Laver JR, Gorringe A, Deasy AM, Dale AP, Morris PD, Didelot X, Maiden MCJ, Read RC. 2018. Microevolution of Neisseria lactamica during nasopharyngeal colonisation induced by controlled human infection. Nat Commun 9:4753–4753.
118. Moradigaravand D, Gouliouris T, Blane B, Naydenova P, Ludden C, Crawley C, Brown NM, Török ME, Parkhill J, Peacock SJ. 2017. Within-host evolution of Enterococcus faecium during longitudinal carriage and transition to bloodstream infection in immunocompromised patients. Genome Med 9:119–119.
119. Klughammer J, Dittrich M, Blom J, Mitesser V, Vogel U, Frosch M, Goesmann A, Müller T, Schoen C. 2017. Comparative Genome Sequencing Reveals Within-Host Genetic Changes in Neisseria meningitidis during Invasive Disease. PloS One 12:e0169892–e0169892.
120. Cao Q, Didelot X, Wu Z, Li Z, He L, Li Y, Ni M, You Y, Lin X, Li Z, Gong Y, Zheng M, Zhang M, Liu J, Wang W, Bo X, Falush D, Wang S, Zhang J. 2015. Progressive genomic convergence of two Helicobacter pylori strains during mixed infection of a patient with chronic gastritis. Gut 64:554.
121. Levade I, Terrat Y, Leducq J-B, Weil AA, Mayo-Smith LM, Chowdhury F, Khan AI, Boncy J, Buteau J, Ivers LC, Ryan ET, Charles RC, Calderwood SB, Qadri F, Harris JB, LaRocque RC, Shapiro BJ. 2017. Vibrio cholerae genomic diversity within and between patients. Microb Genomics 3:e000142.
122. Darch SE, McNally A, Harrison F, Corander J, Barr HL, Paszkiewicz K, Holden S, Fogarty A, Crusz SA, Diggle SP. 2015. Recombination is a key driver of genomic and phenotypic diversity in a Pseudomonas aeruginosa population during cystic fibrosis infection. Sci Rep 5:7649.
123. Diaz Caballero J, Clark ST, Wang PW, Donaldson SL, Coburn B, Tullis DE, Yau YCW, Waters VJ, Hwang DM, Guttman DS. 2018. A genome-wide association analysis reveals a potential role for recombination in the evolution of antimicrobial resistance in Burkholderia multivorans. PLoS Pathog 14:e1007453–e1007453.

124. Khademi SMH, Sazinas P, Jelsbak L. 2019. Within-Host Adaptation Mediated by Intergenic Evolution in Pseudomonas aeruginosa. Genome Biol Evol 11:1385–1397.
125. Price EP, Sarovich DS, Mayo M, Tuanyok A, Drees KP, Kaestli M, Beckstrom-Sternberg SM, Babic-Sternberg JS, Kidd TJ, Bell SC, Keim P, Pearson T, Currie BJ. 2013. Within-Host Evolution of Burkholderia pseudomallei over a Twelve-Year Chronic Carriage Infection. mBio 4:e00388-13.
126. Honsa ES, Cooper VS, Mhaissen MN, Frank M, Shaker J, Iverson A, Rubnitz J, Hayden RT, Lee RE, Rock CO, Tuomanen EI, Wolf J, Rosch JW. 2017. RelA Mutant Enterococcus faecium with Multiantibiotic Tolerance Arising in an Immunocompromised Host. mBio 8:e02124-16.
127. Sherrard LJ, Tai AS, Wee BA, Ramsay KA, Kidd TJ, Ben Zakour NL, Whiley DM, Beatson SA, Bell SC. 2017. Within-host whole genome analysis of an antibiotic resistant Pseudomonas aeruginosa strain sub-type in cystic fibrosis. PLOS ONE 12:e0172179.
128. Stanczak-Mrozek KI, Manne A, Knight GM, Gould K, Witney AA, Lindsay JA. 2015. Within-host diversity of MRSA antimicrobial resistances. J Antimicrob Chemother 70:2191–2198.
129. Mwangi MM, Wu SW, Zhou Y, Sieradzki K, de Lencastre H, Richardson P, Bruce D, Rubin E, Myers E, Siggia ED, Tomasz A. 2007. Tracking the in vivo evolution of multidrug resistance in Staphylococcus aureus by whole-genome sequencing. Proc Natl Acad Sci U S A 104:9451–9456.
130. Howden BP, McEvoy CRE, Allen DL, Chua K, Gao W, Harrison PF, Bell J, Coombs G, Bennett-Wood V, Porter JL, Robins-Browne R, Davies JK, Seemann T, Stinear TP. 2011. Evolution of Multidrug Resistance during Staphylococcus aureus Infection Involves Mutation of the Essential Two Component Regulator WalKR. PLOS Pathog 7:e1002359.
131. Sun G, Luo T, Yang C, Dong X, Li J, Zhu Y, Zheng H, Tian W, Wang S, Barry CE 3rd, Mei J, Gao Q. 2012. Dynamic population changes in Mycobacterium tuberculosis during acquisition and fixation of drug resistance in patients. J Infect Dis 206:1724–1733.
132. Eldholm V, Norheim G, von der Lippe B, Kinander W, Dahle UR, Caugant DA, Mannsåker T, Mengshoel AT, Dyrhol-Riise AM, Balloux F. 2014. Evolution of extensively drug-resistant Mycobacterium tuberculosisfrom a susceptible ancestor in a single patient. Genome Biol 15:490.
133. Saunders NJ, Trivedi UH, Thomson ML, Doig C, Laurenson IF, Blaxter ML. 2011. Deep resequencing of serial sputum isolates of Mycobacterium tuberculosis during therapeutic failure

due to poor compliance reveals stepwise mutation of key resistance genes on an otherwise stable genetic background. J Infect 62:212–217.

134. Merker M, Kohl TA, Roetzer A, Truebe L, Richter E, Rüsch-Gerdes S, Fattorini L, Oggioni MR, Cox H, Varaine F, Niemann S. 2013. Whole genome sequencing reveals complex evolution patterns of multidrug-resistant Mycobacterium tuberculosis Beijing strains in patients. PloS One 8:e82551–e82551.

135. O'Neill MB, Mortimer TD, Pepperell CS. 2015. Diversity of Mycobacterium tuberculosis across Evolutionary Scales. PLOS Pathog 11:e1005257.

136. Gilchrist CA, Turner SD, Riley MF, Petri WA, Hewlett EL. 2015. Whole-Genome Sequencing in Outbreak Analysis. Clin Microbiol Rev 28:541.

137. Knowles LL, Maddison WP. 2002. Statistical Phylogeography. Mol Ecol 11:2623–2635.

138. Lemey P, Rambaut A, Drummond AJ, Suchard MA. 2009. Bayesian phylogeography finds its roots. PLoS Comput Biol 5:e1000520–e1000520.

139. Holmes EC. 2004. The phylogeography of human viruses. Mol Ecol 13:745–756.

140. Waskito L, Yamaoka Y. 2019. The Story of Helicobacter pylori: Depicting Human Migrations from the Phylogeography.

141. Linz B, Balloux F, Moodley Y, Manica A, Liu H, Roumagnac P, Falush D, Stamer C, Prugnolle F, van der Merwe SW, Yamaoka Y, Graham DY, Perez-Trallero E, Wadstrom T, Suerbaum S, Achtman M. 2007. An African origin for the intimate association between humans and Helicobacter pylori. Nature 445:915.

142. Moodley Y, Linz B, Bond RP, Nieuwoudt M, Soodyall H, Schlebusch CM, Bernhöft S, Hale J, Suerbaum S, Mugisha L, van der Merwe SW, Achtman M. 2012. Age of the Association between Helicobacter pylori and Man. PLOS Pathog 8:e1002693.

143. Monot M, Honoré N, Garnier T, Zidane N, Sherafi D, Paniz-Mondolfi A, Matsuoka M, Taylor GM, Donoghue HD, Bouwman A, Mays S, Watson C, Lockwood D, Khamesipour A, Dowlati Y, Jianping S, Rea TH, Vera-Cabrera L, Stefani MM, Banu S, Macdonald M, Sapkota BR, Spencer JS, Thomas J, Harshman K, Singh P, Busso P, Gattiker A, Rougemont J, Brennan PJ, Cole ST. 2009. Comparative genomic and phylogeographic analysis of Mycobacterium leprae. Nat Genet 41:1282.

144. Benjak A, Avanzi C, Singh P, Loiseau C, Girma S, Busso P, Fontes ANB, Miyamoto Y, Namisato M, Bobosha K, Salgado CG, da Silva MB, Bouth RC, Frade MAC, Filho FB, Barreto JG, Nery JAC, Bührer-Sékula S, Lupien A, Al-Samie AR, Al-Qubati Y, Alkubati AS, Bretzel G, Vera-Cabrera L, Sakho F, Johnson CR, Kodio M, Fomba A, Sow SO, Gado M, Konaté O, Stefani MMA, Penna GO, Suffys PN, Sarno EN, Moraes MO, Rosa PS, Baptista IMFD, Spencer

JS, Aseffa A, Matsuoka M, Kai M, Cole ST. 2018. Phylogenomics and antimicrobial resistance of the leprosy bacillus Mycobacterium leprae. Nat Commun 9:352–352.

145. Sánchez-Busó L, Golparian D, Corander J, Grad YH, Ohnishi M, Flemming R, Parkhill J, Bentley SD, Unemo M, Harris SR. 2018. Antimicrobial exposure in sexual networks drives divergent evolution in modern gonococci. bioRxiv 334847.

146. Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, Croucher NJ, Choi SY, Harris SR, Lebens M, Niyogi SK, Kim EJ, Ramamurthy T, Chun J, Wood JLN, Clemens JD, Czerkinsky C, Nair GB, Holmgren J, Parkhill J, Dougan G. 2011. Evidence for several waves of global transmission in the seventh cholera pandemic. Nature 477:462–465.

147. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G, Yeboah-Manu D, Bothamley G, Mei J, Wei L, Bentley S, Harris SR, Niemann S, Diel R, Aseffa A, Gao Q, Young D, Gagneux S. 2013. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nat Genet 45:1176–1182.

148. Kay GL, Sergeant MJ, Zhou Z, Chan JZ-M, Millard A, Quick J, Szikossy I, Pap I, Spigelman M, Loman NJ, Achtman M, Donoghue HD, Pallen MJ. 2015. Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. Nat Commun 6:6717.

149. Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, Forrest SA, Bryant JM, Harris SR, Schuenemann VJ, Campbell TJ, Majander K, Wilbur AK, Guichon RA, Wolfe Steadman DL, Cook DC, Niemann S, Behr MA, Zumarraga M, Bastida R, Huson D, Nieselt K, Young D, Parkhill J, Buikstra JE, Gagneux S, Stone AC, Krause J. 2014. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature advance online publication.

150. Pepperell CS, Granka JM, Alexander DC, Behr MA, Chui L, Gordon J, Guthrie JL, Jamieson FB, Langlois-Klassen D, Long R, Nguyen D, Wobeser W, Feldman MW. 2011. Dispersal of Mycobacterium tuberculosis via the Canadian fur trade. Proc Natl Acad Sci 108:6526–6531.

151. Brynildsrud OB, Pepperell CS, Suffys P, Grandjean L, Monteserin J, Debech N, Bohlin J, Alfsnes K, Pettersson JO-H, Kirkeleite I, Fandinho F, da Silva MA, Perdigao J, Portugal I, Viveiros M, Clark T, Caws M, Dunstan S, Thai PVK, Lopez B, Ritacco V, Kitchen A, Brown TS, van Soolingen D, O'Neill MB, Holt KE, Feil EJ, Mathema B, Balloux F, Eldholm V. 2018. Global expansion of Mycobacterium tuberculosis lineage 4 shaped by colonial migration and local adaptation. Sci Adv 4:eaat5869–eaat5869.

Chapter 2: Adaptation of the Gonococcal Genetic Island

This project was performed under the supervision of Caitlin S. Pepperell in collaboration with Joseph P. Dillard and Melanie Callaghan. CSP and JPD conceived of the study. AS and CSP designed the analyses, analyzed the data and drafted the manuscript.

Abstract

Due to its propensity to engage in lateral gene transfer (LGT), *Neisseria gonorrhoeae*, the causative agent of gonorrhea, is acquiring antibiotic resistance at an alarming rate. *N. gonorrhoeae* possesses a Type IV secretion system encoded on a genetic island, the Gonococcal Genetic Island (GGI), that secretes ssDNA facilitating LGT. The GGI is prevalent among *N. gonorrhoeae*, but little is known about the evolution of the GGI or its role in shaping *N. gonorrhoeae's* genome. Using whole genome sequence data, we characterized the diversity of the GGI and analyzed pan-genome structure and patterns of core genome recombination in GGI+ and GGI- sub-populations. We identified patterns of core genome recombination that differentiate these populations. Additionally, we found evidence of epistatic interactions between the GGI and *N. gonorrhoeae's* core genome, and multiple facets of the GGI's diversity suggest it acts as a mobile element. These findings demonstrate the importance of mobile elements in shaping their host chromosomes.

Introduction

Neisseria gonorrhoeae is an obligate human pathogen and the causative agent of gonorrhea. The Centers for Disease Control estimates 820,000 new infections with *N. gonorrhoeae* occur in the United States each year. The rate of reported cases in the United States rose by 18.6% in 2017 and has increased every year since 2013 [a 67% increase overall; (1)]. *N. gonorrhoeae* has been declared an urgent public health threat, as increasing rates of antibiotic resistance raise the specter of untreaTable 2.forms of the disease.

N. gonorrhoeae's propensity to engage in LGT has facilitated the spread of antibiotic resistant gonorrhea. Transformation, the direct uptake and incorporation of exogenous DNA into the native chromosome, is the primary mechanism of LGT used by *N. gonorrhoeae* (2). *N. gonorrhoeae* possesses two mechanisms by which it releases its DNA into the external milieu: autolysis and Type IV secretion (T4S). *N. gonorrhoeae's* T4S system (T4SS) secretes ssDNA

across the cell envelope (3, 4). In an experimental mating system, T4S conferred 500-fold greater transformation efficiency than autolysis, raising the possibility that T4S enables high rates of natural transformation (3, 5).

Genes encoding the T4S system are located on a 59kb genomic island, the Gonococcal Genetic Island (GGI), which is prevalent among *N. gonorrhoeae* [approximately 64-80% of isolates; (6, 7)]. The GGI, first described in *N. gonorrhoeae* reference strain MS11, contains 66 genes. Genes encoding the T4SS share sequence similarity to the *E. coli* F-plasmid T4SS. Twenty-two genes in the GGI have been identified as essential to secretion and 6 other genes have sequence similarity to DNA binding and processing proteins, but the remaining 38 genes are non-essential for secretion and largely of unknown function (3, 5, 8).

The GGI has distinctly different sequence characteristics from *N. gonorrhoeae's* chromosome, including lower GC-content, differing dinucleotide frequencies and a lower average number of DNA uptake sequences (4, 8). This suggests the GGI may have been laterally acquired from an exogenous source.

Excision of the GGI is mediated by the site-specific recombinase XerCD (9). XerCD recognizes the *dif* site, a repeat sequence flanking both sides of the GGI. Loss, but not transfer, following excision of the GGI has been observed in experimental settings (4, 9). Given that the GGI may have been laterally acquired and is capable of being excised from its host chromosome, it is possible the GGI was or may be a mobile genetic element (MGE).
Here we sought to characterize the evolutionary history of the GGI and elucidate its role in shaping its host chromosome using whole genome sequence data (WGS) from natural populations of *N. gonorrhoeae*.

Methods

Data and *de novo* assembly

For genomes with only raw read data available, we downloaded publicly available WGS for *N. gonorrhoeae* from the European Nucleotide Archive (*n* = 167). We normalized sequences to uniform coverage (150X) using the error correction and read normalization tool BBNorm (10). We *de novo* assembled the subsampled sequences using the iMetAMOS pipeline (11), comparing assemblies from SPAdes (12), MaSurCA (13), and Velvet (14). We downloaded an additional 16 finished/draft genomes from the National Center for Biotechnology Information. The final data set contains 183 *N. gonorrhoeae* isolates (Table 2.S1).

Pan-genome calculation

We used Prokka (15) to annotate the *de novo* assemblies and finished/draft genomes and Roary to calculate the pan-genome, identify core and accessory genes, and generate core genome alignments for our sample of *N. gonorrhoeae* (16). We calculated the pan-genome of our entire sample, isolates with the GGI and isolates without the GGI.

Determining GGI presence/absence

We used the results of the pan-genome analysis to determine GGI presence/absence in our sample of *N. gonorrhoeae*. We identified the bounds of the GGI (as first described in *N. gonorrhoeae* reference strain MS11: *parA* and *yaa* using BLAST and defined GGI presence as possessing ≥ 70% of the gene content within these bounds (4, 17, 18). Using the genes of the GGI described in MS11 as a custom database, we verified genes within these bounds as belonging to the GGI and classified them according to their MS11 nomenclature using BLAST. Additionally, non-syntenic paralogs (genes outside the bounds of the GGI with significant homology to the genes present in the GGI encoded by MS11) were identified in the pan-genome using BLAST.

SNP identification and alignment

We used SNP-sites to convert the core genome alignments to a multi-sample variant call format (VCF) file (19). The core genome of reference strain NCCP11945 was used as a reference to call SNPs.

Maximum likelihood phylogenetic analysis

We used RAxML to infer maximum likelihood phylogenetic trees (20). We used the GTRGAMMA substitution model and performed bootstrapping using the autoMR convergence criteria. We used ggtree to visualize trees (21).

Core genome diversity

We calculated π, Θw, and Tajima's D for the core genome and the core genes of the GGI using egglib (22).

Recombination analyses

We used BratNextGen to identify recombination breakpoints in core genome alignments and visualized these breakpoints using Phandango (23, 24). We used SplitsTree4 to visualize phylogenetic networks and perform the phi-test for recombination (25). We used fineSTRUCTURE to identify co-inherited haplotypes within our sample (26). We inferred the relative recombination intensity (H_i) for each bi-allelic SNP in the core genome alignment using OrderedPainting (27). The recombination intensity of each gene in the core genome was calculated as the mean of H_i for all SNPs within that gene. To identify genes with extreme differences in H_i between GGI+ isolates and GGI- isolates, we calculated the fold-change in H_i between GGI+ and GGI- isolates (i.e. H_i in GGI+/ H_i in GGI-) for each gene in the core genome. We performed z-transformation of these values and calculated a p-value for each gene. We used Bonferroni correction for multiple testing, setting a cutoff of 0.05 to identify outliers.

Ancestral reconstruction of the GGI

We used adegenet (28) to perform principal component analysis on the GGI's core genes (present in ≥ 99% of GGI+ isolates). We reconstructed gains and losses of the GGI groups identified by PCA on the maximum-likelihood phylogeny using APE (29). We defined gains and losses as a change in the majority (> 50%) reconstructed ancestral state from parent to child node.

Trait-phylogeny association test

We used BaTS, a bayesian sampling method that accounts for phylogenetic uncertainty, to test for a significant association between GGI presence/absence as a trait and its location on the maximum-likelihood phylogeny (30). We generated a distribution of phylogenies using MrBayes with the GTRGAMMA substitution model (31). Markov chains were run in duplicate for 200 million generations each with sampling every 20,000 generations, and the first 2 million generations were discarded as burn-in. We randomly sampled 1500 trees from this distribution for the trait-phylogeny association test.

F_{ST} calculation

We calculated Weir and Cockerham's F_{ST} for biallelic SNPs in the core genomes of GGI+ and GGI- populations using vcflib (https://github.com/vcflib/vcflib). We permuted this analysis 100X and used the maximum F_{ST} observed in the null distribution as a cut-off to identify F_{ST} outliers.

Homoplasy identification

We used TreeTime to perform ancestral reconstruction of core genome SNPs on the maximum-likelihood phylogeny (32). We identified homoplastic SNPs (SNPs that arose > 2 times on the phylogeny) from the output of TreeTime using custom python scripts.

Pan-genome wide association study

We used Scoary to calculate the strength of association between accessory gene content and the GGI as a trait (33). We used the Bonferroni correction to account for multiple comparisons (p-value cutoff of < 0.05) and performed 100 permutations of the analysis.

Results

Identification of the Gonococcal Genetic Island among isolates of *N. gonorrhoeae*

The Gonococcal Genetic Island (GGI) was first described in *N. gonorrhoeae* laboratory reference strain MS11: 66 genes were identified between bounds defined by the genes *yaa* and *parA* in this strain (4, 8). We used a core genome analysis to define the GGI in our sample of *N.*

gonorrhoeae: we used Prokka for annotation and Roary to delineate the pan and core genomes among the annotated genes (15, 16). These results indicated there were a total of 77 genes within the bounds of the GGI among the 183 isolates. The isolates could be clearly separated into a group encoding most (i.e. 53 or more genes, ≥ 70%) of the gene content and a smaller group (n = 4) encoding a fragment of the GGI. Of the 183 isolates, 24 encoded less than 1% of the GGI's gene content (Figure 2.S1). Based on these observations, we designated isolates in which at least 53 of the genes within the bounds of the GGI were present as having the genomic island. Applying this threshold definition, we found that 62% of isolates encode the GGI (n = 115). Non-syntenic paralogs (i.e. genes outside the bounds of the GGI with significant homology to the genes present in the GGI encoded by MS11) were identified in the pan-genome using BLAST. Querying the pan-genome in this way, we identified 27 additional genes. Two genes, *yegB* and *yee*, were not annotated by Prokka. Their DNA sequences were obtained from the *de novo* assemblies and draft/complete genomes using BLAST and aligned using MAFFT (17, 18, 34). Among isolates with the GGI, 48 genes defined the "core GGI" (i.e. present in 99% of isolates, Table 2.S2). Our estimate of GGI prevalence contrasts with a previous study, based on PCR identification of the genes *altA* and *traG* applied to 115 isolates, which estimated that 80% of isolates encoded the island (Dillard and Seifert 2001). The strain collection from Dillard and Seifert 2001 was purposefully chosen to include a significant number of isolates from specific disease states e.g. pelvic inflammatory disease and disseminated gonococcal infection (DGI). Dillard and Seifert 2011 found and association between the GGI and DGI, and the abundance of DGI isolates from their strain collection (32/115) likely skewed their estimate of 80%.

The results of core genome analyses for the complete sample, as well as sub-groups defined by GGI presence/absence, are shown in Table 2.1. The average number of genes encoded by each isolate is ~2000, whereas the core genome size ranges from ~1400-1600 genes. This indicates that a substantial proportion of gene content (20-30%) varies from isolate to isolate. While the average number of genes per isolate is similar in subgroups with and

without the GGI, genome configuration varies, with GGI+ isolates encoding more of their gene content in the core, as opposed to accessory, genome (Figure 2.S2). The difference between groups is particularly striking in the size of the cloud genome (genes present in < 15% of isolates), which includes almost 1000 more genes in the GGI- group. This indicates that genes within the larger accessory genome of GGI- isolates are skewed to rare frequencies.

We calculated genome-wide values of π, Θw, and Tajima's D for the core genomes of GGI+ and GGI- isolates (Table 2.1). Genome-wide values of Tajima's D were slightly negative. Genome-wide values of π and Θw were similar between groups, but the genome-wide value of Tajima's D was lower in isolates of *N. gonorrhoeae* with the GGI.

Recombination in the core genome and within the GGI

We used BratNextGen to identify recombinant breakpoints in a core genome alignment from the complete sample of *N. gonorrhoeae* isolates (23). According to this analysis, 63% of sites in the core genome have been affected by recombination. Recombinant fragments, with a median length of 300 bp (Figure 2.S3), are numerous and widely dispersed across *N. gonorrhoeae's* core genome. Some recombinant fragments are shared among closely related isolates, suggesting vertical inheritance following transfer into a common ancestor (Figure 2.1).

The proportion of recombinant sites per genome was not significantly different between groups (Figure 2.2); however, a greater proportion of the alignment of all genomes has been affected by recombination in isolates with the GGI than in those without the GGI (47% versus 35%, respectively). This indicates that recombinant tracts are more likely to be shared among GGI- isolates; there is a greater diversity of recombinant fragments in the GGI+ group, consistent with higher rates of recombination. This is also reflected in a wider distribution of recombinant sites per genome in the GGI+ group (Figure 2.2).

In order to account for the effect of sample size on the differences in the proportion of the alignment affected by recombination in GGI+ and GGI- isolates, we randomly selected samples of 30 isolates from each group and calculated the proportion of the alignment affected

by recombination, repeating this procedure 50X. Results from the subsampling analyses confirmed that a greater proportion of the alignment was affected by recombination in GGI+ isolates (Figure 2.2).

Analysis of the individual core genomes for each group is consistent with differing amounts of recombination in the core genomes of these groups. The proportion of recombinant sites in the core genome of GGI+ isolates is significantly greater than that of GGI- isolates (Figure 2.3).

Network analyses of the individual core genomes were also consistent with differing amounts of recombination in the isolates with and without the GGI: there are more reticulations in the network of GGI+ isolates than in the network of GGI- isolates (Figure 2.4). The PHI test for recombination was significant for both groups ($p < 0.05$).

In addition to differences in the amount of recombination, GGI+ and GGI- isolates differed with respect to regional patterns of recombination. We used OrderedPainting to estimate the relative intensity of recombination (H_i) along the genome and identify 'hotspots' and 'coldspots' (27). These results revealed spatial biases in the amount of recombination, with clear hotspots evident along the core genome (Figure 2.5, Table 2.S3). We identified three genes that were recombination hotspots in an alignment including GGI+ and GGI- isolates. The locations of other hotspots differed between isolates of *N. gonorrhoae* with and without the GGI (Figure 2.5, Table 2.S3).

To further characterize regional patterns of recombination in GGI+ and GGI- isolates, we performed z-score transformation of fold change in H_i values between the two groups (e.g. GGI+ H_i / GGI- H_i) and identified 11 genes with extreme fold changes in recombination intensities (Table 2.S4). Regional differences in the intensity of recombination may reflect the influence of purifying selection (coldspots) and diversifying selection [hotspots; (35)]; the genes with marked differences in recombination intensity among GGI+ and GGI- isolates are candidate loci under distinct selection pressures in the two groups.

Gene flow between GGI+ and GGI- sub-groups

In order to identify potential barriers to gene flow between GGI+ and GGI- populations, we used fineSTRUCTURE to identify co-inherited haplotypes within our sample (26). There was no structured pattern of admixture between GGI+ and GGI- populations in the resulting co-ancestry matrix (Figure 2.6).

Intra-locus recombination in the GGI

The GGI showed evidence of intra-locus recombination, as indicated by reticulations in a network based on a 'GGI core' alignment (Figure 2.7) and a significant PHI test ($p < 0.05$).

Associations between the GGI and loci in the core and accessory genomes of *N. gonorrhoeae*

Phylogenetic analysis showed that isolates concordant for GGI presence/absence form monophyletic groups on the core genome phylogeny (Figure 2.8). Principal component analysis of genetic diversity within the core genome of the GGI (48 genes) delineated five distinct groups, which we refer to as I-V (Figure 2.9). These groups were also highly structured on the core genome phylogeny (Figure 2.10). Ancestral reconstruction of GGI type along the core genome phylogeny indicated that the element has been gained 13 and lost 2 times during the evolutionary history of our sample (Figure 2.10). Additionally, we identified 3 instances in which a monophyletic GGI+ group gained a GGI of a different group. We refer to this as a "switch."

We used BaTS to formally test for an association between genetic background (core genome phylogeny) and GGI type (28). The results of this analysis showed a significant association between trait and phylogeny for each GGI group, with the exception of group IV. Our observation of associations between patterns of diversity in the GGI and core genome suggest there could be epistatic interactions between the core genome and the genetic island (Table 2.2).

We used F_{ST} outlier analysis to identify core genome loci with a potential role interacting with the GGI. GGI+ and GGI- isolates were treated as two distinct populations, and we used

estimates of F_{ST} for polymorphic sites in the core genome to identify variants segregating at markedly different frequencies in these two populations. We identified 653 sites with extreme values of F_{ST} (F_{ST} > 0.19 e.g. the maximum F_{ST} observed in a null distribution of values calculated from 100 permutations of the F_{ST} outlier analysis), 570 of which were homoplastic (i.e. arose more than once on the phylogeny; Table 2.S5). The 570 SNPs correspond to 214 genes, 78 of which had ≥ 2 outliers. We randomly sampled 570 SNPs from the core genome (17299 bi-allelic SNPs) 100X and found the maximum number of genes sampled with ≥ 2 SNPs was 126. The emergence of 409 of these homoplastic F_{ST} outliers coincided with gains of the GGI (Table 2.S6).

In order to permute this analysis, we randomly sampled the same number of gains at nodes and tips of the core genome phylogeny and counted the number of times a homoplastic F_{ST} outlier emerged in concert with these randomly assigned gains. The maximum number of homoplastic F_{ST} outliers that emerged with randomly assigned gains from 100 permutations was 316.

Given the substantial variation in gene content we observed among isolates of *N. gonorrhoeae*, we hypothesized that the accessory genome could show signs of co-adaptation with the GGI. Defining the sub-groups as above, we used Scoary to calculate the strength of association between all genes in the accessory genome and presence/absence of the GGI (31). Genes within the GGI had the greatest strength of association, indicating the method performs as expected ($p < 0.05$). We identified four genes (all annotated as hypothetical proteins) that were strongly associated with GGI presence/absence and arose multiple times on the phylogeny (Table 2.3). These genes are candidate loci involved in epistatic interactions with the genetic island.

Associations between other mobile elements and loci in the core and accessory genomes of *N. gonorrhoeae*

Given the associations we observed between the GGI and *N. gonorrhoeae's core genome*, and the GGI's apparent mobility on the core genome phylogeny, we hypothesized that other mobile genetic elements might show similar patterns of association with the core genome in their evolutionary history. From the results of the pan-genome analysis, we identified 112 genes with annotations corresponding to mobile genetic elements (Table 2.S7). We chose a gene encoding a phage tail protein at segregating frequencies in our sample (n = 101). Ancestral reconstruction of this gene on the core genome phylogeny indicated the gene was gained three times and lost one time during the evolutionary history of our sample and isolates concordant for presence/absence of the phage tail form monophyletic groups on the core genome phylogeny (Figure 2.S4).

The upper quantiles of F_{ST} values were higher using presence/absence of the phage tail protein to define populations than presence/absence of the GGI (Figure 2.S5). This is reflective of the highly structured distribution of the phage tail on the core genome phylogeny. We identified 2342 F_{ST} outliers, 2035 of which were homoplastic. These 2035 homoplastic F_{ST} outliers correspond to 624 genes, 308 of which had ≥ 2 SNPs. The maximum number of multiple SNPs calculated from 100 permutations of randomly sampling 2035 SNPs from the core genome was 436. The number of homoplastic F_{ST} outliers gained in concert with the phage tail protein was 367. We permuted this analysis as described above and calculated a maximum of 663 homoplastic F_{ST} outliers emerged with randomly assigned gains. Repeating the pan-genome wide association study using Scoary, we found 34 genes strongly associated with presence/absence of the phage tail protein.

Discussion

Here we examined the diversity and characterized the pan-genome of natural populations of *N. gonorrhoeae* in order to gain insight into the evolution of the GGI. It's important to understand the adaptation of this locus, because co-culture experiments suggest it

may have a profound impact on the landscape of LGT in *N. gonorrhoeae*. In our sample of 183 *N. gonorrhoeae* isolates, we found the GGI segregates at an intermediate frequency (62%), which is consistent with recent estimates (6, 7). We found the GGI is associated with specific core genome variants and accessory gene content and presence/absence of the genetic island defines distinct sub-populations that form monophyletic groups on a core genome phylogeny. The GGI has been gained and lost several times during the evolutionary history of our sample and this state has been maintained during the evolution of these well-defined sub-groups. These observations suggest that presence or absence of the GGI is stabilized by natural selection and accompanied by adaptation to the element by the core and accessory genomes. Our observation that isolates tend to encode all or none of the GGI's genes is also consistent with a role for natural selection in maintaining the carriage state of the island. We observed the overall frequency of recombination to be high across GGI positive and negative groups, with isolates carrying the genomic island appearing to recombine more frequently than those that don't. Some recombination hotspots differ between these groups, which is consistent with adaptation to the element.

Our analyses indicate that recombination is frequent in natural populations of N. gonorrhoeae. *N. gonorrhoeae* has an open pan-genome; gene content differed substantially from isolate to isolate in our sample, with just 60% of gene content shared among all the isolates analyzed here. We found recombinant fragments to be numerous and widely distributed across *N. gonorrhoeae's* core genome, and the estimated number of sites affected by recombination comprised more than half of the core genome. The patterns of recombination we observed here, both the median fragment length and the distribution of fragment lengths, are characteristic of bacteria that recombine via natural transformation, e.g. *Neisseria meningitidis* [median recombinant fragment length ~400 bp (36)] and *Streptococcus pneumoniae* [mean fragment length ~2000 bp (37, 38)].

Similar to other bacterial species that undergo frequent recombination we observed relatively high levels of genetic diversity in this sample of *N. gonorrhoeae*: our estimates of pairwise differences (π) and segregating polymorphisms (θ) per site in the core genome were both equal to 0.002 (Table 2.1). This is consistent with previous estimates of diversity in *N. gonorrhoeae* (39).

In contrast to *N. gonorrhoeae's* pan-genome, our analyses indicate the genes within the bounds of the GGI are largely at core frequencies (Table 2.S2). This core complement consists of all genes previously identified as essential to secretion, as well as genes of unknown function that are likely essential to the element (39).

Previous studies have shown the GGI has distinctly different sequence characteristics from *N. gonorrhoeae's* core genome and is excised by the site-specific recombinase XerCD at the *dif* sites (4, 9). The *dif* site is highly conserved among prokaryotes (40), and a number of mobile genetic elements exploit XerCD recombinases to facilitate their transfer (41). This suggests the GGI was a mobile element laterally acquired from an exogenous source and may be mobile in extant populations of *N. gonorrhoeae*.

Our results indicate the GGI remains mobile in natural populations of *N. gonorrhoeae*. We inferred several instances of GGI gain and loss during the evolutionary history of our sample (Figure 2.10). The GGI was associated with specific variants in the core genome, which arose repeatedly along with the element. Five types of GGI were identifiable in our sample, and these were highly structured on the core genome phylogeny, suggesting a barrier to transfer in GGI-populations (Figure 2.9, Figure 2.10).

Random integration of MGEs may disrupt the regulatory and/or physiological networks of bacterial genomes, outweighing their potential benefits and imposing a fitness cost on their host (42). Integrating at a "permissive" location in a host's genome, such as the *dif* site recognized by XerCD, may attenuate these costs (43). Fitness costs associated with MGEs may also be mitigated by compensatory mutations in the chromosome, the MGE or both. These epistatic

interactions between laterally acquired elements and genetic background have been demonstrated in experimental evolution experiments of plasmids (44–49). A non-replicating plasmid carrying part of the GGI and its *dif* site can be integrated into isogenic GGI mutants, but complete transfer of the GGI to isogenic GGI mutants has not been observed (4, 9). This could be reflective of a conflict between the genomic island and genetic backgrounds the element has not evolved with in concert.

We found evidence of associations between the element and variants in both the core and accessory genomes of *N. gonorrhoeae*, which is also suggestive of epistatic interactions between the genomic island and the genome of *N. gonorrhoeae (*Table 2.3, Table 2.5*)*. Additionally, we discovered that hot spots and cold spots for recombination differ between the core genomes of GGI+ and GGI- *N. gonorrhoeae* (Table 2.S4). Lastly, we identified accessory gene content that emerged repeatedly in association with the GGI (Table 2.3). These SNPs and accessory gene content are suggestive of compensatory alterations of the genetic landscape of GGI+ populations that have allowed for the maintenance of the GGI in natural populations.

The GGI is not exclusive to *N. gonorrhoeae*, and the GGI of other *Neisseria spp.* may provide insight into the maintenance of *N. gonorrhoeae's* GGI in natural populations. *N. meningitidis* contains variants of the GGI (5, 50, 51), albeit at a much lower frequency (approximately 17% of isolates) and the GGI has been identified in the genome sequences of *Neisseria bacilliformis* (5, 52) and *Neisseria musculi* (unpublished). The GGI of some *N. meningitidis* strains have accumulated indels in genes essential for secretion in *N. gonorrhoeae*, potentially nullifying T4S. As for *N. meningitidis* strains with a complete T4SS, their T4SS does not appear to secrete ssDNA. Further, no connection has been observed between the presence of the T4S of *N. meningitidis* and the infection process. The indels and subsequent degradation of the GGI present in *N. meningitidis* are reflective of compensatory mutations to mitigate the fitness-cost of the island.

The GGI was present at an intermediate frequency (62%) in this sample of 183 isolates of *N. gonorrhoeae*. The GGI's segregation at an intermediate frequency in the presence of abundant recombination and its apparently intact capacity for mobilization suggest that the element is under some form of balancing selection. A recent study of genome dynamics in *S. pneumoniae* found evidence of negative frequency dependent selection, a form of balancing selection, in accessory gene frequency distributions of natural bacterial populations (53). The accessory genes found at intermediate frequencies included MGEs, and the authors hypothesized that dynamics between MGE and protective restriction modification (RM) systems shaped the frequency distribution of these elements in *S. pneumoniae*. Given that the GGI appears mobile in natural populations of *N. gonorrhoeae*, it is possible that dynamics similar to those operating in *S. pneumoniae* have stabilized the element at an intermediate frequency. We identified 49 genes with annotations corresponding to RM systems in our pan-genome analysis, none of which were specific to GGI+ or GGI- isolates (Table 2.S8), and 112 genes with annotations corresponding to MGEs in our sample, largely at core or rare accessory frequencies (Figure 2.S6, Table 2.S7). One of these genes, a phage tail protein at intermediate frequency in the sample, showed similar differentiation between +/- sub-populations.

Another, not mutually exclusive, possibility is that intransitivity [often described as 'rock-paper-scissors' dynamics; (53, 54)] has enabled the sTable 2.circulation of GGI+ and GGI- lineages. Intransitivity has been described in association with inter-strain competition enabled by toxin-antitoxin and other systems; the GGI notably encodes a toxin-antitoxin system as part of its core genome. Excluding the GGI, we found seven genes with annotations corresponding to toxin-antitoxin systems in the pan-genome calculated using Roary. These systems were not unique to either the GGI+ or GGI- sub-populations (Table 2.S9).

Finally, the genes encoded in the GGI have not all been characterized, and it is also possible that some of its gene products are involved in the kind of interactions with the host that result in sTable 2.diversity of the interacting molecules. We have no evidence to suggest that

GGI+ and GGI- bacteria occupy distinct niches, another potential explanation for segregating diversity at this locus. However, previous studies have shown the T4SS, through unknown mechanisms, is involved in *tonB* independent intracellular iron uptake. Interestingly, this phenotype has been observed even in the absence of DNA uptake or secretion (55). Additionally, the single stranded DNA secreted by the T4SS is the involved initial stages of biofilm development: Cells with a functional T4SS form more sTable 2.biofilms cells deficient for secretion (56).

Here we have described patterns of pan-genome structure and recombination that differentiate natural populations of *N. gonorrhoeae* based on presence/absence of the GGI. We found that the island may act as a mobile element and maintains a non-random distribution in natural populations, likely due to epistatic interactions between the element and genetic background. These findings demonstrate the importance of mobile elements in shaping their host chromosomes and provides a new analytical framework for identifying co-adaptation of laterally acquired DNA in natural populations of bacteria.

Acknowledgements

We would like to thank Joe Dillard and Melanie Callaghan for their expertise and insight on this project. AS and CP are supported by the National Institutes of Health (R01AI113287 and R01AI047958). Funding for this project was also provided by the University of Wisconsin School of Medicine and Public Health from the Wisconsin Partnership Program.

References

1. CDC. 2018. Gonorrhea - 2017 Sexually Transmitted Diseases Surveillance.
2. KOOMEY M. 1998. Competence for Natural transformation in Neisseria gonorrhoeae: A Model System for Studies of Horizontal Gene Transfer. APMIS 106:56–61.

3. Dillard JP, Seifert HS. 2001. A variable genetic island specific for Neisseria gonorrhoeae is involved in providing DNA for natural transformation and is found more often in disseminated infection isolates. Molecular Microbiology 41:263–277.

4. Hamilton HL, Domínguez NM, Schwartz KJ, Hackett KT, Dillard JP. 2005. Neisseria gonorrhoeae secretes chromosomal DNA via a novel type IV secretion system. Molecular Microbiology 55:1704–1721.

5. Pachulec E, Siewering K, Bender T, Heller E-M, Salgado-Pabon W, Schmoller SK, Woodhams KL, Dillard JP, van der Does C. 2014. Functional Analysis of the Gonococcal Genetic Island of Neisseria gonorrhoeae. PLOS ONE 9:e109613.

6. Pachulec E, van der Does C. 2010. Conjugative Plasmids of Neisseria gonorrhoeae. PLOS ONE 5:e9962.

7. Wu Z, Xu L, Tu Y, Chen R, Yu Y, Li J, Tan M, Chen H. 2011. The relationship between the symptoms of female gonococcal infections and serum progesterone level and the genotypes of Neisseria gonorrhoeae multi-antigen sequence type (NG-MAST) in Wuhan, China. Eur J Clin Microbiol Infect Dis 30:113–116.

8. Callaghan MM, Heilers J-H, van der Does C, Dillard JP. 2017. Secretion of Chromosomal DNA by the Neisseria gonorrhoeae Type IV Secretion System. Curr Top Microbiol Immunol 413:323–345.

9. Domínguez NM, Hackett KT, Dillard JP. 2011. XerCD-mediated site-specific recombination leads to loss of the 57-kilobase gonococcal genetic island. J Bacteriol 193:377–388.

10. BBMap. SourceForge.

11. Koren S, Treangen TJ, Hill CM, Pop M, Phillippy AM. 2014. Automated ensemble assembly and validation of microbial genomes. BMC Bioinformatics 15:126–126.

12. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477.

13. Zimin A, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome Assembler. Bioinformatics btt476.

14. Zerbino DR, Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829.

15. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics btu153.

16. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691–3693.

17. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421–421.

18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. Journal of Molecular Biology 215:403–410.

19. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microbial Genomics 2.

20. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

21. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. 2017. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol 8:28–36.

22. De Mita S, Siol M. 2012. EggLib: processing, analysis and simulation tools for population genetics and genomics. BMC Genetics 13:27.

23. Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, Corander J. 2012. Detection of recombination events in bacterial genomes from large population samples. Nucleic Acids Res 40:e6.

24. Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, Harris SR. 2017. Phandango: an interactive viewer for bacterial population genomics. Bioinformatics 34:292–293.

25. Huson DH, Bryant D. 2005. Application of Phylogenetic Networks in Evolutionary Studies. Molecular Biology and Evolution 23:254–267.

26. Lawson DJ, Hellenthal G, Myers S, Falush D. 2012. Inference of Population Structure using Dense Haplotype Data. PLOS Genetics 8:e1002453.

27. Yahara K, Didelot X, Ansari MA, Sheppard SK, Falush D. 2014. Efficient inference of recombination hot regions in bacterial genomes. Mol Biol Evol msu082.

28. Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403–1405.

29. Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20:289–290.

30. Parker J, Rambaut A, Pybus OG. 2008. Correlating viral phenotypes with phylogeny: Accounting for phylogenetic uncertainty. Infection, Genetics and Evolution 8:239–246.

31. Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.
32. Sagulenko P, Puller V, Neher RA. 2018. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evolution 4.
33. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. 2016. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biology 17:238.
34. Katoh K, Standley DM. 2014. MAFFT: iterative refinement and additional methods. Methods Mol Biol 1079:131–146.
35. Yahara K, Didelot X, Jolley KA, Kobayashi I, Maiden MCJ, Sheppard SK, Falush D. 2016. The Landscape of Realized Homologous Recombination in Pathogenic Bacteria. Molecular biology and evolution 33:456–471.
36. Alfsnes K, Frye SA, Eriksson J, Eldholm V, Brynildsrud OB, Bohlin J, Harrison OB, Hood DW, Maiden MCJ, Tønjum T, Ambur OH. 2018. A genomic view of experimental intraspecies and interspecies transformation of a rifampicin-resistance allele into Neisseria meningitidis. Microb Genom 4:e000222.
37. Croucher NJ, Harris SR, Barquist L, Parkhill J, Bentley SD. 2012. A high-resolution view of genome-wide pneumococcal transformation. PLoS Pathog 8:e1002745–e1002745.
38. Mortimer TD, Pepperell CS. 2014. Genomic Signatures of Distributive Conjugal Transfer among Mycobacteria. Genome Biology and Evolution 6:2489–2500.
39. Ezewudo MN, Joseph SJ, Castillo-Ramirez S, Dean D, Del Rio C, Didelot X, Dillon J-A, Selden RF, Shafer WM, Turingan RS, Unemo M, Read TD. 2015. Population structure of Neisseria gonorrhoeae based on whole genome data and its relationship with antibiotic resistance. PeerJ 3:e806–e806.
40. Carnoy C, Roten C-A. 2009. The dif/Xer Recombination Systems in Proteobacteria. PLOS ONE 4:e6531.
41. Huber KE, Waldor MK. 2002. Filamentous phage integration requires the host recombinases XerC and XerD. Nature 417:656–659.
42. Baltrus DA. 2013. Exploring the costs of horizontal gene transfer. Trends in Ecology & Evolution 28:489–495.
43. Touchon M, Bobay L-M, Rocha EP. 2014. The chromosomal accommodation and domestication of mobile genetic elements. Current Opinion in Microbiology 22:22–29.
44. Bouma JE, Lenski RE. 1988. Evolution of a bacteria/plasmid association. Nature 335:351–352.

45. De Gelder L, Williams JJ, Ponciano JM, Sota M, Top EM. 2008. Adaptive plasmid evolution results in host-range expansion of a broad-host-range plasmid. Genetics 178:2179–2190.
46. Zwanzig M, Harrison E, Brockhurst MA, Hall JPJ, Berendonk TU, Berger U. 2019. Mobile Compensatory Mutations Promote Plasmid Survival. mSystems 4:e00186-18.
47. Stalder T, Rogers LM, Renfrow C, Yano H, Smith Z, Top EM. 2017. Emerging patterns of plasmid-host coevolution that stabilize antibiotic resistance. Sci Rep 7:4853–4853.
48. Dahlberg C, Chao L. 2003. Amelioration of the cost of conjugative plasmid carriage in Eschericha coli K12. Genetics 165:1641–1649.
49. San Millan A, Toll-Riera M, Qi Q, MacLean RC. 2015. Interactions between horizontally acquired genes create a fitness cost in Pseudomonas aeruginosa. Nat Commun 6:6845–6845.
50. Snyder LAS, Jarvis SA, Saunders NJ. 2005. Complete and variant forms of the 'gonococcal genetic island' in Neisseria meningitidis. Microbiology 151:4005–4013.
51. Woodhams KL, Benet ZL, Blonsky SE, Hackett KT, Dillard JP. 2012. Prevalence and Detailed Mapping of the Gonococcal Genetic Island in Neisseria meningitidis. J Bacteriol 194:2275.
52. 2011. Neisseria bacilliformis ATCC BAA-1200 contig00002, whole genome shotgun sequence.
53. Corander J, Fraser C, Gutmann MU, Arnold B, Hanage WP, Bentley SD, Lipsitch M, Croucher NJ. 2017. Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. Nat Ecol Evol 1:1950–1960.
54. Rainey PB, Buckling A, Kassen R, Travisano M. 2000. The emergence and maintenance of diversity: insights from experimental bacterial populations. Trends in Ecology & Evolution 15:243–247.
55. Zola TA, Strange HR, Dominguez NM, Dillard JP, Cornelissen CN. 2010. Type IV secretion machinery promotes ton-independent intracellular survival of Neisseria gonorrhoeae within cervical epithelial cells. Infect Immun 78:2429–2437.
56. Zweig M, Schork S, Koerdt A, Siewering K, Sternberg C, Thormann K, Albers S-V, Molin S, van der Does C. 2014. Secreted single-stranded DNA is involved in the initial phase of biofilm formation by Neisseria gonorrhoeae. Environmental Microbiology 16:1040–1052.

Figures

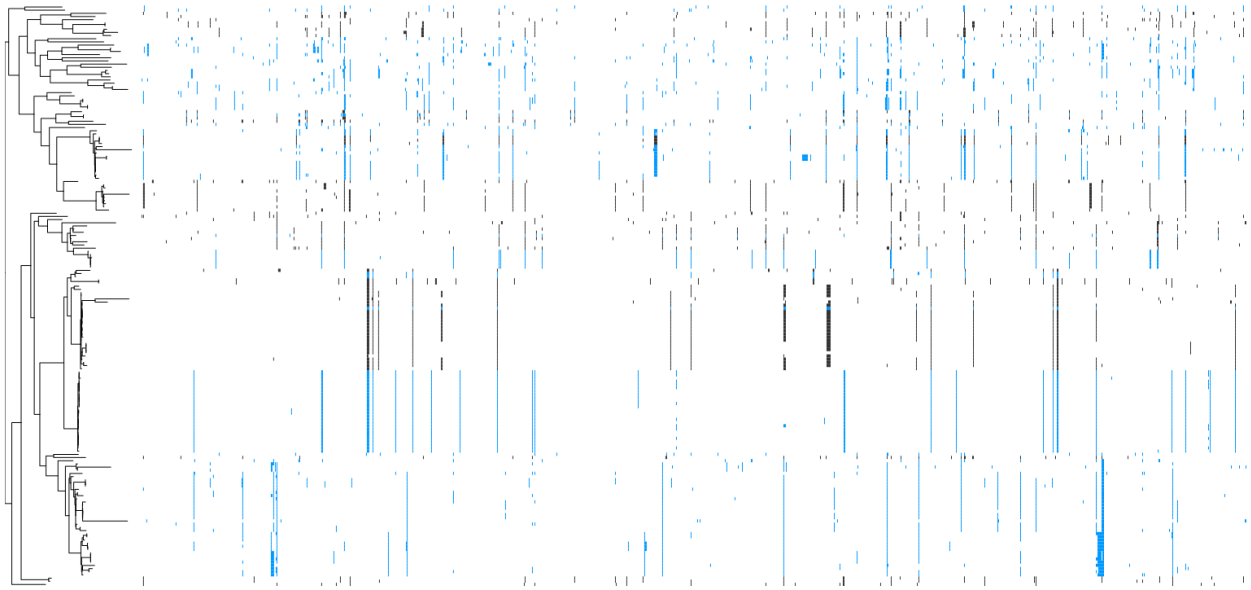

Figure 2.1. Distribution of recombinant fragments across *N. gonorrhoeae* core genome. Recombinant fragments identified by BRATNextGen are colored by GGI presence (blue) /absence (black) and ordered by location on core genome alignment (~1.2 Mb). Maximum likelihood phylogeny of the core genome alignment inferred by RAxML (left). Plot generated with Phandango (22).

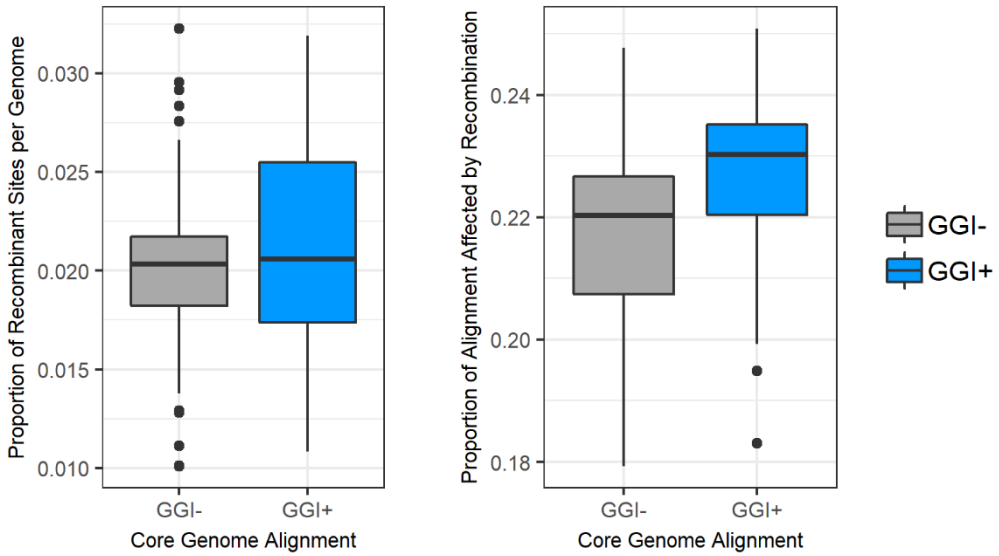

Figure 2.2. Proportion of recombinant sites per genome and proportion of alignment affected by recombination in GGI+ and GGI- isolates. Left: Boxplots of the proportion of recombinant sites per genome identified by the program BratNextGen in GGI+ and GGI- isolates. Right: Boxplots of the proportion of the alignment affected per group across 50 subsamples (*n* = 30 isolates per subsample). A greater proportion of the alignment has been affected by recombination in GGI+ isolates than GGI- isolates (medians = 0.23 and 0.21, respectively). The difference between means was statistically significant ($p < 0.05$). GGI+ shown in blue and GGI- shown in charcoal.

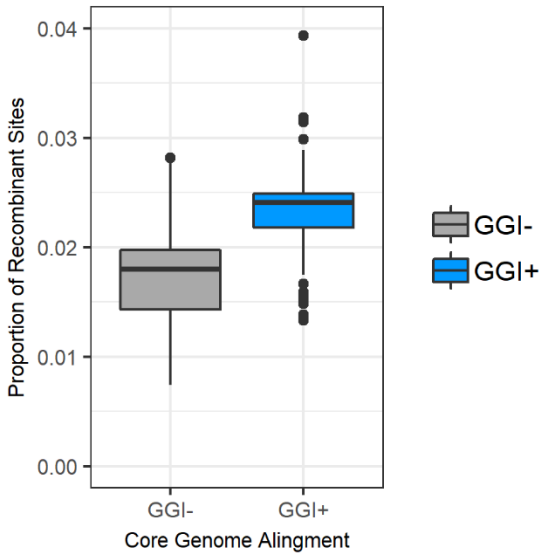

Figure 2.3. Proportion of recombinant sites per genome in individual core genomes of GGI+ and GGI- isolates. Boxplots of the proportion of recombinant sites per genome identified by the program BratNextGen in individual core genomes of GGI+ and GGI- isolates. A greater proportion of sites has been affected by recombination in GGI+ isolates than GGI- isolates (medians = 0.025 and 0.019, respectively). The difference between means was statistically significant ($p < 0.05$). GGI+ shown in blue and GGI- shown in charcoal.

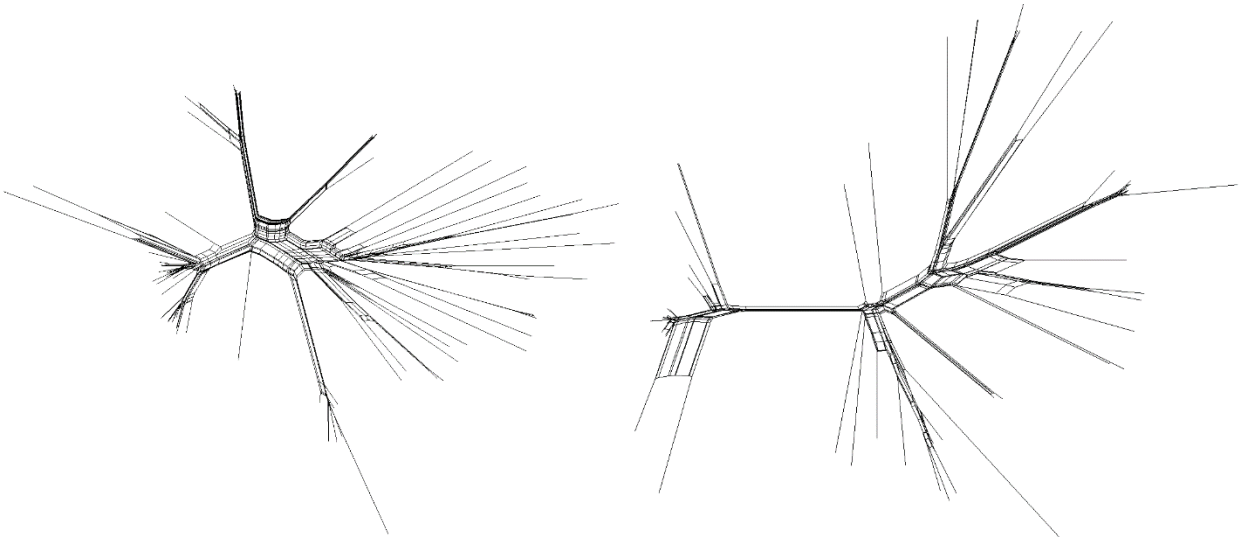

Figure 2.4. Core genome network of isolates with and without the GGI. Networks were created in SplitsTree4 from individual core genome alignments of GGI+ (left) and GGI- (right) isolates of *N. gonorrhoeae*. There are more reticulations in the network of GGI + isolates than in the network of GGI- isolates The PHI test was significant for both groups ($p < 0.05$), indicating there was evidence for recombination in both groups.

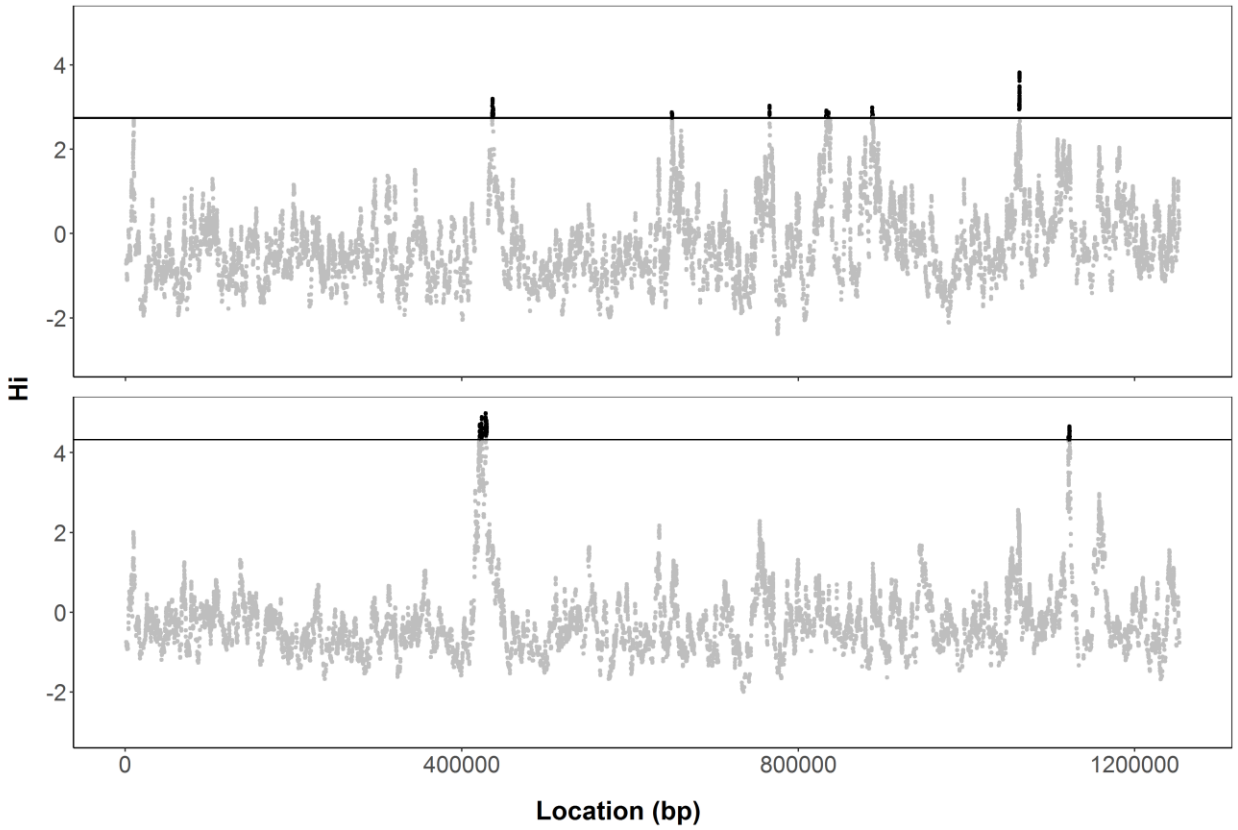

Figure 2.5. Intensity of recombination across *N. gonorrhoeae* genomes. The relative intensity of recombination per site across the core genome of GGI+ (top) and GGI- (bottom) isolates. H_i was calculated using OrderedPainting on a SNP alignment of *N. gonorrhoeae's* core genome. Hot spots (black) were defined as sites within the 99th percentile (black line) of H_i values.

