Development and Application of Genome-Wide Association Studies in Bacteria

Sarah Grace Earle

Nuffield Department of Medicine

Lincoln College

University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

2017

Development and Application of Genome-Wide Association Studies in Bacteria

Sarah Grace Earle, Lincoln College, Oxford University
A thesis submitted for the degree of Doctor of Philosophy
Trinity Term 2017

Abstract

Since the first genome-wide association study (GWAS) applied to humans in 2005, incredible advances have been made in understanding the genetic basis underlying common human diseases and complex traits. Dramatic technological developments have enabled rapid, inexpensive whole-genome sequencing in large numbers of bacteria, creating intense interest in the large-scale application of GWAS to bacteria. However, fundamental differences between the genomes of humans and bacteria mean that although the methodological developments in the human setting are an invaluable starting point, novel methods tailored specifically to bacteria are required. This thesis concerns the development and application of GWAS in bacteria. Taking lessons from the past decade of human GWAS, I began by assessing the feasibility of GWAS in bacteria by investigating the bacterial genetic basis underlying antimicrobial resistance. I aimed to empirically test the feasibility of bacterial GWAS in light of particular challenges posed by bacteria such as strong population structure, genome-wide linkage disequilibrium and the presence of large accessory genomes. Specifically, I performed a detailed investigation into fusidic acid resistance in *Staphylococcus aureus* to assess the impact of controlling for population stratification in highly structured populations. This demonstrated the importance of controlling for population structure in reducing the number of false positives, but also the substantial cost in doing so. Testing for lineage-level associations enabled the inference of important lineage-level differences in phenotypes, typically discarded when controlling for population structure.

I then went on to apply the methods developed to two further phenotypes. The first, carriage vs invasive disease in *Neisseria meningitidis*, identified the known hyperinvasive ST-11 lineage to be associated with invasive disease, and suggested that newly-reported variants in genes involved in capsule production and phase variation play an important role in the virulence potential of meningococci in natural populations. I hypothesised that a combination of two particular variants upstream of the gene encoding the virulence factor fHbp (factor H binding protein), produces a second putative FNR box, a binding site for the global transcriptional regulator FNR, which may affect expression of fHbp. Finally, I investigated wild bird vs chicken colonisation in *Campylobacter jejuni* and identified lineage-level associations in agreement with previously identified host-associated lineage characteristics. I hypothesised that host-associated variants downstream of the CRISPR-Cas region, in genes involved in lipooligosaccharide biosynthesis and the chemotaxis pathway, represent pathways enabling *C. jejuni* to survive bacteriophages encountered upon colonising a new host. To conclude, I discussed the findings of this thesis and suggested areas for future development where new technologies and methods will enable bacterial GWAS to be further advanced.

Acknowledgements

This DPhil was funded by the Nuffield Department of Medicine and the Medical Research Council.

I would like to thank my supervisor Daniel Wilson for his support and guidance over the past four years. I am extremely grateful for all of the time he put into meetings and discussions about my work which have been fundamental to the progress of my thesis. Danny's support and his patience when introducing new concepts have enabled my development as a scientist. I would also like to thank Derrick Crook, Tim Peto and Sarah Walker for the opportunity to work on interesting data and for stimulating discussions.

I am grateful for collaborations with Martin Maiden, Samuel Sheppard and Guillaume Méric which enabled me to apply the methods we developed to interesting datasets. I would also like to thank Chieh-Hsi Wu and Jane Charlesworth who were my co-authors on our GWAS paper, I learnt a lot from working with them.

My office mates and colleagues Jess Hedge, Bernadette Young, Anna Sheppard, Nicola de Maio and Melissa Ward have made the day to day of my DPhil very enjoyable, and I would like to thank them for their friendship. I am also incredibly grateful to my friends away from Oxford, particularly Alex Townsend, Hannah Pollard-Earle, Jenny Davis, Hannah Powell, Katie Pearson and Ellen Rawlins, they have been a great support to me from the beginning. I would like to thank my parents Jonathan and Jo and my brother Daniel for all of their love and encouragement, and for their never-ending enthusiasm for discussing my research.

Finally, I would like to thank my husband Matt, my biggest supporter and encourager. I will always be grateful for his belief in me and his unfailing ability to make me laugh. I can't thank him enough for his endless love and support.

Declaration

I declare that the work described within this thesis is my own, except where explicitly stated below, and that this thesis is my own composition.

Chapter 2 – Methods for applying genome-wide association studies in bacteria

The theory of testing for lineage effects, testing the power of detecting lineage effects by simulating phenotypes and identifying non genome-wide principal components was developed by Assoc. Prof. Daniel Wilson, as described in Section 2.10. The pangenome pipeline described in Section 2.5 was developed by Dr Jane Charlesworth.

Chapter 3 – Proof of principle for bacterial genome-wide association studies

Data collection was performed by Dr Nicole Stoesser for *Escherichia coli* and *Klebsiella pneumoniae*, by Dr Claire Gordon for *Staphylococcus aureus* and by Dr Timothy Walker for *Mycobacterium tuberculosis*.

Development of the theory for testing lineage effects discussed in Section 3.4.6 was performed by Assoc. Prof. Daniel Wilson. Testing the power of detecting lineage effects by simulating phenotypes as discussed in Sections 3.4.4 and 3.4.6.3, was performed by Assoc. Prof. Daniel Wilson. Dr Jane Charlesworth performed the gene presence/absence analyses described in Section 3.4.7. Dr Chieh-Hsi Wu performed the SNP and kmer analyses for *E. coli*, *K. pneumoniae* and *M. tuberculosis* as described in Section 3.4.7.

Figure 3.15 was created based on R code provided by Assoc. Prof. Daniel Wilson. Figure 3.17 was created by Dr Chieh-Hsi Wu.

Chapter 4 – Genome-wide association study of *Neisseria meningitidis* carriage versus invasive disease

Data was obtained from pubMLST (https://pubmlst.org/neisseria/) and the European Nucleotide Archive (http://www.ebi.ac.uk/ena) in collaboration with Prof. Martin Maiden.

Figure 4.14 was created based on R code provided by Assoc. Prof. Daniel Wilson.

Chapter 5 – Genome-wide association study of *Campylobacter jejuni* wild bird versus chicken adaptation

Data was provided by Prof. Samuel Sheppard and Dr Guillaume Méric.

Figure 5.7 was created based on R code provided by Assoc. Prof. Daniel Wilson.

Abbreviations

ANOVA, Analysis of variance model
bp, base pair
CC, Clonal Complex
CDS, coding sequence
CFH, complement factor H
CI, confidence interval
CMH, Cochran-Mantel-Haenszel
CNP, copy number polymorphism
CRISPR, clustered regularly interspaced short palindromic repeats
FDR, false discovery rate
fHbp, factor H binding protein
FNR, fumerate and nitrate reduction regulator
FPR, false positive rate
FWER, family-wise error rate
GRR, genotypic relative risk
GWAS, genome-wide association study
HCV, Hepatitis C virus
HGT, horizontal gene transfer
HLA, human leucocyte antigen
kb, kilobase
LD, linkage disequilibrium
LMM, linear mixed model
LOS, lipooligosaccharide
LRT, likelihood ratio test
MAF, minor allele frequency
Mb, megabase
MDS, multidimensional scaling
MIC, minimim inhibitory concentration
ML, maximum likelihood
MLST, Multi-locus sequence type
MMM, Modernising Medical Microbiology
OR, odds ratio
PC, principal component
PCA, principal component analysis
PCR, polymerase chain reaction
PPV, positive predictive value
QQ, quantile-quantile
SCC, Staphylococcal cassette chromosome
SNP, single nucleotide polymorphism
ST, sequence type
WGS, whole genome sequencing

Table of contents

Chapter 1

Introduction to genome-wide association studies

1 Introduction to genome-wide association studies

Since the first genome-wide association study (GWAS) in 2005 (Klein et al. 2005) incredible progress has been made in understanding the role human genetics play in common human diseases and other complex traits (Price, Spencer & Donnelly 2015). More recently, interest has been building in the application of these analyses to bacteria, aiming to understand the bacterial genetic basis to important diseases and phenotypes of interest (Power, Parkhill & de Oliveira 2017). Developments in large-scale genome sequencing of short fragments of DNA have enabled large volumes of bacterial sequence data to be produced (Read & Massey 2014). The first fully-recognisable bacterial GWAS was published in 2013 (Sheppard et al. 2013) and since then multiple studies have applied or developed GWAS methods in order to validate methods and gain insights into various important phenotypes such as antimicrobial resistance, host association and virulence. However, there are significant differences between human and bacterial genomes which influence the way GWAS are undertaken in the bacterial setting, and the development of analysis methods tailored specifically for bacterial settings is still very much in its infancy. The successes of human GWAS are encouraging to the field of bacterial GWAS, but important lessons can be learnt from a decade's worth of intensive methodological development and application. In this chapter I review the development of human association studies and the progress which has been made in vital methodological areas such as controlling for population stratification. I review the differences between human and bacterial genomes and the progress thus far in applying GWAS to bacteria, before setting out the aims of this thesis and outlining its structure.

1.1 Human genome wide association studies

1.1.1 Early genotype-phenotype studies in humans

Before the GWAS era two approaches were commonly used to understand the genetic basis of common diseases, family-based linkage studies and candidate gene association studies. The aim of linkage studies is to identify markers within families associated with the trait more often than expected by chance, and candidate gene studies assess the significance of a gene for which there is prior evidence of its possible role in the trait of interest. Linkage analyses began with the investigation of fruit flies in 1913 (Sturtevant 1913) but their application to humans was not until a human genetic map was produced based on naturally occurring DNA sequence polymorphisms in 1980 (Botstein et al. 1980), enabling the feasibility of linkage analyses to be established by a study which localised Huntington disease in 1983 to chromosome four (Gusella et al. 1983). A genetic map determines the order and relative distance between sequence polymorphisms, and the use of linkage studies enabled the discovery of thousands of genes for rare Mendelian diseases (Altshuler, Daly & Lander 2008).

Parameter-based linkage methods aim to identify genetic markers flanking a putative disease locus that is hypothesised to segregate with the disease-of-interest within families. Parameter-free linkage methods test whether affected relatives share an excess of haplotypes that are identical by descent in any particular region (Dawn Teare & Barrett 2005). For example, if a marker at a particular site has no linkage to the disease in question, then we would expect siblings to share alleles 50% of the time. If there is linkage however, we would expect affected siblings to share the same allele at a particular site more than 50% of the time. The more frequently affected siblings share the same allele at a site, the more likely it is close to the disease variant (Risch & Merikangas 1996).

Linkage studies had many successes in mapping genes underlying monogenic Mendelian diseases (for example, Knowlton et al. 1985; Corder et al. 1993), but these successes have not translated into discoveries for common diseases, which has been attributed to low power for variants of a small effect (Risch & Merikangas 1996; Risch 2000). Risch & Merikangas (1996) discussed how large the effect size of a gene must be in order for linkage analysis to detect it with 80% power. Genotypic relative risk (GRR) is the increased chance that an individual with a particular genotype has the disease/trait. Risch and Merikangas demonstrated that mutant alleles which when heterozygous confer a GRR of disease of at least 2 times more than the homozygoes wild type, would require at least around 2,500 familes in order to marginally exceed 50% marker sharing for allele frequencies of 50%. More families would be required for lower and higher allele frequencies and lower GRRs. They also demonstrated how association studies have more power in cases where the GRR is low and that the main limitation at the time was technological, because association analyses require a large number of polymorphisms to first be identified (Risch & Merikangas 1996). They suggested that association studies would be particularly fruitful for identifying genes containing common variants conferring a small effect on disease risk, as expected for complex disorders. For genetically complex diseases, risk alleles are more probabilistic than deterministic, as the presence of a risk variant may only slightly increase the probability of disease rather than cause it (Hirschhorn et al. 2002), and in these cases association studies will be more powerful.

Candidate gene studies have also been successful in identifying disease conferring variants. They do not rely on large family pedigrees, but instead work by focusing on a particular gene or region selected because of a prior hypothesis based on biological understanding about its role in disease or the trait of interest. Associations between

variation in the gene/region and the phenotype are then assessed. This means that families with only one affected member can be used in association studies. Among the successes of candidate gene studies was the discovery of associations between the human leukocyte antigen (HLA) complex and many autoimmune diseases such as type 1 diabetes (Nepon & Erlich 1991).

Despite these successes, candidate gene studies are necessarily limited to specific genes, and therefore ignore most of the sequence variation present in the human genome. Many studies failed to replicate (Lohmueller et al. 2003) and initial studies also typically suggested stronger effects than the following replication studies (Ioannidis et al. 2001), a phenomenon known as the winner's curse. Some variants have been validated however in large validation studies and meta-analyses (e.g. Cox et al. 2007; Liao et al. 2010) but the majority have not (Dong et al. 2008), indicating a high false positive rate. This may be due to poor choices in candidate loci, underpowered studies and susceptibility to confounding (Altshuler, Daly & Lander 2008).

1.1.2 Patterns of human genetic variation

The completion of the human genome sequence (International Human Genome Sequencing Consortium 2001; Venter et al. 2001), high density SNP maps (The International SNP Map Working Group 2001), rapid development in genotyping technologies (Ragoussis 2009) and the initiation of the International HapMap project (The International HapMap Consortium 2003) enabled the introduction of large scale genome-wide association studies by facilitating the collection of large cohorts for many common diseases with very densely characterised genetic markers (The Wellcome Trust Case Control Consortium 2007).

As single nucleotide polymorphism (SNP) typing increased in density, it became apparent that the correlations between variants in the human genome form a block-like

structure, where the blocks represent regions within which little recombination has occurred to break down the correlations, inspiring the idea that common SNP genotypes can be inferred using a pared down collection of locally-occurring "tag" SNPs (Daly et al. 2001; Johnson et al. 2001; Gabriel et al. 2002; Zhang et al. 2002). This correlation is known as linkage disequilibrium (LD) and a specific combination of alleles on a chromosome is called a haplotype. LD implies that individuals carrying a particular allele at one site can be accurately predicted to contain other specific alleles in the surrounding haplotype due to the shared ancestry of the regions. In fact, for the SNPs typed by Phase I HapMap, less than 500,000 SNPs were required to capture all observed common variants using an r^2 of 0.8 for all European, Asian and West African samples (The International HapMap Consortium 2005). Since then, reductions in the cost of genotyping have led to a steady increase in the number of SNPs typically typed in a GWAS. By 2005, the HapMap project had genotyped around 1 million common SNPs and by 2007 over 3 million. By 2010, these were integrated with copy number polymorphisms (CNPs) and lower-frequency SNPs (The International HapMap Consortium 2005; The International HapMap Consortium 2007; The International HapMap 3 Consortium 2010).

1.1.3 Standards for human genome-wide association studies

Driven forward by cheaper genotyping arrays and international consortia such as the HapMap project, the first GWAS began to appear in 2005 (Klein et al. 2005), with the stated aim of detecting common variants underlying common diseases (Price, Spencer & Donnelly 2015), and there are now 20,272 published associations at genome-wide significance ($P < 5 \times 10^{-8}$, as of August 2017, http://www.ebi.ac.uk/gwas/). Human GWAS is now a large field that necessarily forms the starting point for a well-informed foray into GWAS in bacteria. Here I discuss standards and conventions that have evolved in human GWAS, with an eye to the challenges and pitfalls that await bacterial GWAS.

1.1.3.1 Quality control

Early on it was established that genotyping arrays provide a rich but imperfect source of genetic data. In order to obtain robust data, once genotypes have been called they must be subjected to stringent quality control filtering. SNPs are typically filtered out due to extreme deviations from theoretical expectations (e.g. Hardy-Weinburg equilibrium) or empirical observations (e.g. high missingness) or to improve power (e.g. low minor allele frequency (MAF) variants) (The Wellcome Trust Case Control Consortium 2007). Association results are frequently visualised by inspecting the region of the association signal and through quantile-quantile (QQ) plots. The association region should reveal an elevated signal surrounding the association, decreasing with increasing distance from the signal due to LD (The Wellcome Trust Case Control Consortium 2007). QQ plots compare the empirical distribution of *P* values to the expected distribution under the null hypothesis. Systematic deviations other than in the extreme tail of the distribution can indicate poorly calibrated test statistics or other issues including uncontrolled population stratification (The Wellcome Trust Case Control Consortium 2007; Mccarthy et al. 2008).

1.1.3.2 Significance thresholds

A GWAS involves performing an extremely large number of tests, so stringent thresholds corrected for multiple testing are required to limit the number of false signals. Conventions concerning the adequacy of significance thresholds for limiting false positives have been agreed primarily in terms of classical *P* value thresholds. Control of the family-wise error rate (FWER) has been generally preferred over the less stringent false discovery rate (FDR; Benjamini & Hochberg 1995). In the former, the probability of a single false signal per study is limited, whereas in the latter the frequency of false positives among signals declared significant is limited. Pe're et al. (2008) estimated the effective number of independent tests in a human GWAS to be around 1 million, on the

basis of which a Bonferroni correction producing a threshold of $P < 5 \times 10^{-8}$ was proposed to limit the FWER to 5%. This has become the standard agreed by the community and journal editors (Fadista et al. 2016), but its derivation depends on a detailed understanding of the LD structure of the human genome.

1.1.3.3 Replication of significant results

The principle that associations must be replicated in a separate validation study of cases and controls has been widely embraced by researchers and editors (NCI-NHGRI Working Group 2007). Replication typically comprises repeating the original study in an independent population or providing experimental validation of significant findings. It is considered imperative for ensuring the reliability of genotype-phenotype associations. Evaluating an association in a population of a different ancestry to the original study in particular adds confidence to the original finding and may help fine-map the signal if the populations have different linkage patterns in the region (NCI-NHGRI Working Group 2007).

1.1.4 Imputation

As variants tested in GWAS do not represent all polymorphisms, but instead genetic markers which also tag untyped variants, increasing the number of genotyped variants can assist in refining the association signal and fine mapping the causal variant. One way of achieving this is by imputation. Imputation is a statistical framework for inferring genotypes which were not directly typed and can be performed genome wide or in particular regions in order to fine map an association signal (Marchini & Howie 2010). Genome wide imputation can increase the precision of an association study by increasing the number of markers available to test for association. Imputation can also be used to combine data sets typed on different genotyping arrays for a meta-analysis by imputing sites which were not typed in any one of the studies, increasing the number of markers

common to all studies (Porcu et al. 2013). Further, imputation can be extended to impute other variation such as copy number variation and to correct genotyping errors (Marchini & Howie 2010). In order to impute untyped variation, first a reference panel is produced from a modest-sized sample of densely genotyped unrelated individuals. This information is then used to impute a larger set of individuals typed at a subset of the SNPs, where the knowledge of the LD patterns of the human genome is exploited to infer the untyped variation. The HapMap Project and the more densely genotyped 1000 Genomes Project have served as reference panels which can be used to infer unknown genotypes (The International HapMap 3 Consortium 2010; The 1000 Genomes Project Consortium 2015). Software are available to perform genotype imputation, including BEAGLE, fastPHASE, IMPUTE2, MaCH+minimac and PLINK (Scheet & Stephens 2006; Purcell et al. 2007; Browning & Browning 2009; Howie, Donnelly & Marchini 2009; Li et al. 2010). BEAGLE and PLINK only use information from neighbouring genotypes when inferring missing genotypes making them more computationally efficient, however fastPHASE, IMPUTE2 and MaCH although more computationally intensive, are more accurate by using information across all markers for inferring missing genotypes, a method which is particularly helpful for rare variants (Porcu et al. 2013). Model-based imputation algorithms phase each individual, assigning haplotypes from the reference panel as a mosaic across the sample, followed by imputing missing genotypes (Marchini & Howie 2010). These two stages can be performed together, however in order to reduce the computational burden, pre-phasing was introduced where haplotypes are first independently estimated followed separately by imputing missing genotypes, and this has been implemented in the MaCh and IMPUTE2 frameworks (Howie et al. 2012). Genotypes are typically imputed with uncertainty, which can be taken into account in

downstream analyses using ‘dosage compensation’ models or by filtering out SNPs with imputation confidence (Zeggini et al. 2008; Marchini & Howie 2010).

1.1.5 Confounding by population stratification

Perhaps the major focus of methodological development in the field of human GWAS has been the avoidance of spurious associations caused by population stratification (see e.g. Price et al. 2010). Spurious associations can occur due to population stratification for several reasons (Voight & Pritchard 2005; Balding 2006; Weir, Anderson & Hepler 2006):

- When the trait differs systematically in a particular genetic subgroup, any variant predictive of the subgroup is likely to be associated with the phenotype. Although some of these variants may be truly causal, such associations are likely to contain many false positives.
- When uncontrolled confounding environmental variables contribute to population stratification of phenotypic variability, entirely artefactual associations may be detected at population-stratified genetic variants.
- When systematic differences in sampling propensity are population stratified, they can manifest as artefactual stratification of phenotypic variability, and may cause entirely spurious associations at population-stratified genetic variants.

Close relatives are usually downsampled because they exacerbate these problems, and cryptic relatedness can therefore be problematic (Voight & Pritchard 2005; Balding 2006). Methods for controlling for population stratification aim to avoid all these forms of spurious associations, and therefore constitute an important aspect of the practice of human GWAS. An overview of these methods follows.

1.1.5.1 Genomic control

Genomic control is a widely-used method to detect the presence of confounding due to population stratification, as per the QQ plot, and then correct for it by diminishing the association statistics by a uniform inflation factor (Devlin & Roeder 1999; Pritchard & Rosenberg 1999; Reich & Goldstein 2001). In the case of subpopulation differences resulting from recent genetic drift, applying a correction using genomic control can sufficiently correct for population stratification (Price et al. 2009). However, more generally it does not provide optimal power to detect associations because markers with stronger or weaker than average stratification receive inadequate or excess correction respectively (Price et al. 2006). Further, the rankings of associations are unchanged by genomic control.

1.1.5.2 Genetic ancestry

More ambitious methods aim to characterise population stratification in terms of the genetic ancestry of sampled individuals and avoid spurious associations by controlling statistically for systematic differences in phenotype between populations. One such method is the structured association, where samples are assigned to "unstructured" subpopulation clusters using software such as STRUCTURE (Pritchard, Stephens & Donnelly 2000; Rosenberg et al. 2002) or ADMIXTURE (Alexander, Novembre & Lange 2009), and markers are tested for association under the assumption that associations within subpopulations cannot be due to structure (Pritchard et al. 2000). While convenient, clearly this approach has limited generality.

Principal components analysis (PCA) has been widely used to characterise genetic ancestry. Including leading principal components, identified by software such as EIGENSTRAT (Price et al. 2006), as covariates in GWAS can correct for population stratification (Zhu et al. 2002; Price et al. 2006; Price et al. 2010). PCA does not

categorise individuals into discrete populations, rather it defines their position along a set of continual axes (Patterson, Price & Reich 2006). This can also be applied using software such as PLINK which uses an equivalent approach based on multidimensional scaling (MDS) (Purcell et al. 2007). The main difficultly with the application of PCA however is deciding on the number of PCs to include, and since family relatedness and the effects of other causal loci might only be captured by low-order PCs, the population stratification control may not be sufficient.

1.1.5.3 Mixed models

The above methods inferring genetic ancestry have limitations, they do not generally capture family structure or cryptic relatedness, which could result in the inflation of test statistics (Price et al. 2010). There has therefore been a shift to using mixed models to correct for population stratification, as they can model population structure, family structure and cryptic relatedness (Yu et al. 2006; Visscher, Hill & Wray 2008). An additional benefit of LMMs is that they can increase power in the presence of multiple causal loci (Yang et al. 2014). The mixed model approach models phenotypes as a mixture of fixed effects and random effects, where fixed effects include the candidate marker plus optional covariates and random effects represent the degree that genetically similar individuals are phenotypically similar by a genetic relationship matrix (GRM), also known as a 'kinship' matrix (Price et al. 2010). Initial exact implementations for GWAS such as EMMA (Kang et al. 2008) were very computationally demanding. Approximations such as EMMAX (Kang et al. 2010) and TASSEL (Zhang et al. 2010) made the application of linear mixed models to GWAS computationally feasible and were shown to outperform both PCA and genomic control in correcting for sample stratification, assessed by genomic control parameters (Kang et al. 2010). Further implementations using approximations such as GRAMMAR-Gamma have since been

developed (Svishcheva et al. 2012). However, approximations may result in a loss of power, dependent on the level of population structure within the data. Zhou & Stephens (2012) found that the approximations used in EMMAX and GRAMMAR led to a loss in power in a mouse dataset, but only made a small difference in a human dataset with little population stratification. Exact, computationally efficient methods are now available however in FaST-LMM (Lippert et al. 2011), GEMMA (Zhou & Stephens 2012) and GCTA (Yang et al. 2011). These advances mean that LMMs have become the favoured approach to controlling for population stratification in human GWAS.

LMMs are not without their limitations. It has been shown that including the candidate marker in the GRM can result in a loss of power due to fitting the candidate marker as both a fixed effect and a random effect (Lippert et al. 2011; Listgarten et al. 2012; Yang et al. 2014). The FaST-LMM software provides an implementation that enables the candidate marker to be excluded from the GRM when testing it for association (Lippert et al. 2011; Listgarten et al. 2012), and GCTA also has an implementation where all SNPs on the chromosome of the candidate SNP are excluded from the calculation of the GRM (Yang et al. 2011).

Rare variants pose a particular challenge when correcting for population stratification. Mathieson & McVean (2012) showed that confounding due to structured populations resulted in inflated test statistics due to rare variants using both PCA and LMMs. Listgarten, Lippert & Heckerman (2013) discuss how FaST-LMM-Select can address the problem of overinflation, however this will still be underpowered in the case of detecting rare variants that are spatially clustered (Mathieson & McVean 2013), therefore further method development and different study designs are required in this area. Nevertheless, these problems are not specific to LMMs, whose use is generally preferred over other methods for controlling for population stratification.

1.1.6 Fine mapping and prioritising variants for follow up

In many important respects, a GWAS is only one step in the process of discovering novel genetic variants underlying traits of interest. Following a GWAS study there are further steps to prioritise variants within highlighted associated regions. The first of these is fine-mapping, which is the attempt to narrow down the signal to the causal variant(s). Although the patterns of local LD structure are useful in designing GWAS, because causal variants do not have to be directly genotyped in order to discover new associations, it makes the identification of the true causal variant a major challenge due to the mapping resolution. Fine-mapping requires denser characterisation of variants through additional genotyping, imputation or sequencing and possibly by combining studies, preferably from different populations (Zaitlen et al. 2010). The second entails statistical prioritisation of variants based on their *P* value (or Bayesian posterior probability), LD with the most significant variant, and functional consequence of variants (Pickrell 2014; Kichaev et al. 2014), which can reduce the number of variants to follow up using functional assays.

1.1.7 What has been learnt from human GWAS studies

Much can be learnt from the human GWAS field in applying association studies to bacteria, from the strict criteria required to call an association, method development in controlling for population stratification, estimating the genetic contribution to common diseases, to efforts in understanding the biological relevance of the many associations that have been discovered. Lessons learnt from human GWAS are an indispensable starting point, however important differences between human and bacterial populations mean that not all developments in the human setting can be applied without adjustment. Below I summarise the major methodological developments in human GWAS that we can learn from when applying GWAS to bacteria. Differences between human and bacterial genomes are also discussed in further detail in Section 1.2.4.

- The LD structure of human genomes resulting in redundancies among local SNPs is crucial to the design and analysis of human GWAS. Correlations between SNPs are exploited and just a subset are tested as the effect of the correlated variants are as a result indirectly tested. However unlike in human GWAS, bacterial data comes from whole genome sequencing, where variants can also be tri- and tetra-alellic as well as biallelic, and bacterial genomes can have large accessory genomes which will require novel analytical tools. The clonal reproduction and facultative recombination of bacteria also means that LD may be stronger in many regions in bacteria and more variants will be population stratified.
- The importance of stringent quality control filtering of genotypes in human GWAS has been well established. Although bacterial genomes can be sequenced directly, Illumina sequencing which is typically used is error-prone, and criteria for variant inclusion in bacterial GWAS will need to be established. The visualisation of QQ plots can be directly applied to bacteria to indicate poorly calibrated test statistics and uncontrolled population stratification.
- In order to limit the number of false signals, the human GWAS community and journal editors have agreed on a stringent significance threshold to correct for multiple testing for classical *P* value thresholds. However, this is calculated based on the effective number of LD blocks in the human genome, and therefore its derivation depends on detailed understanding of the LD structure of the human genome. As bacterial genomes do not typically exhibit the same block like structure of LD as in humans, and can also contain large accessory genomes, deriving appropriate corrections for multiple testing will be challenging.
- The importance of the replication of results in separate validation studies has been widely observed in human GWAS. This is a quality which will also be important in

bacterial studies, particularly as highly structured bacterial populations may lead to high rates of false positives. As bacterial whole genome sequencing (WGS) datasets increase in size, large validation studies and meta-analyses will become easier to apply, and bacterial GWAS results also have the possibility of *in vitro* validation.

- Imputation has been demonstrated to be important in human GWAS because it can increase the precision of an association study and can also be used to combine datasets for meta-analyses. Whole genome sequencing of bacteria means that imputation will not be required for the same reasons, but it will be important to impute SNPs which are filtered out by strict QC filtering in order to maintain the highest possible resolution. However, bacterial populations will not need phasing, which is a necessary first step in human imputation, so novel approaches for imputation may be required.
- Possibly the dominant methodological focus in human GWAS has been the avoidance of spurious associations due to population stratification. This will be important in bacterial studies due to the high number of population stratified variants. The methodological development in the human setting can be taken as a starting point for bacterial studies, where LMMs have become the favoured approach for controlling for population stratification due to their ability to model population structure, family structure and cryptic relatedness. It has also been demonstrated that inclusion of the candidate marker when calculating the kinship matrix can result in a loss of power, however solutions such as removing all variants on the same chromosome as the candidate marker when calculating the kinship matrix will not be applicable in bacteria because bacterial genomes typically reside on a single chromosome.
- Finally, a GWAS is just the first step in discovering novel variants, and the human GWAS field has developed approaches for interpreting significant associations.

Although the block-like LD structure of the human genome is crucial in human GWAS study design, it also makes the identification of true causal variants difficult due to the mapping resolution. It does however enable identification of the genomic region of interest, whereas in bacteria causal variants may be in high LD with variants genome-wide. Bacterial studies will therefore need to learn from human approaches which take into account the functional consequence and biological plausibility of variants in able to prioritise associations for follow-up.

1.2 Bacterial genome-wide association studies

The quest to identify the genetic basis of important traits in bacteria is far from new. Such studies are as old as molecular microbiology as a discipline, for example classical studies demonstrated that microorganisms cause disease (e.g. reviewed in Isenberg 1988), and now we want to understand which genes or variants cause disease, if they do. Since the 1980s, the standards for linking genotype to phenotypes related to virulence (the quantifiable frequency or severity of disease) have been formalised through Falkow's molecular Koch's postulates (Falkow 1988). The first concerns the discovery of associations between phenotypes and virulent strains, and the second and third concern genetic manipulations of bacteria in the laboratory to demonstrate that a gene or marker confers a phenotypic change. In requiring genetic manipulation of the bacteria, this is a stringent set of criteria aimed narrowly at virulence components that are manipulable in the laboratory. Nevertheless, of the three postulates, it is the first which has been neglected and the community has focused on the second and third postulates utilising laboratory manipulations, meaning that natural variation has largely been overlooked until more recently. The two approaches for understanding the molecular basis of traits of interest, molecular and population based, have previously been described as bottom-up and top-down based on whether the starting point is the DNA or the phenotype

	Pros	Cons
Molecular approaches	- Can demonstrate direct links between genotype and phenotype - Can be used to dissect the consequence of mutations identified in population studies	- Most phenotypes cannot be replicated in a laboratory setting - Cannot replicate exact conditions in the laboratory, and may not understand natural conditions - Labour intensive: not feasible to establish a the phenotypic effect of all genetic variants
Population studies	- Studies natural variation, not laboratory constructs - Investigate natural phenotypes - Concerns mean phenotype across many genetic backgrounds and environments - Testing need not be biased by prior hypotheses of what genes/variants may be associated	- Need validation to demonstrate a direct link between genotype and phenotype - Population stratification and other confounders can create spurious associations - Cannot disentangle the effects of strongly linked variants

Table 1.1 Molecular versus population based approaches for linking genotype and phenotype

(Falush & Bowden 2006). The pros and cons of molecular versus population approaches are discussed in Table 1.1, but to fully characterise the bacterial genetic contribution to important phenotypes, both approaches will be important.

There are two ways of viewing the benefits of the study of natural variation using GWAS which are not mutually exclusive. The first is that GWAS is a way of prioritising variants for molecular follow up. In this case GWAS is simply used as a way of narrowing down the full complement of variation to a more manageable list of variants most likely to contribute to the trait of interest, which can then be analysed by molecular work. The second view is that GWAS is a way to ask questions that cannot be asked with molecular work. For example, GWAS allows the investigation of phenotypes which cannot be replicated within the laboratory across many genetic backgrounds on a large scale. It also enables an unbiased, hypothesis free approach unlike candidate gene based analyses.

1.2.1 What traits can GWAS exploit?

GWAS is of particular interest for traits which are difficult to study in the laboratory, but as explained above the investigation of natural variation and the ability to test a phenotype across a wide range of genetic backgrounds and environmental conditions makes it an attractive tool for all phenotypes. Examples of phenotypes which will be especially difficult to study in the laboratory include adaptation to particular hosts and the ability to cause invasive infections. However, even phenotypes which are easily amenable in the laboratory, such as antimicrobial resistance, have the potential to benefit from GWAS.

1.2.1.1 Antimicrobial resistance

The bacterial genetic basis underlying antimicrobial resistance for many drugs and species is well understood, as mirrored by the high predictive accuracy of phenotypic susceptibility using genetic data (Zankari et al. 2013; Stoesser et al. 2013; Gordon et al. 2014; Walker et al. 2015; Bradley et al. 2015). Discovering the mechanisms conferring resistance in the remaining unexplained cases presents a greater challenge however as the underlying variants are likely to be rarer, requiring large population samples to establish sufficient statistical power. For example, a small number of mutations cause the majority of resistance to most first and second line drugs in *Mycobacterium tuberculosis*, but a study using known resistance-conferring mutations to predict drug susceptibility and resistance could not make predictions for 10.8% of phenotypes due to the presence of uncharacterised mutations (Walker et al. 2015). Prediction accuracy varied substantially by drug and was particularly poor for pyrazinamide resistance, where mutations characterised as resistance-conferring accounted for just 25% of resistant isolates (Walker et al. 2015). GWAS applied on a large scale has the potential to assist in increasing antimicrobial resistance catalogues and improve resistance prediction.

1.2.1.2 Virulence

Virulence is defined here as the quantifiable frequency or severity of disease. We define virulence factors as genetic variants that increase the frequency or expected severity of disease. Virulence factors have typically been identified by molecular work, however genotypes shown to confer important phenotypes in the laboratory may not be variable in natural populations, and also may not be naturally occurring. Instead, virulence factors identified in the GWAS setting will discover variants which explain the observable variability in disease frequency or severity in natural populations, which is distinct to the definition of virulence from molecular work. By the definition of virulence in this thesis, a virulence factor need not always cause disease. For example, the importance of the *Neisseria meningitidis* capsule in evading complement-mediated and phagocytic killing has been demonstrated (Jarvis & Vedros 1991; Spinosa et al. 2007) and unencapsulated meningococci only rarely cause invasive disease, typically in immunocompromised patients (Vogel et al. 2004; Hoang et al. 2005; Ganesh et al. 2017). However, the meningococcal capsule can also be found in isolates sampled from human carriers (Claus et al. 2005; Jolley et al. 2000; Yazdankhah et al. 2004). This can make the identification of virulence factors difficult, as their presence may not be split trivially into virulent and avirulent strains. Virulence also might not be a result of the acquisition or gain of particular genes or variants, but due to the loss of genes or genomic regions (Maurelli 2007).

For any practical study, it is necessary to choose what aspect of disease or disease severity to quantify. This choice will bias the measure of virulence and restrict the analysis to particular components of the pathogen-host interaction (Diard & Hardt 2017). Studies investigating the bacterial genetic basis of virulence have taken different approaches to defining the phenotype. Some have investigated the occurrence of invasive

disease (Holt et al. 2015; Lees et al. 2016) whereas others have focused on specific aspects of virulence such as toxicity and biofilm formation (Laabei et al. 2014; Recker et al. 2017).

1.2.1.3 Host adaptation

Understanding how bacterial pathogens adapt to different hosts can be difficult to establish as many hosts will not be experimentally tractable. Comparative genomics however has enabled insights into the evolutionary processes underlying host adaptation. For example, Langridge et al. (2015) investigated the genetic basis for *Salmonella enterica* subspecies *enterica* host adaptation within a lineage exhibiting a range of host specialisations, describing patterns of psuedogene formation across differently host-adapted strains. A study of host tropism in *Staphylococcus aureus* by Viana et al. (2015) reported that a single mutation was sufficient for a human-specific *S. aureus* strain to become infective for rabbits, however this study was able to exploit the fact that rabbits are an experimentally tractable host species. Genome-wide association studies will enable these studies to be applied on a larger scale for multiple host species, to assist in understanding the genetic basis of host adaptation, an area which has already begun with two studies investigating host associations in *Campylobacter jejuni* (Sheppard et al. 2013; Yahara et al. 2017).

1.2.2 Early studies investigating genotype-phenotype associations in bacteria

Like in humans, observational studies into the genetic basis of natural variation in bacterial traits of interest began with candidate genes. Early studies used PCR, low resolution sequence data such as MLST, or comparative genome hybridisation microarray analyses in order to understand the genetic basis of various phenotypes (Maiden et al. 1998; Zhou 2003). For example, Peacock et al. (2002) investigated virulence in *Staphylococcus aureus* by comparing the presence of 33 genes across isolates from

healthy blood donors and patients with invasive disease. Gene presence was determined by either PCR or by phenotypic tests and the effect of population structure was corrected using MLST. Seven of the investigated genes were found to be significantly more common in invasive isolates, with their effects appearing to be cumulative (Peacock et al. 2002). However, a later microarray analysis of a subset of these isolates was undertaken, which had greater sensitivity to sequence variation, and did not find any gene to be associated with invasion (Lindsay et al. 2006). It is not clear if the difference is due to sequence variation in the PCR binding site or due to problems with the reproducability of microarray analyses (Lindsay et al. 2006; Draghici et al. 2006).

Bille et al. (2005) investigated virulence in *N. meningitidis* within a panel of strains assembled using MLST data, comparing the genetic complement of hypervirulent versus non-hypervirulent strains using a microarray. Bille et al. (2005) found that a prophage was statistically overrepresented in the hypervirulent isolates, although this result was not replicated in a further study (Dunning Hotopp et al. 2006) and it is possible that the association represents a marker for particular hypervirulent lineages in young adults rather than virulence more generally (Bille et al. 2008).

1.2.3 Whole genome sequencing

The introduction of high-throughput whole genome sequencing brought in a dramatic change in the way bacterial populations could be studied. Whole genome sequencing has revealed genome dynamics at the level of within-host (Didelot et al. 2016), transmission (Bentley & Parkhill 2015) and antimicrobial resistance (Bradley et al. 2015), and now in understanding the link between genotype and phenotype using GWAS (Power, Parkhill & de Oliveira 2017).

1.2.4 Challenges for bacterial GWAS

Bacteria differ genetically from humans in many important ways, meaning that tools developed for human GWAS cannot necessarily be directly applied to bacteria. Firstly, bacteria are haploid and the bacterial genome is typically on a single chromosome, as opposed to the human genome which is organised over 23 chromosome pairs (International Human Genome Sequencing Consortium 2001). Reproduction is very different in humans and bacteria: in humans, sexual reproduction involving meiosis results in genetic recombination and chromosome segregation every generation, unlike asexual bacterial reproduction where recombination is facultative. The recombination rate varies over the human genome, with the majority of recombination occurring in recombination hotspots (McVean et al. 2004) leading to the linkage patterns and block like structure seen in the human genome (Gabriel et al. 2002; McVean 2002; The International HapMap Consortium 2005).

Genetic exchange among and within bacterial populations is driven by three main mechanisms: transformation of DNA, transduction by bacteriophages, and plasmid-mediated conjugation (Thomas & Nielsen 2005). Bacterial species differ in whether they are naturally competent for DNA uptake (Christie & Dubnau 2005). Major differences arise in the LD landscape between human genomes, shaped by crossing over and independent chromosome segregation, and bacterial genomes, shaped by transformation, transduction and conjugation. In particular, over the 50 kb scale, LD decays to residual levels near zero in humans. In bacteria, LD also decays but strong population stratification mean the residual levels are far from zero. This manifests as genome-wide LD (Figure 1.1). This substantial residual LD in bacterial genomes reflects highly structured populations.