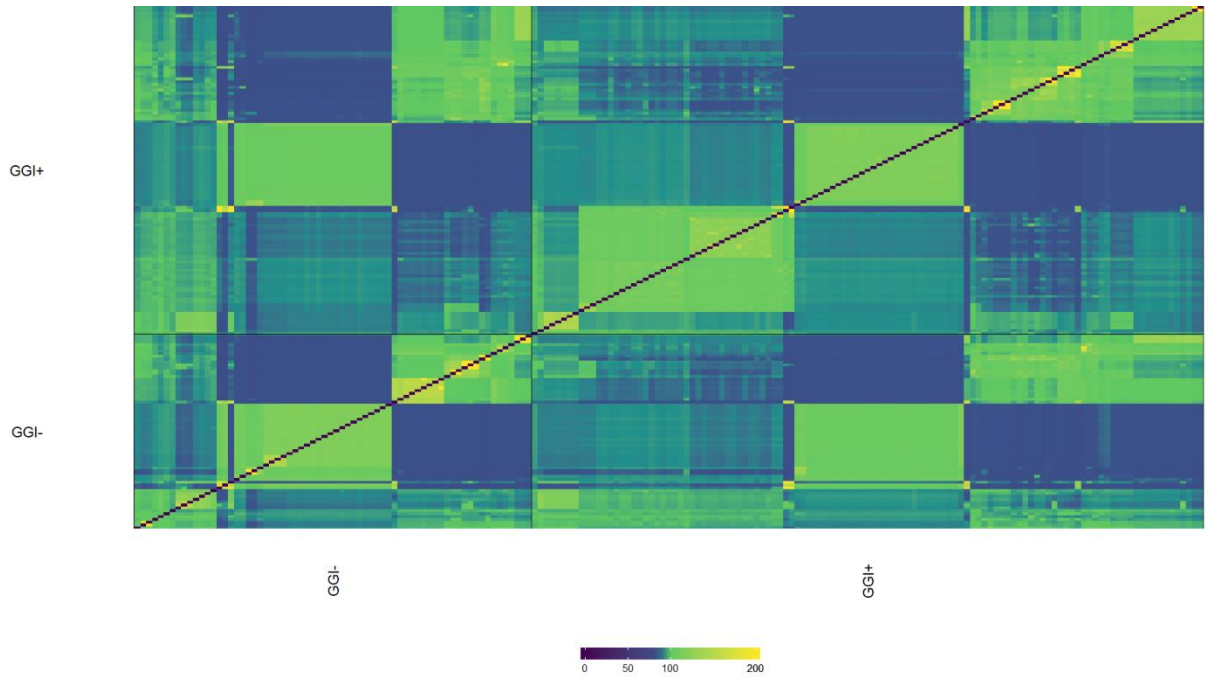

Figure 2.6. Co-ancestry matrix of GGI+ and GGI- populations. Co-ancestry matrix showing population structure and shared recombination chunks between donors (columns) and recipients (rows). Isolates divided by GGI+ and GGI- populations. Heatmap colors correspond to number of shared recombination chunks between isolates. There is no significant pattern of shared ancestry between sub-populations.

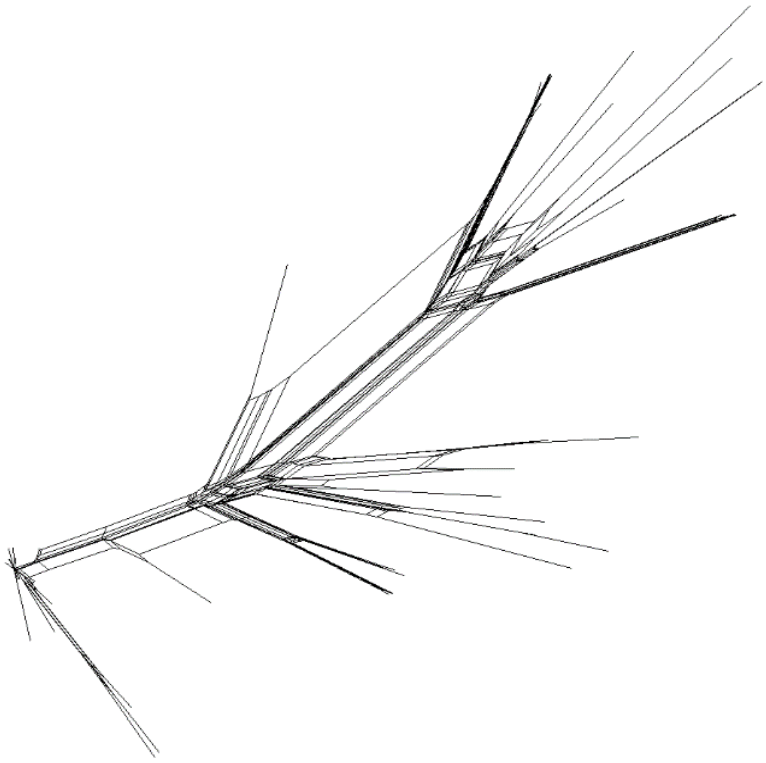

Figure 2.7. Network of the GGI. Network of core GGI genes from GGI+ isolates. The network was created in SplitsTree4 from a concatenated alignment of 48 genes present in ≥ 99% of GGI+ isolates. Results of the PHI test were significant ($p < 0.05$) for this alignment, indicating there was evidence of intra-locus recombination.

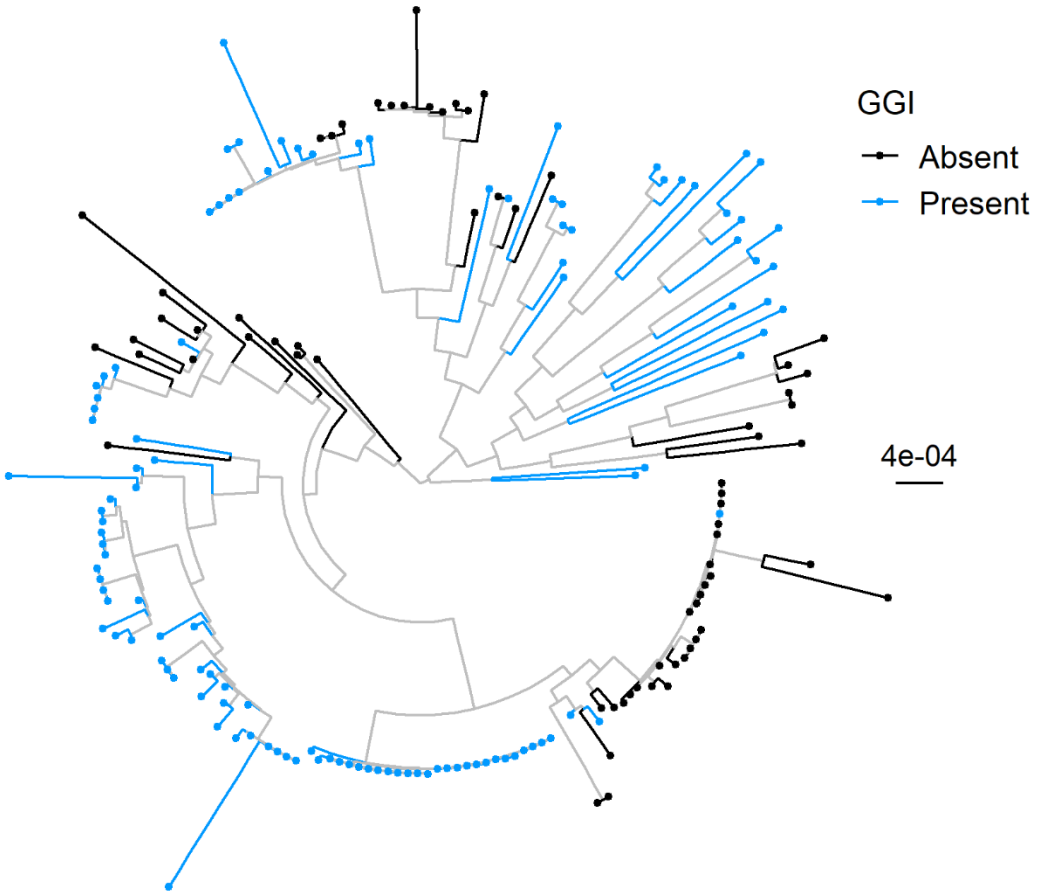

Figure 2.8. Maximum-likelihood phylogeny of *N. gonorrhoeae* isolates with presence/absence of the GGI indicated. RAxML was used for phylogenetic inference from the core genome alignment of our sample of *N. gonorrhoeae* (*n* = 183). The phylogeny is midpoint rooted. GGI+ isolates shown in blue, GGI- isolates shown in black. Tree scale shown on right. Isolates with and without the GGI tend to form monophyletic groups.

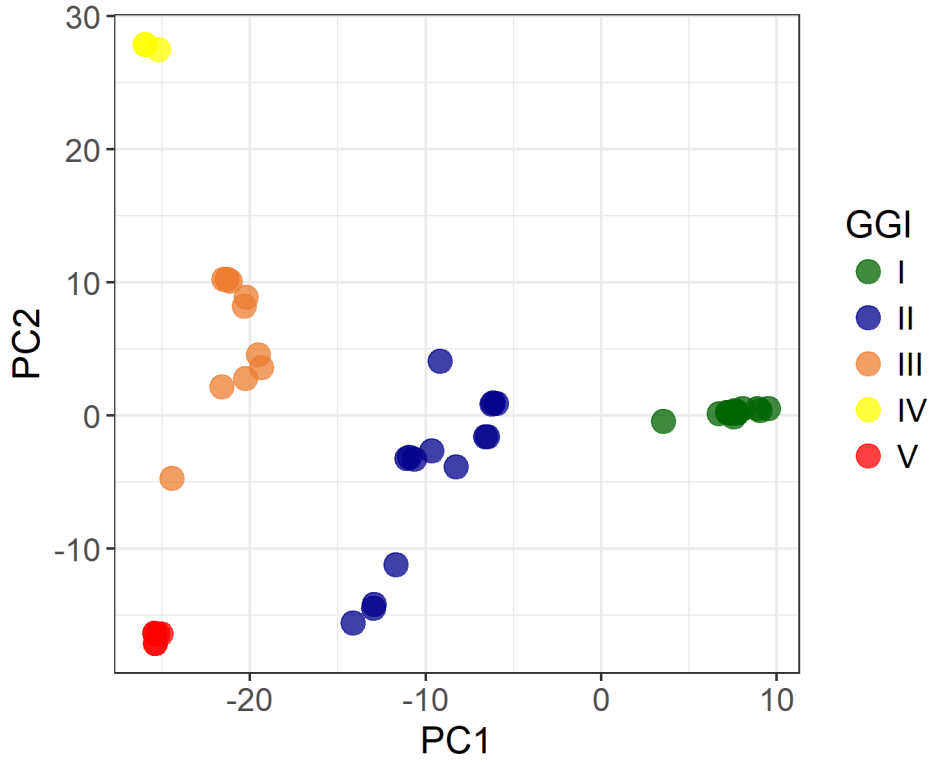

Figure 2.9. Principal component analysis of diversity within the core genome of the GGI. Principal component analysis performed with adegenet based on a concatenated alignment of 48 genes present in ≥ 99% of GGI+ isolates. The GGI clusters into five distinct groups (referred to as I-V). Type I GGI in green, type II in blue, type III in orange, type IV in yellow and type V in red.

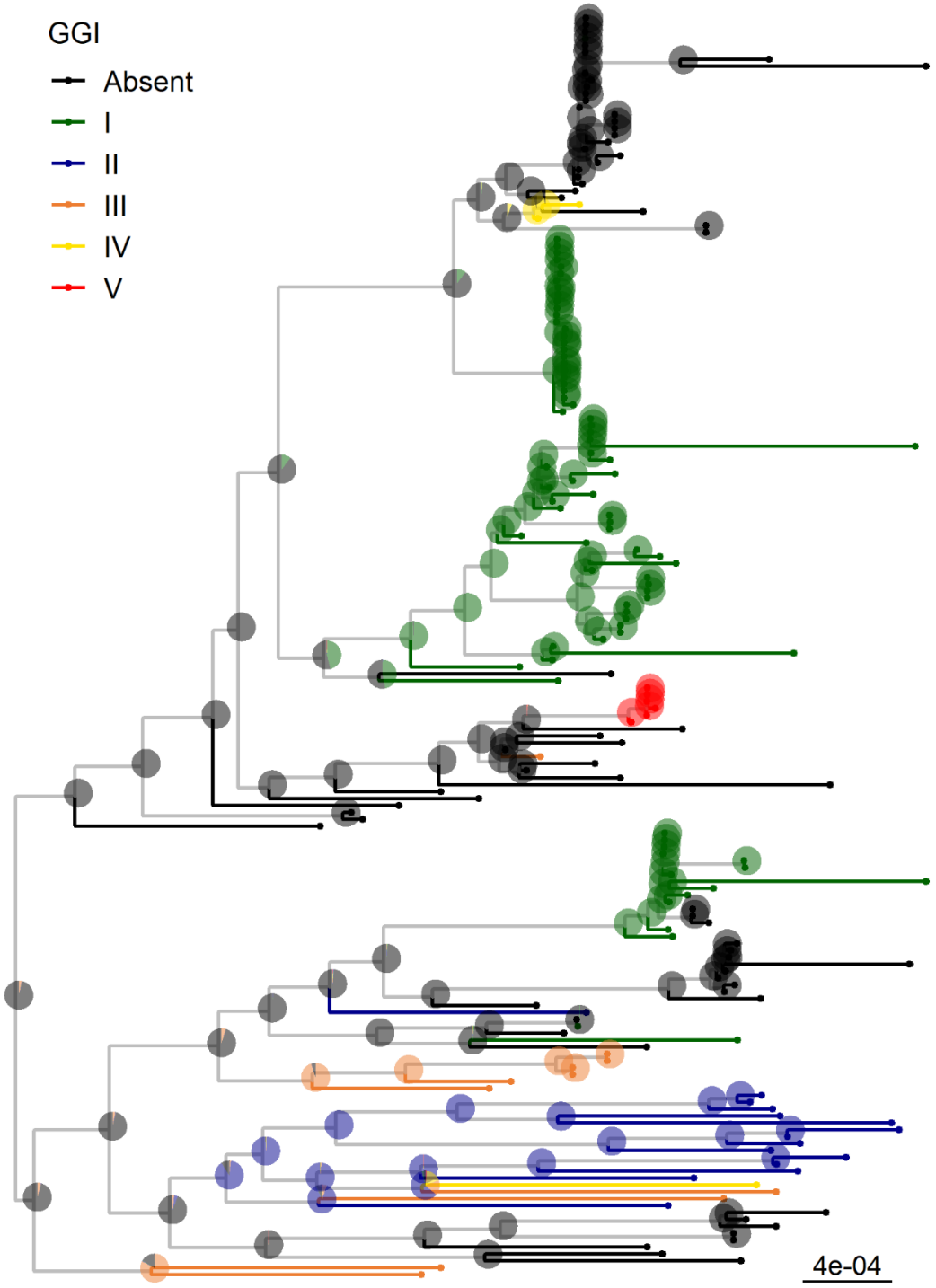

Figure 2.10. Reconstruction of the ancestral state of the GGI locus. Pie charts on nodes colored according to GGI state probabilities. Absence of the GGI shown in black, type I GGI in green, type II in blue, type III in orange, type IV in yellow and type V in red. Tree scale shown on bottom right. This analysis indicates that the GGI has been gained 13 times, switched 3 times and lost 2 times during the evolutionary history of the sample.

Tables

Table 2.1. Core genome analysis and global diversity estimates. Average gene number per isolate and gene content distribution within the core (found in ≥ 99% isolates), soft core (95% ≤ frequency < 99%), shell (15% ≤ frequency < 95%) and cloud genomes (frequency < 15%), as calculated by Roary. Two genes present in the GGI (*yee* and *yegB*) at core frequencies were not annotated by Prokka prior to calculating the pan-genome and are not included in this table. Global diversity estimates of the core genomes for each sample as calculated using egglib.

Sample	N	Average Gene Number	Core Genes	Soft Core Genes	Shell Genes	Cloud Genes	θ	π	Tajima's D
All	183	2081	1434	194	695	3943	2.9e-3	2.0e-3	-0.98
GGI+	115	2093	1644	114	529	1889	2.6e-3	2.0e-3	-0.74
GGI-	68	2066	1495	187	606	2877	2.5e-3	2.1e-3	-0.56
GGI	115	59	46	4	11	9	1.6e-3	1.1e-3	-1.18

Table 2.2. BaTS association statistics. Phylogeny-trait association statistics calculated using BaTS from a posterior distribution of 1500 trees inferred from a core genome alignment using MrBayes. GGI presence (by group) and absence were treated as traits. Statistics include the AI (Association Index), PS (Fitch parsimony score) and MC (maximum exclusive single-state clade size). Observed and null means, as well as the p-value reported for each statistic.

Statistic	Observed Mean	Null Mean	Significance
AI	1.9	13	0.0
PS	19	89	0.0
MC – Absent	10	3.0	<0.01
MC – Group I	37	3.6	<0.01
MC – Group II	8.6	1.3	<0.01
MC – Group III	6.0	1.2	<0.01
MC – Group IV	1.0	1.0	1.0
MC – Group V	6.0	1.0	<0.01

Table 2.3. Accessory gene content associated with the GGI. *N. gonorrhoeae* accessory genes associated with the GGI identified with Scoary. GGI presence and absence were treated as traits. Abundance in GGI+ vs GGI- isolates and Bonferroni corrected p-value reported for each gene.

Gene ID	BLAST Result	Annotation	GGI+	GGI-	Bonferroni Adjusted p-value
group_2619	MS11 locus tag 02312	Hypothetical protein	1	.74	1.8E-05
group_2027	MS11 locus tag 02004	Hypothetical protein	.70	1	4.73E-05
group_3127	NCCP11945 locus tag 1984	Hypothetical protein	.39	0	1.34E-07
group_3126	NCCP11945 locus tag 1985	Hypothetical protein	.33	0	8.43E-06

Supplementary Figures

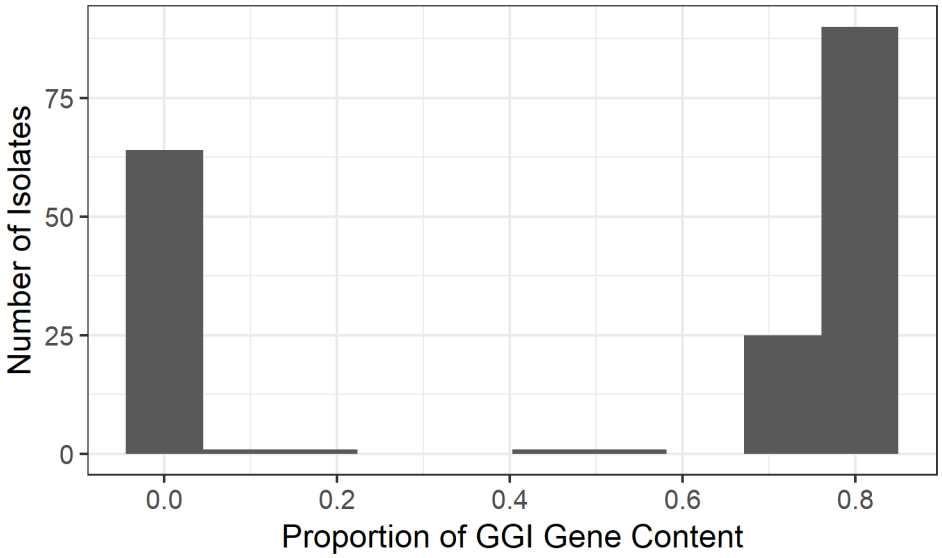

Figure 2.S1. Distribution of gene content within the bounds of the GGI. Histogram of the proportion of gene content within the bounds of the GGI encoded by each isolate in the sample (*n* = 183). The island's bounds, genes *yaa* and *parA,* were first described in laboratory strain MS11. Isolates can be clearly separated into a group encoding most (≥ 60% genes, *n* = 115) of the GGI and a group encoding a fragment of the GGI (≥ 10% and ≤ 60%, *n* = 4). A third group encodes <1% of the GGI (*n* = 64).

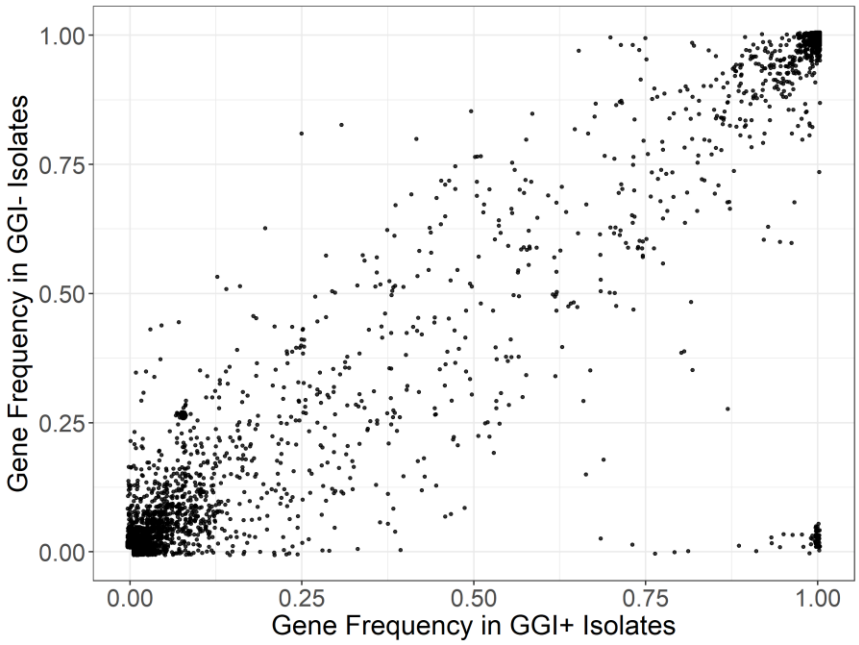

Figure 2.S2. Distribution of gene content in the pan-genome of *N. gonorrhoeae*. Scatterplot of gene frequency in the pan-genome of GGI+ and GGI- isolates. Pan-genome calculated using Roary. Each point corresponds to one gene in the pan-genome of 6266 genes.

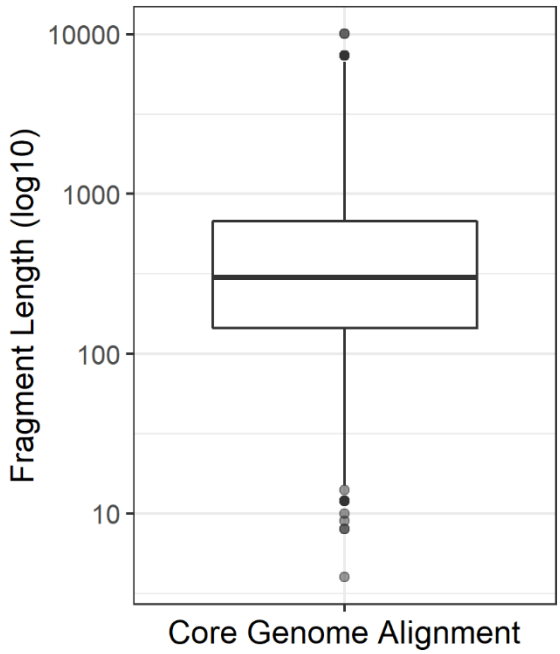

Figure 2.S3. Distribution of recombinant fragment lengths. Boxplot of recombinant fragment lengths (y-axis, log10 transformed) in *N. gonorrhoeae's* core genome identified by BratNextGen. Fragments range from 4-10008 bp in length. Median fragment length is 300 bp.

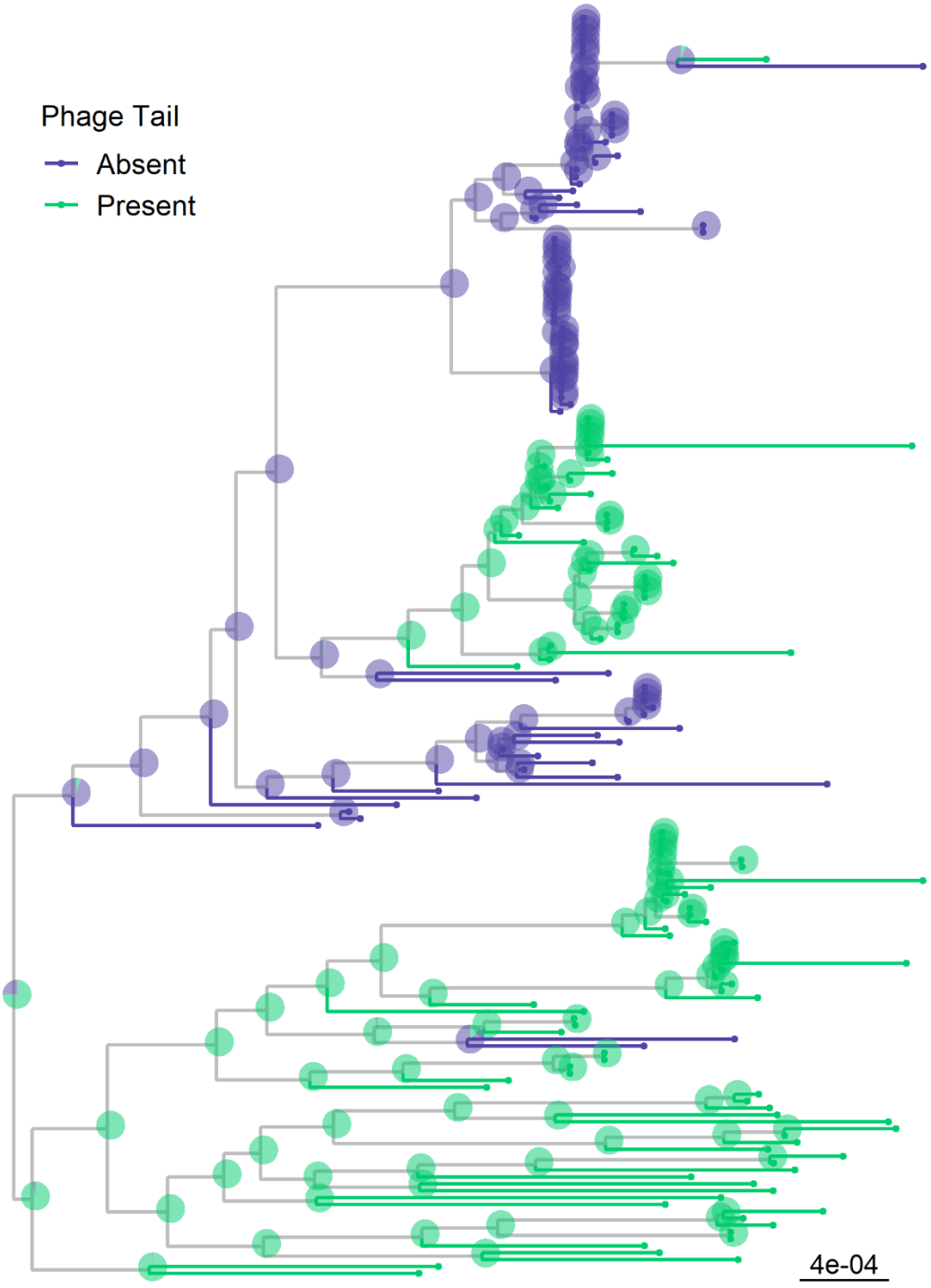

Figure 2.S4. Reconstruction of the ancestral state of the phage tail protein. Pie charts on branches indicate the proportions of log-likelihood for each state, with absence of the phage tail shown in purple and presence in green (*n*= 82 and *n* = 101, respectively). Tree scale shown on right. This analysis indicates that the phage tail protein has been gained 3 times and lost once in the evolutionary history of our sample.

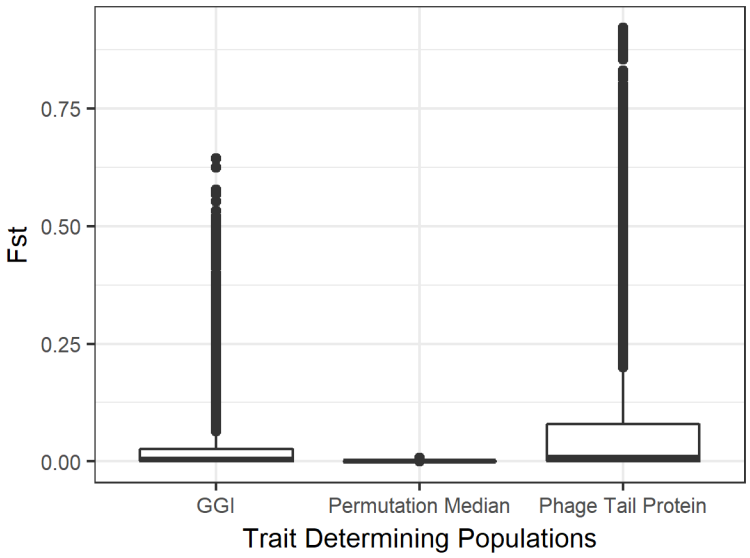

Figure 2.S5. Distribution of genome-wide F_{ST} values. Boxplots of the F_{ST} values calculated using the GGI and phage tail protein to define populations, as well as the median value calculated for each SNP in 100 permutations of the F_{ST} analysis (medians 0.004, 0.009 and 0.00, respectively).

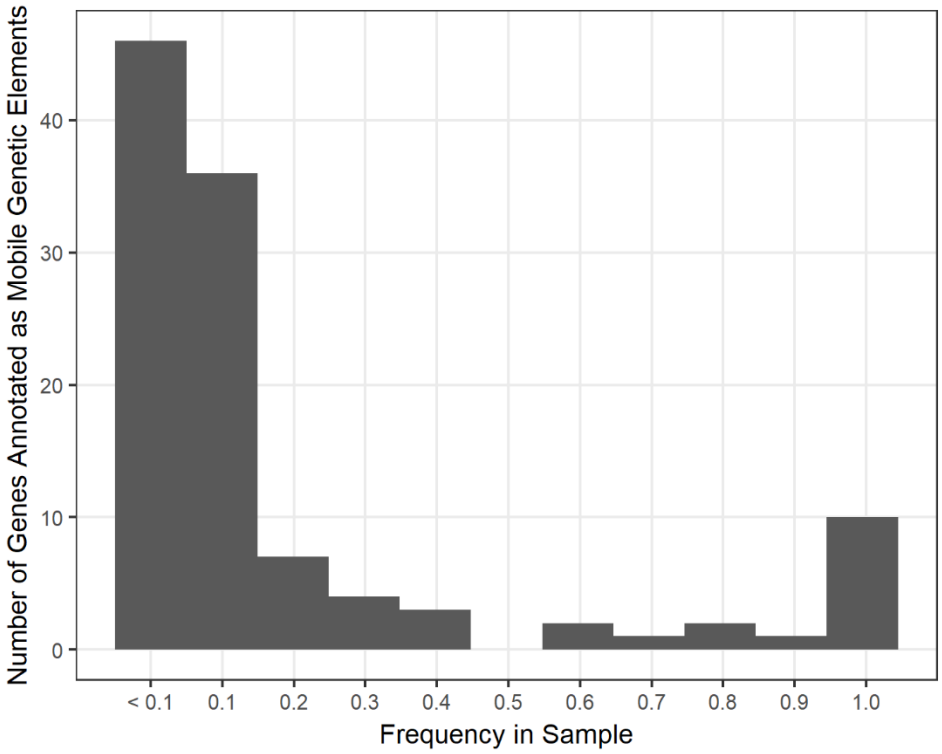

Figure 2.S6. Distribution of mobile genetic elements in the pan-genome. Histogram of mobile genetic elements (MGEs, *n* = 112) in the pan-genome of our *N. gonorrhoeae* sample (*n* = 183). MGEs are largely at core (≥ 99% of isolates) or rare accessory frequencies (< 10% of isolates).

Supplementary Tables

Table 2.S1 List of isolate accessions and GGI presence/absence

Accession	GGI	GGI Group
ERR222939	Present	I
ERR223670	Present	I
ERR191761	NA	Absent
ERR191789	NA	Absent
ERR191773	NA	Absent
SRR969336	Present	I
ERR223640	Present	I
ERR191807	Present	I
ERR223622	Present	I
ERR222905	Present	I
ERR191751	Present	I
ERR223696	Present	I
ERR223676	Present	I
ERR223674	Present	I
ERR191785	Present	I

ERR191755	Present	I
ERR222907	NA	Absent
ERR223618	Present	I
ERR355927	NA	Absent
ERR223660	NA	Absent
ERR223690	Present	I
ERR223656	Present	III
ERR223648	Present	III
ERR191742	Present	III
ERR355924	Present	III
ERR191744	Present	III
ERR355925	Present	III
F62	NA	Absent
fa_1090	NA	Absent
FA6140	NA	Absent
PID24_1	NA	Absent
DGI18	NA	Absent
SRR969045	NA	Absent
SRR969349	NA	Absent
SRR969337	NA	Absent
ERR222899	Present	II
SK931035	Present	III
FA19	Present	II
PID332	Present	II
SRR1248000	Present	II
PID1	Present	II
ERR191801	Present	III
SRR969345	Present	IV
ERR223668	Present	II
SRR960561	Present	II
SRR969341	Present	II
SRR969354	Present	II
ERR191779	Present	II
DGI2	Present	II
PID18	Present	II
ERR351670	Present	II
ERR494649	Present	II
3502	NA	Absent
ERR191803	NA	Absent
SRR959017	NA	Absent
ERR223678	NA	Absent
SRR969343	NA	Absent
SRR969348	NA	Absent

SRR969342	NA	Absent
SRR969340	Present	III
SRR969338	NA	Absent
SRR969344	NA	Absent
SRR969350	NA	Absent
ERR191793	NA	Absent
ERR191740	Present	V
ERR191741	Present	V
ERR223606	Present	V
ERR191745	Present	V
ERR191743	Present	V
ERR355926	Present	V
ERR191738	NA	Absent
ERR223612	Present	IV
SRR969351	Present	IV
ERR223650	NA	Absent
ERR191739	NA	Absent
ERR191821	NA	Absent
ERR191791	NA	Absent
ERR223624	NA	Absent
ERR222917	NA	Absent
ERR223682	NA	Absent
ERR191747	NA	Absent
ERR222925	NA	Absent
ERR222929	Present	I
ERR222923	NA	Absent
ERR222903	NA	Absent
ERR222897	NA	Absent
ERR223626	NA	Absent
ERR222901	NA	Absent
ERR223634	NA	Absent
ERR223644	NA	Absent
ERR223642	NA	Absent
ERR191767	NA	Absent
ERR223616	NA	Absent
ERR191825	NA	Absent
ERR223604	NA	Absent
ERR223662	NA	Absent
ERR223680	NA	Absent
ERR223698	NA	Absent
ERR222927	NA	Absent
ERR191812	NA	Absent
ERR191813	NA	Absent

ERR223632	NA	Absent
ERR222918	Present	I
ERR222916	Present	I
ERR223697	Present	I
ERR222896	Present	I
ERR222906	Present	I
ERR222898	Present	I
ERR222908	Present	I
ERR223679	Present	I
ERR223675	Present	I
ERR223681	Present	I
ERR223673	Present	I
ERR223677	Present	I
ERR223659	Present	I
ERR223667	Present	I
ERR191802	Present	I
ERR223655	Present	I
ERR222912	Present	I
ERR222892	Present	I
ERR222910	Present	I
ERR222922	Present	I
ERR222902	Present	I
ERR222900	Present	I
ERR222920	Present	I
ERR222924	Present	I
ERR222914	Present	I
ERR191800	Present	I
ERR191763	Present	I
ERR191811	NA	Absent
ERR223658	Present	I
ERR191764	Present	I
ERR223607	Present	I
ERR223641	Present	I
ERR191735	Present	I
NCCP11945	Present	I
ERR222893	Present	I
ERR222921	Present	I
ERR191799	Present	I
ERR191765	Present	I
ERR223664	Present	I
ERR223605	Present	I
ERR222937	Present	I
ERR223614	Present	I

ERR223636	Present	I
ERR222933	Present	I
ERR223646	Present	I
ERR223628	Present	I
ERR222931	Present	I
ERR222935	Present	I
ERR191787	Present	I
ERR191733	Present	I
ERR191731	Present	I
ERR223654	Present	I
ERR191815	Present	I
ERR222895	Present	I
ERR191809	Present	I
ERR191771	Present	I
ERR191783	Present	I
ERR222919	Present	I
ERR222913	Present	I
ERR223688	Present	I
ERR191749	Present	I
SRR969352	Present	I
ERR223638	Present	I
ERR191781	Present	I
ERR191823	Present	I
ERR191753	NA	Absent
ERR191769	NA	Absent
SK92679	NA	Absent
SRR955973	Present	III
SRR969353	Present	III
MS11	Present	II
ERR191737	NA	Absent
ERR223686	NA	Absent
ERR223692	NA	Absent
ERR223652	NA	Absent
ERR191819	NA	Absent
ERR222915	NA	Absent
ERR222911	NA	Absent
ERR222909	NA	Absent
ERR223694	NA	Absent
1291	NA	Absent

Table 2.S2 GGI genes in GGI+ isolates

Roary ID	MS11 Gene	# of Isolates	Within GGI Bounds
group_2093	Yaa	62	TRUE

group_3970	traD	115	TRUE
group_1664	traD	115	TRUE
group_2094	traD	88	TRUE
group_1665	traI	115	TRUE
group_1243	Yaf	115	TRUE
group_1666	ltgX	115	TRUE
group_4025	Yag	115	TRUE
group_4025	ltgX	115	TRUE
group_1667	traA	115	TRUE
group_4013	traL	115	TRUE
group_3796	traE	115	TRUE
group_1668	traK	115	TRUE
group_3503	traB	115	TRUE
dsbC_2	dsbC	115	TRUE
group_3144	traV	115	TRUE
group_3806	traC	115	TRUE
group_3999	Ybe	115	TRUE
group_3083	trbI	115	TRUE
group_3879	traW	115	TRUE
group_1244	traU	115	TRUE
group_3979	trbC	115	TRUE
group_2095	Ybi	114	TRUE
group_3516	traN	115	TRUE
group_2703	traF	1	TRUE
group_3202	Ycb	115	TRUE
group_2096	traF	114	TRUE
group_2097	traG	14	TRUE
group_1004	traH	16	TRUE
group_1003	traH	113	TRUE
group_1876	traG	93	TRUE
group_1669	atlA	107	TRUE
group_2098	Ych	107	TRUE
group_422	exp2	23	TRUE
group_422	exp1	23	TRUE
group_2224	cspA	79	TRUE
group_421	Yda	101	TRUE
group_1005	ydbA	115	TRUE
yhaV	ydbB	115	TRUE
prlF	ydcA	115	TRUE
group_1246	Ydd	2	TRUE
group_1672	ydcB	115	TRUE
group_1245	Ydd	114	TRUE
group_1673	ydeA	115	TRUE

group_2099	ydeB	114	TRUE
group_689	Ydf	105	TRUE
group_3335	Ydg	115	TRUE
group_3800	ydhA	115	TRUE
group_1674	ydhB	112	TRUE
group_3966	Ydi	115	TRUE
group_1006	Yea	102	TRUE
mshD	Yeb	115	TRUE
group_2365	yecB	1	TRUE
group_3962	yecA	115	TRUE
group_2222	yecB	84	TRUE
group_2481	yecB	6	TRUE
group_3518	yedA	115	TRUE
group_3852	yedB	115	TRUE
group_3855	yee/yegA	115	TRUE
group_3475	yegA	115	TRUE
group_4086	Yeh	115	TRUE
topB	topB	109	TRUE
group_3137	ssbB	115	TRUE
group_878	Yfa	111	TRUE
group_879	Yfa	5	TRUE
group_3957	Yfb	115	TRUE
group_2482	yfeA	1	TRUE
group_821	Yfd	111	TRUE
group_2101	yfeA	114	TRUE
group_3583	yfeB	115	TRUE
group_1675	parB	115	TRUE
parA	parA	115	TRUE
group_2223	cspA	1	FALSE
group_423	exp2	1	FALSE
group_2925	traN	1	FALSE
group_419	exp1	1	FALSE
group_419	exp2	1	FALSE
group_822	Yfd	19	FALSE
group_1009	topB	7	FALSE
group_2367	yecB	6	FALSE
group_2366	yecB	3	FALSE
group_2480	yecB	4	FALSE
group_1898	yecB	17	FALSE
group_2816	Yea	1	FALSE
group_2817	Yea	1	FALSE
group_1007	Yea	14	FALSE
group_690	Ydf	19	FALSE

group_1247	Ydd	10	FALSE
group_420	Yda	14	FALSE
group_2850	exp1	1	FALSE
group_2850	exp2	1	FALSE
group_1670	cspA	32	FALSE
group_2351	traG	8	FALSE
group_2494	Ybi	1	FALSE
NA	Yee	115	NA
NA	yegB	115	NA

GGI genes present within GGI+ isolates. Gene ID assigned by Roary, gene annotation in MS11 determined by BLAST, number of isolates, and whether the gene was within the bounds of the GGI (*yaa* and *parA*). Paralogs outside the bounds of the GGI, as well as *yee* and *yegB* (not annotated by Prokka), were identified using BLAST. Genes present in > 114 isolates were included in the core GGI alignment.

Table 2.S3. Genes with extreme H_i (≥ 99th percentile) in GGI+ and GGI- Isolates

Roary ID	Prokka Annotation	GGI
group_728	hypothetical protein	Absent
group_2448	hypothetical protein	Absent
group_2921	hypothetical protein	Absent
group_1944	hypothetical protein	Absent
group_2621	hypothetical protein	Absent
carA	Carbamoyl-phosphate synthase small chain	Absent
minD_1	Septum site-determining protein MinD	Absent
patA	Peptidoglycan O-acetyltransferase	Present
sotB_2	Sugar efflux transporter B	Present
sotB_1	Sugar efflux transporter	Present
merP	Mercuric transport protein periplasmic component precursor	Present
topA	DNA topoisomerase 1	Present
ftsY	Signal recognition particle receptor FtsY	Present
nudG	CTP pyrophosphohydrolase	Present
amiC	N-acetylmuramoyl-L-alanine amidase AmiC precursor	Present
tsaE	tRNA threonylcarbamoyladenosine biosynthesis protein TsaE	Present
glyS	Glycine--tRNA ligase beta subunit	Present
group_373	hypothetical protein	Present

Table 2.S4. Genes with extreme fold changes in GGI+ and GGI- Isolates

Gene	Annotation	GGI+ Hi	GGI- Hi	Fold Change	Adjusted p value
aroD	3-dehydroquinate dehydratase	-1.2	0.01	-89	7.18E-05
glnG	Nitrogen regulation protein NR(I)	1.28	0.01	214	9.84E-35
group_1317	hypothetical protein	-0.91	-0.01	137	2.57E-13
group_1758	hypothetical protein	1.52	0.01	137	2.76E-13
group_6026	NnrS protein	0.29	-0.0009	-302	5.06E-71

ispH	4-hydroxy-3-methylbut-2-enyl diphosphate reductase	0.44	0.002	178	1.42E-23
recN	DNA repair protein RecN	-1.39	0.02	-77	0.003
rhaS	HTH-type transcriptional activator RhaS	-0.96	0.01	-80	0.002
rplF	50S ribosomal protein L6	-1.12	-0.01	107	2.09E-07
secB	Protein-export protein SecB	0.58	-0.003	-233	6.27E-42
xerD	Tyrosine recombinase XerD	-1.38	0.02	-76	0.005

Table 2.S5. Homoplastic F_{ST} outliers

SNP	GGI+ Frequency	GGI- Frequency	FST	Roary ID	Prokka Annotation
7240	0.21	0	0.21	ygbN	Inner membrane permease YgbN
8998	0.27	0	0.27	thrB	Homoserine kinase
9016	0.57	0.18	0.28	thrB	Homoserine kinase
9237	0.49	0.16	0.22	ubiG	Ubiquinone biosynthesis O-methyltransferase
10068	0.38	0.72	0.21	mtr	Tryptophan-specific transport protein
10074	0.38	0.72	0.21	mtr	Tryptophan-specific transport protein
27637	0.49	0.82	0.23	ybaN	Inner membrane protein YbaN
68854	0.42	0.85	0.35	group_2522	hypothetical protein
68980	0.42	0.76	0.22	group_2522	hypothetical protein
68998	0.42	0.76	0.22	group_2522	hypothetical protein
69036	0.42	0.76	0.22	group_2522	hypothetical protein
69050	0.42	0.76	0.22	group_2522	hypothetical protein
69059	0.42	0.76	0.22	group_2522	hypothetical protein
69697	0.34	0.68	0.21	nadC	putative nicotinate-nucleotide pyrophosphorylase [carboxylating]
71319	0.12	0.43	0.21	cbbZC	Phosphoglycolate phosphatase, chromosomal
78699	0.15	0.54	0.3	ribD	Riboflavin biosynthesis protein RibD
78714	0.15	0.54	0.3	ribD	Riboflavin biosynthesis protein RibD
78720	0.15	0.54	0.3	ribD	Riboflavin biosynthesis protein RibD
78723	0.15	0.54	0.3	ribD	Riboflavin biosynthesis protein RibD
78738	0.15	0.54	0.3	ribD	Riboflavin biosynthesis protein RibD
78740	0.15	0.54	0.3	ribD	Riboflavin biosynthesis protein RibD
78990	0.32	0	0.33	group_5712	Glutamine amidotransferases class-II
81828	0.01	0.37	0.36	group_5727	Pilus assembly protein, PilO

82058	0.3	0.04	0.21	group_6255	Fimbrial assembly protein (PilN)
82183	0.45	0.79	0.22	group_6255	Fimbrial assembly protein (PilN)
93623	0.3	0.66	0.23	dacC_2	D-alanyl-D-alanine carboxypeptidase DacC precursor
119231	0.68	0.99	0.29	group_1464	hypothetical protein
123571	0.38	0.82	0.35	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
123577	0.38	0.82	0.35	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
123580	0.38	0.82	0.35	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
123583	0.38	0.82	0.35	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
123589	0.38	0.82	0.35	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
123598	0.38	0.82	0.35	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
123604	0.38	0.82	0.35	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
123631	0.38	0.82	0.35	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
123640	0.38	0.82	0.35	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
123646	0.38	0.82	0.35	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
123652	0.38	0.82	0.35	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
123653	0.38	0.82	0.35	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
123714	0.56	0.93	0.31	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
128223	0.5	0.85	0.25	trpA	Tryptophan synthase alpha chain
130411	0.58	0.91	0.26	Tmk	Thymidylate kinase
132960	0.68	0.97	0.26	group_5635	hypothetical protein
140580	0.01	0.32	0.31	trkH	Trk system potassium uptake protein TrkH

154881	0.53	0.85	0.22	lolA	Outer-membrane lipoprotein carrier protein precursor
155602	0.72	1	0.28	group_1250	hypothetical protein
155836	0.57	0.91	0.27	group_1250	hypothetical protein
155839	0.57	0.91	0.27	group_1250	hypothetical protein
155840	0.57	0.91	0.27	group_1250	hypothetical protein
155860	0.56	0.91	0.28	group_1250	hypothetical protein
155875	0.56	0.91	0.28	group_1250	hypothetical protein
157411	0.68	0.99	0.29	ppsR	Phosphoenolpyruvate synthase regulatory protein
165334	0.55	0.91	0.29	group_5506	preprotein translocase subunit SecD
165703	0.61	0.93	0.25	group_5506	preprotein translocase subunit SecD
168666	0.52	0.87	0.25	tatC	Sec-independent protein translocase protein TatC
169688	0.37	0.71	0.21	group_6170	hypothetical protein
179890	0.03	0.41	0.37	pgi_1	Glucose-6-phosphate isomerase
181235	0.03	0.41	0.37	pgi_1	Glucose-6-phosphate isomerase
181251	0.14	0.57	0.35	pgi_1	Glucose-6-phosphate isomerase
182257	0.29	0.9	0.57	ybbH	putative HTH-type transcriptional regulator YbbH
183103	0.17	0.85	0.64	Glk	Glucokinase
209785	0.41	0.75	0.22	serC	Phosphoserine aminotransferase
209818	0.62	0.94	0.27	serC	Phosphoserine aminotransferase
210377	0.57	0.87	0.2	serC	Phosphoserine aminotransferase
210770	0.57	0.91	0.26	serC	Phosphoserine aminotransferase
210773	0.57	0.91	0.26	serC	Phosphoserine aminotransferase
210781	0.57	0.9	0.24	serC	Phosphoserine aminotransferase
210785	0.57	0.9	0.24	serC	Phosphoserine aminotransferase
210787	0.57	0.9	0.24	serC	Phosphoserine aminotransferase
210791	0.57	0.9	0.24	serC	Phosphoserine aminotransferase
210792	0.57	0.9	0.24	serC	Phosphoserine aminotransferase
210800	0.57	0.9	0.24	serC	Phosphoserine aminotransferase
210803	0.57	0.9	0.24	serC	Phosphoserine aminotransferase
212919	0.38	0.96	0.55	group_5905	putative membrane protein
213001	0.54	0.88	0.26	group_5905	putative membrane protein
225745	0.48	0.81	0.22	dapA	4-hydroxy-tetrahydrodipicolinate synthase
227485	0.42	0.79	0.26	rlmB_1	23S rRNA (guanosine-2'-O-)-methyltransferase RlmB
227487	0.38	0.79	0.3	rlmB_1	23S rRNA (guanosine-2'-O-)-methyltransferase RlmB
227735	0.4	0.75	0.23	rlmB_1	23S rRNA (guanosine-2'-O-)-methyltransferase RlmB

242107	0.49	0.82	0.23	sdhB	Succinate dehydrogenase iron-sulfur subunit
274076	0.57	0.88	0.22	rlmJ	Ribosomal RNA large subunit methyltransferase J
274077	0.57	0.88	0.22	rlmJ	Ribosomal RNA large subunit methyltransferase J
274193	0.46	0.82	0.25	rlmJ	Ribosomal RNA large subunit methyltransferase J
274199	0.42	0.81	0.28	rlmJ	Ribosomal RNA large subunit methyltransferase J
274201	0.42	0.81	0.28	rlmJ	Ribosomal RNA large subunit methyltransferase J
274202	0.42	0.81	0.28	rlmJ	Ribosomal RNA large subunit methyltransferase J
274203	0.42	0.81	0.28	rlmJ	Ribosomal RNA large subunit methyltransferase J
274205	0.42	0.81	0.28	rlmJ	Ribosomal RNA large subunit methyltransferase J
277534	0.7	1	0.3	group_6143	hypothetical protein
279764	0	0.21	0.21	ftsK_2	DNA translocase FtsK
281966	0.67	0.94	0.21	proB	Glutamate 5-kinase 1
281992	0.68	0.99	0.29	proB	Glutamate 5-kinase 1
284837	0.64	0.96	0.27	group_5885	hypothetical protein
292403	0.47	0.13	0.24	group_5730	Putative lipoprotein/NMB1164 precursor
301186	0.56	0.88	0.24	accA	Acetyl-coenzyme A carboxylase carboxyl transferase subunit alpha
312618	0.06	0.38	0.27	cysJ_2	Sulfite reductase [NADPH] flavoprotein alpha-component
313081	0.74	1	0.27	cysNC	Bifunctional enzyme CysN/CysC
313687	0.06	0.44	0.33	cysNC	Bifunctional enzyme CysN/CysC
313919	0.01	0.34	0.33	cysNC	Bifunctional enzyme CysN/CysC
319624	0.29	0.62	0.2	Rnr	Ribonuclease R
333066	0.63	0.97	0.32	Lon	Lon protease
344448	0.69	0.99	0.29	Cca	Multifunctional CCA protein
353929	0.47	0.88	0.33	group_6177	hypothetical protein
354098	0.46	0.85	0.3	group_6177	hypothetical protein
354101	0.46	0.85	0.3	group_6177	hypothetical protein
354102	0.46	0.85	0.3	group_6177	hypothetical protein
354104	0.46	0.85	0.3	group_6177	hypothetical protein
354109	0.46	0.85	0.3	group_6177	hypothetical protein
355190	0.46	0.85	0.3	group_5434	Putative O-methyltransferase/MSMEI_4947
355193	0.46	0.85	0.3	group_5434	Putative O-methyltransferase/MSMEI_4947
365148	0.63	0.99	0.34	spoT	Bifunctional (p)ppGpp synthase/hydrolase SpoT

365339	0.5	0.93	0.38	spoT	Bifunctional (p)ppGpp synthase/hydrolase SpoT
365519	0.51	0.94	0.38	spoT	Bifunctional (p)ppGpp synthase/hydrolase SpoT
372666	0.01	0.37	0.36	hgpA	Hemoglobin and hemoglobin-haptoglobin-binding protein A precursor
395445	0.23	0	0.24	group_5868	tetratricopeptide repeat protein
395589	0.37	0.04	0.28	group_5868	tetratricopeptide repeat protein
402820	0.66	0.94	0.22	aniA	Copper-containing nitrite reductase precursor
402821	0.66	0.94	0.22	aniA	Copper-containing nitrite reductase precursor
402824	0.66	0.94	0.22	aniA	Copper-containing nitrite reductase precursor
402827	0.66	0.94	0.22	aniA	Copper-containing nitrite reductase precursor
403526	0.26	0.59	0.2	aniA	Copper-containing nitrite reductase precursor
441438	0.01	0.34	0.33	ttcA	tRNA 2-thiocytidine biosynthesis protein TtcA
441528	0.24	0.63	0.27	ttcA	tRNA 2-thiocytidine biosynthesis protein TtcA
441531	0.24	0.63	0.27	ttcA	tRNA 2-thiocytidine biosynthesis protein TtcA
442744	0.4	0.78	0.26	ygfZ	tRNA-modifying protein YgfZ
443711	0.2	0.53	0.21	dnaJ_1	Chaperone protein DnaJ
466948	0.27	0.68	0.29	uvrC	UvrABC system protein C
468861	0.25	0.6	0.23	copA_1	Copper-exporting P-type ATPase A
469041	0.49	0.85	0.27	copA_1	Copper-exporting P-type ATPase A
473147	0.68	0.96	0.23	group_1567	hypothetical protein
478636	0.2	0.62	0.31	ispG	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase
483399	0.57	1	0.45	group_5865	S-formylglutathione hydrolase
484826	0.56	0.99	0.42	frmA_1	S-(hydroxymethyl)glutathione dehydrogenase
485952	0.64	1	0.37	ihfB	Integration host factor subunit beta
508069	0.03	0.38	0.32	iscU	NifU-like protein
509021	0.01	0.34	0.33	iscS	Cysteine desulfurase
520112	0.48	0.87	0.3	nadE	NH(3)-dependent NAD( ) synthetase
531123	0.49	0.82	0.23	Sad	Succinate semialdehyde dehydrogenase [NAD(P)] Sad
531509	0.6	0.93	0.26	Sad	Succinate semialdehyde dehydrogenase [NAD(P)] Sad

532117	0.39	0.74	0.22	group_6064	hypothetical protein
535613	0.58	0.91	0.26	suhB_1	Inositol-1-monophosphatase
535753	0.12	0.54	0.34	suhB_1	Inositol-1-monophosphatase
536725	0.54	0.97	0.41	hrpB	ATP-dependent RNA helicase HrpB
537867	0.59	1	0.42	hrpB	ATP-dependent RNA helicase HrpB
537947	0.62	1	0.39	hrpB	ATP-dependent RNA helicase HrpB
537948	0.63	1	0.38	hrpB	ATP-dependent RNA helicase HrpB
538869	0.54	0.87	0.23	argJ	Arginine biosynthesis bifunctional protein ArgJ
550647	0.57	0.9	0.25	parC	DNA topoisomerase 4 subunit A
553613	0.02	0.35	0.32	alaS	Alanine--tRNA ligase
555051	0.08	0.51	0.38	alaS	Alanine--tRNA ligase
555224	0.3	0.72	0.3	alaS	Alanine--tRNA ligase
555239	0.12	0.5	0.29	alaS	Alanine--tRNA ligase
555254	0.06	0.34	0.22	alaS	Alanine--tRNA ligase
555267	0.06	0.37	0.25	alaS	Alanine--tRNA ligase
555269	0.06	0.37	0.25	alaS	Alanine--tRNA ligase
555272	0.06	0.37	0.25	alaS	Alanine--tRNA ligase
556646	0.43	0.9	0.4	potF_3	Putrescine-binding periplasmic protein precursor
558155	0.03	0.35	0.28	group_5768	EamA-like transporter family protein
570618	0.07	0.4	0.27	ilvI	Acetolactate synthase isozyme 3 large subunit
571332	0.09	0.49	0.33	ilvI	Acetolactate synthase isozyme 3 large subunit
571749	0.01	0.34	0.33	ilvH	Acetolactate synthase isozyme 3 small subunit
596256	0.08	0.46	0.32	uvrA	UvrABC system protein A
596301	0.2	0.53	0.21	uvrA	UvrABC system protein A
596348	0.2	0.66	0.37	uvrA	UvrABC system protein A
596370	0.2	0.54	0.23	uvrA	UvrABC system protein A
596397	0.2	0.54	0.23	uvrA	UvrABC system protein A
596415	0.2	0.54	0.23	uvrA	UvrABC system protein A
596640	0.12	0.44	0.23	uvrA	UvrABC system protein A
596967	0.01	0.38	0.37	uvrA	UvrABC system protein A
596970	0.01	0.38	0.37	uvrA	UvrABC system protein A
597282	0.06	0.4	0.28	uvrA	UvrABC system protein A
597283	0.06	0.4	0.28	uvrA	UvrABC system protein A
597291	0.06	0.4	0.28	uvrA	UvrABC system protein A
597294	0.06	0.4	0.28	uvrA	UvrABC system protein A
597295	0.06	0.4	0.28	uvrA	UvrABC system protein A

605685	0.42	0.84	0.33	group_6085	Cytochrome c-555 precursor
616065	0.46	0.82	0.25	yjjV	putative deoxyribonuclease YjjV
618150	0.61	0.93	0.25	group_5652	hypothetical protein
623680	0.01	0.37	0.36	group_1579	hypothetical protein
631915	0.01	0.32	0.31	oprM	Outer membrane protein OprM precursor
631979	0.62	0.96	0.3	oprM	Outer membrane protein OprM precursor
632629	0.68	0.97	0.26	oprM	Outer membrane protein OprM precursor
633218	0.65	0.96	0.26	oprM	Outer membrane protein OprM precursor
633418	0.2	0.62	0.31	group_2485	hypothetical protein
633422	0.2	0.62	0.31	group_2485	hypothetical protein
633424	0.2	0.62	0.31	group_2485	hypothetical protein
634301	0.3	0.74	0.33	abgT	p-aminobenzoyl-glutamate transport protein
634435	0.28	0.72	0.33	abgT	p-aminobenzoyl-glutamate transport protein
634867	0.35	0.75	0.29	abgT	p-aminobenzoyl-glutamate transport protein
634870	0.35	0.75	0.29	abgT	p-aminobenzoyl-glutamate transport protein
634930	0.32	0.69	0.24	abgT	p-aminobenzoyl-glutamate transport protein
634936	0.32	0.69	0.24	abgT	p-aminobenzoyl-glutamate transport protein
634939	0.32	0.69	0.24	abgT	p-aminobenzoyl-glutamate transport protein
634996	0.27	0.79	0.44	abgT	p-aminobenzoyl-glutamate transport protein
635005	0.27	0.84	0.5	abgT	p-aminobenzoyl-glutamate transport protein
635011	0.27	0.84	0.5	abgT	p-aminobenzoyl-glutamate transport protein
635014	0.27	0.84	0.5	abgT	p-aminobenzoyl-glutamate transport protein
635020	0.27	0.84	0.5	abgT	p-aminobenzoyl-glutamate transport protein
635022	0.27	0.84	0.5	abgT	p-aminobenzoyl-glutamate transport protein
635023	0.27	0.84	0.5	abgT	p-aminobenzoyl-glutamate transport protein
643413	0.38	0.72	0.21	Def	Peptide deformylase
654155	0.08	0.51	0.38	Ffh	Signal recognition particle protein
655043	0.13	0.49	0.26	group_1149	hypothetical protein
655410	0.07	0.47	0.35	yphA	Inner membrane protein YphA
660148	0.53	0.18	0.25	ygdH	LOG family protein YgdH