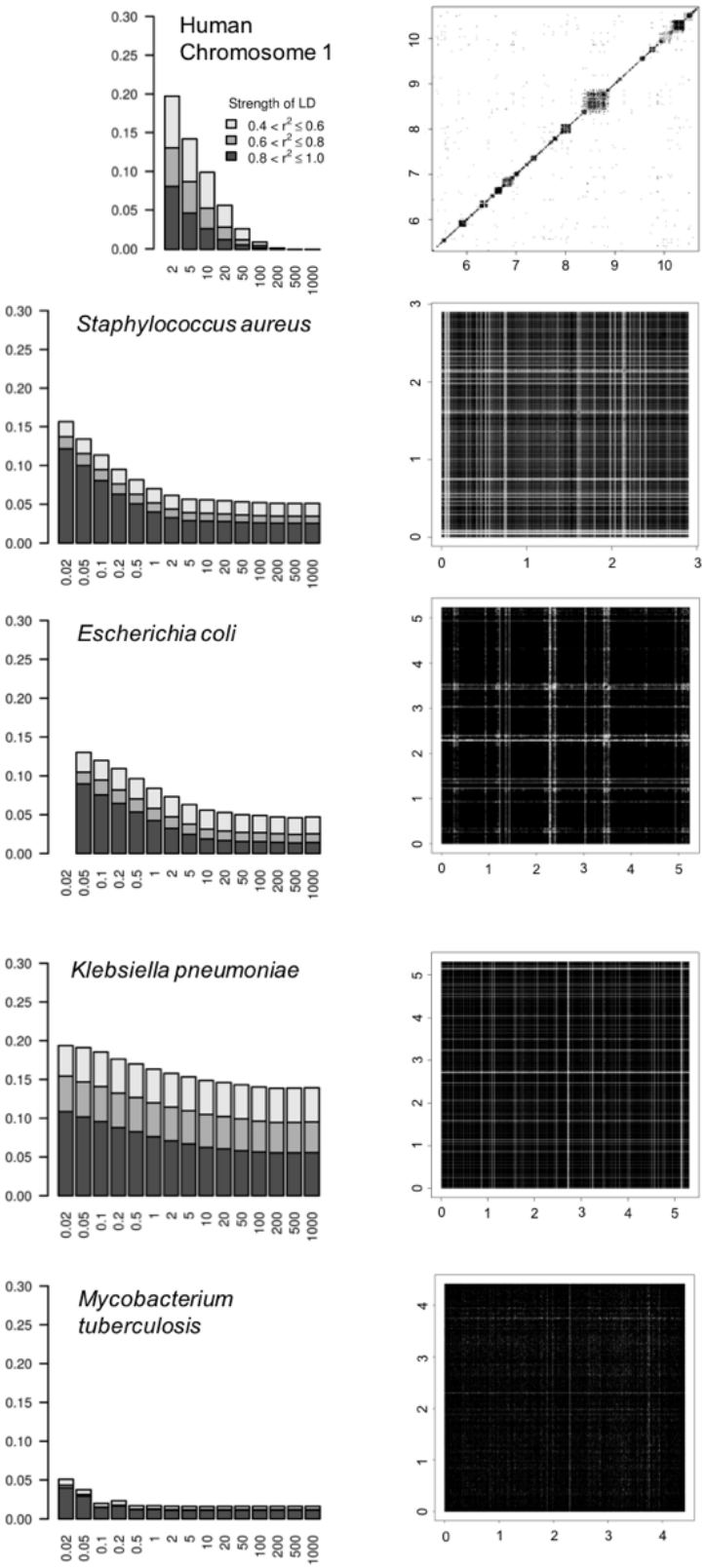

Figure 1.1 Linkage disequilibrium and population structure in bacteria vs humans is strikingly different. Left panels: Proportion of SNPs at different strengths of LD with increasing distance on the chromosome. LD (r^2) decays rapidly with physical distance (kb) in bacteria, as in humans, however it plateaus to non-zero residual levels resulting in genome-wide LD. Right panels: Points show the position of sites (Mb) in high LD, with an r^2 of greater than 0.7. The bacterial genomes comprise one large LD block in comparison to the smaller block like structure of LD in humans due to strong population structure and limited homologous recombination. The human data presented represents a 4Mb region of chromosome 1 from HapMap data.

Non-homologous genetic exchange in bacteria represents a major form of variation which will require purpose-built tools for bacterial GWAS to characterise and exploit this variation. Non-homologous genetic exchange results in extensive accessory

genomes for many bacterial species, which led to the concept of a "pan-genome", all genes present at least once within a species (Lapierre & Gogarten 2009). Previous analyses have reported species with seemingly open pan-genomes, meaning that the rate at which previously unobserved genes increases as a function of sample size does not appear to plateau, even in very large samples, such as in *Streptococcus pneumoniae* (Donati et al. 2010) and *Klebsiella pneumoniae* (Holt et al. 2015). It is crucial that the accessory genome is captured in association studies in order to fully understand the bacterial genetic basis of traits of interest. This means that SNP based studies relying on mapping to a relatively highly conserved reference genome only capture core variation, and new methods to capture the accessory genome, not typically employed in human GWAS, plus insertions and deletions, are required.

Bacteria also import genetic material in place of existing homologous material by homologous recombination, and rates vary across bacterial species (Vos & Didelot 2009) and also across a single genome (Everitt et al. 2014). Import lengths estimated for homologous recombination are typically short, for example a study found an average import length of 591bp and 183bp in *Clostridium difficile* ST6 and *Staphylococcus aureus* species wide diversity, respectively (Didelot & Wilson 2015). However, this may reflect the dominant role of transformation and transduction in bacterial homologous recombination. Large importations can occur, particularly by conjugation, which in some cases have generated hybrid strains in what have been termed chromosomal replacements (Robinson & Enright 2004; Brochet et al. 2008; Chen et al. 2014; Wyres et al. 2015).

As discussed in light of the human population in Section 1.1.5, population stratification is an important potential confounder in GWAS and is expected to present a major challenge in applying GWAS to bacteria. The extent of this effect depends on the species, as bacteria can range from entirely clonal (e.g. *Mycobacterium tuberculosis,*

Achtman 2008) to highly recombining (e.g. *Helicobacter pylori*, Vos & Didelot 2009). Geographically-restricted clonal expansions can also occur in response to selection pressures on the population (Keenan et al. 2015), which can lead to population stratified differences in sampling and therefore confounding. Applying GWAS to bacteria will therefore require stringent controls for population structure to avoid false positive associations which are only predictive of ancestry, sampling, or the environment.

Despite these challenges, the application of GWAS to bacteria does have some advantages. As the genomes are much smaller, the whole genome can be sequenced using next-generation technologies such as Illumina, which have rapidly reduced the cost and increased the throughput of bacterial genome sequencing, rather than requiring SNP chips to represent genome variation. The second advantage is that bacterial genomes are able to be manipulated and have short generation times, therefore for phenotypes which can be tested in a laboratory setting, genetic disruption and reconstruction can be used to validate GWAS findings (Laabei & Massey 2016). Combining GWAS studies with functional genomics has enormous potential to promote the discovery of links between genotype and phenotype.

1.2.5 The application of GWAS to bacteria

A new wave of bacterial GWAS began in 2013, taking lessons from the methodological development in human genetics, beginning with the paper by Sheppard et al. (2013). The study applied a "kmer" approach, where 30bp DNA "words" or "kmers" were counted in a sliding window for each assembled genome. The presence or absence of each variably present kmer was then tested for association with the phenotype, host association in *Campylobacter jejuni*. Representing variation using kmers captures SNPs, gene presence and absence and short insertions and deletions, which enabled Sheppard and colleagues to find multiple variants determining vitamin B_5 synthesis to be associated with host

adaptation. Multiple studies have since applied and developed methods for GWAS in bacteria, as summarised in Section 1.2.6. Some studies have focused on SNP variation (Farhat et al. 2013; Alam et al. 2014; Laabei et al. 2014; Chewapreecha et al. 2014; Farhat et al. 2014; Hall 2014; Chen & Shapiro 2015; Recker et al. 2017; Lees et al. 2017; Collins & Didelot 2017) whereas others have attempted to capture greater variation by using various forms of a kmer approach (Sheppard et al. 2013; Pascoe et al. 2015; Lees et al. 2016; Yahara et al. 2017) and pangenome analyses (Salipante et al. 2015; Holt et al. 2015). The phenotypes investigated have largely been traits under strong selection, such as antimicrobial resistance (Farhat et al. 2013; Alam et al. 2014; Chewapreecha et al. 2014; Salipante et al. 2015; Farhat et al. 2014; Chen & Shapiro 2015; Desjardins et al. 2016; Lees et al. 2016; Collins & Didelot 2017), predominantly as proof of principle, but also aiming to discover novel resistance-conferring mechanisms. Other studies have also begun to investigate host association/host source (Sheppard et al. 2013; Hall 2014; Farhat et al. 2014; Yahara et al. 2017), pathogenicity (Hall 2014) and virulence (Laabei et al. 2014; Holt et al. 2015; Lees et al. 2016; Recker et al. 2017; Lees et al. 2017; Collins & Didelot 2017).

1.2.6 Analytical approaches used in the new wave of bacterial GWAS

In addition to the investigation of multiple phenotypes by bacterial GWAS, multiple analytical approaches have also been applied. Here I summarise the introductions of the analytical approaches applied to bacterial GWAS which, with the exception of the influential study by Sheppard et al. (2013), have all appeared since I begun this thesis, demonstrating the rapid developments of the field. These analytical approaches broadly fall under the following headings: no correction for population stratification, phylogenetic approaches, genetic ancestry approaches, and sample covariance approaches.

1.2.6.1 No correction for population stratification

Hall (2014)

Variation characterised: SNPs identified from *de novo* assemblies or raw sequencing reads without requiring an alignment.

Variant exclusion criteria: None.

Control of population structure: None.

Formal test: χ^2 test.

Multiple testing correction threshold: None.

Benefits: Investigated SNP variation without requiring a genome alignment.

Limitations: χ^2 test assumes independence between observations, whereas the probability of each observation will likely be dependent on population structure. Cannot be applied to datasets with hundreds of genomes and hundreds of thousands of SNPs without subsampling the data.

Organisms/strains: 68 *E. coli* and *Shigella* strains and 116 *E. coli* strains.

Phenotypes: Binary – pathogenic vs commensal (68 *E. coli* and *Shigella*) and human vs non-human host source (116 *E. coli*).

Notable hits: Identified variants highly predictive of both pathogenicity, with an accuracy of 94.4% and host source, with an accuracy of 89.7%.

Phenotype prediction: The data was split into two groups, and the phenotype was modified to "unknown" for one group. SNPs were ordered by their χ^2 probability and were added until a minimum of 50, plus either the positive predictive value (PPV) reached a user-defined maximum or the PPV declined below a user-defined fraction of the maximum PPV identified.

Additional ranking criteria: SNPs identified as diagnostic by the phenotype prediction were ranked by their χ^2 probability that the SNPs changed randomly across the internal

branches where the phenotype changed state, to identify SNPs most likely to be causal. To determine the branches where SNP and phenotype state changes occurred ancestral state reconstruction was performed.

Holt et al. (2015)

Variation characterised: Gene presence vs absence.

Variant exclusion criteria: None.

Control of population structure: None.

Formal test: Fisher's exact test.

Multiple testing correction threshold: Benjamini-Hochberg correction to control the FDR (Benjamini & Hochberg 1995).

Benefits: Gene presence vs absence was tested for the pangenome, therefore capturing the accessory genome. The proportion of the phenotype which could be explained by the strongest associations was characterised.

Limitations: Fisher's exact test assumes independence between observations, whereas the probability of each observation will likely be dependent on population structure.

Organisms/strains: 328 *Klebsiella pneumoniae* strains.

Phenotypes: Binary – virulence (defined by infection vs carriage and invasive vs non-invasive).

Notable hits: *rmpA/2* and siderophore genes and five additional predicted iron-metabolism genes which are present on the virulence plasmid pK2044.

Baines et al. (2015)

Variation characterised: Core genome non-synonymous SNPs, plus SNP groupings by assigning an affected or unaffected status to each gene based on whether the gene

contained any SNPs.

Variant exclusion criteria: None.

Control of population structure: None.

Formal test: Used PLINK which can perform various tests such as χ^2 test, Fisher's exact test, Cochran-Mantel-Haenszel test and linear and logistic models.

Multiple testing correction threshold: Bonferroni correction on the number of SNPs or genes tested to control the FWER.

Benefits: Pooling SNPs by gene may increase power to detect variants underlying polygenic traits where rare variants affect the same phenotype.

Limitations: No correction for population structure.

Organisms/strains: 123 ST-239 *S. aureus* isolates.

Phenotypes: Binary – vancomycin MICs of ≤ 2 and ≥ 3 μg/ml.

Notable hits: A SNP resulting in the amino acid substitution $RpoB_{H481Y}$ when testing SNPs individually, and the genes *walK* and *walR* when pooling SNPs by gene.

Recker et al. (2017)

Variation characterised: SNPs. Tri- and tetra-allelic variants were collapsed into biallelic by assigning 1 if the allele differed to the reference and 0 if it did not. Annotated intergenic elements such as miscellaneous RNAs were considered separate loci.

Variant exclusion criteria: Synonymous SNPs, plus SNPs in known mobile genetic elements and repeat regions. MAF ≤ 5%.

Control of population structure: None.

Formal test: Analysis of variance model (ANOVA).

Multiple testing correction threshold: Associations tested at both uncorrected ($P < 0.05$) and Bonferroni-corrected ($P < 4.6 \times 10^{-5}$, controls the FWER).

Benefits: Functional validation of a subset of the significant associations.

Limitations: High reported false positive rates of 70% for biofilm formation and 95% for toxicity.

Organisms/strains: 135 clonal complex (CC) 22 and 165 CC30 *S. aureus* strains.

Phenotypes: Continuous – biofilm formation and cytolytic activity (assessed as the ability to lyse a monocyte cell line, THP-1).

Notable hits: Functionally verified the effect of variants in five biofilm-associated loci and five toxicity-associated loci using transposon mutants. Verified biofilm-associated loci included the gene NE513 encoding a helicase, NorA (NE1034) a quinolone efflux pump, and NE1455 a peptidase in biofilm formation.

Phenotype prediction: A Random Forests machine learning approach identified the most important determinants for host mortality based on genotype, phenotype and clinical metadata in both CC22 and CC30.

1.2.6.2 Phylogenetic approaches

Sheppard et al. (2013)

Variation characterised: Presence or absence of 20bp, 30bp and 40bp kmers counted from Velvet *de novo* assemblies.

Variant exclusion criteria: None.

Control of population structure: Simulated kmer gain and loss along an estimated phylogenetic tree. Kmers were simulated under the null hypothesis of no association, and a null distribution of association statistics was produced. The correlation of the real kmers was compared with the distribution of the correlations of simulated kmers, producing a phylogenetically corrected *P* value.

Formal test: As above.

Multiple testing correction threshold: $P < 10^{-4}$.

Benefits: Analysing kmers captured variation in both the core and accessory genome, and results were robust to the three kmer lengths tested. Limiting the analysis to ST-45 and ST-283 focused on a lineage with a high frequency of phenotype switching and reduced the level of population stratification.

Limitations: It is not clear how the reliance on a phylogenetic tree will impact species with high recombination rates. The method also shares the limitation of genomic control in that it is only concerned with setting the significance threshold, and does not alter the rankings of association signals.

Organisms/strains: 29 ST-45 clonal complex *C. jejuni* plus two ST-283 complex isolates. The association signal was also tested among 161 *C. jejuni* and *C. coli* isolates outside of the ST-45 complex to validate the results.

Phenotypes: Binary – cattle vs chicken colonisation.

Notable hits: A seven-gene region, of which three encoded vitamin B_5 biosynthesis genes. Isolates from cattle were demonstrated to be better able to grow in vitamin B_5-depleted media.

Farhat et al. (2013)

Variation characterised: SNPs.

Variant exclusion criteria: SNPs within 5bp of an insertion or deletion and SNPs that did not have an adjacent consensus quality score of 20.

Control of population structure: Phylogenetic convergence test for selection (PhyC). Phylogenies were constructed from the SNPs using three methods, Bayesian, parsimony and maximum likelihood, excluding SNPs in repetitive elements and known drug resistance-associated genes. Ancestral state reconstruction was performed for the SNPs

and for the phenotype. Ambiguously reconstructed states were removed. The number of convergent SNPs (changes to the same base) in resistant and sensitive branches was counted for each position and gene in the genome, and compared to the empirical background distribution of convergence. This phylogeny-dependent permutation approach identified sites and genes which mutated more frequently in resistant branches than sensitive branches of the phylogeny than would be expected by chance.

Formal test: PhyC, as above.

Multiple testing correction threshold: A SNP was required to have $P < 0.05$ across all phylogenetic and ancestral reconstruction methods.

Benefits: Identified signatures of positive selection specific to antimicrobial-resistant strains.

Limitations: It is not clear how the reliance on a phylogenetic tree will impact species with high recombination rates.

Organisms/strains: 123 *M. tuberculosis* strains.

Phenotypes: Binary – antimicrobial resistance. Defined a 'broad resistance' phenotype as resistance to any drug tested plus specific resistance phenotypes to the five first-line tuberculosis drugs.

Notable hits: Detected all 11 known resistance determinants plus 39 novel associations. The functional impact of one of the novel mutations was assessed, which revealed that a mutation in the gene *ponA1* conferred a survival advantage in the presence of rifampicin over the wild-type strains.

Farhat et al. (2014)

Variation characterised: SNPs

Variant exclusion criteria: Repetitive, transposon, and phage-related regions.

Control of population structure: A matched sampling scheme. Closely related phylogenetic pairs with contrasting phenotypes were studied.

Formal test: The number of mutations, insertions and deletions between each strain pair was summed across each locus. This was then compared to a Poisson distribution of variants randomly distributed with respect to the branches of the phylogeny. Significantly associated genes were defined by genes with larger than expected counts under the null distribution.

Multiple testing correction threshold: Bonferroni-correction on the number of pairs × the number of variant loci to control the FWER.

Benefits: The paired sampling scheme minimises the impact of population stratification.

Limitations: Only designed for clonal to moderately recombining bacteria. Requiring closely matched pairs will come at the cost of sample size.

Organisms/strains: Eight *M. tuberculosis* pairs plus eight *C. jejuni* and *C. coli* pairs.

Phenotypes: Binary – antimicrobial resistance and host source.

Notable hits: Confirmed known resistance-conferring mechanisms in *Mycobacterium tuberculosis* and found associations with two of the genes previously identified within the vitamin B_5 biosynthesis region affecting host-association in *Campylobacter* by testing a subset of the original data from Sheppard et al. (2013).

Brynildsrud et al. (2016)

Variation characterised: Gene presence vs absence.

Variant exclusion criteria: Correlated variants were collapsed into a single variant. An optional filtering step using a Fisher's exact test removed variants where $P > 0.05$.

Control of population structure: A phylogenetic pairwise comparisons algorithm, aiming to identify the maximum number of phylogenetically non-intersecting pairs of isolates with contrasting genotypes and phenotypes, i.e. a terminal branch with phenotype

0 and wild type allele paired with a closely related terminal branch with phenotype 1 and alternative allele. Under the null hypothesis, the phenotype was randomly assigned with equal probability of being a case or control, irrespective of the genotype. Identifying the maximum number of contrasting pairs defined the number of independent emergences of a gene-trait combination. The method defined the maximum number of pairs which contrasted in both gene and phenotype states (the "best" pairings), and also the minimum number of pairs supporting it (the "worst" pairings). *P* values were calculated using a binomial test for the best and worst pairings, with the test statistic being the maximum number of contrasting pairs divided by the maximum number of pairs. Only pairs which contrasted in both phenotype and genotype were considered informative. A conservative interpretation would require both *P* values to be less than the pre-determined significance threshold. An additional permutation procedure was also implemented, which randomly permuted the phenotype and calculated the associated test statistic.

Formal test: Binomial test.

Multiple testing correction threshold: Either Bonferroni or Benjamini-Hochberg adjustments to control the FWER or FDR, respectively.

Benefits: Gene presence vs absence was tested for the pangenome, therefore capturing both the core and accessory genome.

Limitations: It is not clear how the reliance on a phylogenetic tree will impact species with high recombination rates.

Organisms/strains: 21 *Staphylococcus epidermidis* strains and 3,085 *S. pneumoniae* strains.

Phenotypes: Binary – linezolid resistance (*S. epidermidis*) and erythromycin resistance (*S. pneumoniae*).

Notable hits: Identified the resistance-conferring *cfr* gene as associated with linezolid

resistance in *S. epidermidis*, although the higher end of the *P* value range exceeded 0.5. Also identified two further plasmid-associated genes associated with linezolid resistance in *S. epidermidis*. Identified the resistance-conferring *erm* gene to be significantly associated with erythromycin resistance in *S. pneumoniae*.

1.2.6.3 Genetic ancestry approaches

1.2.6.3.1 Genetic clustering

Laabei et al. (2014)

Variation characterised: SNPs and indels.

Variant exclusion criteria: Missingness ≥ 90%, MAF ≤ 5%.

Control of population structure: Discrete groups and genomic control. Assessed whether significant SNPs were robust to a permutation procedure using PLINK where cluster membership, defined using a hierarchical clustering algorithm in R, was corrected for. A correction for genomic inflation was also applied.

Formal test: Used PLINK which can perform various tests such as χ^2 test, Fisher's exact test, Cochran-Mantel-Haenszel test and linear and logistic models.

Multiple testing correction threshold: None.

Benefits: Investigated a single strain which will reduce the level of population stratification and performed functional validation of significant variants.

Limitations: SNPs were not imputed. Population stratification within clusters will not be controlled for.

Organisms/strains: 90 *S. aureus* ST-239 isolates.

Phenotypes: Continuous (categorised into three non-overlapping groups for phenotype prediction) - toxicity, defined by the expression of alpha toxin, the ability to lyse T cells and the ability to lyse lipid vesicles.

Notable hits: Variants which affected toxicity *in vivo*: a SNP between the genes *tarK* and *tarF*, a SNP between a hypothetical gene and *fmt*, a SNP as intergenic between a TELA-like protein and a putative branched-chain amino acid transporter protein in the reference FPR3757, and finally a SNP annotated in FPR3757 as intergenic between a hypothetical gene and an AcrB/AcrD/AcrF family gene.

Phenotype prediction: Random forest machine learning. Using the associated SNPs all highly toxic and low toxic strains were correctly predicted, although the toxicity of strains with medium toxicity was underestimated.

Chewapreecha et al. (2014)

Variation characterised: SNPs.

Variant exclusion criteria: Missingness by strain >10%, missingness by variant >10%, MAF ≤1%. Just the two most common variants per site were analysed.

Control of population structure: Cluster correction and genomic control. Genetic clusters were calculated using the software BAPS (Corander, Waldmann & Sillanpää 2003; Corander et al. 2008) which were included as factors.

Formal test: Cochran-Mantel-Haenszel (CMH) test conditional on the population cluster.

Multiple testing correction threshold: Bonferroni correction on the number of variants at a *P* value of 0.01, to control the FWER.

Benefits: Applied to a species with a high level of recombination in the genes encoding the peptidoglycan biosynthesis pathway, the target of β-lactams, and tested a large sample size.

Limitations: Population stratification within BAPS clusters will not be corrected for, and the level of control will depend on the number of clusters chosen. SNPs were not imputed.

Organisms/strains: 3,085 *S. pneumonia* genomes followed by a replication study of 616 genomes.

Phenotypes: Binary – β-lactam resistance.

Notable hits: Confirmed known β-lactam resistance mechanisms and identified novel SNPs with the potential to contribute to β-lactam resistance, requiring functional validation to explore their potential contribution to β-lactam resistance.

Predictive power: The 301 SNPs identified in both studies when combined could explain close to 100% of resistance in each population.

1.2.6.3.2 Principal component analysis and multidimensional scaling

Salipante et al. (2015)

Variation characterised: Gene presence vs absence.

Variant exclusion criteria: Genes ≤ 75 amino acids in length, coding sequences relating to repetitive elements, transposons, insertion sequences, plasmid support machinery, resistance mechanisms for drugs other than those tested, unrelated biochemical pathways and coding sequences present in ≥15% of the susceptible isolates.

Control of population structure: PCA.

Formal test: Logistic regression including the top three PCs as covariates, reported to be the optimal number of PCs by visual inspection of QQ plots.

Multiple testing correction threshold: None - the 20 most significant genes were considered for each antimicrobial.

Benefits: PCA to correct for population stratification. Also performed a conditional analysis, where the presence of known mechanisms were included as covariates in order to discover novel variants.

Limitations: PCA does not control for cryptic relatedness. By removing so many genes, their importance in natural populations could not be tested.

Organisms/strains: 312 extraintestinal pathogenic *E. coli*.

Phenotypes: Binary – antimicrobial resistance.

Notable hits: Validated known resistance-conferring genes.

Lees et al. (2016)

Variation characterised: Presence or absence of variable length kmers, between nine and one hundred bases long. Redundancy was reduced by removing any kmer that was a substring of a longer kmer present in the same isolates. Kmers were counted from *de novo* assemblies.

Variant exclusion criteria: MAF ≤ 1%. $P > 10^{-5}$ by a χ^2 test (or Welch two-sample *t*-test in the case of continuous phenotypes). Post-association, kmers with negative effect sizes were removed.

Control of population structure: MDS. A distance matrix was constructed from a small random subsample of kmers (between 0.1-1% of those with MAF ≥ 5%). A pairwise distance matrix was then constructed, with each element representing an estimate of the number of non-shared kmers between a pair of samples. Clustering samples using this distance matrix produced the same clusters as using hierarchical BAPS clustering (Cheng et al. 2013). MDS was then applied to the distance matrix, which projects the distances into a reduced number of dimensions and the normalised eigenvectors of each dimension were used as covariates in a regression model to test for association. It was suggested that the number of dimensions to use can be evaluated by the goodness-of-fit and the magnitude of the eigenvalues.

Formal test: Logistic or linear regression, depending on the phenotype, adjusting for the eigenvectors as described above.

Multiple testing correction threshold: Bonferroni-correction at $P < 0.05$, based on the assumption of every position in the genome having three possible mutations, which resulted in a threshold of $P < 1 \times 10^{-8}$, intended to control the FWER.

Benefits: Kmer length was extended as long as possible up to 100bp, and power was shown to be higher when counting variable kmer lengths over a single kmer length. Variation in the core and accessory genome was explored without requiring alignments or SNP calling.

Limitations: Variants were filtered using a χ^2 test which assumes independence between observations, and filtering kmers with negative effect sizes could lead to missing important biological mechanisms. Using a limited number of MDS dimensions will not control for subtle population structure. Control of FWER not demonstrated.

Organisms/strains: *S. pneumoniae* and *S. pyogenes*.

Phenotypes: Binary – antimicrobial resistance (*S. pneumoniae*) and invasive vs non-invasive (*S. pyogenes*).

Notable hits: Confirmed known resistance mechanisms in *S. pneumoniae* and identified possible compensatory mutations requiring functional validation. Identified significant associations with invasiveness in *S. pyogenes* including SNPs in the intergenic region upstream of IgG-binding protein H (*sph*) and *nrdI* (ribonucleotide reductase) requiring functional validation.

Annotation: Kmers were searched for in the *de novo* assemblies to assist in their annotation, and kmers were assembled to improve the annotation of short kmers. Kmers were also mapped to a reference sequence and SNPs were called.

1.2.6.4 Sample covariance

Alam et al. (2014)

Variation characterised: SNPs identified in core regions.

Variant exclusion criteria: None.

Control of population structure: Accounted for sample covariance using ROADTRIPS or QROADTRIPS for binary and continuous phenotypes, respectively (Thornton & Mcpeek 2010). The genomic control inflation factor was also calculated, and the QQ plot inspected for evidence of systematic inflation.

Formal test: ROADTRIPS applies three tests for association: The *RM* test, the ROADTRIPS extension of the M_{QLS} test of Thornton & McPeek (2007), and the $R\chi$ and *RW* tests, the ROADTRIPS extensions of the corrected χ^2 and W_{QLS} tests, respectively, of Bourgain et al. (2003). For all tests, the *P* value is calculated based on a χ_1^2 asymptotic null distribution.

Multiple testing correction threshold: Bonferroni-correction at $P < 0.05$ corrected by the total number of SNPs, giving a threshold of $P < 8.9 \times 10^{-7}$, to control the FWER.

Benefits: The authors report that ROADTRIPS accounts for both known and unknown population structure.

Limitations: The accessory genome was not investigated and SNPs were not imputed. Wu et al. (2011) and Price et al. (2010) found that the method did not sufficiently control the false positive rate in simulations of human data for highly differentiated SNPs.

Organisms/strains: 75 *S. aureus* strains.

Phenotypes: Binary – vancomycin-sensitive vs vancomycin-intermediate. Continuous – vancomycin minimum inhibitory concentration (MIC) measured by an Etest.

Notable hits: A mutation in the gene *rpoB* (H481Y/N/L) when testing the continuous phenotype.

Phenotype prediction: Multiple methods, including a Random Forest classifier, using rare mutations in candidate genes plus experimentally validated mutations as independent features. Training was performed on 47 strains and prediction on 28, achieving an accuracy of 89% and sensitivity for vancomycin-sensitive of 57%.

1.2.6.5 Summary of the new wave of bacterial GWAS

In summary, it is clear that the field of bacterial GWAS currently consists of a wide variety of analytical approaches. In order to avoid identifying false positives, lessons from the decade of methodological advances in human GWAS should be followed, such as the importance of controlling for population stratification, multiple testing correction and imputation of missing data, however these are not universally applied in bacterial studies. There is also currently an absence of estimations of sample heritability of important bacterial phenotypes. Estimating sample heritability provides an upper limit of the bacterial genetic contribution to the trait of interest, and indicates the likelihood of identifying underlying bacterial genetic factors. Phenotypes which are less heritable and more reliant on host and environmental factors will be challenging to determine a bacterial genetic basis for.

Multiple methods have found significant associations with known antimicrobial resistance-conferring variants, however resistance is a unique phenotype in the strength of selection underlying the evolution of resistance. The effect sizes reported in some bacterial GWAS studies are much larger than those typically seen in human studies, with typical odds ratios of 1.1 or lower and only rarely above 1.3 (Price, Spencer & Donnelly 2015). For example, the bacterial study by Chewapreecha et al. (2014) identified

significant resistance-conferring SNPs with a median odds ratio of 11.09. Lees et al. (2016) also described how many thousands of samples will be needed to have the power to detect causal associations when the odds ratio of the causal locus is low, sizes beyond those which are currently reported in bacterial studies.

1.2.7 Thesis objectives and outline

1.2.7.1 Aims

In contrast to over a decade of GWAS application and methodological development in the human setting, bacterial GWAS is still in its infancy. Significant differences between the genomes of humans and bacteria mean that developments in the human setting cannot be applied unthinkingly to bacteria, but the lessons learnt from human GWAS nevertheless provide a crucial starting point. Building on the methodological development in human GWAS, I aimed to:

- Test empirically the feasibility of applying GWAS to bacteria in light of the challenges faced by strong population structure, genome-wide LD, and the presence of large accessory genomes.
- Develop methods for applying GWAS to bacteria.
- Assess bacterial GWAS feasibility by testing antimicrobial resistance as a phenotype, where the bacterial genetic basis underlying the trait is well understood.
- Apply the bacterial GWAS methods to phenotypes where the underlying bacterial genetic basis is less well understood.
- Investigate whether particular bacterial lineages are associated with important phenotypes.
- Quantify the heritability of important bacterial phenotypes from the perspective of the bacterial genome.

- Search for signals of association between individual bacterial genetic variants and important phenotypes and their implications on our understanding of bacterial pathogens.

1.2.7.2 Approach

Along with colleagues, I developed methods for applying GWAS to bacteria and assessed their feasibility by testing phenotypes where the underlying genetic basis is well understood. I then applied the methods to investigate phenotypes where the bacterial genetic basis is less well understood in order to gain insights into the genetic architecture of important bacterial traits.

I first focused on antimicrobial resistance because the availability of detailed phenotyping and the comprehensive knowledge base of antimicrobial resistance determinants, as mirrored by the high sensitivity and specificity of antimicrobial resistance prediction, make it an ideal candidate for establishing proof-of-principle for GWAS feasibility and identifying potential challenges. I conducted a detailed analysis of fusidic acid resistance in *S. aureus* to investigate the impact of controlling for population structure in highly structured bacterial populations. I demonstrated how a method devised by Daniel Wilson which tests for associations with bacterial lineages, in addition to individual variants, can retrieve power lost when controlling for population structure in bacterial GWAS. Along with colleagues, I also investigated the ability of GWAS to pinpoint causal variants in multiple antimicrobials across four major pathogens, *M. tuberculosis*, *S. aureus*, *E. coli* and *K. pneumoniae*.

I then moved on to applying these methods to investigate two phenotypes where the underlying bacterial genetic basis is less well understood, virulence and host adaptation. Specifically, I investigated carriage vs invasive disease in *N. meningitidis* and wild bird vs chicken colonisation in *C. jejuni*. I aimed to quantify the heritability of these

important phenotypes from the bacterial genetic perspective, investigate whether bacterial lineages are associated with these traits and finally identify loci contributing to natural phenotypic variability in order to understand the genetic architecture of important traits.

Together, these studies demonstrate the feasibility of applying GWAS to bacteria, identify potentially novel findings with respect to antimicrobial resistance, virulence and host adaptation, and identify challenges and limitations in the application of bacterial GWAS in understanding the bacterial genetic basis of traits displayed by important pathogens.

Chapter 2

Methods for applying genome-wide association studies in bacteria

2 Methods for applying genome-wide association studies in bacteria

In this chapter existing and newly-developed methods applied to genome-wide association studies in bacteria are described. Methods common to all chapters are described here, and methods unique to particular chapters are described within the chapter.

2.1 Isolates, Sequencing and Variant Calling

In Chapter 3, 241 *Escherichia coli* and 176 *Klebsiella pneumoniae* UK clinical isolates sampled by Nicole Stoesser were investigated along with 992 *Staphylococcus aureus* and 1,735 *Mycobacterium tuberculosis* isolates previously reported (Gordon et al. 2014; Walker et al. 2015). All isolates were tested for resistance to multiple antimicrobials using routine clinical laboratory protocols and DNA was extracted and sequenced on Illumina platforms as previously described (Stoesser et al. 2013; Gordon et al. 2014; Walker et al. 2015).

In Chapter 4, 261 *Neisseria meningitidis* isolates sampled from the Czech Republic in 1993 were analysed (Jolley et al. 2000). Illumina sequencing reads were downloaded from the European Nucleotide Archive (ENA, http://www.ebi.ac.uk/ena), and Velvet *de novo* assemblies were downloaded from pubMLST (https://pubmlst.org/neisseria/) in collaboration with Martin Maiden.

In Chapter 5, 480 *Campylobacter jejuni* isolates were analysed which were collated by Samuel Sheppard and Guillaume Méric, who provided paired end Illumina sequencing reads.

2.1.1 SNP calling and assembly

Sequence reads were mapped and assembled using the Modernising Medical Microbiology (MMM) bioinformatics pipeline as described previously (Eyre et al. 2012; Everitt et al. 2014; Stoesser et al. 2017). The pipeline was run by members of the MMM bioinformatics team for Chapters 3 and 4 and by me for Chapter 5. Sequence reads were mapped to species specific reference genomes using Stampy (Lunter & Goodson 2011), and repetitive regions were masked prior to variant calling as the true mapping location is difficult to determine when reads map well to multiple regions. SNP calling was performed using SAMtools (Li et al. 2009) under a diploid model as part of the MMM pipeline and variants were required to pass multiple quality filters. To ensure accurate SNP calls, high quality bases were defined as a base quality of 33 or more and only calls passing the following filters were retained (Eyre et al. 2012; Everitt et al. 2014; Stoesser et al. 2017):

- The proportion of high-quality bases supporting the call must be ≥90%
- Base calls must be supported by at least five high-quality bases, including at least one in each direction
- Base calls must be homozygous under a diploid model
- The root mean square mapping quality of reads covering the site must be ≥30
- The Phred scaled quality supporting the call must be ≥25
- Reads spanning the site must be made up of ≥35% high-quality bases

Sequence reads were also assembled *de novo* using the MMM pipeline using Velvet (Zerbino & Birney 2008).

2.2 Phylogenetic inference

Maximum likelihood phylogenies were estimated for visualisation of phylogenetic relationships in each sampling frame and for SNP imputation purposes using RAxML

(Stamatakis 2014) with a general time reversible (GTR) model and no rate heterogeneity, using alignments from the mapped data based on biallelic sites, with non-biallelic sites being set to the reference allele. Rapid bootstrapping was also achieved using RAxML with a GTR model where 100 bootstrap replicates were performed.

2.3 SNP imputation

Due to the error prone nature of Illumina sequencing, strict SNP-calling thresholds needed to be applied to ensure only reliable mapping-based SNP calls were made (Ross et al. 2013). This results in uncalled bases across the genome, some due to genuine deletions relative to the reference genome, but some due to ambiguity in the SNP calling. As the size of datasets increase, the number of sites where at least one individual does not have a call at that site may increase, so restricting analyses to sites with no missing data could potentially result in a huge loss in the number of sites available for association tests. Conceptually, GWAS can be thought of as model comparison on a grand scale in which each variable genetic locus is assessed for its ability to explain patterns of phenotype variability. Sensible model comparison relies on a direct comparison between the competing models (candidate loci) using the same data. While it is possible to ignore missing sites by testing for association at each locus by removing individuals without a SNP call, this violates the principle of basing model comparison on identical data, making it difficult to interpret direct comparisons of *P* values between loci.

For these reasons, SNP imputation is therefore considered necessary for GWAS and is used across human GWAS (Scheet & Stephens 2006; Browning & Browning 2009; Purcell et al. 2007; Howie, Donnelly & Marchini 2009; Li et al. 2010). Within human genetics, imputation is achieved by using a reference panel of haplotypes to impute bases in a sample of individuals typed at some SNPs but not all, allowing for a higher power within the study and also for the resolving of causal variants (Marchini & Howie 2010).

In Chapter 3, we tested the accuracy of imputing missing base calls using two different approaches, ClonalFrameML (Didelot & Wilson 2015) and Beagle (Browning & Browning 2009). ClonalFrameML is a method developed for bacteria which takes the clonal frame estimated by maximum likelihood, in our case RAxML (Stamatakis 2014), and then jointly reconstructs ancestral states plus missing bases using maximum likelihood (Pupko et al. 2000). RAxML infers the clonal frame as recombination affects just a portion of the genome, therefore the signal of the clonal frame dominates and maximum likelihood can accurately reconstruct the topology (Hedge & Wilson 2014). Beagle is a model which locally clusters haplotypes, haplotypes are iteratively grouped at each marker based on the similarity of the haplotypes at the adjacent markers, and the number of clusters at each marker can vary. Missing genotype probabilities are calculated from the model fitted at the final iteration (Browning & Browning 2009).

2.3.1 Testing imputation accuracy

One hundred sequences were randomly sampled from each of the four GWAS data sets in Chapter 3 in order to simulate data to test imputation accuracy. Samples were chosen with reference to the phylogeny in order to maintain phylogenetic diversity. RAxML was used to estimate maximum likelihood phylogenies for the 100 sequences of each species as described in Section 2.2 (Stamatakis 2014). Any columns in the alignments which corresponded entirely to ambiguous bases in the reference genome (for example repeat-masked regions) were excluded from the analysis. Imputation was performed using both ClonalFrameML and Beagle to test imputation accuracy. Imputation accuracy was summarised by the proportion of correctly imputed ambiguities and by Pearson correlation between truth and imputations, broken down by the frequency of Ns per site and the minor allele frequency. As detailed in Chapter 3.4.2, imputation by

ClonalFrameML was generally more accurate than Beagle, so ClonalFrameML was used for all GWAS analyses.

2.4 Counting kmers

Some genetic diversity, such as insertions, deletions and repeats, is difficult to capture using standard variant calling tools. Variant calling by mapping to a reference genome also fails to capture potentially important and informative gene presence and absence. Therefore, to capture non-SNP-based variation, we pursued a kmer or word-based approach where all unique 31 base haplotypes were counted from either sequencing reads or Velvet assemblies. The advantage of a kmer-based method is that it is alignment free and therefore not biased by a chosen reference genome, but can still capture SNP based variation, and even short haplotypes if multiple SNPs fall within the chosen kmer length. This approach was earlier applied to bacterial GWAS by Sheppard et al. (2013) who tested 20bp, 30bp and 40bp kmers counted from *Campylobacter* isolates for association with cattle versus chicken hosts where they found comparable results across the different kmer lengths.

We counted kmers from raw Illumina sequencing reads in Chapters 3 and 4 and from Velvet assemblies in Chapters 4 and 5. In all cases, we counted 31 base kmers per isolate using dsk (Rizk, Lavenier & Chikhi 2013) following adaptor trimming and removal of duplicates and low-quality reads using Trimmomatic in the case of counting kmers from sequencing reads (Bolger, Lohse & Usadel 2014). In-house software *gwas_kmer_pattern* written by Daniel Wilson in C++ was then used to create a deduplicated set of variably present kmers across the data set, with the presence or absence of each determined per isolate. When counting kmers from sequencing reads, a kmer was counted as present if observed five or more times in an isolate. When counting kmers from Velvet assemblies, a kmer was counted as present if observed at least once in

an assembly. The C++ software can be found at https://github.com/jessiewu/bacterialGWAS/blob/master/kmerGWAS/gwas_kmer_pattern.

2.5 Defining the pan genome

To investigate gene presence or absence, we created a pan-genome for each set of isolates investigated in Chapter 3. Open reading frames within the *de novo* assemblies were annotated and a set of protein sequences for each assembly was identified using the Bayesian gene-finding program Prodigal (Hyatt et al. 2010). The annotated protein sequences were then clustered using CD-hit (Li & Godzik 2006) using a clustering threshold of 70% identity across 70% of the longer sequence. The output of CD-hit was then converted into a matrix of binary genotypes indicating the presence or absence of each gene cluster in each genome. This pangenome pipeline was created and applied by Jane Charlesworth and can be found at: https://github.com/jessiewu/bacterialGWAS/tree/master/pangenomeGWAS.

2.6 Calculating association statistics without controlling for population structure

We calculated the significance of associations before and after controlling for population structure in order to determine the effect of controlling for population structure in bacterial populations. For SNPs and gene presence absence data, associations between each SNP or gene and the phenotype was tested by logistic regression in R. For the kmer analyses, an association between the presence or absence of each kmer was tested using a χ^2 test implemented in *gwas_kmer_pattern* by Daniel Wilson. For each variant a *P* value was produced. These tests were performed merely for comparison with methods that control population structure and, occasionally, to downsample kmers for full linear mixed model analysis.

2.7 Correcting for multiple testing

Multiple testing was accounted for in all analyses by applying Bonferroni corrections (Dunn 1959). The unit of correction for all studies of individual loci was taken to be the number of unique "phylopatterns" i.e. the number of unique partitions of individuals according to allele membership. The locus effect of an individual variant was considered to be significant if its P value was smaller than α/n_p, where we took $\alpha = 0.05$ to be the genome-wide false positive rate (i.e. family-wide error rate, FWER) and n_p to be the number of unique phylopatterns. In Chapter 3 where both SNPs and a pangenome were defined, the number of unique phylopatterns was calculated based on both SNPs and variably present genes. In SNP-only analyses in Chapters 4 and 5, n_p was taken to be the number of unique SNP phylopatterns, and in all kmer analyses n_p was taken to be the number of unique kmer presence/absence phylopatterns. Adjusting for the number of unique phylopatterns rather than the number of loci avoids incurring a punitive Bonferroni correction when there is no advantage to be gained because loci with identical phylopatterns cannot be distinguished statistically. When testing for lineage effects through the Bayesian Wald test (Section 2.10) for principal component-phenotype associations, a Bonferroni-correction was applied for the number of non-redundant principal components which is equal to the sample size n when no two genomes are identical.

2.8 Calculating approximate posterior probabilities

In Chapters 4 and 5 we calculated approximate posterior probabilities for the biallelic SNPs and for the kmer analyses, conservatively assuming one causal variant. Let M_i indicate that variant i is the causal variant, assuming just one causal variant out of the L variants. Let R_i be the maximized likelihood ratio from the test for variant i. Then the

posterior probability that variant *i* is the causal variant, assuming equal prior probabilities among the variants, is approximately

$$\Pr(M_i \mid \text{Data}) \approx R_i / (R_1 + \ldots + R_L) \quad (1)$$

The motivation for this approximation, which is used as indicative only, is that the maximized likelihood ratio provides an upper bound for the Bayes factor, and is more appropriate when all tests have the same degrees of freedom (Wilson 2017), for example when comparing among biallelic SNPs or among kmer presence vs absence. We subsequently calculated the 95% credibility set of variants defined to be the smallest such set for which the sum of the posterior probabilities met or exceeded the threshold of 95%.

2.9 Linear Mixed Models

In order to control for population structure, Linear mixed model (LMM) analysis implemented within the program GEMMA (Zhou & Stephens 2012) was applied. LMMs have been shown to control for close relatedness within samples by capturing the fine structure of populations more accurately than other approaches (Price et al. 2010) and are more applicable than phylogenetic approaches due to the presence of recombination in most bacterial species (Pérez-Losada et al. 2006; Vos & Didelot 2009). The development of the application of LMMs in human GWAS is described in Chapter 1.1.5.3. In LMMs the phenotype is modelled as depending on the fixed effects of covariates including an intercept, the 'foreground' fixed effect of the locus whose individual contribution is to be tested, the 'background' random effects of all the loci, and the random effect of the environment: (Yu et al. 2006; Kang et al. 2008; Kang et al. 2010; Lippert et al. 2011; Listgarten et al. 2012). The LMM can be expressed verbally as

phenotype = covariates + foreground locus + background loci + environment

Formally,

$$y_i = W_{i1}\alpha_1 + \ldots + W_{ic}\alpha_c + X_{il}\beta_l + X_{i1}\gamma_1 + \ldots + X_{iL}\gamma_L + \varepsilon_i, \quad (2)$$

where there are n individuals, c covariates, L loci, l is the foreground locus, y_i is the phenotype in individual i, W_{ij} is covariate j in individual i, α_j is the effect of covariate j, X_{ij} is the genotype of locus j in individual i, β_l is the foreground effect of locus l, γ_j is the background effect of locus j and ε_i is the effect of the environment (or error) on individual i. Biallelic genotypes are numerically encoded as as $-f_j$ (common allele) or $1-f_j$ (rare allele), where f_j is the frequency of the rare allele at locus j. This ensures that the mean value of X_{ij} over individuals i is zero for any locus j. As tri- and tetra-alellic loci are rare, we use only biallelic loci to model background effects. When the foreground locus is tri-allelic ($K = 3$) or tetra-allelic ($K = 4$), the genotype in individual i is encoded as a vector indicating the presence (1) or absence (0) of the first ($K - 1$) alleles and β_l becomes a vector of length ($K-1$).

Treating the background effects of the loci as random effects means the precise values of the coefficients γ_j are averaged over. The γ_js are assumed to follow independent normal distributions with common mean 0 and variance $\lambda\tau^{-1}$ to be estimated. Since most loci are expected to have little or no effect on a particular phenotype, this tends to constrain the magnitude of the background effect sizes to be small. The environmental effects are also treated as random effects assumed to follow independent normal distributions with mean 0 and variance τ^{-1}. The model can be rewritten in matrix form as

$$\mathbf{y} = \mathbf{W}\,\boldsymbol{\alpha} + \mathbf{X}_{\bullet l}\,\beta_l + \mathbf{u} + \boldsymbol{\varepsilon} \tag{3}$$

with

$$\mathbf{u} = \mathbf{X}_{\bullet 1}\,\gamma_1 + \ldots + \mathbf{X}_{\bullet L}\,\gamma_L$$

$$\mathbf{u} \sim \mathrm{MVN}_n(0, \lambda\,\tau^{-1}\,\mathbf{K})$$

$$\boldsymbol{\varepsilon} \sim \mathrm{MVN}_n(0, \tau^{-1}\,\mathbf{I}_n)$$

where $\mathbf{u}$ represents the cumulative background effects of the loci, MVN denotes the multivariate normal distribution, $\mathbf{I}_n$ is an $n \times n$ identity matrix and $\mathbf{K}$ is an $n \times n$

relatedness matrix defined as $\mathbf{K} = \mathbf{X}\,\mathbf{X}'$, which captures the genetic covariance between individuals. $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2 \ldots \mathbf{w}_c)$ is an $n \times c$ matrix of covariates (fixed effects) including a column vector of 1, the intercept. $\boldsymbol{\alpha}$ is a $c \times 1$ vector of covariate coefficients including the intercept. $\mathbf{X}_{\bullet l}$ is the marker genotype being tested, β_l is the effect size of the marker. $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of errors, τ^{-1} is the variance of the residual errors, λ is the ratio between the two variance components.

2.9.1 Estimating sample heritability

Heritability of the sample, the proportion of the phenotypic variation that can be explained by the bacterial genotype, was estimated using the LMM null model in GEMMA where there is no foreground SNP, along with an estimate of the standard error (Zhou & Stephens 2012). Sample relatedness was calculated from either biallelic SNPs or kmers counted from Velvet assemblies, which was used to infer the narrow-sense heritability that is explained by these variants. This measure of heritability captures the additive effect of the variants. Heritability is characterised relative to the population it is calculated from at the time the data was sampled. Estimating heritability in case control studies is dependent on the prevalence of cases and the sampling scheme (Zaitlen & Kraft 2012). Ascertainment strategy will therefore impact on the relationship between the estimate of sample heritability and the true population heritability. In case-control studies, cases are typically very over-enriched, which will cause sample heritability to be much higher than true heritability (Browning & Browning 2011; Lee et al. 2011).

2.9.2 Testing for locus effects

To assess the significance of the effect of an individual locus l on the phenotype whilst controlling for population structure and and background genetic effects, the parameters of the linear mixed model $\alpha_1 \ldots \alpha_c$, β_l, λ and τ were estimated by maximum likelihood and a

likelihood ratio test with ($K-1$) degrees of freedom was performed against the null hypothesis that $\beta_l = 0$ using the software GEMMA (Zhou & Stephens 2012) version 0.93b available at https://github.com/sgearle/bugwas/tree/master/gemma.

In all chapters, for the analyses of SNPs, genes and kmers, we computed the relatedness matrix **K** from biallelic SNPs only using the option "-gk 1" in GEMMA to calculate the centered relatedness matrix. In Chapters 4 and 5 we also computed the relatedness matrix **K** from kmers counted from Velvet *de novo* assemblies. We tested for foreground effects at all biallelic, triallelic and tetrallelic SNPs and kmers in all chapters, plus genes in Chapter 3. GEMMA was run using no minor allele frequency cut-off to include all variants. GEMMA was modified by Daniel Wilson to output the maximized log-likelihood under the null and alternative models and $-\log_{10} P$ values were then calculated by a likelihood ratio test using R.

To perform LMM on tri- and tetra-allelic SNPs, each SNP was encoded as $K - 1$ binary columns corresponding to the first $K - 1$ alleles. For each column, an individual was encoded 1 if it contained that allele and 0 otherwise. The first column was input as the genotype, and the others as covariates into GEMMA. The log-likelihood of the null from the biallelic SNPs, together with the log-likelihood under the alternative for each of the SNPs, was used to calculate the *P* value per SNP.

Due to the large number of kmers present within each data set, it was not computationally feasible to run LMM on all kmers for all analyses. For the 26 antibiotic resistance studies in Chapter 3, the LMM was applied to the top 200,000 most significant kmers from the χ^2 test plus 200,000 randomly selected kmers of those remaining. The randomly selected kmers were used to indicate whether some were becoming relatively more significant than the top 200,000, providing a warning in the case where large numbers of kmers became significant only after controlling for population structure. This

however was not an issue for the 26 antibiotic resistance GWAS. For the analyses in Chapters 4 and 5, all kmers were tested for association whilst controlling for population structure. This was computationally feasible because each chapter involved one or a small number of GWAS.

In Chapters 4 and 5, LMM analyses were also applied to the biallelic SNPs including principal components as additional fixed effects. This was investigated because in the human setting it has been shown that the inclusion of principal components as additional fixed effects in the LMM can improve power and reduce false positives (Tucker, Price & Berger 2014; Widmer et al. 2014). Therefore, principal components were accounted for in order to ensure that population structure had been controlled for sufficiently and to test the robustness of the results. PCs 1, 1-5, 1-10, and 1-20 were all separately included, calculated in R using the function prcomp() (R Core Team 2015).

2.10 Testing for lineage effects

Since controlling for population structure significantly reduces power at population-stratified variants, and since bacterial genomes consist of a large proportion of population-stratified variants, we recovered information from the LMM by capturing lineage-level differences in phenotype.

2.10.1 Principal Components Analysis

We performed Principal Components Analysis (PCA) in R using the function prcomp() (R Core Team 2015) in order to define bacterial lineages using principal components (PCs). This is because Daniel Wilson observed that PCs tend to trace paths through the clonal frame genealogy, corresponding to recognisable lineages. This can be visualised by branch colouring in Figure 3.7, Chapter 3. As PCs are mutually uncorrelated, defining lineages in this way is expected to reduce loss of power to detect lineage associations due

to correlations, compared to the alternative of defining lineages using genetic clustering approaches.

Computing the PCs produces an L by n loading matrix $\mathbf{D}$ and an n by n score matrix $\mathbf{T}$ where $\mathbf{T} = \mathbf{X}\,\mathbf{D}$ and $\mathbf{X}$ is defined as above. D_{ij} records the contribution of biallelic SNP i to the definition of PC j while T_{ij} records the projection of individual i on to PC j.

2.10.2 Wald test for lineage effects

The following Wald test for lineage effects was devised by Daniel Wilson. The point estimates and standard errors for the background locus effects are assumed to follow normal distributions with common mean 0 and variance $\lambda\tau^{-1}$. This tends to result in 'shrinkage' in that the estimated effects are small in magnitude relative to the standard errors and therefore not significantly different from zero. For this reason, the estimates of these coefficients are typically ignored. However, the background locus effects can cumulatively capture systematic phenotypic lineage differences, particularly because there can be many loci with identical phylopatterns. We therefore reinterpreted the 'post-data' (equivalently, the empirical Bayes posterior) distribution of the background locus random effects, $\boldsymbol{\gamma}$, from the LMM, in terms of the lineage-level phenotypic differences.

We calculated the mean and variance-covariance matrix of the multivariate normal post-data distribution of $\boldsymbol{\gamma}$ in the LMM null model. These are equivalent to those of a ridge regression (O'Hagan & Forster 2010) and were calculated as:

$$\boldsymbol{\mu} = (\mathbf{X}'\mathbf{X} + 1/\lambda\,\mathbf{I}_L)^{-1}\,\mathbf{X}'\mathbf{y} \text{ and } \boldsymbol{\Sigma} = (\tau\,\mathbf{X}'\mathbf{X} + 1/\lambda\,\mathbf{I}_L)^{-1} \quad (4)$$

respectively. Both λ and τ were estimated by GEMMA under the LMM null model.

The background random effects can be rewritten in terms of the contribution of the n PCs using the inverse transformation of the biallelic variants from the PCA, $\mathbf{X} = \mathbf{T}\,\mathbf{D}^{-1}$

$$\mathbf{u} = \mathbf{X}_{\bullet 1}\gamma_1 + \ldots + \mathbf{X}_{\bullet L}\gamma_L \quad (5)$$

$$= \mathbf{X}\boldsymbol{\gamma} = \mathbf{T}\,\mathbf{D}^{-1}\boldsymbol{\gamma} = \mathbf{T}\,\mathbf{g}$$

$$= \mathbf{T}_{\bullet 1}g_1 + \ldots + \mathbf{T}_{\bullet n}g_n$$

where $\mathbf{g} = \mathbf{D}^{-1}\boldsymbol{\gamma}$, g_j being the background effect of PC j on the phenotype. We therefore computed the mean and variance of the post-data distribution of $\mathbf{g}$ as $\mathbf{m} = \mathbf{D}^{-1}\boldsymbol{\mu}$ and $\mathbf{S} = \mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}$ respectively using the affine transformation for a multivariate normal distribution. To test the null hypothesis of no background effect of PC j (i.e. $g_j = 0$) we employed a Wald test with test statistic $w_j = m_j^2/S_{jj}$, which we compared against a χ^2 distribution with one degree of freedom to obtain a P value.

2.10.3 Phenotype prediction

Phenotypes were predicted using bacterial genetic data using a ridge regression in which every SNP is used for prediction, and there is no foreground SNP, which is the null Linear Mixed Model. Phenotypes were coded as cases (1) and controls (0) and mean centred, so that positive values represented cases. The phenotype was predicted as follows:

$$\mathbf{y} = \mathbf{W}\,\boldsymbol{\alpha} + \mathbf{u} \quad (6)$$

$$\mathbf{u} = \mathbf{X}_{\bullet 1}\mu_1 + \ldots + \mathbf{X}_{\bullet L}\mu_L$$

where α and $\boldsymbol{\mu}$ are estimated by the LMM null model. Binary case vs control predictions were made by identifying whether the predictions were negative or positive. The continuous predictions were also analysed.

2.10.4 Identifying non genome-wide PCs

Some PCs capture variation localised to particular regions of the genome. Devised by Daniel Wilson, we identified non genome-wide PCs by testing for spatial heterogeneity of the loading matrix $\mathbf{W}$ for biallelic SNPs across the genome. SNPs were grouped into 20 contiguous bins (indexed by j) of nearly equal sizes N_j and the mean O_{ij} and variance V_{ij}

in the absolute value of the SNP loadings for PC *i* in bin *j* were calculated, along with the mean absolute value E_i of the SNP loadings for PC *i* across all SNPs. The null hypothesis of no heterogeneity was assessed by comparing the test statistic $\chi_i^2 = \Sigma_j (O_{ij} - E_i)^2/(V_{ij}/N_j)$ to a χ^2 distribution with degrees of freedom equal to the number of bins minus one to obtain a *P* value.

2.10.5 Assigning loci to lineages

Locus-specific effects were reassessed in light of the lineage effects by assigning variants to lineages according to the principal component to which they were most correlated. To achieve this, correlations were calculated between each of the variant phylopatterns and the projections of the isolates onto each of the principal components individually by Pearson correlation using the R function cor() with default settings. Variants were assigned to the principal component to which their absolute correlation was highest.

2.10.6 Testing power by simulating phenotypes

To assess the performance of controlling for population structure and testing for lineage effects, Daniel Wilson performed 100 simulations per species for the four datasets in Chapter 3. A biallelic SNP was chosen randomly (MAF>20%) to be the causal SNP. Binary phenotypes (case or control) were simulated for each genome with case probabilities of 0.25 and 0.5, respectively, in individuals with the common and rare allele at the causal SNP (odds ratio of 3). Locus and lineage effects were tested for each simulated dataset. The power to detect locus effects was defined as the proportion of simulations in which the causal SNP was found to be significant. This was compared to a theoretically optimum power computed as the proportion of simulations in which the causal SNP was found to be significant when population structure and multiple testing were not controlled for. The power to detect lineage effects was defined as the proportion

of simulations in which the principal component most strongly correlated to the causal SNP was found to be significant.

2.11 Variant annotation

SNPs were annotated in R using the reference fasta and genbank files in order to determine SNP type (synonymous, non-synonymous, nonsense, read-through and intergenic), codon, codon position, reference and non-reference amino acids, gene name and gene product, on the assumption of a single change to the reference sequence.

Unlike the SNP approach where we can easily refer to the reference to find what gene the SNP is in and the effect that it may have, the kmer approach lacks a simple annotation approach. We used BLAST (Camacho et al. 2009) to identify the kmers in databases of annotated sequences. Kmers were first annotated against a BLAST database created of all refseq genomes of the relevant genus on NCBI. This enabled automatic annotation of all kmers with high identity against the genus-specific database as the gene annotations are also available to download. All kmers were also searched against the whole nucleotide NCBI database, first to compare and confirm the matches made against the first genus database and second to annotate the kmers that did not find a match in the genus database. Kmers were also mapped to relevant reference genomes using Bowtie2 (Langmead & Salzberg 2012) using the options "-r -D 24 -R 3 -N 0 -L 18 -i S,1,0.30" to identify a single best mapping position for each kmer and "-r -D 24 -R 3 -N 0 -L 18 -i S,1,0.30" to identify all possible mapping positions. This was used to determine whether significant kmers contained particular SNPs to further investigate the type of variation they were capturing, and also to visualise the results genome-wide using Manhattan plots.

Where pangenome analyses were used, genes were annotated for each CD-hit cluster using BLAST (Camacho et al. 2009). A representative sequence from each cluster

was searched against a database of curated protein sequences downloaded from UNIPROT (The UniProt Consortium 2015).

2.12 Pairwise SNP tests of association

In Chapters 4 and 5, to determine whether pairs of significant SNP associations represented independent signals, every pair of unique biallelic SNP phylopatterns above a given significance threshold were tested using LMM including both SNPs as fixed effects, thereby assuming additivity between the two loci. To account for the additional number of tests performed, we applied a more stringent Bonferroni correction calculated as 0.05/(total number of phylopatterns × number of phylopattern pairs exceeding the single SNP analysis threshold).

2.13 Software

The *bacterialGWAS* software my colleagues and I developed for applying GWAS to bacteria, plus the R package *bugwas* I developed to test for associations whilst controlling for population structure and identify lineage effects, can be found at the following github pages:

http://github.com/jessiewu/bacterialGWAS

http://github.com/sgearle/bugwas

Chapter 3

Proof of principle for bacterial genome-wide association studies

3 Proof of principle for bacterial genome-wide association studies

The results of this chapter have been published in the following paper:

Earle SG*, Wu CH*, Charlesworth J*, Stoesser N, Gordon NC, Walker TM, Spencer CCA, Iqbal Z, Clifton DA, Hopkins KL, Woodford N, Smith EG, Ismail N, Llewelyn MJ, Peto TE, Crook DW, McVean G, Walker AS, Wilson DJ. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat. Microbiol.* **1**, 16041 (2016).

* Indicates joint authorship

3.1 Introduction

3.1.1 Antimicrobial resistance is an increasing problem

The evolution and spread of antimicrobial resistance poses a growing threat to public health. Antimicrobial resistance in many bacterial pathogens is rising in frequency worldwide, with antimicrobial therapy failures threatening the safe provision of healthcare and causing the deaths of hundreds of thousands every year (Davies & Davies 2010; World Health Organisation 2012). One way of combating this would be the discovery of new antimicrobials, however the rate of antimicrobial discovery is low, and resistance will emerge to new antibiotics like it has to existing ones (Brown & Wright 2016).

Antimicrobial susceptibility testing is crucial in the treatment of bacterial infections in order to determine which antimicrobials will be effective in treating disease on an individual patient basis. However, phenotypic tests are not instant, taking 1-2 days for *Staphylococcus aureus* for example and weeks for the slower growing *Mycobacterium tuberculosis* (Bradley et al. 2015). Whole genome sequencing of bacteria therefore has

the potential to transform routine clinical microbiology as it is rapid and becoming increasingly more cost effective (Didelot et al. 2012). By using next generation sequencing to identify known resistance determining variants, the speed of antibiotic resistance prediction can be greatly expedited. Techniques to hasten the speed of DNA extraction ready for whole genome sequencing are improving (Votintseva et al. 2017), leaving the unfinished challenge of creating a comprehensive catalogue of resistance-conferring mutations able to rapidly predict resistance versus susceptibility.

3.1.2 Genetic classes of antimicrobial resistance

The molecular and genetic mechanisms of resistance have been studied extensively (see e.g. Alekshun & Levy 2007; Palmer & Kishony 2013). Antibiotics can be categorised into various classes, and classes of commonly used antimicrobials include β-lactams, aminoglycosides, glycopeptides, tetracyclines, macrolides and quinolones (Davies & Davies 2010). The various classes each have different bacterial targets, resulting in the evolution of different modes of resistance (Davies & Davies 2010). Resistance can be intrinsic, due to genes already present in the host chromosome, but it can also be acquired through mutations and the transfer of resistance determinants on mobile genetic elements (Alekshun & Levy 2007). For example the carriage of efflux pumps can be an intrinsic mechanism enabling resistance to multiple antimicrobials (Li, Livermore & Nikaidol 1994; Webber & Piddock 2003). Chromosomal mutations can also induce resistance by altering antimicrobial targets, for example specific single nucleotide polymorphisms (SNPs) in the genes *gyrA*, *grlA* and *grlB* confer resistance to the fluoroquinolone ciprofloxacin in *Staphylococcus aureus* and other species (Kaatz & Seo 1997; Tanaka et al. 2000). Resistance can also be acquired by the transfer of resistance genes on plasmids, bacteriophages, transposons and other mobile genetic elements, such as the acquisition of plasmid-borne β-lactamases (Jacoby & Munoz-Price 2005). In order to gain a full

understanding of the genetic basis of antimicrobial resistance it is therefore crucial that both the core and accessory genomes are investigated.

3.1.3 Predicting antimicrobial resistance

Studies using whole genome sequencing to predict phenotypic resistance have been shown to be effective. Zankari et al. (2013) and Stoesser et al. (2013) investigated 200 isolates originating from Danish pigs covering four bacterial species and 143 clinical isolates of *E. coli* and *K. pneumoniae* from Oxfordshire, respectively. These studies used simple BLAST methods to identify resistance determinants across their isolates. Although their predictions were not validated using replication datasets and were small in scale, they found high concordance between their genotypic predictions and the phenotypic measurements. Zankari and colleagues achieved 99.74 concordance and Stoesser and colleagues achieved a sensitivity and specificity of 0.96 and 0.97, providing an indication that resistance prediction from genotypic data is a possible alternative to routinely used phenotypic methods. A more recent study by Moran et al. (2017) predicting antimicrobial resistance in 51 *E. coli* strains from genes identified using PCR and WGS demonstrated the importance of aditionally identifying when resistance-conferring genes are incomplete or lack a promoter when predicting resistance from genotypic data, as the presence of a gene does not necessarily mean it is expressed.

Gordon et al. (2014) investigated the performance of phenotypic resistance prediction of 12 antimicrobials using genotypic data across 501 unrelated *S. aureus* isolates, performing an independent validation after algorithm refinement in 491 also unrelated isolates. A BLAST approach was again used against a reference panel of resistance determinants. Gordon et al. (2014) found an overall sensitivity and specificity of 0.97 (95% confidence interval [95% CI], 0.95 to 0.98) and 0.99 (95% CI, 0.98 to 1), respectively, in comparison to standard methods for susceptibility testing, which is as

sensitive and specific as routine susceptibility testing methods (Nonhoff, Rottiers & Struelens 2005; Carroll et al. 2006). This shows a very high concordance with phenotypic sensitivity, demonstrating that we have a good understanding of the common resistance mechanisms for *S. aureus*. Many of the errors were due to problems with prediction for particular drugs, in particular penicillin and ciprofloxacin. Focusing on increasing our understanding of how these drugs work and creating a comprehensive panel of resistance determinants will improve prediction overall.

The use of WGS for *M. tuberculosis* susceptibility prediction would be particularly transformative due to the many weeks that phenotypic testing can take (Bradley et al. 2015; Votintseva et al. 2017). Genotypic assays do exist for susceptibility testing in *M. tuberculosis* (Drobniewski et al. 2013; Mao et al. 2015) however they target a small number of genetic loci associated with resistance to common drugs. The use of whole genome sequencing to predict antimicrobial resistance for *M. tuberculosis* was reported by Köser et al. (2013) who investigated a patient with extensively drug-resistant (XDR) tuberculosis, revealing the importance of being able to predict resistance to drugs beyond those measured by standard phenotypic assays.

Walker et al. (2015) assessed the performance of genotypic prediction of phenotypic resistance for *M. tuberculosis* across 3,651 isolates, of which 2,099 were used to train the algorithm and 1,552 to validate it. Walker and colleagues defined a resistance-conferring catalogue of mutations as well as a benign catalogue of mutations. Using these catalogues, they were able to predict 89.2% phenotypes during validation, with mean sensitivity and specifiy of 92.3 (95% CI 90.7–93.7) and 98.4 (98.1–98.7) respectively. Although sensitivity and specificity were high for mutations observed when training the model, 10.8% of the phenotypes could not be predicted during validation because the isolates contained uncharacterised mutations, revealing the importance of discovering all

possible resistance-conferring mechanisms. Many common *M. tuberculosis* resistance determinants are well understood, as often a single mutation causes the majority of resistance to a particular drug such as the *katG* S315T mutation which confers resistance to isoniazid (Walker et al. 2015). However, it is likely that there are many less frequent resistance-conferring variants yet to be discovered. Genotypic resistance prediction is only as good as the catalogue of mutations it uses for prediction, so it is crucial that novel resistance-conferring mechanisms are identified. Given the high sensitivity and specificity of prediction, WGS is now being used as a diagnostic tool for the prediction of resistance for tuberculosis infections by Public Health England, enabling patients to be treated with the right antimicrobials quickly (Public Health England 2017).

3.1.4 Previous applications of GWAS to investigate antimicrobial resistance

Due to the increasing numbers of available whole genome sequences, studies are now beginning to investigate the utility of genome-wide association studies for identifying antimicrobial resistance determinants in bacteria. Since the onset of this thesis, early studies investigating associations between bacterial genotypes and antimicrobial resistance phenotypes have identified known resistance-conferring variants as well as potentially novel resistance-conferring variants (Farhat et al. 2013; Alam et al. 2014; Chewapreecha et al. 2014; Salipante et al. 2015; Farhat et al. 2014; Chen & Shapiro 2015; Lees et al. 2016). The analytical approaches taken have been widely varied, as described in Chapter 1.2.6, but have shown the power of applying GWAS to the investigation of antimicrobial resistance. As well as providing evidence of GWAS identifying known causal variants in bacteria, these studies have also highlighted important concepts such as the potential challenge of correlated phenotypes (Chewapreecha et al. 2014), the difficulty in fine-mapping variants when the phenotype is lineage-associated (Lees et al. 2016), the inflation of test-statistics when not controlling

for population structure (Chen & Shapiro 2015) and the decreasing ability to detect causal variants when sample size is reduced (Brynildsrud et al. 2016).

3.1.5 Challenges we expect to encounter when applying GWAS to bacteria

3.1.5.1 Defining the phenotype

The phenotype in the current study was defined by the routine clinical phenotypic measurements of resistance vs. susceptibility. Like any phenotype, resistance is not measured without error. Repeat phenotyping of the same isolate can sometimes result in the resistance status switching (Gordon et al. 2014). Further, by testing a binary resistance versus sensitive phenotype, power is lost in comparison to the true underlying phenotype, which is more accurately represented on a continuous scale. Resistance measured by minimum inhibitory concentration (MIC) would be a more precise phenotype, as seen by Alam et al. (2014), however we were limited by the results obtained from the routine laboratory.

Given that the isolates were collected by a routine clinical laboratory, we were also limited by the proportion of phenotypically resistant isolates collected. The prevalence of resistance for some drugs collected by routine clinical work may be low. The power to detect causal variants depends on their frequency in the population, which will be reflected in the frequencies of phenotypic resistance and sensitivity. In the case of antimicrobial resistance, however, penetrance is high for many resistance-conferring variants which may mitigate this problem (Walker et al. 2015).

3.1.5.2 Genome-wide LD

As discussed in the introduction Chapter 1.2.4, we expect to encounter problems with applying GWAS to bacteria due to their clonal reproduction and typically limited rates of recombination which results in genome-wide linkage disequilibrium (Falush & Bowden

2006). The extent to which this will impact analyses will depend on the species, as bacteria can range from entirely clonal (e.g. *Mycobacterium tuberculosis*, Achtman 2008) to highly recombining (e.g. *Helicobacter pylori* Vos & Didelot 2009). Larger sample sizes, populations with elevated levels of recombination and homoplasy would all help to alleviate this problem. We wished to test the feasibility of bacterial GWAS in the knowledge of these difficulties across a taxonomically diverse range of bacteria.

3.1.5.3 Capturing the accessory genome

As well as point mutations generating diversity, bacterial populations also have large accessory genomes (Lapierre & Gogarten 2009; Jackson et al. 2011). This is quantified using the concept of the pan-genome, the set of all genes present in a group of organisms (Lapierre & Gogarten 2009) and it is important that the accessory genome is captured in association studies as they can contain important antimicrobial resistance determinants and virulence factors (Jackson et al. 2011). SNP-based association studies only capture the core genome as they rely on sequencing reads being mapped to a common reference genome. However, investigating SNPs alongside a pangenome analysis and a kmer-based analysis where either sequencing reads or *de novo* assembled contigs are split into sequences of length k, has the potential to target all variation present. The kmer approach was first applied to bacterial GWAS by Sheppard et al. (2013) where they investigated host association in *Campylobacter jejuni* by testing 20bp, 30bp and 40bp kmers for association with host colonisation, finding consistent results across the different kmer lengths. Therefore in addition to testing SNPs and gene presence/absence for association, we tested 31bp kmers counted from sequencing reads, aiming to capture SNPs in both the core and accessory genome, the presence of mobile accessory genes and the presence of short insertions and deletions.

3.1.5.4 Testing variants individually

Although not a problem unique to bacteria, a potential challenge with GWAS is that many phenotypes of interest are polygenic. We know that many antimicrobial resistance phenotypes are determined by more than one genetic variant (Stoesser et al. 2013; Gordon et al. 2014; Walker et al. 2015). Typically multiple variants or genes can confer resistance to a particular antibiotic but just one is required to confer resistance (Stoesser et al. 2013; Gordon et al. 2014; Walker et al. 2015), however some resistance-conferring variants are additive, and higher MICs are achieved when multiple variants occur together. For example, high-level resistance to fluoroquinolones in *E. coli* requires the accumulation of multiple mutations (Marcusson, Frimodt-Møller & Hughes 2009). Often multiple genetic mechanisms can also be involved, for example in *S.aureus* intermediate level vancomycin resistance is conferred by mutations arising in the pathogen, however full resistance is achieved by carriage of the plasmid-borne transposon Tn1546 originally acquired from vancomycin-resistant *Enterococcus faecalis* (Gardete & Tomasz 2014). The potential existence of multiple variants in the population conferring the same resistance phenotype may mean the power to identify any one of them as significantly associated with resistance is reduced. We aimed to explore whether despite this, if GWAS can identify any one of the resistance-conferring variants as the most significantly associated variant in each of our studies. Testing for association using Linear Mixed Models (LMM) which can increase power in the presence of polygenic effects may assist in finding these variants (Yang et al. 2014).

3.2 Chapter aims

Our aim, like several of the papers which have been published since the outset of this thesis reviewed in Chapter 1.2.5, was to test empirically the feasibility of bacterial GWAS in light of the special challenges bacteria pose in terms of strong population

structure and genome-wide LD. Like the other studies, we focused on antimicrobial resistance because of the availability of detailed phenotyping, the keen interest in antimicrobial resistance prediction, and the special status of antimicrobial resistance as one of the best-understood bacterial phenotypes. In this study we conducted 26 GWAS of 17 antimicrobials across the four major pathogens *M. tuberculosis, S. aureus*, *E. coli* and *K. pneumoniae* (Stoesser et al. 2013; Gordon et al. 2014; Walker et al. 2015). The four species are taxonomically diverse and range from completely clonal to highly recombining (Pérez-Losada et al. 2006; Dos Vultos et al. 2008; Vos & Didelot 2009), giving us insights into the feasibility of GWAS across a gamut of bacterial species.

Besides testing broadly whether genuine antimicrobial resistance-conferring variants were detected as the top hits in bacterial GWAS, we employed detailed analysis of fusidic acid resistance in *S. aureus* and simulations to investigate the potential problem that controlling for population structure in bacterial GWAS risks substantial loss of statistical power because the majority of variants tend to be population-stratified in bacteria, as discussed in Chapter 1.2.4. We introduced a new method that tests for associations not only with individual genetic variants, but also with particular strains, which we show can be characterised using PCA. We demonstrated that this test for lineage effects retrieves much of the power lost to controlling population structure in bacterial GWAS. We discuss the caveats of testing for lineage-level associations and the implications of our results for bacterial GWAS in general.

3.3 Methods

3.3.1 Sampling frames

We investigated 241 *Escherichia coli* and 176 *Klebsiella pneumoniae* UK clinical isolates sampled by Nicole Stoesser along with 992 *Staphylococcus aureus* and 1,735 *Mycobacterium tuberculosis* isolates previously reported (Gordon et al. 2014; Walker et

Species	# Isolates	GWAS studies	Reference genome for mapping	Number of variants				
				Bialellic SNPs	Triallelic SNPs	Tetra-allelic SNPs	Kmers	Gene clusters
S. aureus	992	Ciprofloxacin, Erythromycin, Fusidic acid, Gentamicin, Penicillin, Methicillin, Tetracycline, Rifampicin	MRSA252 (BX571856.1)	264604	14731	519	24154606	13881
	323	Trimethoprim	MRSA252 (BX571856.1)	196996	8712	269	15840354	10261
E. coli	241	β-lactam: Ampicillin, Ceftazidime, Cefuroxime, Ceftriaxone; Quinolone: Ciprofloxacin; Amincoglycoside: Gentamicin, Tobramycin	CFT073 (AE014075.1)	417645	25298	1287	39918870	23502
K. pneumoniae	176	β-lactam: Ampicillin, Ceftazidime, Cefuroxime, Ceftriaxone; Quinolone: Ciprofloxacin; Amincoglycoside: Gentamicin, Tobramycin	MGH 78578 (CP000647.1)	654425	63639	5029	53816250	21382
M. tuberculosis	1630	Ethambutol	H37Rv (NC_000962.2)	107480	954	8	15680376	-
	1709	Isoniazid	H37Rv (NC_000962.2)	110400	1020	10	15941713	-
	1707	Pyrazidamide	H37Rv (NC_000962.2)	110162	1012	10	15963479	-
	1573	Rifampicin	H37Rv (NC_000962.2)	101968	864	8	15554437	-

Table 3.1 Number of SNP, kmer and gene presence/absence variants for all GWAS studies

al. 2015). All isolates had been tested for resistance to multiple antimicrobials using routine clinical laboratory protocols (Table 3.1). DNA was extracted and sequenced as previously described (Stoesser et al. 2013; Gordon et al. 2014; Walker et al. 2015) and SNPs were called using standard methods, using Stampy to map reads to the reference strains CFT073, MGH 78578, H37Rv and MRSA252 for *E. coli*, *K. pneumoniae*, *M. tuberculosis* and *S. aureus*, respectively. The distributions of the biallelic SNP frequencies for the four species are provided in Table 3.2. The total numbers of variable SNPs, genes and kmers are provided in Table 3.1 and the Bonferroni-corrected significance thresholds are provided in Table 3.3.

Species	Minor allele frequency (%) 0 – 1%	1 – 2%	2 – 5%	5 – 10%	10 – 20%	20 – 50%
E. coli	26.5	9.0	11.4	10.7	12.9	29.5
K. pneumoniae	21.2	11.3	17.1	17.1	16.5	16.8
M. tuberculosis	93.9	1.8	1.6	1.0	1.1	0.7
S. aureus	49.2	6.9	6.5	9.7	5.2	22.5

Table 3.2 Distribution of minor allele frequencies for biallelic SNPs for the four species investigated. Originally published in Earle et al. *Nat Microbiol*. 2016.

Species	Bonferroni-corrected significance threshold ($-\log_{10}P$) SNPs and genes	Kmers	Lineages
Escherichia coli	6.5	7.3	3.7
Klebsiella pneumoniae	6.6	7.3	3.5
Mycobacterium tuberculosis	5.0	7.6	4.5
Staphylococcus aureus			
Ciprofloxacin, erythromycin, fusidic acid, gentamicin, penicillin, methicillin, tetracycline and rifampicin	6.1	7.3	4.3
Trimethoprim	5.9	6.7	3.8

Table 3.3 Bonferroni-corrected significance thresholds for combined SNPs and genes and for kmers. Bonferroni-correction used to control the FWER to 0.05. Originally published in Earle et al. *Nat Microbiol*. 2016.

3.4 Results

3.4.1 Prevalence of resistance

The prevalence of resistance for some drugs studied in this chapter was very low, as seen in Figure 3.1. The lowest proportion of resistance was that of rifampicin resistance in *S. aureus* at only 0.8% of the sampling frame, followed closely by gentamicin resistance in *S. aureus* and ethambutol and pyrazinamide resistance in *M. tuberculosis* at 1.1%, 2.5% and 2.6%, respectively (Figure 3.1). Therefore, power to detect causal mechanisms for these antimicrobials may be low.

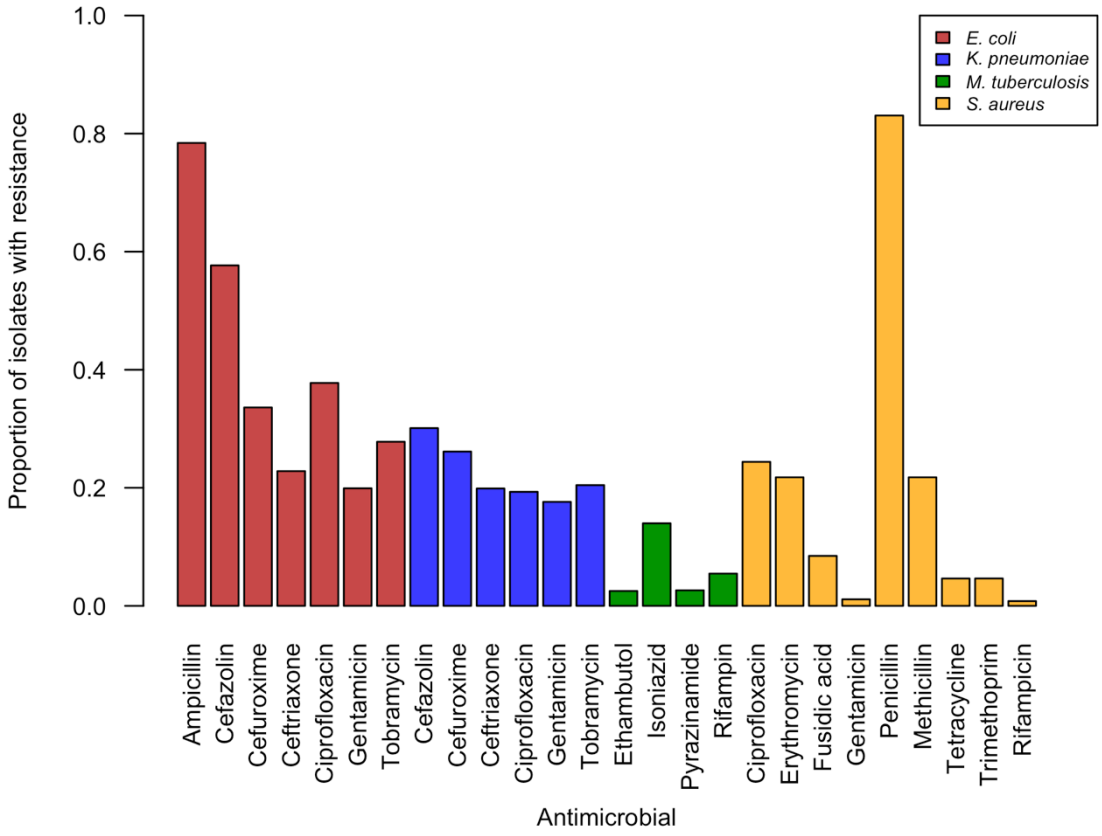

Figure 3.1 Distribution of resistance phenotypes across the four species.