660154	0.53	0.18	0.25	ygdH	LOG family protein YgdH
669416	0.59	0.91	0.25	recQ	ATP-dependent DNA helicase RecQ
671479	0.32	0.75	0.32	group_6187	hypothetical protein
673047	0.64	0.97	0.3	group_2577	RlpA-like protein precursor
690336	0.13	0.49	0.26	ndhC	NAD(P)H-quinone oxidoreductase subunit 3
690588	0.03	0.38	0.33	ndhC	NAD(P)H-quinone oxidoreductase subunit 3
694104	0.7	0.96	0.2	gabD	Succinate-semialdehyde dehydrogenase [NADP( )] GabD
694465	0.41	0.76	0.23	group_5401	hypothetical protein
701262	0.03	0.32	0.25	group_6035	Nickel uptake substrate-specific transmembrane region
702422	0.03	0.34	0.27	alaA	Glutamate-pyruvate aminotransferase AlaA
709115	0.31	0.76	0.35	Rph	Ribonuclease PH
710719	0.37	0.72	0.23	scpA	Segregation and condensation protein A
710725	0.37	0.72	0.23	scpA	Segregation and condensation protein A
710730	0.37	0.72	0.23	scpA	Segregation and condensation protein A
710739	0.37	0.72	0.23	scpA	Segregation and condensation protein A
710742	0.37	0.72	0.23	scpA	Segregation and condensation protein A
710772	0.06	0.37	0.25	scpA	Segregation and condensation protein A
710775	0.06	0.37	0.25	scpA	Segregation and condensation protein A
710781	0.06	0.37	0.25	scpA	Segregation and condensation protein A
710787	0.06	0.37	0.25	scpA	Segregation and condensation protein A
710793	0.06	0.37	0.25	scpA	Segregation and condensation protein A
710796	0.06	0.37	0.25	scpA	Segregation and condensation protein A
710799	0.06	0.37	0.25	scpA	Segregation and condensation protein A
711618	0.19	0.6	0.31	pncB2	Nicotinate phosphoribosyltransferase 2
712776	0.3	0.68	0.25	argS	Arginine--tRNA ligase
712809	0.3	0.71	0.28	argS	Arginine--tRNA ligase
713022	0.23	0.69	0.37	argS	Arginine--tRNA ligase
713034	0.23	0.69	0.37	argS	Arginine--tRNA ligase
713040	0.23	0.69	0.37	argS	Arginine--tRNA ligase

713076	0.27	0.71	0.33	argS	Arginine--tRNA ligase
713088	0.27	0.69	0.31	argS	Arginine--tRNA ligase
713097	0.27	0.69	0.31	argS	Arginine--tRNA ligase
713133	0.28	0.69	0.3	argS	Arginine--tRNA ligase
713139	0.28	0.69	0.3	argS	Arginine--tRNA ligase
713184	0.33	0.68	0.22	argS	Arginine--tRNA ligase
713205	0.33	0.68	0.22	argS	Arginine--tRNA ligase
713301	0.08	0.44	0.3	argS	Arginine--tRNA ligase
713559	0.05	0.38	0.28	argS	Arginine--tRNA ligase
713622	0.26	0.71	0.34	argS	Arginine--tRNA ligase
715347	0.54	0.88	0.26	rpiA	Ribose-5-phosphate isomerase A
725453	0.03	0.26	0.21	group_6164	putative FAD-linked oxidoreductase
735176	0.71	1	0.29	dnaG	DNA primase
749431	0.07	0.38	0.25	fumC	Fumarate hydratase class II
750589	0.19	0.75	0.49	yhbE	putative inner membrane transporter YhbE
752398	0.07	0.38	0.25	group_2350	murein transglycosylase C
752692	0.05	0.54	0.46	group_2350	murein transglycosylase C
752905	0.06	0.34	0.22	group_321	Transposase DDE domain protein
752911	0.06	0.34	0.22	group_321	Transposase DDE domain protein
753760	0.29	0	0.29	metG	Methionine--tRNA ligase
754269	0.33	0.68	0.22	metG	Methionine--tRNA ligase
754694	0.3	0.01	0.27	metG	Methionine--tRNA ligase
754698	0.3	0.01	0.27	metG	Methionine--tRNA ligase
755128	0.52	0.87	0.25	metG	Methionine--tRNA ligase
755129	0.53	0.87	0.24	metG	Methionine--tRNA ligase
755602	0.04	0.35	0.27	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
755611	0.04	0.35	0.27	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
755612	0.04	0.35	0.27	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
755614	0.04	0.35	0.27	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
755620	0.04	0.35	0.27	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
755623	0.04	0.35	0.27	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
755638	0.04	0.34	0.25	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
755639	0.04	0.34	0.25	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
755640	0.04	0.34	0.25	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]

755642	0.04	0.34	0.25	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
755644	0.04	0.34	0.25	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
755646	0.04	0.34	0.25	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
755651	0.04	0.34	0.25	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
755653	0.04	0.34	0.25	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
755654	0.04	0.35	0.27	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
755656	0.04	0.34	0.25	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
755665	0.04	0.34	0.25	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
755674	0.04	0.34	0.25	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
755719	0.04	0.34	0.25	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
755722	0.04	0.34	0.25	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
755723	0.04	0.34	0.25	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
755724	0.04	0.34	0.25	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
755731	0.04	0.34	0.25	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
755740	0.17	0.51	0.24	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
755743	0.04	0.34	0.25	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
757737	0.53	0.84	0.2	mltA	Membrane-bound lytic murein transglycosylase A precursor
760331	0.11	0.57	0.39	glmU	Bifunctional protein GlmU
760579	0.3	0.69	0.27	glmU	Bifunctional protein GlmU
760615	0.35	0.91	0.52	glmU	Bifunctional protein GlmU
765753	0.53	0.84	0.2	ftsY	Signal recognition particle receptor FtsY
765988	0.36	0.69	0.2	ftsY	Signal recognition particle receptor FtsY
766238	0.37	0.71	0.21	ftsY	Signal recognition particle receptor FtsY
768228	0.09	0.47	0.32	ileS	Isoleucine--tRNA ligase
768291	0.47	0.87	0.31	ileS	Isoleucine--tRNA ligase
768369	0.04	0.43	0.35	ileS	Isoleucine--tRNA ligase
768370	0.04	0.43	0.35	ileS	Isoleucine--tRNA ligase
768375	0.04	0.43	0.35	ileS	Isoleucine--tRNA ligase

768384	0.04	0.43	0.35	ileS	Isoleucine--tRNA ligase
768715	0.15	0.69	0.48	ileS	Isoleucine--tRNA ligase
770085	0.21	0.53	0.2	ileS	Isoleucine--tRNA ligase
772983	0.63	0.96	0.29	nqrB	Na( )-translocating NADH-quinone reductase subunit B
787096	0.15	0.62	0.39	cmpB	Bicarbonate transport system permease protein CmpB
787188	0.05	0.43	0.33	ssuB	Aliphatic sulfonates import ATP-binding protein SsuB
787217	0.12	0.44	0.23	ssuB	Aliphatic sulfonates import ATP-binding protein SsuB
787271	0.07	0.57	0.46	ssuB	Aliphatic sulfonates import ATP-binding protein SsuB
788960	0.03	0.4	0.33	dsbC_1	putative thiol:disulfide interchange protein DsbC precursor
792106	0.42	0.79	0.26	Smc	Chromosome partition protein Smc
796614	0.62	0.97	0.33	lysS	Lysine--tRNA ligase
796617	0.62	0.97	0.33	lysS	Lysine--tRNA ligase
796631	0.46	0.81	0.23	lysS	Lysine--tRNA ligase
796665	0.63	0.99	0.35	lysS	Lysine--tRNA ligase
798160	0.44	0.88	0.36	group_1984	Natural resistance-associated macrophage protein
798161	0.44	0.88	0.36	group_1984	Natural resistance-associated macrophage protein
798162	0.44	0.88	0.36	group_1984	Natural resistance-associated macrophage protein
798226	0.45	0.88	0.35	group_1984	Natural resistance-associated macrophage protein
798231	0.45	0.88	0.35	group_1984	Natural resistance-associated macrophage protein
798240	0.45	0.88	0.35	group_1984	Natural resistance-associated macrophage protein
798243	0.45	0.88	0.35	group_1984	Natural resistance-associated macrophage protein
798258	0.57	0.94	0.33	group_1984	Natural resistance-associated macrophage protein
798276	0.63	0.96	0.28	group_1984	Natural resistance-associated macrophage protein
798288	0.63	0.96	0.28	group_1984	Natural resistance-associated macrophage protein
798302	0.63	0.96	0.28	group_1984	Natural resistance-associated macrophage protein
798408	0.63	0.96	0.28	group_1984	Natural resistance-associated macrophage protein
798732	0.43	0.9	0.4	group_1984	Natural resistance-associated macrophage protein
798735	0.43	0.9	0.4	group_1984	Natural resistance-associated macrophage protein

798738	0.43	0.9	0.4	group_1984	Natural resistance-associated macrophage protein
798784	0.43	0.88	0.37	group_1984	Natural resistance-associated macrophage protein
798855	0.44	0.9	0.39	group_1984	Natural resistance-associated macrophage protein
798863	0.43	0.88	0.37	group_1984	Natural resistance-associated macrophage protein
798871	0.43	0.87	0.35	group_1984	Natural resistance-associated macrophage protein
798873	0.43	0.87	0.35	group_1984	Natural resistance-associated macrophage protein
800210	0.22	0.6	0.27	comM	Competence protein ComM
801354	0.34	0.94	0.58	group_5630	cell division protein FtsN
802304	0.62	0.94	0.27	dsbA_1	Thiol:disulfide interchange protein DsbA precursor
802973	0.69	0.97	0.25	uppP	Undecaprenyl-diphosphatase
812331	0.09	0.56	0.42	murG	UDP-N-acetylglucosamine--N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase
813930	0.27	0.01	0.24	murC	UDP-N-acetylmuramate--L-alanine ligase
813939	0.27	0.01	0.24	murC	UDP-N-acetylmuramate--L-alanine ligase
813940	0.27	0.01	0.24	murC	UDP-N-acetylmuramate--L-alanine ligase
821022	0.1	0.6	0.44	prpC	2-methylcitrate synthase
821025	0.1	0.6	0.44	prpC	2-methylcitrate synthase
821054	0.07	0.62	0.51	prpC	2-methylcitrate synthase
821055	0.07	0.62	0.51	prpC	2-methylcitrate synthase
821057	0.07	0.62	0.51	prpC	2-methylcitrate synthase
821064	0.07	0.62	0.51	prpC	2-methylcitrate synthase
821067	0.07	0.62	0.51	prpC	2-methylcitrate synthase
821068	0.07	0.62	0.51	prpC	2-methylcitrate synthase
821082	0.1	0.62	0.45	prpC	2-methylcitrate synthase
821088	0.1	0.62	0.45	prpC	2-methylcitrate synthase
821127	0.26	0.72	0.36	prpC	2-methylcitrate synthase
821428	0.14	0.65	0.44	prpC	2-methylcitrate synthase
821583	0.09	0.62	0.48	prpC	2-methylcitrate synthase
821586	0.09	0.62	0.48	prpC	2-methylcitrate synthase
832393	0.36	0.07	0.21	acpS	Holo-[acyl-carrier-protein] synthase
833318	0.43	0.04	0.35	nudG	CTP pyrophosphohydrolase
833364	0.77	0.4	0.25	nudG	CTP pyrophosphohydrolase
836901	0.36	0.07	0.21	race	Glutamate racemase 1
837373	0.33	0.04	0.24	race	Glutamate racemase 1

860939	0.42	0.76	0.22	feuB	Iron-uptake system permease protein FeuB
861388	0.17	0.5	0.23	feuB	Iron-uptake system permease protein FeuB
874204	0.54	0.9	0.28	group_5969	Thiol-disulfide oxidoreductase ResA
886948	0.26	0.01	0.23	glyQ	Glycine--tRNA ligase alpha subunit
887842	0.07	0.43	0.3	glyS	Glycine--tRNA ligase beta subunit
887845	0.07	0.43	0.3	glyS	Glycine--tRNA ligase beta subunit
887848	0.07	0.43	0.3	glyS	Glycine--tRNA ligase beta subunit
887854	0.06	0.43	0.31	glyS	Glycine--tRNA ligase beta subunit
887912	0.15	0.53	0.29	glyS	Glycine--tRNA ligase beta subunit
887947	0.07	0.41	0.28	glyS	Glycine--tRNA ligase beta subunit
887950	0.07	0.41	0.28	glyS	Glycine--tRNA ligase beta subunit
887953	0.07	0.41	0.28	glyS	Glycine--tRNA ligase beta subunit
888010	0.05	0.4	0.3	glyS	Glycine--tRNA ligase beta subunit
888053	0.07	0.4	0.27	glyS	Glycine--tRNA ligase beta subunit
888409	0.3	0.94	0.62	glyS	Glycine--tRNA ligase beta subunit
888454	0.26	0	0.27	glyS	Glycine--tRNA ligase beta subunit
888475	0.39	0.01	0.37	glyS	Glycine--tRNA ligase beta subunit
888478	0.39	0.01	0.37	glyS	Glycine--tRNA ligase beta subunit
888559	0.27	0	0.27	glyS	Glycine--tRNA ligase beta subunit
888560	0.27	0	0.27	glyS	Glycine--tRNA ligase beta subunit
888568	0.3	0.01	0.28	glyS	Glycine--tRNA ligase beta subunit
888622	0.32	0.01	0.29	glyS	Glycine--tRNA ligase beta subunit
888706	0.27	0	0.27	glyS	Glycine--tRNA ligase beta subunit
888754	0.1	0.66	0.51	glyS	Glycine--tRNA ligase beta subunit
888757	0.1	0.66	0.52	glyS	Glycine--tRNA ligase beta subunit
888778	0.49	0.9	0.34	glyS	Glycine--tRNA ligase beta subunit
889342	0.37	0.75	0.27	glyS	Glycine--tRNA ligase beta subunit
889434	0.16	0.54	0.29	glyS	Glycine--tRNA ligase beta subunit
889444	0.17	0.59	0.33	glyS	Glycine--tRNA ligase beta subunit
889457	0.56	0.94	0.34	glyS	Glycine--tRNA ligase beta subunit
889555	0.17	0.57	0.3	glyS	Glycine--tRNA ligase beta subunit
890669	0.5	0.88	0.3	rsmE	Ribosomal RNA small subunit methyltransferase E
892490	0.11	0.68	0.51	group_6025	hypothetical protein
896140	0.12	0.71	0.53	msbA	Lipid A export ATP-binding/permease protein MsbA
897404	0.68	0.97	0.26	fabD	Malonyl CoA-acyl carrier protein transacylase
897668	0.28	0	0.28	fabD	Malonyl CoA-acyl carrier protein transacylase

897671	0.28	0	0.28	fabD	Malonyl CoA-acyl carrier protein transacylase
897691	0.28	0.01	0.25	fabD	Malonyl CoA-acyl carrier protein transacylase
899785	0.34	0.69	0.22	plsX	Phosphate acyltransferase
911562	0.23	0	0.24	group_5902	Prolyl endopeptidase
914151	0.62	0.91	0.22	argA	Amino-acid acetyltransferase
920001	0.7	0.99	0.28	xerC	Tyrosine recombinase XerC
921694	0.38	0.81	0.32	Dxs	1-deoxy-D-xylulose-5-phosphate synthase
926940	0.03	0.32	0.25	pyrC	Dihydroorotase
944128	0.05	0.54	0.46	bamD	Outer membrane protein assembly factor BamD precursor
945016	0.31	0.68	0.24	rluD	Ribosomal large subunit pseudouridine synthase D
945164	0.04	0.32	0.23	rluD	Ribosomal large subunit pseudouridine synthase D
945221	0.03	0.32	0.25	rluD	Ribosomal large subunit pseudouridine synthase D
945466	0.04	0.49	0.41	rluD	Ribosomal large subunit pseudouridine synthase D
946198	0.06	0.51	0.41	group_2272	Sodium Bile acid symporter family protein
946709	0.05	0.5	0.41	group_2272	Sodium Bile acid symporter family protein
947619	0.25	0.66	0.3	yfiH	Laccase domain protein YfiH
947789	0.27	0.68	0.29	lptE	LPS-assembly lipoprotein LptE precursor
948679	0.35	0.75	0.29	holA	DNA polymerase III subunit delta
959122	0.57	0.97	0.38	thrS	Threonine--tRNA ligase
959128	0.57	0.97	0.38	thrS	Threonine--tRNA ligase
959131	0.57	0.97	0.38	thrS	Threonine--tRNA ligase
959439	0.31	0.66	0.22	thrS	Threonine--tRNA ligase
969195	0.68	0.99	0.29	bioA	Adenosylmethionine-8-amino-7-oxononanoate aminotransferase
972371	0.7	1	0.31	hprK	HPr kinase/phosphorylase
996962	0.3	0.75	0.34	group_5982	hypothetical protein
997067	0.36	0.79	0.34	group_5982	hypothetical protein
997136	0.28	0.79	0.43	group_5982	hypothetical protein
997151	0.21	0.66	0.35	group_5982	hypothetical protein
997156	0.21	0.66	0.35	group_5982	hypothetical protein
997211	0.37	0.75	0.27	group_5982	hypothetical protein
997295	0.03	0.49	0.43	group_5982	hypothetical protein
997436	0.29	0.63	0.22	group_5982	hypothetical protein
997437	0.29	0.63	0.22	group_5982	hypothetical protein
997439	0.29	0.63	0.22	group_5982	hypothetical protein

998454	0.31	0.88	0.52	hemE	Uroporphyrinogen decarboxylase
998455	0.31	0.88	0.52	hemE	Uroporphyrinogen decarboxylase
998458	0.31	0.88	0.52	hemE	Uroporphyrinogen decarboxylase
998932	0.33	0.9	0.52	hemE	Uroporphyrinogen decarboxylase
1000130	0.12	0.56	0.36	group_1834	Z1 domain protein
1000670	0.32	0.81	0.4	group_1834	Z1 domain protein
1001600	0.41	0.87	0.38	group_5402	NgoFVII restriction endonuclease
1002997	0.23	0.01	0.2	bspRIM	Modification methylase BspRI
1003602	0.45	0.82	0.26	group_5735	hypothetical protein
1007173	0.29	0.69	0.29	group_6052	putative amino-acid ABC transporter-binding protein precursor
1008083	0.03	0.49	0.44	tcyB	L-cystine transport system permease protein TcyB
1008432	0.23	0.59	0.23	algC	Phosphomannomutase/phosphoglucomutase
1015541	0.05	0.56	0.48	rlmE	Ribosomal RNA large subunit methyltransferase E
1016207	0.37	0.79	0.31	group_1840	RNA-binding protein
1018091	0.33	0.71	0.25	metC	Cystathionine beta-lyase
1021092	0.21	0.59	0.27	ppnK	putative inorganic polyphosphate/ATP-NAD kinase
1022364	0.23	0	0.24	group_2301	hypothetical protein
1023318	0.23	0	0.24	fadR	Fatty acid metabolism regulator protein
1023321	0.23	0	0.24	fadR	Fatty acid metabolism regulator protein
1023802	0.23	0	0.24	murB	UDP-N-acetylenolpyruvoylglucosamine reductase
1024251	0.23	0	0.24	murB	UDP-N-acetylenolpyruvoylglucosamine reductase
1024771	0.25	0	0.26	murB	UDP-N-acetylenolpyruvoylglucosamine reductase
1025772	0.49	0.81	0.21	mdtK	Multidrug resistance protein MdtK
1025863	0.23	0	0.24	mdtK	Multidrug resistance protein MdtK
1025868	0.23	0	0.24	mdtK	Multidrug resistance protein MdtK
1025928	0.36	0.79	0.34	mdtK	Multidrug resistance protein MdtK
1026163	0.56	0.9	0.26	mdtK	Multidrug resistance protein MdtK
1030374	0.3	0.68	0.25	pyrF	Orotidine 5'-phosphate decarboxylase
1032460	0.7	0.99	0.28	hldD	ADP-L-glycero-D-manno-heptose-6-epimerase
1034382	0.4	0.79	0.28	group_1845	putative type I restriction enzymeP M protein

1034460	0.4	0.78	0.26	group_1845	putative type I restriction enzymeP M protein
1034461	0.41	0.78	0.25	group_1845	putative type I restriction enzymeP M protein
1034875	0.39	0.79	0.29	group_5775	DNA-damage-inducible protein D
1035093	0.39	0.81	0.31	group_5775	DNA-damage-inducible protein D
1043844	0.74	1	0.27	group_5400	MORN repeat protein
1058398	0.4	0.75	0.23	cysA	Sulfate/thiosulfate import ATP-binding protein CysA
1058793	0.44	0.82	0.27	cysW_1	Sulfate transport system permease protein CysW
1059138	0.47	0.79	0.21	cysW_1	Sulfate transport system permease protein CysW
1059260	0.43	0.79	0.25	cysW_1	Sulfate transport system permease protein CysW
1063020	0.23	0.01	0.2	group_373	hypothetical protein
1063062	0.23	0.01	0.2	group_373	hypothetical protein
1063165	0.02	0.38	0.35	group_373	hypothetical protein
1068923	0.47	0.79	0.21	group_2316	hypothetical protein
1069074	0.16	0.69	0.46	group_2316	hypothetical protein
1070219	0.42	0.75	0.21	group_5750	hypothetical protein
1070914	0.35	0.69	0.21	group_1499	hypothetical protein
1071002	0.37	0.71	0.2	group_1499	hypothetical protein
1071032	0.35	0.69	0.21	group_1499	hypothetical protein
1071158	0.14	0.49	0.25	group_1499	hypothetical protein
1071176	0.35	0.69	0.21	group_1499	hypothetical protein
1071215	0.38	0.75	0.25	group_1499	hypothetical protein
1083160	0.43	0.79	0.24	sstT	Serine/threonine transporter SstT
1085290	0.52	0.88	0.27	tamB	Translocation and assembly module TamB
1088567	0.54	0.9	0.28	group_2332	hypothetical protein
1089020	0.47	0.9	0.36	prfB	Peptide chain release factor 2
1089178	0.57	0.96	0.35	prfB	Peptide chain release factor 2
1089179	0.57	0.96	0.35	prfB	Peptide chain release factor 2
1089184	0.57	0.96	0.35	prfB	Peptide chain release factor 2
1089310	0.46	0.87	0.32	prfB	Peptide chain release factor 2
1089313	0.46	0.87	0.32	prfB	Peptide chain release factor 2
1089322	0.46	0.87	0.32	prfB	Peptide chain release factor 2
1089574	0.54	0.9	0.28	prfB	Peptide chain release factor 2
1090159	0.36	0.79	0.34	rssA_1	NTE family protein RssA
1093718	0.68	0.99	0.29	group_5922	hypothetical protein
1094428	0.5	0.84	0.23	tsaC_2	Threonylcarbamoyl-AMP synthase
1099565	0.19	0.68	0.4	group_1868	nicotinamidase/pyrazinamidase
1099574	0.19	0.66	0.38	group_1868	nicotinamidase/pyrazinamidase
1099575	0.19	0.66	0.38	group_1868	nicotinamidase/pyrazinamidase

1099705	0.23	0.68	0.34	group_1868	nicotinamidase/pyrazinamidase
1099739	0.01	0.5	0.49	group_1868	nicotinamidase/pyrazinamidase
1100601	0.1	0.5	0.32	tsaA	putative tRNA (adenine(37)-N6)-methyltransferase
1115720	0.54	0.85	0.21	group_5680	Twitching mobility protein
1120553	0.11	0.43	0.22	gndA	6-phosphogluconate dehydrogenase, NADP( )-dependent, decarboxylating
1120815	0.01	0.31	0.29	gndA	6-phosphogluconate dehydrogenase, NADP( )-dependent, decarboxylating
1122277	0.23	0	0.24	minD_1	Septum site-determining protein MinD
1122279	0.23	0	0.24	minD_1	Septum site-determining protein MinD
1122283	0.23	0	0.24	minD_1	Septum site-determining protein MinD
1122285	0.23	0	0.24	minD_1	Septum site-determining protein MinD
1122288	0.23	0	0.24	minD_1	Septum site-determining protein MinD
1122290	0.23	0	0.24	minD_1	Septum site-determining protein MinD
1122708	0.1	0.46	0.27	minD_1	Septum site-determining protein MinD
1122759	0.13	0.54	0.33	minD_1	Septum site-determining protein MinD
1123190	0.31	0.69	0.25	minD_1	Septum site-determining protein MinD
1123302	0.42	0.79	0.26	minD_1	Septum site-determining protein MinD
1124624	0.23	0	0.24	hpaC	4-hydroxyphenylacetate 3-monooxygenase reductase component
1125176	0.26	0	0.27	rmlA	Glucose-1-phosphate thymidylyltransferase
1127388	0.23	0	0.24	Fhs	Formate--tetrahydrofolate ligase
1127837	0.23	0	0.24	ychF	Ribosome-binding ATPase YchF
1129161	0.11	0.51	0.32	tyrS	Tyrosine--tRNA ligase
1157187	0.01	0.34	0.33	oxyR	Hydrogen peroxide-inducible genes activator
1157862	0.35	0.07	0.21	truA	tRNA pseudouridine synthase A
1157865	0.35	0.07	0.21	truA	tRNA pseudouridine synthase A
1157869	0.35	0.07	0.21	truA	tRNA pseudouridine synthase A
1157871	0.35	0.07	0.21	truA	tRNA pseudouridine synthase A
1163035	0.65	0.96	0.26	alsT_2	Amino-acid carrier protein AlsT
1184164	0.39	0.72	0.2	pdxA1	4-hydroxythreonine-4-phosphate dehydrogenase 1

1195623	0.68	0.96	0.23	gyrB	DNA gyrase subunit B
1197770	0.5	0.87	0.27	prlC	Oligopeptidase A
1197785	0.53	0.87	0.24	prlC	Oligopeptidase A
1197834	0.45	0.81	0.24	prlC	Oligopeptidase A
1198109	0.5	0.87	0.28	prlC	Oligopeptidase A
1198121	0.5	0.88	0.3	prlC	Oligopeptidase A
1198127	0.5	0.87	0.28	prlC	Oligopeptidase A
1198157	0.5	0.88	0.3	prlC	Oligopeptidase A
1198167	0.5	0.88	0.3	prlC	Oligopeptidase A
1198169	0.5	0.88	0.3	prlC	Oligopeptidase A
1198172	0.5	0.88	0.3	prlC	Oligopeptidase A
1198259	0.43	0.84	0.32	prlC	Oligopeptidase A
1198265	0.43	0.85	0.34	prlC	Oligopeptidase A
1198280	0.4	0.78	0.26	prlC	Oligopeptidase A
1198289	0.4	0.78	0.26	prlC	Oligopeptidase A
1198301	0.33	0.66	0.2	prlC	Oligopeptidase A
1198304	0.33	0.66	0.2	prlC	Oligopeptidase A
1204240	0.62	0.97	0.33	pyrD	Dihydroorotate dehydrogenase (quinone)
1209643	0.03	0.41	0.37	Mqo	Malate:quinone oxidoreductase
1209647	0.03	0.41	0.37	Mqo	Malate:quinone oxidoreductase
1209664	0.03	0.41	0.35	Mqo	Malate:quinone oxidoreductase
1209754	0.03	0.41	0.37	Mqo	Malate:quinone oxidoreductase
1209883	0.03	0.41	0.37	Mqo	Malate:quinone oxidoreductase
1209895	0.03	0.41	0.37	Mqo	Malate:quinone oxidoreductase
1209910	0.03	0.41	0.37	Mqo	Malate:quinone oxidoreductase
1209922	0.03	0.41	0.37	Mqo	Malate:quinone oxidoreductase
1209953	0.03	0.41	0.37	Mqo	Malate:quinone oxidoreductase
1209988	0.03	0.41	0.37	Mqo	Malate:quinone oxidoreductase
1210078	0.03	0.41	0.37	Mqo	Malate:quinone oxidoreductase
1210189	0.03	0.41	0.37	Mqo	Malate:quinone oxidoreductase
1210228	0.03	0.41	0.37	Mqo	Malate:quinone oxidoreductase
1210246	0.03	0.41	0.37	Mqo	Malate:quinone oxidoreductase
1210453	0.03	0.43	0.38	group_5560	hypothetical protein
1210495	0.03	0.43	0.38	group_5560	hypothetical protein
1210498	0.03	0.43	0.38	group_5560	hypothetical protein
1210508	0.03	0.43	0.38	group_5560	hypothetical protein
1227446	0.63	0.29	0.2	group_1742	bifunctional tRNA (mnm(5)s(2)U34)-methyltransferase/FAD-dependent cmnm(5)s(2)U34 oxidoreductase
1227598	0.03	0.34	0.27	group_1742	bifunctional tRNA (mnm(5)s(2)U34)-

					methyltransferase/FAD-dependent cmnm(5)s(2)U34 oxidoreductase
1229331	0.39	0.76	0.25	codB	Cytosine permease
1239436	0.1	0.49	0.32	petA	Ubiquinol-cytochrome c reductase iron-sulfur subunit
1241205	0.68	0.97	0.26	petB	Cytochrome b
1241269	0.13	0.49	0.26	petB	Cytochrome b
1241323	0.17	0.5	0.23	petC	Cytochrome c1 precursor
1241360	0.14	0.5	0.27	petC	Cytochrome c1 precursor
1241410	0.21	0.57	0.25	petC	Cytochrome c1 precursor
1244524	0.05	0.38	0.28	rmuC	DNA recombination protein RmuC
1251406	0.55	0.85	0.2	thiC	Phosphomethylpyrimidine synthase

Table 2.S6. Homoplastic F_{ST} outliers gained with the GGI on the core genome phylogeny

SNP	Frequency Gained with GGI	Roary ID	Prokka Annotation
7240	3	ygbN	Inner membrane permease YgbN
8998	1	thrB	Homoserine kinase
9016	5	thrB	Homoserine kinase
9237	4	ubiG	Ubiquinone biosynthesis O-methyltransferase
10068	2	mtr	Tryptophan-specific transport protein
10074	2	mtr	Tryptophan-specific transport protein
27637	3	ybaN	Inner membrane protein YbaN
68854	3	group_2522	hypothetical protein
68980	6	group_2522	hypothetical protein
68998	6	group_2522	hypothetical protein
69036	6	group_2522	hypothetical protein
69050	6	group_2522	hypothetical protein
69059	6	group_2522	hypothetical protein
69697	5	nadC	putative nicotinate-nucleotide pyrophosphorylase [carboxylating]
71319	2	cbbZC	Phosphoglycolate phosphatase, chromosomal
78699	3	ribD	Riboflavin biosynthesis protein RibD
78714	3	ribD	Riboflavin biosynthesis protein RibD
78720	3	ribD	Riboflavin biosynthesis protein RibD
78723	3	ribD	Riboflavin biosynthesis protein RibD
78738	3	ribD	Riboflavin biosynthesis protein RibD
78740	3	ribD	Riboflavin biosynthesis protein RibD
78990	3	group_5712	Glutamine amidotransferases class-II
82058	2	group_6255	Fimbrial assembly protein (PilN)
93623	1	dacC_2	D-alanyl-D-alanine carboxypeptidase DacC precursor
123571	2	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB

123577	2	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
123580	2	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
123583	2	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
123589	2	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
123598	2	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
123604	2	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
123631	2	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
123640	2	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
123646	2	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
123652	2	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
123653	2	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
123714	4	group_5590	putative 3'-5' exonuclease related to the exonuclease domain of PolB
128223	3	trpA	Tryptophan synthase alpha chain
130411	2	tmk	Thymidylate kinase
154881	3	lolA	Outer-membrane lipoprotein carrier protein precursor
155602	1	group_1250	hypothetical protein
155836	3	group_1250	hypothetical protein
155839	3	group_1250	hypothetical protein
155840	3	group_1250	hypothetical protein
155860	2	group_1250	hypothetical protein
155875	2	group_1250	hypothetical protein
165334	2	group_5506	preprotein translocase subunit SecD
165703	2	group_5506	preprotein translocase subunit SecD
168666	2	tatC	Sec-independent protein translocase protein TatC
182257	3	ybbH	putative HTH-type transcriptional regulator YbbH
183103	2	glk	Glucokinase
209785	2	serC	Phosphoserine aminotransferase
209818	2	serC	Phosphoserine aminotransferase
210377	1	serC	Phosphoserine aminotransferase
210770	2	serC	Phosphoserine aminotransferase
210773	2	serC	Phosphoserine aminotransferase
210781	2	serC	Phosphoserine aminotransferase
210785	2	serC	Phosphoserine aminotransferase
210787	2	serC	Phosphoserine aminotransferase

210791	2	serC	Phosphoserine aminotransferase
210792	2	serC	Phosphoserine aminotransferase
210800	2	serC	Phosphoserine aminotransferase
210803	2	serC	Phosphoserine aminotransferase
212919	1	group_5905	putative membrane protein
225745	2	dapA	4-hydroxy-tetrahydrodipicolinate synthase
227485	1	rlmB_1	23S rRNA (guanosine-2'-O-)-methyltransferase RlmB
227487	1	rlmB_1	23S rRNA (guanosine-2'-O-)-methyltransferase RlmB
227735	4	rlmB_1	23S rRNA (guanosine-2'-O-)-methyltransferase RlmB
274076	1	rlmJ	Ribosomal RNA large subunit methyltransferase J
274077	1	rlmJ	Ribosomal RNA large subunit methyltransferase J
274193	3	rlmJ	Ribosomal RNA large subunit methyltransferase J
274199	3	rlmJ	Ribosomal RNA large subunit methyltransferase J
274201	3	rlmJ	Ribosomal RNA large subunit methyltransferase J
274202	3	rlmJ	Ribosomal RNA large subunit methyltransferase J
274203	3	rlmJ	Ribosomal RNA large subunit methyltransferase J
274205	3	rlmJ	Ribosomal RNA large subunit methyltransferase J
277534	1	group_6143	hypothetical protein
292403	2	group_5730	Putative lipoprotein/NMB1164 precursor
312618	1	cysJ_2	Sulfite reductase [NADPH] flavoprotein alpha-component
319624	1	rnr	Ribonuclease R
344448	1	cca	Multifunctional CCA protein
353929	3	group_6177	hypothetical protein
354098	3	group_6177	hypothetical protein
354101	3	group_6177	hypothetical protein
354102	3	group_6177	hypothetical protein
354104	3	group_6177	hypothetical protein
354109	3	group_6177	hypothetical protein
355190	3	group_5434	Putative O-methyltransferase/MSMEI_4947
355193	3	group_5434	Putative O-methyltransferase/MSMEI_4947
365148	1	spoT	Bifunctional (p)ppGpp synthase/hydrolase SpoT
365339	3	spoT	Bifunctional (p)ppGpp synthase/hydrolase SpoT
365519	1	spoT	Bifunctional (p)ppGpp synthase/hydrolase SpoT
395445	2	group_5868	tetratricopeptide repeat protein
395589	3	group_5868	tetratricopeptide repeat protein
441528	2	ttcA	tRNA 2-thiocytidine biosynthesis protein TtcA
441531	2	ttcA	tRNA 2-thiocytidine biosynthesis protein TtcA
443711	1	dnaJ_1	Chaperone protein DnaJ
469041	1	copA_1	Copper-exporting P-type ATPase A
483399	2	group_5865	S-formylglutathione hydrolase
520112	1	nadE	NH(3)-dependent NAD( ) synthetase
531509	1	sad	Succinate semialdehyde dehydrogenase [NAD(P)] Sad

532117	2	group_6064	hypothetical protein
535613	2	suhB_1	Inositol-1-monophosphatase
535753	1	suhB_1	Inositol-1-monophosphatase
537867	2	hrpB	ATP-dependent RNA helicase HrpB
537947	3	hrpB	ATP-dependent RNA helicase HrpB
537948	2	hrpB	ATP-dependent RNA helicase HrpB
550647	1	parC	DNA topoisomerase 4 subunit A
555254	1	alaS	Alanine--tRNA ligase
555267	1	alaS	Alanine--tRNA ligase
555269	1	alaS	Alanine--tRNA ligase
555272	1	alaS	Alanine--tRNA ligase
570618	2	ilvI	Acetolactate synthase isozyme 3 large subunit
571332	4	ilvI	Acetolactate synthase isozyme 3 large subunit
596256	2	uvrA	UvrABC system protein A
596301	2	uvrA	UvrABC system protein A
596348	2	uvrA	UvrABC system protein A
596370	2	uvrA	UvrABC system protein A
596397	2	uvrA	UvrABC system protein A
596415	2	uvrA	UvrABC system protein A
596640	1	uvrA	UvrABC system protein A
597282	1	uvrA	UvrABC system protein A
597283	1	uvrA	UvrABC system protein A
597291	1	uvrA	UvrABC system protein A
597294	1	uvrA	UvrABC system protein A
597295	1	uvrA	UvrABC system protein A
605685	3	group_6085	Cytochrome c-555 precursor
618150	1	group_5652	hypothetical protein
631979	1	oprM	Outer membrane protein OprM precursor
633218	3	oprM	Outer membrane protein OprM precursor
633418	3	group_2485	hypothetical protein
633422	3	group_2485	hypothetical protein
633424	3	group_2485	hypothetical protein
634867	2	abgT	p-aminobenzoyl-glutamate transport protein
634870	2	abgT	p-aminobenzoyl-glutamate transport protein
634930	2	abgT	p-aminobenzoyl-glutamate transport protein
634936	2	abgT	p-aminobenzoyl-glutamate transport protein
634939	2	abgT	p-aminobenzoyl-glutamate transport protein
634996	3	abgT	p-aminobenzoyl-glutamate transport protein
635005	3	abgT	p-aminobenzoyl-glutamate transport protein
635011	3	abgT	p-aminobenzoyl-glutamate transport protein
635014	3	abgT	p-aminobenzoyl-glutamate transport protein
635020	3	abgT	p-aminobenzoyl-glutamate transport protein
635022	3	abgT	p-aminobenzoyl-glutamate transport protein

635023	3	abgT	p-aminobenzoyl-glutamate transport protein
643413	1	def	Peptide deformylase
655043	2	group_1149	hypothetical protein
660148	1	ygdH	LOG family protein YgdH
660154	1	ygdH	LOG family protein YgdH
669416	1	recQ	ATP-dependent DNA helicase RecQ
671479	1	group_6187	hypothetical protein
673047	2	group_2577	RlpA-like protein precursor
690336	3	ndhC	NAD(P)H-quinone oxidoreductase subunit 3
690588	1	ndhC	NAD(P)H-quinone oxidoreductase subunit 3
694104	1	gabD	Succinate-semialdehyde dehydrogenase [NADP( )] GabD
701262	1	group_6035	Nickel uptake substrate-specific transmembrane region
702422	1	alaA	Glutamate-pyruvate aminotransferase AlaA
709115	2	rph	Ribonuclease PH
710772	1	scpA	Segregation and condensation protein A
710775	1	scpA	Segregation and condensation protein A
710781	1	scpA	Segregation and condensation protein A
710787	1	scpA	Segregation and condensation protein A
710793	1	scpA	Segregation and condensation protein A
710796	1	scpA	Segregation and condensation protein A
710799	1	scpA	Segregation and condensation protein A
711618	1	pncB2	Nicotinate phosphoribosyltransferase 2
713184	1	argS	Arginine--tRNA ligase
713205	1	argS	Arginine--tRNA ligase
713301	2	argS	Arginine--tRNA ligase
713559	2	argS	Arginine--tRNA ligase
715347	4	rpiA	Ribose-5-phosphate isomerase A
725453	2	group_6164	putative FAD-linked oxidoreductase
735176	1	dnaG	DNA primase
749431	2	fumC	Fumarate hydratase class II
750589	1	yhbE	putative inner membrane transporter YhbE
752398	3	group_2350	murein transglycosylase C
752692	1	group_2350	murein transglycosylase C
752905	1	group_321	Transposase DDE domain protein
752911	1	group_321	Transposase DDE domain protein
753760	3	metG	Methionine--tRNA ligase
754269	3	metG	Methionine--tRNA ligase
754694	3	metG	Methionine--tRNA ligase
754698	3	metG	Methionine--tRNA ligase
755128	4	metG	Methionine--tRNA ligase
755129	3	metG	Methionine--tRNA ligase

755740	2	glmS	Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
757737	3	mltA	Membrane-bound lytic murein transglycosylase A precursor
760331	3	glmU	Bifunctional protein GlmU
760579	3	glmU	Bifunctional protein GlmU
760615	1	glmU	Bifunctional protein GlmU
765753	2	ftsY	Signal recognition particle receptor FtsY
765988	3	ftsY	Signal recognition particle receptor FtsY
766238	3	ftsY	Signal recognition particle receptor FtsY
768228	2	ileS	Isoleucine--tRNA ligase
768291	2	ileS	Isoleucine--tRNA ligase
768369	1	ileS	Isoleucine--tRNA ligase
768370	1	ileS	Isoleucine--tRNA ligase
768375	1	ileS	Isoleucine--tRNA ligase
768384	1	ileS	Isoleucine--tRNA ligase
768715	1	ileS	Isoleucine--tRNA ligase
770085	1	ileS	Isoleucine--tRNA ligase
772983	3	nqrB	Na( )-translocating NADH-quinone reductase subunit B
787096	1	cmpB	Bicarbonate transport system permease protein CmpB
787188	1	ssuB	Aliphatic sulfonates import ATP-binding protein SsuB
787217	2	ssuB	Aliphatic sulfonates import ATP-binding protein SsuB
788960	1	dsbC_1	putative thiol:disulfide interchange protein DsbC precursor
792106	3	smc	Chromosome partition protein Smc
796631	3	lysS	Lysine--tRNA ligase
798258	1	group_1984	Natural resistance-associated macrophage protein
798276	2	group_1984	Natural resistance-associated macrophage protein
798288	2	group_1984	Natural resistance-associated macrophage protein
798302	2	group_1984	Natural resistance-associated macrophage protein
798408	2	group_1984	Natural resistance-associated macrophage protein
798855	1	group_1984	Natural resistance-associated macrophage protein
800210	1	comM	Competence protein ComM
801354	3	group_5630	cell division protein FtsN
802304	2	dsbA_1	Thiol:disulfide interchange protein DsbA precursor
802973	1	uppP	Undecaprenyl-diphosphatase
812331	1	murG	UDP-N-acetylglucosamine--N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase
813930	2	murC	UDP-N-acetylmuramate--L-alanine ligase
813939	2	murC	UDP-N-acetylmuramate--L-alanine ligase
813940	2	murC	UDP-N-acetylmuramate--L-alanine ligase
821022	2	prpC	2-methylcitrate synthase

821025	2	prpC	2-methylcitrate synthase
821054	2	prpC	2-methylcitrate synthase
821055	2	prpC	2-methylcitrate synthase
821057	2	prpC	2-methylcitrate synthase
821064	2	prpC	2-methylcitrate synthase
821067	2	prpC	2-methylcitrate synthase
821068	2	prpC	2-methylcitrate synthase
821082	2	prpC	2-methylcitrate synthase
821088	2	prpC	2-methylcitrate synthase
821127	2	prpC	2-methylcitrate synthase
821428	2	prpC	2-methylcitrate synthase
821583	2	prpC	2-methylcitrate synthase
821586	2	prpC	2-methylcitrate synthase
832393	1	acpS	Holo-[acyl-carrier-protein] synthase
833318	2	nudG	CTP pyrophosphohydrolase
833364	3	nudG	CTP pyrophosphohydrolase
836901	1	racE	Glutamate racemase 1
837373	1	racE	Glutamate racemase 1
860939	1	feuB	Iron-uptake system permease protein FeuB
861388	3	feuB	Iron-uptake system permease protein FeuB
874204	1	group_5969	Thiol-disulfide oxidoreductase ResA
886948	2	glyQ	Glycine--tRNA ligase alpha subunit
887842	1	glyS	Glycine--tRNA ligase beta subunit
887845	1	glyS	Glycine--tRNA ligase beta subunit
887848	1	glyS	Glycine--tRNA ligase beta subunit
887854	1	glyS	Glycine--tRNA ligase beta subunit
887912	1	glyS	Glycine--tRNA ligase beta subunit
887947	1	glyS	Glycine--tRNA ligase beta subunit
887950	1	glyS	Glycine--tRNA ligase beta subunit
887953	1	glyS	Glycine--tRNA ligase beta subunit
888010	1	glyS	Glycine--tRNA ligase beta subunit
888053	1	glyS	Glycine--tRNA ligase beta subunit
888409	2	glyS	Glycine--tRNA ligase beta subunit
888454	2	glyS	Glycine--tRNA ligase beta subunit
888475	3	glyS	Glycine--tRNA ligase beta subunit
888478	3	glyS	Glycine--tRNA ligase beta subunit
888559	2	glyS	Glycine--tRNA ligase beta subunit
888560	2	glyS	Glycine--tRNA ligase beta subunit
888568	2	glyS	Glycine--tRNA ligase beta subunit
888622	3	glyS	Glycine--tRNA ligase beta subunit
888706	3	glyS	Glycine--tRNA ligase beta subunit
888754	1	glyS	Glycine--tRNA ligase beta subunit
888757	1	glyS	Glycine--tRNA ligase beta subunit

888778	4	glyS	Glycine--tRNA ligase beta subunit
889342	2	glyS	Glycine--tRNA ligase beta subunit
889434	2	glyS	Glycine--tRNA ligase beta subunit
889444	2	glyS	Glycine--tRNA ligase beta subunit
889457	1	glyS	Glycine--tRNA ligase beta subunit
889555	2	glyS	Glycine--tRNA ligase beta subunit
890669	1	rsmE	Ribosomal RNA small subunit methyltransferase E
892490	2	group_6025	hypothetical protein
896140	2	msbA	Lipid A export ATP-binding/permease protein MsbA
897404	2	fabD	Malonyl CoA-acyl carrier protein transacylase
897668	2	fabD	Malonyl CoA-acyl carrier protein transacylase
897671	2	fabD	Malonyl CoA-acyl carrier protein transacylase
897691	2	fabD	Malonyl CoA-acyl carrier protein transacylase
899785	2	plsX	Phosphate acyltransferase
911562	2	group_5902	Prolyl endopeptidase
914151	1	argA	Amino-acid acetyltransferase
921694	2	dxs	1-deoxy-D-xylulose-5-phosphate synthase
926940	1	pyrC	Dihydroorotase
945016	2	rluD	Ribosomal large subunit pseudouridine synthase D
947619	3	yfiH	Laccase domain protein YfiH
947789	3	lptE	LPS-assembly lipoprotein LptE precursor
948679	4	holA	DNA polymerase III subunit delta
959122	3	thrS	Threonine--tRNA ligase
959128	3	thrS	Threonine--tRNA ligase
959131	3	thrS	Threonine--tRNA ligase
959439	1	thrS	Threonine--tRNA ligase
972371	1	hprK	HPr kinase/phosphorylase
996962	2	group_5982	hypothetical protein
997067	2	group_5982	hypothetical protein
997136	2	group_5982	hypothetical protein
997151	1	group_5982	hypothetical protein
997156	1	group_5982	hypothetical protein
997211	1	group_5982	hypothetical protein
997295	1	group_5982	hypothetical protein
997436	1	group_5982	hypothetical protein
997437	1	group_5982	hypothetical protein
997439	1	group_5982	hypothetical protein
998454	2	hemE	Uroporphyrinogen decarboxylase
998455	2	hemE	Uroporphyrinogen decarboxylase
998458	2	hemE	Uroporphyrinogen decarboxylase
998932	3	hemE	Uroporphyrinogen decarboxylase
1000670	3	group_1834	Z1 domain protein
1001600	1	group_5402	NgoFVII restriction endonuclease

1002997	2	bspRIM	Modification methylase BspRI
1003602	1	group_5735	hypothetical protein
1007173	3	group_6052	putative amino-acid ABC transporter-binding protein precursor
1008083	1	tcyB	L-cystine transport system permease protein TcyB
1008432	1	algC	Phosphomannomutase/phosphoglucomutase
1015541	1	rlmE	Ribosomal RNA large subunit methyltransferase E
1016207	4	group_1840	RNA-binding protein
1018091	2	metC	Cystathionine beta-lyase
1021092	3	ppnK	putative inorganic polyphosphate/ATP-NAD kinase
1022364	2	group_2301	hypothetical protein
1023318	2	fadR	Fatty acid metabolism regulator protein
1023321	2	fadR	Fatty acid metabolism regulator protein
1023802	2	murB	UDP-N-acetylenolpyruvoylglucosamine reductase
1024251	2	murB	UDP-N-acetylenolpyruvoylglucosamine reductase
1024771	2	murB	UDP-N-acetylenolpyruvoylglucosamine reductase
1025772	1	mdtK	Multidrug resistance protein MdtK
1025863	2	mdtK	Multidrug resistance protein MdtK
1025868	2	mdtK	Multidrug resistance protein MdtK
1025928	2	mdtK	Multidrug resistance protein MdtK
1026163	1	mdtK	Multidrug resistance protein MdtK
1030374	1	pyrF	Orotidine 5'-phosphate decarboxylase
1032460	1	hldD	ADP-L-glycero-D-manno-heptose-6-epimerase
1035093	1	group_5775	DNA-damage-inducible protein D
1058398	1	cysA	Sulfate/thiosulfate import ATP-binding protein CysA
1058793	2	cysW_1	Sulfate transport system permease protein CysW
1059138	4	cysW_1	Sulfate transport system permease protein CysW
1059260	1	cysW_1	Sulfate transport system permease protein CysW
1063020	2	group_373	hypothetical protein
1063062	1	group_373	hypothetical protein
1063165	1	group_373	hypothetical protein
1069074	2	group_2316	hypothetical protein
1070914	1	group_1499	hypothetical protein
1071002	1	group_1499	hypothetical protein
1071032	1	group_1499	hypothetical protein
1071158	2	group_1499	hypothetical protein
1071176	1	group_1499	hypothetical protein
1071215	1	group_1499	hypothetical protein
1085290	1	tamB	Translocation and assembly module TamB
1088567	2	group_2332	hypothetical protein
1089020	1	prfB	Peptide chain release factor 2
1089178	1	prfB	Peptide chain release factor 2
1089179	1	prfB	Peptide chain release factor 2

1089184	1	prfB	Peptide chain release factor 2
1089310	2	prfB	Peptide chain release factor 2
1089313	2	prfB	Peptide chain release factor 2
1089322	2	prfB	Peptide chain release factor 2
1089574	2	prfB	Peptide chain release factor 2
1090159	2	rssA_1	NTE family protein RssA
1094428	1	tsaC_2	Threonylcarbamoyl-AMP synthase
1099565	2	group_1868	nicotinamidase/pyrazinamidase
1099574	2	group_1868	nicotinamidase/pyrazinamidase
1099575	2	group_1868	nicotinamidase/pyrazinamidase
1099705	1	group_1868	nicotinamidase/pyrazinamidase
1100601	1	tsaA	putative tRNA (adenine(37)-N6)-methyltransferase
1115720	1	group_5680	Twitching mobility protein
1122277	2	minD_1	Septum site-determining protein MinD
1122279	2	minD_1	Septum site-determining protein MinD
1122283	2	minD_1	Septum site-determining protein MinD
1122285	2	minD_1	Septum site-determining protein MinD
1122288	2	minD_1	Septum site-determining protein MinD
1122290	2	minD_1	Septum site-determining protein MinD
1122708	2	minD_1	Septum site-determining protein MinD
1122759	1	minD_1	Septum site-determining protein MinD
1123302	2	minD_1	Septum site-determining protein MinD
1124624	2	hpaC	4-hydroxyphenylacetate 3-monooxygenase reductase component
1125176	2	rmlA	Glucose-1-phosphate thymidylyltransferase
1127388	2	fhs	Formate--tetrahydrofolate ligase
1127837	2	ychF	Ribosome-binding ATPase YchF
1129161	4	tyrS	Tyrosine--tRNA ligase
1157862	1	truA	tRNA pseudouridine synthase A
1157865	1	truA	tRNA pseudouridine synthase A
1157869	1	truA	tRNA pseudouridine synthase A
1157871	1	truA	tRNA pseudouridine synthase A
1184164	2	pdxA1	4-hydroxythreonine-4-phosphate dehydrogenase 1
1197770	6	prlC	Oligopeptidase A
1197785	4	prlC	Oligopeptidase A
1197834	4	prlC	Oligopeptidase A
1198109	4	prlC	Oligopeptidase A
1198121	4	prlC	Oligopeptidase A
1198127	4	prlC	Oligopeptidase A
1198157	4	prlC	Oligopeptidase A
1198167	4	prlC	Oligopeptidase A
1198169	4	prlC	Oligopeptidase A
1198172	4	prlC	Oligopeptidase A

1198259	2	prlC	Oligopeptidase A
1198265	2	prlC	Oligopeptidase A
1198280	2	prlC	Oligopeptidase A
1198289	2	prlC	Oligopeptidase A
1198301	2	prlC	Oligopeptidase A
1198304	2	prlC	Oligopeptidase A
1204240	1	pyrD	Dihydroorotate dehydrogenase (quinone)
1209664	1	mqo	Malate:quinone oxidoreductase
1227446	2	group_1742	bifunctional tRNA (mnm(5)s(2)U34)-methyltransferase/FAD-dependent cmnm(5)s(2)U34 oxidoreductase
1227598	2	group_1742	bifunctional tRNA (mnm(5)s(2)U34)-methyltransferase/FAD-dependent cmnm(5)s(2)U34 oxidoreductase
1229331	2	codB	Cytosine permease
1241269	3	petB	Cytochrome b
1241323	3	petC	Cytochrome c1 precursor
1241360	4	petC	Cytochrome c1 precursor
1241410	2	petC	Cytochrome c1 precursor
1251406	2	thiC	Phosphomethylpyrimidine synthase

Table 2.S7. Mobile genetic element genes identified in N. gonorrhoeae pan-genome

Roary ID	Prokka Annotation	# of Isolates
intA_2	Prophage CP4-57 integrase	183
group_1984	Natural resistance-associated macrophage protein	183
group_2280	phage T7 F exclusion suppressor FxsA	183
intA_1	Prophage CP4-57 integrase	183
group_321	Transposase DDE domain protein	183
group_5836	Phage integrase family protein	183
group_6190	Transposase IS200 like protein	183
sunS	SPBc2 prophage-derived glycosyltransferase SunS	182
group_2923	Phage Mu protein F like protein	182
group_1256	bacteriophage N4 receptor, outer membrane subunit	180
group_2112	Transposase IS116/IS110/IS902 family protein	172
group_2615	Transposase IS116/IS110/IS902 family protein	151
group_1068	Integrase core domain protein	147
group_431	ISXO2-like transposase domain protein	130
group_239	Bacteriophage replication protein O	106
group_2401	Phage tail protein (Tail_P2_I)	101
group_5798	Integrase core domain protein	76
group_421	Phage terminase large subunit	75
group_5799	Integrase core domain protein	73
group_149	ISXO2-like transposase domain protein	61
group_425	Phage terminase large subunit	55

group_296	Transposase IS116/IS110/IS902 family protein	48
group_301	Transposase IS116/IS110/IS902 family protein	46
group_230	Bacteriophage replication protein O	39
group_433	ISXO2-like transposase domain protein	37
group_2624	Phage tail fibre repeat protein	32
group_424	Phage terminase large subunit	31
group_298	Transposase IS116/IS110/IS902 family protein	30
group_129	ISXO2-like transposase domain protein	26
group_126	ISXO2-like transposase domain protein	24
group_143	ISXO2-like transposase domain protein	23
group_320	Transposase IS116/IS110/IS902 family protein	21
group_134	ISXO2-like transposase domain protein	20
group_311	Transposase IS116/IS110/IS902 family protein	20
group_135	ISXO2-like transposase domain protein	19
group_145	ISXO2-like transposase domain protein	19
group_1383	Integrase core domain protein	17
group_312	Transposase IS116/IS110/IS902 family protein	17
group_5801	Integrase core domain protein	17
group_5803	Integrase core domain protein	17
group_1001	Conjugative relaxosome accessory transposon protein	16
group_144	ISXO2-like transposase domain protein	16
group_299	Transposase IS116/IS110/IS902 family protein	16
group_147	ISXO2-like transposase domain protein	15
group_138	ISXO2-like transposase domain protein	14
group_304	Transposase IS116/IS110/IS902 family protein	14
group_316	Transposase IS116/IS110/IS902 family protein	14
group_136	ISXO2-like transposase domain protein	13
group_1384	Integrase core domain protein	12
group_140	ISXO2-like transposase domain protein	12
group_2614	Transposase IS116/IS110/IS902 family protein	12
group_314	Transposase IS116/IS110/IS902 family protein	11
group_2113	Transposase IS116/IS110/IS902 family protein	10
group_139	ISXO2-like transposase domain protein	9
group_306	Transposase IS116/IS110/IS902 family protein	8
group_5804	Integrase core domain protein	7
group_130	ISXO2-like transposase domain protein	6
group_2616	Transposase IS116/IS110/IS902 family protein	6
group_237	Bacteriophage replication protein O	5
group_1257	bacteriophage N4 receptor, outer membrane subunit	4
group_148	ISXO2-like transposase domain protein	4
group_303	Transposase IS116/IS110/IS902 family protein	3
group_142	ISXO2-like transposase domain protein	2
group_308	Transposase IS116/IS110/IS902 family protein	2

group_315	Transposase IS116/IS110/IS902 family protein	2
group_3307	ISXO2-like transposase domain protein	2
group_5802	Integrase core domain protein	2
group_127	ISXO2-like transposase domain protein	1
group_133	ISXO2-like transposase domain protein	1
group_1497	Prophage CP4-57 integrase	1
group_302	Transposase IS116/IS110/IS902 family protein	1
intS	Putative prophage CPS-53 integrase	1
group_4332	ISXO2-like transposase domain protein	1
group_5800	Integrase core domain protein	1
tnpR	Transposon Tn3 resolvase	18
tnpR_2	Transposon Tn3 resolvase	1
group_3939	IS1 transposase	1
group_3944	Tn3 transposase DDE domain protein	1
group_3969	Transposase DDE domain protein	1
group_3960	IS2 transposase TnpB	1
group_3961	Transposase	1
group_3962	Transposase, Mutator family	1
group_3957	Putative transposase	1
group_3964	Integrase core domain protein	1
group_3974	Transposase, Mutator family	1
group_3970	Transposase	1
group_3959	Transposase, Mutator family	1
group_3965	Transposase DDE domain protein	1
group_2138	conjugal transfer mating pair stabilization protein TraN	1
group_5156	Integrase core domain protein	1
pepF1	Oligoendopeptidase F, plasmid	1
group_4265	Bacteriophage peptidoglycan hydrolase	1
group_2685	Putative conjugal transfer protein/MT3759	28
trfA	Plasmid replication initiator protein TrfA	28
virB4	Type IV secretion system protein virB4	27
traG	Conjugal transfer protein TraG	27
ptlG	Type IV secretion system protein virB10	27
group_2683	Conjugal transfer protein	27
group_2684	VirB8 protein	27
group_3182	Conjugal transfer protein TrbH	27
group_3183	Type IV secretory pathway, VirB3-like protein	27
group_3184	TrbC/VIRB2 family protein	27
group_2173	TrbL/VirB6 plasmid conjugal transfer protein	26
tetM	Tetracycline resistance protein TetM from transposon TnFO1	17
group_4830	Type IV secretory system Conjugative DNA transfer	1
Int-Tn	Transposase from transposon Tn916	1
group_5004	Excisionase from transposon Tn916	1

group_132	ISXO2-like transposase domain protein	1
group_2115	Transposase IS116/IS110/IS902 family protein	1
group_2116	Transposase IS116/IS110/IS902 family protein	1
group_3432	Integrase core domain protein	4
group_1002	Conjugative relaxosome accessory transposon protein	3

Table 2.S8. Restriction modification genes identified in pan-genome of *N. gonorrhoeae*

Roary ID	Prokka Annotation	GGI+ Frequency	GGI- Frequency
bcgIA	Restriction enzyme BgcI subunit alpha	1	1
bcgIB	Restriction enzyme BgcI subunit beta	1	1
dpnC	Type-2 restriction enzyme DpnI	0.99	0.97
group_1086	NgoBV restriction endonuclease	0	0.03
group_1087	NgoBV restriction endonuclease	0.38	0.26
group_1088	NgoBV restriction endonuclease	0.01	0
group_1089	NgoBV restriction endonuclease	0.99	0.97
group_1737	HaeII restriction endonuclease	0.99	0.99
group_1738	HaeII restriction endonuclease	0.02	0.01
group_1845	putative type I restriction enzymeP M protein	1	1
group_2077	NgoPII restriction endonuclease	1	1
group_2078	NgoPII restriction endonuclease	0.01	0
group_2079	NgoPII restriction endonuclease	0	0.01
group_2080	NgoPII restriction endonuclease	0.4	0.51
group_2911	Type I restriction modification DNA specificity domain protein	0.99	1
group_2918	Type III restriction enzyme, res subunit	0.97	1
group_2974	Eco29kI restriction endonuclease	0.96	0.94
group_3828	Type I restriction modification DNA specificity domain protein	0.01	0
group_408	putative type I restriction enzymeP M protein	0.05	0.03
group_413	putative type I restriction enzymeP M protein	0.03	0
group_415	Putative type-1 restriction enzyme specificity protein MPN_089	0.73	0.71
group_416	Putative type-1 restriction enzyme specificity protein MPN_089	0.01	0.01
group_417	Putative type-1 restriction enzyme specificity protein MPN_089	0.35	0.24
group_418	Putative type-1 restriction enzyme specificity protein MPN_089	0.47	0.43
group_419	Putative type-1 restriction enzyme specificity protein MPN_089	0.21	0.22
group_420	Putative type-1 restriction enzyme specificity protein MPN_089	0.01	0
group_4272	putative type I restriction enzymeP M protein	0	0.01
group_4375	Type III restriction enzyme, res subunit	0	0.01
group_4434	Type I restriction enzyme EcoR124II R protein	0	0.01

group_4435	Putative type-1 restriction enzyme specificity protein MPN_089	0	0.01
group_4436	Putative type-1 restriction enzyme specificity protein MPN_089	0	0.01
group_4438	putative type I restriction enzymeP M protein	0	0.01
group_513	Type I restriction enzyme EcoR124II R protein	0.23	0.12
group_514	Type I restriction enzyme EcoR124II R protein	0.04	0.07
group_5402	NgoFVII restriction endonuclease	1	1
group_649	Type III restriction enzyme, res subunit	0.91	0.96
group_650	Type III restriction enzyme, res subunit	0.01	0
group_652	Type III restriction enzyme, res subunit	0.08	0.04
group_815	Type-1 restriction enzyme R protein	0.01	0.01
group_816	Type-1 restriction enzyme R protein	0.11	0.04
group_817	Type-1 restriction enzyme R protein	0.1	0.07
group_818	Type-1 restriction enzyme R protein	0.01	0
hsdM	putative type I restriction enzymeP M protein	1	0.96
hsdR_1	Type I restriction enzyme EcoR124II R protein	1	0.96
hsdR_2	Type I restriction enzyme EcoR124II R protein	0.8	0.9
hsdS	Type-1 restriction enzyme EcoKI specificity protein	0	0.01
ngoMIVR	Type-2 restriction enzyme NgoMIV	1	1
ngoMIVR_2	Type-2 restriction enzyme NgoMIV	0.03	0
ydiO	putative BsuMI modification methylase subunit YdiO	0.98	1

Table 2.S9. Toxin-antitoxin genes identified in pan-genome of *N. gonorrhoeae*

Roary ID	Prokka Annotation	GGI+ Frequency	GGI- Frequency
dinJ	Antitoxin DinJ	0	0.01
fitA	Antitoxin FitA	1	1
fitB	Toxin FitB	1	1
group_2681	Zeta toxin	0.08	0.26
group_4511	Zeta toxin	0	0.01
mazE	Antitoxin MazE	0.99	1
sinR	antitoxin HipB	1	1
vapD	Endoribonuclease VapD	0.87	0.87
vapD_2	Endoribonuclease VapD	0.08	0.13
ywqK	Putative antitoxin YwqK	1	1

Chapter 3: Effects of Host, Sample, and *in vitro* Culture on Genomic Diversity of Pathogenic Mycobacteria

Published as:

Abigail C. Shockey, Jesse Dabney and Caitlin S Pepperell

Frontiers in Genetics, June 2019, doi:10.3389/fgene.2019.00477

Author Contributions:

CSP conceived of the study. JD performed the sample preparation and processing. AS, JD and CSP designed the analyses, analyzed the data and drafted the manuscript.