3.4.2 Imputing missing SNP calls using ClonalFrameML was highly accurate

As discussed in the Methods Chapter 2.3, SNP imputation is necessary for GWAS. The error prone nature of Illumina sequencing means that strict SNP-calling thresholds must be applied to ensure reliable mapping-based SNP calls. Restricting analyses to sites with no missing data could potentially drastically reduce the number of variants available for association tests, and comparing loci tested using different isolates violates the principle of basing model comparison on identical data. Imputation is used across human GWAS GWAS (Scheet & Stephens 2006; Browning & Browning 2009; Purcell et al. 2007; Howie, Donnelly & Marchini 2009; Li et al. 2010) and we tested the accuracy of imputing missing and uncalled bases by simulations using two different approaches, ClonalFrameML (Didelot & Wilson 2015) and Beagle (Browning & Browning 2009).

One hundred sequences were randomly sampled from each of the four GWAS

data sets in order to simulate data to test imputation accuracy. Samples were taken from the phylogeny in order to maintain phylogenetic diversity. RAxML (Stamatakis 2014) was used to estimate maximum likelihood phylogenies for the 100 sequences of each species as described in the Methods Chapter 2.2 (Stamatakis 2014). For each data set of 100 sequences the empirical distribution of Ns across individuals per site in the alignment was determined, and these were then sampled with replacement and imposed on randomly chosen variable sites in the alignment to reintroduce Ns. Imputation was then performed on each simulated dataset using both ClonalFrameML and Beagle to test imputation accuracy. Accuracy was measured as the proportion of simulated ambiguities correctly imputed, as well as the correlation between the truth and imputed data both at only sites simulated as Ns plus all sites. Overall, ClonalFrameML was more accurate than Beagle (Figure 3.2-Figure 3.4;Table 3.4). As expected, accuracy measured as the proportion of simulated ambiguities dropped with the % of Ns per site for both methods, but accuracy was very high overall, with Beagle achieving 69.1% correctly imputed sites in *S. aureus* even when 90% or more bases at a site were masked with Ns (Figure 3.2A). The lowest accuracy by ClonalFrameML was 77.5% correctly imputed sites at sites with between 90-100% Ns in *S. aureus* and overall accuracy was 93.2%; 91.5% in *E. coli*, 92.0% in *K. pneumoniae*, 98.2% in *M. tuberculosis* and 91.3% in *S. aureus*. Again, as expected imputation accuracy increased with increasing allele frequency with the exception of *K. pneumoniae* which had a lower proportion of high frequency SNPs than *E. coli* and *S. aureus* (Figure 3.2; Table 3.4). Similar patterns were seen when computing accuracy using Pearson correlation (Figure 3.3; Figure 3.4). Due to the higher imputation accuracy of ClonalFrameML over Beagle, ClonalFrameML was used for all analyses.

% Ns at the site	% Correct imputation		Correlation (all individuals)		MAF	% Correct imputation		Correlation (all individuals)	
	CFML	Beagle	CFML	Beagle		CFML	Beagle	CFML	Beagle
Escherichia coli									
0-10%	96.5	97.6	0.929 (0.993)	0.958 (0.994)	0.00-0.01	93.2	91.9	0.732 (0.942)	0.703 (0.932)
10-20%	94.9	95.6	0.845 (0.962)	0.885 (0.965)	0.01-0.05	90	88.1	0.754 (0.931)	0.761 (0.925)
20-30%	94	94.3	0.814 (0.935)	0.841 (0.938)	0.05-0.10	91.8	90.5	0.821 (0.96)	0.84 (0.96)
30-40%	93.7	93.9	0.799 (0.907)	0.822 (0.91)	0.10-0.15	91.9	91.6	0.839 (0.968)	0.866 (0.969)
40-50%	91.9	90.9	0.767 (0.87)	0.761 (0.86)	0.15-0.20	92.9	92.8	0.879 (0.98)	0.894 (0.98)
50-60%	93	92.1	0.769 (0.855)	0.773 (0.844)	0.20-0.25	91.9	92.1	0.865 (0.98)	0.891 (0.981)
60-70%	91.7	89	0.733 (0.814)	0.712 (0.779)	0.25-0.30	92.3	93.1	0.874 (0.983)	0.903 (0.985)
70-80%	90.1	87.7	0.696 (0.752)	0.665 (0.717)	0.30-0.35	93.3	93.4	0.892 (0.988)	0.916 (0.988)
80-90%	87.1	82.8	0.647 (0.668)	0.575 (0.599)	0.35-0.40	94.6	94.8	0.911 (0.992)	0.93 (0.992)
90-100%	81.7	73.4	0.584 (0.564)	0.409 (0.426)	0.40-0.45	96.5	96.9	0.952 (0.997)	0.96 (0.997)
					0.45-0.50	96.1	97	0.953 (0.998)	0.97 (0.998)
Klebsiella pneumoniae									
0-10%	95.8	94.3	0.962 (0.99)	0.944 (0.987)	0.00-0.01	92.1	88.6	0.777 (0.958)	0.677 (0.939)
10-20%	92	90.1	0.895 (0.94)	0.86 (0.923)	0.01-0.05	92.3	89.6	0.89 (0.964)	0.833 (0.952)
20-30%	91.3	89.2	0.828 (0.9)	0.789 (0.87)	0.05-0.10	92.3	91.2	0.918 (0.977)	0.906 (0.974)
30-40%	95.6	94.6	0.857 (0.916)	0.82 (0.893)	0.10-0.15	94.9	94.3	0.935 (0.981)	0.914 (0.979)
40-50%	95.8	95.1	0.835 (0.904)	0.803 (0.883)	0.15-0.20	94.3	92.9	0.959 (0.987)	0.931 (0.985)
50-60%	96.6	96	0.851 (0.903)	0.813 (0.881)	0.20-0.25	96	94.5	0.966 (0.994)	0.947 (0.993)
60-70%	95.6	94.3	0.824 (0.859)	0.767 (0.814)	0.25-0.30	84.5	87.2	0.812 (0.979)	0.858 (0.982)
70-80%	94	90.6	0.791 (0.814)	0.689 (0.73)	0.30-0.35	82.1	85.7	0.782 (0.975)	0.841 (0.979)
80-90%	79.7	76.1	0.586 (0.633)	0.485 (0.546)	0.35-0.40	82.2	85.9	0.783 (0.979)	0.852 (0.983)
90-100%	83.9	71.6	0.65 (0.631)	0.419 (0.416)	0.40-0.45	86	89.3	0.819 (0.988)	0.883 (0.99)
					0.45-0.50	88.4	91.7	0.869 (0.994)	0.923 (0.995)

Table 3.4 Summary of the simulations testing imputation accuracy using ClonalFrameML and Beagle. Results across categories of percentage of Ns to impute plus minor allele frequencies are shown. Accuracy has been defined here by the proportion of correctly imputed sites per stratified group and also by correlation between truth and imputed data, both only including individuals per site simulated as Ns and including all individuals as shown in brackets. Overall, imputation by ClonalFrameML was more accurate. Originally published in part in Earle et al. *Nat Microbiol*. 2016. (*Continued on the next page*).

% Ns at the site	% Correct imputation		Correlation (all individuals)		MAF	% Correct imputation		Correlation (all individuals)	
	CFML	Beagle	CFML	Beagle		CFML	Beagle	CFML	Beagle
Mycobacterium tuberculosis									
0-10%	99.8	99	0.998 (0.999)	0.946 (0.997)	0.00-0.01	96.8	94.9	0.841 (0.989)	0.382 (0.981)
10-20%	99.8	99	0.995 (0.993)	0.916 (0.982)	0.01-0.05	98.3	90.2	0.967 (0.995)	0.789 (0.969)
20-30%	99.8	98.8	0.994 (0.989)	0.901 (0.969)	0.05-0.10	99.8	93.8	0.993 (0.999)	0.847 (0.986)
30-40%	99.7	98.3	0.979 (0.984)	0.844 (0.953)	0.10-0.15	99.6	94.3	0.998 (1.0)	0.875 (0.989)
40-50%	99.7	98.4	0.991 (0.975)	0.825 (0.94)	0.15-0.20	100	96.7	1.0 (1.0)	0.912 (0.988)
50-60%	99.6	98.4	0.959 (0.964)	0.853 (0.914)	0.20-0.25	99.7	96.2	0.998 (1.0)	0.975 (0.997)
60-70%	99	96.2	0.97 (0.946)	0.737 (0.877)	0.25-0.30	99.9	96.3	0.995 (1.0)	0.938 (0.994)
70-80%	98.4	95.3	0.948 (0.917)	0.694 (0.814)	0.30-0.35	100,0	96.4	1.0 (1.0)	0.952 (0.997)
80-90%	97.8	91.5	0.919 (0.881)	0.631 (0.715)	0.35-0.40	100	96	1.0 (1.0)	0.929 (0.998)
90-100%	88.3	78.1	0.776 (0.739)	0.269 (0.495)	0.40-0.45	100	98.5	1.0 (1.0)	1.0 (1.0)
					0.45-0.50	NA	NA	NA	NA
Staphylococcus aureus									
0-10%	98	97.9	0.968 (0.995)	0.966 (0.994)	0.00-0.01	92.5	89.8	0.77 (0.968)	0.706 (0.955)
10-20%	97.7	97.1	0.935 (0.982)	0.912 (0.978)	0.01-0.05	91.4	85.9	0.877 (0.957)	0.822 (0.935)
20-30%	94.9	94.4	0.888 (0.955)	0.86 (0.948)	0.05-0.10	92.7	88.8	0.875 (0.978)	0.835 (0.97)
30-40%	91.7	88.2	0.824 (0.917)	0.749 (0.887)	0.10-0.15	91.5	87.8	0.879 (0.971)	0.87 (0.963)
40-50%	93.7	89.6	0.845 (0.914)	0.765 (0.869)	0.15-0.20	92.9	90.5	0.915 (0.979)	0.905 (0.974)
50-60%	90.7	84.4	0.766 (0.869)	0.669 (0.794)	0.20-0.25	92.9	90.8	0.948 (0.987)	0.915 (0.983)
60-70%	89	82.3	0.757 (0.825)	0.644 (0.728)	0.25-0.30	94.3	91.4	0.951 (0.988)	0.926 (0.984)
70-80%	91.1	84.6	0.777 (0.826)	0.674 (0.71)	0.30-0.35	95.3	93.6	0.953 (0.993)	0.93 (0.991)
80-90%	88.4	82.9	0.707 (0.757)	0.594 (0.639)	0.35-0.40	95.2	95.4	0.959 (0.994)	0.953 (0.994)
90-100%	77.5	69.1	0.571 (0.593)	0.396 (0.405)	0.40-0.45	95.9	96.2	0.949 (0.997)	0.952 (0.996)
					0.45-0.50	95	96.4	0.955 (0.997)	0.965 (0.998)

Table 3.4 (Contd.) Summary of the simulations testing imputation accuracy using ClonalFrameML and Beagle. Results across categories of percentage of Ns to impute plus minor allele frequencies are shown. Accuracy has been defined here by the proportion of correctly imputed sites per stratified group and also by correlation between truth and imputed data, both only including individuals per site simulated as Ns and including all individuals as shown in brackets. Overall, imputation by ClonalFrameML was more accurate. Originally published in part in Earle et al. *Nat Microbiol*. 2016.

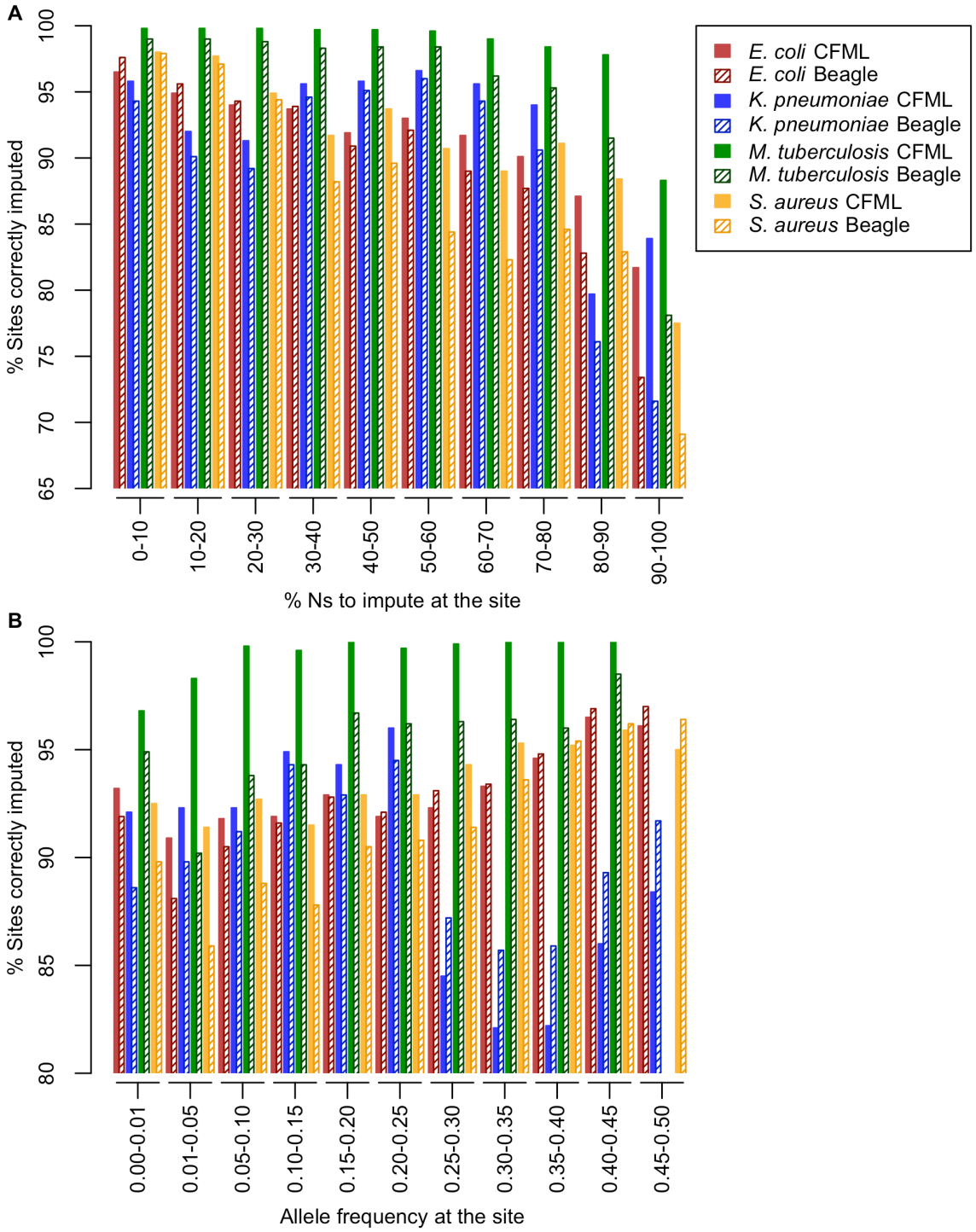

Figure 3.2 Imputation accuracy for ClonalFrameML and Beagle per species. A Accuracy per category of percentage of Ns to impute at the site. **B** Accuracy across minor allele frequency categories. Accuracy has been defined here by the proportion of correctly imputed sites per stratified group. Overall, imputation by ClonalFrameML was more accurate than Beagle.

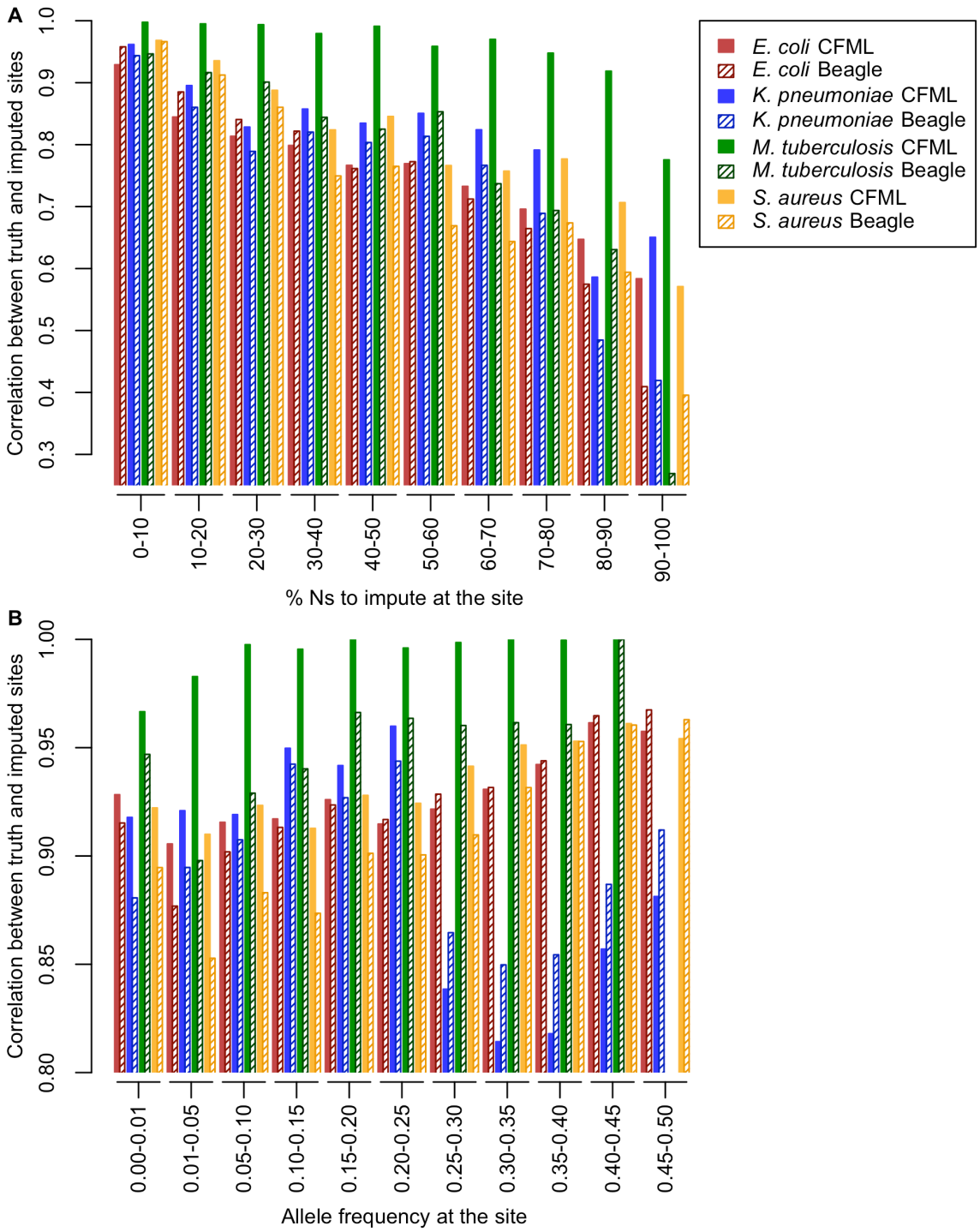

Figure 3.3 Imputation correlations for ClonalFrameML and Beagle per species for positions simulated as Ns. A Correlations between truth and imputed data for sites simulated as Ns per category of percentage of Ns to impute at the site. **B** Correlations between truth and imputed data for sites simulated as Ns across minor allele frequency categories. Overall, imputation by ClonalFrameML was more accurate than Beagle.

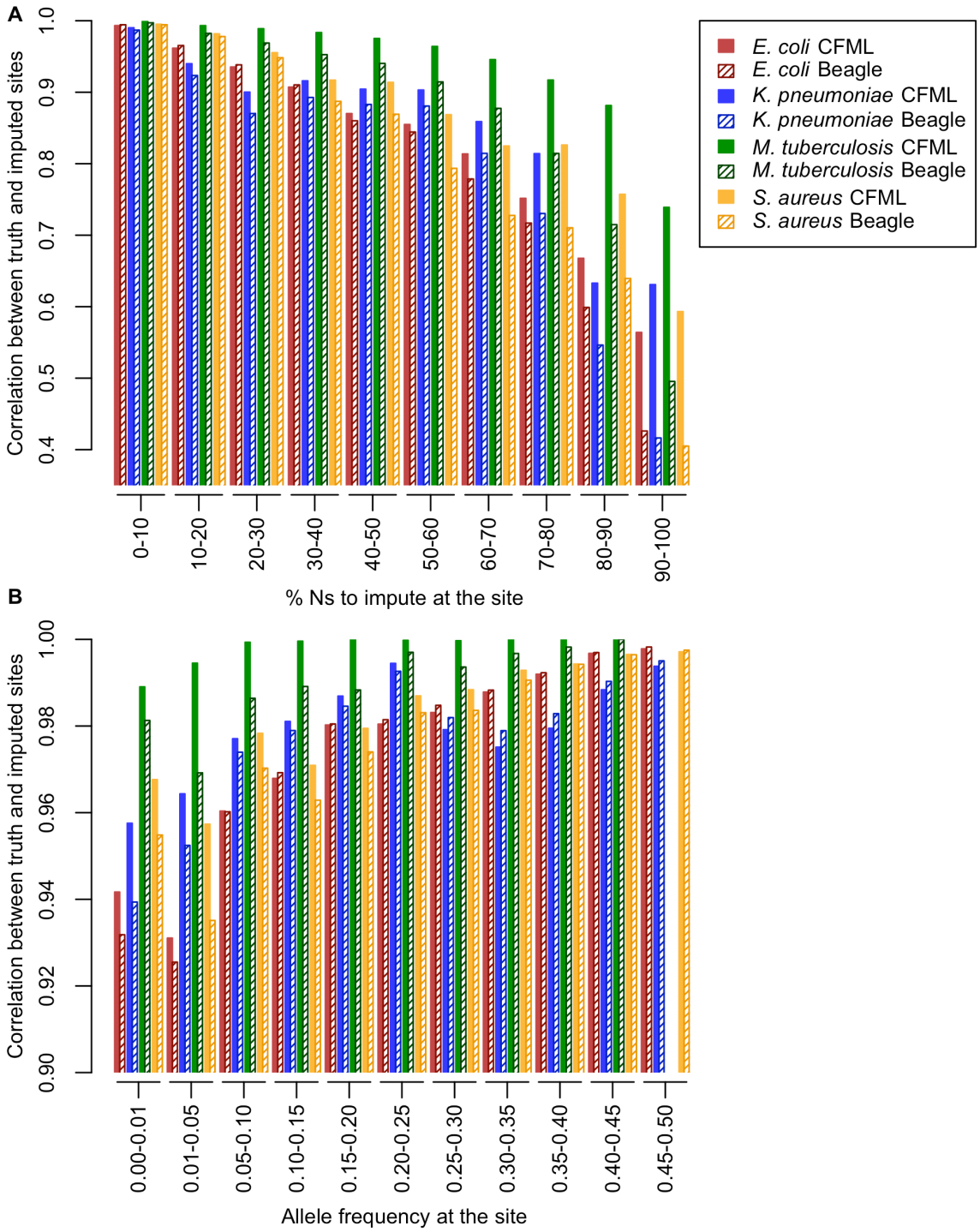

Figure 3.4 Imputation correlations for ClonalFrameML and Beagle per species for all positions. A Correlations between truth and imputed data for all sites (regardless of whether they were imputed) per category of percentage of Ns to impute at the site. **B** Correlations between truth and imputed data for all sites (regardless of whether they were imputed) across minor allele frequency categories. Overall, imputation by ClonalFrameML was more accurate than Beagle.

3.4.3 A case study of fusidic acid resistance in *S. aureus*: failure to control for population structure resulted in a large number of false positives

Fusidic acid is an antibiotic commonly used to treat *S. aureus* and other gram-positive infections, typically skin and soft tissue infections and it can be taken topically or systemically (Dobie & Gray 2004). Fusidic acid interferes with bacterial protein synthesis by preventing the release of elongation factor G (EF-G) from the ribosome post translocation (Tanaka, Kinoshita & Masukawa 1968; Gao et al. 2009). Fusidic acid resistance is encoded by three different genetic elements: certain substitutions in the chromosomal gene *fusA*, the gene encoding EF-G, which prevent fusidic acid from binding to EF-G (Besier et al. 2003), and the presence of accessory genes *fusB* or *fusC* which bind to EF-G and prevent translation from being inhibited by fusidic acid (O'Neill & John Chopra 2006; O'Neill et al. 2007). When *fusC* is present in *S. aureus* it is always found on SCC-type elements in isolates sampled globally, indicating that SCC-mediated horizontal transfer is the principal mechanism enabling the spread of *fusC* (Baines et al. 2016). The dissemination of *fusB* is due to a different mechanism, as it is typically transferred via plasmids such as pUB101, pUB102 (O'Brien et al. 2002; Monecke et al. 2006; Stegger et al. 2012). *fusB* has also been found chromosomally, albeit causing comparably lower minimum inhibitory concentrations (MICs) possibly due to a much lower gene copy number than when present on plasmids which are typically present in multiple copies (O'Neill et al. 2004).

We investigated associations between fusidic acid resistance and the presence or absence of short 31bp haplotypes or kmers in *S. aureus* in order to test the feasibility of the application of GWAS in bacteria. By testing kmers, we captured resistance encoded by substitutions in the core genome, the presence of mobile accessory genes and insertions and deletions. We first tested each kmer for association with fusidic acid

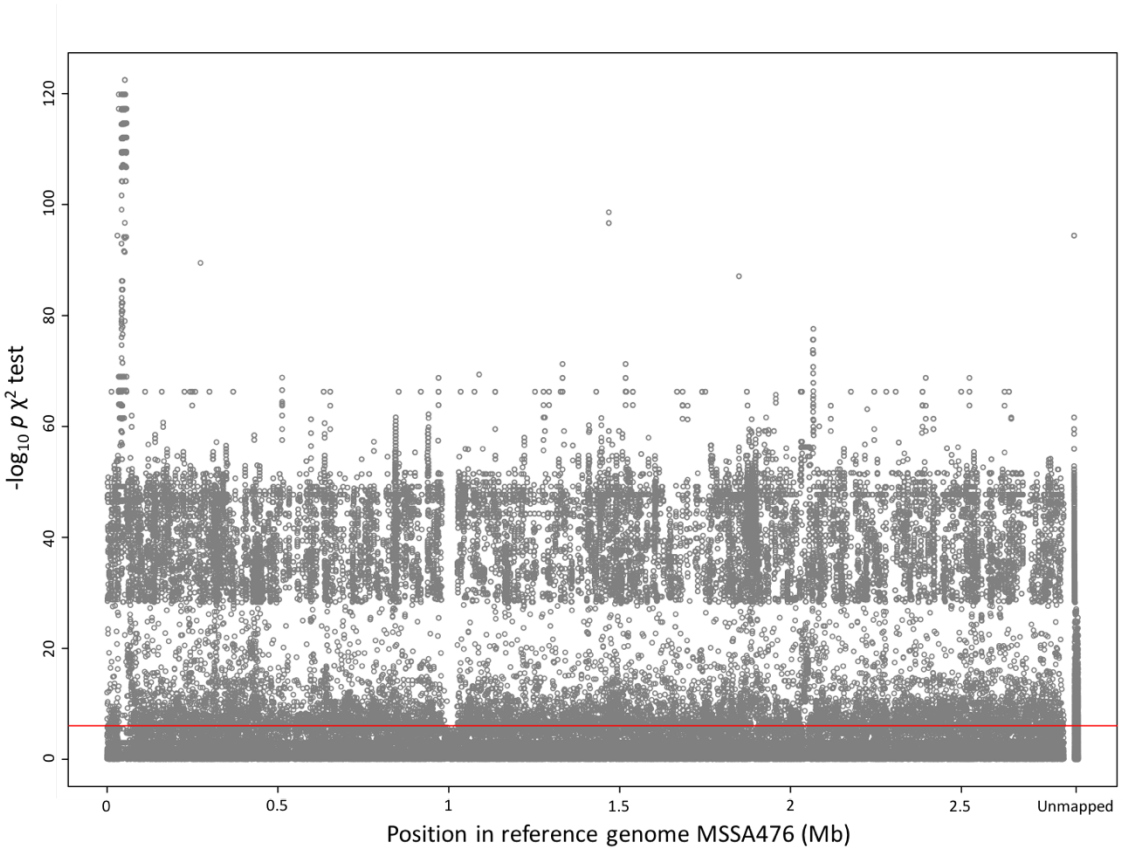

Figure 3.5 Significance of association of the presence or absence of 31bp *S. aureus* kmers with fusidic acid resistance by χ^2 test. All kmers above $-\log_{10} P = 30$ are plotted plus a selection of those below 30. The red line denotes the Bonferroni-corrected significance threshold accounting for the number of kmer patterns tested. Significance is hugely inflated, but the most significant region correctly identifies the region within SCC carrying the resistance determinant *fusC*.

resistance using a χ^2 test and found that kmers in the gene SAS0040 (MSSA476, accession NC_002953), which is adjacent to and linked to the presence of *fusC*, showed the strongest genome-wide association ($P = 10^{-122}$, odds ratio (OR) = ∞, Figure 3.5). Kmers mapping to *fusC* were also highly significant at $P = 10^{-120}$ and were only found in resistant isolates, resulting in an infinite OR. *fusC* is located within the SCC region of *S. aureus*, and other kmers within this region showed the strongest association signals along with *fusC*. The most significant resistance causing SNP in *fusA* was much less significant than *fusC* at $P = 10^{-42}$ (OR=0.004). This was still highly significant and far above the Bonferroni-corrected significance threshold of $P = 10^{-7}$, however it is in the 29th percentile of results, thus was lost within the large number of more significant

associations.

The Manhattan plot revealed bands of variants in genome-wide LD, for example at around $P = 10^{-65}$. These bands corresponded to particular lineages revealing the presence of population structure, a known confounder in association studies (Marchini et al. 2004) (Figure 3.5). Therefore, despite the χ^2 test identifying the correct region of the genome as significantly associated with fusidic acid resistance, 208,092 other variants genome-wide were also incorrectly found to be significantly associated (Figure 3.5). As the mechanisms of fusidic acid resistance are well understood, we calculated an approximate false positive rate (FPR) under the assumption that all kmers which did not map to the *fusB* containing plasmid pUB101 (AY373761.1), *fusC* containing SCC_{476} (NC_002953.3), or *fusA* (NC_002952.2) were false positives, resulting in an FPR of 89.2%.

3.4.4 Accounting for population structure resulted in widespread loss of significance because many bacterial variants are population stratified

We controlled for population structure using Linear Mixed Models (LMM) implemented in the software GEMMA (Zhou & Stephens 2012). Doing so reduced the significance of the most significant kmers within *fusC* from $P = 10^{-119}$ to $P = 10^{-39}$. However, kmers which captured resistance-conferring variants in *fusA* greatly increased in significance from $P = 10^{-42}$ by the χ^2 test to $P = 10^{-157}$ by LMM. This could be because these variants were low frequency, and it has been shown that rare variants can have inflated significance even after applying LMMs (Mathieson & McVean 2012), but also LMMs can improve power in the presence of polygenic effects (Yang et al. 2014). Importantly, the benefit of this widespread loss of significance due to controlling for population structure was the greatly reduced number of false positives. Using LMM to control for population structure resulted in an eight-fold reduction in the total number of estimated

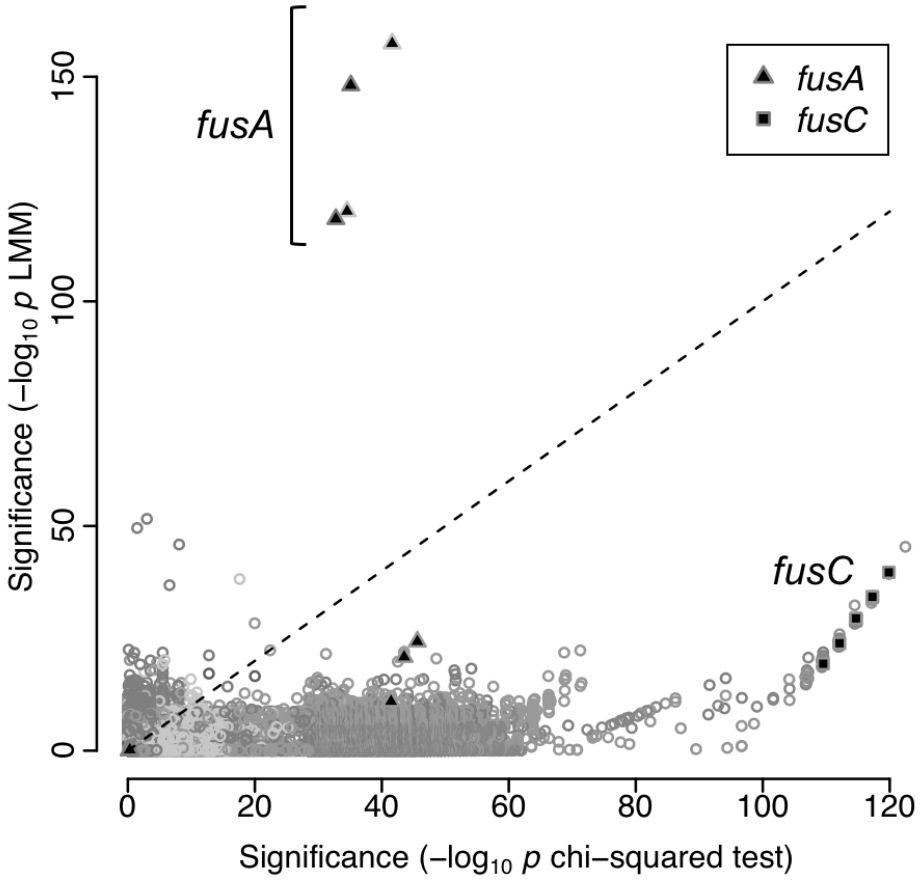

Figure 3.6 The effect of controlling for population structure using LMM, in comparison to a χ^2 test, on the significance of associations between the presence or absence of 31bp kmers and fusidic acid resistance in *S. aureus*. The 200,000 most significant kmers prior to control for population structure plus a random 200,000 are plotted. The significance of kmers capturing the resistance determinant *fusC* are greatly reduced in significance but kmers capturing resistance-conferring variants in *fusA* are propelled to greater significance.

false positives, although the FPR was still reasonably high at 51.9%. As expected, the FPR was better controlled in simulations. Simulations performed by Daniel Wilson revealed that testing for association by LMM as opposed to a logistic regression for SNP based analyses in the four species reduced the number of false positives by 30-fold (*K. pneumoniae*) to 3600-fold (*S. aureus*).

Although the most significant kmer was in *fusA*, the minor allele captured by the kmer was present in just one isolate and therefore explained only a small proportion of resistance in comparison to the presence of *fusC*. Of the 84 resistant isolates, 44 contained *fusC* determined by using BLAST to query the assemblies of all resistant isolates, and 24

contained a resistance determining SNP in *fusA* as determined by examination of the mapped data. Thus *fusC* explained almost twice as much resistance as *fusA*. In order to explain a large proportion of resistance, the kmers capturing *fusC* are of most interest.

The gene SAS0040 (MSSA476, accession NC_002953) immediately upstream of *fusC* remained more significant than *fusC* after controlling for population structure ($P = 10^{-45}$). Baines et al. (2016) found that *fusC* was invariably found within a conserved ~1.3kb region, sharing an upstream putative coding sequence (CDS), SAS0040. However, SAS0040 is annotated as a hypothetical protein, and Baines and colleagues could not assign any putative functional roles to the resultant protein using either BLASTX or Pfam searches. Nevertheless, given that SAS0040 is consistently present with *fusC*, and that we found it as the most significant association in the region both before and after controlling for population structure, the protein may be involved in regulation of transcription of *fusC* and should be investigated futher (Baines et al. 2016).

The association between fusidic acid resistance and kmers capturing *fusC* was greatly reduced in significance after controlling for population structure, but did not suffer outright loss of significance (Figure 3.6). This loss of significance in a genuine resistance determinant is suggestive of a cost to controlling population structure, in terms of power to detect true associations. Although in this case there was no outright loss of significance for the causal kmers, in cases of weaker association there could be. Simulations performed by Daniel Wilson revealed that for phenotypes with much more modest effect sizes, with odds ratios of 3, controlling for population structure risks loss of genome-wide significance in causal variants and therefore the power to detect true associations at 59, 75, 99 and 99% of high-frequency causal variants in *M. tuberculosis* ($n = 1573$), *S. aureus* ($n = 992$), *E. coli* ($n = 241$) and *K. pneumoniae* ($n = 176$), respectively.

Power loss was greatest in *E. coli* and *K. pneumoniae*, where the sample size was low and the number of variants and unique variant patterns was high.

3.4.5 Population stratification of fusidic acid resistance explained the loss of significance when controlling for population structure

Given the widespread loss of power when controlling for population structure, it was important to understand the genetic basis underlying the loss of power. Observing the distribution of resistance and sensitivity to fusidic acid in the *S. aureus* population on the phylogeny of the data revealed that fusidic acid resistance was population stratified in the *S. aureus* dataset. In fact, the *fusC* gene was found exclusively in strains ST-1 and ST-8 by multi-locus sequence typing (MLST). This is something that has been observed previously, as Lannergård, Norström & Hughes (2009) found some *fusC* S. aureus strains sampled from Denmark to be closely related and possibly clonal based on their spa type, *fusC* and *fusA* sequences and context. Further studies of strains sampled from New Zealand and Denmark revealed that large proportions of fusidic acid resistance could be attributed to ST-5 *fusC* containing clones (McLaws et al. 2011; Baines et al. 2016). The ST-1 lineage is somewhat divergent from other lineages, differing by more than 8,000 SNPs from the rest of the *S. aureus* phylogeny (yellow branch in Figure 3.7). This is typical of the strong population structure observed in *S. aureus*, which implies that a large proportion of the total genetic variability of the species is population-stratified in the sense that it distinguishes lineages from one another. The substantial divergence between lineages makes it difficult to attribute lineage-to-lineage phenotypic variability to individual loci. This is why we lose power to detect locus effects when we control for population structure (i.e. stratification), and why, in the case of *fusC*, the significance of the true causal variant was reduced.

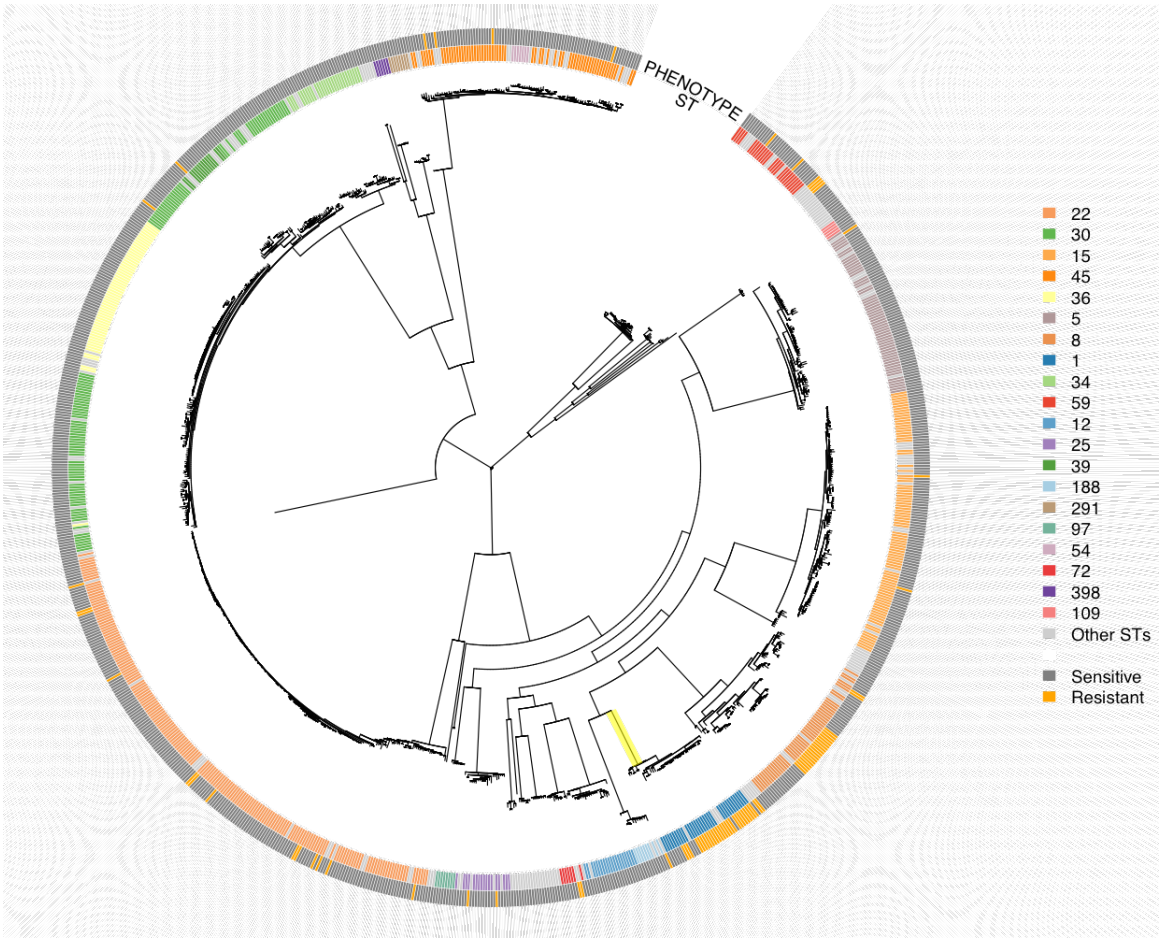

Figure 3.7 *Staphylococcus aureus* phylogeny. Maximum likelihood phylogeny built using RAxML based on biallelic SNPs and annotated with the 20 most common STs and resistance (orange) or sensitivity (grey) to fusidic acid. Resistance determined by *fusC* is found exclusively in ST-1 and ST-8. Bootstrap supports are shown in Appendix B Figure B.1.