Abstract

Mycobacterium tuberculosis (*M. tb*), an obligate human pathogen and the etiological agent of tuberculosis (TB), remains a major threat to global public health. Comparative genomics has been invaluable for monitoring the emergence and spread of TB and for gaining insight into adaptation of *M. tb*. Most genomic studies of *M. tb* are based on single bacterial isolates that have been cultured for several weeks *in vitro*. However, in its natural human host, *M. tb* comprises complex, in some cases massive bacterial populations that diversify over the course of infection and cannot be wholly represented by a single genome. Recently, enrichment via hybridization capture has been used as a rapid diagnostic tool for TB, circumventing culturing protocols and enabling the recovery of *M. tb* genomes directly from sputum. This method has further applicability to the study of *M. tb* adaptation, as it enables a higher resolution and more direct analysis of *M. tb* genetic diversity within hosts with TB. Here we analyzed genomic material from *M. tb* and *M. bovis* populations captured directly from sputum and from cultured samples using metagenomic and Pool-Seq approaches. We identified effects of sampling, patient, and sample type on bacterial genetic diversity. Bacterial genetic diversity was more variable and on average higher in sputum than in culture samples, suggesting that manipulation in the laboratory reshapes the bacterial population. Using outlier analyses, we identified candidate bacterial genetic loci mediating adaptation to these distinct environments. The study of *M. tb* in its natural human host is a powerful tool for illuminating host pathogen interactions and understanding the bacterial genetic underpinnings of virulence.

Introduction

Tuberculosis (TB) is the leading cause of death worldwide due to an infectious disease (World Health Organization, 2018). Among the tools brought to bear to understand and tackle the TB pandemic, comparative genomics has received increased attention following the

development of affordable, high throughput sequencing technologies. For example, comparative genomic methods have been used to investigate the spread of TB at regional [e.g. (2)] and global [e.g. (3)] scales and to identify drug resistance loci [e.g. (4–6)].

As a first step to performing analyses of *Mycobacterium tuberculosis* (*M. tb*) whole genome sequence data, isolation of bacterial DNA typically begins with decongestion of putatively infected sputum and transfer to artificial media. Sputum samples that harbor *M. tb* are then enriched for *M. tb* through growth in axenic culture for multiple weeks before DNA extraction can be performed (Nor, Ng, and Fong 2009). Although routine, the microevolutionary dynamics of this process are not well characterized.

Recent methodological advances have enabled enrichment of target DNA molecules from within complex backgrounds. DNA enrichment *via* hybridization capture has become a standard procedure for recovering genomic regions of interest from genetically homogenous mixtures, and even full genomes from complex metagenomic backgrounds (7–9). Indeed, enrichment via hybridization capture has recently been investigated as a rapid TB diagnostic, circumventing lengthy culturing procedures and enabling the recovery of *M. tb* genomes directly from infected sputum (10, 11).

In addition to TB diagnostics, the ability to recover genomes directly from infected tissues has important implications for the field of *M. tb* comparative genomics. Most studies have relied on 1:1 comparisons of representative genomes from single bacterial strains isolated via axenic culture. However, *M. tb* populations within hosts are composed of potentially billions of bacterial cells that diversify over the course of infection and cannot be wholly represented by a single genome (12, 13). Capturing bacterial genomic material directly from sputum enables a more direct analysis of *M. tb* genetic diversity during infection.

Here we used metagenomic and Pool-Seq approaches to compare genome-wide sequence data from *M. tb* and *M. bovis* populations isolated from paired sputum and culture samples. Our results suggest genetic diversity is reshaped during *in vitro* culture of bacterial

populations and we propose candidate loci mediating differential adaptation to these distinct environments.

Methods

Sample collection

We analyzed whole genome sequence data from a previously published study in which *M. tuberculosis* DNA was captured directly from infected sputum samples (accession code PRJEB9206) (10). TB treatment data were not provided for samples included in Brown et al; according to the text, at least some of the patients had received TB treatment prior to sample collection. We obtained an additional five residual sputum samples from the Wisconsin State Laboratory of Hygiene. Four of these samples were taken from a single TB patient over a 48-hour period. These were the first samples collected from this patient. The presence of *Mycobacterium bovis* (*M. bovis*) in these four had been confirmed through positive MGIT cultures. The fifth sputum sample was a pool taken from numerous TB negative patients.

Sample preparation and sequencing

Two 500 ul aliquots were taken from each positive sputum sample. One aliquot was used directly for DNA extraction, while the second aliquot was used for inoculation into 10ml of Middlebrook 7H11 broth in T-25 culture flasks. Cultures were incubated at 37°C for 3 weeks or longer until growth was visible.

DNA extractions from all samples, including a 500 ul aliquot from the negative sputum sample, were performed following the protocol described in Brown et al, with some modifications. Briefly, the 500 ul aliquots of decongested sputum, and 5 ml of culture were spun down for 5 minutes at maximum speed on a benchtop centrifuge to pellet cells. The supernatant was discarded, and the sediment then resuspended in 300ul TE buffer and transferred to 2ml tubes with 250 ug of 0.1mm glass beads. Samples were incubated at 80°C for 50 minutes, and then frozen at -80°C overnight. After thawing, tubes were vortexed for 3 minutes and spun

down, followed by the addition of 10 ul Mutanolysin and 1-hour incubation at 37C. Following incubation, samples were centrifuged at max speed on a benchtop centrifuge and purified with the DNeasy Blood and Tissue kit (Qiagen) and eluted in 100 ul volumes. All extractions were performed in a BSL-3 laboratory.

For the sputum samples, 50 ul of each extract was sheared using the Covaris M220 on 250bp setting. 10 ul of sheared DNA was then used as input for sequencing library preparation using the NEBNext Ultra II kit (New England Biolabs) following manufacturer's instructions. 1:100 dilutions of each library were quantified via qPCR using Maxima master mix (ThermoFisher). Double indexes (NEB) were then added using AccuPrime Pfx polymerase (ThermoFisher) and the following PCR heating profile: 2min at 95°C, 15 cycles of 20s at 95°C, 30s at 65°C and 1min20s at 68°C, followed by 5min at 68°C. PCR reactions were purified using MinElute spin columns (Qiagen). 1 ul of each resulting indexed library was then run through 1 cycle of PCR to remove heteroduplices and purified with MinElute columns. Samples were pooled in equal volumes and negative controls in 1:10 volumes. This pool was quantified on a BioAnalyzer using a DNA 1000 chip and sequenced on 1 lane of a 2x125bp run on a HiSeq 2500.

The culture samples were prepared according the TruSeq Nano DNA LT Library Prep Kit (Illumina Inc., San Diego, California, USA) with minor modifications. Samples were sheared using a Covaris M220 Ultrasonicator (Covaris Inc, Woburn, MA, USA), and were size selected for an average insert size of 550 bp using SPRI bead-based size exclusion. Quality and quantity of the finished libraries were assessed using an Agilent DNA1000 chip and Qubit® dsDNA HS Assay Kit, respectively. Libraries were standardized to 2nM.

Sequence processing

We processed fastq files for all samples using our reference guided assembly pipeline (https://github.com/pepperell-lab/RGAPepPipe). Briefly, adapters and low-quality bases were trimmed using Trim Galore! (https://github.com/FelixKrueger/TrimGalore) and aligned to either the *M. tb* H37Rv (Brown et al. 2015 samples and the negative sputum sample) or *M. bovis*

AF2122 (samples from the Wisconsin State Laboratory of Hygiene) reference genomes using BWA mem (14, 15). SAM files were converted to bam format and sorted using Samtools followed by duplicate removal with Picard (https://broadinstitute.github.io/picard/) and local realignment with GATK (16, 17).

Aligned sequences were taxonomically classified using Kraken and the RefSeq bacterial, viral and archaea databases as implemented in Kraken's standard database build (18, 19). Paired-end sequences where one or both sequences were not assigned to the *Mycobacterium* genus or lower were removed from the aligned sequences using Picard. Indels, repetitive regions (including PE/PPE genes), mobile elements, as well as rRNA and tRNA genes were removed from samples using PoPoolation2 (20) (Table 3.S1). We identified and removed indels present in each sample using PoPoolation (21). We used VCF files generated with Samtools to identify strand-bias positions in each sample, which were removed across all samples.

Estimates of nucleotide diversity

We estimated genetic diversity for each sample independently using the Pool-seq approach implemented in PoPoolation2. Following O'Neill et al. 2015, we randomly subsampled (n = 10) read data from each sample to a uniform 50X coverage to limit the effects of differential coverage across samples. Using these subsampled data with uniform coverage, we then calculated nucleotide diversity (π), Watterson's theta (θ_w) and Tajima's D in 100 kb sliding windows across the genome in 10 kb steps (12). Additionally, we calculated π and θ_w for each gene using gene annotations based on the *M. tb* H37Rv and *M. bovis* AF2122 reference genomes. Following recommendations and rationale described in O'Neill et al, we required at least 50% coverage of each region and a minimum allele count of 2. Pool-size was set at 10,000. We calculated the genome-wide averages of nucleotide diversity in sputum and culture samples for each patient as the mean of the sliding windows of diversity (100 kb windows, 10 kb

steps). We performed a paired t-test on these genome-wide values of nucleotide diversity in sputum and culture for each patient.

Identification of windows of overlap in nucleotide diversity

We identified regional peaks in nucleotide diversity across the genome. Using nucleotide diversity from the sliding-window analysis, we calculated a z-score and p-value for each window in sputum and culture for each patient. We performed FDR correction, setting a p-value cutoff of 0.05. Windows were defined as overlapping if they were present in > 1 patient. Code available on https://github.com/.

Identification of outlier genes

We identified genes with significant changes in nucleotide diversity (π) from sputum to culture in each patient using three different approaches. Method 1: we performed linear regression of nucleotide diversity per gene in sputum versus culture for each patient (22). We calculated Cook's distance (D_i) from the regression line for each gene and used a threshold of > 4 times the mean of D_i to define outlier genes in each patient. Method 2: for each patient and each gene we calculated the fold-change in nucleotide diversity between sputum and culture samples (i.e. nucleotide diversity in sputum/nucleotide diversity in culture). We performed z-transformation of these values and calculated a p-value for each gene. For genes with non-zero diversity in sputum and zero diversity in culture, we calculated a z-score and p-value for the difference in nucleotide diversity between these sample types. We used FDR correction for multiple testing, setting a cutoff of 0.05 to identify outliers. Method 3: treating sputum and culture pairs as two different populations, we calculated F_{ST} per gene using PoPoolation2. We required a minimum allele count of 3, minimum coverage of 10 and maximum coverage of 350. Pool-size was set at 10,000. Genes that were masked (either insufficient coverage or within the bounds of removed regions described in Table 3.S1) in the gene-wise estimates of nucleotide diversity described above were excluded from these analyses. We performed a Fisher's exact test with

FDR correction to assess significance for F_{ST} values from each gene, setting a p-value cutoff of 0.01. Code available on https://github.com/AbigailShockey/sputum.

Lineage Typing

We used SNP-IT (https://github.com/samlipworth/snpit) to perform lineage typing for our sample of *M. tb*. Briefly, we used bcftools to call consensus sequences from our sputum and culture samples of *M. tb*. We required a minimum read and mapping quality of 20. We masked indels, repetitive regions (including PE/PPE genes), mobile elements, as well as rRNA and tRNA genes in the consensus sequences using in-house scripts (Table 3.S1). We performed lineage typing on the masked consensus sequences.

Identification of Mixed Infections

In order to investigate the possibility of mixed infection, we looked for overlap between sites defined as variable in our analyses and lineage-defining positions from (23). Of the 6,915 positions proposed by Coll et al., 47 were masked in our analyses due to not meeting quality control thresholds. We did not observe consistent patterns of variation at the remaining 6,868 positions to suggest that the samples derived from infections that contained mixtures of lineages (Table 3.S2).

Data availability

The *M. tuberculosis* sequence data from Brown et al 2015 is publicly available in the Sequence Read Archive under BioProject Accession Code PRJEB9206. The *M. bovis* sequence data will be submitted to the Sequence Read Archive.

Statement of Ethics

Newly sequenced data in this study was obtained from residual clinical samples at the State Lab of Hygiene. We did not collect any data or samples for research purposes nor was routine clinical care altered by this study. This study was reviewed and approved by the UW-Madison Health Sciences Institutional Review Board.

Results

Removal of putative contaminating sequences with metagenomic filtering

To address the possibility of background contamination with high sequence similarity to the *M. tb* H37Rv or *M. bovis* AF2122 reference genomes, we performed metagenomic filtering on aligned reads from all samples. We used Kraken (18) to assign each read to a taxon and removed reads not assigned to the *Mycobacterium* genus or a species within it.

Between 8-99% of aligned sequences were removed from the sputum samples, with 9 of the 35 samples losing more than 50% of aligned sequences (Table 3.S3). The *M. tb* sputum samples from Brown *et al.* were published with associated smear scores ranging from negative to 3+ (10). The percent of sequences remaining after filtering increased with smear score (Figure 3.1). This increase was significant for samples with a smear score ≥ 1+ when compared to samples with a negative sputum score (ANOVA, p-value < 0.01; Table 3.S4) suggesting that some of the variation in the number of sequences removed can be attributed to the severity of infection. However, the variation in the *M. bovis* sputum samples taken over a 48-hour period (1-31% sequences retained after filtering) indicate the degree of stochasticity when sampling repeatedly from a single patient (Table 3.S3).

We applied this filter to culture samples, which allowed us to assess the stringency of this step, as these samples should have minimal contaminating sequences. In samples with > 50X starting coverage, less than 3% of aligned sequences were removed by the filter. The filtered sequences likely arise from low levels of contamination or sequences in conserved regions that can't be confidently assigned (Table 3.S3).

Similarly, we applied the filter to a pool of negative sputum from patients without TB. Since there should be no sequences belonging to *Mycobacterium* genus in this sample, any sequences carried through must come from background contamination not detecTable 3.by alignment or the metagenomic filter. Only 0.3% of starting sequences could be aligned to either the *M. tb* or *M. bovis* reference genome. From those, 98% were removed in the metagenomic

filter step, indicating that the filter, together with alignment, is efficient at removing potential contaminating sequences contributed by metagenomic background (Table 3.S3).

In conjunction with the metagenomic filter, we also removed indels, repetitive regions (including PE/PPE genes), rRNA and tRNA genes, and mobile elements (Table 3.S1). Together these filters lead to an average reduction in genome wide coverage of 14% in culture samples, and 34% in sputum samples (Table 3.S2). Only samples with 50X or greater final coverage were included in subsequent analyses.

Effect of sampling and sample type on bacterial genetic diversity

Diversity of *M. tb* samples varies among patients (Figure 3.2). We did not find any evidence to suggest this was driven by bacterial lineage (Table 3.S3). Nucleotide diversity is more variable among sputum samples, where genome-wide values span an order of magnitude, whereas culture samples are more homogenous. The distribution of pairwise differences among samples from the same patient is nested within the distribution for differences between patients (Figure 3.3). This is consistent with a substantial impact of sample to sample variation on bacterial genetic diversity, similar to the observed effect of sampling on the amount of target sequencing data recovered (Table 3.S3). The distribution of windows of nucleotide diversity (π) across the genome varied across comparisons from the same patient, further reflecting the effects of sampling (ANOVA p-value < 0.05 for sample 1 to 3 & sample 2 to 3 comparison; NS for comparison of sample 1 & 2).

Despite inter- and intra-patient variability, there is a consistent pattern of greater diversity in sputum versus culture: genome-wide π is significantly greater in sputum (paired t-test, p-value = 0.029, 0.028 for inter-patient and intra-patient samples, respectively; Figure 3.4). This is indicative of a systematic loss of diversity during growth in culture.

Values of θw are generally higher than π in all samples, indicating an abundance of low frequency variants (Figures 3.S1 & 3.S2). Tajima's D is uniformly low across the genome (Figure 3.S3). Average Tajiima's D values are lower in sputum samples for all but one patient.

It is possible that background contamination not removed during alignment or metagenomics filtering contributed to observed differences in diversity between sputum and culture. Homologous sequences from non-mycobacteria present in the lungs or respiratory tract would not be present in culture and could artificially inflate nucleotide diversity in sputum samples. To address this problem, we sequenced a pool of sputum from multiple patients not displaying symptoms of TB and processed these sequences identically to infected samples. Although 99.9% of sequences were removed, approximately 3,000 sequences passed through our filters. We merged these sequences with those from Patient 14's culture sample and calculated nucleotide diversity (π) in sliding-windows as described above. Average π in this composite sample was slightly higher than the culture sample alone, but less than the paired sputum sample (π = 2.69e-05, 6.74e-05, and 1.47e-04 in culture, composite, and sputum, respectively). Increases in nucleotide diversity in the composite sample did not mirror the topology of Patient 14's sputum sample (Figure 3.S4). These findings indicate minimal background contamination passes our filters, and this contamination does not drive the patterns of nucleotide diversity seen in sputum samples.

Regional patterns of diversity

To identify regional peaks in diversity across the genome, we calculated a z-score and p-value for π per window in sputum and culture for each patient. We identified 35 regions of overlapping high π (i.e. present in > 1 patient) between culture samples from different patients, and 34 windows of overlap among sputum samples. There are 17 windows found in multiple patients in both sputum and culture. These windows correspond to two genomic regions (Table 3.S5). From the intra-patient samples, there were five windows of overlap in sputum and 12 in culture. No windows were shared between culture and sputum samples from the same patient. To assess whether changes in diversity between sputum and culture samples occur in specific genomic regions, we calculated the fold-change across the genome as the ratio of π in sputum to π in culture in sliding windows across the genome. For four patients, the diversity of culture

and sputum samples was similar. Patterns of diversity in the other patients did not reveal any obvious “hotspot regions” across patients or samples in which culture and sputum exhibited consistent differences (Figure 3.S5).

Patterns of variation at the individual gene level

We categorized each gene based on differences in nucleotide diversity between sputum and culture. The majority of genes in each patient maintained zero diversity or decreased in diversity (Fig 5, 54% and 32% of total gene content, respectively). Among genes that decreased in diversity, the majority lost all diversity in culture. As with the findings described above, these results suggest a significant loss in bacterial diversity occurs following growth in culture.

To identify specific genes with marked changes in diversity between sputum and culture, we performed linear regression of gene diversity in the two sample types, for each patient and sample. We identified 49 outlier genes in *M. tb*, 17 of which were found across more than one patient (Figure 3.6, Table 3.1, Table 3.S6). Remarkably, Rv2020c was an outlier in all 13 patients. As an alternate method of identifying genes with significant differences in diversity, we calculated the fold-change in nucleotide diversity (π sputum/π culture) per gene in each patient and sample. Fold changes vary among intra- and inter-patient samples, and the fold changes from sputum to culture can span orders of magnitude (Figure 3.7). We calculated a z-score and p-value for the fold change per gene in each patient and identified three *M. tb* genes with significant fold change in > 1 patient; an additional 71 genes had an extreme fold change in a single patient (Figure 3.7, Table 3.1, Table 3.S6).

Of these outliers, *nrdE* has the highest fold-change in 4 patients and *rpoB (*Rv0667) in two patients. *nrdE* is seen in the extremes of the fold-change distributions in a total of 6 patients, and *rpoB* in two. Although it is only found in the extremes of diversity in a single patient, *plcC* is the only gene with higher diversity in sputum in all 13 TB patients.

To assess genes with high diversity in sputum that have zero diversity in culture, which aren’t amenable to fold-change calculation, we examined differences rather than fold-changes.

For genes with non-zero diversity in sputum and zero diversity in culture we calculated a z-score and p-value for each difference in nucleotide diversity (π) and identified 127 *M. tb* genes with significant differences. Fifteen of these were found across more than one patient (Table 3.1, Table 3.S6).

As another method of identifying genes with major changes in diversity, we calculated F_{ST} per gene treating sputum and culture as two populations. We used Fisher's exact test (with FDR correction) to assess the significance of F_{ST} per gene and found 63 *M. tb* genes to have significant differences in more than one patient; an additional 292 genes were outliers in a single patient (Table 3.1, Table 3.S6, Figure 3.S6).

Discussion

Although *Mycobacterium tuberculosis* can be grown axenically in the lab, its natural niche is within human tissues. Here we compared patterns of genetic diversity of *M. tb* in expectorated sputum to bacteria grown *in vitro*, in order to gain insight on differences between evolutionary pressures encountered within the host and those imposed by *ex vivo* manipulation of bacterial populations. It's important to understand bacterial adaptation to both settings, as the former is informative of host-pathogen interactions and the latter is vital in distinguishing signal from noise in bacterial sequencing data. We found that diversity of *M. tb* in sputum samples varies substantially within and among hosts, and that diversity of these populations is higher than it is for *M. tb* grown *in vitro*. Using outlier analyses, we further identified a group of genes that exhibit consistent shifts in diversity between culture and sputum. These are candidate loci mediating differential adaptation to the two environments.

We found overall diversity of *M. tb* populations to be higher in sputum samples than in culture (Figures 2 & 4). At a gene by gene level, the most common pattern observed was for genes with measurable diversity in sputum to lose all diversity in culture (Figure 3.5). This pattern could arise from bacterial population bottlenecks that occur during processing of sputum

samples for *in vitro* culture. Sputum and culture samples from Brown *et al* were produced from different input volumes of initial suspension [1900ul and 100ul respectively (10)]. We controlled for this potential bias in the processing of samples from patient 35, using equal volumes of initial suspension for direct DNA extraction and inoculation into culture media. The results from patient 35 mirror those of the Brown samples, with greater diversity of *M. tb* in sputum than in culture (Figures 2 & 4). This suggests that the reduction in diversity observed in culture samples is not an artifact of sample processing.

An alternate explanation of the observed difference between sputum and culture samples is that apparent diversity of *M. tb* in sputum samples is inflated by DNA sequences from organisms other than *M. tb*, i.e. bacteria present in the upper respiratory tract. We applied stringent filters to remove off-target sequences (*Methods*), and our analyses of uninfected sputa (Table 3.S3) and culture samples spiked with TB-negative sputa (*Results*, Figure 3.S1) showed that the patterns of *M. tb* diversity observed in sputum samples did not arise from contamination of *M. tb* sequencing data.

Explanations of the relatively high diversity in sputum include relaxed purifying selection and/or diversifying selection that is specific to this environment, mutation rate variation, and bacterial sub-populations within hosts that have variable fitness *in vitro*. The degree of differentiation between *M. tb* populations in sputum and culture varies substantially among patients (Figures 2 & 6): overall diversity of bacterial populations in the two environments is nearly identical for some patients (e.g. patient 4) and an order of magnitude different for others (e.g. patient 14). This suggests that the evolutionary pressures driving genome wide differences in diversity vary from patient to patient.

We found previously, using pooled culture-based samples, that overall diversity of within-host *M. tb* populations varies among patients and that patients with pre-terminal TB can harbor extremely diverse populations of bacteria (12). It's possible that *M. tb* populations within hosts occasionally undergo massive expansions associated with relaxation of purifying selection, and

that this becomes evident in comparisons with bacterial populations cultured under relatively uniform conditions. Pulmonary cavitation is one plausible condition under which such an expansion could occur: cavitation results in a shift from a hypoxic to an oxygen-rich environment and the interior of the cavity is relatively inaccessible to the immune system. Trauner *et al* reported an observable shift in *M. tb* population structure following cavitation of a large granuloma (13), demonstrating that *M. tb* within-host population diversity reflects the ongoing evolution of disease in the host. Clinical metadata from the patients whose samples we analyzed here do not point to any obvious reasons for observed differences in *M. tb* sputum diversity [e.g. patient 4, with low diversity, has 3+ smear positivity and MDR TB and patient 14, with high diversity, has 1+ smear and MDR TB; (10)], but these data are limited.

Host immune responses impose a range of stresses on *M. tb* populations, including DNA damage [reviewed in (24, 25)]. Host imposed mutagenic stressors are likely to vary over time and among patients, and high *M. tb* sputum diversity could reflect more mutagenic environments within certain hosts (and host states) versus the relatively uniform conditions of *in vitro* culture. Relatively high diversifying selection is an alternative explanation for high *M. tb* sputum diversity within a subset of TB patients. However, given that the pattern of elevated diversity is genome-wide (Figure 3.S2), this seems less likely than relaxed purifying selection and/or variation in within-host mutation rates.

Beyond its potential instructiveness about the varied adaptive milieu within hosts with TB, the uneven accumulation of *M. tb* genetic diversity across TB patients has implications for the reconstruction of TB transmission networks from bacterial genetic data. *M. tb* genetic distances have been used as evidence of epidemiological links among TB patients and method development is active in this area [e.g. (26)]. Our finding here and in prior published work that *M. tb* diversity varies dramatically within patients with TB implies that epidemiological links can be obscured in pathogen genetic data. In a recently published study comparing *M. tb* outbreak strains with endemically circulating strains, we found evidence suggesting that bacterial

diversification is uneven, characterized by long periods of stasis and punctuated bursts (27). This pattern could arise from occasional, exceptionally large bacterial population expansions and/or mutation rate variation within hosts.

Our finding of increased *M. tb* diversity in sputum relative to culture is consistent with results of other studies using capture based methods (11, 28). Votinsteva et al, who used shotgun sequencing to compare *M. tb* in sputum and culture, did not identify a difference in overall diversity between these sample types (29). The shotgun and capture-based studies are not directly comparable, as shotgun sequencing is less sensitive and was applied to smear positive samples only. In addition, coverage was inadequate to allow diversity to be estimated for several of the samples in Votintseva et al.

Variant calling and quantification of diversity was also performed differently across studies. Nimmo et al estimated the number of heterozygous sites, as did Votintsteva et al, but Votinsteva et al used a distinct variant calling method and restricted their analysis to a subset of 68,695 loci at which they had previously identified segregating polymorphisms in a large sample of clinical isolates. Culture-based studies of intra-host diversity suggest that most *M. tb* variants are segregating at rare frequencies (12, 13), which parallels findings at the between-host scale (30). Our findings here also suggest that within-host diversity is skewed to rare variation, and that this skew is more pronounced in sputum than in culture (Figure 3.S4). There is no *a priori* reason to expect that the same rare mutations will be encountered in individual clinical isolates, culture-based surveys of within-host diversity, and clinical samples. Based on these observations, we posit that restricting the estimation of *M. tb* diversity in sputum to loci at which variants were observed in clinical isolates is likely to result in an underestimate of the amount of bacterial variation present in sputum.

Results from several studies suggest that the *M. tb* population within hosts is structured into genetically distinct sub-populations (13, 31–33). Consistent with these prior studies, our

results here demonstrate sample to sample variation in sputa collected from a single patient (Table 3.S2, Figure 3.3).

Published data demonstrate that sputum from TB patients contains phenotypically distinct sub-populations of *M. tb* and that these phenotypes are not recovered during *in vitro* culture (34). *In vitro* culture of mixtures of genetically and phenotypically distinct *M. tb* has been shown to result in a loss of diversity (35–37) and *M. tb* adaptation to laboratory conditions is a well described phenomenon (38–42). Taken together, these findings show that the population of *M. tb* within hosts is genetically and phenotypically diverse, and that *in vitro* culture imposes distinct evolutionary pressures on *M. tb* that reshape the bacterial population. It follows that the full diversity of *M. tb* found in sputum is unlikely to survive the transition to growth *in vitro*; this offers a complementary/ alternative explanation of observed differences in *M. tb* genetic diversity between sputum and culture.

In order to gain insight on evolutionary pressures in sputum and culture, we performed outlier analyses of gene-wise patterns of variation (Figures 6 & 7; Table 3.1). We identified two major groups of outlier genes. The first group, typified by Rv2020c (encoding a conserved hypothetical protein), exhibited high diversity in both sputum and culture without significant differences between environments (Figure 3.6). Genes with a similar pattern include two predicted adenylate cyclases (Rv1318c & 1319c), molybdenum cofactor biosynthesis protein *moaA1* (Rv3109), transcriptional regulatory protein *embR* (Rv1267c), membrane-associated phospholipases C1 & C2 (*plcB*/Rv2350c & *plcA*/Rv2351c), Rv2081c (conserved transmembrane protein) and Rv2082 (conserved hypothetical). We previously found Rv2020c to be in the 99th percentile of diversity in a sample of 201 globally extant strains of *M. tb* (12). Several other genes in this group exhibited similarly high diversity in our previous study: *plcA*/Rv2351c, Rv1319c, Rv2081c and Rv2082 were also in the 99th percentile of gene-wise diversity, whereas *plcB*/Rv2350c was in the 81st and *moaA1*/Rv3109 in the 87th percentile of gene-wise diversity in the global sample. With the exception of *moaA1*/Rv3109, for which data are conflicting, none of

the genes in this grouping is essential for growth *in vitro* (43–45). Deletions affecting *plcA*/Rv2351c have been identified in clinical *M. tb* isolates, suggesting its function is dispensable in certain settings and/or genetic backgrounds (46). Collectively, these results suggest the genes are under relaxed purifying selection or diversifying selection.

Interestingly, for three of the six genes in this group (*plcA*/Rv2351c, *plcB*/Rv2350c, Rv2020c), growth *in vitro* is actually enhanced when the gene is disrupted by transposon insertion (45); this is also true of *plcC*/Rv2349c, which was not an outlier but exhibited consistent differences between sputum and culture. Gene expression studies suggest that *M. tb* in sputum are in a slowly replicating or nonreplicating state relative to *M. tb* in culture (34, 47, 48). Non replicating persistence is likely adaptive, as *M. tb* in this physiological state is able to survive a wide range of stressors and becomes progressively enriched in the sputa of TB patients (47, 49, 50). A recent, detailed investigation of persistent *M. tb* in sputum identified several distinct sub-populations of bacteria, suggesting that selection for this trait maintains diversity in natural populations of *M. tb* (51). We found previously that a subset of positively selected loci in *M. tb* are characterized by high diversity and numerous rare mutations; we referred to these loci as “sloppy targets” (6). Here we propose that Rv2020c, Rv1318c, Rv1319c, Rv2081c, Rv2082, *moaA1*/Rv3100g, *embR*/Rv1267c, *plcA*/Rv2351c and *plcB*/Rv2350c are sloppy targets. Of note, similar to *plcA*/Rv2351c, the canonical sloppy target *pncA* is deleted in commonly circulating sub-lineages of *M. tb* (52, 53). The nine putative sloppy targets identified in this study were all FST outliers in multiple patients, indicating that while these genes are similarly diverse in sputum and culture, variants within them differ between environments (Table 3.1). This is consistent with positive selection in at least one of these environments as an explanation of high diversity, as opposed to global relaxation of purifying selection. Selection for persistence is a possible example of differential selective pressure in sputum and culture as this trait is unlikely to be adaptive during growth in antibiotic free media.

We identified a second group of genes, typified by ribonucleoside diphosphate reductase *nrdE* (Rv3051c), characterized by marked changes in diversity between sputum and culture. Genes in this group, which we will hereafter refer to as “shifting targets”, include RNA polymerases *rpoB* (Rv0667) and *rpoC* (Rv0668), elongation factor *fusA1* (Rv0684), ribosomal protein *rpsA* (Rv1630), iron sulfur binding reductase Rv0338c, respiratory nitrate reductase *narI* (Rv1164), and maltosyltransferase *glgE* (Rv1327c). Genes in this grouping are annotated as either intermediary metabolism (n=3) or information pathways (n=5); all but one (*narI*/Rv1164) is essential for *in vitro* growth. *RpoB* and *rpoC* are known to mediate resistance to rifamycins, which are first line TB treatments; as expected, signatures of positive selection have been identified previously at these loci (6, 54). TB treatment details were not provided for the samples included in Brown et al, but at least some of the patients included in the study and analyzed here had been treated previously (10). We clearly expect selection pressures on drug resistance loci to shift between sputum and culture in antibiotic free media, and thus the identification of *rpoB* and *rpoC* provides support for the use of our outlier method to identify genes under differential selection pressures *in vivo* and *in vitro*. As with the putative sloppy targets, the eight genes listed above were F_{ST} outliers across multiple patients (Table 3.1), further supporting the idea that they are under distinct selection pressures in the two environments. Of note, seven of eight genes in this group (*glgE*/Rv1327c is the exception) appear to be expressed differently in sputum versus culture (34, 47, 48, 55). As described above, broad patterns of gene expression suggest that the shift of *M. tb* from sputum to culture involves an increase in metabolic activity and replication. We hypothesize that shifting targets are under relatively strong purifying selection *in vitro*, as bacteria compete in an environment in which it is no longer advantageous to suspend growth. This transition to relatively strong purifying selection is expected to result in a decrease in diversity during culture, as observed here.

In this analysis of *M. tb* and *M. bovis* genomic data recovered directly from sputum and from cultured samples, we identified intra- and inter-patient variability, as well as an effect of

sample type on bacterial genetic diversity. We hypothesize that this variability reflects differences in the milieu within hosts, the nature of host pathogen interactions, and the distinct evolutionary pressures experienced by these bacteria in natural and laboratory environments.

Acknowledgements

We would like to thank Julie Tans-Kersten and the staff at the Wisconsin State Laboratory of Hygiene for curating the residual sputum specimens used in this study. We also thank Mary O'Neill (Institut Pasteur) and members of the Pepperell Lab for their input on this study. AS and CP are supported by the National Institutes of Health (R01AI113287). This work was supported in part by NIH/NHGRI training (Grant No. T32 HG002760). Funding for this project was also provided by the University of Wisconsin School of Medicine and Public Health from the Wisconsin Partnership Program.

References

1. World Health Organization. 2018. WHO | Global tuberculosis report 2018. WHO.
2. Martin MA, Lee RS, Cowley LA, Gardy JL, Hanage WP. 2018. Within-host Mycobacterium tuberculosis diversity and its utility for inferences of transmission. Microb Genomics.
3. O'Neill MB, Shockey AC, Zarley A, Aylward W, Eldholm V, Kitchen A, Pepperell CS. 2018. Lineage specific histories of Mycobacterium tuberculosis dispersal in Africa and Eurasia. bioRxiv 210161.
4. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, Warren RM, Streicher EM, Calver A, Sloutsky A, Kaur D, Posey JE, Plikaytis B, Oggioni MR, Gardy JL, Johnston JC, Rodrigues M, Tang PKC, Kato-Maeda M, Borowsky ML, Muddukrishna B, Kreiswirth BN, Kurepina N, Galagan J, Gagneux S, Birren B, Rubin EJ, Lander ES, Sabeti PC, Murray M. 2013. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. Nat Genet advance online publication.
5. Zhang H, Li D, Zhao L, Fleming J, Lin N, Wang T, Liu Z, Li C, Galwey N, Deng J, Zhou Y, Zhu Y, Gao Y, Wang T, Wang S, Huang Y, Wang M, Zhong Q, Zhou L, Chen T, Zhou J, Yang R, Zhu G, Hang H, Zhang J, Li F, Wan K, Wang J, Zhang X-E, Bi L. 2013. Genome sequencing of

161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance. Nat Genet advance online publication.

6. Mortimer TD, Weber AM, Pepperell CS. 2018. Signatures of Selection at Drug Resistance Loci in Mycobacterium tuberculosis. mSystems 3:e00108-17.

7. Christiansen MT, Brown AC, Kundu S, Tutill HJ, Williams R, Brown JR, Holdstock J, Holland MJ, Stevenson S, Dave J, Tong CYW, Einer-Jensen K, Depledge DP, Breuer J. 2014. Whole-genome enrichment and sequencing of Chlamydia trachomatis directly from clinical samples. BMC Infect Dis 14:591–591.

8. Enk J, Devault A, Kuch M, Murgha Y, Rouillard J-M, Poinar H. 2014. Ancient whole genome enrichment using baits built from modern DNA. Mol Biol Evol msu074.

9. Clark SA, Doyle R, Lucidarme J, Borrow R, Breuer J. 2018. Targeted DNA enrichment and whole genome sequencing of Neisseria meningitidis directly from clinical specimens. Int J Med Microbiol IJMM 308:256–262.

10. Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, Chan JZM, Depledge DP, Nikolayevskyy V, Broda A, Stone MJ, Christiansen MT, Williams R, McAndrew MB, Tutill H, Brown J, Melzer M, Rosmarin C, McHugh TD, Shorten RJ, Drobniewski F, Speight G, Breuer J. 2015. Rapid Whole-Genome Sequencing of Mycobacterium tuberculosis Isolates Directly from Clinical Samples. J Clin Microbiol 53:2230–2237.

11. Doyle RM, Burgess C, Williams R, Gorton R, Booth H, Brown J, Bryant JM, Chan J, Creer D, Holdstock J, Kunst H, Lozewicz S, Platt G, Romero EY, Speight G, Tiberi S, Abubakar I, Lipman M, McHugh TD, Breuer J. 2018. Direct Whole-Genome Sequencing of Sputum Accurately Identifies Drug-Resistant Mycobacterium tuberculosis Faster than MGIT Culture Sequencing. J Clin Microbiol 56:e00666-18.

12. O'Neill MB, Mortimer TD, Pepperell CS. 2015. Diversity of Mycobacterium tuberculosis across Evolutionary Scales. PLOS Pathog 11:e1005257.

13. Trauner A, Liu Q, Via LE, Liu X, Ruan X, Liang L, Shi H, Chen Y, Wang Z, Liang R, Zhang W, Wei W, Gao J, Sun G, Brites D, England K, Zhang G, Gagneux S, Barry CE, Gao Q. 2017. The within-host population dynamics of Mycobacterium tuberculosis vary with treatment efficacy. Genome Biol 18:71.

14. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10–12.

15. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio.

16. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,

Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

17. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498.

18. Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:R46.

19. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–D745.

20. Schlötterer C, Pandey RV, Kofler R. 2011. PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics 27:3435–3436.

21. Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A, Kosiol C, Schlötterer C. 2011. PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PloS One 6:e15925–e15925.

22. Yahara K, Didelot X, Jolley KA, Kobayashi I, Maiden MCJ, Sheppard SK, Falush D. 2016. The Landscape of Realized Homologous Recombination in Pathogenic Bacteria. Mol Biol Evol 33:456–471.

23. Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, Portugal I, Pain A, Martin N, Clark TG. 2014. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. Nat Commun 5.

24. Stallings CL, Glickman MS. 2010. Is Mycobacterium tuberculosis stressed out? A critical assessment of the genetic evidence. Microbes Infect Inst Pasteur 12:1091–1101.

25. Flentie K, Garner AL, Stallings CL. 2016. Mycobacterium tuberculosis Transcription Machinery: Ready To Respond to Host Attacks. J Bacteriol 198:1360–1373.

26. Stimson J, Gardy J, Mathema B, Crudu V, Cohen T, Colijn C. 2019. Beyond the SNP

Threshold: Identifying Outbreak Clusters Using Inferred Transmissions. Mol Biol Evol 36:587–603.

27. Doroshenko A, Pepperell CS, Heffernan C, Egedahl ML, Mortimer TD, Smith TM, Bussan HE, Tyrrell GJ, Long R. 2018. Epidemiological and genomic determinants of tuberculosis outbreaks in First Nations communities in Canada. BMC Med 16:128.

28. Nimmo C, Shaw LP, Doyle R, Williams R, Brien K, Burgess C, Breuer J, Balloux F, Pym AS. 2018. Whole genome sequencing Mycobacterium tuberculosis directly from sputum identifies more genetic diversity than sequencing from culture. bioRxiv 446849.

29. Votintseva AA, Bradley P, Pankhurst L, Elias C del O, Loose M, Nilgiriwala K, Chatterjee A, Smith EG, Sanderson N, Walker TM, Morgan MR, Wyllie DH, Walker AS, Peto TEA, Crook DW, Iqbal Z. 2017. Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. J Clin Microbiol 55:1285–1298.

30. Pepperell CS, Casto AM, Kitchen A, Granka JM, Cornejo OE, Holmes EC, Birren B, Galagan J, Feldman MW. 2013. The Role of Selection in Shaping Diversity of Natural M. tuberculosis Populations. PLoS Pathog 9:e1003543.

31. Liu Q, Via LE, Luo T, Liang L, Liu X, Wu S, Shen Q, Wei W, Ruan X, Yuan X, Zhang G, Barry CE, Gao Q. 2015. Within patient microevolution of *Mycobacterium tuberculosis* correlates with heterogeneous responses to treatment. Sci Rep 5:17507.

32. Lieberman TD, Wilson D, Misra R, Xiong LL, Moodley P, Cohen T, Kishony R. 2016. Genomic diversity in autopsy samples reveals within-host dissemination of HIV-associated *Mycobacterium tuberculosis*. Nat Med 22:1470–1474.

33. Martin CJ, Cadena AM, Leung VW, Lin PL, Maiello P, Hicks N, Chase MR, Flynn JL, Fortune SM. 2017. Digitally Barcoding Mycobacterium tuberculosis Reveals In Vivo Infection Dynamics in the Macaque Model of Tuberculosis. mBio 8:e00312-17.

34. Garton NJ, Waddell SJ, Sherratt AL, Lee S-M, Smith RJ, Senner C, Hinds J, Rajakumar K, Adegbola RA, Besra GS, Butcher PD, Barer MR. 2008. Cytological and Transcript Analyses Reveal Fat and Lazy Persister-Like Bacilli in Tuberculous Sputum. PLOS Med 5:e75.

35. Martín A, Herranz M, Ruiz Serrano MJ, Bouza E, García de Viedma D. 2010. The clonal composition of Mycobacterium tuberculosis in clinical specimens could be modified by culture. Tuberculosis 90:201–207.

36. Hanekom M, Streicher EM, Berg DV de, Cox H, McDermid C, Bosman M, Pittius NCG van, Victor TC, Kidd M, Soolingen D van, Helden PD van, Warren RM. 2013. Population Structure of Mixed Mycobacterium tuberculosis Infection Is Strain Genotype and Culture Medium Dependent. PLOS ONE 8:e70178.

37. Metcalfe JZ, Streicher E, Theron G, Colman RE, Penaloza R, Allender C, Lemmer D, Warren RM, Engelthaler DM. 2017. Mycobacterium tuberculosis Subculture Results in Loss of Potentially Clinically Relevant Heteroresistance. Antimicrob Agents Chemother 61.
38. Domenech P, Reed MB. 2009. Rapid and spontaneous loss of phthiocerol dimycocerosate (PDIM) from Mycobacterium tuberculosis grown in vitro: implications for virulence studies. Microbiology 155:3532–43.
39. Ioerger TR, Feng Y, Ganesula K, Chen X, Dobos KM, Fortune S, Jacobs WR, Mizrahi V, Parish T, Rubin E, Sassetti C, Sacchettini JC. 2010. Variation among Genome Sequences of H37Rv Strains of Mycobacterium tuberculosis from Multiple Laboratories. J Bacteriol 192:3645–3653.
40. Molina-Torres CA, Castro-Garza J, Ocampo-Candiani J, Monot M, Cole ST, Vera-Cabrera L. 2010. Effect of serial subculturing on the genetic composition and cytotoxic activity of Mycobacterium tuberculosis. J Med Microbiol 59:384–391.
41. Domenech P, Rog A, Moolji J, Radomski N, Fallow A, Leon-Solis L, Bowes J, Behr MA, Reed MB. 2014. Origins of a 350-Kilobase Genomic Duplication in Mycobacterium tuberculosis and Its Impact on Virulence. Infect Immun 82:2902–2912.
42. De Majumdar S, Sikri K, Ghosh P, Jaisinghani N, Nandi M, Gandotra S, Mande S, Tyagi JS. 2019. Genome analysis identifies a spontaneous nonsense mutation in ppsD leading to attenuation of virulence in laboratory-manipulated Mycobacterium tuberculosis. BMC Genomics 20:129.
43. Sassetti CM, Boyd DH, Rubin EJ. 2003. Genes required for mycobacterial growth defined by high density mutagenesis. Mol Microbiol 48:77–84.
44. Griffin JE, Gawronski JD, Dejesus MA, Ioerger TR, Akerley BJ, Sassetti CM. 2011. High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. PLoS Pathog 7:e1002251.
45. DeJesus MA, Gerrick ER, Xu W, Park SW, Long JE, Boutte CC, Rubin EJ, Schnappinger D, Ehrt S, Fortune SM, Sassetti CM, Ioerger TR. 2017. Comprehensive Essentiality Analysis of the Mycobacterium tuberculosis Genome via Saturating Transposon Mutagenesis. mBio 8.
46. Tsolaki AG, Hirsh AE, DeRiemer K, Enciso JA, Wong MZ, Hannan M, Goguet de la Salmoniere YO, Aman K, Kato-Maeda M, Small PM. 2004. Functional and evolutionary genomics of Mycobacterium tuberculosis: insights from genomic deletions in 100 strains. Proc Natl Acad Sci U A 101:4865–70.
47. Honeyborne I, McHugh TD, Kuittinen I, Cichonska A, Evangelopoulos D, Ronacher K, van Helden PD, Gillespie SH, Fernandez-Reyes D, Walzl G, Rousu J, Butcher PD, Waddell SJ.

2016. Profiling persistent tubercule bacilli from patient sputa during therapy predicts early drug efficacy. BMC Med 14:68.

48. Sharma S, Ryndak MB, Aggarwal AN, Yadav R, Sethi S, Masih S, Laal S, Verma I. 2017. Transcriptome analysis of mycobacteria in sputum samples of pulmonary tuberculosis patients. PLOS ONE 12:e0173508.

49. Betts JC, Lukey PT, Robb LC, McAdam RA, Duncan K. 2002. Evaluation of a nutrient starvation model of Mycobacterium tuberculosis persistence by gene and protein expression profiling. Mol Microbiol 43:717–731.

50. Voskuil MI, Visconti KC, Schoolnik GK. 2004. Mycobacterium tuberculosis gene expression during adaptation to stationary phase and low-oxygen dormancy. Tuberculosis 84:218–227.

51. Jain P, Weinrick BC, Kalivoda EJ, Yang H, Munsamy V, Vilcheze C, Weisbrod TR, Larsen MH, O'Donnell MR, Pym A, Jacobs WR. 2016. Dual-Reporter Mycobacteriophages (Φ2DRMs) Reveal Preexisting Mycobacterium tuberculosis Persistent Cells in Human Sputum. mBio 7:e01023-16.

52. Nguyen D, Brassard P, Westley J, Thibert L, Proulx M, Henry K, Schwartzman K, Menzies D, Behr MA. 2003. Widespread pyrazinamide-resistant Mycobacterium tuberculosis family in a low-incidence setting. J Clin Microbiol 41:2878–2883.

53. Nguyen D, Brassard P, Menzies D, Thibert L, Warren R, Mostowy S, Behr M. 2004. Genomic characterization of an endemic Mycobacterium tuberculosis strain: evolutionary and epidemiologic implications. J Clin Microbiol 42:2573–2580.

54. Wilson DJ, Consortium TCr. 2019. GenomegaMap: within-species genome-wide dN/dS estimation from over 10,000 genomes. bioRxiv 523316.

55. Garcia BJ, Loxton AG, Dolganov GM, Van TT, Davis JL, de Jong BC, Voskuil MI, Leach SM, Schoolnik GK, Walzl G, Strong M, Walter ND. 2016. Sputum is a surrogate for bronchoalveolar lavage for monitoring Mycobacterium tuberculosis transcriptional profiles in TB patients. Tuberculosis 100:89–94.

Figures

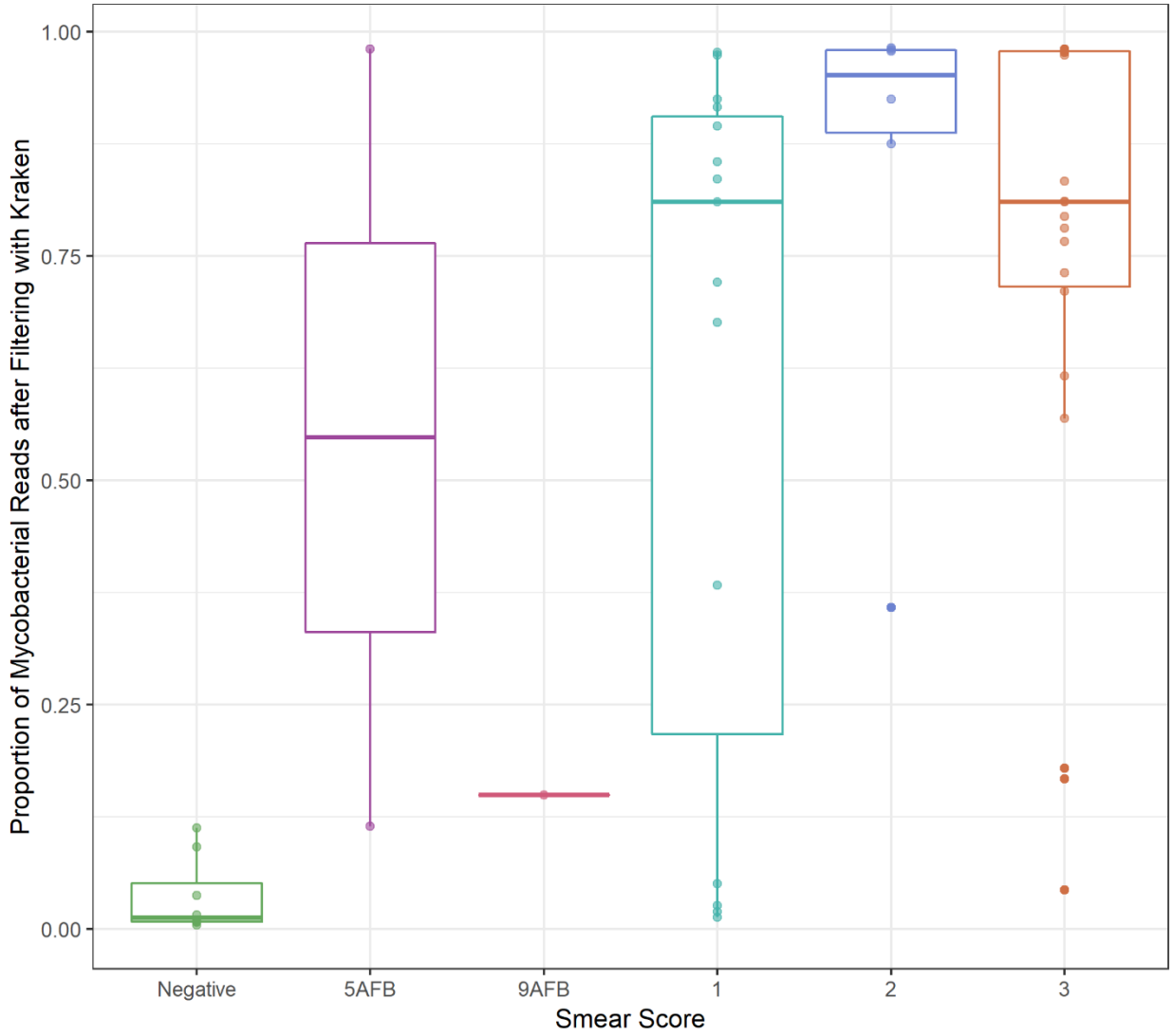

Figure 3.1. Smear score. Boxplot of proportion of sequencing data retained after filtering (y-axis) versus sputum smear score (x-axis) for *M. tb* sputum samples from Brown *et al*. Thresholds for smear score of potentially infected sputum are 0 AFB/100 fields: smear negative, 1-9 AFB/100 fields: actual number of AFB seen on slide, 10-99 AFB/100 fields: 1+, 1-10 AFB/field in 50 fields: 2+, >10 AFB/field in 20 fields: 3+, where AFB corresponds to the number of acid-fast bacilli present.

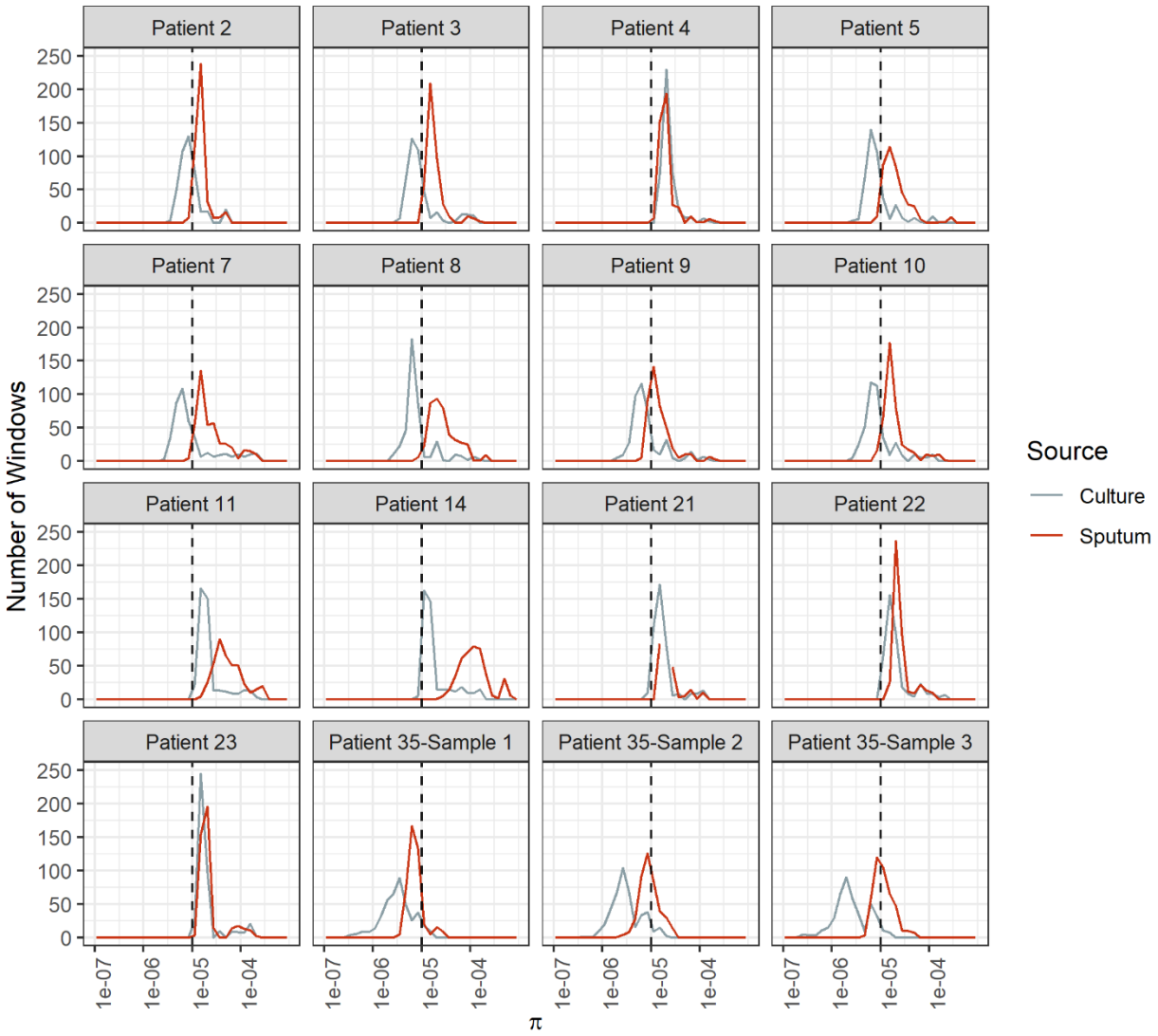

Figure 3.2. Bacterial diversity across genome windows in sputum and culture samples. Frequency (y-axis) of nucleotide diversity (π, x-axis) in sliding windows across the genome (windows = 100Kb, step-size = 10 Kb). Diversity varies among patients as well as among samples from the same patient, as shown by the differences in the shapes of these distributions. Sputum samples exhibit more variability among patients, and diversity is generally higher than it is for culture samples. Sputum and culture in red and grey, respectively. Dotted line at π = 1e-5.

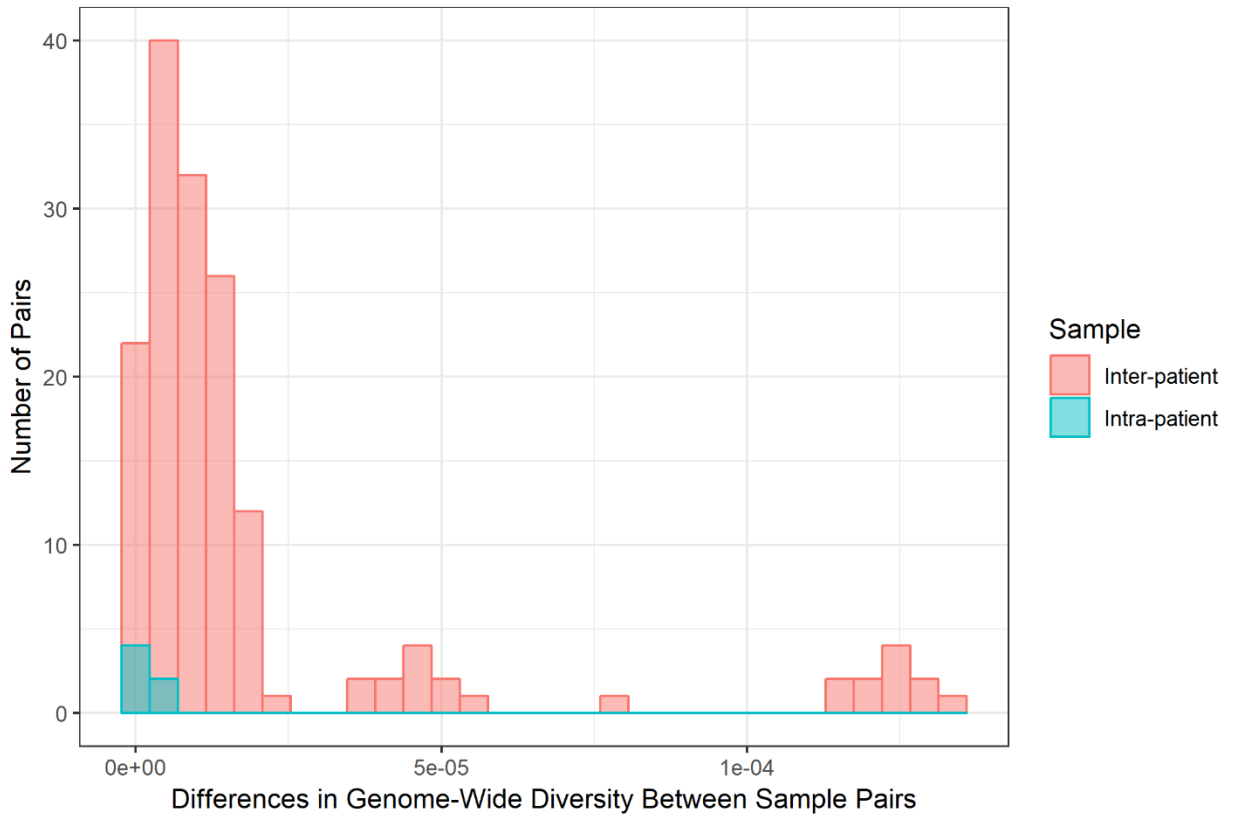

Figure 3.3. Differences in bacterial diversity among inter-patient and intra-patient pairs of samples. Frequency (y-axis) of absolute difference in genome-wide nucleotide diversity (average π across sliding windows, x-axis) between samples. We calculated the absolute difference in genome-wide π for all possible pairs of sputum samples and all possible pairs of culture samples for inter- and intra-patient samples (sputum versus sputum and culture versus culture). Inter- and intra-patient differences pictured in coral and teal, respectively. Differences among samples from the same patient are similar to the bulk of comparisons between patients. Some between-patient pairs exhibit extreme differences in diversity.