One way to visualise the strong stratification of bacterial species is by quantifying the proportion of genetic variability the leading principal components (PCs) explain. The majority of genetic variance in the bacterial species we examined, unlike in the human chromosome, could be explained by the leading PCs. We found that the first ten principal components explained between 70-93% of genetic variation in the four bacterial species, in comparison to 27% in human chromosome one (Figure 3.8). Therefore, the cost of controlling the false positive rate through accounting for population structure in bacterial association studies is to risk a huge loss of power to detect genuine associations as the majority of loci will be population stratified.

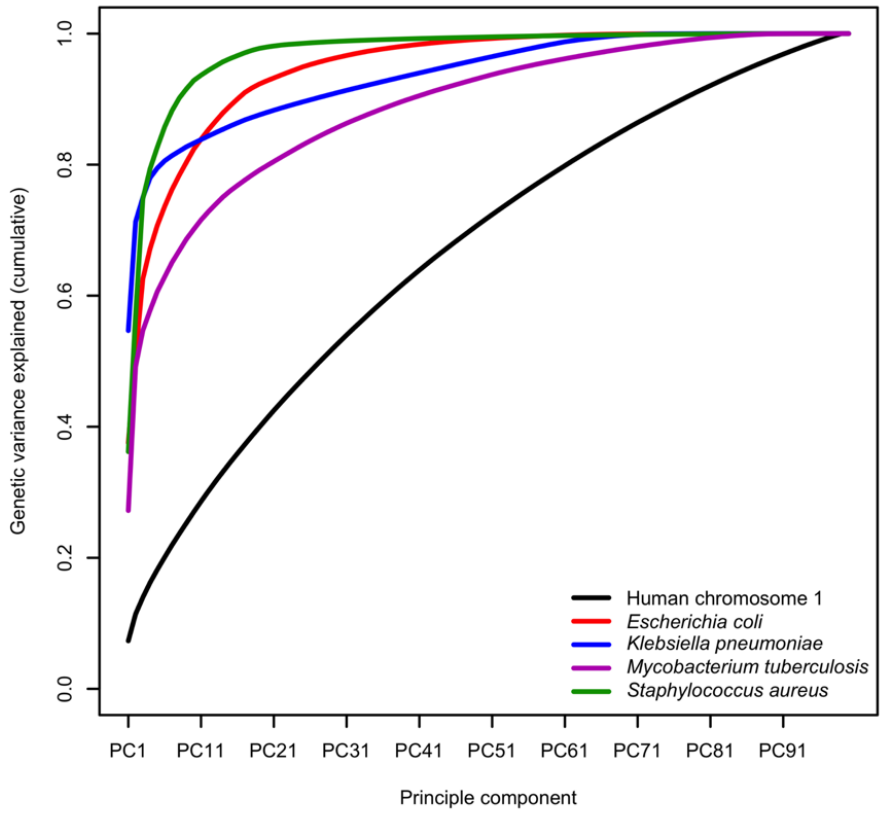

Figure 3.8 The cumulative proportion of genetic variability explained by leading principal components in the four bacterial species investigated plus human chromosome one, each calculated from a subsample of 100 genomes. Leading principal components in bacteria explain much larger genetic variability than in humans due to strong population structure and limited homologous recombination.

3.4.6 Regaining the lost power in bacterial GWAS by testing for lineage effects

We have developed a method to recover information which is discarded when controlling for population structure. In studies where there is limited power to detect locus specific associations due to population stratification, we can instead infer lineage-specific associations without sacrificing any power to detect locus-specific associations when able to do so. However, this approach is not risk-free because lineage-level differences in phenotype are vulnerable to artefacts caused by phenomena including differential sampling intensities of cases and controls in different lineages and differences in unmeasured confounders in different lineages.

3.4.6.1 Defining lineages by Principal Components

Principal Components Analysis (PCA) was performed on the biallelic imputed SNPs. We observed that despite large differences in recombination rates across the four species studied (Dos Vultos et al. 2008; Vos & Didelot 2009), that leading principal components tend to correspond to major lineages in bacterial phylogenies. Figure 3.9 depicts the leading five principal components on the phylogeny of the four species we investigated. One can see that the leading PCs trace paths through the deepest branches of the tree which reflects an underlying relationship between genealogical history and principal component analysis, as has been previously shown by McVean (2009).

In what follows, we define lineages in terms of PCs. Besides the empirical and theoretical motivation that PCs can be interpreted in terms of major lineages in the species' genealogical history, PCs have the advantage that they are uncorrelated, meaning they capture different axes of genetic variability in the population. We expect this to reduce the statistical uncertainty associated with estimating the relationship between phenotypic variability and each PC, which should in turn lead to a better powered test for lineage effects.

3.4.6.2 Wald test for lineage effects

In an LMM, every locus is included in the regression as a random effect, which is equivalent to including every principal component in the regression as a random effect because of the linear relationship between PCs and the genetic variability they capture (Astle & Balding 2009). Principal components are frequently used to control for population structure in human GWAS studies by including a number of leading principal components as fixed effects in a regression (Price et al. 2006). Therefore, the regression coefficients which are estimated for the principal components could be considered as

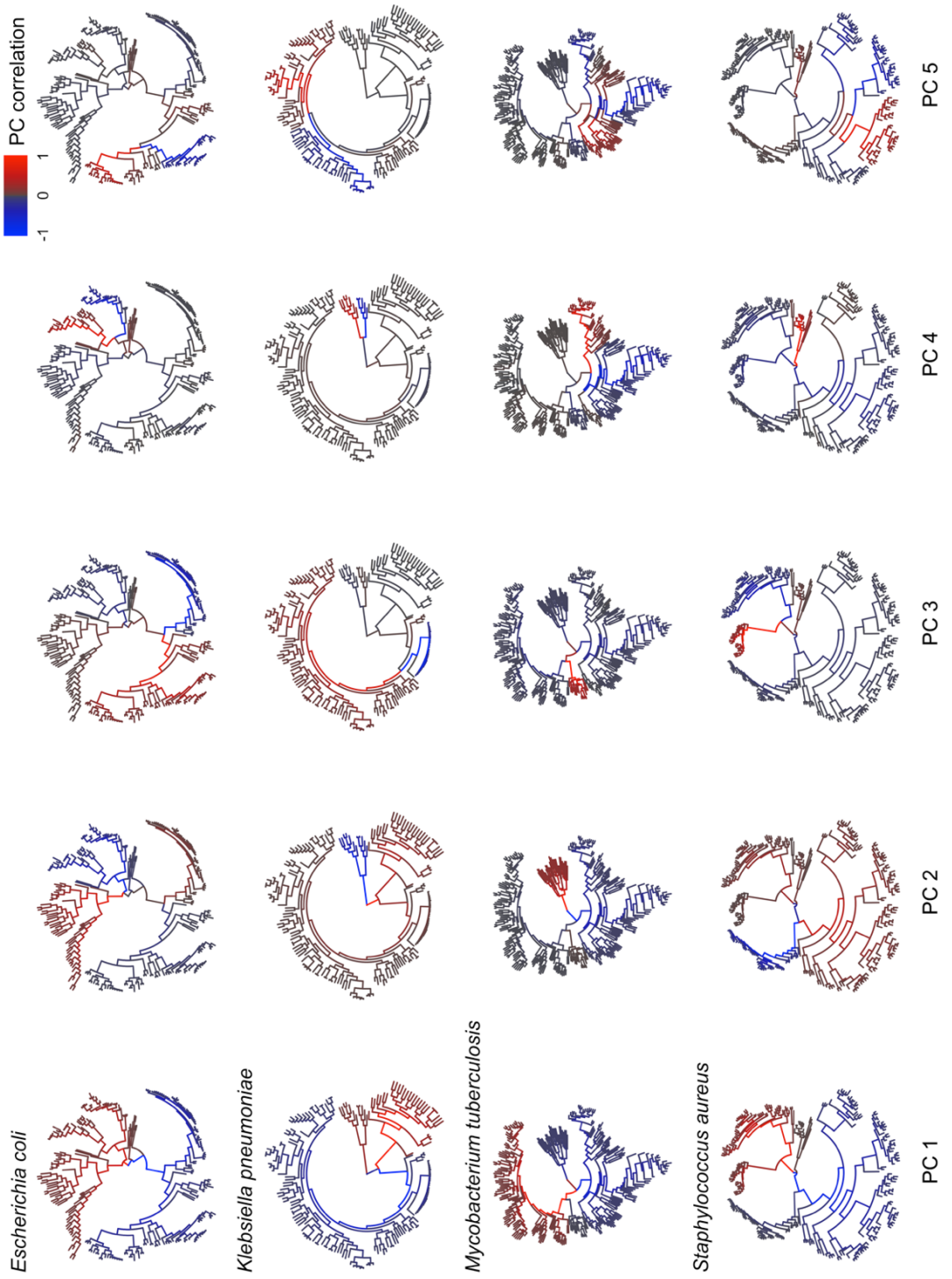

Figure 3.9 Leading principal components correspond to major lineages in bacterial phylogenies. For each of the four species investigated, 100 isolates were sampled from the phylogeny. Each branch was assigned a pattern based on how the branch split the isolates: 1 for all isolates one side of the branch, 0 for all isolates the other side of the branch. The colours reflect the correlation of the pattern with the projection of the isolates onto each PC. Positive versus negative correlations are shown in red vs. blue, with brighter colours representing stronger correlations. A square root transform was applied to the branch lengths to visualise fine-scale structure at the tips of the phylogenies. Originally published in Earle et al. *Nat Microbiol*. 2016.

representing phenotypic differences at the level of the lineages, and each lineage/principal component is therefore implicitly tested for an association with the phenotype. Traditionally in GWAS these coefficients have not been interrogated.

The equivalence between including each principal component vs each locus in an LMM means that the random effects of the loci estimated by the LMM can be decomposed in order to obtain an estimate and standard error for the coefficient of each principal component. Whereas the point estimates and standard errors of the random effects estimated for individual loci are typically ignored, because their assumed normal distribution with a common variance constrains ('shrinks') them to be small and not significantly different from zero, cumulatively, these background locus effects can represent substantial phenotypic differences at the level of lineages (PCs). We therefore assessed the significance of the association between each PC and the phenotype using a Wald test (Wald 1943), in which the test statistic W = (point estimate – theta0)^2/standard error^2 follows a χ^2 distribution with 1df under the null hypothesis of no lineage effect. We implemented our method in the R package *bugwas* which is freely available on github (github.com/sgearle/bugwas).

3.4.6.3 Wald test for lineage effects revealed significant lineage associations with fusidic acid resistance

Our Wald test for lineage effects revealed strong signals of association between fusidic acid resistance in *S. aureus* and lineages including PC-6 and PC-9 ($P = 10^{-138}$ and $P = 10^{-106}$, respectively) which was comparable to the significance of the low-frequency kmers representing resistance causing variants in *fusA* (Figure 3.10**;** Figure 3.6). After correcting for the number of lineages tested (equivalent to the number of isolates at 992) using a Bonferroni-corrected significance threshold, 60 PCs were above or equal to the $-\log_{10} P$ value threshold of 4.3. Figure 3.10 displays the significance of the top 20 most significant

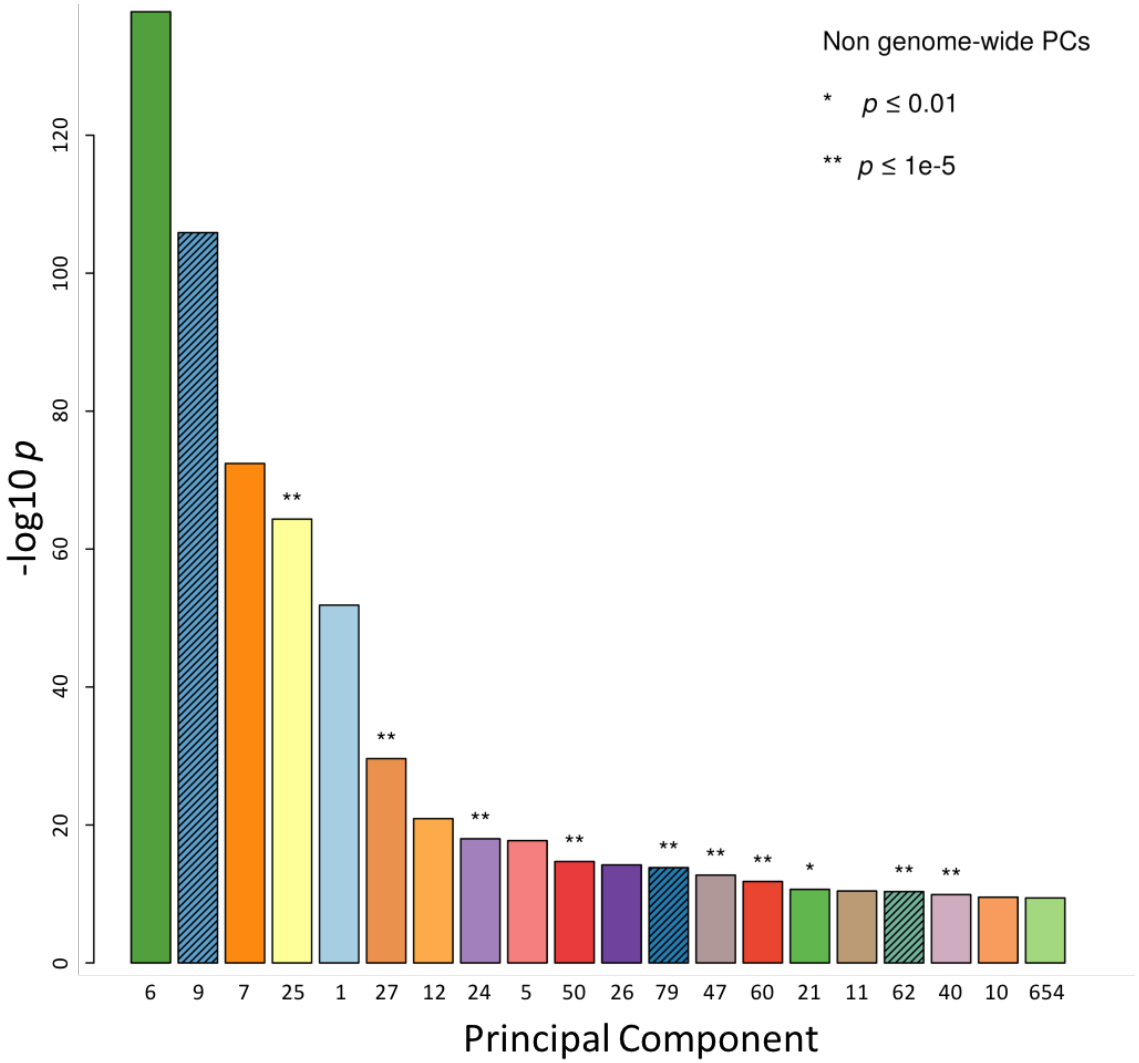

Figure 3.10 Wald test of significance of lineage associations of *S. aureus* principal components with fusidic acid resistance. The first 20 most significant PCs are shown. Some PCs such as PC-9 are hashed to indicate that no branch in the phylogeny was most strongly correlated with the PC. Asterisks above the bars such as PC-25 indicate the strength of evidence for the principal component being associated with particular regions of the genome as opposed to being genome wide. Many lineages were significantly associated with fusidic resistance with the strongest signals being seen for PC-6 and PC-9. Originally published in Earle et al. *Nat Microbiol*. 2016.

PCs. Some PCs are hashed to indicate that no branch in the clonal genealogy was most strongly correlated with it (Figure 3.18). Simulations performed by Daniel Wilson found a widespread loss of genome-wide significance when testing for locus effects as discussed in Section 3.4.4, however by identifying lineage-level associations, power was increased from 2.5-fold (*M. tuberculosis*) to 22.0-fold (*E. coli*).

As detailed in the Methods Chapter 2.10.4, asterisks above the bars, for example PC-25 and PC-27, indicate evidence for lineages associated with particular genomic regions (Figure 3.11). This was assessed by testing for spatial hererogeneity of the SNP loadings, which are the correlations of each SNP to a given PC (Price et al. 2010). This revealed that although many of the most significant lineages were genome wide, meaning that the SNP loadings for these PCs were evenly distributed across the genome (e.g. Figure 3.11A), some captured variation localised to particular areas of the genome (e.g. Figure 3.11B). We found that when a PC captured variation localised to particular genomic regions, these regions typically represented known mobile elements in the *S. aureus* genome, including SCC, SaPI4, a transposon in ICE6013, ϕSa2, Tn*552* and ϕSa3 (Holden et al. 2004; Everitt et al. 2014) (Figure 3.11B).

3.4.6.4 The significant lineages corresponded to resistance-associated *S. aureus* lineages

Viewing the top three most significant lineages on the phylogeny separately made clear the lineages that they represented (Figure 3.12). This revealed that lineage PC-6 represented the split between the resistance-associated lineage ST-8 from the fully sensitive lineage ST-5. Lineage PC-9 represented the split between the resistance-associated ST-1 lineage from the mostly sensitive lineage including ST-12. Lineage PC-7 represented the split between the two resistance-associated lineages ST-8 and ST-1, splitting the more resistant lineage ST-1 from the more phenotypically mixed lineage ST-8.

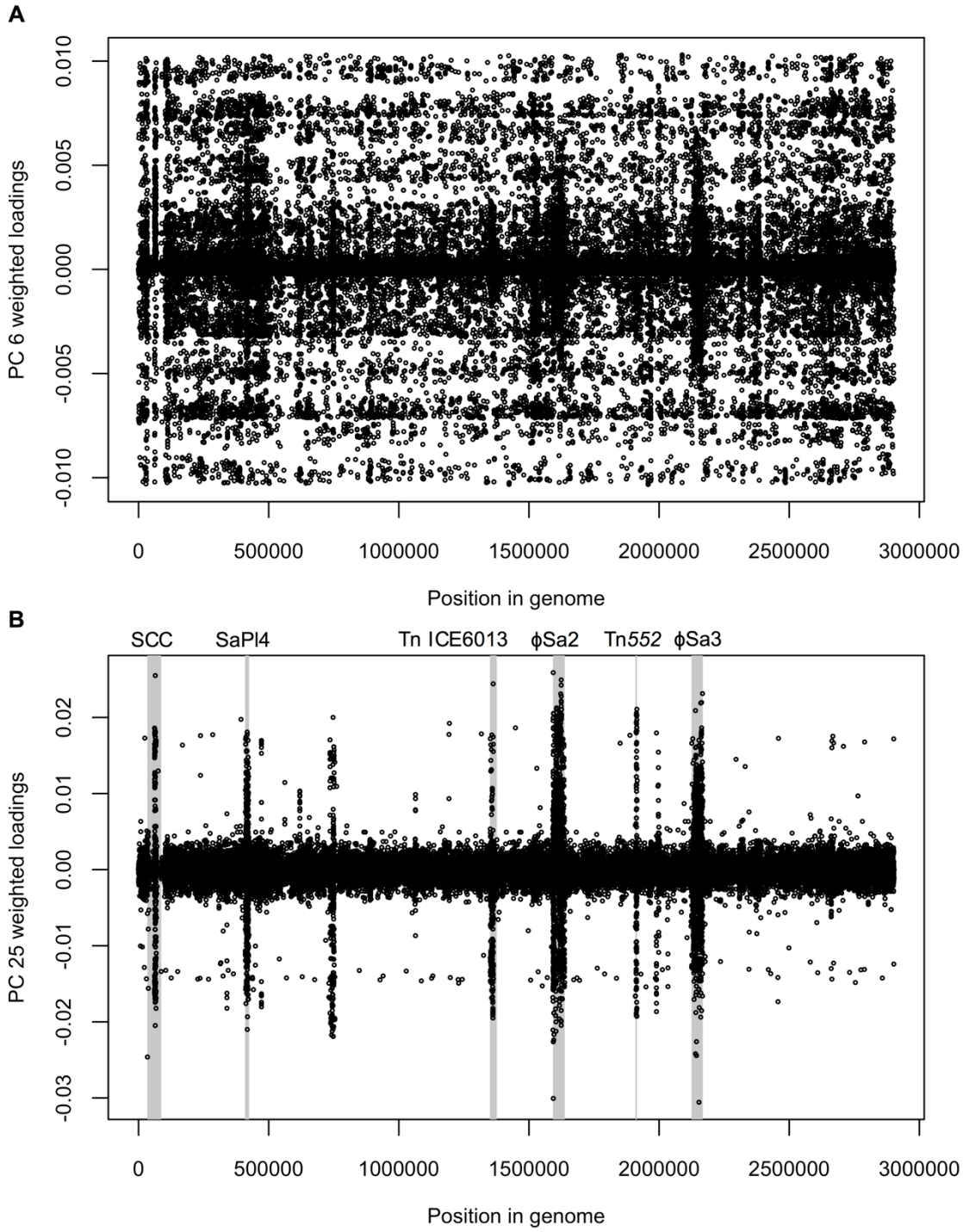

Figure 3.11 SNP loadings of principal components A PC-6 and B PC-25. Some principal components capture variation localised to particular genomic regions. PC-6 represents a PC capturing a deep branch of the tree and therefore the SNPs contributing to the PC are genome wide. PC-25 however appears to be capturing variation in particular regions corresponding to known *S. aureus* mobile elements, highlighted in grey. Tn = Transposon.

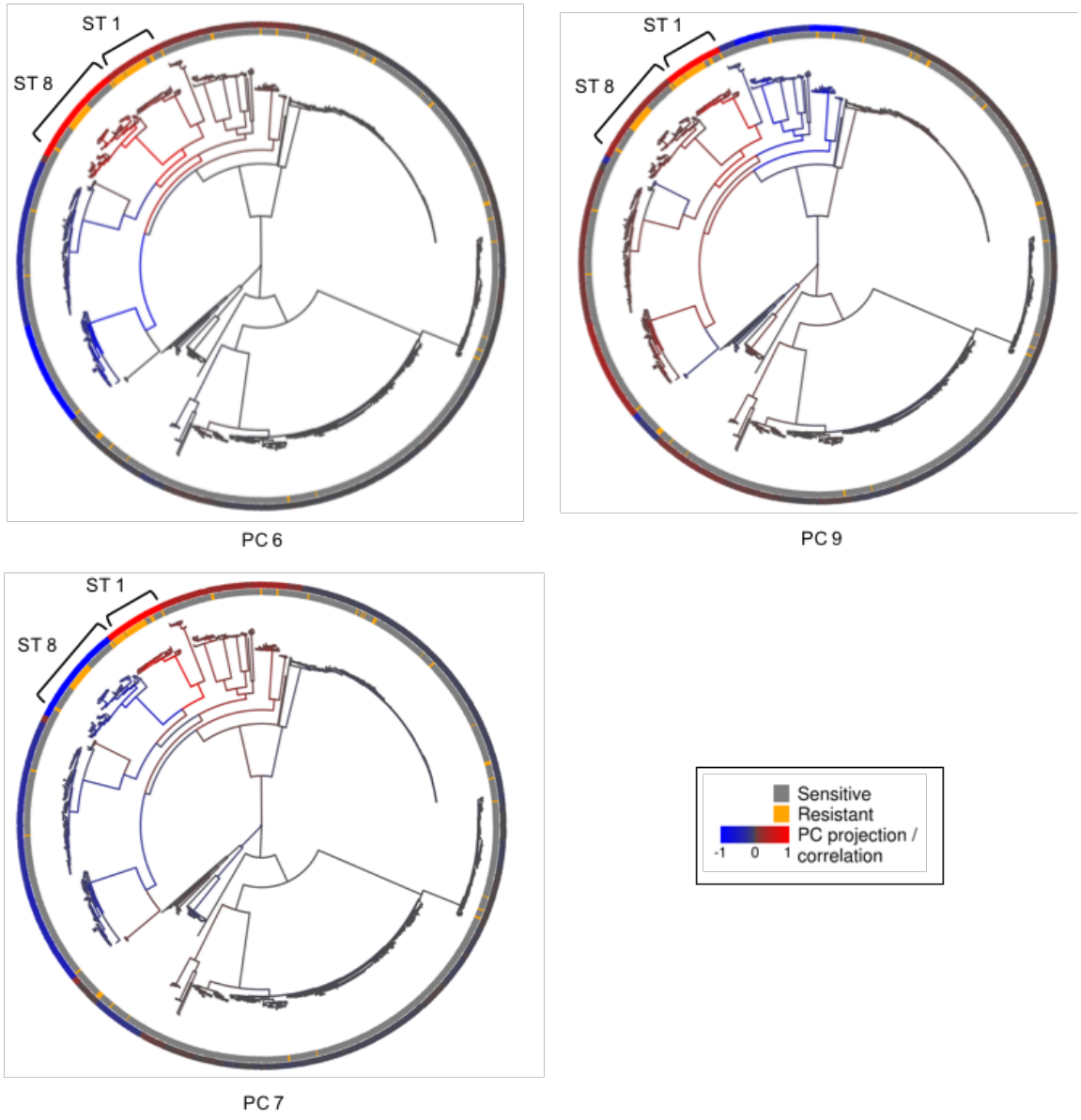

Figure 3.12 The *S. aureus* phylogeny annotated by fusidic acid resistance status and principal component projections of the most significant PCs by the Wald test. Tips are annotated by fusidic acid resistance, grey for sensitive and orange for resistant on the inner ring and by the projection of the isolates onto each of the PCs on the outer ring rescaled between -1 and 1 (blue to red). Branches of the tree are also coloured according to their correlation with the projections of the isolates onto each PC. Positive correlations are coloured red and negative correlations blue, with brighter colours representing stronger correlations. Each PC appears to trace a path through the clonal genealogy, contrasting one lineage (red branches) with another (blue branches) enabling the interpretation of what the significant lineage effects represent. . Bootstrap supports are shown in Appendix B Figure B.1. Originally published in Earle et al. *Nat Microbiol*. 2016.

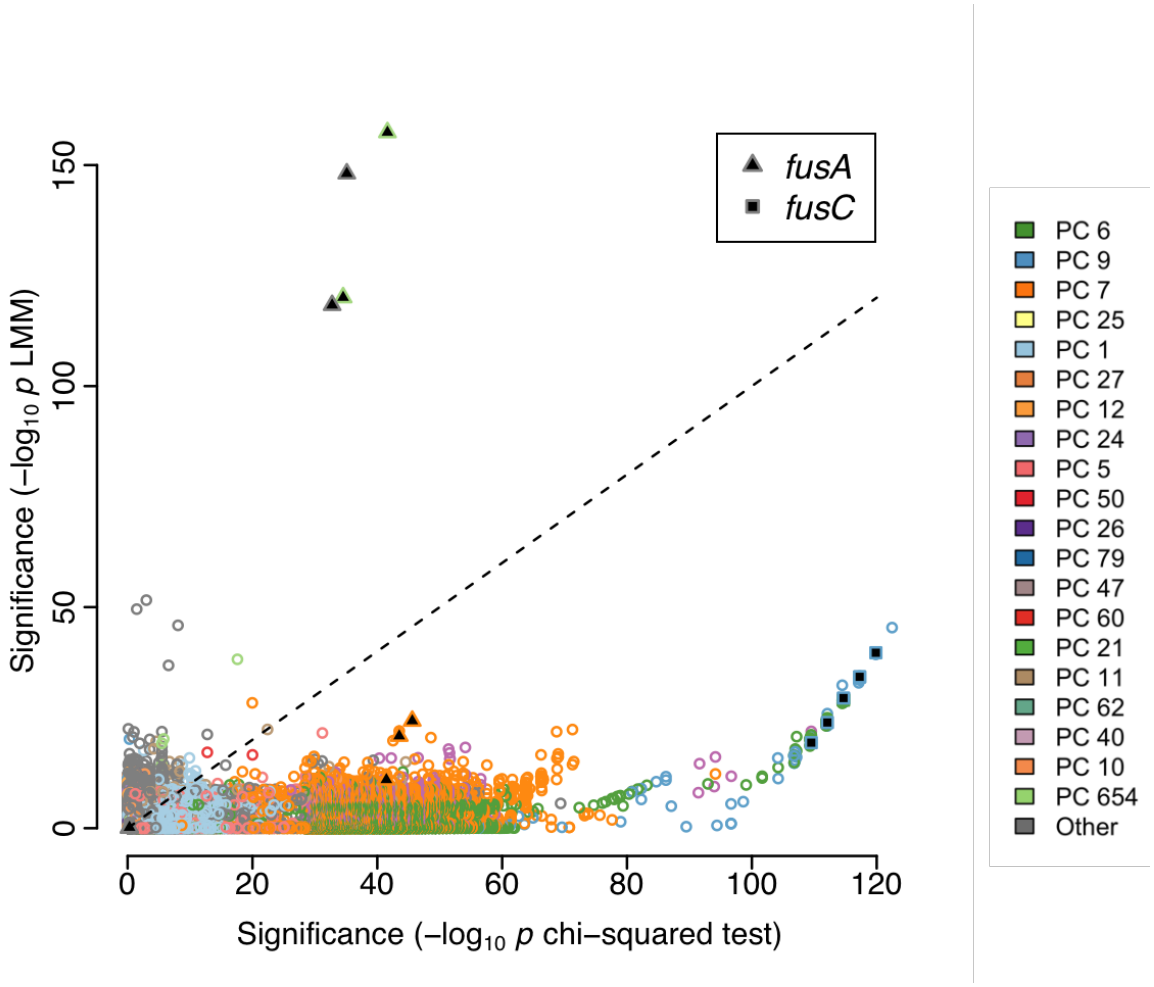

Figure 3.13 The effect of controlling for population structure using LMM on the significance of the presence or absence of 31bp *S. aureus* kmers with fusidic acid resistance coloured by lineage associations. The 200,000 most significant kmers prior to control for population structure plus a random 200,000 were plotted. Each kmer was coloured according to the PC to which it was most strongly correlated, grey if not one of the 20 most significant PCs by the Wald test. Kmers capturing resistance conferring *fusC* were amongst the most significant variants most correlated to PCs 6 and 9. Originally published in Earle et al. *Nat Microbiol*. 2016.

3.4.6.5 Kmers capturing the fusidic acid resistance determinant *fusC* were amongst the strongest signals of association correlated with PC-6 and P-9

We next reassessed the locus-specific effects by assigning variants to lineages according to the principal component to which they were most correlated. To achieve this, we calculated the correlation between the SNP patterns and the projections of the isolates onto the principal components, and assigned variants to the principal component to which the absolute correlation was highest. Figure 3.13 depicts the results of the kmer

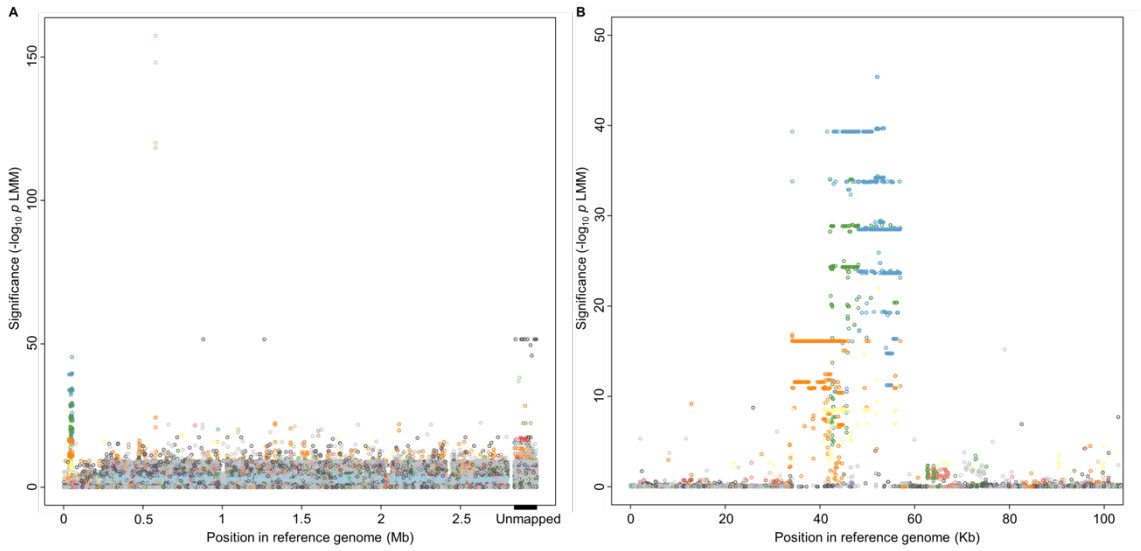

Figure 3.14 Significance of *S. aureus* 31bp kmers with fusidic acid resistance after controlling for population structure using LMM, mapped to MSSA476 and coloured by lineage effects. A All kmers which mapped to MSSA476 with unmapped kmers plotted to the right. **B** The SCC region of MSSA476 containing the resistance determinant *fusC*. The strongest locus-specific associations of the top lineage effects localised to the SCC region.

association tests using χ^2 tests versus LMM, coloured by the lineage associations if the variant was most correlated to one of the top 20 most significant PCs by the Wald test, grey otherwise. This revealed that kmers mapping to *fusC* and the SCC region in linkage disequilibrium with *fusC* accounted for the strongest signals within PC-6 and PC-9 ($P = 10^{-34}$ and 10^{-45} respectively).

In fact, visualising the mapping of the kmers to reference genome MSSA476 (accession NC_002953) in Figure 3.14 revealed that the majority of the strongest locus-specific associations localised to a 20kb region containing the staphylococcal cassette chromosome (SCC), and the most significant kmers in the region mapped to the gene adjacent to the resistance-conferring *fusC*, SAS0040.

Figure 3.15 depicts the differential effect of controlling for population structure using LMM on the significance of association between fusidic acid resistance and kmers that were correlated to any of the significant lineages in *S. aureus* versus those which

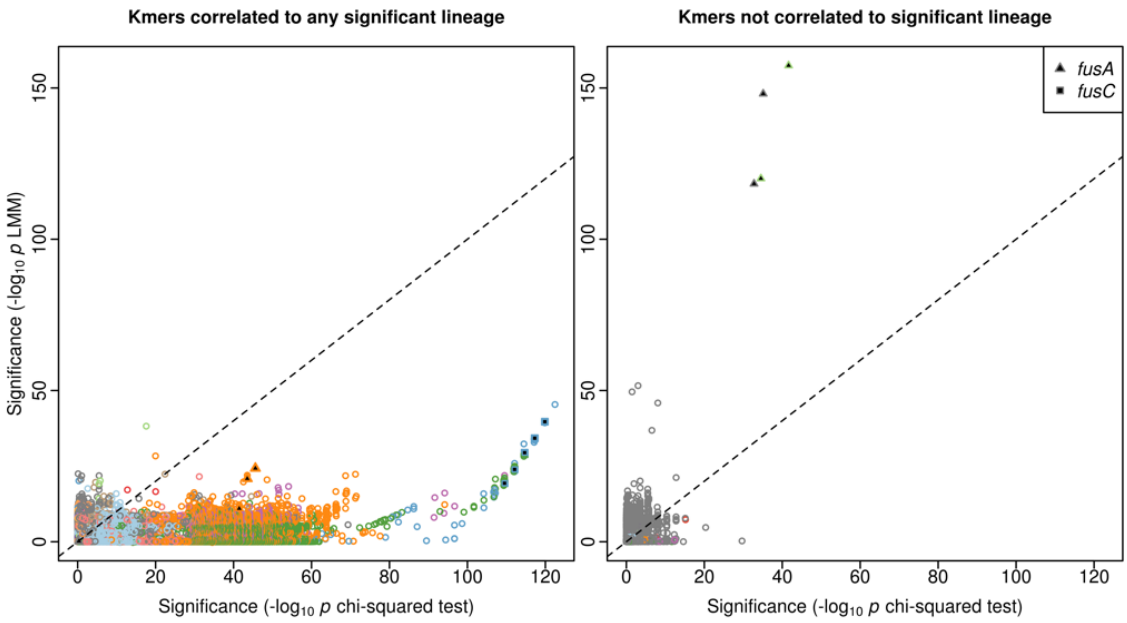

Figure 3.15 Differential effect of controlling for population structure using LMM on significance of association between kmers that are correlated vs. uncorrelated with any significant variant in *S. aureus* and resistance to fusidic acid. A Kmers with absolute correlation > 0.25 to any PC significantly associated with fusidic acid resistance by the Wald test. **B** Kmers with no absolute correlation > 0.25 to any PC significantly associated with fusidic acid resistance by the Wald test. The 200,000 most significant kmers prior to population structure control plus a random 200,000 were plotted. Kmers were coloured according to the PC to which they were most strongly correlated to, grey if not one of the most significant 20 PCs by the Wald test. Population-stratified kmers suffered greater loss of significance than non-stratified kmers using LMM. Originally published in Earle et al. *Nat Microbiol*. 2016.

were not correlated to any of the significant lineages. The plot shows the results of not controlling for population structure on the x-axis versus controlling for population structure using LMM on the y-axis. The plot is split by the kmers which had an absolute correlation > 0.25 to any PC which showed a significant association to fusidic acid resistance by the Wald test (Figure 3.15A) versus kmers which had no absolute correlation > 0.25 to any of the PCs which showed significant association to fusidic acid resistance by the Wald test (Figure 3.15B). This revealed that variants which were lineage associated dropped in significance when controlling for population structure much more than unstratified variants, which was expected if the control for population structure was working effectively. This also shows how unstratified causal variants can be propelled to

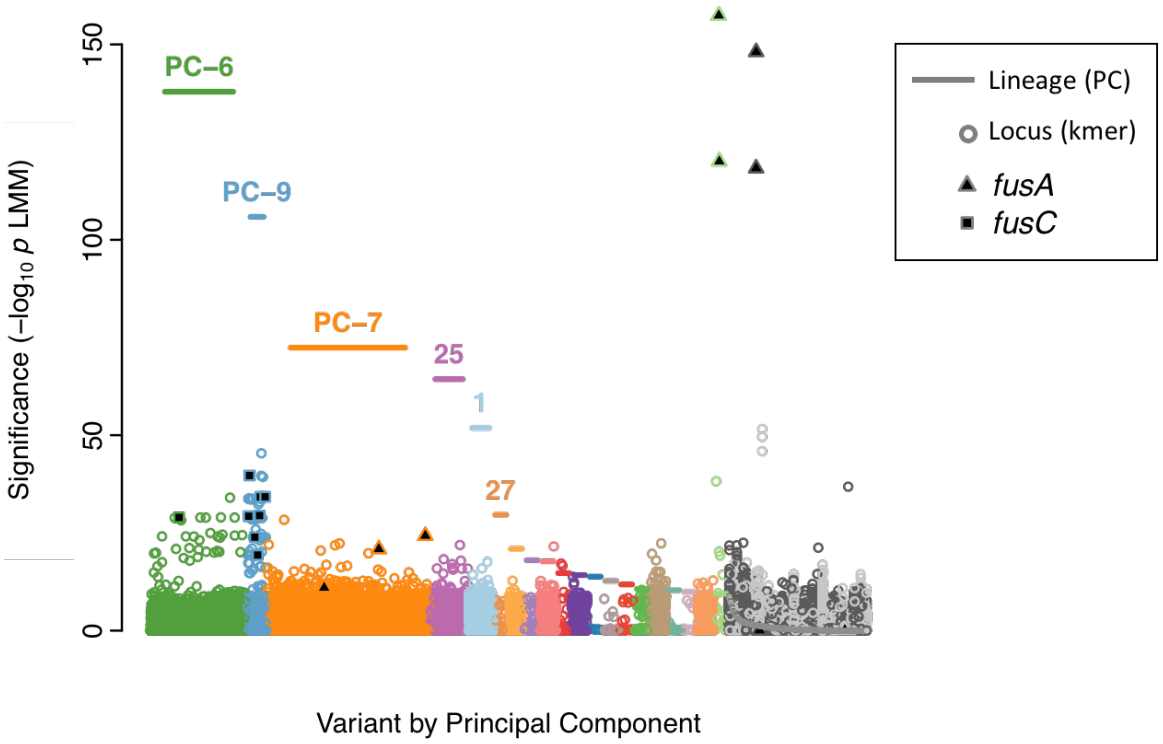

Figure 3.16 Significance of association of *S. aureus* 31bp kmers with fusidic acid resistance after controlling for population structure using LMM, stratified by the lineage to which they were most correlated. Unique variant patterns were plotted, with the horizontal ordering within each PC being randomised. On the y-axis, the significance of lineage effects (bars) is shown alongside locus effects (circles). Variants were coloured according to the PC to which they were most correlated, allowing for the identification of variants within the top lineage associations. This revealed that the filled squares representing kmers capturing resistance conferring *fusC* were among the most significant variants correlated to PCs 6, 9 and 7. Originally published in Earle et al. *Nat Microbiol*. 2016.

greater significance using LMM, which can improve power in the presence of polygenic effects (Yang et al. 2014).