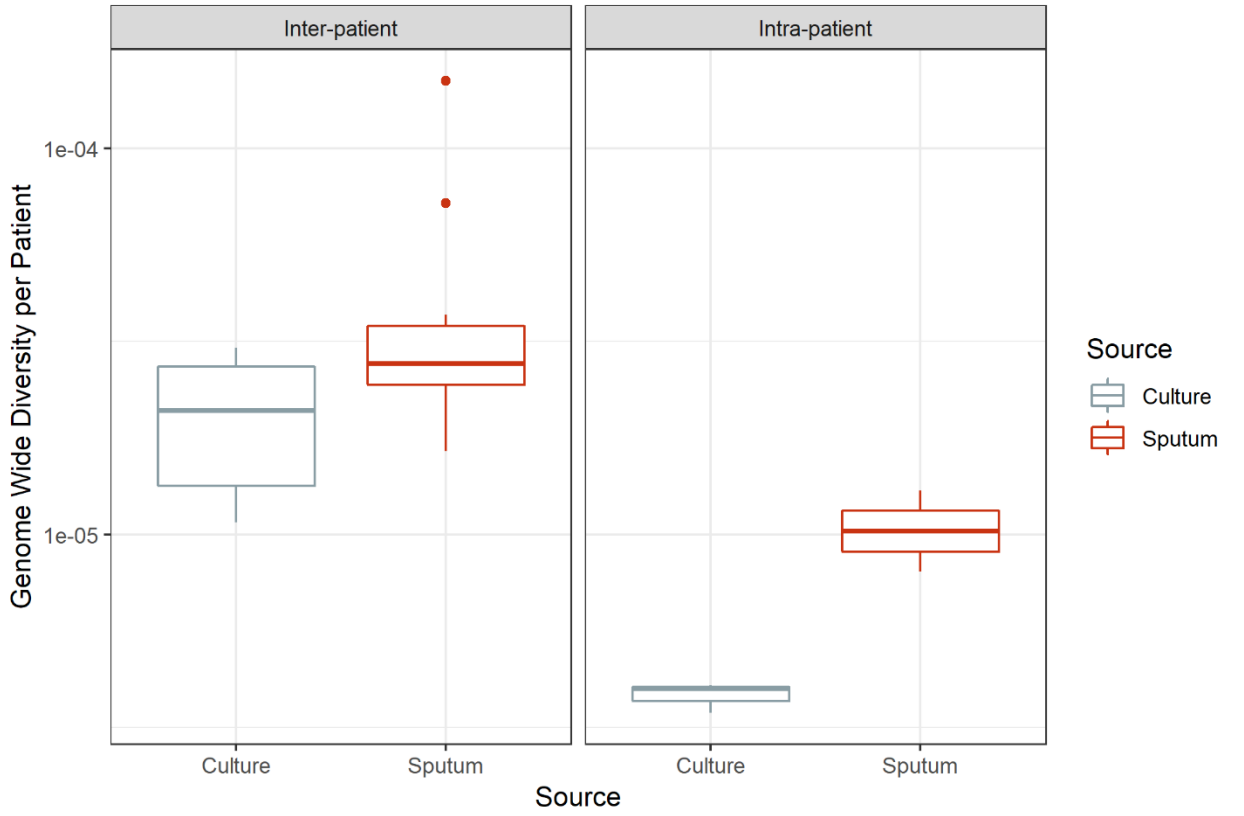

Figure 3.4. Genome-wide diversity in sputum and culture. Average genome-wide nucleotide diversity (π) in sputum and culture samples. Sputum and culture pictured in red and grey, respectively. Left box: inter-patient samples; right box: intra-patient samples. Medians 2.77e-05 and 2.10e-05, 1.02e-05 and 3.98e-06 for sputum and culture in inter-patient and intra-patient samples, respectively.

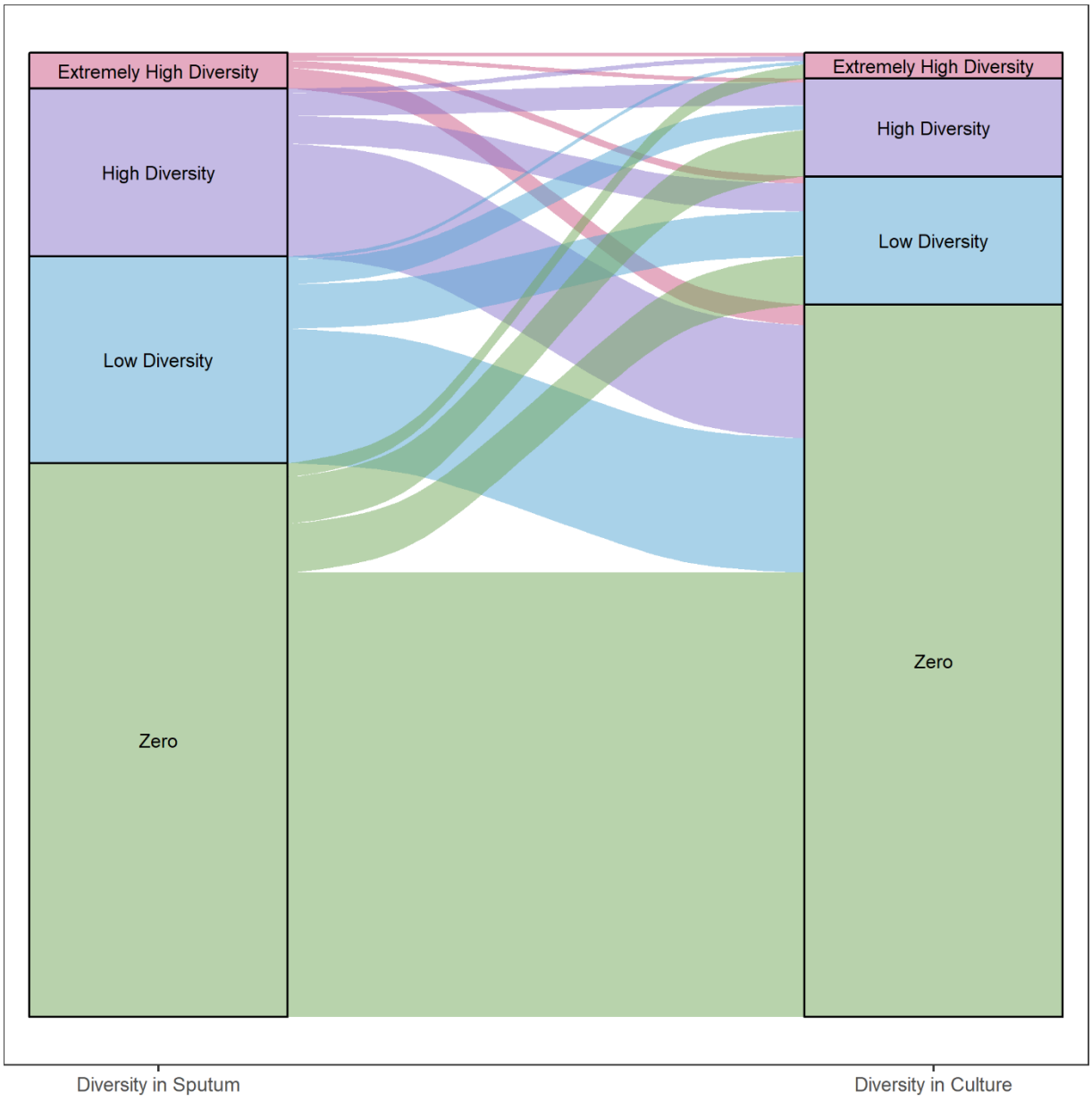

Figure 3.5. Differences in gene diversity of sputum and culture samples. Alluvial plot of changes in nucleotide diversity (π) per gene in sputum and culture samples (left and right, respectively). Inter- and intra-patient samples were all included. Categories defined by diversity quartiles: Extremely high ($\geq Q_3 + 1.5*IQR$), high ($\geq Q_2$ and $< Q_3 + 1.5*IQR$), low ($\geq Q_1 - 1.5*IQR$ and $< Q_2$) and zero diversity. Strata colored by each category. Strata width corresponds to the total number of genes in each category.

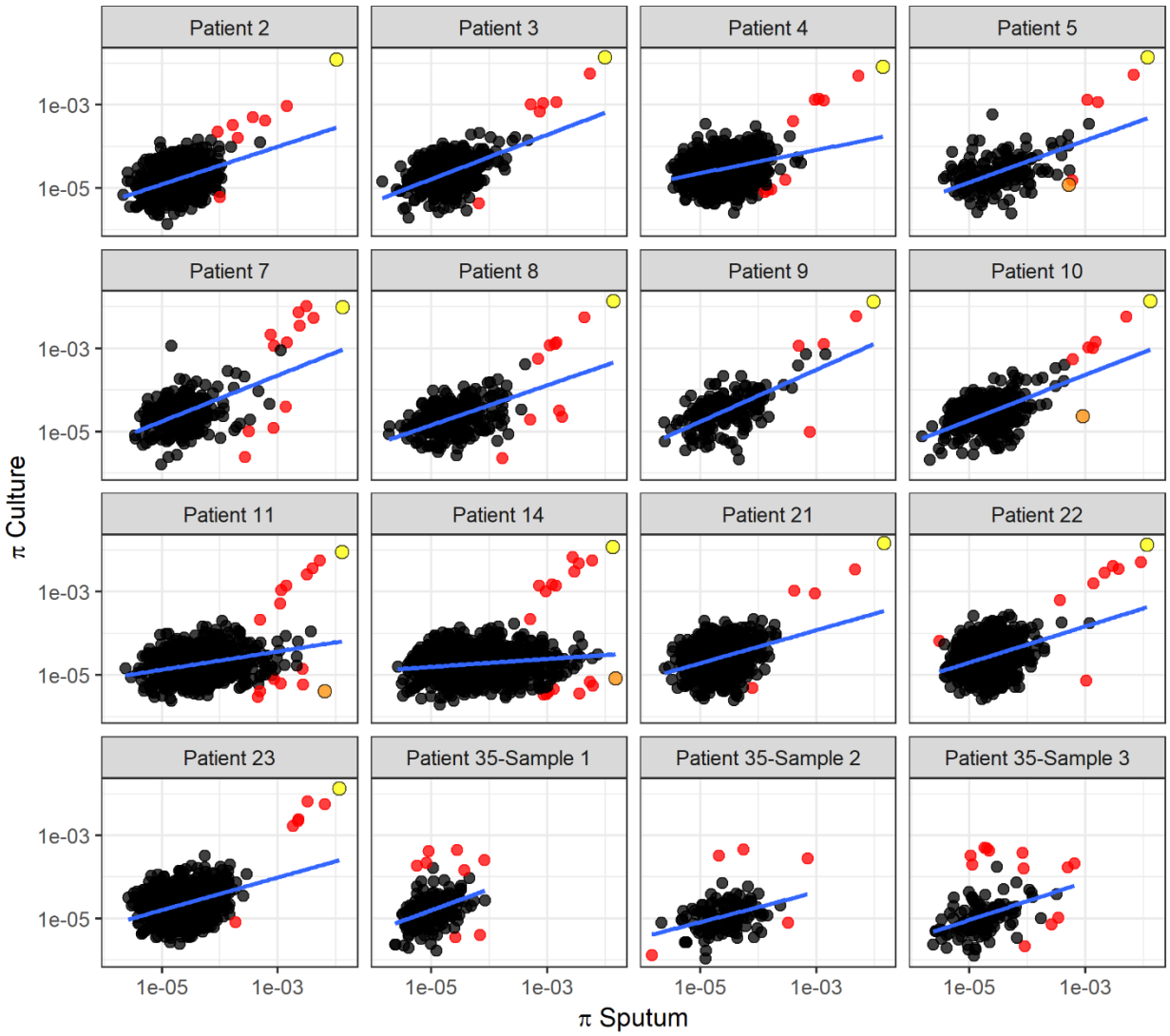

Figure 3.6. Linear regression of gene diversity in culture vs sputum. Nucleotide diversity (π) per gene in sputum (x-axis) vs culture (y-axis) for each patient. Each dot corresponds to a gene identified in sputum and culture samples. Regression line is shown in blue. Red dots mark outliers as identified with Cook's distance from the regression line (i.e. genes with Cook's distance > 4 times the mean of observed distances). We used the F-test of overall significance to assess the fit of a linear regression model to the observed data; no F-test had a p-value > 0.05. The slope of the regression line varies substantially among patients. Two genes identified as outliers in this linear regression analysis are specifically highlighted: Rv2020c, marked in yellow, appeared as an outlier in all 13 TB patients. *nrdE*, marked in orange, also appeared repeatedly as an outlier but in a different part of the distribution.

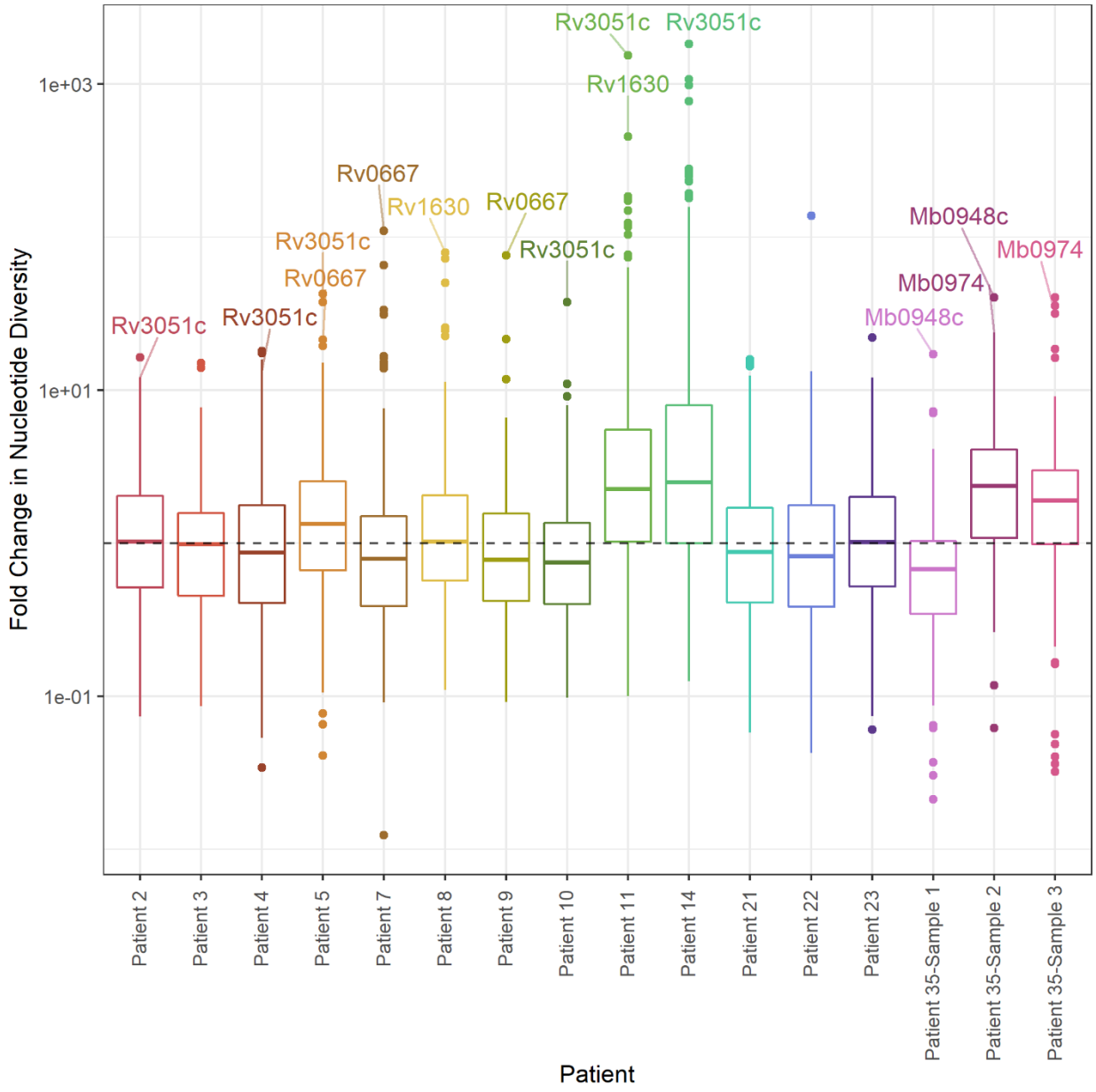

Figure 3.7. Distribution of fold change in nucleotide diversity per gene. Boxplots of fold change in nucleotide diversity per gene (π sputum/π culture). Y-axis log10 scaled. Outliers identified by z-score present in > 1 patient labelled (p-value < 0.05).

Tables

Table 3.1. *M. tb* genes with extreme patterns of variation across multiple patients and multiple measures. The number of patients in whom each gene was identified as an outlier is shown for linear regression of diversity in sputum versus culture (lm), fold change in diversity (z-score) or absolute difference (Zo) and F_{ST} analyses.

Gene	Lm	z-score	Zo	F_{ST}
Rv2020c (hypothetical)	13	0	0	6
Rv1318c	11	0	0	11
Rv1319c	9	0	0	6
Rv3109 (moaA1)	11	0	0	8
Rv1267c (embR)	10	0	0	5
Rv2351c (plcA)	5	0	0	5
Rv2350c (plcB)	5	0	0	4
Rv2082 (hypothetical)	3	0	0	3
Rv2081c (Conserved transmembrane protein)	2	0	0	2
Rv3051c (nrdE)	4	6	2	7
Rv0338c	2	1	0	2
Rv0684 (fusA1)	2	1	2	3
Rv1164 (narI)	2	1	2	5
Rv1327c (glgE)	2	1	0	2
Rv1630 (rpsA)	2	2	1	2
Rv0667 (rpoB)	2	3	1	4
Rv0668 (rpoC)	1	1	2	4

Supplementary Figures

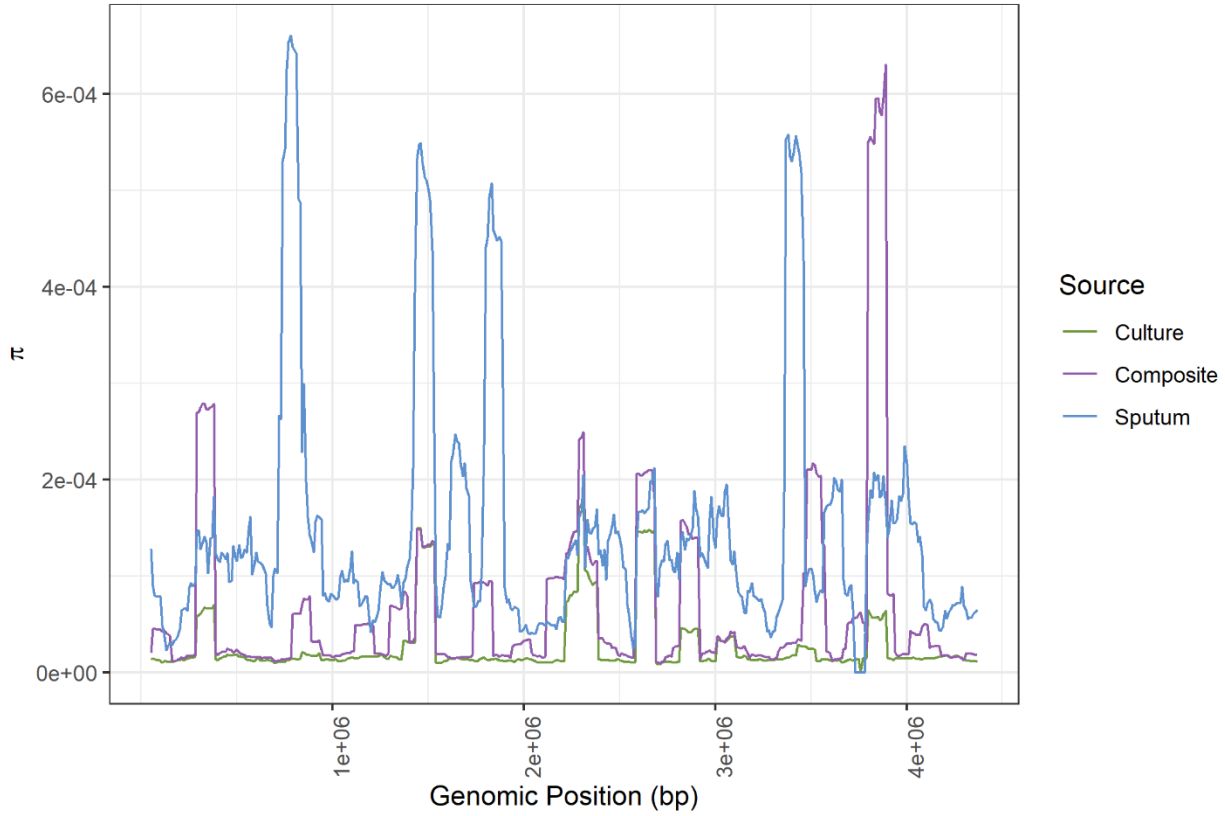

Figure 3.S1. Sliding window diversity of Patient 14 in sputum, culture and culture spiked with negative sputum. Nucleotide diversity (π, y-axis) calculated in sliding windows (10 kb steps, 100 kb windows) across the genome (x-axis) for the culture, composite, and sputum samples from Patient 14. Composite sample contains sequences from Patient 14's culture sample and ~3,000 sequences from TB negative sputum that passed metagenomic filtering. Culture, composite and sputum pictured in green, purple and blue, respectively. Background contamination does increase nucleotide diversity across the genome, but the patterns produced do not mirror those seen in the sputum sample.

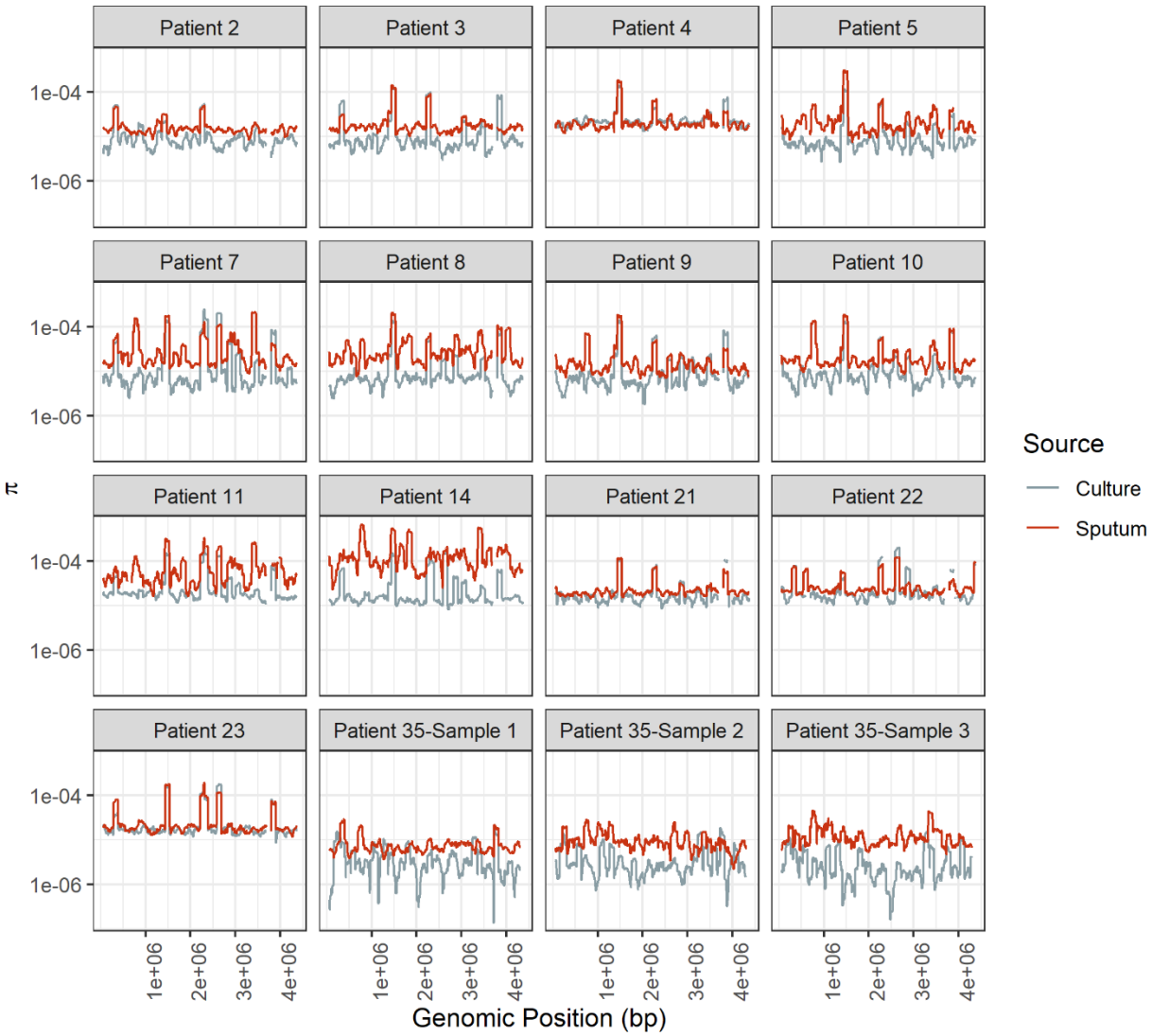

Figure 3.S2. Patterns of nucleotide diversity in composite patient sample. Nucleotide diversity (π, y-axis log10 transformed) calculated in sliding windows (10 kb steps, 100 kb windows) across the genome (x-axis) for sputum and culture samples from each patient.

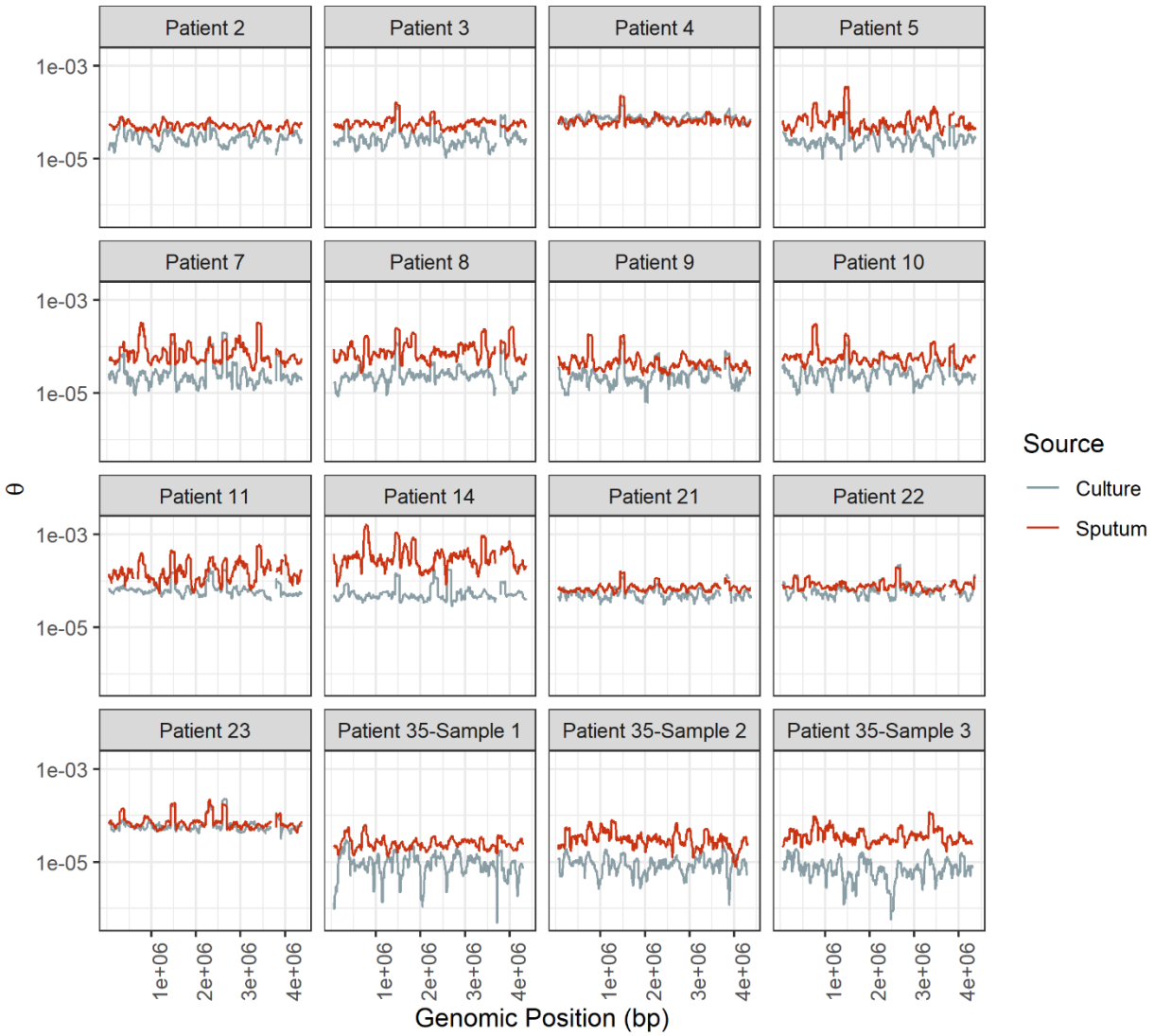

Figure 3.S3. Patterns of Watterson's theta in sputum and culture. Watterson's theta (θ_W, y-axis log10 transformed) calculated in sliding windows (10 kb steps, 100 Kb windows) across the genome (x-axis, in bp) for sputum and culture samples from each patient.

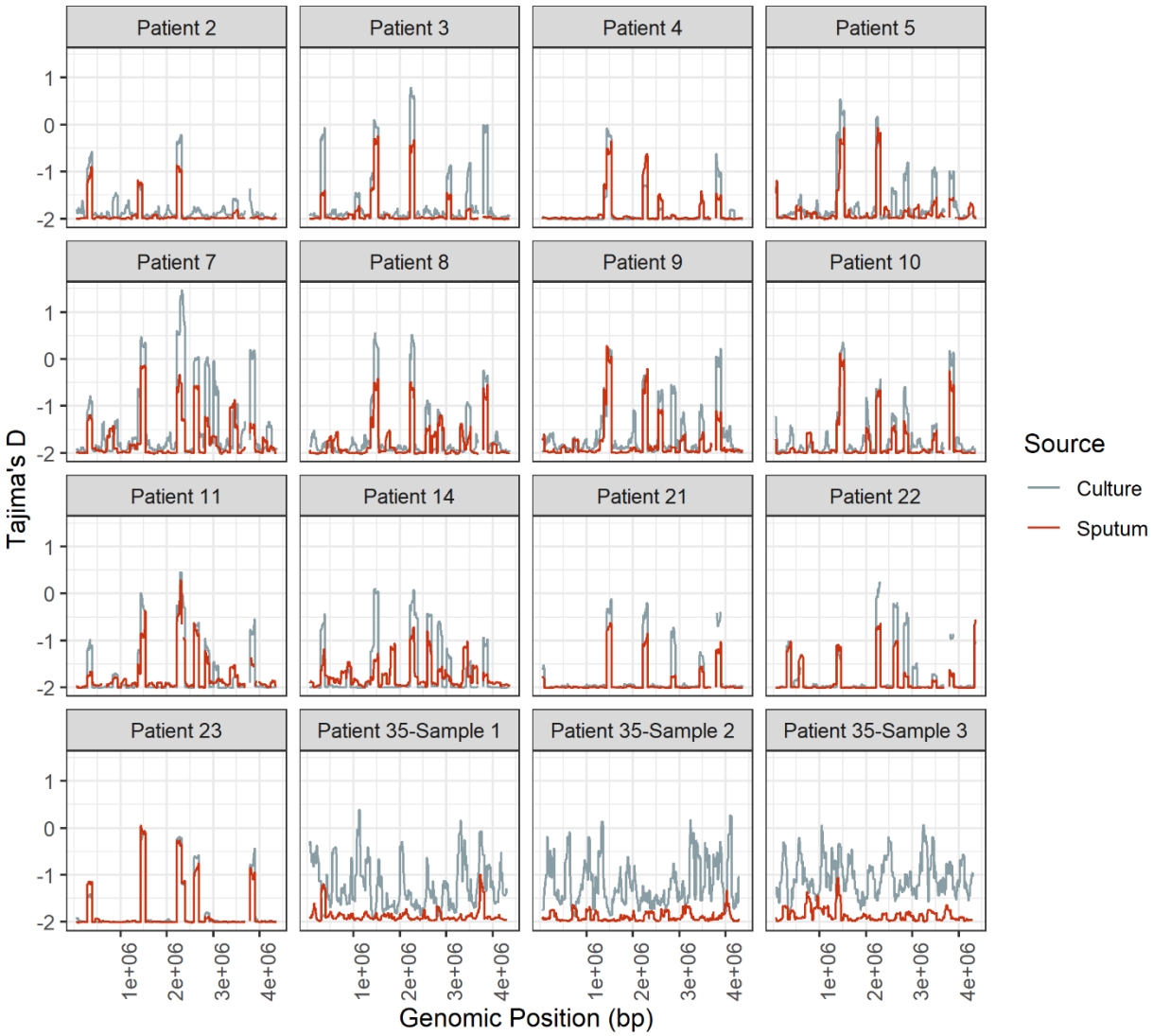

Figure 3.S4. Patterns of Tajima's D in sputum and culture. Tajima's D (y-axis) calculated in sliding windows (10 kb steps, 100 Kb windows) across the genome (x-axis, in bp) for sputum and culture samples from each patient.

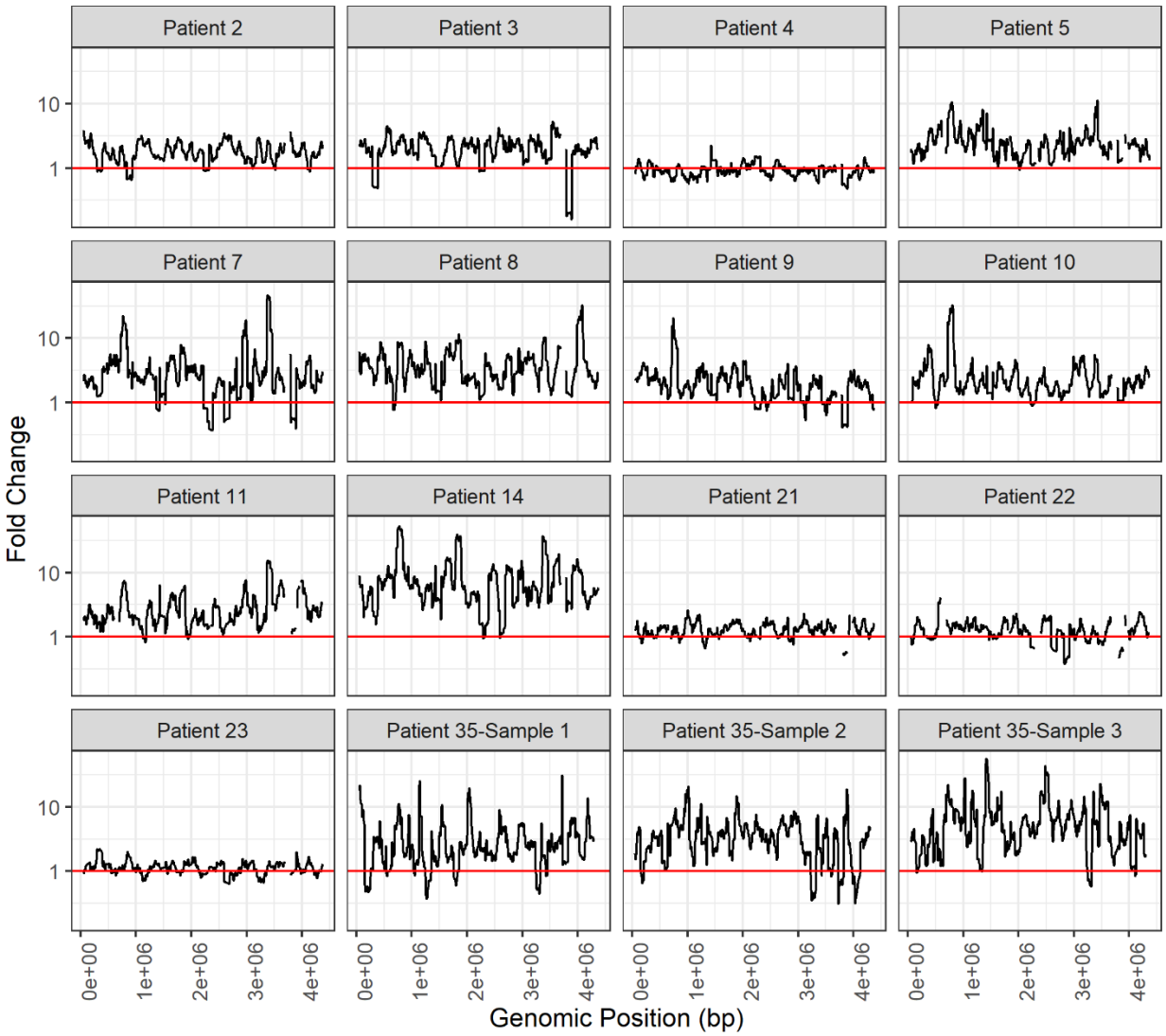

Figure 3.S5. Patterns of fold change in nucleotide diversity across the genome. Fold change in nucleotide diversity (y-axis log10 transformed) from sputum to culture across the genome (x-axis, in bp) for each patient. Fold change calculated from sliding-window analysis of nucleotide diversity as π per window in sputum/π per window in culture for each patient.

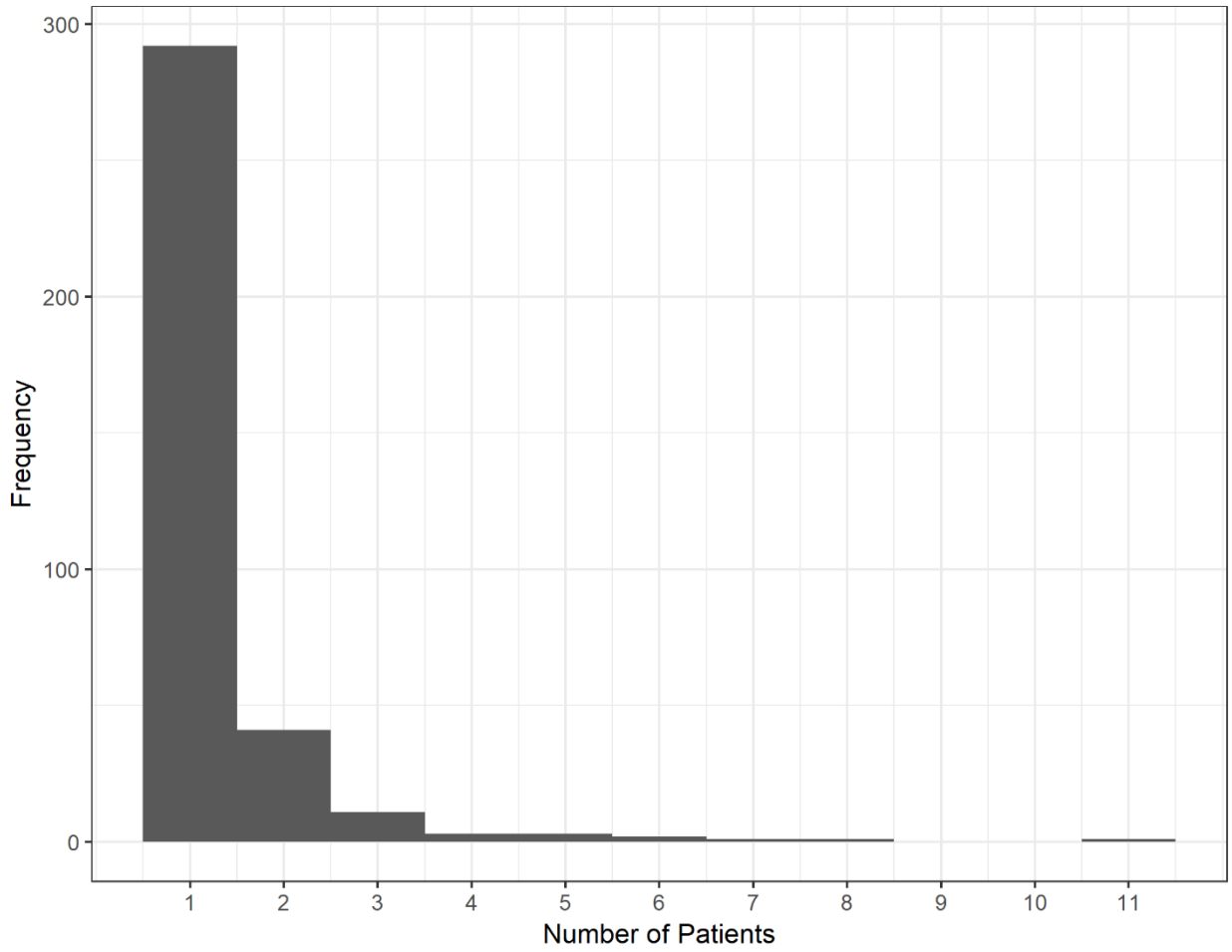

Figure 3.S6. Histogram of F_{ST} outliers in *M. tuberculosis* samples. Frequency (y-axis) of F_{ST} outliers versus number of patients in which an outlier is present (x-axis). F_{ST} outliers identified using Fisher's exact test (adjusted p-value < 0.01).

Supplementary Tables

Table 3.S1. Regions Removed

M. tb Start	*M. tb* Stop		*Mb* Start	*Mb* Stop
4154	4214		10887	10960
4553	4642		10887	10960
7246	7339		10887	10960
10881	10966		11112	11184
11108	11185		11112	11184
17865	17972		11112	11184
18025	18077		25625	25707
21859	21921		25625	25707
22454	22495		25625	25707
23168	23278		28918	29187
24686	24741		33566	33778
32335	32390		43546	46455
33572	33804		80272	80586
36215	36277		103748	105251
53525	53589		105360	106751
55524	55565		105360	106751
55640	55705		131420	132910
79495	79567		131420	132910
80175	80555		149571	151187
84622	84703		149571	151187
95259	95328		177733	179499
103688	105225		177733	179499
105314	106725		179509	181086
112318	112388		179509	181086
131372	132882		187623	189029
142555	142635		187623	189029
149523	151006		189121	190629
154068	154243		189121	190629
161054	161104		194334	194513
163655	163714		248311	249102
167964	168053		306202	307872
177230	177288		306202	307872
177533	180906		331762	332004
185874	185953		331762	332004
187423	188849		331827	334433
188921	190449		331827	334433
194944	194999		334697	337330

206807	206855		334697	337330
206864	206912		337580	340075
207922	208017		337580	340075
210664	210733		340366	341976
215931	215988		340366	341976
222792	222883		350626	350934
234453	234489		350626	350934
242264	242315		350937	352478
243516	243571		350937	352478
249237	249307		362336	364156
253412	253463		362336	364156
260351	260407		367197	370640
261971	262094		367197	370640
267591	267648		370784	376741
271297	271350		370784	376741
272850	272960		387234	387304
279528	279598		387234	387304
307867	309557		387234	387304
309891	310111		400565	401080
327549	327601		400565	401080
332797	332874		425173	435696
333427	336384		425173	435696
336550	339148		467689	469020
339354	340984		467689	469020
344015	344071		531770	533209
349614	351486		531770	533209
353796	353853		544193	545749
361324	363119		544193	545749
366140	372774		601592	602998
372810	375721		623943	625688
379626	379674		623943	625688
382034	382320		625594	625821
382638	382705		625594	625821
399525	400060		659352	659432
400142	401713		659352	659432
410224	410306		659352	659432
421182	421250		673239	677159
423325	423377		673239	677159
424006	424153		702629	704012
424259	424704		702651	703259

424767	434689		703261	704004
467449	468011		722975	723679
472702	472759		732730	732802
472771	474116		732730	732802
475800	476194		732730	732802
484274	484318		732839	732912
488730	488777		732839	732912
490194	490250		732839	732912
498220	498292		734760	734832
507871	507947		734760	734832
514323	514382		734760	734832
530741	532224		834176	834692
543164	544740		834358	834672
547483	547522		834814	835332
554164	554232		834814	835332
569811	569988		837525	839918
577276	577506		837525	839918
580200	580267		840317	843046
580567	580837		840317	843046
587449	587507		848349	850103
595338	595378		848349	850103
595388	595449		850293	852230
606501	608072		850293	852230
611944	611994		852532	852717
616823	616883		852832	852903
621354	641400		852832	852903
642749	642816		852832	852903
645201	645267		891218	892312
671986	675926		891218	892309
678392	678481		922405	922695
681570	681624		922405	922692
686593	686660		924633	924705
690544	690607		924633	924705
698082	698145		924633	924705
701242	701374		924829	924902
701396	702769		924829	924902
703514	703581		924829	924902
703907	703990		924940	925013
706785	706868		924940	925013
707923	707983		924940	925013

708386	708450		925043	925116
709420	709553		925043	925116
709580	709668		925043	925116
710248	710320		925781	926194
710348	710422		925781	926194
711619	711707		926191	928512
714406	714477		926191	928512
715452	715511		928739	931234
721083	721155		928739	931234
725545	725642		948066	948396
725971	726023		948067	948399
731600	731678		968884	970710
733885	733937		968884	970710
736228	736273		977352	978668
742596	742639		977352	978668
745321	745381		979827	980693
748930	749021		1020524	1021795
754128	754208		1020524	1021795
755258	755322		1021810	1022109
791755	791811		1021810	1022109
795462	795523		1025787	1025859
802424	802482		1025787	1025859
812830	812980		1025787	1025859
822548	822608		1025924	1027359
832035	832366		1025924	1025938
832524	832858		1025963	1027282
832971	833518		1027345	1027359
835150	835164		1027507	1029829
835691	838062		1027507	1027555
838441	840866		1027526	1027542
841490	841566		1027570	1028151
842010	842341		1028151	1029803
846149	847923		1029782	1029829
848093	850050		1029811	1029827
850332	850537		1090812	1093583
850631	850721		1090812	1093583
863150	863260		1093800	1094807
863625	863666		1093800	1094807
867252	867313		1095529	1096902
885603	885669		1095529	1096902

888947	891492		1113962	1114034
898219	898322		1113962	1114034
908171	909356		1113962	1114034
912083	912139		1125895	1127454
915545	915596		1133785	1134360
917568	917740		1138523	1138594
921565	921903		1138523	1138594
923800	923885		1138523	1138594
924100	924196		1159365	1159754
924941	927620		1159368	1160880
927827	930495		1159822	1160508
947302	947654		1160542	1160880
950010	950080		1161744	1162919
960157	960338		1161744	1162919
961382	961438		1162996	1163823
964575	965049		1162996	1163823
966558	966620		1165019	1165997
968414	970254		1165019	1165570
971856	971911		1165539	1165946
976862	978213		1169747	1171181
993355	993460		1169820	1169834
995579	995650		1169872	1171119
1004771	1004822		1171129	1171143
1010775	1010827		1177376	1177690
1020048	1021653		1177844	1177917
1025314	1025398		1177844	1177917
1025453	1025477		1177844	1177917
1025487	1026826		1188869	1190884
1026874	1026898		1188869	1190884
1027056	1027081		1191217	1193508
1027094	1029365		1191217	1193508
1043368	1043419		1212920	1215244
1049817	1049874		1212920	1215244
1056807	1056877		1215894	1216328
1064045	1064108		1215894	1216328
1082311	1082375		1216207	1216512
1085274	1085336		1216207	1216512
1090363	1093154		1217850	1220402
1093351	1094602		1217850	1220402
1095068	1096702		1252988	1254343

1108773	1108833		1263643	1265499
1112913	1112958		1263643	1265499
1121656	1121732		1277672	1279120
1135370	1135433		1279215	1280192
1142279	1142400		1279215	1279218
1145250	1145302		1279219	1279235
1146444	1146524		1279265	1279672
1150190	1150246		1279641	1280189
1158908	1159317		1280173	1280188
1159365	1160071		1280189	1280192
1160085	1160443		1300139	1300681
1160529	1161177		1300139	1300681
1161287	1162482		1300635	1301177
1162539	1163386		1300635	1301177
1164562	1165559		1301195	1301497
1169287	1170743		1301195	1301497
1171795	1171872		1303128	1304054
1176918	1177383		1303128	1304054
1179340	1179400		1340253	1340552
1179545	1179633		1340253	1340552
1188411	1190434		1340599	1341771
1190747	1192158		1340599	1341771
1195719	1195773		1342535	1342542
1203486	1203549		1342543	1343977
1211550	1213873		1342581	1342595
1214503	1215141		1342605	1343852
1216459	1219040		1343890	1343904
1236930	1236988		1343978	1343985
1240520	1240578		1358493	1358867
1251607	1252982		1358493	1358867
1256506	1256572		1386233	1387921
1262262	1264138		1386233	1387921
1275919	1276025		1444242	1444314
1276283	1277807		1444242	1444314
1277833	1278830		1444242	1444314
1282832	1282893		1444428	1446080
1283452	1283655		1466189	1467688
1284355	1284431		1466189	1466208
1287965	1288019		1467669	1467688
1296263	1296315		1469882	1471418

1297712	1297785		1469882	1471418
1298754	1300134		1469882	1471418
1301745	1302691		1471694	1474831
1305485	1305671		1471694	1474831
1318254	1318305		1471694	1474831
1330540	1330591		1474935	1475049
1338993	1339312		1474935	1475049
1339339	1342740		1474935	1475049
1351216	1351284		1486189	1488000
1353148	1353208		1486189	1488000
1353242	1353333		1510763	1510846
1357075	1357127		1510763	1510846
1357170	1357254		1510763	1510846
1357283	1357635		1530478	1531668
1357929	1358003		1530478	1531668
1367316	1367379		1558143	1558451
1369247	1369302		1558143	1558451
1369610	1369670		1558448	1560067
1383874	1383933		1558448	1560067
1384979	1386687		1568806	1570164
1410877	1410952		1568806	1570164
1418863	1418932		1578473	1579411
1420193	1420262		1579408	1580781
1421128	1421181		1602693	1604279
1421891	1421949		1602693	1604279
1425283	1425335		1614537	1616012
1428005	1428088		1614537	1616012
1434753	1434809		1626985	1631211
1441417	1441550		1626985	1631211
1441576	1441702		1632588	1634975
1443728	1443780		1632588	1634975
1446182	1446273		1652355	1653467
1451769	1452005		1652355	1653467
1456580	1456632		1720532	1723657
1457448	1457562		1735310	1737346
1461088	1461145		1735310	1737346
1461955	1462014		1764680	1765090
1468138	1469626		1765296	1765607
1469628	1469656		1765712	1766065
1472713	1472961		1766009	1767430

1473022	1473257		1767438	1767950
1474157	1474274		1768124	1768594
1474436	1474618		1768675	1768989
1475761	1476033		1768986	1769258
1476107	1476522		1769272	1769667
1479770	1479831		1769863	1771278
1481384	1481432		1771278	1771676
1481444	1481504		1771673	1771894
1483103	1483150		1771950	1772465
1483213	1483265		1772462	1773871
1488144	1489975		1773873	1775235
1507526	1507586		1813381	1813454
1508673	1508728		1813381	1813454
1525560	1525616		1813381	1813454
1532433	1533663		1833883	1837401
1535288	1535336		1838971	1839753
1536126	1536182		1841129	1842061
1541939	1543328		1841129	1842061
1548976	1549022		1844098	1845123
1553244	1553318		1845123	1847618
1559385	1559448		1847712	1850768
1561454	1563398		1847712	1850768
1572117	1573867		1898942	1900216
1578616	1578703		1903282	1904088
1584142	1584204		1912553	1913931
1588812	1588865		1916838	1917995
1595912	1595966		1916838	1917995
1606376	1607982		1918035	1919219
1611868	1611912		1918035	1919219
1612553	1612667		1931953	1932026
1618199	1619694		1931953	1932026
1625361	1625423		1931953	1932026
1630628	1634800		1932189	1932363
1635994	1638402		1967012	1970167
1644243	1644375		1967012	1970167
1649393	1649480		1989024	1990253
1655599	1656731		1989024	1989037
1658572	1658665		1989582	1989797
1678953	1679018		1990240	1990253
1679052	1679102		1991054	1992928

1681860	1681922		1991054	1992928
1683757	1683812		2015759	2016853
1685479	1685546		2015759	2016853
1702401	1702448		2016932	2017231
1702469	1702516		2016932	2017231
1702667	1702853		2017245	2018426
1708592	1708658		2017245	2018426
1724527	1724577		2018880	2019932
1725801	1725849		2018880	2019932
1751287	1753343		2020359	2020658
1776751	1776830		2020359	2020658
1777604	1777671		2029908	2031875
1779184	1779734		2029908	2031875
1779799	1782594		2032456	2033727
1782613	1783238		2032456	2033727
1783299	1784311		2033839	2035230
1784487	1786538		2033839	2035230
1786574	1789875		2035378	2037000
1800175	1800228		2035378	2037000
1809654	1809759		2037034	2037297
1810323	1810370		2037034	2037297
1814843	1815138		2038527	2038826
1821850	1821933		2038527	2038826
1830069	1830130		2038853	2040052
1831729	1831818		2038853	2040052
1849453	1849541		2040376	2041605
1855754	1856706		2040376	2041605
1858834	1858888		2041737	2042300
1859876	1859936		2041737	2042300
1862337	1865392		2042301	2043149
1867241	1867318		2042301	2043149
1868407	1868455		2051633	2053138
1869201	1869258		2051633	2053138
1870485	1870589		2078323	2079723
1871815	1871870		2078323	2079723
1876454	1876469		2134686	2135117
1876570	1876635		2153213	2158945
1882571	1882617		2153213	2158945
1883749	1883790		2159283	2160287
1889452	1889481		2159283	2160287

1895343	1895603		2160370	2162247
1900022	1900100		2160370	2162247
1902728	1902778		2187623	2188984
1907454	1907580		2205160	2206836
1918531	1918585		2205160	2206836
1923414	1923470		2210596	2211135
1926192	1927147		2239359	2240586
1927201	1928594		2239440	2240060
1931487	1932664		2240014	2240604
1932684	1933888		2324842	2326310
1938088	1938150		2338572	2340047
1944751	1944813		2338572	2340047
1946608	1946701		2347842	2348138
1946839	1947033		2347842	2348138
1955662	1955677		2348194	2348925
1958367	1958428		2348194	2348925
1969339	1969408		2361496	2362890
1970094	1970143		2361496	2362890
1981604	1984785		2367600	2368667
1987689	1989069		2367600	2368667
1989823	1992587		2371613	2372857
1994964	1995032		2382501	2382586
1996089	1998622		2382501	2382586
1999132	1999367		2382501	2382586
1999795	1999818		2403758	2405230
2000604	2002480		2403758	2405230
2004918	2004976		2408627	2409661
2014109	2014161		2418179	2418982
2016296	2016362		2418316	2418981
2017877	2017938		2489746	2489817
2025291	2026408		2489746	2489817
2026467	2027981		2489746	2489817
2028415	2029487		2559359	2559432
2029894	2030213		2559359	2559432
2030322	2030993		2559359	2559432
2039443	2041430		2578326	2579474
2041991	2043282		2578326	2579474
2043374	2044785		2595264	2596505
2044913	2046852		2595264	2596505
2048062	2048381		2597004	2597076

2048388	2049607
2049911	2051160
2051272	2052698
2052700	2052789
2059424	2059605
2061168	2062684
2074426	2074643
2087961	2089528
2094907	2094922
2101328	2101392
2108587	2108665
2122554	2122610
2124851	2124898
2136995	2137048
2152758	2152836
2153712	2153746
2155718	2155774
2156801	2156869
2157480	2157549
2162922	2167321
2167639	2170622
2186009	2186068
2187106	2187155
2195847	2197370
2198718	2198819
2204586	2204694
2206523	2206574
2211306	2211371
2226234	2227930
2237389	2237442
2242369	2242422
2244105	2244158
2244624	2244691
2254188	2254243
2260655	2261698
2262000	2263179
2266507	2266517
2270877	2270928
2289864	2289909
2291348	2291418

2597004	2597076
2597004	2597076
2603651	2605024
2603651	2605024
2605762	2607606
2605762	2607606
2607744	2609135
2619824	2620009
2619824	2620009
2660871	2661953
2660871	2661953
2674201	2674920
2674201	2674920
2688829	2690039
2688829	2688841
2688961	2689653
2689667	2689960
2690027	2690039
2695499	2696083
2695499	2696083
2696130	2696429
2696130	2696429
2715758	2718388
2733494	2733567
2733494	2733567
2733494	2733567
2733704	2733774
2733704	2733774
2733704	2733774
2760978	2761051
2760978	2761051
2760978	2761051
2762102	2762926
2762102	2762926
2762605	2763777
2762605	2763777
2763566	2764186
2763566	2764186
2768055	2771507
2768055	2771507

2292286	2292363
2295160	2296432
2296964	2297034
2299187	2300832
2300897	2300984
2301250	2302522
2303051	2303120
2305256	2306899
2309562	2309640
2315203	2315270
2318466	2318551
2321642	2321693
2324325	2324376
2330142	2330230
2337797	2337863
2337927	2338104
2339366	2339430
2339444	2339508
2340116	2340252
2340520	2340697
2341363	2341425
2343017	2343926
2343984	2344234
2345421	2345490
2347230	2347298
2347385	2347623
2350382	2350430
2353918	2353972
2356719	2358216
2358379	2360051
2365394	2366781
2367349	2367665
2367701	2368452
2372430	2372554
2381061	2382502
2387192	2387982
2401812	2401880
2404876	2404938
2407507	2407548
2412567	2412616

2771504	2773033
2771504	2773033
2794651	2794723
2794651	2794723
2794651	2794723
2795291	2796725
2795329	2795343
2795353	2796600
2796638	2796652
2802291	2802363
2802291	2802363
2802291	2802363
2802582	2804060
2802582	2804060
2840710	2841222
2841223	2843937
2863416	2865251
2872269	2873540
2888999	2890639
2888999	2890639
2892274	2892840
2902503	2904245
2902503	2904245
2908646	2910724
2911057	2912451
2911057	2912451
2927567	2929903
2927567	2929903
2936959	2937030
2936959	2937030
2936959	2937030
2937215	2937287
2937215	2937287
2937215	2937287
2937317	2937387
2937317	2937387
2937317	2937387
2937404	2937475
2937404	2937475
2937404	2937475

2412945	2413001		2939425	2940353
2419690	2419776		2939498	2939512
2421412	2421468		2939550	2940353
2421695	2421762		2996461	2997405
2423230	2424848		3001622	3003160
2427954	2428013		3010528	3012042
2430104	2431484		3010528	3012042
2439130	2439958		3033468	3034652
2445687	2445745		3033468	3034652
2458387	2458454		3034732	3035559
2461310	2461514		3034732	3035559
2462549	2462623		3035883	3037031
2468266	2468328		3035883	3037031
2471703	2471771		3045594	3046910
2476267	2476321		3056724	3058755
2482559	2482660		3056751	3058130
2482819	2482904		3058130	3058711
2484934	2485012		3058913	3059809
2493796	2493823		3072293	3072694
2494047	2494116		3072296	3072694
2498273	2498340		3073369	3074775
2512061	2512121		3073370	3074050
2517089	2517151		3073980	3074777
2521728	2521776		3075735	3080138
2522099	2522153		3075735	3075770
2522168	2522235		3075806	3075841
2523179	2523242		3075879	3075914
2528654	2528710		3075953	3075988
2531883	2532219		3076032	3076067
2536672	2536727		3076109	3076144
2537212	2537270		3076183	3076218
2549994	2551380		3076257	3076292
2581753	2582308		3076330	3076365
2583425	2583789		3076404	3076439
2597855	2597898		3076475	3076510
2598756	2598818		3076548	3076583
2600721	2601889		3076621	3076656
2617657	2618918		3076693	3076728
2619400	2619487		3076765	3076800
2622065	2622148		3076839	3076874