3.4.6.6 Understanding lineage effects by assigning loci to lineages

As we assigned our variants to the lineages defined by principal components, we could then rank our results first by lineage, and then by locus effects. This enabled us to directly compare the significance of variants within lineages, as an alternative to the standard Manhattan plot. This again made clear that kmers capturing the *fusC* resistance-conferring gene were among the most significant variants of the top three lineage effects, along with variants in the surrounding SCC region (Figure 3.16). As a group, lineage PC-

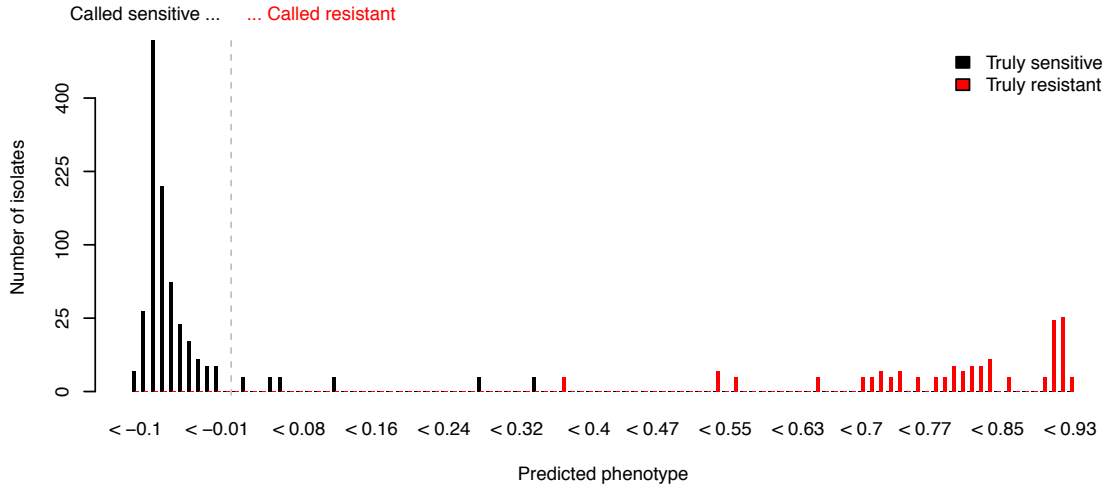

Figure 3.17 Phenotype prediction distribution. Histogram of the continuous phenotype predictions using SNP data. Bars were coloured by their true phenotype, and those plotted to the right of the grey dotted line were predicted to be resistant, and those to the left of the line were predicted to be sensitive. This revealed that most phenotype predictions were strong, and the six incorrect phenotype predictions were the weakest of those predicted to be resistant. This figure was created based on code written by Daniel Wilson.

6 was of a similar level of significance to the low-frequency resistance-conferring variants in *fusA*, which cumulatively were only present in 24 isolates in comparison to *fusC* which was present in 44 isolates. Therefore, by identifying the loci which contribute the most to significant lineage effects, we gain flexibility in interpreting GWAS results and an alternative method for prioritising variants for experimental follow-up, which is typically based entirely on locus-specific significance. This introduces a trade off, by interpreting lineage-level effects we must acknowledge the increased risk of confounding by population-stratified differences in environment or sampling. However, with the high risk can also come high gain, as identifying the most significant lineage-associated loci permits the pursuit of functional validation within groups of lineage-associated variants. In this example, the most significant locus effects within the top lineage effects are significant in their own right.

3.4.6.7 Heritability of fusidic acid resistance

Fusidic acid resistance was highly heritable, with a heritability estimate of 99.8% with a

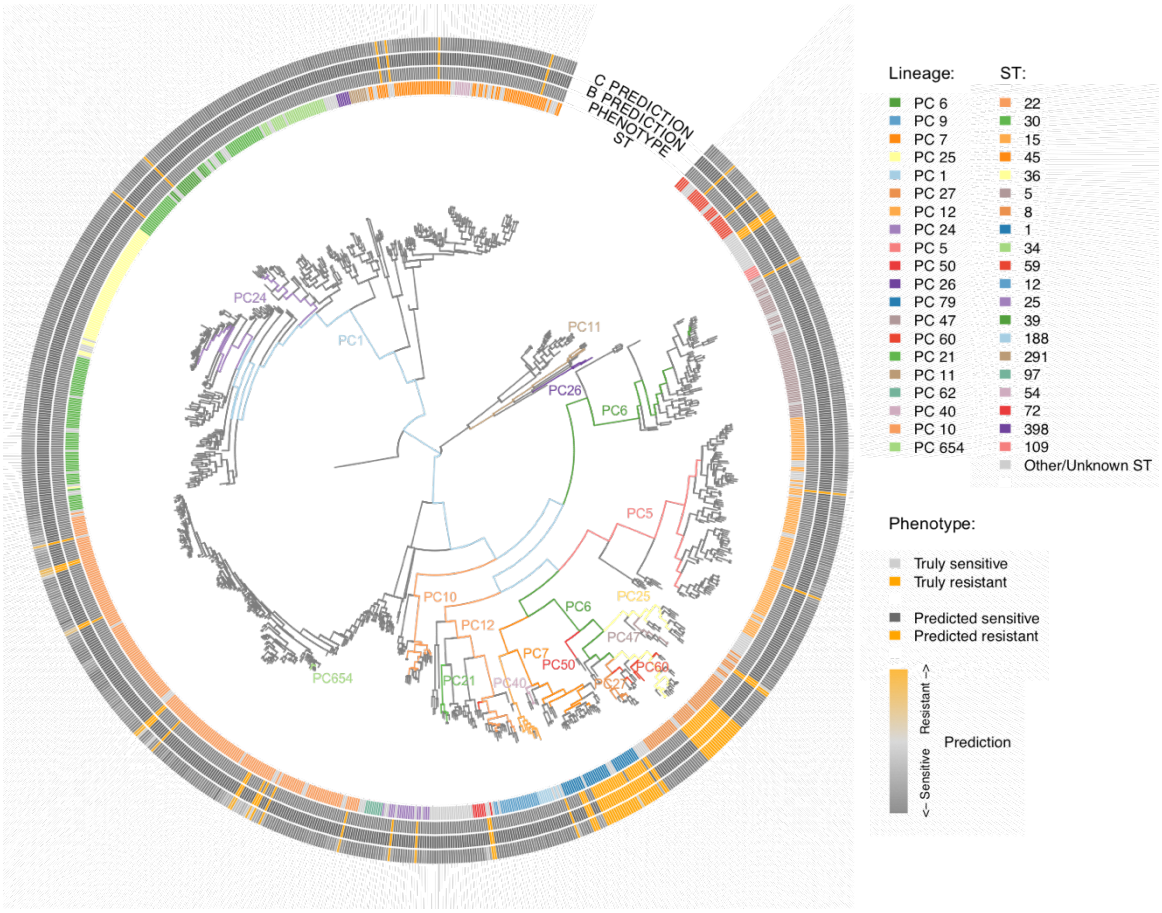

Figure 3.18 Phylogenetic phenotype prediction distribution. Maximum likelihood phylogeny annotated with the 20 most common STs and the true and predicted phenotypes. Branches are coloured by the PC they were most strongly correlated with. Due to high heritability, the predicted phenotype was highly accurate. Bootstrap supports are shown in Appendix B Figure B.1.

standard error of 4×10^{-4}. This means that the proportion of phenotypic variation explained by the genetic variation was extremely high, resulting in almost perfectly being able to predict the phenotype from the genetic data (Figure 3.17). The phenotype was predicted using a ridge regression in which every SNP is used for prediction, which is the null Linear Mixed Model. The phenotype was coded as resistant (1) and sensitive (0) and mean centred, so that positive values represented resistant strains and negative values sensitive strains. We made a binary sensitive versus resistant prediction for each isolate by identifying whether the prediction was negative or positive, resulting in correctly predicting 986/992 isolates. Only six isolates were incorrectly predicted, and these were all truly sensitive isolates that were predicted to be resistant. The continuous phenotype

prediction for those which were incorrect were the weakest predictions amongst the resistant predictions (Figure 3.17**;** Figure 3.18). Of the incorrectly predicted isolates, two were from the resistance dominated ST-8 and ST-1 lineages explaining their incorrect predictions, and four were ST-22 isolates (Figure 3.18). The incorrectly predicted ST-22 isolates were all phylogenetically close to a resistant isolate, again explaining the incorrect, albeit weak, predictions (Figure 3.18).

3.4.7 GWAS identified genuine causal variants or variants in close physical linkage with causal variants in 25/26 studies of 17 antimicrobials across four species

Given the strong population structure and genome-wide linkage disequilibrium in bacteria, we aimed to empirically test the ability of GWAS to identify genuine causal variants more generally. We conducted 26 GWAS for resistance to 17 antimicrobials across the major pathogens *E. coli* (241 isolates), *K. pneumonia* (176 isolates), *M. tuberculosis* (≤1735 isolates) and *S. aureus* (≤ 992 isolates) (Figure 3.19). As well as testing 31bp kmers for association, we also tested SNP variation and pangenome gene presence/absence (Methods Chapter 2.5). Pangenome analyses were performed by Jane Charlesworth and SNP and kmer analyses for *E. coli*, *K. pneumoniae* and *M. tuberculosis* were performed by Chieh-Hsi Wu. We imputed missing SNPs using ClonalFrameML as discussed in Section 3.4.2 (Didelot & Wilson 2015). For each drug and species, we evaluated whether the most significant variant identified by the GWAS matched a known causal variant or was in physical linkage with a known causal variant, combining the SNP and gene results as we typically expect resistance to be conferred by either a SNP or a gene, and assessing the kmers separately. Causal variants were assessed using the resistance catalogues of Stoesser et al. (2013), Gordon et al. (2014) and Walker et al. (2015). Using this measure, the GWAS performance was very high across species.

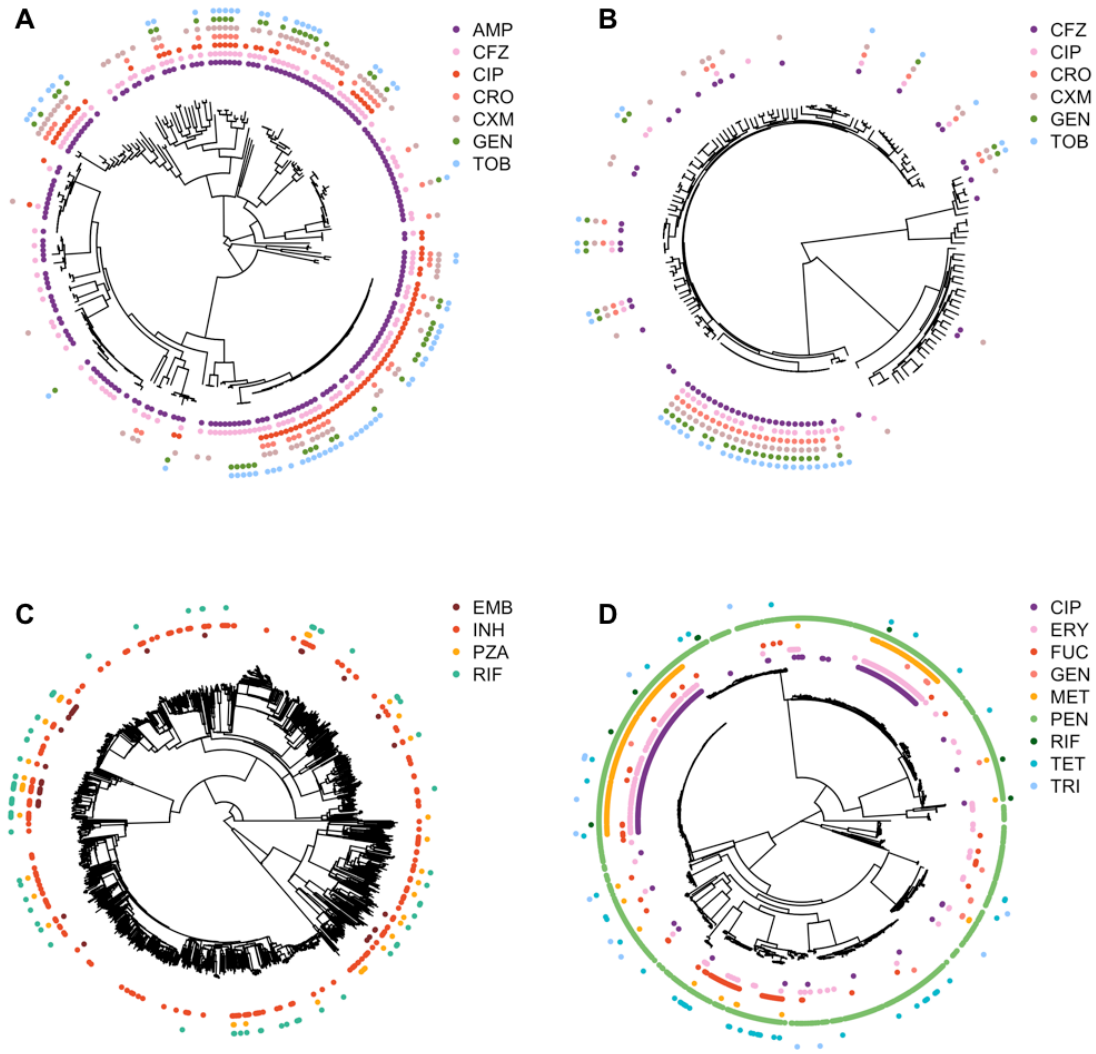

Figure 3.19 Distribution of antibiotic resistance phenotypes on the phylogeny of the four species investigated: A *E. coli* B *K. pneumoniae* C *M. tuberculosis* D *S. aureus*. The midpoint rooted RAxML maximum likelihood phylogeny is shown. AMP = Ampicillin, CFZ = Cefazolin, CIP = Ciprofloxacin, CRO = Ceftriaxone, CXM = Cefuroxime, GEN = Gentamicin, TOB = Tobramycin, EMB = Ethambutol, INH = Isoniazid, PZA = Pyrazinamide, RIF = Rifampicin, ERY = Erythromycin, FUC = Fusidic acid, GEN = Gentamicin, MET = Methicillin, PEN = Penicillin, RIF = Rifampicin, TET = Tetracycline, TRI = Trimethoprim. Figure created by Chieh-Hsi Wu. Originally published in Earle et al. *Nat Microbiol*. 2016.

Genuine causal loci or regions in physical linkage with causal loci were identified as the most significant variant in 25/26 cases for the SNP and gene approach combined and the kmer approach after controlling for population structure (Table 3.5). Therefore, genuine resistance-conferring variants were detected in all but one study. The

Antibiotic	# R	# S	Resistance mechanism	SNP / gene rank	SNP / gene LMM rank	Kmer rank	Kmer LMM rank
E. coli							
Ampicillin	189	52	β-lactamase genes bla_{TEM}	1	1	6 (tnp)	6 (tnp)
Cefazolin	139	102	β-lactamase genes bla_{CTX-M}	2 (*nmpC*)	3 (*nmpC*)	121710 (nmpC)	3690 (nmpC)
Cefuroxime	81	160	β-lactamase genes bla_{CTX-M}	1	1	1598 (162-192 upstream bla_{CMY-2})	470 (162-192 upstream bla_{CMY-2})
Ceftriaxone	55	186	β-lactamase genes bla_{CTX-M}	1	1	1403 (tnp)	470 (tnp)
Ciprofloxacin	91	150	SNPs in *gyrA*[a], *gyrB*, *parC*[a] or *parE* or presence of PMQR	1[b]	1[b]	1[b]	1[a]
Gentamicin	48	193	*aac* (*aac(3)-II*), *ant*, *aph* or rRNA methylase	1	1	1	1
Tobramycin	67	174	*aac* (*aac(3)-II*), *ant* or rRNA methylase	1	1	1	1
K. pneumoniae							
Cefazolin	53	123	β-lactamase genes bla_{CTX-M}	1 + HP + *wbuC*	1	762 (tnp)	837 (tnp)
Cefuroxime	46	130	β-lactamase genes bla_{CTX-M}	1 + HP + *wbuC*	1+ HP + *wbuC*	762 (tnp)	1480 (tnp)
Ceftriaxone	35	141	β-lactamase genes bla_{CTX-M}	1 + HP + *wbuC*	1 + HP + *wbuC*	771 (tnp)	812 (tnp)
Ciprofloxacin	34	142	SNPs in *gyrA*, *gyrB*, *parC* or *parE* or presence of PMQR (*qnr-B1*[a], *qnr-B19*[b])	2[a] (tnp)	2[a] (tnp)	1853[b] (tnp)	4427[b] (tnp)
Gentamicin	31	145	*aac* (*acc(3)-II*), *ant*, *aph* or rRNA methylase	1	1	1	79 (*tmrB_2*)
Tobramycin	36	140	*aac* (*acc(3)-II*), *ant* or rRNA methylase	1	1	1	1
M. tuberculosis							
Ethambutol	41	1589	*embB*	2 (*rpoB*)	1	1	1
Isoniazid	239	1470	*katG*, *fabG1*	1	1	1	1
Pyrazinamide	45	1662	*pncA*	142 (*rpoB*)	1	126 (*rpoB*)	1
Rifampin	86	1487	*rpoB*	1	1	1	1
S. aureus							
Ciprofloxacin	242	750	*grlA* or *gyrA*	1	1	1	1
Erythromycin	216	776	*ermA*, *ermC*, *ermT* or *msrA*	1	1	1	1
Fusidic acid	84	908	SNPs in *fusA*[a] or presence of *fusB* or *fusC*[b]	4[a] (*SAS0037*)	1[a]	75[a] (*SAS0040*)	1[a]
Gentamicin	11	981	*aacA*/*aphD*	1 + GNAT acetyltransferase	1 + GNAT acetyltransferase	1 + 415 bases upstream to 100 bases downstream	1 + 415 bases upstream to 100 bases downstream
Penicillin	824	168	*blaZ*	1	1	2 (*blaI*)	2 (*blaI*)
Methicillin	216	776	*mecA*	1	1 + *mecR1*	1 + SCC*mec* genes	1 + SCC*mec* genes
Tetracycline	46	946	*tetK*, *tetL* or *tetM*	2 (*repC*)	2 (*repC*)	1 + plasmid genes	1 + plasmid genes
Trimethoprim	15	308	*SNPs in dfrB, presence of dfrG or dfrA*	1	1	1	1
Rifampicin	8	984	*rpoB*	1	1	1	1

Table 3.5 Results of 26 GWAS across 17 antimicrobials in four bacterial species. For each study, known resistance mechanisms are listed and coloured red if resistance is conferred by gene presence, blue if conferred by particular SNP variants and purple if by both. If more than one resistance mechanism is listed, the mechanism identified was underlined or is referred to by [a] or [b]. Results are coloured white if the most significant variant was expected, light grey if it was in close physical linkage with the expected mechanism, dark grey if other. R, resistant; S, sensitive; HP, hypothetical protein; tnp, transposase; PMQR, plasmid mediated quinolone resistance. Originally published in Earle et al. *Nat Microbiol*. 2016.

–log10 *P* values for the most significant result for each GWAS, plus the causal variant if not the most significant, are detailed in Appendix A, Tables A.1-A.4. For cefazolin resistance in *E. coli*, we identified the variable presence of an unexpected gene to be the variant most strongly associated with resistance, the gene *nmpC* ($P = 10^{-12.4}$ for gene presence vs absence). *nmpC* encodes an outer membrane porin which was more prevalent

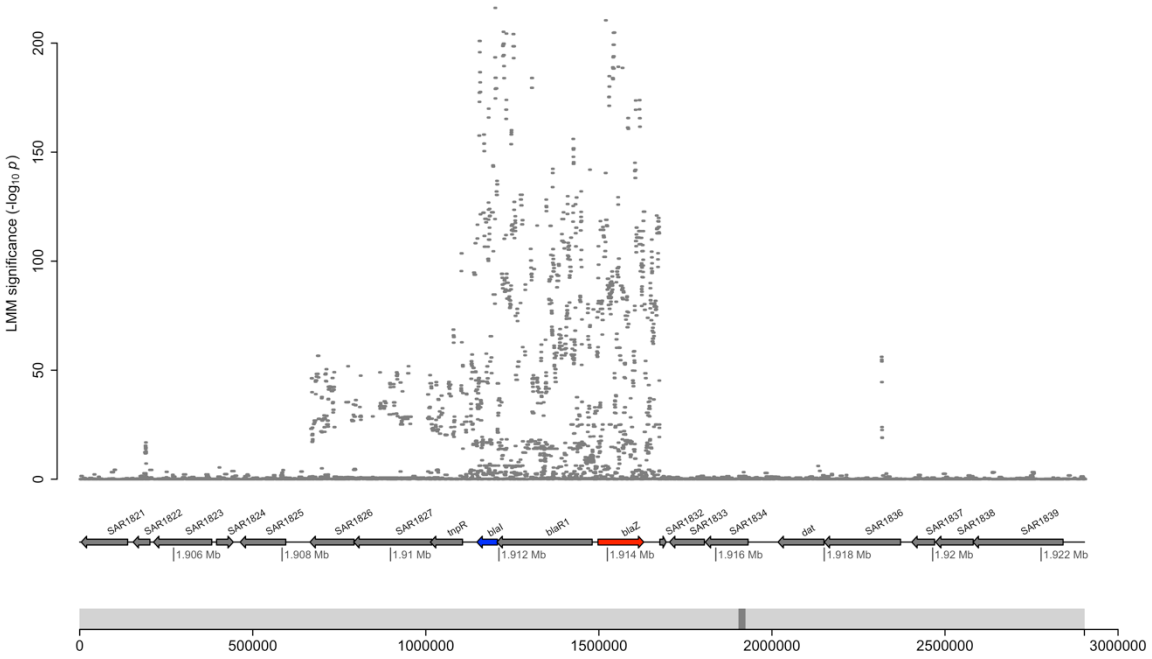

Figure 3.20 The significance of *S. aureus* 31bp kmers with penicillin resistance after controlling for population structure using LMM mapped to the region of *S. aureus* MRSA252 containing the penicillin resistance conferring *blaZ* gene. The causal gene *blaZ* is coloured red, but the most significant kmers were found in *blaI* coloured in blue. The mobile element associated region of LD was detected along with the true causal mechanism in this study. Originally published in Earle et al. *Nat Microbiol*. 2016.

in susceptible isolates. Further investigation revealed that this is a strong candidate for a novel resistance-conferring mechanism in *E. coli* as permeability in the *Salmonella typhimurium* homologue enables resistance to other cephalosporin β-lactams (Sun et al. 2009).

3.4.7.1 Non-causal variants in LD with causal variants were often significantly associated

We often found that non-causal variants in linkage disequilibrium with causal variants would be detected along with the causal variant, particularly for accessory genes such as β-lactamases where mobile element associated regions were often found associated along with the causal locus. Figure 3.20 depicts the region containing the penicillin resistance-conferring *blaZ* gene in *S. aureus* where the kmers have been mapped to the reference genome MRSA252. Here the most significant kmers in the region were actually in *blaI*,

the *blaZ* repressor, but the whole mobile element-associated region of LD was detected along with *blaZ* the causal mechanism.

We encountered false positives as a result of correlated phenotypes due to multi-drug resistant isolates. First-line treatment for *M. tuberculosis* infection involves taking a combination of drugs, therefore multi-drug resistance frequently evolves (Dorman & Chaisson 2007). Before controlling for population structure, this resulted in spurious associations between ethambutol and pyrazinamide resistance and SNPs and kmers in rifampicin resistance-conferring *rpoB*. After controlling for population structure resistance-conferring variants in *embB* and *pncA* were the top results for ethambutol and pyrazinamide ($P = 10^{-83}$, $P = 10^{-60}$ for the SNP GWAS) respectively, however variants within *rpoB* remained highly significant at $P = 10^{-45}$ and $P = 10^{-54}$. This could reflect the patterns of the emergence of drug resistance seen globally, where rifampicin resistance-conferring mutations are typically preceded by resistance-conferring mutations to other drugs (Manson et al. 2017). Therefore, it will be important to be aware of phenotype correlations and resistance mutation emergence patterns in future studies aiming to discover novel resistance-conferring mechanisms. Multivariate analyses may be beneficial in the presence of correlated antimicrobial resistance phenotypes as it has been shown that they can improve power to detect both pleiotropic effects but also variants affecting just one of the phenotypes (Morris et al. 2010; Stephens 2013).

3.4.7.2 Testing kmers for association can increase power over testing SNPs

We found that testing kmers for association had some advantages over testing SNPs, even when a causal variant was SNP based. In the phylogeny of *M. tuberculosis* antimicrobial resistance has arisen more than 20 times for each of the four antimicrobials by frequent convergent evolution (Figure 3.19). For example, *rpoB* contains multiple targets for selection, as resistance to rifampicin can occur due to amino acid substitutions anywhere

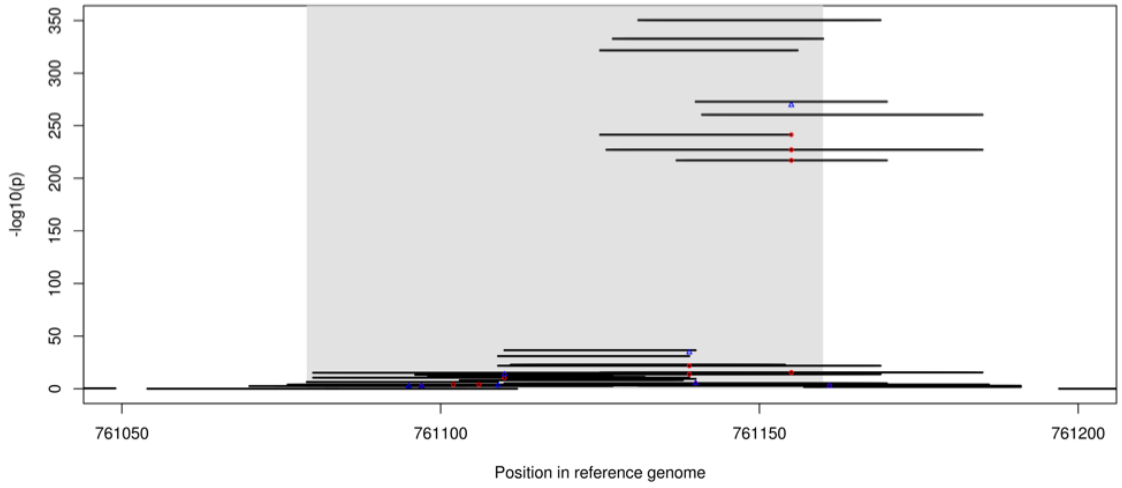

Figure 3.21 The significance of *M. tuberculosis* 31bp kmers with rifampicin resistance after controlling for population structure using LMM mapped to rifampicin resistance conferring *rpoB* of H37Rv. The grey shaded region represents the codons within *rpoB* where amino acid substitutions change resistance status. Kmers are plotted as black lines with red circles representing non-synonymous, nonsense or read-through SNPs annotated on the kmers within the shaded region. Blue triangles represent significance of non-synonymous, nonsense and read-through SNPs based on mapped data within the shaded region after controlling for population structure. As kmers can pool over SNPs if within 31bp, this can increase power above testing for SNP variants individually. Originally published in Earle et al. *Nat Microbiol*. 2016.

between codons 425-452 (with respect to reference genome H37Rv, NC_000962.2) in the rifampicin-resistance determining region (Telenti et al. 1993; Andre et al. 2017). Multiple *rpoB* resistance-conferring variants were present within the *M. tuberculosis* population,with each distinct variant only present at low frequency. Although the SNP analysis correctly identified known variants in resistance-causing codons to be the most significant variants after controlling for population structure, greater significance was achieved in the kmer-based analysis. This is because many of the variants were within 31bp of one another and therefore a single kmer could capture multiple causal variants. More specifically, kmers which captured the (sensitive) wild-type haplotype (31mer), and therefore the absence of all of the resistance-conferring variants, were found to be the most significant kmers ($P = 10^{-322}$, OR = 0.001). This effect of pooling over the alternative resistance-conferring alleles when testing for association between resistance and the presence or absence of the wild type haplotype led to an increase in significance

over testing the SNPs individually. Figure 3.21 depicts both the SNP and kmer results in the region where amino acid substitutions change resistance status. This revealed that the most significant kmers did not contain any resistance-conferring SNPs, but did in fact cover multiple resistance-conferring SNPs present in the population. Although one resistance-conferring SNP was highly significant, by pooling over multiple variants the wild-type kmers benefited from an increase in significance and were found to be more significant than the SNP variant. We may expect to see this in other settings where disruptions to particular protein domains by multiple amino acid substitutions can cause the same phenotype. We in fact found the same effect for the most significant kmer in *fusA* when testing fusidic acid resistance in *S. aureus*, where the most significant *fusA* kmer was more prevalent within the sensitive isolates with an OR of 0.004.

3.5 Discussion

3.5.1 Summary

In this chapter, I developed methods for applying GWAS to bacteria and along with colleagues, tested its feasibility through the application of 26 GWAS to investigate the underlying genetic basis of 17 antimicrobials in *E. coli*, *K. pneumoniae*, *S. aureus* and *M. tuberculosis*, four species which are taxonomically diverse and range from clonal to recombining. The principal findings were:

- ClonalFrameML (Didelot & Wilson 2015) provided a highly accurate method for imputing missing SNP calls in bacteria, much more accurate than Beagle, which was developed for imputation in humans (Browning & Browning 2009). Overall accuracy was 91.5%, 92.0%, 98.2% and 91.3% in *E. coli*, *K. pneumoniae*, *M. tuberculosis* and *S. aureus*, respectively. As expected, accuracy decreased with increasing proportion of missing sites and decreasing MAF.
- Failure to control for population structure resulted in a large number of false

positives. In testing kmers for association with fusidic acid resistance in *S. aureus*, an FPR of 89.2% was estimated before controlling for population structure. Controlling for population structure reduced the number of false positives eight-fold to an FPR of 51.9%, however the trade-off to this was the reduction of significance in the resistance-conferring *fusC*. Although this is necessary to address the inadmissibly high number of false positives before correction, for many phenotypes this may result in outright loss of significance.

- Lost power to detect locus effects in bacterial GWAS can be regained by testing for lineage effects. By defining lineages in terms of principal components (PCs), coefficients can be estimated that capture lineage-level phenotypic differences and a Wald test was employed to assess the significance of the association between each PC and the phenotype. Applying the Wald test to fusidic acid resistance in *S. aureus* revealed that the most significant lineages corresponded to fusidic acid-resistant *S. aureus* lineages ST-8 and ST-1.
- A greater understanding of the lineage effects can be gained by reassessing the locus effects in light of them. By assigning loci to the *S. aureus* lineages we found that kmers capturing the *fusC* resistance-conferring gene were amongst the most significant variants contributing to the top three most significant lineages. By comparing the significance of variants within a significant lineage, this provides an alternative to the standard Manhattan plot and a different way to prioritise variants for experimental follow up.
- Genuine causal loci or regions in physical linkage with causal loci were identified as the most significant association in 25/26 GWAS for antimicrobial resistance.
- Often non-causal variants in LD with causal variants were identified as significantly associated with the phenotype, particularly mobile element-

associated genes linked to accessory resistance-determining genes such as the β-lactamases.

- False positive associations were encountered as a result of correlations between resistance phenotypes due to multi-drug resistance.
- Testing kmers for association sometimes increased significance over testing SNPs, even when causal variants were SNP-based. When multiple variants were within 31bp of one another, a single kmer often captured a wild type haplotype against which multiple resistance-causing mutations arose, for example in rifampicin resistance in *M. tuberculosis*. By pooling over the alternative resistance-conferring alleles when testing for association between resistance and the presence or absence of the wild type haplotype, an increase in significance over testing the SNPs individually was found.
- GWAS identified putative novel resistance-conferring mechanisms. For cefazolin resistance in *E. coli*, the variable presence of the gene *nmpC* was identified to be associated with resistance. *nmpC* encodes an outer membrane porin more prevalent in susceptible individuals. As the permeability of the homologue in *S. typhimurium* confers resistance to other cephalosporin β-lactamases, we believe that this is a strong candidate for a novel mechanism (Sun et al. 2009).
- I developed a pipeline for applying GWAS in bacteria and an R package *bugwas* for testing variants for association whilst controlling for population structure and testing lineage effects.

3.5.2 Revisiting the challenges with applying GWAS to bacteria

3.5.2.1 Defining the phenotype

Although we found that applying GWAS to bacteria to detect antimicrobial resistance-

conferring variants was extremely effective, resistance is a special case. The high accuracy demonstrated in predicting antimicrobial resistance phenotypes genotypically is reflected by the good power to map the genotypes conferring resistance using GWAS (Zankari et al. 2013; Stoesser et al. 2013; Gordon et al. 2014; Walker et al. 2015; Bradley et al. 2015). However due to frequent convergent evolution, antimicrobial resistance has arisen 20 or more times per drug in the *E. coli* tree, 9 or more times for *K. pneumoniae*, over 20 times for *M. tuberculosis* and over 20 times for *S. aureus* (if excluding rifampicin and gentamicin which have a low frequency of resistance). The resistance determinants are highly homoplastic as the selection pressure underlying the evolution of antimicrobial resistance is extremely high. As a result of which, the phenotypes are also highly homoplastic. High homoplasy as a result of repeat mutation and recombination at resistance-conferring loci breaks down LD which assists greatly with fine-mapping the signal and pinpointing resistance-conferring loci.

The binary sensitive/resistant status or MIC phenotypes measured in the laboratory for resistance studies are also likely to be more closely correlated to the phenotypes under selection imposed by antimicrobials in natural populations than for other phenotypes; other traits amenable to laboratory testing are likely to capture just one component of the phenotype, for example biofilm formation in lieu of virulence, and are likely to be less tightly associated with biologically important fitness traits (Falush 2016).

Strains were often resistant to multiple antimicrobials, and we identified confounding between phenotypes as a potential pitfall for future studies. This may be an issue for trying to identify novel resistance-conferring mechanisms, although multivariate GWAS may help to address these cases and improve power (Morris et al. 2010; Stephens 2013). The phenotypic measurements were also limited by what was taken in the routine clinical laboratory. The phenotypes were not taken for a scientific study but for routine

clinical practice, therefore the phenotypes were measured as a binary sensitive/resistant trait. It is possible that defining phenotypes as MICs and performing a multivariate analysis may help to disentangle the genetic contribution to the different phenotypes.

Some of the phenotypes also had a very small proportion of resistant samples, again as we made use of the samples that came through the routine laboratory. Sample frequency determines power, however despite the low proportion of resistant cases for gentamicin and rifampicin resistance in *S. aureus* and ethambutol and pyrazinamide resistance in *M. tuberculosis*, we were able to identify genuine causal mechanisms as the most significant variants for all four analyses after controlling for population structure. Study design will also be important in understanding sample heritability estimates made in GWAS studies, and will influence how closely the sample heritability corresponds to the true heritability.

3.5.2.2 The accessory genome and strong LD

We found that if the dominant resistance-conferring mechanism within the population was a substitution, then substitutions and kmers capturing them were always the most significant variants after controlling for population structure. However, when gene presence conferred resistance, non-causal variants in LD with the causal gene were often associated at equal or higher significance than the causal gene. In these cases, the signal was still concentrated to one genomic region, for example the *blaI*, *blaR1*, *blaZ* region in *S. aureus* (Figure 3.20), or often a transposase found adjacent to a resistance-conferring gene was equally or more significant than the causal gene (Table 3.5). We hypothesise that this could be because there are multiple gene variants which can cause resistance within the dataset, but the plasmid or mobile element associated genes are less variable resulting in there potentially being more power to detect them. In other words, the conserved elements effectively pool signal across multiple resistance-conferring cargo

genes or alleles. Nevertheless, when testing phenotypes under less strong selection pressures, linked variants may be genome-wide and fine-mapping a challenge, highlighting the need for experimental validation going forward.

A concern with applying GWAS to identify antimicrobial resistance-conferring mechanisms was that variants were tested individually for an effect on the phenotype, when we know that the phenotypes are polygenic, each containing multiple resistance-conferring mechanisms. Despite this, we were able to identify a genuine causal variant or the surrounding region as the top hit in all but one study. This is likely to be a result of the dominance of particular resistance-conferring mechanisms within the population, and also the advantage of LMM of improved power when traits are polygenic. This may however be a challenge for future studies aiming to uncover all resistance-conferring mechanisms, but we have shown that when multiple causal variants fall within a kmer length, a kmer analysis can increase significance by pooling over these variants.

3.5.3 Application

If planning a study of this kind, it would be preferable to design a collection of an equal number of cases and controls for optimal power. In order to perform a full investigation into all significant variants, prospective studies would be the most powerful as sequencing could be designed with experimental randomisation of cases and controls. For example, the CRyPTIC project which aims to uncover all antimicrobial resistance-conferring mutations in *M. tuberculosis* launched in 2017 (University of Oxford 2017). The project aims to collect 100,000 samples from across the world, to sequence and measure their drug resistance profiles in order discover the rare mutations underlying resistance in cases where currently resistance cannot be predicted for. Large scale sample collections such as the CRyPTIC project aim to systematically investigate population-wide genomic variation in order to uncover novel resistance-conferring variants and

provide greater insights into the genetic variants underlying antimicrobial resistance.

The discovery of novel antimicrobial resistance and virulence determinants has the ability to transform clinical practice. Genotypic prediction of antimicrobial resistance and sensitivity is already applied routinely by PHE for *M. tuberculosis* (Public Health England 2017), but the more we can increase the catalogue of resistance-conferring mutations for all drugs and pathogens, the more accurately and rapidly resistance can be predicted clinically. Advances are already being made in software which use these catalogues to predict resistance. Bradley et al. (2015) created the software package 'Mykrobe predictor' which takes sequencing reads as input, currently from *S. aureus* and *M. tuberculosis*, and generates a report detailing resistance predictions for multiple antimicrobials within three minutes, in a user-friendly format that is simple to understand.

The use of emerging, more rapid real-time sequencing technology such as the single-molecule nanopore sequencer also promises to increase clinical diagnosis speed (Quick et al. 2015). Pilot studies using nanopore sequencing for resistance prediction have shown that resistance variants can be accurately and rapidly genotyped, however larger feasibility studies are needed due to the current high error rate of nanopore sequencing (Judge et al. 2015; Bradley et al. 2015; Votintseva et al. 2017). As the technology to sequence and predict resistance and other phenotypes is continually developing and improving, we now have the opportunity to exploit GWAS to discover variants contributing to important phenotypes such as resistance and virulence on a large scale.

The methods development and analyses in this chapter provided the basis for the subsequent analyses performed in this thesis, the genome-wide association studies of carriage versus invasive disease in *Neisseria meningitidis* and host adaptation to chickens and wild birds in *Campylobacter jejuni*.

Chapter 4

Genome-wide association study of *Neisseria meningitidis* carriage versus invasive disease

4 Genome wide-association study of *Neisseria meningitidis* carriage versus invasive disease

4.1 Introduction

The exclusively human pathogen *Neisseria meningitidis*, the meningococcus, is a major cause of bacterial septicaemia and meningitis worldwide (Stephens, Greenwood & Brandtzaeg 2007; Halperin et al. 2012). However, the meningococcus is principally a commensal, colonising the nasopharynx of between 10-35% of young adults estimated from point prevalence surveys in Europe and the United States (Caugant & Maiden 2009). Although the commensal nature of the majority of meningococcal infections is well understood, and host factors such as carriage state, complement deficiencies, social behaviour and geographic location have been associated with increased disease risk, the bacterial genetic factors underlying invasive disease remain to be fully elucidated (Caugant & Maiden 2009).

4.1.1 The meningococcal genome

There are several notable features of the genome of *N. meningitidis* enabling it to increase genome plasticity, and these may be important for evading the immune system, effectively colonising the human nasopharynx and invading to cause bacteraemia and meningitis. Most notable is the extent and variety of repetitive elements in the genome (Parkhill et al. 2000; Tettelin et al. 2000; Bentley et al. 2007). These range from simple sequence repeats to gene cluster duplications. The most common repeats are the *Neisseria* DNA uptake sequences (5'-GCCGTCTGAA-3) enabling the recognition and uptake of DNA during transformation (Goodman & Scocca 1988). Other repetitive elements include Neisserial intergenic mosaic elements (NIMES) and the Correia repeat enclosed elements (CREE) (Correia, Inouye & Inouye 1988; Parkhill et al. 2000).

Neisseria are naturally competent for DNA uptake; the meningococcus has a high rate of recombination relative to mutation, quantified by an *r*/*m* of 80, and a large accessory genome (Feil et al. 1999; Joseph et al. 2011). However, despite the substantial meningococcal diversity, the population is also highly structured and can be categorised into clonal complexes defined by related seven-locus MLSTs and further into phylogenetic clades consisting of related clonal complexes (Budroni et al. 2011).

Another mechanism of diversity generation in meningococci is that of phase variation at contingency loci. Phase variation is the switching of gene expression states purported to enable adaptation to changes in environmental conditions in a clonal population. This typically occurs by slipped strand mispairing: the shortening and lengthening of repetitive tracts during replication (Levinson & Gutman 1987; Murphy et al. 1989; Henderson, Owen & Nataro 1999). Many genes have been identified as putative phase variable genes, although not all have been experimentally proven (Saunders et al. 2000; Martin et al. 2003; Siena et al. 2016).

4.1.2 The meningococcal capsule

The genetic diversity and dynamic gene content of the meningococcus is therefore thought to allow it to escape host responses. Molecular work has identified the importance of the meningococcal capsular polysaccharide (CPS) in enabling the bacterium to evade complement-mediated and phagocytic killing (Jarvis & Vedros 1991; Spinosa et al. 2007). Meningococci are divided into 13 serogroups based on the serological differences of their capsular polysaccharide. Of these, six (A, B, C, W, X, Y) cause the majority of invasive meningococcal disease (Rosenstein et al. 2001; Halperin et al. 2012). Carriage isolates are frequently unencapsulated, either because they lack the genetic island encoding capsule synthesis where it has been replaced by a noncoding region termed the capsule null locus (Claus et al. 2002), or because they downregulate

expression of the capsule (Hammerschmidt et al. 1996). Rare cases of unencapsulated meningococci causing invasive disease have been reported in immunodeficient patients (Vogel et al. 2004; Ganesh et al. 2017) and even more rarely in immunocompetent patients (Hoang et al. 2005). In addition, many carriage isolates have also been demonstrated to express capsules associated with disease (Claus et al. 2005; Jolley et al. 2000; Yazdankhah et al. 2004). Taken together, these observations suggest that the capsule is necessary but not sufficient to cause invasive disease.