2622780	2622843		3076913	3076948
2624357	2624427		3076984	3077003
2625868	2626693		3077001	3077003
2632913	2634108		3077004	3078358
2634518	2637034		3077004	3077031
2637248	2637543		3077046	3077882
2637678	2639618		3077981	3078307
2651743	2651948		3078331	3078358
2664274	2664368		3078359	3078361
2672054	2672134		3078362	3078377
2674204	2674275		3078418	3078453
2681308	2681351		3078493	3078528
2686476	2686524		3078566	3078601
2687123	2687262		3078642	3078677
2690666	2690739		3078719	3078754
2692789	2693894		3078791	3078826
2706007	2706746		3078863	3078898
2708313	2708367		3078936	3078971
2716310	2716396		3078997	3079032
2720625	2721787		3079074	3079109
2721839	2721861		3079146	3079181
2727326	2727930		3079221	3079256
2727957	2728276		3079295	3079330
2738102	2738161		3079369	3079404
2744460	2744521		3079440	3079475
2744529	2744610		3079515	3079550
2753317	2753403		3079589	3079624
2760389	2760475		3079664	3079699
2762757	2763066		3079736	3079771
2763392	2763701		3079807	3079842
2779812	2779942		3079880	3079915
2780739	2780788		3079952	3079987
2784591	2785987		3080029	3080064
2794168	2794256		3080103	3080138
2795291	2797395		3097058	3097885
2797856	2797917		3107772	3109520
2798635	2798738		3118784	3120631
2800666	2800923		3118784	3120631
2801244	2806246		3146099	3147193
2806363	2806630		3150681	3152948

2815022	2815063		3150745	3152064
2824615	2824667		3152061	3152948
2827847	2827955		3157310	3158536
2828484	2829940		3157310	3158536
2835491	2835568		3172181	3172873
2835775	2837273		3172877	3173407
2837402	2837471		3244934	3244938
2842038	2842151		3244939	3247151
2844099	2844152		3244939	3244992
2850923	2850974		3245037	3246278
2866496	2866573		3246278	3246808
2866581	2866759		3247098	3247151
2866917	2866996		3247152	3247156
2867032	2867099		3269972	3270361
2867154	2867234		3290455	3292479
2867236	2867406		3290455	3290460
2867456	2867485		3290472	3291851
2867505	2867664		3291851	3292435
2867688	2867765		3292474	3292479
2873069	2873124		3305107	3305179
2880988	2881048		3305107	3305179
2887933	2887982		3305107	3305179
2905135	2905221		3305219	3305290
2915149	2915214		3305219	3305290
2921541	2923192		3305219	3305290
2935036	2936798		3305365	3306837
2943590	2945236		3323879	3325408
2953453	2953502		3326565	3328049
2960095	2962451		3328046	3328345
2969498	2969572		3333554	3334867
2969750	2969827		3333554	3334867
2969936	2970017		3334888	3334974
2970541	2971559		3335595	3335888
2971649	2972037		3335935	3337242
2972095	2973475		3335935	3337242
2973785	2976564		3337239	3337553
2976576	2976919		3337239	3337553
2976979	2980828		3337865	3337872
2982934	2983884		3337873	3339307
2990579	2990703		3337911	3337925

2995994	2996178		3337935	3339182
3007028	3007230		3339220	3339234
3013607	3013692		3339308	3339315
3022125	3022199		3339345	3340448
3038448	3038517		3388387	3388459
3041651	3041715		3388387	3388459
3053904	3055501		3388387	3388459
3072037	3072100		3422325	3423638
3073050	3073118		3422325	3423638
3076884	3078088		3437865	3437872
3078148	3078995		3437873	3439307
3079299	3080467		3437946	3437960
3082456	3082664		3437998	3439245
3100169	3102172		3439255	3439269
3104974	3105042		3439308	3439315
3105553	3105611		3444248	3445423
3107355	3107404		3444248	3445423
3113032	3113092		3455106	3456251
3113699	3113776		3455106	3456251
3113790	3113850		3456903	3458045
3115731	3116152		3456903	3458045
3116713	3116741		3465197	3466426
3116808	3118237		3465197	3466426
3119178	3123581		3482500	3484269
3129945	3130081		3482500	3484269
3132882	3133465		3484517	3486415
3135778	3136361		3484517	3486415
3137381	3137453		3511796	3513185
3142822	3142887		3511796	3511858
3155869	3156099		3511845	3512879
3156294	3156405		3513123	3513185
3159422	3159492		3513905	3513978
3160517	3160588		3513905	3513978
3162258	3164125		3513905	3513978
3171458	3171637		3668986	3670071
3173805	3173834		3668998	3670710
3181789	3181841		3679870	3680880
3181847	3181902		3686614	3690630
3189478	3189524		3686614	3690630
3189533	3189559		3690679	3694692

3189560	3189591		3690679	3694692
3192185	3192365		3694689	3696308
3194128	3196442		3694689	3696308
3200784	3202030		3697245	3700430
3203384	3204300		3697245	3700430
3219058	3219102		3700427	3706717
3225734	3225838		3700427	3706717
3232641	3232876		3706864	3707893
3239465	3239795		3707297	3707788
3247574	3247642		3707825	3708565
3247810	3247911		3707825	3708565
3247923	3248158		3709479	3713138
3252989	3253055		3709479	3713138
3253227	3253574		3713132	3719266
3258327	3258396		3713132	3719266
3263691	3263777		3719324	3720628
3266553	3266625		3719324	3720628
3285770	3285855		3731263	3731727
3288454	3290516		3732094	3733974
3291414	3291441		3732094	3733974
3291452	3291508		3752268	3752269
3293192	3293238		3752270	3753837
3313273	3313682		3752270	3752294
3316409	3316475		3752375	3753079
3317524	3317749		3753069	3753746
3318830	3318894		3753813	3753837
3319463	3319671		3753838	3753839
3319788	3319843		3753936	3756185
3320753	3320810		3753936	3756185
3324329	3324395		3794867	3795403
3333132	3333190		3794867	3795403
3333748	3335758		3795344	3796507
3335782	3335797		3795344	3796507
3335992	3336036		3796996	3797841
3336494	3336812		3796998	3797842
3339306	3339359		3823752	3824645
3351086	3351218		3831221	3832613
3353997	3354049		3840431	3840727
3361693	3361749		3840431	3840727
3364843	3364889		3840764	3841945

3370927	3370986		3840764	3841945
3376917	3378279		3872192	3876274
3378319	3378460		3872192	3876274
3379354	3381036		3876565	3880947
3381302	3382759		3876565	3880947
3387531	3387588		3883854	3889670
3392126	3392193		3883854	3889670
3393617	3393668		3890501	3893479
3394362	3394418		3890501	3893479
3415176	3415199		3912558	3913778
3426042	3426094		3912558	3913778
3429293	3429356		3913920	3915668
3430941	3431043		3913920	3915668
3444275	3444343		3921274	3922713
3454826	3454903		3921274	3922713
3456333	3456385		3941195	3942853
3459654	3459737		3941195	3942853
3462136	3462191		3964401	3965609
3462589	3462645		3965610	3967019
3465768	3467101		3974619	3976373
3467364	3467477		3974619	3976373
3473535	3473641		3979813	3981291
3481315	3482774		3979813	3981291
3483936	3485121		3984664	3986181
3489875	3490012		4012903	4015020
3490466	3491661		4012903	4012951
3493380	3493445		4012989	4013336
3501324	3501742		4013721	4014221
3501784	3502946		4014221	4014967
3510078	3511327		4014972	4015020
3521364	3521417		4015743	4017035
3527381	3529173		4015743	4015755
3532762	3532827		4015757	4016986
3535192	3535273		4017023	4017035
3536560	3536627		4018602	4018674
3543311	3543386		4018602	4018674
3551202	3552604		4018602	4018674
3552695	3554080		4028470	4028754
3557301	3558355		4028470	4028754
3559355	3559452		4030869	4031183

3560027	3560095		4030869	4031183
3573437	3573509		4031177	4031890
3579515	3579581		4031177	4031890
3587740	3587789		4063888	4063961
3589231	3589361		4063888	4063961
3590655	3590702		4063888	4063961
3591488	3591574		4105805	4105890
3594253	3594468		4105805	4105890
3598479	3598571		4105805	4105890
3598716	3598776		4126745	4127101
3614755	3614823		4126745	4127101
3617109	3617185		4132465	4132800
3617921	3618006		4132465	4132800
3626609	3626671		4135168	4135383
3630276	3630327		4135425	4135511
3632992	3633046		4135425	4135511
3650939	3650989		4135425	4135511
3653782	3664005		4153159	4153231
3679071	3679163		4153159	4153231
3686809	3686868		4153159	4153231
3690946	3691064		4153262	4153350
3695686	3695751		4153262	4153350
3704890	3705009		4153262	4153350
3710369	3713471		4155980	4157227
3717556	3717622		4158876	4158962
3720078	3720213		4158876	4158962
3729354	3736945		4158876	4158962
3736974	3742784		4189149	4190642
3743193	3743610		4189149	4189175
3743701	3753194		4189288	4189905
3753321	3755247		4189880	4190611
3755942	3767112		4190616	4190642
3769509	3769725		4212844	4214358
3769749	3769867		4212844	4214358
3770989	3771096		4237816	4239688
3774348	3774401		4237836	4239062
3775671	3775715		4239059	4239670
3778558	3780344		4243928	4245187
3781201	3781300		4254615	4255644
3781956	3782005		4255048	4255539

3790337	3790391
3795034	3796426
3799982	3800016
3800044	3801473
3801525	3801559
3801643	3803918
3806693	3806750
3807003	3807073
3808835	3808906
3809848	3809934
3810075	3810130
3810457	3810509
3810824	3810905
3810916	3810989
3811202	3811253
3811787	3811841
3820312	3820599
3830720	3830784
3841523	3842813
3842836	3842898
3843026	3843744
3843875	3844650
3844728	3845980
3846451	3848815
3849284	3850149
3851798	3851933
3858351	3858456
3873454	3873518
3877773	3877837
3878577	3878622
3883515	3884931
3890762	3892150
3894083	3895617
3908955	3909002
3909764	3909823
3914556	3914607
3918208	3936720
3937852	3937930
3939607	3944973
3945088	3945607

4287065	4287361
4287065	4287361
4287395	4288501
4287395	4288501
4308441	4309640
4308441	4309640
4309719	4309952
4309719	4309952
4341488	4342162

3945768	3950273
3950820	3951339
3953579	3953638
3955425	3955482
3969333	3970573
3970695	3972463
3978049	3979508
3991563	3991630
3997970	3999648
4006907	4006951
4020196	4020254
4021386	4021458
4026554	4026622
4031394	4033168
4036721	4038060
4042648	4042761
4045224	4045280
4046637	4046698
4052490	4052552
4052944	4053559
4059974	4060601
4060638	4062208
4063123	4063184
4064556	4064612
4070236	4070325
4075610	4075635
4075742	4076109
4076474	4077755
4077787	4077859
4077989	4078211
4078501	4079759
4079781	4079803
4087491	4087625
4088863	4088911
4091027	4091078
4091223	4091527
4093622	4094537
4095859	4095942
4096666	4096719
4118107	4118171

4118868	4118922
4120910	4121058
4134596	4134730
4146550	4146605
4149398	4149471
4149622	4149675
4151038	4151100
4155454	4155588
4156789	4156977
4157003	4157076
4168436	4168490
4168936	4168987
4175865	4175925
4189275	4190242
4190274	4190527
4196161	4196516
4198195	4198607
4198864	4199099
4204922	4204981
4207124	4207178
4215871	4216279
4216861	4216942
4222395	4222458
4227017	4227073
4238008	4238074
4252872	4254342
4266551	4266627
4268843	4268894
4269949	4270012
4274489	4274560
4276561	4278095
4279321	4279378
4280060	4280139
4283411	4283467
4287745	4287808
4301533	4303407
4311220	4311292
4317978	4318023
4318331	4319381
4320039	4320111

4343218	4343275
4346675	4346730
4348716	4348831
4350735	4351054
4351065	4352191
4353189	4353441
4353846	4353912
4359121	4359179
4369594	4369716
4374474	4375693
4375752	4376005
4378078	4378150
4383361	4383438
4390041	4390147
4390341	4390403
4400961	4401025
4408688	4408739

Table 3.S2. Variation at lineage defining loci

Patient	Lineage	# of lineage specific loci with multiple allele states	Total # of lineage specific loci	Proportion of lineage-specific loci with >1 allele state
10	L2	9	665	0.013533835
10	L3	1	590	0.001694915
10	L4	5	2843	0.001758706
10	L6	1	220	0.004545455
10	L7	2	898	0.002227171
11	L1	6	1113	0.005390836
11	L2	23	665	0.034586466
11	L3	60	590	0.101694915
11	L4	164	2843	0.057685543
11	L5	1	372	0.002688172
11	L6	1	220	0.004545455
11	L7	3	898	0.003340757
11	BOVAFRI	1	167	0.005988024
14	L1	14	1113	0.012578616
14	L2	46	665	0.069172932
14	L3	8	590	0.013559322
14	L4	165	2843	0.058037285
14	L5	4	372	0.010752688
14	L7	8	898	0.008908686
14	BOVAFRI	1	167	0.005988024
2	L2	1	665	0.001503759

2	L4	6	2843	0.002110447
21	L1	5	1113	0.004492363
21	L2	5	665	0.007518797
21	L7	1	898	0.001113586
22	L1	20	1113	0.017969452
22	L3	1	590	0.001694915
22	L4	3	2843	0.001055223
22	L5	1	372	0.002688172
22	L7	3	898	0.003340757
23	L2	1	665	0.001503759
23	L3	3	590	0.005084746
23	L4	1	2843	0.000351741
23	L7	1	898	0.001113586
3	L4	1	2843	0.000351741
4	L1	1	1113	0.000898473
4	L2	22	665	0.033082707
4	L4	11	2843	0.003869152
5	L2	2	665	0.003007519
5	L4	1	2843	0.000351741
5	L7	2	898	0.002227171
7	L2	2	665	0.003007519
7	L4	6	2843	0.002110447
7	L7	3	898	0.003340757
7	BOVAFRI	1	167	0.005988024
8	L1	1	1113	0.000898473
8	L2	21	665	0.031578947
8	L4	11	2843	0.003869152
8	L7	2	898	0.002227171
9	L2	1	665	0.001503759
9	L4	1	2843	0.000351741
9	L7	2	898	0.002227171

Table 3.S3.1 Coverage statistics

Sample	Patient	Source	Starting Seq.	Aligned Seq. (%)	Final Seq. (%)	Initial Cov.	Final Cov.
ERR867528	1	culture	9148963	98.7	97.9	190.3	165.2
ERR867519	2	sputum	6646608	94.7	92.5	168.7	140.5
ERR867529	2	culture	5953855	98.6	98	131.1	115.2
ERR867520	3	sputum	5294384	89.5	87.5	128.7	108.5
ERR867530	3	culture	3979300	98.9	98.2	83.2	73.8
ERR867521	4	sputum	5517011	75.2	73.1	109.6	91.4
ERR867531	4	culture	4281752	98	97.4	87	76.6
ERR867522	5	sputum	5230228	60.7	56.9	80.3	64.9
ERR867532	5	culture	3796847	98.6	97.9	78.4	68.7

ERR867523	6	sputum	626100	88.8	79.4	15	12.1
ERR867533	6	culture	4280737	98.8	98	90.8	79.9
ERR867524	7	sputum	5517225	82.9	76.6	121	99.4
ERR867534	7	culture	3807941	98.3	97.7	76.9	68
ERR867525	8	sputum	5737989	76.7	71.1	110.8	90.4
ERR867535	8	culture	4046977	98.7	98.1	83.4	73.5
ERR867526	9	sputum	4605729	83.8	81.1	96.9	80.7
ERR867536	9	culture	3304710	98.7	98.1	69.3	61.4
ERR867527	10	sputum	5980880	85.1	81	133.3	109.8
ERR867537	10	culture	3554521	98.6	98	76	67
ERR867538	11	sputum	8556954	39.9	35.8	81.4	67.6
ERR867546	11	culture	7389662	99.1	97.8	164.2	142.3
ERR867547	12	culture	8538092	98.8	97.7	178	153.5
ERR867548	13	culture	7580863	99.1	97.4	163	139.5
ERR867541	14	sputum	13575009	46.3	38.3	135.4	107.8
ERR867549	14	culture	7381945	98.7	97.7	156.2	135.4
ERR867542	15	sputum	13778926	15.6	11.4	38.8	29.9
ERR867550	15	culture	8763567	99.3	98.1	190.2	164.4
ERR867543	16	sputum	1847742	18.7	16.7	7.5	6.3
ERR867551	16	culture	16341030	85	83.3	317	258.7
ERR867544	17	sputum	5218464	3.2	1.9	2.7	1.7
ERR867552	17	culture	1897822	69.4	67.6	28.9	24.2
ERR867545	18	sputum	5133337	2.4	1.3	1.8	1.1
ERR867553	18	culture	2004349	73.5	72.1	33.4	28
ERR867554	19	sputum	2289017	24.2	17.9	11.2	8.7
ERR867560	19	culture	2934225	63.6	61.6	44.1	37.6
ERR867555	20	sputum	10002308	5.9	4.3	11	8.7
ERR867561	20	culture	2812090	79.9	78.1	54.5	45.9
ERR867556	21	sputum	7552762	87	85.5	155.4	123.4
ERR867562	21	culture	3554297	94.5	92.5	81.1	68.9
ERR867557	22	sputum	8510346	85.1	83.6	169.5	134.7
ERR867563	22	culture	3531961	93.3	91.6	79.1	67.5
ERR867558	23	sputum	7562433	82.5	81	146.1	113.2
ERR867564	23	culture	5744613	91.1	89.5	125.7	104.8
ERR867559	24	sputum	9709567	6.4	5	12.1	10.1
ERR867566	25	sputum	3084512	4.1	0.8	1	0.4
ERR867567	26	sputum	2961547	22.2	11.2	9.9	6.1
ERR867568	27	sputum	1407520	17.9	14.9	5.4	4.2
ERR867569	28	sputum	4142037	22.3	0.4	4	0.2
ERR867570	29	sputum	4555597	32	0.7	7.4	0.4

ERR867571	30	sputum	614931	12.4	0.9	0.9	0.1
ERR867572	31	sputum	418642	43.8	2.6	2.5	0.2
ERR867573	32	sputum	769048	30.6	3.7	2.5	0.5
ERR867574	33	sputum	998353	11.9	1.5	1.4	0.3
ERR867575	34	sputum	1154247	25.4	9.1	4.3	1.8
MbWH01c	35	culture	3477683	99.6	99.2	97.6	85.7
MbWH01s	35	sputum	42506627	32	31.1	217.1	191.4
MbWH02c	35	culture	3736764	99.4	98.9	104.4	97.6
MbWH02s	35	sputum	33145172	1.9	0.9	7.2	5.2
MbWH03c	35	culture	4061325	99.3	98.8	113.4	99.8
MbWH03s	35	sputum	25292925	17.3	16.5	78.9	70.1
MbWH04c	35	culture	5117264	98.6	98	141.6	122
MbWH04s	35	sputum	34599011	17.6	16.6	104.4	92.2
Negative Sputum	NA	sputum	41013676	0.3	0	0.6	0

Table 3.S3.2. Lineage typing of *M. tb* samples

Sample	Patient	Source	Lineage
ERR867519	2	sputum	Lineage 4
ERR867529	2	culture	Lineage 4
ERR867520	3	sputum	Lineage 4
ERR867530	3	culture	Lineage 4
ERR867521	4	sputum	Lineage 2
ERR867531	4	culture	Lineage 2
ERR867522	5	sputum	Lineage 2
ERR867532	5	culture	Lineage 2
ERR867524	7	sputum	Lineage 4
ERR867534	7	culture	Lineage 4
ERR867525	8	sputum	Lineage 2
ERR867535	8	culture	Lineage 2
ERR867526	9	sputum	Lineage 2
ERR867536	9	culture	Lineage 2
ERR867527	10	sputum	Lineage 2
ERR867537	10	culture	Lineage 2
ERR867538	11	sputum	Lineage 4
ERR867546	11	culture	Lineage 4
ERR867541	14	sputum	Lineage 4
ERR867549	14	culture	Lineage 4
ERR867556	21	sputum	Lineage 2
ERR867562	21	culture	Lineage 2
ERR867557	22	sputum	Lineage 1

ERR867563	22	culture	Lineage 1
ERR867558	23	sputum	Lineage 3
ERR867564	23	culture	Lineage 3

Table 3.S4. Smear Score ANOVA of *M. tb* Samples

Pairs	Adjusted p-value
5AFB-Negative	0.2983938
9AFB-Negative	0.9992703
1-Negative	0.0013321
2-Negative	0.0001493
3-Negative	0.0000108
9AFB-5AFB	0.8941813
1-5AFB	0.9998597
2-5AFB	0.8313938
3-5AFB	0.9399742
1-9AFB	0.7028359
2-9AFB	0.2969369
3-9AFB	0.3947471
1-2	0.5697069
1-3	0.6880947
2-1	0.9846854

Table 3.S5.1. *M. tb* windows of diversity (π) overlap

Sputum

Window	Frequency
740000	2
750000	2
760000	2
770000	2
780000	2
790000	2
800000	2
810000	2
1440000	11
1450000	11
1460000	11
1470000	11
1480000	10
1490000	10
1500000	10
1510000	10

Culture

Window	Frequency
1440000	10
1450000	10
1460000	10
1470000	10
1480000	10
1490000	10
1500000	10
1510000	10
1520000	10
1530000	10
2220000	2
2230000	2
2240000	2
2250000	2
2260000	2
2270000	2

1520000	10
1530000	8
2250000	2
2260000	2
2270000	2
2280000	2
2290000	4
2300000	4
2310000	5
3370000	2
3380000	2
3390000	2
3400000	2
3410000	2
3420000	3
3430000	3
3440000	3
3450000	2

2280000	2
2290000	6
2300000	7
2310000	7
2320000	2
2330000	2
2590000	5
2600000	5
2610000	5
2620000	5
2630000	5
2640000	5
2650000	5
2660000	5
2670000	5
2680000	4
3800000	4
3810000	2
3880000	2
3890000	4

Table 3.S5.2. *Mb* windows of diversity (π) overlap

Culture

Window	Frequency
210000	2
220000	2
230000	2
240000	2
3260000	2
3270000	2
3280000	2
3290000	2
3300000	2
3310000	2
3730000	2
3740000	2
3750000	2

Sputum

Window	Frequency
740000	3
750000	3
760000	3
770000	3
780000	3

Table 3.S6.1 Outlier genes: lm and fold change

Gene	Number of patients/samples: lm outlier
Rv2020c	13
Rv3109	11
Rv1318c	11
Rv1267c	10
Rv1319c	9
Rv2351c	5
Rv2350c	5
Rv3051c	4
Rv2082	3
Rv2319c	2
Rv2081c	2
Rv1630	2
Rv1327c	2
Rv1164	2
Rv0684	2
Rv0667	2
Rv0338c	2
Mb3111	2
Mb0999c	2
Mb0948c	2
Mb0508	2
Mb0217	3
Mb0177	2
Mb0174	1
Mb0704	1
Mb0826	1
Mb0974	1
Mb1048	1
Mb1131c	1
Mb1554c	1
Mb2118c	1
Mb2521	1
Mb2981	1
Mb3284c	1
Mb3324	1
Rv0006	1
Rv0041	1
Rv0211	1
Rv0668	1

Gene	Number of patients/samples: Fold change outlier
Mb0948c	2
Mb0974	2
Rv0667	3
Rv1630	2
Rv3051c	6
Mb0826	1
Mb1048	1
Mb1695c	1
Mb2049c	1
Mb3324	1
Rv0037c	1
Rv0058	1
Rv0089	1
Rv0101	1
Rv0182c	1
Rv0271c	1
Rv0319	1
Rv0338c	1
Rv0585c	1
Rv0591	1
Rv0630c	1
Rv0668	1
Rv0684	1
Rv0685	1
Rv0731c	1
Rv0753c	1
Rv0800	1
Rv0803	1
Rv0839	1
Rv1007c	1
Rv1020	1
Rv1164	1
Rv1165	1
Rv1166	1
Rv1249c	1
Rv1250	1
Rv1266c	1
Rv1280c	1
Rv1315	1

Rv0685	1
Rv0753c	1
Rv0803	1
Rv0950c	1
Rv0965c	1
Rv1020	1
Rv1165	1
Rv1250	1
Rv1266c	1
Rv1322	1
Rv1380	1
Rv1413	1
Rv1475c	1
Rv1485	1
Rv1629	1
Rv1638	1
Rv2023A	1
Rv2245	1
Rv2318	1
Rv2572c	1
Rv2610c	1
Rv3113	1
Rv3148	1
Rv3526	1
Rv3573c	1
Rv3885c	1
Rv3900c	1
Rv3901c	1

Rv1327c	1
Rv1336	1
Rv1365c	1
Rv1380	1
Rv1410c	1
Rv1425	1
Rv1453	1
Rv1467c	1
Rv1485	1
Rv1530	1
Rv1564c	1
Rv1629	1
Rv1638	1
Rv1659	1
Rv1698	1
Rv1769	1
Rv1970	1
Rv1997	1
Rv2006	1
Rv2097c	1
Rv2150c	1
Rv2163c	1
Rv2201	1
Rv2230c	1
Rv2391	1
Rv2531c	1
Rv2572c	1
Rv2590	1
Rv2610c	1
Rv2845c	1
Rv2888c	1
Rv2940c	1
Rv2942	1
Rv2967c	1
Rv3261	1
Rv3305c	1
Rv3447c	1
Rv3448	1
Rv3612c	1
Rv3762c	1
Rv3885c	1

Rv3903c	1

Table 3.S6.2 Outlier genes: Absolute change and F_{ST}

Gene	Number of patients/samples: Absolute change outlier
Rv0708	6
Rv3648c	4
Rv0640	3
Rv0378	2
Rv0668	2
Rv0684	2
Rv0717	2
Rv0748	2
Rv1164	2
Rv1824	2
Rv2272	2
Rv2703	2
Rv2904c	2
Rv3051c	2
Rv3604c	2
Mb3177	2
Mb1047	3
Mb0659	3
Rv0024	1
Rv0049	1
Rv0053	1
Rv0081	1
Rv0084	1
Rv0158	1
Rv0253	1
Rv0329c	1
Rv0371c	1
Rv0392c	1
Rv0417	1
Rv0444c	1
Rv0467	1
Rv0492A	1
Rv0500B	1
Rv0512	1
Rv0514	1
Rv0525	1
Rv0569	1
Rv0580c	1
Rv0653c	1

Gene	Number of patients/samples: F_{ST} outlier
Rv1318c	11
Rv3109	8
Rv3051c	7
Rv1319c	6
Rv2020c	6
Rv1164	5
Rv1267c	5
Rv2351c	5
Rv0667	4
Rv0668	4
Rv2350c	4
Rv0001	3
Rv0005	3
Rv0684	3
Rv1248c	3
Rv1475c	3
Rv1536	3
Rv2082	3
Rv2357c	3
Rv2823c	3
Rv3240c	3
Rv3859c	3
Rv0006	2
Rv0050	2
Rv0206c	2
Rv0244c	2
Rv0282	2
Rv0338c	2
Rv0350	2
Rv0410c	2
Rv0440	2
Rv0642c	2
Rv0655	2
Rv0663	2
Rv0711	2
Rv0777	2
Rv0896	2
Rv0987	2
Rv1327c	2

Rv0658c	1
Rv0667	1
Rv0685	1
Rv0700	1
Rv0745	1
Rv0847	1
Rv0871	1
Rv0889c	1
Rv0900	1
Rv0956	1
Rv1013	1
Rv1014c	1
Rv1087A	1
Rv1106c	1
Rv1107c	1
Rv1116	1
Rv1143	1
Rv1177	1
Rv1218c	1
Rv1305	1
Rv1323	1
Rv1398c	1
Rv1417	1
Rv1535	1
Rv1536	1
Rv1630	1
Rv1744c	1
Rv1767	1
Rv1799	1
Rv1852	1
Rv1856c	1
Rv1863c	1
Rv1893	1
Rv1953	1
Rv1988	1
Rv2010	1
Rv2023c	1
Rv2057c	1
Rv2063	1
Rv2111c	1
Rv2185c	1

Rv1527c	2
Rv1630	2
Rv1638	2
Rv2023A	2
Rv2081c	2
Rv2115c	2
Rv2349c	2
Rv2435c	2
Rv2444c	2
Rv2477c	2
Rv2583c	2
Rv2606c	2
Rv2737c	2
Rv2783c	2
Rv2939	2
Rv3048c	2
Rv3060c	2
Rv3148	2
Rv3411c	2
Rv3436c	2
Rv3597c	2
Rv3783	2
Rv3884c	2
Rv3894c	2
Rv0014c	1
Rv0015c	1
Rv0032	1
Rv0037c	1
Rv0041	1
Rv0054	1
Rv0055	1
Rv0058	1
Rv0064	1
Rv0066c	1
Rv0071	1
Rv0126	1
Rv0171	1
Rv0173	1
Rv0189c	1
Rv0202c	1
Rv0207c	1

Rv2198c	1
Rv2319c	1
Rv2422	1
Rv2444c	1
Rv2520c	1
Rv2530A	1
Rv2530c	1
Rv2606c	1
Rv2621c	1
Rv2728c	1
Rv2744c	1
Rv2765	1
Rv2799	1
Rv2804c	1
Rv2806	1
Rv2830c	1
Rv2840c	1
Rv2848c	1
Rv2862c	1
Rv2869c	1
Rv2901c	1
Rv2907c	1
Rv2922A	1
Rv3048c	1
Rv3066	1
Rv3171c	1
Rv3175	1
Rv3183	1
Rv3209	1
Rv3319	1
Rv3341	1
Rv3353c	1
Rv3418c	1
Rv3461c	1
Rv3462c	1
Rv3489	1
Rv3516	1
Rv3527	1
Rv3583c	1
Rv3587c	1
Rv3596c	1

Rv0211	1
Rv0247c	1
Rv0248c	1
Rv0252	1
Rv0267	1
Rv0272c	1
Rv0284	1
Rv0322	1
Rv0337c	1
Rv0384c	1
Rv0394c	1
Rv0407	1
Rv0418	1
Rv0423c	1
Rv0430	1
Rv0457c	1
Rv0462	1
Rv0470c	1
Rv0500A	1
Rv0510	1
Rv0512	1
Rv0529	1
Rv0563	1
Rv0592	1
Rv0640	1
Rv0644c	1
Rv0647c	1
Rv0676c	1
Rv0683	1
Rv0685	1
Rv0703	1
Rv0704	1
Rv0705	1
Rv0707	1
Rv0714	1
Rv0718	1
Rv0719	1
Rv0732	1
Rv0753c	1
Rv0761c	1
Rv0782	1

Rv3642c	1
Rv3656c	1
Rv3688c	1
Rv3692	1
Rv3723	1
Rv3753c	1
Rv3771c	1
Rv3777	1
Rv3911	1
Mb1109c	1
Mb0920	1
Mb3077c	1
Mb3703c	1

Rv0785	1
Rv0803	1
Rv0806c	1
Rv0821c	1
Rv0822c	1
Rv0831c	1
Rv0839	1
Rv0859	1
Rv0861c	1
Rv0885	1
Rv0913c	1
Rv0946c	1
Rv0949	1
Rv0950c	1
Rv0951	1
Rv0969	1
Rv0974c	1
Rv0985c	1
Rv0988	1
Rv1013	1
Rv1020	1
Rv1069c	1
Rv1092c	1
Rv1097c	1
Rv1112	1
Rv1118c	1
Rv1130	1
Rv1133c	1
Rv1143	1
Rv1161	1
Rv1162	1
Rv1165	1
Rv1173	1
Rv1187	1
Rv1194c	1
Rv1213	1
Rv1223	1
Rv1237	1
Rv1246c	1
Rv1266c	1
Rv1272c	1

Rv1286	1

Chapter 4: Bayesian and Site Frequency Based Inference of *Mycobacterium tuberculosis'* Migratory Histories

SFS based inference results published in:
Mary B O'Neill, Abigail Shockey, Alex Zarley, William Aylward, Vegard Eldholm, Andrew Kitchen and Caitlin S Pepperell 2019
May 2019, Molecular Ecology, doi:10.1111/mec.15120

Bayesian based inference results pre-print in:
Claire V. Mulholland, Abigail C. Shockey, Htin L. Aung, Ray T. Cursons, Ronan F. O'Toole, Sanjay S. Gautam, Daniela Brites, Sebastien Gagneux, Sally A. Roberts, Noel Karalus, Gregory M. Cook, Caitlin S. Pepperell, and Vickery L. Arcus
May 2019, bioRxiv, doi:10.1101/631937

O'Neill et al. 2019 Authorship Contributions:
CSP, MBO, and AK conceived and designed the project. MBO, AK, AS, and AZ performed the analyses and all authors interpreted the data. CSP and MBO drafted the manuscript and all authors provided critical feedback, reviewed, and edited the manuscript.

Mulholland et al. 2019 Authorship Contributions:
CVM: Formal Analysis, Investigation, Methodology, Visualization, Writing – Original Draft Preparation. ACS: Formal Analysis, Investigation, Methodology, Visualization. HLA: Investigation. RTC: Conceptualization, Supervision. RFO: Investigation. SSG: Investigation. DB: Investigation, Writing – Review & Editing. SG: Investigation, Writing – Review & Editing. SAR: Resources. NK: Resources. GMC: Conceptualization, Funding Acquisition. CSP: Conceptualization, Supervision, Methodology, Writing – Review & Editing. VLA: Conceptualization, Supervision, Methodology, Funding Acquisition, Writing – Review & Editing

Abstract

Mycobacterium tuberculosis (*M. tb*) is a globally distributed obligate human pathogen that can be divided into seven distinct lineages. Recently, Bayesian reconstruction of *M. tb* migration patterns from a sample of 552 globally distributed *M. tb* isolates suggest the current distribution of Lineage 1 (L1) arose from migrations out of India. Additionally, phylogenetic analysis of 236 *M. tb* Lineage 4 (L4) isolates revealed sub-lineage L4.4.1.1 isolates from Indigenous Maori and Pacific people in New Zealand nest within the diversity of L4.4.1.1 isolates from Indigenous populations in Canada. Previous studies have demonstrated *M. tb* was introduced to Canadian Indigenous populations by French-Canadian fur traders. It is possible the introduction of *M. tb* to the South Pacific was driven by similar European colonial migrations and trade expansion, but little is known about the dispersal of *M. tb* in this region. Here, we sought to reconstruct the migratory history of *M. tb* L1 using SFS based methods of migration inference and L4.4.1.1 using Bayesian based methods of migration interference. Reconstructing the migratory history of *M. tb* L1, we found that SFS based methods of inference perform poorly when two population migration models are used. Our Bayesian inference of *M. tb* L4.4.1.1's migratory history is indicative of multiple migrations of *M. tb* from Europe to Indigenous populations in Canada and New Zealand. We linked these patterns of migration to historical phenomena, specifically colonial migrations and trade expansion. Our results demonstrate the power of using a pathogen's genomic data in concert with spatial and temporal data to trace its origins and contribute to the growing body of work showing a clear effect of human migration on the current global distribution of *M. tb*.

Introduction

As bacterial pathogens diversify during infection and spread to new hosts, they acquire informative mutations that can be used to reconstruct their migratory history. When combined, the phylogeography of bacterial pathogens and historical phenomena associated with their

hosts can elucidate the dynamics underlying epidemics, which can be used to inform control strategies.

Mycobacterium tuberculosis (*M. tb*), the causative agent of tuberculosis (TB), is a globally distributed obligate human pathogen that can be divided into seven distinct lineages. These lineages differ in their global distribution and have strong phylogeographic structure (1–3).

Lineage 1 (L1) of *M. tb*, also known as the “Indo-Oceanic” lineage, primarily occurs in regions bordering the Indian Ocean and is common in areas of East Africa and Southeast Asia with a high burden of TB (4). Bayesian reconstruction of *M. tb* migration patterns from a sample of 552 globally distributed *M. tb* isolates (see Appendix 1) suggest the current distribution of L1 arose from migrations out of India. Despite research suggesting inference of bacterial demography using the site frequency spectrum (SFS) is less sensitive to selection than more commonly used Bayesian methods (5), few studies have used SFS based methods to infer the migratory history of bacterial pathogens. We sought to reconstruct the migratory history of *M. tb* L1 described above using SFS based methods.

M. tb Lineage 4 (L4), also known as the “Euro-American” lineage, is the most widely globally dispersed lineage of *M. tb* (2). Ten sub-lineages of L4 have been described (6), and recent reconstructions of L4’s dispersal suggest it was spread through European colonial migrations to Africa and the Americas (1, 2, 7–9). Phylogenetic analysis of 236 *M. tb* L4.4 isolates (see Appendix 1) revealed *M. tb* isolates from Indigenous Maori and Pacific people in New Zealand nest within the diversity of *M. tb* isolates from Indigenous populations in Canada. *M. tb* from both populations belong to the L4.4 sub-lineage L4.4.1.1. The Canadian *M. tb* isolates belong to the $DS6^{Quebec}$ (DS6Q) lineage, which was dispersed to Western Aboriginal Canadians by French-Canadian fur traders in the 18th–19th centuries (9). It is possible the introduction of DS6Q to the South Pacific was driven by similar European colonial migrations and trade expansion. Contemporary *M. tb* genotypes in New Zealand Europeans, Māori and

Pacific people in New Zealand are dominated by L4 strains, consistent with introduction of contemporary strains by Europeans, but little is known about the dispersal of *M. tb* in this region. In order to trace the origins and dispersal of *M. tb* in New Zealand, we reconstructed the migratory history of *M. tb* L4.4.1.1 described above using Bayesian methods of migration inference.

Methods

Sample descriptions

Lineage 1: Our sample of *M. tb* L1 consists of 89 isolates from a globally distributed sample of *M. tb* L1-L7 (*n* = 552; See Appendix 1 for details of sampling strategy, assembly, alignment and variant calling). These 89 *M. tb* L1 isolates span 22 countries and 6 United Nations (UN) subregions (Table 4.S1). The alignment of *M. tb* L1 consists of 60787 SNPs.

Lineage 4.4.1.1: Our sample of *M. tb* L4.4.1.1 consists of 117 isolates from a globally distributed sample of *M. tb* L4.4 (*n* = 236; See Appendix 1 for details of sampling strategy, assembly, alignment and variant calling). These 117 *M. tb* L4.4.1.1 isolates span 19 countries and 5 UN subregions (Table 4.S2). The alignment of *M. tb* L4.4.1.1 consists of 3161 SNPs.

SFS based inference of v for *M. tb* L1 sub-populations

We used SNP-sites (10) to convert our alignment of *M. tb* L1 isolates to a multi-sample VCF and SnpEff (11) to annotate variants with respect to H37Rv [NC_000962.3; (12)] as synonymous, non-synonymous, or intergenic. We removed loci at which any sequence in the population had a gap or unknown character from the data set. We used EasySFS (https://github.com/isaacovercast/easySFS) to convert synonymous variants in the multi-sample VCF of L1 to two-dimensional synonymous SFS (sSFS) and one-dimensional sSFS. We defined the two extant sub-populations as India and the rest of the world (RoW) and projected population sizes from *n* = 31, 58 to *n* = 25, 25 (India and RoW, respectively). In order to determine v (N_e contemporary/N_e ancestral; the magnitude of effective population size change)

for India and RoW, we performed demographic inference with the one-dimensional sSFS of India and RoW using ∂a∂i (13). For each sub-population, we modeled constant population size (standard neutral model) and instantaneous expansion. We optimized our parameter estimates, v and τ (time since expansion).

SFS based migration inference of *M. tb* L1

We performed migration inference with the two-dimensional sSFS of India and RoW using ∂a∂I. We modeled no split (standard neutral model), a split with no migration, a split with symmetric migration, a split with unidirectional migration (India to RoW), and a split with asymmetric migration. We identified the best-fit model and maximal likelihood parameters of the migration model given our observed data. Parameters v1 and v2 (magnitude of effective population size change for India and RoW, respectively) were fixed according to their values estimated from the model of instantaneous expansion (described above). Our parameter estimates m and τ (migration and time since population split, respectively), were optimized for each migration model. We used the Akaike information criterion (AIC) estimator for model selection and calculated the Poisson residuals between model and data for each best-fit model.

Phylogeographic inference of *M. tb* L4.4.1.1.

With the SNP alignment of *M. tb* L4.4.1.1, we performed ancestral reconstruction using BEAST2 (14), with UN region for each isolate modelled as a discrete trait. Analyses were performed using the GTR model of nucleotide substitution, a strict molecular clock with an estimated substitution rate of 6.28×10^{-8} s/s/y and BSP demographic models (see Appendix 1 for description of rate estimation and model and clock selection). For the Bayesian skyline model, we deselected the Jeffrey's (1/X) prior for the population size parameter; this is an improper prior unsuitable for model evaluation using path sampling (see Appendix 1). We used default priors for the remaining parameters. We ran three independent chains for 350 million states sampling every 10000 states. We discarded the first 10% of states as burn-in and assessed chains for convergence and sufficient mixing (effective sample size > 200 for all

parameters). We combined samples from the three independent chains and based parameter estimation on the combined chain. We inferred the maximum clade credibility (MCC, median heights) from the combined tree samples in TreeAnnotator.

We inferred migration rates over time from the MCC tree. As described in O'Neill et al. 2019 (15), migration events were defined as a change in the most probable reconstructed state from parent to child node. Only nodes with a posterior probability > 80% were considered. Median heights of the parent and child nodes were treated as the range of time in which a migration event could occur. Migration rates through time were inferred by summing the number of migration events during each year of the phylogeny, divided by the total number of branches in existence during each year of the phylogeny. We used the Bayesian stochastic search variable selection method (BSSVS) implemented in BEAST2 to identify well-supported (Bayes factor > 5) migration rates between UN regions in the phylogeographic analyses and SpreaD3 (16) to calculate Bayes factor for each pairwise rate.

Results

SFS Based migration inference of L1

Results of the SFS based migration interference of L1 suggest a symmetric model of migration best fit our observed data (Figure 4.1). However, all tested models fit poorly, as indicated by clustering of residuals from the two-dimensional sSFS. The method was therefore unreliable for model selection (Figure 4.2, Table 1).

Bayesian based migration Inference of L4.4.1.1

The BSP suggests L4.4.1.1 underwent a rapid population expansion following its emergence, and corresponding migration analyses show a spike in migration at this time (Figure 4.3). This was followed by a period where the population size remained consistent until another period of population growth in the 19th century, during which time migration tapers off.

Our phylogeographic reconstruction of L4.4.1.1 is indicative of migration from Africa to Asia, as well as Europe to Canada and Oceania (Figure 4.4, Figure 4.5). Our estimated time to most recent common ancestor (TMRCA) of the DS6Q clade is 1652 (95% HPD, 1535–1741). We inferred two separate introductions of *M. tb* to New Zealand (Figure 4.4). The estimated TMRCA of the first, older introduction is 1813 (95% HPD, 1746–1868). The estimated TMRCA of the second, more recent introduction is 1980 (95% HPD, 1969–1988).

Discussion

Using SFS and Bayesian based methods, we have inferred the migratory histories of two distinct lineages of *M. tb*. In reconstructing the migratory history of *M. tb* L1, we found that SFS based methods of inference perform poorly when complex models are used. Our Bayesian inference of *M. tb* L4.4.1.1's migratory history is indicative of multiple migrations of closely related DS6Q strains from Europe to Canada and the South Pacific that are likely the result of colonial migrations.

All tested models for the SFS based inference of *M. tb* L1's migratory history fit poorly, as indicated by clustering of residuals from the two-dimensional sSFS (Figure 4.2). These results are consistent with prior research indicating that SFS based methods perform well for fully linked genomes when inference is done under very simple models but not more complex models (17).

The polytomy at the root of the DS6Q clade and the polyphyletic nature of the New Zealand and Canadian isolates (Figure 4.4) suggests dispersal of several closely related strains from a common origin. The TMRCA estimate for the DS6Q clade of M. tb L4.4.1.1, (mid-17th century) coincides with French migration to Quebec. This TMRCA is consistent with a French origin as previously reported for the DS6Q lineage [Figure 4.4; (9)]. The early 19th century TMRCA estimate for the introduction of *M. tb* to New Zealand fits with an introduction to Polynesia by French/European whalers or other traders. Whaling was the only French economic

activity of any scale in the South Pacific during the early 19th century (18–20). During the whaling era, Polynesian men were often recruited as crew on whaling ships and accounted for up to one-fifth of European whaling crews (21, 22). These interactions would have been conducive for the dispersal of *M. tb*. Additionally, unlike New Zealand, which received a large ingress of European migrants after British annexation in 1840 (24), other Polynesian islands did not experience the same influx of European emigrants. This lends further support to a trade-associated introduction of *M. tb* to New Zealand.

Our results indicate the younger cluster of *M. tb* isolates in New Zealand arose from a relatively recent clonal expansion. This could be the result of a more recent introduction of *M. tb* or emergence from a previously introduced, unsampled DS6Q strain. The TMRCA of this cluster (late 20th century) follows a period of mass migration to New Zealand from the Pacific Islands in the 1950s–1970s, a plausible route of introduction. Although it is evident that this strain ultimately emerged from a strain of European origin, more in-depth sampling of L4.4.1.1 isolates from both New Zealand and the Pacific may provide a clearer picture of the dispersal of this strain from Europe to New Zealand and will shed additional light on the dispersal of this sub-lineage in this region.

The TMRCA of the younger cluster of *M. tb* isolated in New Zealand coincides with major demographic changes in New Zealand's Māori population that occurred in the mid-20th century. Māori experienced a fast rate of urbanization between 1945–1980 (25). Urbanization was accompanied by overcrowded housing and increased incarceration rates, both of which are TB risk factors (26, 27). The temporal association between emergence of the younger New Zealand *M. tb* cluster and the urbanization of Māori suggests that human social phenomena are important contributors to the demographic history of *M. tb*.

A recent reconstruction of the migratory history of L4, which did not include isolates from the South Pacific, suggest global dispersal of L4 was dominated by historical migrations out of Europe and connected with European colonial migrations (7). We observe a similar scenario in

our reconstruction of the migratory history of the DS6Q clade: Several closely related *M. tb* strains dispersed from Europe in the 17th–19th centuries to remote and unconnected populations via European colonial migrations and expanding trade networks.

Our data indicate L4.4.1.1 expanded in the 19th century (Figure 4.3), which could be attributed to colonial activities during this time involving countries in our sample; the French-Canadian fur trade [1710–1870; (28)], the South Pacific whaling trade [1790-1860; (20)], and the rapid occupation and colonization of much of the African continent during the New Imperialism period [1876–1912; (29)]. The decline in effective population size in the late 20th century of our sample of *M. tb* observed in the BSP (Figure 4.3) coincides with the dramatic decline in TB incidence in the developed world over the last century. Our results highlight the power of phylogeographic methods and the utilization of WGS data to reconstruct a pathogen's migratory history in high resolution and reasserts the role of European migrations in the dispersal of *M. tb* L4.

Acknowledgements

We would like to thank Mary O'Neill and Claire Mulholland for providing the data, the work they have done, and their expertise and insight on this project. AS and CP are supported by the National Institutes of Health (R01AI113287). Funding for this project was also provided by the University of Wisconsin School of Medicine and Public Health from the Wisconsin Partnership Program.

References

1. Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW, Small PM. 2004. Stable association between strains of Mycobacterium tuberculosis and their human host populations. Proceedings of the National Academy of Sciences of the United States of America 101:4871–6.

2. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, Jong BC de, Narayanan S, Nicol M, Niemann S, Kremer K, Gutierrez MC, Hilty M, Hopewell PC, Small PM. 2006. Variable host–pathogen compatibility in Mycobacterium tuberculosis. PNAS 103:2869–2873.
3. Gagneux S. 2018. Ecology and evolution of Mycobacterium tuberculosis. Nature Reviews Microbiology 16:202.
4. Palittapongarnpim P, Ajawatanawong P, Viratyosin W, Smittipat N, Disratthakit A, Mahasirimongkol S, Yanai H, Yamada N, Nedsuwan S, Imasanguan W, Kantipong P, Chaiyasirinroje B, Wongyai J, Toyo-oka L, Phelan J, Parkhill J, Clark TG, Hibberd ML, Ruengchai W, Palittapongarnpim P, Juthayothin T, Tongsima S, Tokunaga K. 2018. Evidence for Host-Bacterial Co-evolution via Genome Sequence Analysis of 480 Thai Mycobacterium tuberculosis Lineage 1 Isolates. Scientific Reports 8:11597.
5. Lapierre M, Blin C, Lambert A, Achaz G, Rocha EPC. 2016. The Impact of Selection, Gene Conversion, and Biased Sampling on the Assessment of Microbial Demography. Mol Biol Evol 33:1711–1725.
6. Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, Fenner L, Rutaihwa L, Borrell S, Luo T, Gao Q, Kato-Maeda M, Ballif M, Egger M, Macedo R, Mardassi H, Moreno M, Tudo Vilanova G, Fyfe J, Globan M, Thomas J, Jamieson F, Guthrie JL, Asante-Poku A, Yeboah-Manu D, Wampande E, Ssengooba W, Joloba M, Henry Boom W, Basu I, Bower J, Saraiva M, Vaconcellos SEG, Suffys P, Koch A, Wilkinson R, Gail-Bekker L, Malla B, Ley SD, Beck H-P, de Jong BC, Toit K, Sanchez-Padilla E, Bonnet M, Gil-Brusola A, Frank M, Penlap Beng VN, Eisenach K, Alani I, Wangui Ndung'u P, Revathi G, Gehre F, Akter S, Ntoumi F, Stewart-Isherwood L, Ntinginya NE, Rachow A, Hoelscher M, Cirillo DM, Skenders G, Hoffner S, Bakonyte D, Stakenas P, Diel R, Crudu V, Moldovan O, Al-Hajoj S, Otero L, Barletta F, Jane Carter E, Diero L, Supply P, Comas I, Niemann S, Gagneux S. 2016. Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. Nat Genet 48:1535–1543.
7. Brynildsrud OB, Pepperell CS, Suffys P, Grandjean L, Monteserin J, Debech N, Bohlin J, Alfsnes K, Pettersson JO-H, Kirkeleite I, Fandinho F, da Silva MA, Perdigao J, Portugal I, Viveiros M, Clark T, Caws M, Dunstan S, Thai PVK, Lopez B, Ritacco V, Kitchen A, Brown TS, van Soolingen D, O'Neill MB, Holt KE, Feil EJ, Mathema B, Balloux F, Eldholm V. 2018. Global expansion of Mycobacterium tuberculosis lineage 4 shaped by colonial migration and local adaptation. Sci Adv 4:eaat5869–eaat5869.

8. Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, Roach JC, Kremer K, Petrov DA, Feldman MW, Gagneux S. 2008. High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography. PLoS Biol 6:e311.
9. Pepperell CS, Granka JM, Alexander DC, Behr MA, Chui L, Gordon J, Guthrie JL, Jamieson FB, Langlois-Klassen D, Long R, Nguyen D, Wobeser W, Feldman MW. 2011. Dispersal of Mycobacterium tuberculosis via the Canadian fur trade. PNAS 108:6526–6531.
10. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. biorxiv;038190v1.
11. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly 6:80–92.
12. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, III CEBI, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S. 1998. Erratum: Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature; London 396:190.
13. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLoS Genet 5:e1000695.
14. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Comput Biol 10:e1003537.
15. O'Neill MB, Shockey A, Zarley A, Aylward W, Eldholm V, Kitchen A, Pepperell CS. 2019. Lineage specific histories of Mycobacterium tuberculosis dispersal in Africa and Eurasia. Molecular Ecology 0.
16. Bielejec F, Baele G, Vrancken B, Suchard MA, Rambaut A, Lemey P. 2016. SpreaD3: Interactive Visualization of Spatiotemporal History and Trait Evolutionary Processes. Mol Biol Evol 33:2167–2169.
17. Pepperell CS, Casto AM, Kitchen A, Granka JM, Cornejo OE, Holmes EC, Birren B, Galagan J, Feldman MW. 2013. The Role of Selection in Shaping Diversity of Natural M. tuberculosis Populations. PLoS Pathog 9:e1003543.
18. Maclellan N, Chesneaux J. 1998. After Moruroa: France in the South Pacific. Ocean Press.
19. Foucrier A. 2005. The French and the Pacific World, 17th-19th Centuries: Explorations, Migrations, and Cultural Exchanges. Ashgate.

20. Haines D. 2010. Lighting up the World? Empires and Islanders in the Pacific Whaling Industry, 1790-1860Maritime History as Global History. Liverpool University Press.
21. Chappell DA. 1997. Double Ghosts: Oceanian Voyagers on Euroamerican Ships. M.E. Sharpe.
22. Fischer SR. 2013. A history of the Pacific Islands / Steven Roger Fischer. Palgrave Macmillan, Basingstoke, Hampshire [England] ; New York.
23. Lange R. 1984. PLAGUES AND PESTILENCE IN POLYNESIA: THE NINETEENTH-CENTURY COOK ISLANDS EXPERIENCE. Bulletin of the History of Medicine 58:325–346.
24. United Nations. Economic and Social Commission for Asia and the Pacific. 1985. Population of New Zealand. UN.
25. R J-L, Pool I. 1991. Te Iwi Maori, a New Zealand Population, Past, Present and Projected.
26. Baussano I, Williams BG, Nunn P, Beggiato M, Fedeli U, Scano F. 2010. Tuberculosis Incidence in Prisons: A Systematic Review. PLOS Medicine 7:e1000381.
27. Clark M, Riben P, Nowgesic E. 2002. The association of housing density, isolation and tuberculosis in Canadian First Nations communities. International Journal of Epidemiology 31:940–945.
28. Innis HA, Ray AJ. 1999. The Fur Trade in Canada: An Introduction to Canadian Economic History. University of Toronto Press.
29. Pakenham T. 2015. The Scramble For Africa. Little, Brown Book Group.

Figures

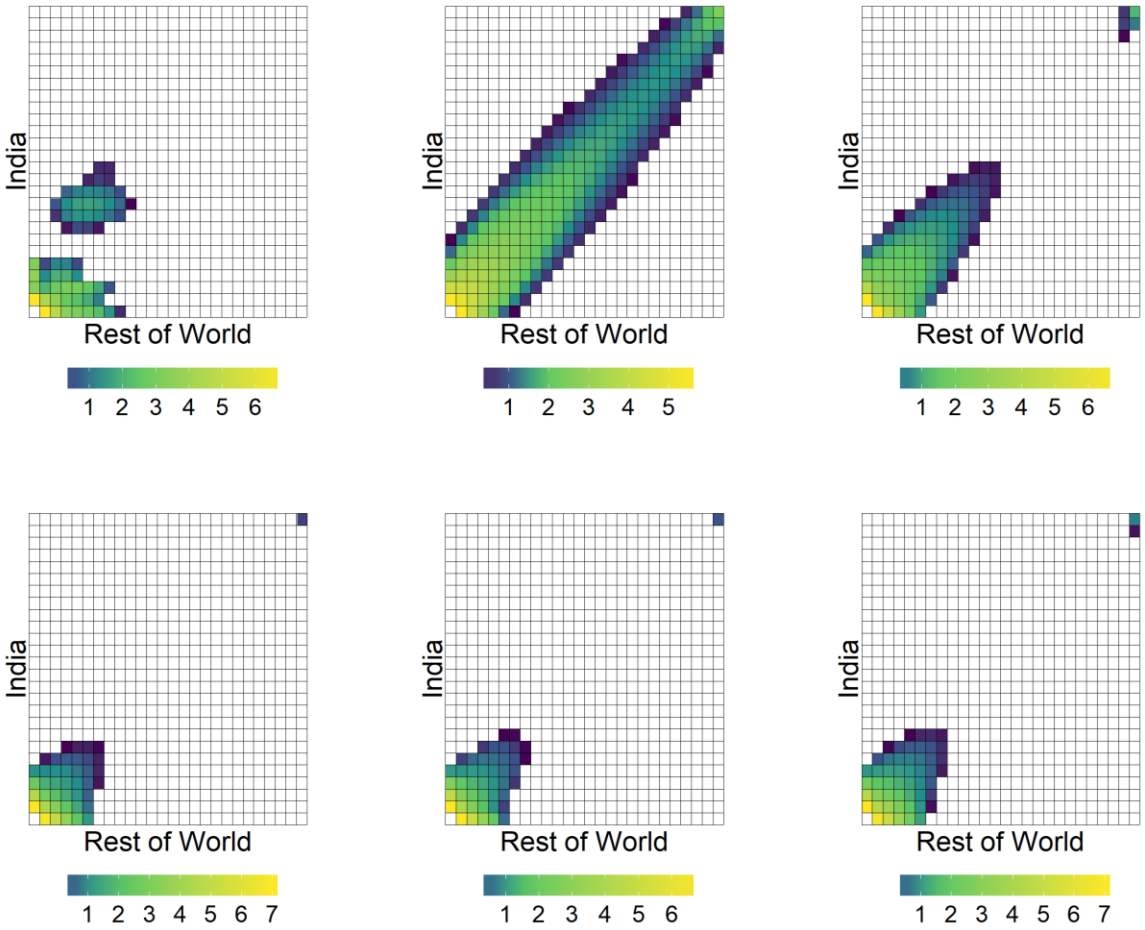

Figure 4.1. Observed and inferred two-dimensional sSFS of India and RoW. Heatmaps of two-dimensional spectra colored by number of SNPs at each frequency in the population (log-transformed). Masked data and SNP counts = 0 in white. From left to right, top to bottom: Observed, no split and no migration, no migration, symmetric migration, unidirectional migration, and asymmetric migration.

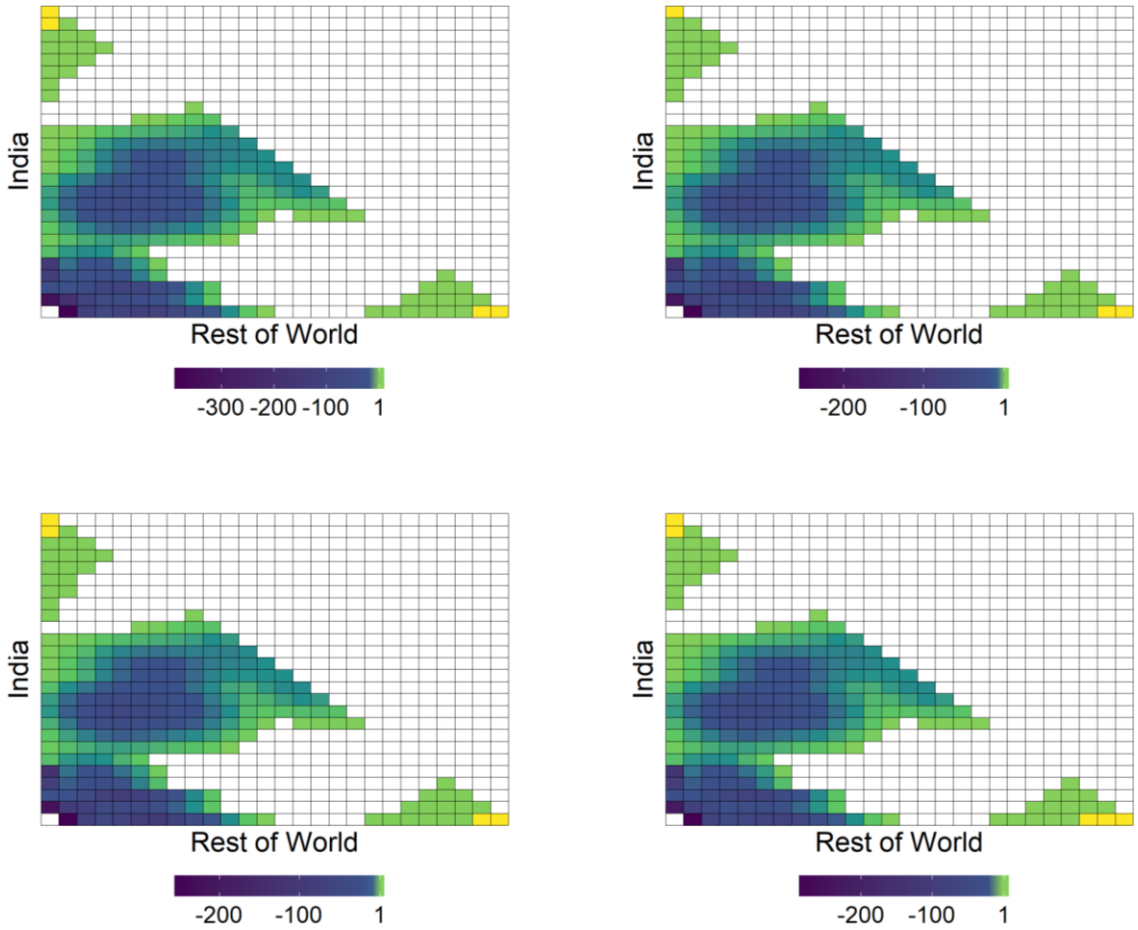

Figure 4.2. Poisson residuals for best-fit migration models. Heatmaps of Poisson residuals between data and model for the spectra calculated using ∂a∂l. Masked data and SNP counts = 0 colored in white. Positive or negative residuals indicate the model predicts too many or too few SNPs at a given frequency, respectively. From left to right, top to bottom: No migration, symmetric migration, unidirectional migration, and asymmetric migration.

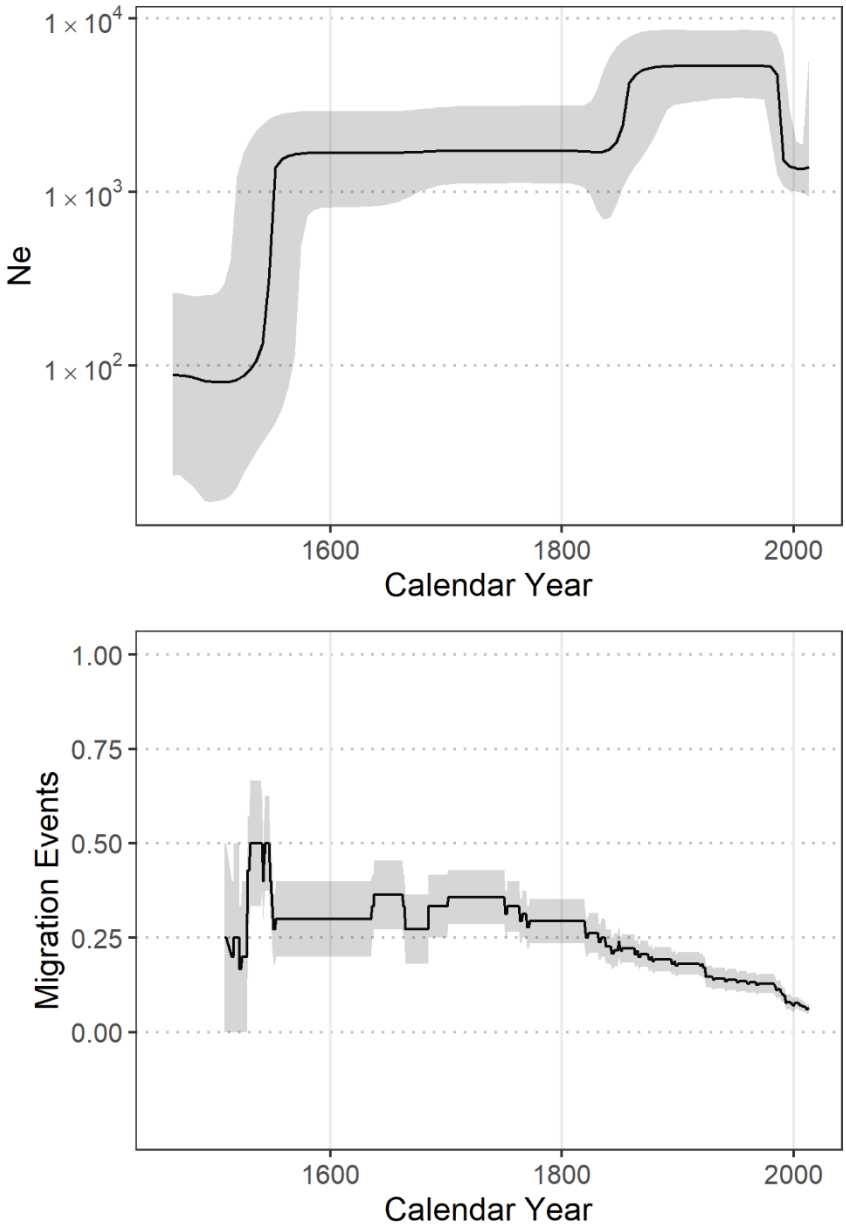

Figure 4.3. Demography of *M. tb* L4.4.1.1. Top: Effective population size (N_e) through time of L4.4.1.1. Median N_e and 95% highest posterior density pictured as black line and grey shading, respectively. X-axis in calendar years. Bottom: Migration events through time of L4.4.1.1. Black line depicts the rate of migration through time, calculated as the sum of migration events occurring across every year of the phylogeny divided by the total number of branches during each year of the phylogeny. Grey shading depicts the rates inferred after the addition or subtraction of a single migration event. X-axis in calendar years.

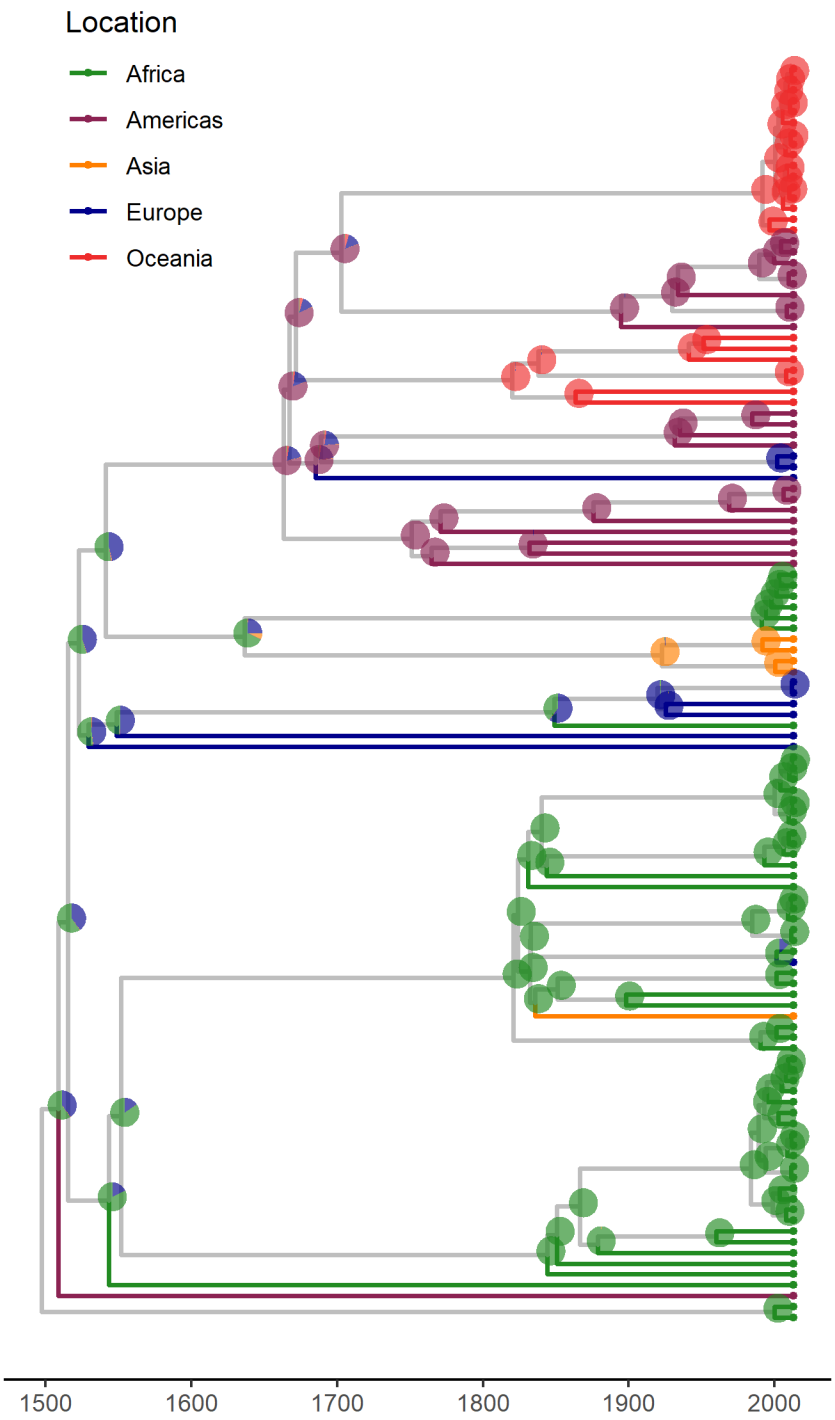

Figure 4.4. MCC phylogeny of *M. tb* L4.4.1.1. Tips and terminal branches colored according to UN region of isolation. Pie charts on nodes colored according to geographic state probabilities. Only nodes with > 0.8 posterior probability shown. X-axis in calendar years.

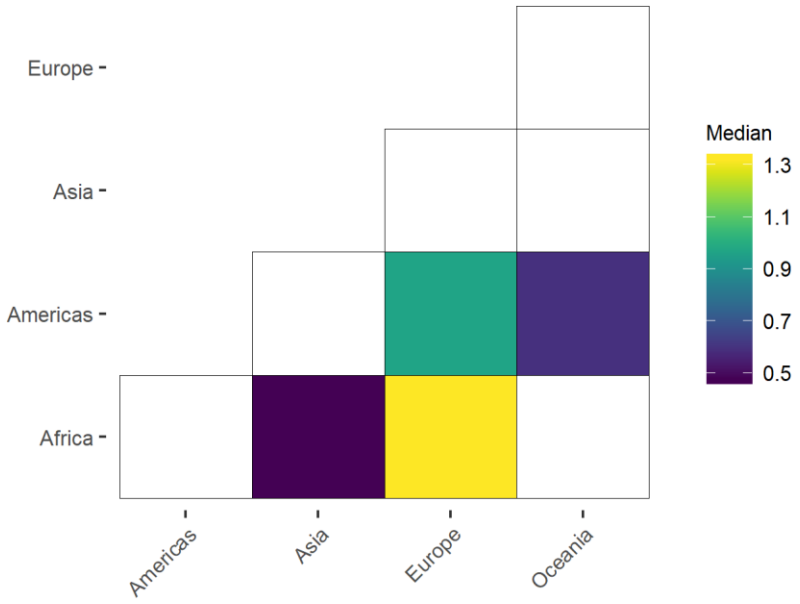

Figure 4.5. Migration matrix of *M. tb* L4.4.1.1. Heatmap of pairwise relative migration rates between UN regions. Only relative rates with Bayes factor > 5 shown.

Tables

Table 4.1. Summary of ∂a∂i migration analyses for *M. tb* L1

Model	ν_{India}	ν_{RoW}	$m_{IndiaToRoW}$	$m_{RoWToIndia}$	Generations	LL	AIC
Neutral	-	-	-	-	-	-2244	-
No Migration	101	30	0.00	0.00	0.65	-507	1020
Symmetric Migration	101	30	0.33	0.33	1.93	-477	1017
Unidirectional Migration	101	30	0.72	0.00	1.68	-503	963
Asymmetric Migration	101	30	0.01	0.65	1.56	-465	940

Values for migration model parameters N_e/N_{anc} (ν), migration (m) and generations since population split (τ), and log likelihood values (LL) and AIC of each migration model

Supplementary Tables

Table 4S.1. *M. tb* L1 accessions and metadata

BioSample	Runs	Assembly Method	Country	UN
ERS019896	ERR245843	reference guided alignment	Malawi	Eastern-Africa
ERS019934	ERR245677	reference guided alignment	Malawi	Eastern-Africa
ERS023444	ERR040121	reference guided alignment	Uganda	Eastern-Africa
ERS1028701	ERR1200626	reference guided alignment	Ethiopia	Eastern-Africa
ERS1028704	ERR1200629	reference guided alignment	Ethiopia	Eastern-Africa
ERS1028707	ERR1200632	reference guided alignment	Ethiopia	Eastern-Africa
ERS107959	ERR163963	reference guided alignment	Malawi	Eastern-Africa
ERS141551	ERR176483	reference guided alignment	Malawi	Eastern-Africa
ERS142071	ERR181936	reference guided alignment	Malawi	Eastern-Africa
ERS142242	ERR181723	reference guided alignment	Malawi	Eastern-Africa
ERS142323	ERR181798	reference guided alignment	Malawi	Eastern-Africa
ERS142332	ERR181807	reference guided alignment	Malawi	Eastern-Africa
ERS142428	ERR190365	reference guided alignment	Malawi	Eastern-Africa
ERS142442	ERR190377	reference guided alignment	Malawi	Eastern-Africa
ERS142454	ERR190389	reference guided alignment	Malawi	Eastern-Africa

ERS153838	ERR211991	reference guided alignment	Malawi	Eastern-Africa
ERS153982	ERR212131	reference guided alignment	Malawi	Eastern-Africa
ERS217639	ERR233351	reference guided alignment	Ethiopia	Eastern-Africa
ERS217641	ERR233353	reference guided alignment	Ethiopia	Eastern-Africa
ERS218215	ERR234164	reference guided alignment	Uganda	Eastern-Africa
ERS218216	ERR234165	reference guided alignment	Tanzania	Eastern-Africa
ERS218323	ERR234272	reference guided alignment	Somalia	Eastern-Africa
ERS217644	ERR233356	reference guided alignment	China	Eastern-Asia
ERS217652	ERR233364	reference guided alignment	China	Eastern-Asia
ERS218313	ERR234262	reference guided alignment	China	Eastern-Asia
SRS490604	SRR1019186, SRR1011511	reference guided alignment	Taiwan	Eastern-Asia
SRS490607	SRR1019190, SRR1011516	reference guided alignment	Taiwan	Eastern-Asia
AOMG02	AOMG00000000.2	multiple genome alignment	Malaysia	South-Eastern-Asia
ATNF01	ATNF00000000.1	multiple genome alignment	Thailand	South-Eastern-Asia
ERS217633	ERR238746	reference guided alignment	Vietnam	South-Eastern-Asia
ERS217651	ERR233363	reference guided alignment	Cambodia	South-Eastern-Asia
ERS217653	ERR233365	reference guided alignment	Vietnam	South-Eastern-Asia
ERS217660	ERR233372	reference guided alignment	Malaysia	South-Eastern-Asia
ERS218206	ERR234155	reference guided alignment	Thailand	South-Eastern-Asia
ERS218207	ERR234156	reference guided alignment	Thailand	South-Eastern-Asia
ERS218208	ERR234157	reference guided alignment	Vietnam	South-Eastern-Asia
ERS218236	ERR234185	reference guided alignment	Singapore	South-Eastern-Asia
ERS218258	ERR234207	reference guided alignment	Vietnam	South-Eastern-Asia
ERS218265	ERR234214	reference guided alignment	Thailand	South-Eastern-Asia
ERS218275	ERR234224	reference guided alignment	The Philippines	South-Eastern-Asia

ERS218289	ERR234238	reference guided alignment	Vietnam	South-Eastern-Asia
ERS218291	ERR234240	reference guided alignment	Vietnam	South-Eastern-Asia
ERS218292	ERR234241	reference guided alignment	Vietnam	South-Eastern-Asia
ERS218315	ERR234264	reference guided alignment	The Philippines	South-Eastern-Asia
ERS218317	ERR234266	reference guided alignment	The Philippines	South-Eastern-Asia
ALYG01	ALYG00000000.1	multiple genome alignment	India	Southern-Asia
ERS217662	ERR233374	reference guided alignment	Sri Lanka	Southern-Asia
ERS217664	ERR233376	reference guided alignment	Nepal	Southern-Asia
ERS217665	ERR233377	reference guided alignment	Nepal	Southern-Asia
ERS217668	ERR233380	reference guided alignment	Nepal	Southern-Asia
ERS218220	ERR234169	reference guided alignment	India	Southern-Asia
ERS218242	ERR234191	reference guided alignment	Sri Lanka	Southern-Asia
ERS218248	ERR234197	reference guided alignment	Afghanistan	Southern-Asia
ERS218286	ERR234235	reference guided alignment	India	Southern-Asia
ERS218288	ERR234237	reference guided alignment	India	Southern-Asia
ERS611795	ERR688020	reference guided alignment	Pakistan	Southern-Asia
ERS611799	ERR688024	reference guided alignment	Pakistan	Southern-Asia
ERS611813	ERR688038	reference guided alignment	Pakistan	Southern-Asia
JMEK01	JMEK00000000.1	multiple genome alignment	India	Southern-Asia
JMIM01	JMIM00000000.1	multiple genome alignment	India	Southern-Asia
JMJH01	JMJH00000000.1	multiple genome alignment	India	Southern-Asia
JNVI02	JNVI00000000.2	multiple genome alignment	India	Southern-Asia
JQGH01	JQGH00000000.1	multiple genome alignment	India	Southern-Asia
SRS557806	SRR1169488, SRR1183033, SRR1180217, SRR1172309	reference guided alignment	India	Southern-Asia

SRS557855	SRR1172226, SRR1169563, SRR1180160, SRR1182980	reference guided alignment	India	Southern-Asia
SRS557872	SRR1180224, SRR1172022, SRR1183159, SRR1169586	reference guided alignment	India	Southern-Asia
SRS559605	SRR1173036, SRR1172709	reference guided alignment	India	Southern-Asia
SRS559606	SRR1172710	reference guided alignment	India	Southern-Asia
SRS559664	SRR1175088, SRR1172787	reference guided alignment	India	Southern-Asia
SRS559665	SRR1172788, SRR1175027	reference guided alignment	India	Southern-Asia
SRS559695	SRR1172900, SRR1172828	reference guided alignment	India	Southern-Asia
SRS559720	SRR1172860, SRR1172872	reference guided alignment	India	Southern-Asia
SRS559736	SRR1175036, SRR1172881	reference guided alignment	India	Southern-Asia
SRS559747	SRR1172905, SRR1175116	reference guided alignment	India	Southern-Asia
SRS559754	SRR1172915, SRR1175053	reference guided alignment	India	Southern-Asia
SRS559766	SRR1175155, SRR1172942	reference guided alignment	India	Southern-Asia
SRS559772	SRR1172950, SRR1175123	reference guided alignment	India	Southern-Asia
SRS559777	SRR1175071, SRR1172955	reference guided alignment	India	Southern-Asia
SRS559783	SRR1172965, SRR1173489	reference guided alignment	India	Southern-Asia
SRS559811	SRR1173198, SRR1173010	reference guided alignment	India	Southern-Asia
SRS559817	SRR1173136, SRR1173031	reference guided alignment	India	Southern-Asia
SRS559825	SRR1173043, SRR1173178	reference guided alignment	India	Southern-Asia
SRS559850	SRR1173611, SRR1173098	reference guided alignment	India	Southern-Asia
SRS559867	SRR1173129, SRR1173673	reference guided alignment	India	Southern-Asia
SRS560015	SRR1173787, SRR1173551	reference guided alignment	India	Southern-Asia
ERS218233	ERR234182	reference guided alignment	Serbia	Southern-Europe
ERS218245	ERR234194	reference guided alignment	Burkina Faso	Western-Africa

ERS218256	ERR234205	reference guided alignment	Ghana	Western-Africa
SRS485067	SRR998721, SRR998720, SRR998722	reference guided alignment	Mali	Western-Africa

Table 4.S2. *M. tb* L4.4.1.1 accessions and metadata

Name	Run ID	Project ID	Year	Country	UN Region
ERR024359	ERR024359	PRJEB2057	1999	Netherlands	Europe
ERR040124	ERR040124	PRJEB2424	2005	Uganda	Africa
ERR040130	ERR040130	PRJEB2424	2004	Uganda	Africa
ERR046901	ERR046901	PRJEB2221	2007	United Kingdom	Europe
ERR047013	ERR047013	PRJEB2221	2006	United Kingdom	Europe
ERR108481	ERR108481	PRJEB2138	2009	Russia	Europe
ERR133974	ERR133974	PRJEB2138	2009	Russia	Europe
ERR161047	ERR161047	PRJEB2794	2003	Malawi	Africa
ERR163930	ERR163930	PRJEB2794	2004	Malawi	Africa
ERR163942	ERR163942	PRJEB2794	2004	Malawi	Africa
ERR163946	ERR163946	PRJEB2794	2004	Malawi	Africa
ERR164012	ERR164012	PRJEB2794	1997	Malawi	Africa
ERR176454	ERR176454	PRJEB2794	2004	Malawi	Africa
ERR176628	ERR176628	PRJEB2794	1997	Malawi	Africa
ERR176725	ERR176725	PRJEB2794	1998	Malawi	Africa
ERR181836	ERR181836	PRJEB2794	2009	Malawi	Africa
ERR181853	ERR181853	PRJEB2794	2009	Malawi	Africa
ERR181946	ERR181946	PRJEB2794	2005	Malawi	Africa
ERR182011	ERR182011	PRJEB2794	2006	Malawi	Africa
ERR190410	ERR190410	PRJEB2794	1997	Malawi	Africa
ERR211992	ERR211992	PRJEB2794	2002	Malawi	Africa
ERR212117	ERR212117	PRJEB2794	1999	Malawi	Africa
ERR212132	ERR212132	PRJEB2794	2009	Malawi	Africa
ERR212152	ERR212152	PRJEB2794	2009	Malawi	Africa
ERR212159	ERR212159	PRJEB2794	2009	Malawi	Africa
ERR221551	ERR221551	PRJEB2794	2002	Malawi	Africa
ERR221554	ERR221554	PRJEB2794	2002	Malawi	Africa
ERR221600	ERR221600	PRJEB2794	2002	Malawi	Africa
ERR228025	ERR228025	PRJEB2138	2010	Russia	Europe
ERR228067	ERR228067	PRJEB2138	2010	Russia	Europe
ERR245800	ERR245800	PRJEB2358	2001	Malawi	Africa
NZO1	SRR5074294	PRJNA356104	2008	New Zealand	Oceania
NZO2	SRR5074712	PRJNA356104	2011	New Zealand	Oceania
NZO3	SRR5074713	PRJNA356104	2008	New Zealand	Oceania
NZO4	NA	NA	2013	New Zealand	Oceania
NZO5	NA	NA	2003	New Zealand	Oceania

NZO6	NA	NA	2013	New Zealand	Oceania
NZO7	NA	NA	2006	New Zealand	Oceania
NZR22	SRR8420474	PRJNA513885	2010	New Zealand	Oceania
NZR278	SRR8420475	PRJNA513885	2010	New Zealand	Oceania
NZR486	SRR8420472	PRJNA513885	2011	New Zealand	Oceania
NZR494	SRR8420473	PRJNA513885	2011	New Zealand	Oceania
NZRA	NA	NA	1992	New Zealand	Oceania
NZRB	NA	NA	1992	New Zealand	Oceania
NZRC	NA	NA	1999	New Zealand	Oceania
NZRE	NA	NA	2001	New Zealand	Oceania
NZRF	NA	NA	2002	New Zealand	Oceania
NZRH	NA	NA	2006	New Zealand	Oceania
NZRI	NA	NA	2006	New Zealand	Oceania
NZRJ	NA	NA	2008	New Zealand	Oceania
NZRK	NA	NA	2008	New Zealand	Oceania
NZRL	NA	NA	1996	New Zealand	Oceania
NZRM	NA	NA	2009	New Zealand	Oceania
NZRN	NA	NA	1991	New Zealand	Oceania
SK1	NA	NA	1994	Canada	Americas
SK111	NA	NA	2003	Canada	Americas
SK153	NA	NA	1996	Canada	Americas
SK160	NA	NA	1999	Canada	Americas
SK192	NA	NA	1990	Canada	Americas
SK196	NA	NA	1987	Canada	Americas
SK198	NA	NA	1987	Canada	Americas
SK199	NA	NA	1989	Canada	Americas
SK20	NA	NA	1988	Canada	Americas
SK201	NA	NA	2001	Canada	Americas
SK202	NA	NA	1994	Canada	Americas
SK204	NA	NA	2002	Canada	Americas
SK211	NA	NA	1997	Canada	Americas
SK215	NA	NA	1999	Canada	Americas
SK254	NA	NA	2004	Canada	Americas
SK256	NA	NA	1987	Canada	Americas
SK260	NA	NA	1996	Canada	Americas
SK287	NA	NA	2002	Canada	Americas
SK288	NA	NA	2003	Canada	Americas
SK30	NA	NA	2003	Canada	Americas
SK54	NA	NA	1990	Canada	Americas
SRR1019141	SRR1019141	PRJNA183624	2013	Kenya	Africa
SRR1019142	SRR1019142	PRJNA183624	2013	South Africa	Africa
SRR1019150	SRR1019150	PRJNA183624	2013	South Africa	Africa
SRR1019153	SRR1019153	PRJNA183624	2013	Kenya	Africa

SRR1019155	SRR1019155	PRJNA183624	2013	South Africa	Africa
SRR1019159	SRR1019159	PRJNA183624	2013	South Africa	Africa
SRR1019165	SRR1019165	PRJNA183624	2013	South Africa	Africa
SRR1019168	SRR1019168	PRJNA183624	2013	South Africa	Africa
SRR1140926	SRR1140926	PRJNA183624	2013	South Africa	Africa
SRR1162884	SRR1162884	PRJNA229360	2010	Sweden	Europe
SRR1163077	SRR1163077	PRJNA229360	2010	Sweden	Europe
SRR1166253	SRR1166253	PRJNA233386	2013	Romania	Europe
SRR1172724	SRR1172724	PRJNA235618	2012	South Africa	Africa
SRR1172876	SRR1172876	PRJNA235618	2013	South Africa	Africa
SRR1172935	SRR1172935	PRJNA235618	2013	South Africa	Africa
SRR1173087	SRR1173087	PRJNA235618	2013	South Africa	Africa
SRR1173181	SRR1173181	PRJNA235618	2013	South Africa	Africa
SRR1173353	SRR1173353	PRJNA235618	2013	South Africa	Africa
SRR1173499	SRR1173499	PRJNA235618	2013	South Africa	Africa
SRR1173522	SRR1173522	PRJNA235618	2013	South Africa	Africa
SRR1173637	SRR1173637	PRJNA235852	2003	India	Asia
SRR1175041	SRR1175041	PRJNA235618	2013	South Africa	Africa
SRR1180189	SRR1180189	PRJNA235618	2013	South Africa	Africa
SRR1180314	SRR1180314	PRJNA235618	2013	South Africa	Africa
SRR1181100	SRR1181100	PRJNA191021	2009	Colombia	Americas
SRR1181216	SRR1181216	PRJNA235618	2013	South Africa	Africa
SRR1184309	SRR1184309	PRJNA235618	2013	South Africa	Africa
SRR5065416	SRR5065416	PRJNA355614	2011	Vietnam	Asia
SRR5067392	SRR5067392	PRJNA355614	2009	Vietnam	Asia
SRR5073887	SRR5073887	PRJNA355614	2009	Vietnam	Asia
SRR5073966	SRR5073966	PRJNA355614	2009	Vietnam	Asia
SRR832977	SRR832977	PRJNA183624	2008	South Africa	Africa
SRR832988	SRR832988	PRJNA183624	2009	South Africa	Africa
SRR833034	SRR833034	PRJNA183624	2008	South Africa	Africa
SRR833044	SRR833044	PRJNA183624	2010	South Africa	Africa
SRR833119	SRR833119	PRJNA183624	2008	South Africa	Africa
SRR833134	SRR833134	PRJNA183624	2010	South Africa	Africa
SRR833144	SRR833144	PRJNA183624	2008	South Africa	Africa
SRR833165	SRR833165	PRJNA183624	2008	South Africa	Africa
SRR847795	SRR847795	PRJNA183624	2011	South Africa	Africa
SRR847797	SRR847797	PRJNA183624	2011	South Africa	Africa
SRR847802	SRR847802	PRJNA183624	2011	South Africa	Africa

Chapter 5: Conclusions and Future Directions

In this thesis I have examined the evolution of bacterial pathogens across three scales using WGS data from natural populations. In Chapter 2, I characterized the evolution of the gonococcal genetic island (GGI), examined its effects on the landscape of LGT in *Neisseria gonorrhoeae* and investigated co-adaptation between the GGI and *N. gonorrhoeae's* core genome. In Chapter 3, I compared patterns of genetic diversity of pathogenic mycobacteria in sputum to bacteria grown *in vitro* and elucidated differences between evolutionary pressures encountered within the host and those imposed by *ex vivo* manipulation of bacterial populations. In Chapter 4, I characterized the migratory histories of two lineages of *Mycobacterium tuberculosis* (*M. tb*) using different methods of demographic inference. This work provides new insights into the evolution of bacterial pathogens and demonstrates the power of WGS data and associated analytical methods to characterize the ecology and evolution of infectious disease.

The gonococcal genetic island and identifying intragenomic interactions

The aim of Chapter 2 was to characterize the evolutionary history of the GGI, which encodes a T4SS, and examine its effects on the landscape of LGT in natural populations of *N. gonorrhoeae*. I found evidence of more core genome recombination in GGI+ populations than in GGI- populations. My results are the first computational prediction of increased core genome recombination in natural populations of GGI+ *N. gonorrhoeae.* These results are consistent with co-culture experiments of laboratory strains of *N. gonorrhoeae* that demonstrated GGI+ donor cells produce up to 500X more transformants than donors deficient in secreting DNA (1). Additionally, I identified recombination hotspots and genes with disparate recombination intensities in the core genomes of GGI+ and GGI- populations. These results further demonstrate the impact the GGI and the T4SS have on the landscape of LGT in *N. gonorrhoeae* and are a valuable resource for researchers investigating the differential selective pressures that shape GGI+ and GGI- populations.

Through my work, I found evidence of associations between the GGI and variants in the core and accessory genomes of *N. gonorrhoeae,* and multiple facets of the GGI's diversity indicate it acts as a mobile element. Additionally, I found that presence/absence of the GGI, as well as specific groups of the GGI were highly structured on *N. gonorrhoeae's* core genome phylogeny. My results suggest the GGI and *N. gonorrhoeae*'s core genome have co-adapted, and we hypothesized that interactions between the GGI and *N. gonorrhoeae's* core genome impose a barrier to transfer of the GGI to a GGI- genetic background. A non-replicating plasmid carrying part of the GGI can be integrated into isogenic GGI mutants (1), but complete transfer of the GGI to isogenic GGI mutants has not been observed. This could be reflective of intragenomic conflict. Co-culture experiments between GGI+ clinical isolates, as well as GGI+ and GGI- clinical isolates, could potentially address this barrier to transfer observed in experimental settings and my analyses of WGS data from natural populations. Additionally, the variants associated with GGI presence/absence I identified are candidates for *in vitro* experiments to determine what mutations and/or gene content allow for the maintenance of the GGI in GGI+ populations and prevent its transfer into GGI- populations.

Co-adaptation between MGEs and bacterial chromosomes has been observed during experimental evolution, and the mutations associated with co-adaptation in these experiments have been identified using WGS data sequenced from the ancestral and evolved strains. I developed an analytical framework for identifying candidates of co-adaptation between MGEs and bacterial core genomes using WGS data from natural populations (2–6). This framework is relevant to researchers studying the evolution of bacterial MGEs in natural populations, independent of species. Additionally, methods are frequently being developed to analyze bacterial pan-genomes, and new techniques could be readily incorporated into this framework (7–9). For example, the recently developed software Piggy (7) specifically characterizes intergenic regions in the core genome of bacteria, which were not included in this analysis but could provide further insight into the evolution of MGEs in *N. gonorrhoeae* and other species.

Analysis of the GGI in other *Neisseria spp.* may give us further insight into the evolutionary history of the GGI in *N. gonorrhoeae*. *N. meningitidis* contains variants of the GGI, albeit at a much lower frequency (approximately 17% of isolates) and the GGI has been identified in one isolate of *Neisseria bacilliformis,* a commensal *Neisseria*; (10–13). The GGI of some *N. meningitidis* strains have accumulated indels in genes essential for secretion in *N. gonorrhoeae*, potentially nullifying their function. As for *N. meningitidis* strains with a functional T4SS, these strains do not appear to secrete DNA and no association has been observed between the presence of the T4SS of *N. meningitidis* and the infection process.

Currently, there are ~1900 *N. meningitidis* genomes deposited on the National Center for Biotechnology Information's genome database (14). Using the computational analyses of *N. gonorrhoeae's* GGI I developed to characterize the diversity of the GGI in natural populations, we may determine if the GGI remains mobile in these populations. Additionally, these analyses can be used to identify candidate variants that provide insight into the maintenance of the GGI in *N. meningitidis* populations, as well as mutual targets of co-adaptation in the core genomes of these pathogenic *Neisseria*.

Within-host evolution of pathogenic Mycobacteria

The aim of Chapter 3 was to examine the effects of host, sample and *in vitro* culture on the genetic diversity of pathogenic Mycobacteria. Our approach integrated molecular techniques and population genomics by analyzing sequences from *M. tb* and *M. bovis* populations captured directly from sputum and from cultured samples. We found that genome-wide diversity of *M. tb* in sputum samples varies substantially within and among hosts, and that diversity of these populations is higher than it is for the same strains grown *in vitro*. Elucidating the effect of *ex vivo* manipulation on the genetic diversity of bacterial populations using capture based methods is an active area of research in the *M. tb* community (15–17), and our finding of increased *M. tb*

diversity in sputum relative to culture is consistent with results of other studies using similar methods (15, 16).

Our estimates of diversity in sputum vs culture for individual genes were consistent with our genome-wide estimates: when genes changed in diversity from sputum to culture, they were more likely to decrease in diversity. This catalog of changes in diversity for each of *M. tb's* ~4000 genes is a valuable resource for the *M. tb* community; It's important to understand how the diversity of *M. tb's* genes changes from sputum to culture, as the former is informative of host-pathogen interactions and the latter is vital in distinguishing signal from noise in bacterial sequencing data.

Using three outlier analyses, we identified a group of genes characterized by marked changes in diversity between sputum and culture. For example, two of these genes, *rpoB* and *rpoC,* are known to mediate resistance to rifamycins, which are first line antibiotic treatments for TB. We expect selection pressures on drug resistance loci to shift between sputum and culture in antibiotic free media, and thus the identification of *rpoB* and *rpoC* provides support for the use of our outlier method to identify genes under differential selection pressures *in vivo* and *in vitro*.

While we did identify outliers in our three samples from a single *M. bovis* patient, we did not discuss these genes and their role in within-host adaptation, because we could not cross validate our findings. Analyses of *in vivo* versus *in vitro* adaptation of *M. tb* and/or *M. bovis* with a greater number of longitudinal samples from a larger group of patients would allow for cross validation of these results and further characterization of within-host genetic diversity during infection.

Our results show clear effects of *in vitro* culture on the genetic diversity of bacterial populations and demonstrate the utility of using capture-based techniques to elucidate differences in the evolutionary pressures encountered *in vivo* versus *in vitro*. The outlier analyses presented here serve as an analytical framework for future studies seeking to examine the effects of *in vitro* culture on the genetic diversity of other bacterial pathogens and are

especially valuable for researchers studying other obligate human pathogens. For example, like *M. tb*, *N. gonorrhoeae* is an obligate human pathogen that can be grown axenically in the lab. Few studies have examined within-host evolution of *N. gonorrhoeae*. Capture of *N. gonorrhoeae* genomic material directly from infected host tissues and characterization of its *in vivo* genetic diversity using the analyses described here could offer insight into within-host adaptation of this species.

Migratory history of *Mycobacterium tuberculosis*

The aim of Chapter 4 was to characterize the migratory history of two lineages of *M. tb* using different methods of migration inference. Reconstructing the migratory history of *M. tb* L1, we found that site frequency spectrum (SFS) based methods of inference perform poorly when two population migration models are used. These results are consistent with prior research indicating that SFS based methods perform well for fully linked genomes when inference is done under very simple models but not when using more complex models (18).

Despite research suggesting inference of bacterial demography using the SFS is less sensitive to selection than Bayesian methods (19), few studies have used SFS-based methods to infer the migratory history of bacterial pathogens. Unlike *M. tb, H. pylori* is highly recombinogenic, and its SFS is unlikely to be skewed by the effects of linkage. As described in Chapter 1, the migratory history of *H. pylori* is well characterized, and WGS data from numerous phylogeographic studies of *H. pylori* are publicly available (20–23). Publicly available data from these studies of *H. pylori* populations could be used to further explore the utility of SFS based methods for reconstructing the migratory history of bacterial pathogens.

Our Bayesian inference of *M. tb* L4.4.1.1's migratory history is indicative of multiple migrations of *M. tb* from Europe to Indigenous populations in Canada and New Zealand. We linked these patterns of migration to historical phenomena, specifically colonial migrations and trade expansion that occurred during the time periods *M. tb* was introduced to these populations

according to our reconstructions. These results are consistent with previous analyses of L4's migratory history (24), which notably did not include isolates from the South Pacific, indicating L4's current distribution was influenced by colonial migrations. Our results demonstrate the power of using a pathogen's genomic data in concert with spatial and temporal data to trace its origins and contribute to the growing body of work showing a clear effect of human migration on the current global distribution of *M. tb*.

In our sample of *M. tb* L4.4.1.1, we observed a cluster of *M. tb* circulating in Indigenous populations of New Zealand that is much younger than the cluster first introduced from Europe. This cluster could be an expansion from a previously introduced, unsampled DS6Q strain or a second, more recent introduction of *M. tb*. More samples of *M. tb* L4.4.1.1 from both New Zealand and the South Pacific may elucidate the route this strain took from Europe to New Zealand.

Finally, bacterial phylogeography is an active area of research and as we continue to sequence more whole genomes, new and increasingly sophisticated models to reconstruct the transmission of bacterial species are being developed. Phylogeographic analysis of *M. tb* using these methods will be key in elucidating the historical dispersal of this globally distributed pathogen and critical for informing strategies to control the current pandemic.

References

1. Hamilton HL, Domínguez NM, Schwartz KJ, Hackett KT, Dillard JP. 2005. Neisseria gonorrhoeae secretes chromosomal DNA via a novel type IV secretion system. Molecular Microbiology 55:1704–1721.
2. Dahlberg C, Chao L. 2003. Amelioration of the cost of conjugative plasmid carriage in Eschericha coli K12. Genetics 165:1641–1649.
3. De Gelder L, Williams JJ, Ponciano JM, Sota M, Top EM. 2008. Adaptive plasmid evolution results in host-range expansion of a broad-host-range plasmid. Genetics 178:2179–2190.

4. Stalder T, Rogers LM, Renfrow C, Yano H, Smith Z, Top EM. 2017. Emerging patterns of plasmid-host coevolution that stabilize antibiotic resistance. Sci Rep 7:4853–4853.
5. Bottery MJ, Wood AJ, Brockhurst MA. 2017. Adaptive modulation of antibiotic resistance through intragenomic coevolution. Nat Ecol Evol 1:1364–1369.
6. Dionisio F, Zilhão R, Gama JA. 2019. Interactions between plasmids and other mobile genetic elements affect their transmission and persistence. Plasmid 102:29–36.
7. Thorpe HA, Bayliss SC, Sheppard SK, Feil EJ. 2018. Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. GigaScience 7.
8. Ding W, Baumdicker F, Neher RA. 2018. panX: pan-genome analysis and exploration. Nucleic Acids Res 46:e5–e5.
9. Abudahab K, Prada JM, Yang Z, Bentley SD, Croucher NJ, Corander J, Aanensen DM. 2018. PANINI: Pangenome Neighbour Identification for Bacterial Populations. Microb Genom 5:e000220.
10. Pachulec E, Siewering K, Bender T, Heller E-M, Salgado-Pabon W, Schmoller SK, Woodhams KL, Dillard JP, van der Does C. 2014. Functional Analysis of the Gonococcal Genetic Island of Neisseria gonorrhoeae. PLOS ONE 9:e109613.
11. Snyder LAS, Jarvis SA, Saunders NJ. 2005. Complete and variant forms of the ‘gonococcal genetic island’ in Neisseria meningitidis. Microbiology 151:4005–4013.
12. Woodhams KL, Benet ZL, Blonsky SE, Hackett KT, Dillard JP. 2012. Prevalence and Detailed Mapping of the Gonococcal Genetic Island in Neisseria meningitidis. J Bacteriol 194:2275.
13. 2011. Neisseria bacilliformis ATCC BAA-1200 contig00002, whole genome shotgun sequence.
14. NCBI. Genomes - Neisseria meningitidis - NCBI.
15. Doyle RM, Burgess C, Williams R, Gorton R, Booth H, Brown J, Bryant JM, Chan J, Creer D, Holdstock J, Kunst H, Lozewicz S, Platt G, Romero EY, Speight G, Tiberi S, Abubakar I, Lipman M, McHugh TD, Breuer J. 2018. Direct Whole-Genome Sequencing of Sputum Accurately Identifies Drug-Resistant Mycobacterium tuberculosis Faster than MGIT Culture Sequencing. Journal of clinical microbiology 56:e00666-18.
16. Nimmo C, Shaw LP, Doyle R, Williams R, Brien K, Burgess C, Breuer J, Balloux F, Pym AS. 2018. Whole genome sequencing Mycobacterium tuberculosis directly from sputum identifies more genetic diversity than sequencing from culture. bioRxiv 446849.

17. Votintseva AA, Bradley P, Pankhurst L, Elias C del O, Loose M, Nilgiriwala K, Chatterjee A, Smith EG, Sanderson N, Walker TM, Morgan MR, Wyllie DH, Walker AS, Peto TEA, Crook DW, Iqbal Z. 2017. Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. Journal of Clinical Microbiology 55:1285–1298.
18. Pepperell CS, Casto AM, Kitchen A, Granka JM, Cornejo OE, Holmes EC, Birren B, Galagan J, Feldman MW. 2013. The Role of Selection in Shaping Diversity of Natural M. tuberculosis Populations. PLoS Pathog 9:e1003543.
19. Lapierre M, Blin C, Lambert A, Achaz G, Rocha EPC. 2016. The Impact of Selection, Gene Conversion, and Biased Sampling on the Assessment of Microbial Demography. Mol Biol Evol 33:1711–1725.
20. Linz B, Balloux F, Moodley Y, Manica A, Liu H, Roumagnac P, Falush D, Stamer C, Prugnolle F, van der Merwe SW, Yamaoka Y, Graham DY, Perez-Trallero E, Wadstrom T, Suerbaum S, Achtman M. 2007. An African origin for the intimate association between humans and Helicobacter pylori. Nature 445:915.
21. Moodley Y, Linz B, Bond RP, Nieuwoudt M, Soodyall H, Schlebusch CM, Bernhöft S, Hale J, Suerbaum S, Mugisha L, van der Merwe SW, Achtman M. 2012. Age of the Association between Helicobacter pylori and Man. PLOS Pathogens 8:e1002693.
22. Waskito L, Yamaoka Y. 2019. The Story of Helicobacter pylori: Depicting Human Migrations from the Phylogeography.
23. Didelot X, Nell S, Yang I, Woltemate S, van der Merwe S, Suerbaum S. 2013. Genomic evolution and transmission of Helicobacter pylori in two South African families. Proc Natl Acad Sci U S A 110:13880–13885.
24. Brynildsrud OB, Pepperell CS, Suffys P, Grandjean L, Monteserin J, Debech N, Bohlin J, Alfsnes K, Pettersson JO-H, Kirkeleite I, Fandinho F, da Silva MA, Perdigao J, Portugal I, Viveiros M, Clark T, Caws M, Dunstan S, Thai PVK, Lopez B, Ritacco V, Kitchen A, Brown TS, van Soolingen D, O'Neill MB, Holt KE, Feil EJ, Mathema B, Balloux F, Eldholm V. 2018. Global expansion of Mycobacterium tuberculosis lineage 4 shaped by colonial migration and local adaptation. Sci Adv 4:eaat5869–eaat5869.

Appendix 1

Analyses and results presented in this appendix were performed by Mary O'Neill and Claire Mulholland

Sample Description of the Old World Collection

We assembled or aligned publicly available whole genome sequences (WGS) of thousands of *M. tb* isolates from recently published studies and databases for which country of origin information were known that fell within regions traditionally defined as the Old World. We obtained geographic locations for each of the 552 samples in the Old World collection from NCBI and/or the publications in which the isolates were first described. We assembled isolates using a reference guided approach (RGA) when FASTQ data were available and by multiple genome alignment (MGA) when only draft genome assemblies were accessible (described below). As we were interested in reconstructing historical migrations of the pathogen, we excluded countries where the majority of contemporary TB cases are identified in recent immigrants (1–6). Due to computational limitations, we limited our dataset to <600 isolates. We implemented a sub-sampling strategy for countries with large numbers of available genomes (7), whereby phylogenetic lineage diversity was captured thus minimizing the overrepresentation of clonal complexes (e.g., outbreaks): we performed phylogenetic inference on all isolates available from a country with Fasttree (8) and selected a random isolate from each clade extending from *n* branches, where *n* was the desired number of isolates from the country. The number of isolates per country was selected based on the availability of appropriate genome sequence data as well as relative TB prevalence (9). The final Old World collection consisted of the WGS of previously published *M. tb* isolates (n = 552) collected from 51 countries spanning 13 UN geoscheme subregions (Table 1).

Reference guided assembly of the Old World Collection

We retrieved previously published FASTQ data from the National Center for Biotechnology Information (NCBI) sequence read archive [SRA; (10)]. We trimmed low-quality bases using a threshold quality of 15 and discarded reads resulting in less than 20bp length using Trim Galore! (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), which is a wrapper tool around Cutadapt (11) and FastQC

(http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Using the MEM algorithm (12), we mapped reads to H37Rv [NC_000962.3; (13)]. We removed duplicates using Picard Tools (http://picard.sourceforge.net), and performed local realignment with GATK (14). We discarded sequencing runs for which <80% of the H37Rv genome was covered by at least 20X coverage and runs for which <70% of the reads mapped [determined by Qualimap (15)]. Using Pilon (16), we called variants with the following parameters: --variant --mindepth 10 --minmq 40 --minqual 20.

Multiple genome alignment of the Old World Collection

Draft genome assemblies were aligned to H37Rv [NC_000962.3 (13)] with Mugsy v1.2.3 (17). Regions not present in H37Rv were removed and merged with the reference-guided assembly.

SNP alignment of the Old World Collection

We converted variant calls (VCFs) to FASTAs with in-house scripts (available at https://github.com/ONeillMB1/Mtb_Phylogeography). We masked transposable elements, phage elements, and repetitive families of genes (PE, PPE, and PE-PGRS gene families) that are poorly resolved with short read sequencing to missing data. We excluded isolates with > 20% missing sites from the Old World collection. Using SNP-sites (18), we extracted variant positions with respect to H37Rv. We included sites where at least half of the isolates had confident data (i.e., non-missing) in migration inference (60787 variant sites; 3838249 bp).

Genomic data and sample description of L4.4

L4.4 genomes include 12 published and 4 unpublished *M. tb* genomes from New Zealand (21, 22), 23 unpublished genomes from Canada and 190 publicly available genomes from recently published studies (23–33) and Broad Institute sequencing initiatives (broadinstitute.org). We assembled publicly available genomes from a list of L4 genomes (34). We identified and selected L4.4 using KvarQ (35). We identified additional L4.4 genomes through literature searches and screening with KvarQ. We excluded genomes with low or mixed

coverage, and if more than one genome sequence was available for a sample only the first listed was used. Country and year of isolation were obtained from the NCBI BioSample database.

Reference guided assembly and variant calling of L4.4

We trimmed low-quality bases using a threshold quality of 15 and discarded reads resulting in less than 20bp length using Trim Galore! (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), which is a wrapper tool around Cutadapt (11) and FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Using the BWA-MEM algorithm (12), we mapped reads to H37Rv [NC_000962.3; (13)]. We removed duplicates using Picard Tools (http://picard.sourceforge.net), and performed local realignment with GATK (14). We excluded genomes if the depth of coverage was <25X or if <75% of trimmed reads mapped to the reference genome [determined by Qualimap; (15)]. Using Pilon (16), we called variants with the following parameters: --variant --mindepth 10 --minmq 40 --minqual 20.

We converted VCF files generated by Pilon to FASTA format using in house scripts that treat ambiguous calls and deletions as missing data (https://github.com/pepperell-lab/RGAPepPipe). We masked transposable elements, phage elements, and repetitive families of genes (PE, PPE, and PE-PGRS gene families) that are poorly resolved with short read sequencing to missing data.

SNP alignment of L4.4

Using SNP-sites (18), we extracted variant positions with respect to H37Rv. We excluded genomes with missing data at > 10% of sites and included only sites where at least 90% of isolates had high quality base calls in phylogenetic and molecular dating analyses.

Bayesian phylogenetic analysis of L4.4.1.1

We performed Bayesian evolutionary analysis of the L4.4.1.1 sub-lineage was performed in BEAST2 (36) using 3161 variant sites extracted from a 3949977 bp alignment of

117 L4.4.1.1 genomes with known year of isolation (Table S2 in Chapter 4). We manually modified XML-input files to specify the number of invariant sites calculated by scaling the number of non-SNP sites in the full alignment by the frequency of each base.

Assessment of temporal signal for tip-based calibration for L4.4.1.1

We calibrated the molecular clock using tip dates covering a 26-year period (1987–2013). To determine if the temporal signal was sufficient for accurate molecular dating, we performed root-to-tip regression and date randomization. Using PhyML, we constructed a maximum likelihood tree, and we used Tempest to determine root-to-tip distance for regression analysis against tip date. This revealed a modest temporal signal in the data (R^2 = 0.229). The Canada-New Zealand-Russia clade sample subset (n = 47) showed weaker temporal signal (R^2 = 0.139) but similar slope (4.6×10^{-4}) to the full L4.4.1.1 sample (2.3×10^{-4}). To further validate the temporal signal, we randomized sampling dates (20X) and analyzed the data using BEAST2 with a strict clock, constant demographic model and the same parameters for the random and true dates. Substitution rate and TMRCA estimates showed no overlap in the 95% HPD between the true and randomized dates. This indicates the data contains sufficient temporal signal for tip-based calibration.

Molecular dating of L4.4.1.1

We estimated mutation rates and divergence times using MCMC sampling in BEAST2 with the BEAGLE library (36, 37). We performed analyses with the GTR substitution model, strict and relaxed molecular clocks (uncorrelated relaxed clock with a log-normal distribution (UCLD)), and coalescent constant, exponential and Bayesian skyline demographic models (38, 39). We specified two monophyletic taxon sets to ensure the root was correctly placed (determined with high confidence bootstrap support in the maximum likelihood phylogeny). We used a uniform prior distribution for the substitution rate (1×10^{-10}–1×10^{-6} s/s/y) and effective population size (upper bound = 1×10^{10}). For the Bayesian skyline model, we deselected the Jeffrey's (1/X) prior for the population size parameter; this is an improper prior unsuitable for

model evaluation using path sampling. We used default priors for the remaining parameters. We ran three independent chains for 100–350 million states sampling every 10000 states. We discarded the first 10% of states as burn-in and assessed chains for convergence and sufficient mixing (effective sample size > 200 for all parameters). We combined samples from the three independent chains and based parameter estimation on the combined chain. We inferred the maximum clade credibility (MCC) from the combined tree samples in TreeAnnotator.

We evaluated the performance of various clock and demographic models by path sampling analysis (40). For each model, we specified 100 path steps using the proportions of a $\beta(0.3, 1.0)$ distribution. To check for consistency, we performed two separate runs per model. Additionally, we ran the MCMC in the absence of data to sample prior distributions for each model. Comparison of marginal posterior and prior distributions showed a strong a strong signal from the data indicating our results are just not an artefact reflecting the prior. The effect of the prior on parameter estimation was also examined by using different upper bounds and the default 1/X prior for the effective population size. Congruent rate and date estimates were obtained when the varying prior parameters on population size demonstrating the robustness of our estimates to this prior specification. The GTR + strict clock + BSP had the best model performance.

References

1. Australian Government Department of Health and Ageing. Tuberculosis notifications in Australia, 2008 and 2009. Australian Government Department of Health and Ageing.
2. Government of Canada PHA of C. 2005. TUBERCULOSIS PREVENTION AND CONTROL IN CANADA A FEDERAL FRAMEWORK FOR ACTION.
3. White Z, Painter J, Douglas P, Abubakar I, Njoo H, Archibald C, Halverson J, Robson J, Posey DL. 2017. Immigrant Arrival and Tuberculosis among Large Immigrant- and Refugee-Receiving Countries, 2005–2009. Tuberculosis Research and Treatment. Research article.
4. Centers for Disease Control. CDC - Reported Tuberculosis in the United States, 2015 - TB.

5. Public Health England. Tuberculosis in England: annual report - GOV.UK.
6. Institute of Environmental Science and Research Limited. Tuberculosis in New Zealand: Annual Report 2014.
7. Thorpe HA, Bayliss SC, Hurst LD, Feil EJ. 2017. Comparative Analyses of Selection Operating on Nontranslated Intergenic Regions of Diverse Bacterial Species. Genetics 206:363–376.
8. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE 5:e9490.
9. World Health Organization. 2017. Global Tuberculosis Report 2017. World Health Organization.
10. Leinonen R, Sugawara H, Shumway M. 2011. The Sequence Read Archive. Nucleic Acids Res 39:D19–D21.
11. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10–12.
12. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997 [q-bio].
13. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, III CEBI, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S. 1998. Erratum: Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature; London 396:190.
14. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498.
15. García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, Dopazo J, Meyer TF, Conesa A. 2012. Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics 28:2678–2679.
16. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLOS ONE 9:e112963.
17. Angiuoli SV, Salzberg SL. 2011. Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics 27:334–342.

18. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. biorxiv;038190v1.
19. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly 6:80–92.
20. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLoS Genet 5:e1000695.
21. Gautam SS, Mac Aogáin M, Bower JE, Basu I, O'Toole RF. 2017. Differential carriage of virulence-associated loci in the New Zealand Rangipo outbreak strain of Mycobacterium tuberculosis. Infectious Diseases 49:680–688.
22. Mulholland CV, Ruthe A, Cursons RT, Durrant R, Karalus N, Coley K, Bower J, Permina E, Coleman MJ, Roberts SA, Arcus VL, Cook GM, Aung HL. 2017. Rapid molecular diagnosis of the Mycobacterium tuberculosis Rangipo strain responsible for the largest recurring TB cluster in New Zealand. Diagnostic Microbiology and Infectious Disease 88:138–140.
23. Bjorn-Mortensen K, Soborg B, Koch A, Ladefoged K, Merker M, Lillebaek T, Andersen AB, Niemann S, Kohl TA. 2016. Tracing Mycobacterium tuberculosis transmission by whole genome sequencing in a high incidence setting: a retrospective population-based study in East Greenland. Sci Rep 6:33180–33180.
24. Bryant JM, Grogono DM, Greaves D, Foweraker J, Roddick I, Inns T, Reacher M, Haworth CS, Curran MD, Harris SR, Peacock SJ, Parkhill J, Floto RA. 2013. Whole-genome sequencing to identify transmission of Mycobacterium abscessus between patients with cystic fibrosis: a retrospective cohort study. Lancet 381:1551–1560.
25. Bryant JM, Harris SR, Parkhill J, Dawson R, Diacon AH, van Helden P, Pym A, Mahayiddin AA, Chuchottaworn C, Sanne IM, Louw C, Boeree MJ, Hoelscher M, McHugh TD, Bateson ALC, Hunt RD, Mwaigwisya S, Wright L, Gillespie SH, Bentley SD. 2013. Whole-genome sequencing to establish relapse or re-infection with Mycobacterium tuberculosis: a retrospective observational study. Lancet Respir Med 1:786–792.
26. Brynildsrud OB, Pepperell CS, Suffys P, Grandjean L, Monteserin J, Debech N, Bohlin J, Alfsnes K, Pettersson JO-H, Kirkeleite I, Fandinho F, da Silva MA, Perdigao J, Portugal I, Viveiros M, Clark T, Caws M, Dunstan S, Thai PVK, Lopez B, Ritacco V, Kitchen A, Brown TS, van Soolingen D, O'Neill MB, Holt KE, Feil EJ, Mathema B, Balloux F, Eldholm V. 2018. Global expansion of Mycobacterium tuberculosis lineage 4 shaped by colonial migration and local adaptation. Sci Adv 4:eaat5869–eaat5869.

27. Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, Kontsevaya I, Corander J, Bryant J, Parkhill J, Nejentsev S, Horstmann RD, Brown T, Drobniewski F. 2014. Evolution and transmission of drug-resistant tuberculosis in a Russian population. Nat Genet 46:279–286.
28. Clark TG, Mallard K, Coll F, Preston M, Assefa S, Harris D, Ogwang S, Mumbowa F, Kirenga B, O'Sullivan DM, Okwera A, Eisenach KD, Joloba M, Bentley SD, Ellner JJ, Parkhill J, Jones-López EC, McNerney R. 2013. Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing. PLoS One 8:e83012–e83012.
29. Holt KE, McAdam P, Thai PVK, Thuong NTT, Ha DTM, Lan NN, Lan NH, Nhu NTQ, Hai HT, Ha VTN, Thwaites G, Edwards DJ, Nath AP, Pham K, Ascher DB, Farrar J, Khor CC, Teo YY, Inouye M, Caws M, Dunstan SJ. 2018. Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. Nature Genetics 50:849–856.
30. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW, Parkhill J, Harris D, Walker AS, Bowden R, Monk P, Smith EG, Peto TE. 2013. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis 13:137–146.
31. Zhang H, Li D, Zhao L, Fleming J, Lin N, Wang T, Liu Z, Li C, Galwey N, Deng J, Zhou Y, Zhu Y, Gao Y, Wang T, Wang S, Huang Y, Wang M, Zhong Q, Zhou L, Chen T, Zhou J, Yang R, Zhu G, Hang H, Zhang J, Li F, Wan K, Wang J, Zhang X-E, Bi L. 2013. Genome sequencing of 161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance. Nat Genet advance online publication.
32. Guerra-Assunção JA, Crampin AC, Houben RMGJ, Mzembe T, Mallard K, Coll F, Khan P, Banda L, Chiwaya A, Pereira RPA, McNerney R, Fine PEM, Parkhill J, Clark TG, Glynn JR. 2015. Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. eLife 4:e05166.
33. Guerra-Assunção JA, Houben RMGJ, Crampin AC, Mzembe T, Mallard K, Coll F, Khan P, Banda L, Chiwaya A, Pereira RPA, McNerney R, Harris D, Parkhill J, Clark TG, Glynn JR. 2015. Recurrence due to relapse or reinfection with `1Mycobacterium tuberculosis: a whole-genome sequencing approach in a large, population-based cohort with a high HIV infection prevalence and active follow-up. J Infect Dis 211:1154–1163.
34. Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, Fenner L, Rutaihwa L, Borrell S, Luo T, Gao Q, Kato-Maeda M, Ballif M, Egger M, Macedo R, Mardassi H, Moreno M, Tudo

Vilanova G, Fyfe J, Globan M, Thomas J, Jamieson F, Guthrie JL, Asante-Poku A, Yeboah-Manu D, Wampande E, Ssengooba W, Joloba M, Henry Boom W, Basu I, Bower J, Saraiva M, Vaconcellos SEG, Suffys P, Koch A, Wilkinson R, Gail-Bekker L, Malla B, Ley SD, Beck H-P, de Jong BC, Toit K, Sanchez-Padilla E, Bonnet M, Gil-Brusola A, Frank M, Penlap Beng VN, Eisenach K, Alani I, Wangui Ndung'u P, Revathi G, Gehre F, Akter S, Ntoumi F, Stewart-Isherwood L, Ntinginya NE, Rachow A, Hoelscher M, Cirillo DM, Skenders G, Hoffner S, Bakonyte D, Stakenas P, Diel R, Crudu V, Moldovan O, Al-Hajoj S, Otero L, Barletta F, Jane Carter E, Diero L, Supply P, Comas I, Niemann S, Gagneux S. 2016. Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. Nat Genet 48:1535–1543.

35. Steiner A, Stucki D, Coscolla M, Borrell S, Gagneux S. 2014. KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. BMC Genomics 15:881.
36. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Comput Biol 10:e1003537.
37. Ayres DL, Darling A, Zwickl DJ, Beerli P, Holder MT, Lewis PO, Huelsenbeck JP, Ronquist F, Swofford DL, Cummings MP, Rambaut A, Suchard MA. 2012. BEAGLE: An Application Programming Interface and High-Performance Computing Library for Statistical Phylogenetics. Syst Biol 61:170–173.
38. Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology 7:214.
39. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. Mol Biol Evol 22:1185–1192.
40. Lartillot N, Philippe H. 2006. Computing Bayes Factors Using Thermodynamic Integration. Systematic Biology 55:195–207.