4.1.3 The role of factor H binding protein in meningococcal disease

The meningococcus also evades immune responses by recruitment of the human complement factor H (CFH) via interactions with bacterial factor H binding protein (fHbp) (Lo, Tang & Exley 2009). CFH is the major negative regulator of the alternative complement pathway (Walport 2001). Complement deficiencies have been shown to be risk factors for invasive meningococcal disease (Densen 1989) and the complement system has further been shown to be crucial in protection against invasive infection by a GWAS performed in humans investigating host susceptibility to meningococcal disease (Davila et al. 2010). The GWAS of 475 cases and 4,703 population controls from the United Kingdom reported significant associations at SNPs within *CFH* and CFH-related protein 3 (Davila et al. 2010). The most significant SNPs were also validated within two replication studies performed in Western European and South European cohorts of 968 cases and 1,376 controls, suggesting that host variation within these two regulators of the complement pathway are important in determining the development of meningococcal invasive disease (Davila et al. 2010). This was further validated within a Central European cohort of 248 patients with invasive meningococcal disease and 835 healthy population controls which also reported that carriers of the minor alleles of the two SNPs

were at lower risk of invasive meningococcal disease, strengthening the association (Biebl et al. 2015).

fHbp is a component of the two licensed serogroup B meningococcal vaccines (Carter 2013; Shirley & Dhillon 2015). Sequence analysis has revealed that the protein can be divided into either three variants or two subfamiles (Masignani et al. 2003; Fletcher et al. 2004). Levels of fHbp expression vary across strains, but this has thus far not been explained (Oriente, Scarlato & Delany 2010). Biagini et al. (2016) aimed to quantify the expression of fHbp across strains representative of meningococcal serogroup B and associate that with genetic background. They used selected reaction monitoring to quantify the protein levels of 105 serogroup B isolates and found that expression of the protein is genetically determined and linked to the promoter sequence, and also that strains carrying fHbp var1 express significantly more fHbp compared with var2 and var3 strains. However, the different categories of promoter sequence are highly connected with the three fHbp variants, so it is not clear that the differential expression across the promoter sequence variants represents any furthur association above that of the fHbp variants. There was also no sequence variation across strains within the promoter sequence identified by Oriente, Scarlato & Delany (2010).

4.1.4 Early association studies investigating virulence in *N. meningitidis*

Colonisation with certain so-called hyperinvasive strains is a risk factor for invasive disease, but the bacterial genetic basis underlying hyperinvasiveness is not well understood (Caugant & Maiden 2009). Importantly, the existence of hyperinvasive strains suggests a genetic basis to invasive disease which is as yet not fully elucidated. A growing number of early association studies have investigated the genetic basis behind carriage versus invasive disease and have found that there are no genes present in all hyperinvasive strains that are absent in carriage strains or non-pathogenic *Neisseria*

species (Snyder & Saunders 2006; Schoen et al. 2008). Nor is there convincing evidence of sequence variation distinguishing carriage from invasive isolates at a population level; Katz et al. (2011) investigated using SNPs as a way to discriminate between eight disease-associated genomes and four carriage-associated genomes. Although they found discriminating SNPs, the study was small in scale and did not control for population structure.

Bille et al. (2005) investigated a larger strain collection of 29 disease-associated isolates representing the major hyperinvasive lineages and 20 isolates belonging to lineages with no disease associations. Isolates were classed as disease-associated if part of an ST-complex containing an invasive strain. Likewise, isolates were defined as belonging to a noninvasive group if no invasive isolates were in the same ST-complex. Whole genome comparisons were perfomed by DNA arrays to portions of the reference strain Z2491 and although no gene was universally present in the invasive isolates but absent in the carriage isolates, the presence of a group of genes corresponding to a prophage was associated with hyperinvasive lineages. It was later shown that the prophage encodes T and B cell stimulating protein (TspB), an immunoglobulin-binding protein specific for the Fc portion of a human IgG2 paraprotein shown to be essential for the survival of a serogroup B strain in normal human serum (Müller et al. 2013; Müller et al. 2015). The gene cluster was present in all disease-associated isolates but only 10% of noninvasive isolates. However, the phenotype was not clearly defined as some carriage isolates were classed as disease-associated due to being within an ST-complex that contained invasive isolates.

The presence of the phage was however further tested by PCR in a larger strain collection of 293 isolates from the Czech Republic where it was found to be significantly associated with disease even after accounting for the confounding clonal association with

invasive complexes and the region was designated the "meningococcal disease associated island" (Bille et al. 2005). This however did not replicate in a study by Dunning Hotopp et al. (2006) of 48 *N. meningitidis* isolates where the island was present in 60% of virulent strains and 42% of carriage strains. Thus it does not appear to be characteristic of invasiveness generally but may be a marker for particular hypervirulent strains such as the ST-11s examined in Bille et al. (2005). The presence of the phage was further investigated by Bille et al. (2008) in 1,288 carriage and disease isolates sampled from South East England. The phage was more likely to be identified in disease isolates than carriage isolates, although this association was weakened once clonal complex had been adjusted for with confidence intervals of the odds ratio spanning 1. A stronger association was determined however between the phage and disease isolates in young adults between 13-28 years after adjusting for clonal complex (odds ratio 3.9; 95% CI 1.1-13.6) (Bille et al. 2008). Therefore, the phage may be a marker for particular hypervirulent lineages in young adults (Dunning Hotopp et al. 2006; Bille et al. 2008).

A recent GWAS investigating the genetic basis of carriage versus invasive disease in *N. meningitidis* was performed by Collins & Didelot (2017) who tested core SNPs and accessory gene presence/absence of 129 European serogroup C isolates for association with disease status. The presence of three genes were found to be associated with invasive disease, and nine were associated with carriage. These included *NadA*, *MafA2*, *hmbR*, NEIS0956, NEIS1574, NEIS1880 and NEIS1996. Seven core SNPs were also associated with either carriage or invasive disease including variants in *porA* and *gapA2*. Among the associated genes and variants some have roles which may explain their association to invasive disease, such as roles in adhesion and colonisation of mucosal cells, and others will require further investigation (Collins & Didelot 2017).

4.1.5 Within host evolution

Investigating invasive disease from a slightly different angle, Klughammer et al. (2017) applied ultra-deep whole genome sequencing to pairs of throat and blood strains isolated from four patients with acute invasive meningococcal disease. Klughammer and colleagues aimed to test the within host evolution hypothesis that commensal pathogens evolve virulence/host damage within hosts as a result of shortsighted within-host competition, particularly due to variation at phase variable loci. Isolates were sequenced on the Illumina HiSeq 2000 platform recovering average coverage of 1500-fold per sample and reads were aligned to assemblies generated from the corresponding throat isolates which were additionally sequenced on the Roche 454 GS FLX platform in order to achieve longer read lengths. After filtering out low quality variants, any which were unique to either the throat or blood were sequenced by Sanger sequencing, which validated 13 SNPs and eight indels. The variants identified predominantly affected phase variable genes and in particular genes with roles in type IV pilus biogenesis, however the mutations identified were not found uniformly across the blood isolates from the four patients. The variants were also not associated with increased serum resistance, higher fitness in human blood *ex vivo* or interaction strengths with human epithelial cells and endothelial cells *in vitro*. Klughammer and colleagues therefore hypothesised that meningococcal virulence is due to the accidental emergence of invasive variants during carriage rather than adaptive within host evolution during disease progression.

GWAS offer new opportunities to understand the genetic basis underlying important traits, and here we investigated the bacterial genetic basis to invasive disease in *N. meningitidis*. We applied a GWAS aiming to identify bacterial genetic elements associated with invasive disease by investigating 261 isolates sampled from the Czech Republic, representing the meningococcal population structure present at the time of

sampling and representing the wider meningococcal population structure. Understanding the factors behind the rare transition from asymptomatic carriage to invasive meningococcal disease in natural populations is key to tackling the disease. Potentially novel virulence factors identified by these methods could provide candidate targets for vaccine development, assisting approaches in treating and preventing meningococcal disease.

4.2 Chapter aims

- Investigate whether particular meningococcal lineages are associated with either carriage or invasive disease and whether we can identify hypervirulent strains.
- Quantify the heritability of invasive meningococcal disease from the bacterial genetic perspective.
- Search for signals of association between individual bacterial genetic variants and invasive meningococcal disease.

4.3 Methods

In this chapter we applied a SNP GWAS and a kmer GWAS to investigate the bacterial genetic basis of carriage versus invasive disease in *N. meningitidis*. By testing both SNPs and kmers, we hoped to capture SNP variation, gene presence/absence plus small insertions and deletions. Kmers were counted from both sequencing reads and Velvet assemblies. We built a maximum likelihood phylogeny using RAxML and imputed missing SNPs using ClonalFrameML. We tested for lineage associations by defining lineages as Principal Components (PCs) and applied a Wald test to investigate their association with the phenotype. Variants were tested for association using LMM in the software GEMMA. Methods detailing the association analyses can be found in Chapter 2.

4.3.1 Sampling frame

The dataset consisted of 261 isolates representative of the *N. meningitidis* population sampled from the Czech Republic in 1993 (Jolley et al. 2000). All isolates had a known carriage or invasive status which was used as the phenotype, and the dataset contained 209 carriage isolates and 52 invasive disease isolates.

4.3.2 Whole genome sequencing and variant calling

Illumina sequencing reads were downloaded from the European Nucleotide Archive (ENA, http://www.ebi.ac.uk/ena), and Velvet *de novo* assemblies were downloaded from pubMLST (https://pubmlst.org/neisseria/) in collaboration with Martin Maiden. SNPs were called using standard methods, using Stampy (Lunter & Goodson 2011) to map reads to the reference strain FAM18 (accession AM421808) as described in Chapter 2.1.1 and were imputed using ClonalFrameML (Didelot & Wilson 2015) as described in Chapter 2.3. Kmers were counted from both Illumina sequencing reads and from Velvet contigs using dsk (Rizk, Lavenier & Chikhi 2013) as described in Chapter 2.4.

4.3.3 Determining the proportion of reads assigned to *Neisseria meningitidis*

Before proceeding with the analysis, quality control measures were taken. Kraken (Wood & Salzberg 2014) was used to determine the proportion of kmers belonging to all species in its database. This revealed that 97-99.7% of kmers matched to *Neisseria*, 86.9-94.2% to *Neisseria meningitidis*, 5.3-10% to unclassified *Neisseria* species and 0.0029-0.39% to non-*Neisseria* taxa (Figure 4.1). Non-*Neisseria*-matching kmers were very few in number, although genomes identified as ST-11 contained a higher proportion of unclassified *Neisseria* kmers, which could reflect their greater divergence to the *N. meningitidis* genomes represented in the Kraken database. All genomes were considered to have passed QC and were therefore used in the analyses.

Variant type	Biallelic SNPs	Tri-allelic SNPs	Tetra-allelic SNPs	Kmer variants identified from sequencing reads	Kmer variants identified from Velvet assemblies
Number of variants	150502	6063	239	14680679	7806583
Number of unique variant phylopatterns	73888	5973	238	1430549	307830

Table 4.1 Total numbers of variants and unique variant phylopatterns across individuals for all variant types investigated.

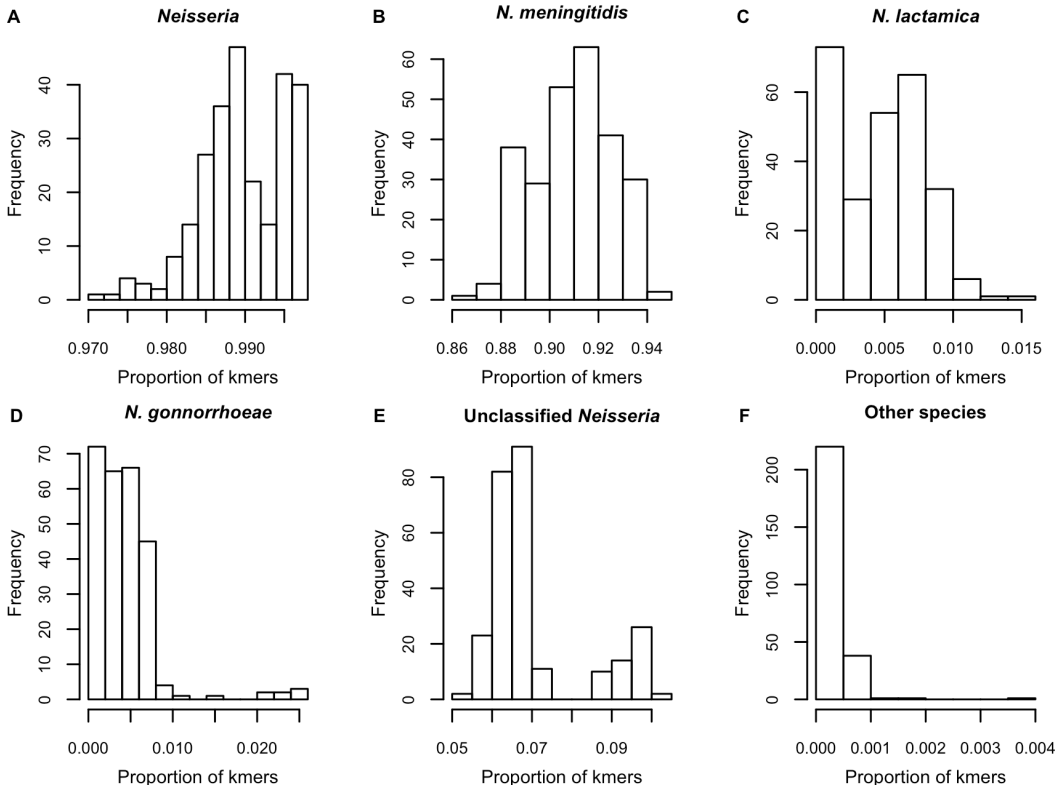

Figure 4.1 The proportion of reads assigned to *Neisseria* species using Kraken. A *Neisseria* genus; **B** *N. meningitidis*; **C** *N. lactamica*; **D** *N. gonnorrhoeae*; **E** Kmers which could be assigned to the *Neisseria* genus but could not be further classified to a species; **G** Non-*Neisseria* species.

4.3.4 Variant counts

Identifying kmers from sequencing reads revealed 14,680,679 variably present or absent kmers defined by 1,430,549 unique phylopatterns. Identifying kmers from Velvet assemblies revealed 7,806,583 variably present kmers, defined by 307,830 unique phylopatterns. The numbers of all variants are summarised in Table 4.1. Bonferroni-corrected significance thresholds were 6.2, 6.9, 6.8 and 3.1 for SNPs, kmers counted from

reads, kmers counted from Velvet assemblies and lineages, respectively.

4.3.5 Calculating standardised kmer counts

Standardised kmer counts were calculated for the kmers defined from sequencing reads. Total sequence length covered by the kmers was calculated by multiplying the number of unique kmers per genome by the kmer length. This was then divided by the Velvet assembly length to produce a mean kmer depth per genome. Standardised kmer counts were then calculated by dividing the raw kmer count by the expected kmer count based on the mean kmer depth.

4.3.6 Predicting stop codon position and phase variable state in phase variable genes

We used BLAST (Camacho et al. 2009) to identify genes of interest in the Velvet assemblies of each isolate. Stop codon positions were predicted by examination of the BLAST alignments. Phase variable state was predicted by defining whether a premature stop codon was present within the gene, or by the presence of frameshift mutations.

4.4 Results

4.4.1 Population structure of the sampling frame

We built a maximum likelihood phylogeny of the Czech Republic data from the biallelic SNPs using RAxML (Stamatakis 2014) as described in Chapter 2.2. The midpoint rooted phylogeny is depicted in Figure 4.2 annotated by clonal complex (CC) designation, serogroup and carriage versus invasive disease status. Bootstrap supports are shown in Appendix B Figure B.2. The population segregates into clonal lineages, as defined by clonal complexes, however is quite star like at the root of the tree representing the high level of recombination within the species. The most common CC was the ST-11 complex at 52 isolates. Other common CCs included the ST-41/44 complex, ST-92 complex, ST-106 complex and ST-116 complex at 30, 21, 18 and 13 isolates respectively. There was a

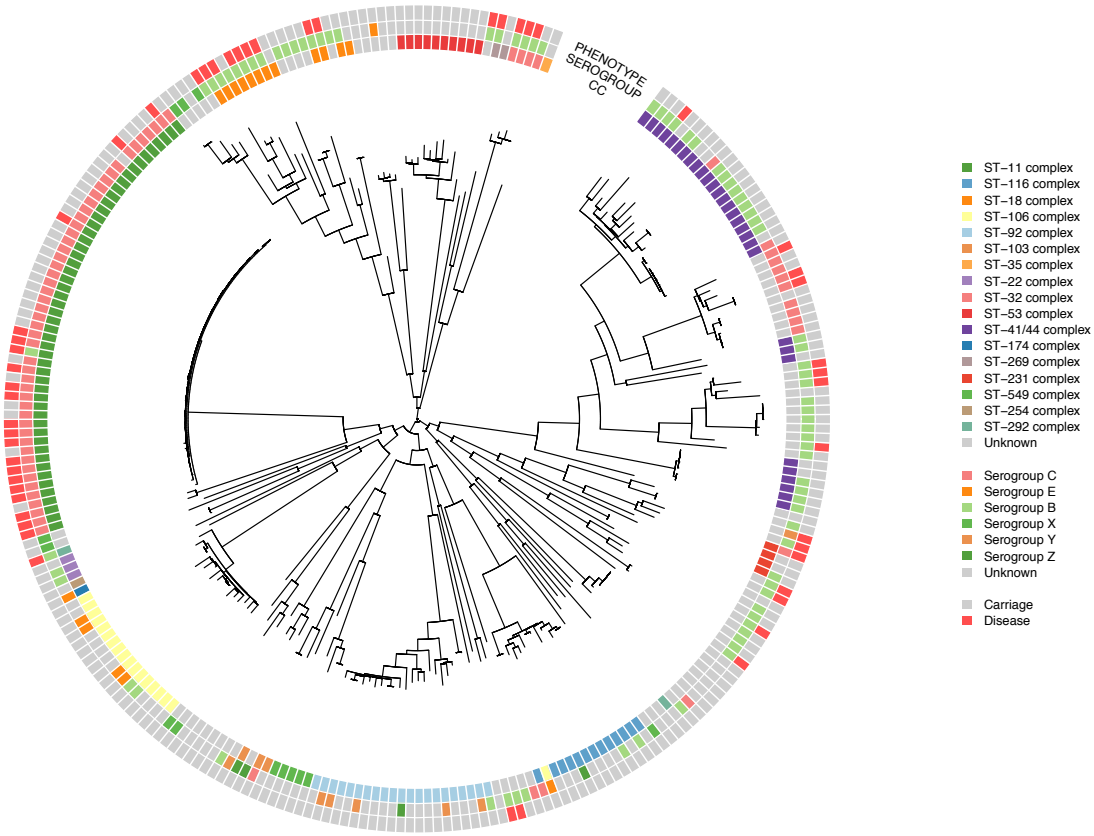

Figure 4.2 ***Neisseria meningitidis*** **phylogeny.** Maximum likelihood phylogeny built using RAxML based on biallelic SNPs and annotated with ST-complex, serogroup and carriage/invasive disease status. Bootstrap supports are shown in Appendix B Figure B.2.

high sampling frequency of low diversity ST-11 isolates within the sampling frame due to an outbreak of ST-11 isolates, a so-called hyperinvasive strain, which occurred at the time of sampling within the Czech Republic (Krízová, Musílek & Kalmusová 1997; Jolley et al. 2000).

4.4.2 The ST-11 lineage was associated with invasive disease

In this section, we investigated the proportion of phenotypic variation explained by the bacterial genotype, and based on this identified how well the phenotype could be predicted from the genetic information. We identified lineages defined by principal components, and tested for their association with the phenotype.

4.4.2.1 Heritability of the phenotype and making phenotype predictions

Heritability, the proportion of the phenotypic variation that can be explained by the

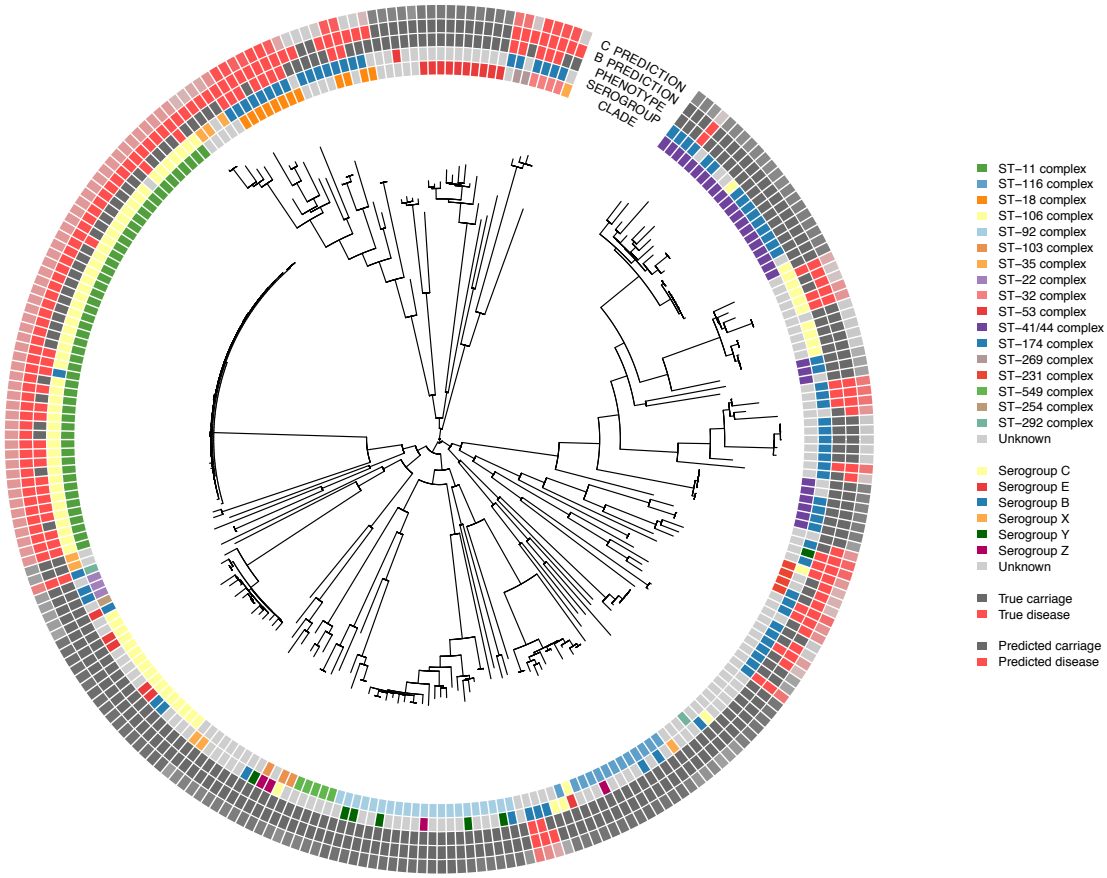

Figure 4.3 Predicting carriage/disease phenotype using SNP data. Phylogeny annotated by ST-complex, serogroup, true phenotype, binary predicted phenotype ('B PREDICTION') and continuous predicted phenotype ('C PREDICTION'). The continuous prediction goes from dark grey (strongly predicted carriage) through light grey to red (strongly predicted invasive).

bacterial genotype, was estimated from biallelic SNPs using the LMM null model in GEMMA (Zhou & Stephens 2012) as described in Chapter 2.9.1. The heritability of invasive disease/carriage status in the sample was estimated to be 36% with a standard error of 10%, so there was a reasonably strong genotype to phenotype relationship. This enabled reasonably accurate predictions of the phenotype using the bacterial genetic data (outer rings, Figure 4.3). The phenotype was predicted using the null LMM in which every SNP is used for prediction, which is equivalent to a ridge regression (O'Hagan & Forster 2010) as described in Chapter 2.10.3. The phenotypes were coded as invasive disease (1) and carriage (0) and centred, meaning positive values represented invasive disease isolates and negative values represented carriage isolates. We made binary

invasive disease versus carriage predictions by identifying whether the predictions were positive or negative, and by doing so correctly predicted all truly invasive isolates to be invasive. Fifty-eight of the true carriers were also predicted to be invasive, of which 32 were of the ST-11 complex, as all ST-11 complex isolates were predicted to be invasive. A finer scaled prediction was achieved by interpreting the continuous predicted value itself. Often when the prediction was incorrect, the prediction was weaker than for correctly predicted phenotypes (outer rings, Figure 4.3).

4.4.2.2 Wald test for lineage effects

Principal components analysis (PCA) was performed on the biallelic imputed SNPs and a Wald test was applied to assess the significance of associations between principal components (PCs) and carriage versus invasive disease. PCs 1-20 were visualised on the phylogeny of the dataset in Figure 4.4. One can see that the leading PCs trace paths through the deepest branches of the tree, validating using PCs to define lineages in this dataset.

After correcting for the number of lineages tested using a Bonferroni correction on the number of PCs, we found PC-1 to be significantly associated with the phenotype (Figure 4.5C). Visualising PC-1 on the phylogeny of the dataset (Figure 4.4 and Figure 4.5B) revealed that the lineage represented the ST-11 complex, the hyperinvasive outbreak lineage at the time and location of sampling. This shows that testing for lineage effects can identify known hyperinvasive lineages, and therefore could be used to identify unknown invasive lineages.

We have therefore shown that 36% (standard error = 10%) of the phenotypic variation within the Czech Republic data could be explained by the bacterial genetic data. Because of this, we were able to make reasonable predictions of the phenotype using the bacterial genetic data, where all truly invasive isolates were predicted to be invasive. The

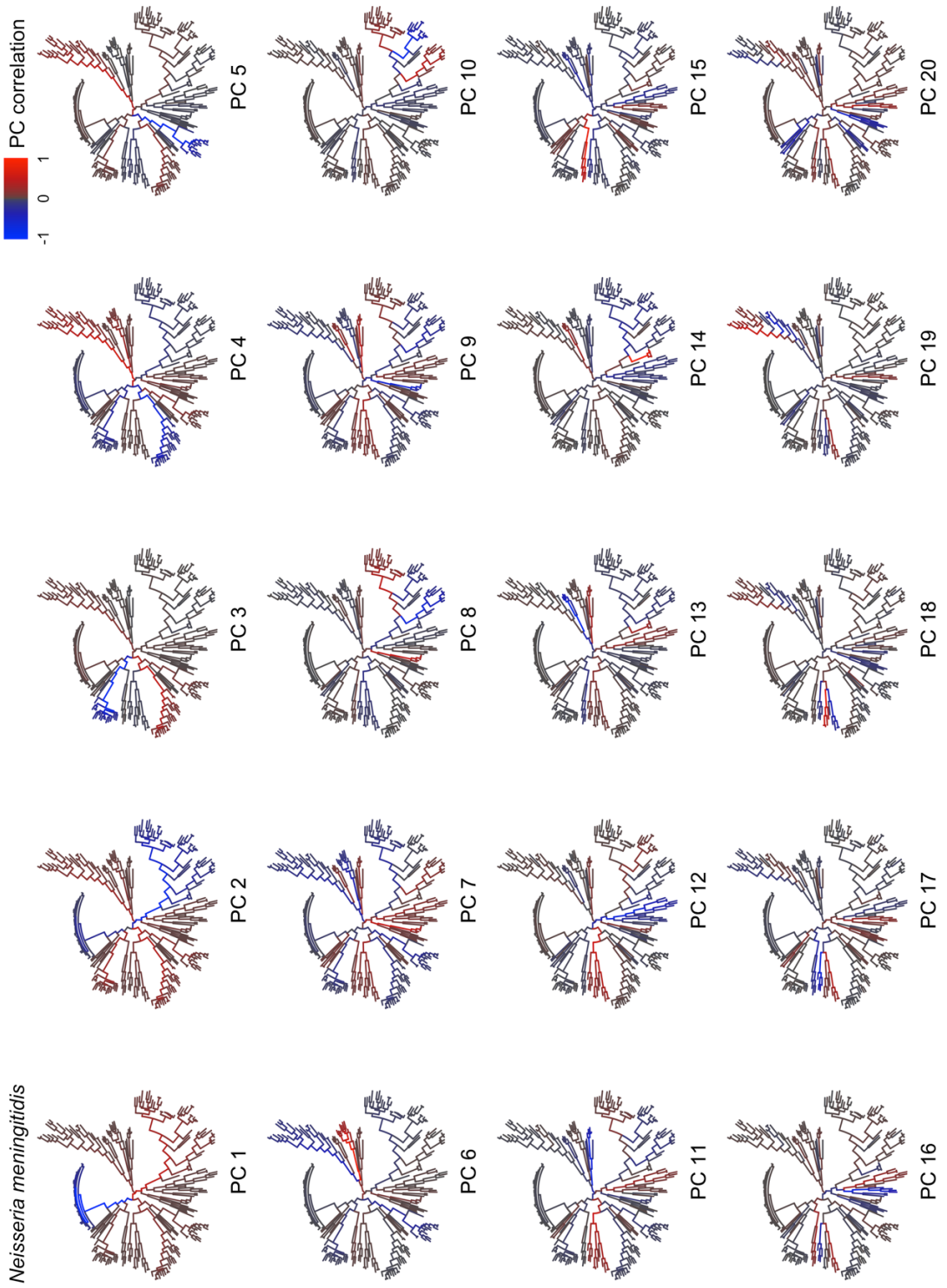

Figure 4.4 *N. meningitidis* principal components annotated on the phylogeny. Branches are coloured by the correlation of the branch pattern with the projections of the individuals onto the principal components. The branch patterns were defined by assigning a 1 to all isolates one side of the branch and a 0 to all isolates the other side of the branch. The colour depicts the correlation of this pattern with the projection of the isolates onto each PC. Branch lengths were square rooted to enable visualisation of the branch colours at the tips of the phylogeny. Tracing from blue to red through the tree shows how the leading principal components trace paths through the tree and correspond to major lineages.

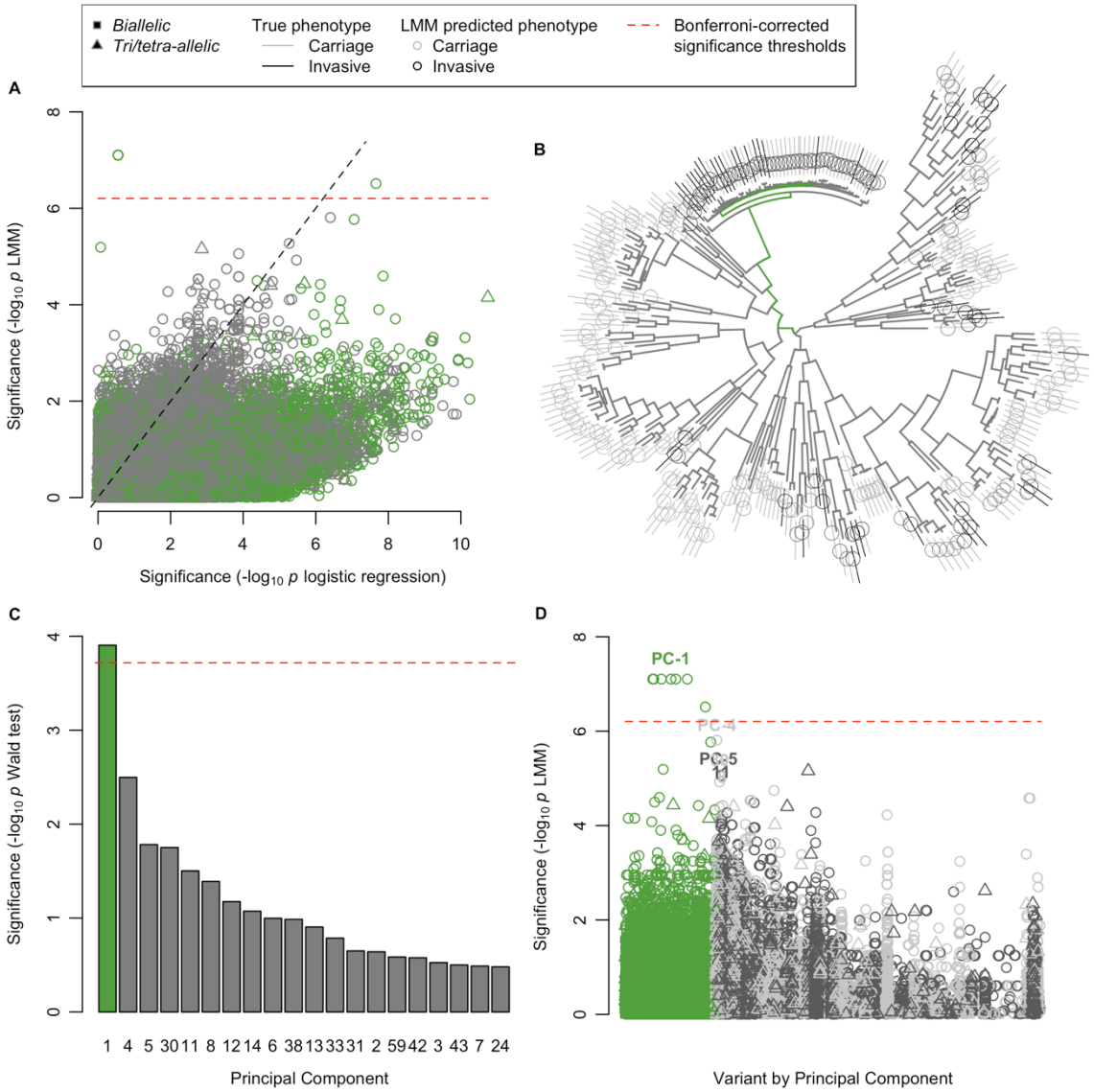

Figure 4.5 Results of applying 'bugwas' to test for lineage effects. A SNP $-\log_{10} P$ values before and after controlling for population structure. Variants associated with lineage PC-1 tended to decrease in significance after controlling for population structure. **B** Maximum likelihood phylogeny of the dataset with the branches coloured by whether they were most correlated to PC-1. Tips show true and predicted phenotypes. **C** Results of the Wald Test for lineage effects, testing for association between PCs and the phenotype. Only the coloured PC-1 was significant after controlling for the number of lineages tested. **D** SNP $-\log_{10} P$ values after controlling for population structure ordered on the x-axis by which principal component they were most correlated to in the order of the Wald test results.

Wald test for lineage effects further revealed lineage PC-1, corresponding to the ST-11 hyperinvasive lineage, to be significantly associated with the phenotype, demonstrating how we can identify known hypervirulent lineages using genetic data.

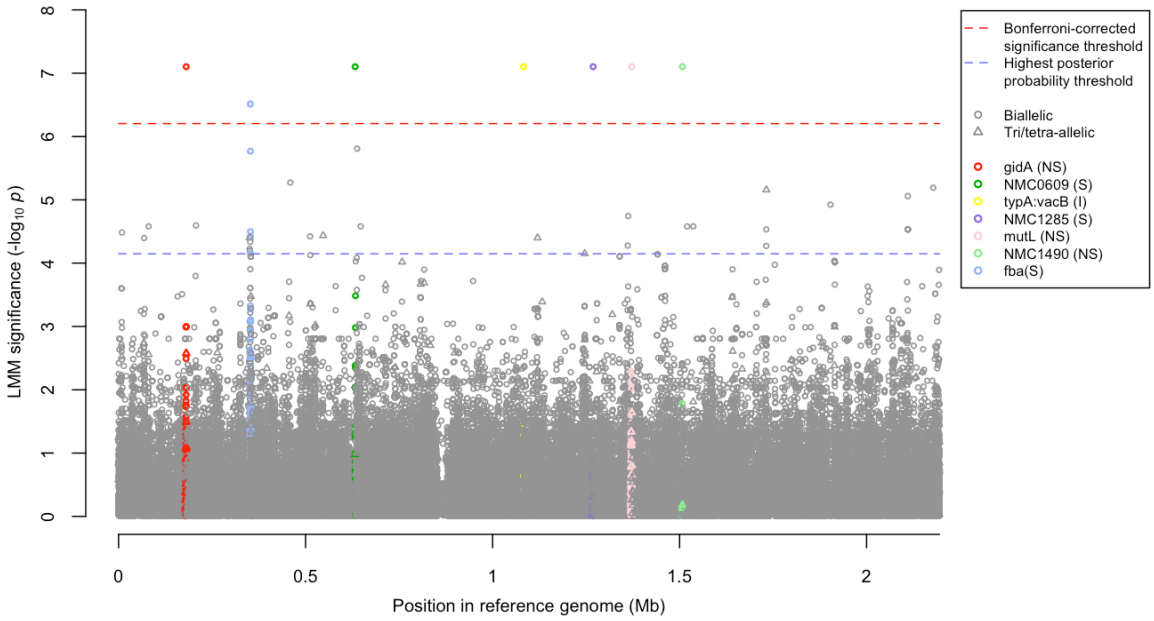

Figure 4.6 SNP significance after controlling for population structure using LMM. SNP positions refer to their position in the reference genome FAM18. Genes and intergenic regions containing significant SNPs are coloured. Brackets following the gene names indicate the type of SNP which is significant within the gene/intergenic region: NS = Non-synonymous; (S) = Synonymous; I = Intergenic.

4.4.3 Seven SNPs were associated with carriage versus invasive disease

4.4.3.1 Identification of SNPs associated with invasive disease

We found seven SNPs to be significantly associated with carriage versus invasive disease by LMM, using a Bonferroni adjusted significance threshold on the number of unique SNP phylopatterns across individuals to account for multiple testing (Figure 4.6). As LMM can increase the significance of low frequency SNPs and not sufficiently control for population structure for low frequency SNPs (Mathieson & McVean 2012), the frequency of the significant SNPs was assessed. All significant SNPs had a minor allele frequency of >16%. The top six SNPs were in the genes *gidA*, NMC0609, NMC1285, *mutL*, NMC1490, and the intergenic region between *typA* and *vacB*. The minor alleles for these six SNPs were present in the exact same individuals, meaning they had the same SNP phylopattern across individuals and were in complete genome-wide linkage disequilibrium (LD), making it difficult to fine map the signal and determine which could

be causal. We assessed the localised regions to look for signatures of local LD surrounding the SNPs to add support to the associations. For the six most significant SNPs in perfect LD, none showed the expected signature of inflated significance around the significant SNP and decay of significance around it (Figure 4.7) (The Wellcome Trust Case Control Consortium 2007; as described in Section 1.1.3.1).

The other significant SNP was a synonymous SNP in the gene *fba* adjacent to the gene which codes for factor H binding protein (fHbp). This SNP had the signature of inflated significance surrounding the SNP and decay in significance with increasing distance from the SNP (Figure 4.8). A second synonymous SNP in *fba* 3bp from the significant SNP was also just below the genome-wide significance threshold (Figure 4.8).

Of the seven significant SNPs, just those in NMC1285 and *fba* were imputed in some individuals. The major allele for NMC1285 was imputed into 22 individuals, and the major allele for *fba* was imputed into 1 individual, therefore imputation did not drastically change the underlying signal.

We calculated approximate posterior probabilities for the biallelic SNPs, conservatively assuming one causal SNP. We subsequently calculated the highest posterior probability threshold such that the sum of the posterior probabilities exceeding the threshold was ≥95%. Only 38 SNPs fell above this threshold, out of a possible 150,502 SNPs, giving confidence that the signal of association was concentrated among the small number of most significant SNPs (Figure 4.6). The six most significant SNPs each had a posterior probability of 14.6% and the SNP in *fba* had a posterior probability of 3.9%.