Tables

Table 1. Accessions and metadata for the Old World Collection

BioSample	Runs	Assembly Method	Country	UN	Lineage
ERS019896	ERR245843	reference guided alignment	Malawi	Eastern-Africa	1
ERS019934	ERR245677	reference guided alignment	Malawi	Eastern-Africa	1
ERS023444	ERR040121	reference guided alignment	Uganda	Eastern-Africa	1
ERS1028701	ERR1200626	reference guided alignment	Ethiopia	Eastern-Africa	1
ERS1028704	ERR1200629	reference guided alignment	Ethiopia	Eastern-Africa	1
ERS1028707	ERR1200632	reference guided alignment	Ethiopia	Eastern-Africa	1
ERS107959	ERR163963	reference guided alignment	Malawi	Eastern-Africa	1
ERS141551	ERR176483	reference guided alignment	Malawi	Eastern-Africa	1
ERS142071	ERR181936	reference guided alignment	Malawi	Eastern-Africa	1
ERS142242	ERR181723	reference guided alignment	Malawi	Eastern-Africa	1
ERS142323	ERR181798	reference guided alignment	Malawi	Eastern-Africa	1
ERS142332	ERR181807	reference guided alignment	Malawi	Eastern-Africa	1
ERS142428	ERR190365	reference guided alignment	Malawi	Eastern-Africa	1
ERS142442	ERR190377	reference guided alignment	Malawi	Eastern-Africa	1
ERS142454	ERR190389	reference guided alignment	Malawi	Eastern-Africa	1

ERS153838	ERR211991	reference guided alignment	Malawi	Eastern-Africa	1
ERS153982	ERR212131	reference guided alignment	Malawi	Eastern-Africa	1
ERS217639	ERR233351	reference guided alignment	Ethiopia	Eastern-Africa	1
ERS217641	ERR233353	reference guided alignment	Ethiopia	Eastern-Africa	1
ERS218215	ERR234164	reference guided alignment	Uganda	Eastern-Africa	1
ERS218216	ERR234165	reference guided alignment	Tanzania	Eastern-Africa	1
ERS218323	ERR234272	reference guided alignment	Somalia	Eastern-Africa	1
ERS217644	ERR233356	reference guided alignment	China	Eastern-Asia	1
ERS217652	ERR233364	reference guided alignment	China	Eastern-Asia	1
ERS218313	ERR234262	reference guided alignment	China	Eastern-Asia	1
SRS490604	SRR1019186, SRR1011511	reference guided alignment	Taiwan	Eastern-Asia	1
SRS490607	SRR1019190, SRR1011516	reference guided alignment	Taiwan	Eastern-Asia	1
AOMG02	AOMG00000000.2	multiple genome alignment	Malaysia	South-Eastern-Asia	1
ATNF01	ATNF00000000.1	multiple genome alignment	Thailand	South-Eastern-Asia	1
ERS217633	ERR238746	reference guided alignment	Vietnam	South-Eastern-Asia	1
ERS217651	ERR233363	reference guided alignment	Cambodia	South-Eastern-Asia	1

ERS217653	ERR233365	reference guided alignment	Vietnam	South-Eastern-Asia	1
ERS217660	ERR233372	reference guided alignment	Malaysia	South-Eastern-Asia	1
ERS218206	ERR234155	reference guided alignment	Thailand	South-Eastern-Asia	1
ERS218207	ERR234156	reference guided alignment	Thailand	South-Eastern-Asia	1
ERS218208	ERR234157	reference guided alignment	Vietnam	South-Eastern-Asia	1
ERS218236	ERR234185	reference guided alignment	Singapore	South-Eastern-Asia	1
ERS218258	ERR234207	reference guided alignment	Vietnam	South-Eastern-Asia	1
ERS218265	ERR234214	reference guided alignment	Thailand	South-Eastern-Asia	1
ERS218275	ERR234224	reference guided alignment	The Philippines	South-Eastern-Asia	1
ERS218289	ERR234238	reference guided alignment	Vietnam	South-Eastern-Asia	1
ERS218291	ERR234240	reference guided alignment	Vietnam	South-Eastern-Asia	1
ERS218292	ERR234241	reference guided alignment	Vietnam	South-Eastern-Asia	1
ERS218315	ERR234264	reference guided alignment	The Philippines	South-Eastern-Asia	1
ERS218317	ERR234266	reference guided alignment	The Philippines	South-Eastern-Asia	1
ALYG01	ALYG00000000.1	multiple genome alignment	India	Southern-Asia	1
ERS217662	ERR233374	reference guided alignment	Sri Lanka	Southern-Asia	1

ERS217664	ERR233376	reference guided alignment	Nepal	Southern-Asia	1
ERS217665	ERR233377	reference guided alignment	Nepal	Southern-Asia	1
ERS217668	ERR233380	reference guided alignment	Nepal	Southern-Asia	1
ERS218220	ERR234169	reference guided alignment	India	Southern-Asia	1
ERS218242	ERR234191	reference guided alignment	Sri Lanka	Southern-Asia	1
ERS218248	ERR234197	reference guided alignment	Afghanistan	Southern-Asia	1
ERS218286	ERR234235	reference guided alignment	India	Southern-Asia	1
ERS218288	ERR234237	reference guided alignment	India	Southern-Asia	1
ERS611795	ERR688020	reference guided alignment	Pakistan	Southern-Asia	1
ERS611799	ERR688024	reference guided alignment	Pakistan	Southern-Asia	1
ERS611813	ERR688038	reference guided alignment	Pakistan	Southern-Asia	1
JMEK01	JMEK00000000.1	multiple genome alignment	India	Southern-Asia	1
JMIM01	JMIM00000000.1	multiple genome alignment	India	Southern-Asia	1
JMJH01	JMJH00000000.1	multiple genome alignment	India	Southern-Asia	1
JNVI02	JNVI00000000.2	multiple genome alignment	India	Southern-Asia	1
JQGH01	JQGH00000000.1	multiple genome alignment	India	Southern-Asia	1

SRS557806	SRR1169488, SRR1183033, SRR1180217, SRR1172309	reference guided alignment	India	Southern-Asia	1
SRS557855	SRR1172226, SRR1169563, SRR1180160, SRR1182980	reference guided alignment	India	Southern-Asia	1
SRS557872	SRR1180224, SRR1172022, SRR1183159, SRR1169586	reference guided alignment	India	Southern-Asia	1
SRS559605	SRR1173036, SRR1172709	reference guided alignment	India	Southern-Asia	1
SRS559606	SRR1172710	reference guided alignment	India	Southern-Asia	1
SRS559664	SRR1175088, SRR1172787	reference guided alignment	India	Southern-Asia	1
SRS559665	SRR1172788, SRR1175027	reference guided alignment	India	Southern-Asia	1
SRS559695	SRR1172900, SRR1172828	reference guided alignment	India	Southern-Asia	1
SRS559720	SRR1172860, SRR1172872	reference guided alignment	India	Southern-Asia	1
SRS559736	SRR1175036, SRR1172881	reference guided alignment	India	Southern-Asia	1
SRS559747	SRR1172905, SRR1175116	reference guided alignment	India	Southern-Asia	1
SRS559754	SRR1172915, SRR1175053	reference guided alignment	India	Southern-Asia	1
SRS559766	SRR1175155, SRR1172942	reference guided alignment	India	Southern-Asia	1
SRS559772	SRR1172950, SRR1175123	reference guided alignment	India	Southern-Asia	1
SRS559777	SRR1175071, SRR1172955	reference guided alignment	India	Southern-Asia	1

SRS559783	SRR1172965, SRR1173489	reference guided alignment	India	Southern-Asia	1
SRS559811	SRR1173198, SRR1173010	reference guided alignment	India	Southern-Asia	1
SRS559817	SRR1173136, SRR1173031	reference guided alignment	India	Southern-Asia	1
SRS559825	SRR1173043, SRR1173178	reference guided alignment	India	Southern-Asia	1
SRS559850	SRR1173611, SRR1173098	reference guided alignment	India	Southern-Asia	1
SRS559867	SRR1173129, SRR1173673	reference guided alignment	India	Southern-Asia	1
SRS560015	SRR1173787, SRR1173551	reference guided alignment	India	Southern-Asia	1
ERS218233	ERR234182	reference guided alignment	Serbia	Southern-Europe	1
ERS218245	ERR234194	reference guided alignment	Burkina Faso	Western-Africa	1
ERS218256	ERR234205	reference guided alignment	Ghana	Western-Africa	1
SRS485067	SRR998721, SRR998720, SRR998722	reference guided alignment	Mali	Western-Africa	1
ERS456774	ERR550659, ERR550658	reference guided alignment	Uzbekistan	Central-Asia	2
ERS456825	ERR550724	reference guided alignment	Uzbekistan	Central-Asia	2
ERS456865	ERR550778	reference guided alignment	Uzbekistan	Central-Asia	2
ERS456867	ERR550783, ERR550782	reference guided alignment	Turkmenistan	Central-Asia	2
ERS456944	ERR550887	reference guided alignment	Uzbekistan	Central-Asia	2

ERS457011	ERR550984	reference guided alignment	Turkmenistan	Central-Asia	2
ERS457145	ERR551159	reference guided alignment	Uzbekistan	Central-Asia	2
ERS457153	ERR551167, ERR551168	reference guided alignment	Turkmenistan	Central-Asia	2
ERS457174	ERR551201	reference guided alignment	Turkmenistan	Central-Asia	2
ERS457242	ERR551293	reference guided alignment	Kazakhstan	Central-Asia	2
ERS457251	ERR551305	reference guided alignment	Uzbekistan	Central-Asia	2
ERS457289	ERR551360, ERR551361	reference guided alignment	Uzbekistan	Central-Asia	2
ERS457384	ERR551494	reference guided alignment	Turkmenistan	Central-Asia	2
ERS457427	ERR551556	reference guided alignment	Uzbekistan	Central-Asia	2
ERS457487	ERR551636	reference guided alignment	Turkmenistan	Central-Asia	2
ERS457488	ERR551638	reference guided alignment	Uzbekistan	Central-Asia	2
ERS457584	ERR551772	reference guided alignment	Kazakhstan	Central-Asia	2
ERS457607	ERR551805, ERR551804	reference guided alignment	Uzbekistan	Central-Asia	2
ERS457618	ERR551821, ERR551822	reference guided alignment	Uzbekistan	Central-Asia	2
ERS457664	ERR551879	reference guided alignment	Turkmenistan	Central-Asia	2
ERS457709	ERR551945, ERR551944	reference guided alignment	Uzbekistan	Central-Asia	2

ERS457995	ERR552358	reference guided alignment	Uzbekistan	Central-Asia	2
ERS458078	ERR552479	reference guided alignment	Kazakhstan	Central-Asia	2
ERS458089	ERR552493, ERR552494	reference guided alignment	Turkmenistan	Central-Asia	2
ERS458146	ERR552580	reference guided alignment	Uzbekistan	Central-Asia	2
ERS458286	ERR552760	reference guided alignment	Turkmenistan	Central-Asia	2
ERS458394	ERR552907	reference guided alignment	Uzbekistan	Central-Asia	2
ERS458395	ERR552910, ERR552911	reference guided alignment	Turkmenistan	Central-Asia	2
ERS458418	ERR552940, ERR552939	reference guided alignment	Uzbekistan	Central-Asia	2
ERS458515	ERR553068	reference guided alignment	Uzbekistan	Central-Asia	2
ERS458536	ERR553098	reference guided alignment	Uzbekistan	Central-Asia	2
ERS458568	ERR553139	reference guided alignment	Uzbekistan	Central-Asia	2
ERS458650	ERR553251	reference guided alignment	Uzbekistan	Central-Asia	2
ERS019949	ERR245692	reference guided alignment	Malawi	Eastern-Africa	2
ERS023476	ERR038746	reference guided alignment	Uganda	Eastern-Africa	2
ERS141674	ERR176606	reference guided alignment	Malawi	Eastern-Africa	2
ERS142685	ERR190369	reference guided alignment	Malawi	Eastern-Africa	2

ERS456995	ERR550957	reference guided alignment	Zimbabwe	Eastern-Africa	2
ERS458528	ERR553086	reference guided alignment	Kenya	Eastern-Africa	2
ERS217647	ERR233359	reference guided alignment	South Korea	Eastern-Asia	2
ERS218149	ERR234098	reference guided alignment	China	Eastern-Asia	2
ERS218154	ERR234103	reference guided alignment	China	Eastern-Asia	2
ERS218159	ERR234108	reference guided alignment	China	Eastern-Asia	2
ERS218165	ERR234114	reference guided alignment	China	Eastern-Asia	2
ERS218167	ERR234116	reference guided alignment	China	Eastern-Asia	2
ERS218179	ERR234128	reference guided alignment	China	Eastern-Asia	2
ERS218183	ERR234132	reference guided alignment	China	Eastern-Asia	2
ERS218184	ERR234133	reference guided alignment	China	Eastern-Asia	2
ERS218189	ERR234138	reference guided alignment	China	Eastern-Asia	2
ERS218299	ERR234248	reference guided alignment	China	Eastern-Asia	2
ERS218303	ERR234252	reference guided alignment	China	Eastern-Asia	2
ERS218304	ERR234253	reference guided alignment	China	Eastern-Asia	2
ERS218320	ERR234269	reference guided alignment	China	Eastern-Asia	2

ERS218321	ERR234270	reference guided alignment	China	Eastern-Asia	2
ERS456971	ERR550927	reference guided alignment	China	Eastern-Asia	2
ERS457866	ERR552177	reference guided alignment	South Korea	Eastern-Asia	2
ERS458344	ERR552838	reference guided alignment	China	Eastern-Asia	2
SRS490597	SRR1011503, SRR1019178	reference guided alignment	Taiwan	Eastern-Asia	2
SRS490603	SRR1011510, SRR1019185	reference guided alignment	Taiwan	Eastern-Asia	2
SRS490606	SRR1011514, SRR1019188	reference guided alignment	Taiwan	Eastern-Asia	2
SRS490612	SRR1019194, SRR1011524	reference guided alignment	Taiwan	Eastern-Asia	2
ERS003236	ERR015614	reference guided alignment	Russia	Eastern-Europe	2
ERS003237	ERR015616	reference guided alignment	Russia	Eastern-Europe	2
ERS094104	ERR133815	reference guided alignment	Russia	Eastern-Europe	2
ERS094247	ERR133958	reference guided alignment	Russia	Eastern-Europe	2
ERS181456	ERR228062	reference guided alignment	Russia	Eastern-Europe	2
ERS181519	ERR229970	reference guided alignment	Russia	Eastern-Europe	2
ERS181548	ERR229999	reference guided alignment	Russia	Eastern-Europe	2
ERS457030	ERR551007	reference guided alignment	Poland	Eastern-Europe	2

ERS457094	ERR551090	reference guided alignment	Russia	Eastern-Europe	2
ERS457179	ERR551212	reference guided alignment	Russia	Eastern-Europe	2
ERS457252	ERR551311	reference guided alignment	Russia	Eastern-Europe	2
ERS457716	ERR551956, ERR551957	reference guided alignment	Russia	Eastern-Europe	2
ERS457840	ERR552136	reference guided alignment	Russia	Eastern-Europe	2
ERS457843	ERR552141	reference guided alignment	Russia	Eastern-Europe	2
ERS458044	ERR552429	reference guided alignment	Russia	Eastern-Europe	2
SRS555885	SRR1169151, SRR1166145	reference guided alignment	Moldova	Eastern-Europe	2
SRS555911	SRR1169036, SRR1166187	reference guided alignment	Moldova	Eastern-Europe	2
ERS456985	ERR550947, ERR550946	reference guided alignment	Papua New Guinea	Melanesia	2
ERS457514	ERR551681, ERR551680	reference guided alignment	Kiribati	Micronesia	2
ERS457522	ERR551688	reference guided alignment	Kiribati	Micronesia	2
ERS458055	ERR552444, ERR552445	reference guided alignment	Kiribati	Micronesia	2
ERS458387	ERR552894, ERR552895	reference guided alignment	Kiribati	Micronesia	2
AMXW02	AMXW00000000.2	multiple genome alignment	Malaysia	South-Eastern-Asia	2
ERS217645	ERR233357	reference guided alignment	Vietnam	South-Eastern-Asia	2

ERS217658	ERR233370	reference guided alignment	Vietnam	South-Eastern-Asia	2
ERS218241	ERR234190	reference guided alignment	Vietnam	South-Eastern-Asia	2
ERS218244	ERR234193	reference guided alignment	Vietnam	South-Eastern-Asia	2
ERS218259	ERR234208	reference guided alignment	Vietnam	South-Eastern-Asia	2
ERS218260	ERR234209	reference guided alignment	Vietnam	South-Eastern-Asia	2
ERS218261	ERR234210	reference guided alignment	Vietnam	South-Eastern-Asia	2
ERS218262	ERR234211	reference guided alignment	Vietnam	South-Eastern-Asia	2
ERS218264	ERR234213	reference guided alignment	Indonesia	South-Eastern-Asia	2
ERS218267	ERR234216	reference guided alignment	Laos	South-Eastern-Asia	2
ERS218293	ERR234242	reference guided alignment	Vietnam	South-Eastern-Asia	2
ERS218296	ERR234245	reference guided alignment	Cambodia	South-Eastern-Asia	2
ERS218297	ERR234246	reference guided alignment	Vietnam	South-Eastern-Asia	2
ERS218298	ERR234247	reference guided alignment	Indonesia	South-Eastern-Asia	2
ERS218300	ERR234249	reference guided alignment	Vietnam	South-Eastern-Asia	2
ERS218301	ERR234250	reference guided alignment	Singapore	South-Eastern-Asia	2
ERS218314	ERR234263	reference guided alignment	Cambodia	South-Eastern-Asia	2

ERS456836	ERR550738	reference guided alignment	Vietnam	South-Eastern-Asia	2
ERS456982	ERR550941, ERR550940	reference guided alignment	Thailand	South-Eastern-Asia	2
ERS457090	ERR551086	reference guided alignment	Vietnam	South-Eastern-Asia	2
ERS457142	ERR551156, ERR551155	reference guided alignment	Thailand	South-Eastern-Asia	2
ERS457191	ERR551225	reference guided alignment	Thailand	South-Eastern-Asia	2
ERS457528	ERR551693, ERR551694	reference guided alignment	Vietnam	South-Eastern-Asia	2
ERS457697	ERR551928, ERR551927	reference guided alignment	Indonesia	South-Eastern-Asia	2
ERS457699	ERR551930	reference guided alignment	Thailand	South-Eastern-Asia	2
ERS457876	ERR552190	reference guided alignment	Thailand	South-Eastern-Asia	2
ERS458128	ERR552550, ERR552549	reference guided alignment	Thailand	South-Eastern-Asia	2
ERS458225	ERR552689	reference guided alignment	Thailand	South-Eastern-Asia	2
ERS458282	ERR552755	reference guided alignment	Vietnam	South-Eastern-Asia	2
ERS458337	ERR552830	reference guided alignment	Vietnam	South-Eastern-Asia	2
ERS458396	ERR552912	reference guided alignment	Vietnam	South-Eastern-Asia	2
ERS458667	ERR553274, ERR553275	reference guided alignment	Thailand	South-Eastern-Asia	2
ERS458670	ERR553277	reference guided alignment	Vietnam	South-Eastern-Asia	2

ERS458692	ERR553304	reference guided alignment	Myanmar	South-Eastern-Asia	2
JWIA01	JWIA00000000.1	multiple genome alignment	Thailand	South-Eastern-Asia	2
ERS456837	ERR550739	reference guided alignment	South Africa	Southern-Africa	2
ERS457077	ERR551071	reference guided alignment	Swaziland	Southern-Africa	2
ERS457084	ERR551079	reference guided alignment	Swaziland	Southern-Africa	2
ERS457264	ERR619080	reference guided alignment	Swaziland	Southern-Africa	2
ERS457643	ERR551854, ERR551855	reference guided alignment	Swaziland	Southern-Africa	2
ERS458268	ERR552743	reference guided alignment	Swaziland	Southern-Africa	2
ERS458554	ERR553116	reference guided alignment	Swaziland	Southern-Africa	2
ERS458582	ERR553156	reference guided alignment	Swaziland	Southern-Africa	2
ERS458700	ERR553313	reference guided alignment	Swaziland	Southern-Africa	2
ERS458709	ERR553324	reference guided alignment	Swaziland	Southern-Africa	2
SRS454488	SRR958234, SRR924709	reference guided alignment	South Africa	Southern-Africa	2
SRS494386	SRR1140947, SRR1019148	reference guided alignment	South Africa	Southern-Africa	2
ERS217663	ERR233375	reference guided alignment	Nepal	Southern-Asia	2
ERS217674	ERR233386	reference guided alignment	India	Southern-Asia	2

ERS217676	ERR233388	reference guided alignment	Nepal	Southern-Asia	2
ERS457399	ERR551520	reference guided alignment	Nepal	Southern-Asia	2
ERS457835	ERR552130	reference guided alignment	Afghanistan	Southern-Asia	2
ERS458527	ERR553082, ERR553081	reference guided alignment	Iran	Southern-Asia	2
ERS458544	ERR553107	reference guided alignment	India	Southern-Asia	2
ERS458590	ERR553171	reference guided alignment	Nepal	Southern-Asia	2
JDVY01	JDVY00000000.1	multiple genome alignment	India	Southern-Asia	2
SRS559431	SRR1172034, SRR1172091, SRR1183088	reference guided alignment	Iran	Southern-Asia	2
SRS559457	SRR1172183, SRR1172288, SRR1180141, SRR1183064	reference guided alignment	Iran	Southern-Asia	2
SRS559458	SRR1172188, SRR1172348, SRR1180211, SRR1183029	reference guided alignment	Iran	Southern-Asia	2
SRS559475	SRR1172289, SRR1174894, SRR1180230, SRR1183028	reference guided alignment	Iran	Southern-Asia	2
SRS559476	SRR1172305, SRR1174896, SRR1180185, SRR1183041	reference guided alignment	Iran	Southern-Asia	2
SRS559607	SRR1172711, SRR1173722	reference guided alignment	Iran	Southern-Asia	2
SRS559672	SRR1172798, SRR1175131	reference guided alignment	India	Southern-Asia	2
SRS559869	SRR1173131, SRR1173875	reference guided alignment	Iran	Southern-Asia	2

SRS559873	SRR1173140	reference guided alignment	Iran	Southern-Asia	2
SRS559892	SRR1173192, SRR1173793	reference guided alignment	Iran	Southern-Asia	2
SRS559947	SRR1173347, SRR1173868	reference guided alignment	Iran	Southern-Asia	2
SRS559963	SRR1173393, SRR1174307	reference guided alignment	Iran	Southern-Asia	2
SRS559981	SRR1173446, SRR1173842	reference guided alignment	Iran	Southern-Asia	2
SRS560014	SRR1173552, SRR1174308	reference guided alignment	Iran	Southern-Asia	2
SRS560036	SRR1173628, SRR1173792	reference guided alignment	Iran	Southern-Asia	2
SRS560053	SRR1173675, SRR1173750	reference guided alignment	Iran	Southern-Asia	2
SRS560058	SRR1173684, SRR1174328	reference guided alignment	Iran	Southern-Asia	2
SRS560065	SRR1173695	reference guided alignment	Iran	Southern-Asia	2
ERS248010	ERR275215	reference guided alignment	Portugal	Southern-Europe	2
ERS248011	ERR275216	reference guided alignment	Portugal	Southern-Europe	2
ERS457295	ERR551369, ERR551370	reference guided alignment	Bosnia	Southern-Europe	2
ERS457325	ERR551412	reference guided alignment	Sierra Leone	Western-Africa	2
ERS457703	ERR551934	reference guided alignment	Nigeria	Western-Africa	2
SRS703247	SRR1577806, SRR1577830	reference guided alignment	Mali	Western-Africa	2

ERS456782	ERR550670	reference guided alignment	Georgia	Western-Asia	2
ERS457425	ERR551554	reference guided alignment	Georgia	Western-Asia	2
ERS457879	ERR552194	reference guided alignment	Azerbaijan	Western-Asia	2
ERS457894	ERR552219	reference guided alignment	Georgia	Western-Asia	2
ERS458079	ERR552482	reference guided alignment	Georgia	Western-Asia	2
ERS458638	ERR553237	reference guided alignment	Georgia	Western-Asia	2
JHUF01	JHUF00000000.1	multiple genome alignment	Georgia	Western-Asia	2
ERS456765	ERR550643, ERR550644	reference guided alignment	Germany	Western-Europe	2
ERS457161	ERR551184, ERR551185	reference guided alignment	Germany	Western-Europe	2
ERS457420	ERR551549	reference guided alignment	Germany	Western-Europe	2
ERS457438	ERR551572	reference guided alignment	Germany	Western-Europe	2
ERS457737	ERR551990	reference guided alignment	Germany	Western-Europe	2
ERS457807	ERR552090	reference guided alignment	Germany	Western-Europe	2
ERS458132	ERR552555	reference guided alignment	Germany	Western-Europe	2
ERS458211	ERR552668	reference guided alignment	Germany	Western-Europe	2
ERS458427	ERR552949	reference guided alignment	Germany	Western-Europe	2

ERS458431	ERR552954	reference guided alignment	Germany	Western-Europe	2
ERS218246	ERR234195	reference guided alignment	Turkmenistan	Central-Asia	3
ERS019846	ERR245793	reference guided alignment	Malawi	Eastern-Africa	3
ERS023446	ERR040123	reference guided alignment	Uganda	Eastern-Africa	3
ERS023459	ERR040136	reference guided alignment	Uganda	Eastern-Africa	3
ERS1028682	ERR1200607	reference guided alignment	Ethiopia	Eastern-Africa	3
ERS1028684	ERR1200609	reference guided alignment	Ethiopia	Eastern-Africa	3
ERS1028705	ERR1200630	reference guided alignment	Ethiopia	Eastern-Africa	3
ERS1028713	ERR1200638	reference guided alignment	Ethiopia	Eastern-Africa	3
ERS108133	ERR161200	reference guided alignment	Malawi	Eastern-Africa	3
ERS142212	ERR181693	reference guided alignment	Malawi	Eastern-Africa	3
ERS142694	ERR190394	reference guided alignment	Malawi	Eastern-Africa	3
ERS217635	ERR233347	reference guided alignment	Ethiopia	Eastern-Africa	3
ERS218204	ERR234153	reference guided alignment	Eritrea	Eastern-Africa	3
ERS218213	ERR234162	reference guided alignment	Tanzania	Eastern-Africa	3
ERS218217	ERR234166	reference guided alignment	Tanzania	Eastern-Africa	3

ERS218218	ERR234167	reference guided alignment	Tanzania	Eastern-Africa	3
ERS218269	ERR234218	reference guided alignment	Ethiopia	Eastern-Africa	3
ERS003250	ERR015615	reference guided alignment	Russia	Eastern-Europe	3
ERS218270	ERR234219	reference guided alignment	Vietnam	South-Eastern-Asia	3
SRS490577	SRR1062848, SRR1011477	reference guided alignment	South Africa	Southern-Africa	3
SRS494366	SRR1019128, SRR1140935	reference guided alignment	South Africa	Southern-Africa	3
ERS217648	ERR233360	reference guided alignment	Pakistan	Southern-Asia	3
ERS217669	ERR233381	reference guided alignment	Nepal	Southern-Asia	3
ERS217671	ERR233383	reference guided alignment	Nepal	Southern-Asia	3
ERS217673	ERR233385	reference guided alignment	Nepal	Southern-Asia	3
ERS217675	ERR233387	reference guided alignment	India	Southern-Asia	3
ERS217679	ERR233391	reference guided alignment	Nepal	Southern-Asia	3
ERS218150	ERR234099	reference guided alignment	Nepal	Southern-Asia	3
ERS218160	ERR234109	reference guided alignment	Nepal	Southern-Asia	3
ERS218162	ERR234111	reference guided alignment	Nepal	Southern-Asia	3
ERS218209	ERR234158	reference guided alignment	Iran	Southern-Asia	3

ERS218219	ERR234168	reference guided alignment	Afghanistan	Southern-Asia	3
ERS218232	ERR234181	reference guided alignment	Pakistan	Southern-Asia	3
ERS218239	ERR234188	reference guided alignment	Sri Lanka	Southern-Asia	3
ERS218240	ERR234189	reference guided alignment	Afghanistan	Southern-Asia	3
ERS218243	ERR234192	reference guided alignment	Afghanistan	Southern-Asia	3
ERS218247	ERR234196	reference guided alignment	Afghanistan	Southern-Asia	3
ERS218283	ERR234232	reference guided alignment	India	Southern-Asia	3
ERS218284	ERR234233	reference guided alignment	India	Southern-Asia	3
ERS611783	ERR688008	reference guided alignment	Pakistan	Southern-Asia	3
ERS611784	ERR688009	reference guided alignment	Pakistan	Southern-Asia	3
ERS611785	ERR688010	reference guided alignment	Pakistan	Southern-Asia	3
ERS611786	ERR688011	reference guided alignment	Pakistan	Southern-Asia	3
ERS611787	ERR688012	reference guided alignment	Pakistan	Southern-Asia	3
ERS611791	ERR688016	reference guided alignment	Pakistan	Southern-Asia	3
ERS611792	ERR688017	reference guided alignment	Pakistan	Southern-Asia	3
ERS611793	ERR688018	reference guided alignment	Pakistan	Southern-Asia	3

ERS611796	ERR688021	reference guided alignment	Pakistan	Southern-Asia	3
ERS611797	ERR688022	reference guided alignment	Pakistan	Southern-Asia	3
ERS611801	ERR688026	reference guided alignment	Pakistan	Southern-Asia	3
ERS611802	ERR688027	reference guided alignment	Pakistan	Southern-Asia	3
ERS611804	ERR688029	reference guided alignment	Pakistan	Southern-Asia	3
ERS611806	ERR688031	reference guided alignment	Pakistan	Southern-Asia	3
ERS611807	ERR688032	reference guided alignment	Pakistan	Southern-Asia	3
ERS611808	ERR688033	reference guided alignment	Pakistan	Southern-Asia	3
ERS611811	ERR688036	reference guided alignment	Pakistan	Southern-Asia	3
ERS611815	ERR688040	reference guided alignment	Pakistan	Southern-Asia	3
ERS611819	ERR688044	reference guided alignment	Pakistan	Southern-Asia	3
ERS611820	ERR688045	reference guided alignment	Pakistan	Southern-Asia	3
ERS611822	ERR688047	reference guided alignment	Pakistan	Southern-Asia	3
SRS559442	SRR1172065, SRR1172075	reference guided alignment	Iran	Southern-Asia	3
SRS559734	SRR1172879, SRR1173771	reference guided alignment	India	Southern-Asia	3
SRS559742	SRR1172895	reference guided alignment	Iran	Southern-Asia	3

ERS217672	ERR233384	reference guided alignment	Portugal	Southern-Europe	3
ERS248000	ERR275205	reference guided alignment	Portugal	Southern-Europe	3
ERS023435	ERR040112	reference guided alignment	Uganda	Eastern-Africa	4
ERS023438	ERR040115	reference guided alignment	Uganda	Eastern-Africa	4
ERS023445	ERR040122	reference guided alignment	Uganda	Eastern-Africa	4
ERS023448	ERR040125	reference guided alignment	Uganda	Eastern-Africa	4
ERS023450	ERR040127	reference guided alignment	Uganda	Eastern-Africa	4
ERS023451	ERR040128	reference guided alignment	Uganda	Eastern-Africa	4
ERS023455	ERR040132	reference guided alignment	Uganda	Eastern-Africa	4
ERS023461	ERR040138	reference guided alignment	Uganda	Eastern-Africa	4
ERS023466	ERR038736	reference guided alignment	Uganda	Eastern-Africa	4
ERS023468	ERR038738	reference guided alignment	Uganda	Eastern-Africa	4
ERS023469	ERR038739	reference guided alignment	Uganda	Eastern-Africa	4
ERS023470	ERR038740	reference guided alignment	Uganda	Eastern-Africa	4
ERS023471	ERR038741	reference guided alignment	Uganda	Eastern-Africa	4
ERS023472	ERR038742	reference guided alignment	Uganda	Eastern-Africa	4

ERS023474	ERR038744	reference guided alignment	Uganda	Eastern-Africa	4
ERS023477	ERR038747	reference guided alignment	Uganda	Eastern-Africa	4
ERS023478	ERR038748	reference guided alignment	Uganda	Eastern-Africa	4
ERS023479	ERR038749	reference guided alignment	Uganda	Eastern-Africa	4
ERS023480	ERR038750	reference guided alignment	Uganda	Eastern-Africa	4
ERS023481	ERR038751	reference guided alignment	Uganda	Eastern-Africa	4
ERS078252	ERR124634	reference guided alignment	Malawi	Eastern-Africa	4
ERS1028680	ERR1200605	reference guided alignment	Ethiopia	Eastern-Africa	4
ERS1028686	ERR1200611	reference guided alignment	Ethiopia	Eastern-Africa	4
ERS1028689	ERR1200614	reference guided alignment	Ethiopia	Eastern-Africa	4
ERS1028690	ERR1200615	reference guided alignment	Ethiopia	Eastern-Africa	4
ERS1028691	ERR1200616	reference guided alignment	Ethiopia	Eastern-Africa	4
ERS1028695	ERR1200620	reference guided alignment	Ethiopia	Eastern-Africa	4
ERS1028696	ERR1200621	reference guided alignment	Ethiopia	Eastern-Africa	4
ERS1028702	ERR1200627	reference guided alignment	Ethiopia	Eastern-Africa	4
ERS1028703	ERR1200628	reference guided alignment	Ethiopia	Eastern-Africa	4

ERS1028712	ERR1200637	reference guided alignment	Ethiopia	Eastern-Africa	4
ERS141583	ERR176515	reference guided alignment	Malawi	Eastern-Africa	4
ERS141816	ERR176484	reference guided alignment	Malawi	Eastern-Africa	4
ERS142107	ERR181972	reference guided alignment	Malawi	Eastern-Africa	4
ERS142122	ERR181987	reference guided alignment	Malawi	Eastern-Africa	4
ERS142343	ERR181818	reference guided alignment	Malawi	Eastern-Africa	4
ERS153902	ERR212051	reference guided alignment	Malawi	Eastern-Africa	4
ERS154027	ERR212176	reference guided alignment	Malawi	Eastern-Africa	4
ERS163344	ERR221528	reference guided alignment	Malawi	Eastern-Africa	4
ERS217636	ERR233348	reference guided alignment	Ethiopia	Eastern-Africa	4
ERS217638	ERR233350	reference guided alignment	Ethiopia	Eastern-Africa	4
ERS217640	ERR233352	reference guided alignment	Ethiopia	Eastern-Africa	4
ERS217657	ERR233369	reference guided alignment	Ethiopia	Eastern-Africa	4
ERS218212	ERR234161	reference guided alignment	Tanzania	Eastern-Africa	4
ERS218214	ERR234163	reference guided alignment	Tanzania	Eastern-Africa	4
ERS218222	ERR234171	reference guided alignment	Uganda	Eastern-Africa	4

ERS218223	ERR234172	reference guided alignment	Uganda	Eastern-Africa	4
ERS218224	ERR234173	reference guided alignment	Uganda	Eastern-Africa	4
ERS218280	ERR234229	reference guided alignment	China	Eastern-Asia	4
SRS490605	SRR1011512, SRR1019187	reference guided alignment	Taiwan	Eastern-Asia	4
SRS490609	SRR1019191, SRR1011517	reference guided alignment	Taiwan	Eastern-Asia	4
SRS490610	SRR1019192, SRR1011521	reference guided alignment	Taiwan	Eastern-Asia	4
SRS515999	SRR1051198, SRR1051199	reference guided alignment	Taiwan	Eastern-Asia	4
ERS050935	ERR067732	reference guided alignment	Russia	Eastern-Europe	4
ERS053656	ERR067629	reference guided alignment	Russia	Eastern-Europe	4
ERS053685	ERR067658	reference guided alignment	Russia	Eastern-Europe	4
ERS066661	ERR108481	reference guided alignment	Russia	Eastern-Europe	4
ERS066698	ERR117451	reference guided alignment	Russia	Eastern-Europe	4
ERS094089	ERR133800	reference guided alignment	Russia	Eastern-Europe	4
ERS094163	ERR133874	reference guided alignment	Russia	Eastern-Europe	4
ERS094226	ERR133937	reference guided alignment	Russia	Eastern-Europe	4
ERS094227	ERR133938	reference guided alignment	Russia	Eastern-Europe	4

ERS094241	ERR133952	reference guided alignment	Russia	Eastern-Europe	4
ERS096358	ERR137228	reference guided alignment	Russia	Eastern-Europe	4
ERS181386	ERR227992	reference guided alignment	Russia	Eastern-Europe	4
ERS181392	ERR227998	reference guided alignment	Russia	Eastern-Europe	4
ERS181428	ERR228034	reference guided alignment	Russia	Eastern-Europe	4
ERS181536	ERR229987	reference guided alignment	Russia	Eastern-Europe	4
SRS555850	SRR1166100	reference guided alignment	Romania	Eastern-Europe	4
SRS555858	SRR1166108	reference guided alignment	Romania	Eastern-Europe	4
SRS555863	SRR1169082, SRR1166116	reference guided alignment	Moldova	Eastern-Europe	4
SRS555869	SRR1166122	reference guided alignment	Romania	Eastern-Europe	4
SRS555879	SRR1166137	reference guided alignment	Romania	Eastern-Europe	4
SRS555896	SRR1166162	reference guided alignment	Romania	Eastern-Europe	4
SRS555906	SRR1166178	reference guided alignment	Romania	Eastern-Europe	4
SRS555924	SRR1166253	reference guided alignment	Romania	Eastern-Europe	4
SRS555929	SRR1166282, SRR1169055	reference guided alignment	Moldova	Eastern-Europe	4
SRS555952	SRR1166325	reference guided alignment	Romania	Eastern-Europe	4

SRS561126	SRR1175470	reference guided alignment	Romania	Eastern-Europe	4
ERS218203	ERR234152	reference guided alignment	Vietnam	South-Eastern-Asia	4
ERS218221	ERR234170	reference guided alignment	Vietnam	South-Eastern-Asia	4
ERS218231	ERR234180	reference guided alignment	Vietnam	South-Eastern-Asia	4
ERS218234	ERR234183	reference guided alignment	Vietnam	South-Eastern-Asia	4
ERS218290	ERR234239	reference guided alignment	Vietnam	South-Eastern-Asia	4
ERS218294	ERR234243	reference guided alignment	Vietnam	South-Eastern-Asia	4
ERS218295	ERR234244	reference guided alignment	Vietnam	South-Eastern-Asia	4
SRS415168	SRR832986, SRR832994	reference guided alignment	South Africa	Southern-Africa	4
SRS415216	SRR833084, SRR833049	reference guided alignment	South Africa	Southern-Africa	4
SRS419330	SRR847783, SRR924228	reference guided alignment	South Africa	Southern-Africa	4
SRS419334	SRR847780, SRR924239	reference guided alignment	South Africa	Southern-Africa	4
SRS454481	SRR958216, SRR924703	reference guided alignment	South Africa	Southern-Africa	4
SRS490579	SRR1062850, SRR1011479	reference guided alignment	South Africa	Southern-Africa	4
SRS490590	SRR1062861, SRR1011490	reference guided alignment	South Africa	Southern-Africa	4
SRS494362	SRR1019125, SRR1140925	reference guided alignment	South Africa	Southern-Africa	4

SRS494389	SRR1140965, SRR1019151	reference guided alignment	South Africa	Southern-Africa	4
AUXC01	AUXC00000000.1	multiple genome alignment	India	Southern-Asia	4
AVQJ01	AVQJ00000000.1	multiple genome alignment	India	Southern-Asia	4
ERS217667	ERR233379	reference guided alignment	Nepal	Southern-Asia	4
ERS217670	ERR233382	reference guided alignment	Nepal	Southern-Asia	4
ERS217677	ERR233389	reference guided alignment	Nepal	Southern-Asia	4
ERS217678	ERR233390	reference guided alignment	Nepal	Southern-Asia	4
ERS611805	ERR688030	reference guided alignment	Pakistan	Southern-Asia	4
ERS611810	ERR688035	reference guided alignment	Pakistan	Southern-Asia	4
ERS611816	ERR688041	reference guided alignment	Pakistan	Southern-Asia	4
ERS611818	ERR688043	reference guided alignment	Pakistan	Southern-Asia	4
SRS559444	SRR1180218, SRR1183118	reference guided alignment	Iran	Southern-Asia	4
SRS559452	SRR1172151, SRR1172252, SRR1180184, SRR1183085	reference guided alignment	Iran	Southern-Asia	4
SRS559453	SRR1172164, SRR1172203, SRR1180156, SRR1183110	reference guided alignment	Iran	Southern-Asia	4
SRS559455	SRR1172171, SRR1172359, SRR1180243, SRR1183100	reference guided alignment	Iran	Southern-Asia	4

SRS559461	SRR1172189, SRR1172210, SRR1180146, SRR1182972	reference guided alignment	Iran	Southern-Asia	4
SRS559628	SRR1172738, SRR1172979	reference guided alignment	India	Southern-Asia	4
SRS559656	SRR1172774, SRR1172887	reference guided alignment	India	Southern-Asia	4
SRS559768	SRR1172944, SRR1174323	reference guided alignment	Iran	Southern-Asia	4
SRS559794	SRR1172989, SRR1175160	reference guided alignment	India	Southern-Asia	4
SRS559976	SRR1173428	reference guided alignment	Iran	Southern-Asia	4
SRS560028	SRR1173580, SRR1173859	reference guided alignment	Iran	Southern-Asia	4
SRS560033	SRR1173615	reference guided alignment	Iran	Southern-Asia	4
SRS560040	SRR1173802, SRR1173637	reference guided alignment	India	Southern-Asia	4
ERS247976	ERR275181	reference guided alignment	Portugal	Southern-Europe	4
ERS247990	ERR275195	reference guided alignment	Portugal	Southern-Europe	4
ERS247999	ERR275204	reference guided alignment	Portugal	Southern-Europe	4
ERS248003	ERR275208	reference guided alignment	Portugal	Southern-Europe	4
ERS248013	ERR275218	reference guided alignment	Portugal	Southern-Europe	4
ERS248031	ERR275236	reference guided alignment	Portugal	Southern-Europe	4
ERS218238	ERR234187	reference guided alignment	Ghana	Western-Africa	4

ERS218249	ERR234198	reference guided alignment	Liberia	Western-Africa	4
ERS218251	ERR234200	reference guided alignment	Ghana	Western-Africa	4
ERS218252	ERR234201	reference guided alignment	Ghana	Western-Africa	4
ERS218254	ERR234203	reference guided alignment	Ghana	Western-Africa	4
ERS218257	ERR234206	reference guided alignment	Ghana	Western-Africa	4
SRS485045	SRR998634, SRR998633, SRR998632	reference guided alignment	Mali	Western-Africa	4
SRS485055	SRR998675, SRR998674, SRR998672	reference guided alignment	Mali	Western-Africa	4
SRS485057	SRR998683, SRR998681	reference guided alignment	Mali	Western-Africa	4
SRS485060	SRR1049965, SRR998693, SRR998694	reference guided alignment	Mali	Western-Africa	4
SRS485068	SRR998725, SRR998726, SRR998724	reference guided alignment	Mali	Western-Africa	4
SRS485077	SRR998763, SRR998760	reference guided alignment	Mali	Western-Africa	4
SRS485082	SRR998783, SRR998782	reference guided alignment	Mali	Western-Africa	4
SRS485084	SRR998791, SRR998789, SRR998788, SRR998790	reference guided alignment	Mali	Western-Africa	4
SRS526927	SRR1162746, SRR1162473	reference guided alignment	Mali	Western-Africa	4
ERS218205	ERR234154	reference guided alignment	Germany	Western-Europe	4
ERS218225	ERR234174	reference guided alignment	Germany	Western-Europe	4

ERS218226	ERR234175	reference guided alignment	Germany	Western-Europe	4
ERS218228	ERR234177	reference guided alignment	Germany	Western-Europe	4
ERS218148	ERR234097	reference guided alignment	Sierra Leone	Western-Africa	5
ERS218164	ERR234113	reference guided alignment	Ghana	Western-Africa	5
ERS218193	ERR234142	reference guided alignment	Ghana	Western-Africa	5
ERS218194	ERR234143	reference guided alignment	Ghana	Western-Africa	5
ERS218195	ERR234144	reference guided alignment	Ghana	Western-Africa	5
ERS218197	ERR234146	reference guided alignment	Sierra Leone	Western-Africa	5
ERS218198	ERR234147	reference guided alignment	Ghana	Western-Africa	5
ERS218200	ERR234149	reference guided alignment	Ghana	Western-Africa	5
ERS218211	ERR234160	reference guided alignment	Ghana	Western-Africa	5
ERS218227	ERR234176	reference guided alignment	Ghana	Western-Africa	5
ERS218250	ERR234199	reference guided alignment	Ghana	Western-Africa	5
ERS218253	ERR234202	reference guided alignment	Ghana	Western-Africa	5
ERS218255	ERR234204	reference guided alignment	Ghana	Western-Africa	5
SRS485033	SRR998584, SRR998587	reference guided alignment	Mali	Western-Africa	5

SRS485041	SRR998616, SRR998617	reference guided alignment	Mali	Western-Africa	5
ERS217650	ERR233362	reference guided alignment	Senegal	Western-Africa	6
ERS217654	ERR233366	reference guided alignment	The Gambia	Western-Africa	6
ERS217655	ERR233367	reference guided alignment	The Gambia	Western-Africa	6
ERS218157	ERR234106	reference guided alignment	Ghana	Western-Africa	6
ERS218199	ERR234148	reference guided alignment	Ghana	Western-Africa	6
ERS218201	ERR234150	reference guided alignment	Ghana	Western-Africa	6
ERS218235	ERR234184	reference guided alignment	Ghana	Western-Africa	6
ERS218237	ERR234186	reference guided alignment	Ghana	Western-Africa	6
ERS218305	ERR234254	reference guided alignment	The Gambia	Western-Africa	6
ERS218306	ERR234255	reference guided alignment	The Gambia	Western-Africa	6
ERS218312	ERR234261	reference guided alignment	The Gambia	Western-Africa	6
SRS485031	SRR998576, SRR998578, SRR998579	reference guided alignment	Mali	Western-Africa	6
SRS485032	SRR998580, SRR998583	reference guided alignment	Mali	Western-Africa	6
SRS485035	SRR1049958, SRR998594, SRR998592, SRR998593	reference guided alignment	Mali	Western-Africa	6
SRS485036	SRR998598, SRR998599, SRR998597	reference guided alignment	Mali	Western-Africa	6

SRS485037	SRR998602, SRR998600	reference guided alignment	Mali	Western-Africa	6
SRS485038	SRR998605, SRR998606, SRR998607	reference guided alignment	Mali	Western-Africa	6
SRS485039	SRR998610, SRR998608	reference guided alignment	Mali	Western-Africa	6
SRS485040	SRR998614, SRR998612	reference guided alignment	Mali	Western-Africa	6
SRS485042	SRR998620, SRR998622	reference guided alignment	Mali	Western-Africa	6
SRS485044	SRR998629, SRR998630, SRR998628	reference guided alignment	Mali	Western-Africa	6
SRS485046	SRR998636, SRR998639	reference guided alignment	Mali	Western-Africa	6
SRS485047	SRR998643, SRR998640, SRR998641	reference guided alignment	Mali	Western-Africa	6
SRS485048	SRR998646, SRR998647, SRR998645	reference guided alignment	Mali	Western-Africa	6
SRS485049	SRR998651, SRR998650, SRR1049960, SRR998649	reference guided alignment	Mali	Western-Africa	6
SRS485050	SRR998652, SRR998655, SRR998654, SRR998653	reference guided alignment	Mali	Western-Africa	6
SRS485072	SRR998742, SRR998741	reference guided alignment	Mali	Western-Africa	6
SRS526842	SRR1162479, SRR1162789, SRR1103387	reference guided alignment	Mali	Western-Africa	6
SRS526933	SRR1162477, SRR1103499, SRR1162738	reference guided alignment	Mali	Western-Africa	6
SRS526956	SRR1162470, SRR1103551, SRR1162788	reference guided alignment	Mali	Western-Africa	6
SRS703269	SRR1577831, SRR1577834, SRR1577820	reference guided alignment	Mali	Western-Africa	6

ERS1028677	ERR1200602	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS1028678	ERR1200603	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS1028679	ERR1200604	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS1028681	ERR1200606	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS1028683	ERR1200608	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS1028685	ERR1200610	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS1028687	ERR1200612	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS1028688	ERR1200613	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS1028692	ERR1200617	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS1028693	ERR1200618	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS1028694	ERR1200619	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS1028698	ERR1200623	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS1028699	ERR1200624	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS1028700	ERR1200625	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS1028706	ERR1200631	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS1028708	ERR1200633	reference guided alignment	Ethiopia	Eastern-Africa	7

ERS1028709	ERR1200634	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS1028710	ERR1200635	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS1028711	ERR1200636	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS1028714	ERR1200639	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS1028715	ERR1200640	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS158317	ERR159956	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS158319	ERR159958	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS158320	ERR159959	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS217634	ERR233346	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS217637	ERR233349	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS217642	ERR233354	reference guided alignment	Ethiopia	Eastern-Africa	7
ERS217643	ERR233355	reference guided alignment	Ethiopia	Eastern-Africa	7
L4_N0101	ERR234220	reference guided alignment	Nicaragua	Central-America	4
L4_N0120	ERR234221	reference guided alignment	Puerto Rico	Central-America	4
L4_N0137	ERR234226	reference guided alignment	Mexico	Central-America	4
L4_N0046	ERR234251	reference guided alignment	Mexico	Central-America	4

L4_N0103	ERR234258	reference guided alignment	Nicaragua	Central-America	4
L4_N0107	ERR234259	reference guided alignment	Guatemal	Central-America	4
L4_N0109	ERR234260	reference guided alignment	El Salvador	Central-America	4
L4_N0131	ERR234223	reference guided alignment	USA	Northern-America	4
L4_N0135	ERR234225	reference guided alignment	USA	Northern-America	4
L4_N0142	ERR234227	reference guided alignment	USA	Northern-America	4
L4_N0143	ERR234228	reference guided alignment	USA	Northern-America	4
L4_N0149	ERR234230	reference guided alignment	USA	Northern-America	4
L4_N0136	ERR234265	reference guided alignment	USA	Northern-America	4
L4_N0146	ERR234267	reference guided alignment	USA	Northern-America	4
L4_N0148	ERR234268	reference guided alignment	USA	Northern-America	4

Table 2. Accessions and metadata for *M. tb* L4.4

Name	Run ID	Project ID	Country	Region	Sub-lineage	Year
ERR024359	ERR024359	PRJEB2057	Netherlands	Europe	4.4.1.1	1999
ERR040124	ERR040124	PRJEB2424	Uganda	Africa	4.4.1.1	2005
ERR040130	ERR040130	PRJEB2424	Uganda	Africa	4.4.1.1	2004
ERR046901	ERR046901	PRJEB2221	United Kingdom	Europe	4.4.1.1	2007
ERR047013	ERR047013	PRJEB2221	United Kingdom	Europe	4.4.1.1	2006
ERR108481	ERR108481	PRJEB2138	Russia	Europe	4.4.1.1	2009
ERR133974	ERR133974	PRJEB2138	Russia	Europe	4.4.1.1	2009
ERR161047	ERR161047	PRJEB2794	Malawi	Africa	4.4.1.1	2003

ERR163930	ERR163930	PRJEB2794	Malawi	Africa	4.4.1.1	2004
ERR163942	ERR163942	PRJEB2794	Malawi	Africa	4.4.1.1	2004
ERR163946	ERR163946	PRJEB2794	Malawi	Africa	4.4.1.1	2004
ERR164012	ERR164012	PRJEB2794	Malawi	Africa	4.4.1.1	1997
ERR176454	ERR176454	PRJEB2794	Malawi	Africa	4.4.1.1	2004
ERR176628	ERR176628	PRJEB2794	Malawi	Africa	4.4.1.1	1997
ERR176725	ERR176725	PRJEB2794	Malawi	Africa	4.4.1.1	1998
ERR181836	ERR181836	PRJEB2794	Malawi	Africa	4.4.1.1	2009
ERR181853	ERR181853	PRJEB2794	Malawi	Africa	4.4.1.1	2009
ERR181946	ERR181946	PRJEB2794	Malawi	Africa	4.4.1.1	2005
ERR182011	ERR182011	PRJEB2794	Malawi	Africa	4.4.1.1	2006
ERR190410	ERR190410	PRJEB2794	Malawi	Africa	4.4.1.1	1997
ERR211992	ERR211992	PRJEB2794	Malawi	Africa	4.4.1.1	2002
ERR212117	ERR212117	PRJEB2794	Malawi	Africa	4.4.1.1	1999
ERR212132	ERR212132	PRJEB2794	Malawi	Africa	4.4.1.1	2009
ERR212152	ERR212152	PRJEB2794	Malawi	Africa	4.4.1.1	2009
ERR212159	ERR212159	PRJEB2794	Malawi	Africa	4.4.1.1	2009
ERR221551	ERR221551	PRJEB2794	Malawi	Africa	4.4.1.1	2002
ERR221554	ERR221554	PRJEB2794	Malawi	Africa	4.4.1.1	2002
ERR221600	ERR221600	PRJEB2794	Malawi	Africa	4.4.1.1	2002
ERR228025	ERR228025	PRJEB2138	Russia	Europe	4.4.1.1	2010
ERR228067	ERR228067	PRJEB2138	Russia	Europe	4.4.1.1	2010
ERR245800	ERR245800	PRJEB2358	Malawi	Africa	4.4.1.1	2001
NZO1	SRR5074294	PRJNA356104	New Zealand	Oceania	4.4.1.1	2008
NZO2	SRR5074712	PRJNA356104	New Zealand	Oceania	4.4.1.1	2011
NZO3	SRR5074713	PRJNA356104	New Zealand	Oceania	4.4.1.1	2008
NZO4	NA	NA	New Zealand	Oceania	4.4.1.1	2013
NZO5	NA	NA	New Zealand	Oceania	4.4.1.1	2003
NZO6	NA	NA	New Zealand	Oceania	4.4.1.1	2013
NZO7	NA	NA	New Zealand	Oceania	4.4.1.1	2006
NZR22	SRR8420474	PRJNA513885	New Zealand	Oceania	4.4.1.1	2010
NZR278	SRR8420475	PRJNA513885	New Zealand	Oceania	4.4.1.1	2010
NZR486	SRR8420472	PRJNA513885	New Zealand	Oceania	4.4.1.1	2011

NZR494	SRR8420473	PRJNA513885	New Zealand	Oceania	4.4.1.1	2011
NZRA	NA	NA	New Zealand	Oceania	4.4.1.1	1992
NZRB	NA	NA	New Zealand	Oceania	4.4.1.1	1992
NZRC	NA	NA	New Zealand	Oceania	4.4.1.1	1999
NZRE	NA	NA	New Zealand	Oceania	4.4.1.1	2001
NZRF	NA	NA	New Zealand	Oceania	4.4.1.1	2002
NZRH	NA	NA	New Zealand	Oceania	4.4.1.1	2006
NZRI	NA	NA	New Zealand	Oceania	4.4.1.1	2006
NZRJ	NA	NA	New Zealand	Oceania	4.4.1.1	2008
NZRK	NA	NA	New Zealand	Oceania	4.4.1.1	2008
NZRL	NA	NA	New Zealand	Oceania	4.4.1.1	1996
NZRM	NA	NA	New Zealand	Oceania	4.4.1.1	2009
NZRN	NA	NA	New Zealand	Oceania	4.4.1.1	1991
SK1	NA	NA	Canada	Americas	4.4.1.1	1994
SK111	NA	NA	Canada	Americas	4.4.1.1	2003
SK153	NA	NA	Canada	Americas	4.4.1.1	1996
SK160	NA	NA	Canada	Americas	4.4.1.1	1999
SK192	NA	NA	Canada	Americas	4.4.1.1	1990
SK196	NA	NA	Canada	Americas	4.4.1.1	1987
SK198	NA	NA	Canada	Americas	4.4.1.1	1987
SK199	NA	NA	Canada	Americas	4.4.1.1	1989
SK20	NA	NA	Canada	Americas	4.4.1.1	1988
SK201	NA	NA	Canada	Americas	4.4.1.1	2001
SK202	NA	NA	Canada	Americas	4.4.1.1	1994
SK204	NA	NA	Canada	Americas	4.4.1.1	2002
SK211	NA	NA	Canada	Americas	4.4.1.1	1997
SK215	NA	NA	Canada	Americas	4.4.1.1	1999
SK254	NA	NA	Canada	Americas	4.4.1.1	2004
SK256	NA	NA	Canada	Americas	4.4.1.1	1987
SK260	NA	NA	Canada	Americas	4.4.1.1	1996
SK287	NA	NA	Canada	Americas	4.4.1.1	2002
SK288	NA	NA	Canada	Americas	4.4.1.1	2003

SK30	NA	NA	Canada	Americas	4.4.1.1	2003
SK54	NA	NA	Canada	Americas	4.4.1.1	1990
SRR1019141	SRR1019141	PRJNA183624	Kenya	Africa	4.4.1.1	2013
SRR1019142	SRR1019142	PRJNA183624	South Africa	Africa	4.4.1.1	2013
SRR1019150	SRR1019150	PRJNA183624	South Africa	Africa	4.4.1.1	2013
SRR1019153	SRR1019153	PRJNA183624	Kenya	Africa	4.4.1.1	2013
SRR1019155	SRR1019155	PRJNA183624	South Africa	Africa	4.4.1.1	2013
SRR1019159	SRR1019159	PRJNA183624	South Africa	Africa	4.4.1.1	2013
SRR1019165	SRR1019165	PRJNA183624	South Africa	Africa	4.4.1.1	2013
SRR1019168	SRR1019168	PRJNA183624	South Africa	Africa	4.4.1.1	2013
SRR1140926	SRR1140926	PRJNA183624	South Africa	Africa	4.4.1.1	2013
SRR1162884	SRR1162884	PRJNA229360	Sweden	Europe	4.4.1.1	2010
SRR1163077	SRR1163077	PRJNA229360	Sweden	Europe	4.4.1.1	2010
SRR1166253	SRR1166253	PRJNA233386	Romania	Europe	4.4.1.1	2013
SRR1172724	SRR1172724	PRJNA235618	South Africa	Africa	4.4.1.1	2012
SRR1172876	SRR1172876	PRJNA235618	South Africa	Africa	4.4.1.1	2013
SRR1172935	SRR1172935	PRJNA235618	South Africa	Africa	4.4.1.1	2013
SRR1173087	SRR1173087	PRJNA235618	South Africa	Africa	4.4.1.1	2013
SRR1173181	SRR1173181	PRJNA235618	South Africa	Africa	4.4.1.1	2013
SRR1173353	SRR1173353	PRJNA235618	South Africa	Africa	4.4.1.1	2013
SRR1173499	SRR1173499	PRJNA235618	South Africa	Africa	4.4.1.1	2013
SRR1173522	SRR1173522	PRJNA235618	South Africa	Africa	4.4.1.1	2013
SRR1173637	SRR1173637	PRJNA235852	India	Asia	4.4.1.1	2003
SRR1175041	SRR1175041	PRJNA235618	South Africa	Africa	4.4.1.1	2013
SRR1180189	SRR1180189	PRJNA235618	South Africa	Africa	4.4.1.1	2013
SRR1180314	SRR1180314	PRJNA235618	South Africa	Africa	4.4.1.1	2013
SRR1181100	SRR1181100	PRJNA191021	Colombia	Americas	4.4.1.1	2009
SRR1181216	SRR1181216	PRJNA235618	South Africa	Africa	4.4.1.1	2013
SRR1184309	SRR1184309	PRJNA235618	South Africa	Africa	4.4.1.1	2013
SRR5065416	SRR5065416	PRJNA355614	Vietnam	Asia	4.4.1.1	2011
SRR5067392	SRR5067392	PRJNA355614	Vietnam	Asia	4.4.1.1	2009
SRR5073887	SRR5073887	PRJNA355614	Vietnam	Asia	4.4.1.1	2009
SRR5073966	SRR5073966	PRJNA355614	Vietnam	Asia	4.4.1.1	2009
SRR832977	SRR832977	PRJNA183624	South Africa	Africa	4.4.1.1	2008
SRR832988	SRR832988	PRJNA183624	South Africa	Africa	4.4.1.1	2009
SRR833034	SRR833034	PRJNA183624	South Africa	Africa	4.4.1.1	2008
SRR833044	SRR833044	PRJNA183624	South Africa	Africa	4.4.1.1	2010
SRR833119	SRR833119	PRJNA183624	South Africa	Africa	4.4.1.1	2008
SRR833134	SRR833134	PRJNA183624	South Africa	Africa	4.4.1.1	2010

SRR833144	SRR833144	PRJNA183624	South Africa	Africa	4.4.1.1	2008
SRR833165	SRR833165	PRJNA183624	South Africa	Africa	4.4.1.1	2008
SRR847795	SRR847795	PRJNA183624	South Africa	Africa	4.4.1.1	2011
SRR847797	SRR847797	PRJNA183624	South Africa	Africa	4.4.1.1	2011
SRR847802	SRR847802	PRJNA183624	South Africa	Africa	4.4.1.1	2011