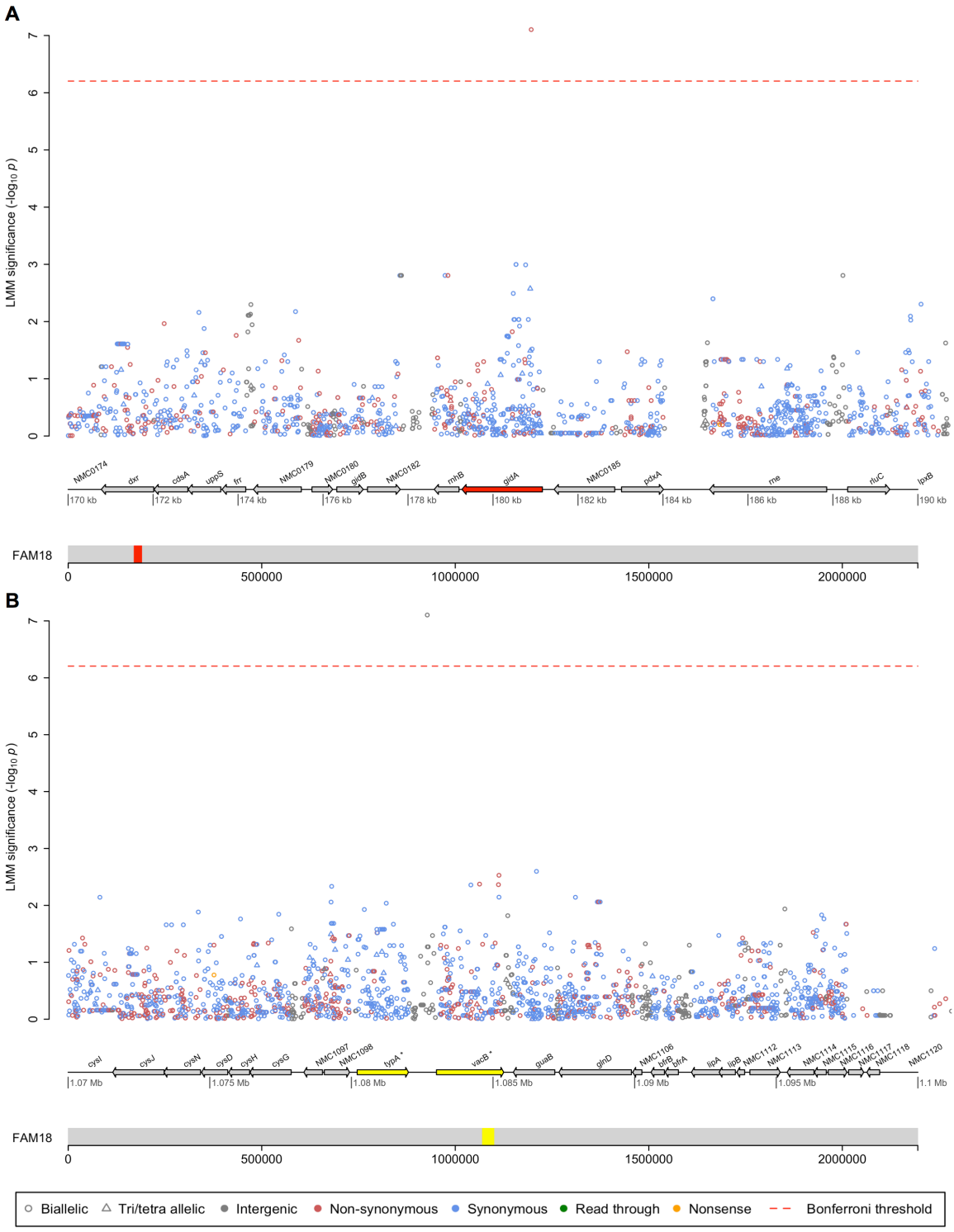

Figure 4.7 Close ups of the SNP Manhattan plot focusing on the top 6 most significant SNPs. SNPs are coloured: grey = intergenic; red = non-synonymous; blue = synonymous; green = read-through; yellow = nonsense. The coloured portion of the lower bar depicts the region of the reference genome that is being shown. **A** *gidA*; **B** *typA*:*vacB*; **C** *mutL*; **D** NMC0609; **E** NMC1285; **F** NMC1490. (*Continued on the next page).*

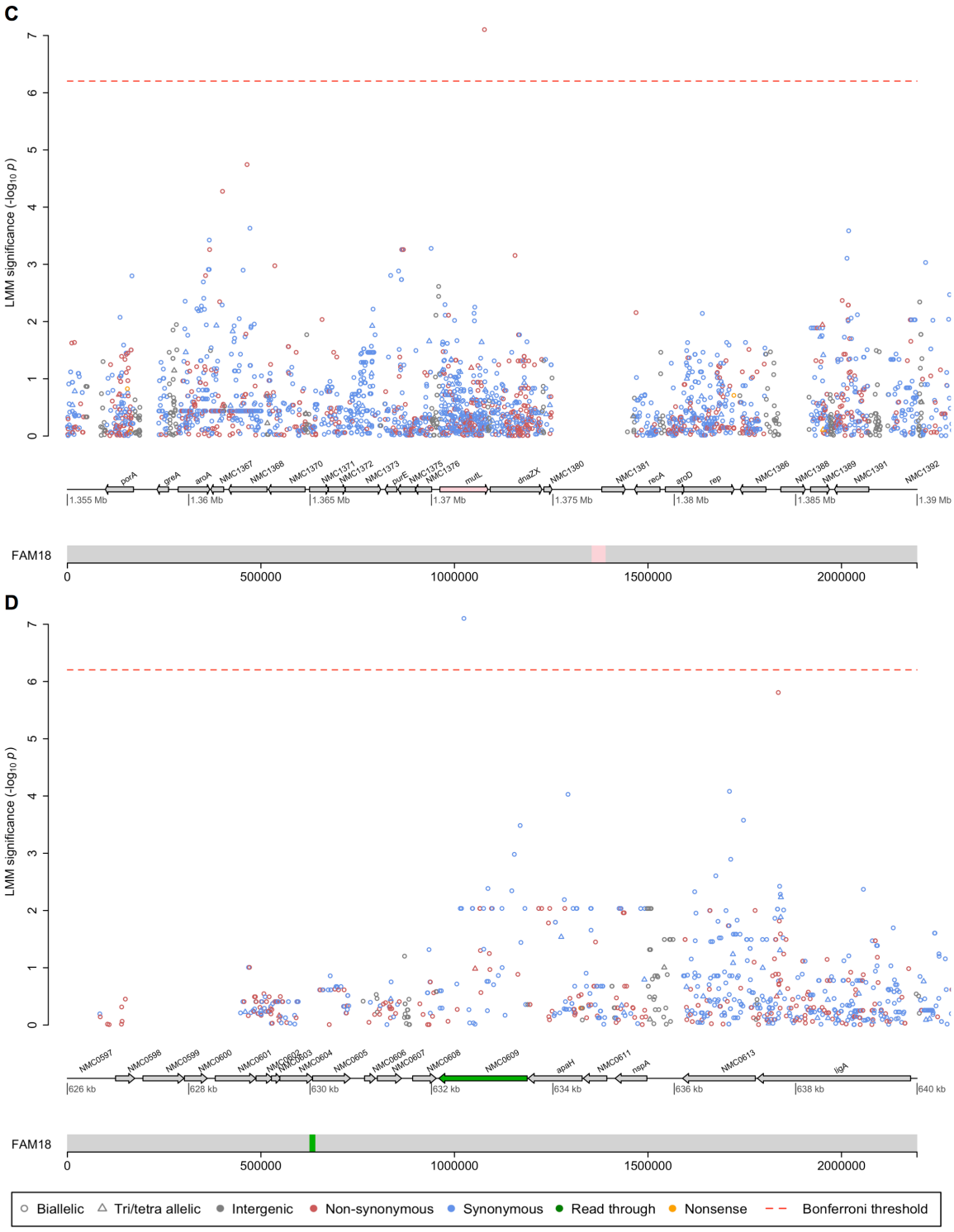

Figure 4.7 (Contd.) Close ups of the SNP Manhattan plot focusing on the top 6 most significant SNPs. SNPs are coloured: grey = intergenic; red = non-synonymous; blue = synonymous; green = read-through; yellow = nonsense. The coloured portion of the lower bar depicts the region of the reference genome that is being shown. **A** *gidA*; **B** *typA*:*vacB*; **C** *mutL*; **D** NMC0609; **E** NMC1285; **F** NMC1490. (*Continued on the next page)*.

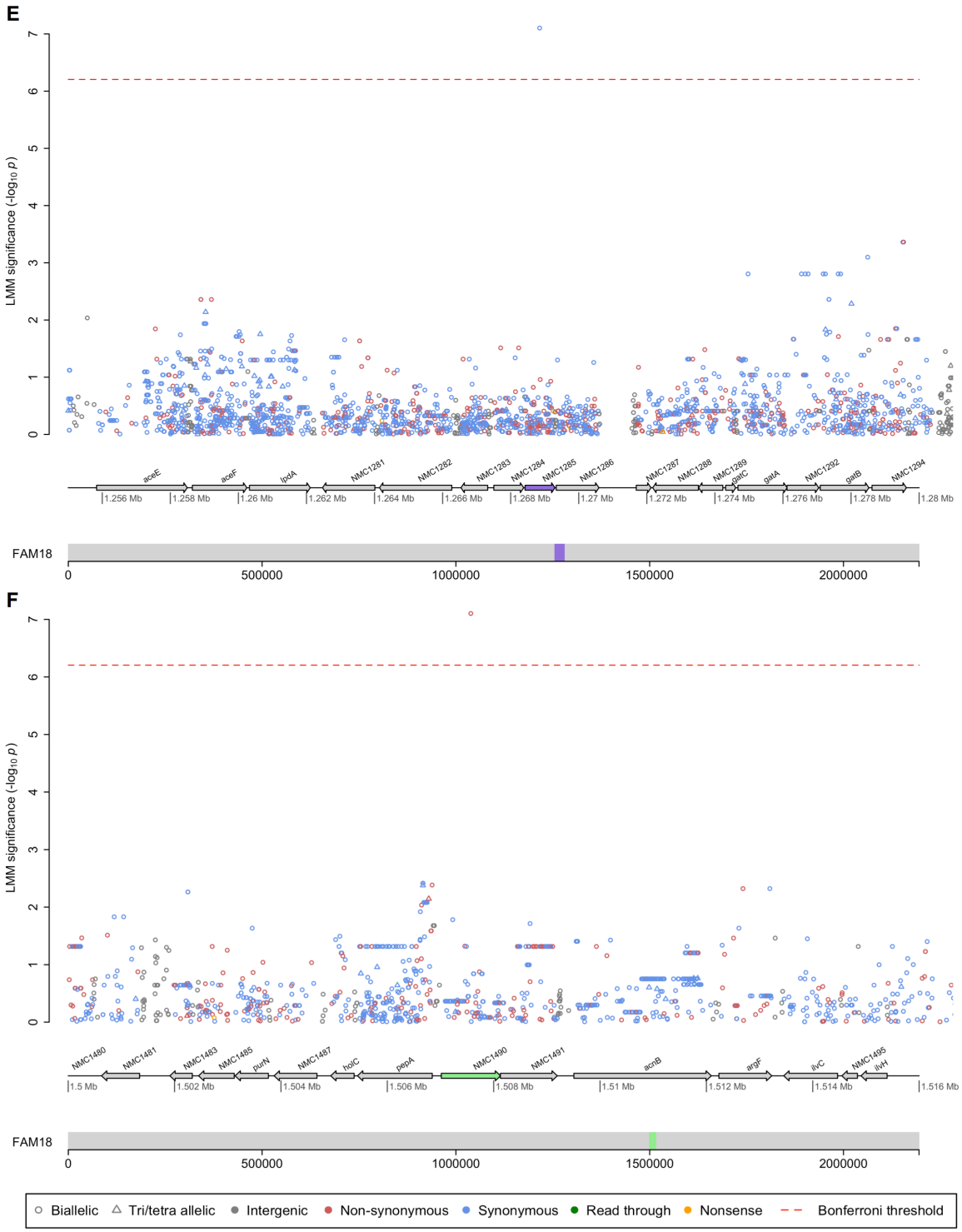

Figure 4.7 (Contd.) Close ups of the SNP Manhattan plot focusing on the top 6 most significant SNPs. SNPs are coloured: grey = intergenic; red = non-synonymous; blue = synonymous; green = read-through; yellow = nonsense. The coloured portion of the lower bar depicts the region of the reference genome that is being shown. **A** *gidA*; **B** *typA*:*vacB*; **C** *mutL*; **D** NMC0609; **E** NMC1285; **F** NMC1490.

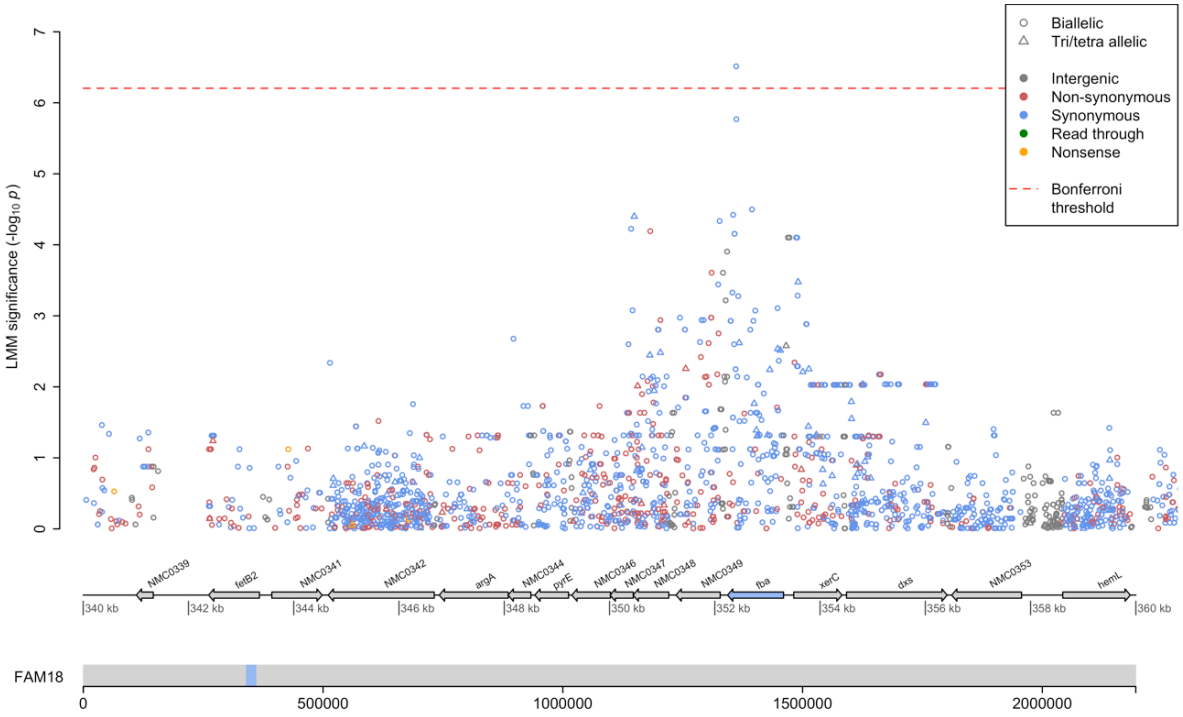

Figure 4.8 Close up of the SNP Manhattan plot focusing on the fba/fHbp region. SNPs are coloured: grey = intergenic; red = non-synonymous; blue = synonymous; green = read-through; yellow = nonsense. The coloured portion of the lower bar depicts the region of the reference genome that is being shown.

In the human setting, it has been shown that including PCs as additional fixed effects in the LMM can improve power and reduce false positives (Tucker, Price & Berger 2014; Widmer et al. 2014). Therefore, in order to ensure that population structure had been controlled for sufficiently and to check the robustness of the results, PCs were included as additional fixed effects in the LMM and results compared with the original LMM results. PCs 1, 1-5, 1-10 and 1-20 were all separately included. All originally significant SNPs remained significant in all tests giving confidence that the control for population structure was sufficient (Figure 4.9).

4.4.3.2 A paired SNP analysis indicated that the significant SNP phylopatterns were independent signals

In order to determine whether the significant SNPs represented independent signals we conducted a paired analysis. All biallelic SNPs were tested using LMM including each of

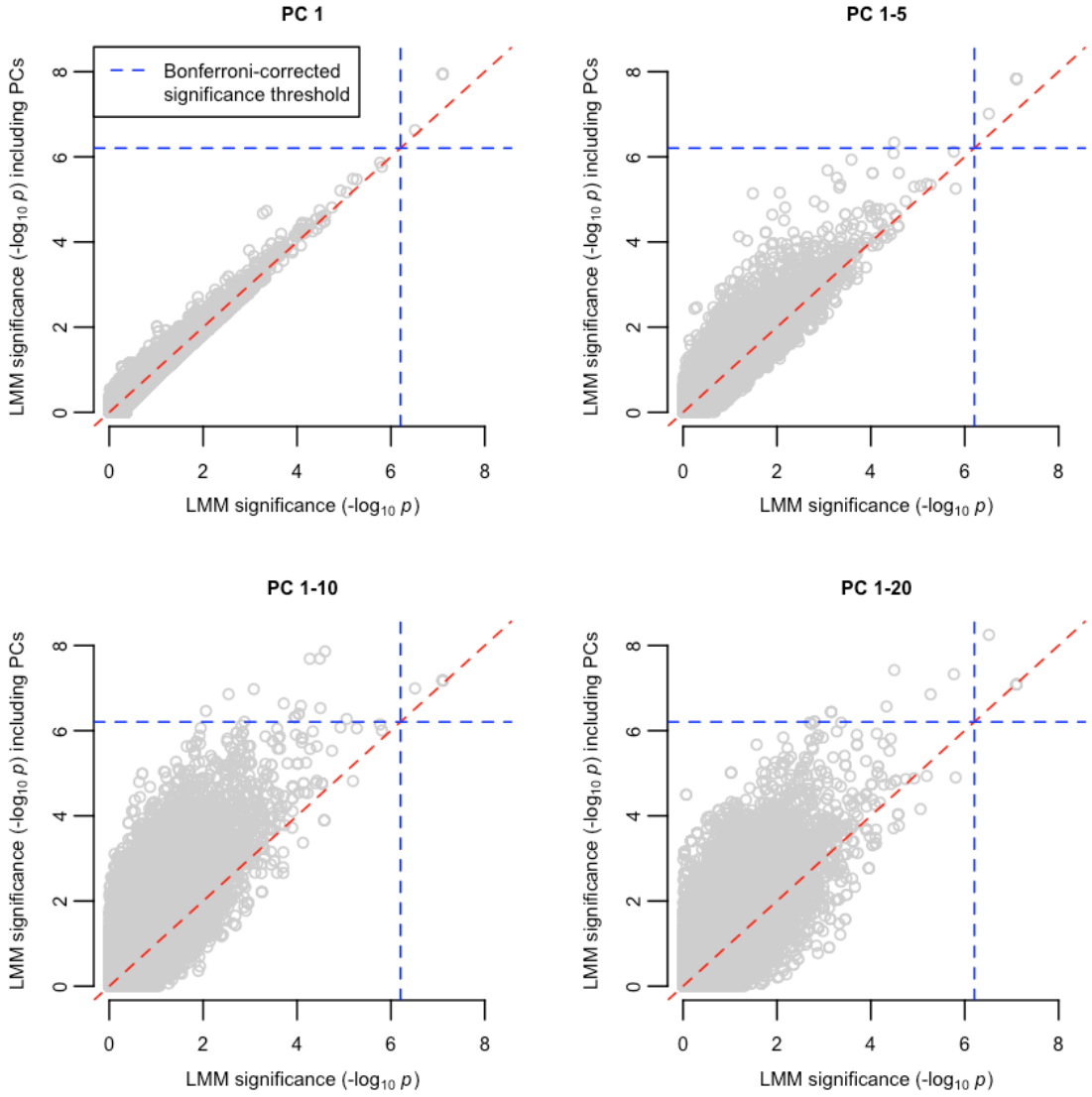

Figure 4.9 Significance by LMM with and without including Principal Components as additional fixed effects. A PC 1. **B** PCs 1-5. **C** PCs 1-10. **D** PCs 1-20. The originally significant SNPs remained significant in all tests.

the unique SNP phylopatterns which fell above the highest posterior probability threshold as an additional fixed effect individually (Figure 4.10). We assumed that loci combined additively in their effect on the phenotype. This measure was additional to the inclusion of all SNPs as random effects in the control for population structure. A new highest posterior probability threshold was calculated which contained 66 SNP pairs. All pairs consisted of the two unique SNP phylopatterns of the seven SNPs above the original Bonferroni-corrected significance threshold.

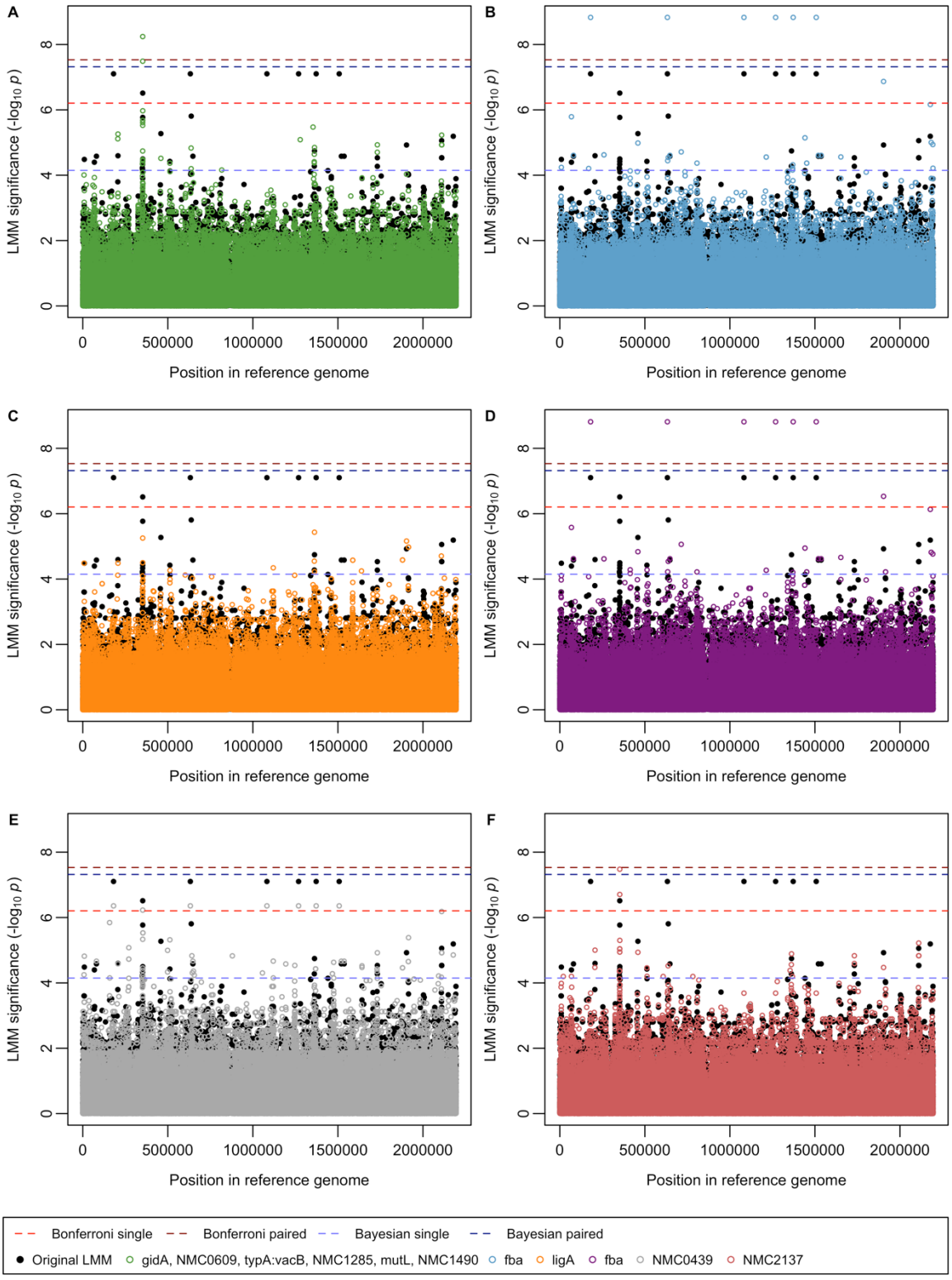

Figure 4.10 Results of the paired SNP analysis. Black dots depict the original LMM results not including any additional SNP patterns as fixed effects on top of the focal SNP which is included as a fixed effect. The coloured dots are the results of including that SNP pattern as an additional fixed effect in the LMM. The original and paired Bonferroni-corrected significance thresholds are shown along with the highest Bayesian posterior probability thresholds.

Including the significant SNP in *fba* as an additional fixed effect caused the

largest increase in significance for the most significant SNP phylopattern, from $-\log_{10} P$ = 7.1 to $-\log_{10} P$ = 8.8. Likewise, including the one representative of the significant SNP band (all members of which had the same phylopattern) also increased the significance of the significant *fba* SNP from $-\log_{10} P$ = 6.5 to $-\log_{10} P$ = 8.2. In both cases, when testing the two significant phylopatterns whilst conditioning on the other, the significance of the SNPs was above the new Bonferroni-corrected significance threshold which took into account the number of pairs tested. This gives support to the two top SNP phylopatterns corresponding to *fba* and the band of six top hits being independent signals.

4.4.3.3 The significant SNPs were ST-11 associated

We visualised the phylopatterns of the significant SNPs on the phylogeny of the dataset (Figure 4.11). This revealed that for the most significant SNP phylopattern, the minor allele was only present within ST-11s, and was also more prevalent in invasive isolates with an odds ratio of 1.5. The minor allele for the significant SNP in *fba* was present in all ST-11 isolates but was also present in multiple other parts of the phylogeny. This was also more prevalent in invasive isolates but with a greater odds ratio of 6.3. Therefore, both significant phylopatterns were associated with the ST-11 hypervirulent lineage. We reassessed the locus effects in light of the lineage effects presented in 4.4.2.2 by assigning variants to lineages. This was achieved by taking the variant phylopattern and determining the correlation of the phylopattern with the projections of the individuals onto each principal component. Variants were assigned to the lineages to which their absolute correlation was highest. This showed that for the SNP locus results, all significant variants were most strongly correlated to lineage PC-1, the ST-11 hypervirulent lineage (Figure 4.5A,D).

Here we have shown that seven SNPs were significantly associated with carriage versus invasive disease. Further evidence is provided by the highest posterior probability

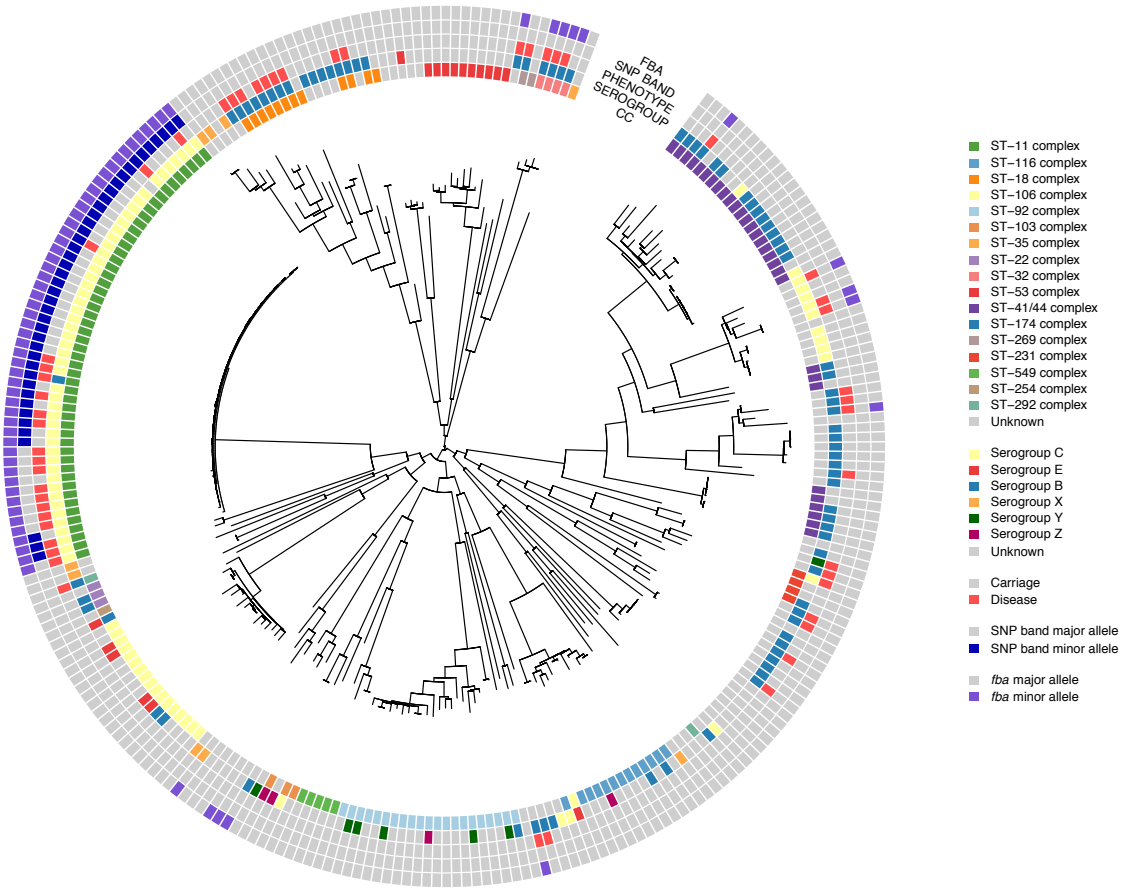

Figure 4.11 Phylogenetic distribution of the two phylopatterns of the seven significant SNPs. The minor alleles of both SNPs were more prevalent in invasive disease isolates than carriage isolates, and were largely associated with the ST-11 hypervirulent complex.

threshold which contained just 38 SNPs, giving confidence that the signal of association is concentrated among the small number of most significant SNPs. The significant SNPs were robust to the inclusion of principal components as fixed effects in the LMM, and also to the inclusion of the other significant SNP phylopattern as an additional fixed effect. All significant SNPs were most associated with lineage PC-1, the ST-11 hypervirulent lineage, as shown by the Wald test for lineage effects and by viewing the presence of the SNP minor alleles on the phylogeny.

4.4.4 An aside – why counting kmers from sequencing reads is not always appropriate

An alternative to the SNP-based analysis is to test for associations between the phenotype and the presence or absence of short kmers, the presence or absence of which can be

defined in different ways: from raw sequencing reads or from *de novo* assembly contigs. Previous bacterial association studies investigating kmers as the genetic unit of interest defined kmers from *de novo* assemblies (Sheppard et al. 2013; Pascoe et al. 2015; Lees et al. 2016; Yahara et al. 2017). We however aimed to define kmers from raw sequencing reads as it has the advantage of capturing variation in reads which do not assemble. Here, we tested the association of 31bp kmers counted from sequencing reads with the phenotype and consider why counting kmers from sequencing reads may not always be appropriate.

4.4.4.1 The total number of kmers counted from sequencing reads per isolate was not uniform

31bp kmers were counted directly from raw Illumina sequencing reads using dsk (Rizk, Lavenier & Chikhi 2013) as discussed in Chapter 2.4. When counting kmers from sequencing reads, a kmer was counted as present if found five or more times within an isolate, absent if not. However, this was potentially too low a threshold, so quality control measures were taken to check whether the results from counting kmers from reads could be trusted.

Counting kmers from sequencing reads produced 14,680,679 variant kmers across the dataset, defined by 1,430,549 unique phylopatterns across isolates. Experimental design is not always ideal in WGS studies. In this study, the ST-11s were sequenced prior to the other isolates and therefore had 76bp reads, all other isolates were sequenced with 100bp reads. Such batching can produce statistical artefacts. Therefore, measures were taken to identify possible batch effects. Specifically, the number of kmers per isolate was investigated, revealing a non-uniform distribution across isolates (Figure 4.12) although this effect did not appear to be strongly confounded with phylogenetic position (Figure 4.13).

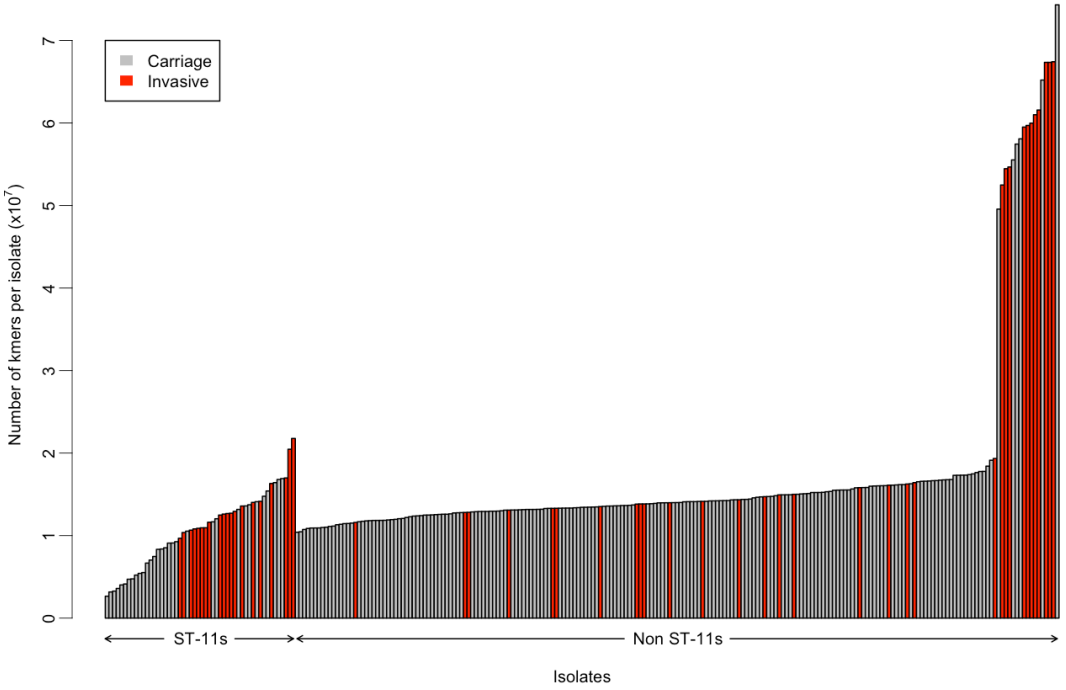

Figure 4.12 Number of kmers per isolate when counting kmers from sequencing reads. Carriage isolates are coloured grey, and invasive isolates are coloured red. The isolates are split by ST-11 vs non ST-11.

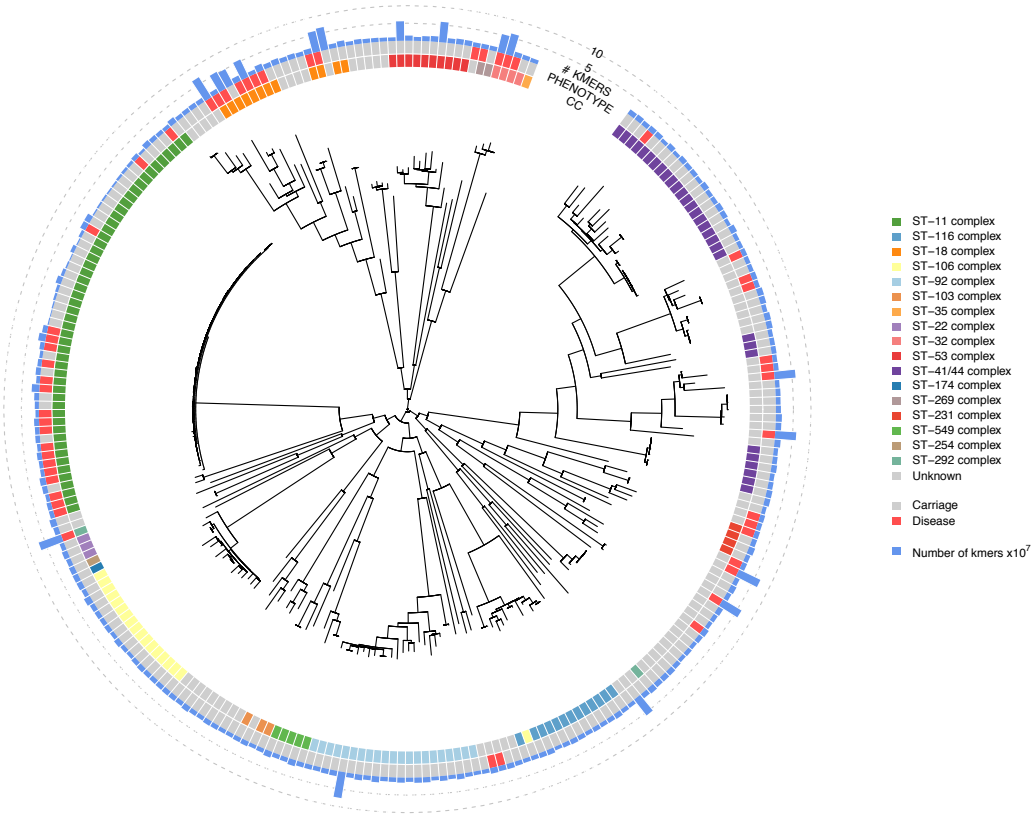

Figure 4.13 The phylogenetic distribution of the number of kmers counted per isolate from sequencing reads. A small subset of the isolates had many more reads than the others, but they were phylogenetically distributed.

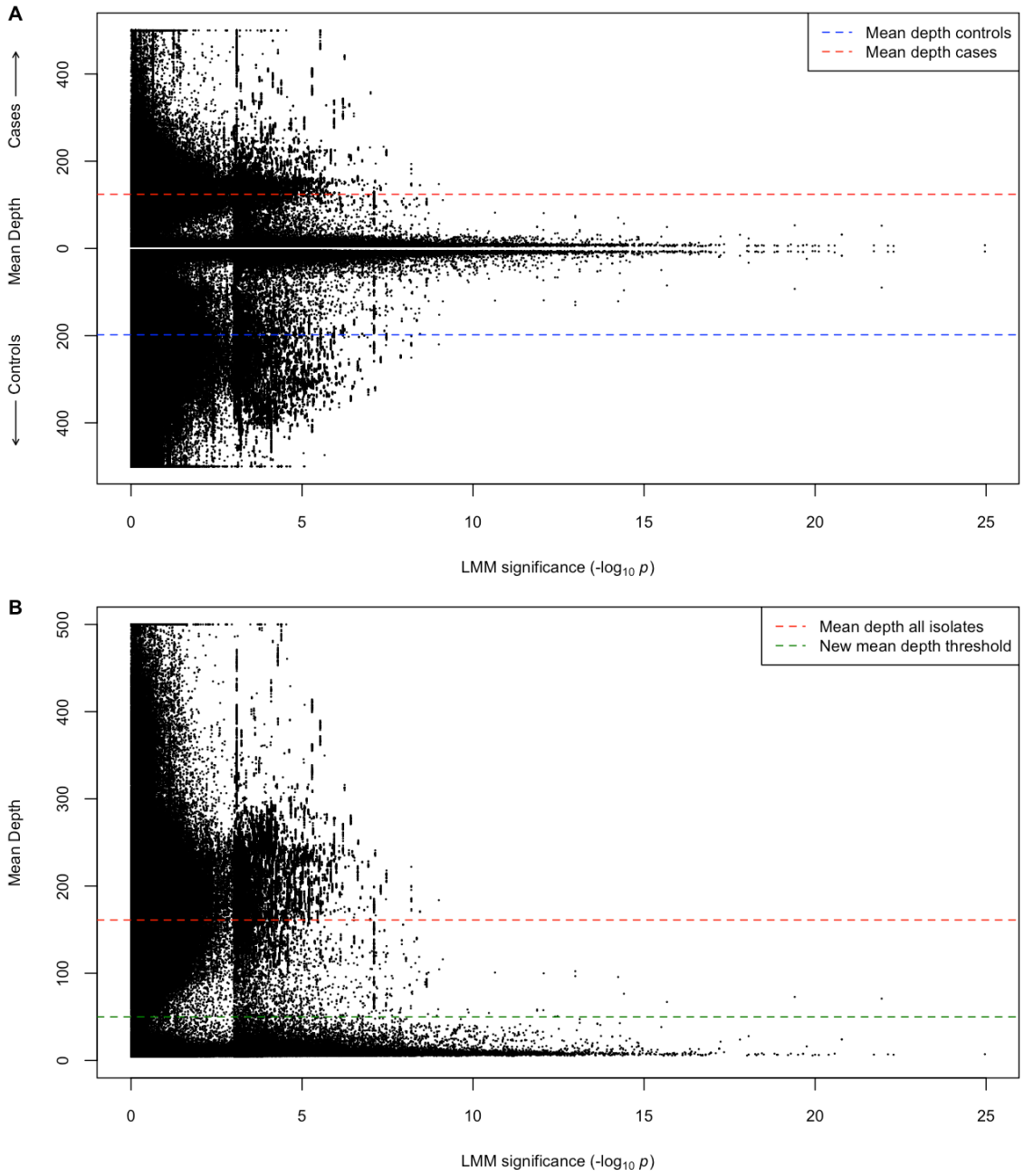

Figure 4.14 Kmer LMM significance against raw depth. A cases and controls split; **B** all isolates. The most significant kmers were all at very low mean depth, therefore a post-counting filtering step was introduced, where only kmers with a mean depth of 50 or greater were kept in the analysis (shown by the green line in **B**). This figure was produced based on code written by Daniel Wilson.

4.4.4.2 Significant kmers counted from sequencing reads had low mean depth

Kmers were tested for association with carriage versus invasive disease status using LMM. However, in performing post-hoc quality control, we found that when plotting the $-\log_{10} P$ values from the LMM against the mean depth per kmer, the most significant kmers were all found at low mean depths (Figure 4.14). This indicated that potentially

some or all of these kmers could have been errors, possibly just exceeding the presence threshold of five counts per isolate. We therefore introduced a filter post kmer counting, only keeping kmers with a mean depth of 50 or greater across all isolates, as this appeared to remove the tail of the distribution. Using a mean depth of 50 is relatively arbitrary however, and results would change depending on the threshold used. Ideally to count kmers from sequencing reads the mean depth of each isolate would be taken into consideration to determine an individual presence threshold per isolate. This is something that requires further work.

4.4.4.3 The raw-read kmer association results were not robust to thresholding

In order to annotate the kmers, we mapped all remaining kmers to the reference genome FAM18 using Bowtie2 (Langmead & Salzberg 2012), reporting up to five mapping positions per kmer. All significant kmers which mapped more than once were then mapped again but all mapping positions were reported. Of the 153 significant kmers counted from reads, 85 mapped more than once to the reference genome (Figure 4.15). This large proportion of significant kmers which did not map uniquely may therefore represent repetitive regions of the genome. This is concerning, as although repetitive regions are often in interesting genes and regions of the genome, it means that the kmers could be present due to sequencing errors.

Sequencing errors are more likely to occur at particular motifs, such as repetitive regions of the genome (Ross et al. 2013). The greater the copy number of a repetitive sequence, the greater the opportunity for a sequencing error to be replicated. For a kmer to be defined as present by error rather than true presence, a sequencing error would only need to have occurred five times within an isolate. Applying a post-counting filtering step may not have solved this problem, as a kmer could be truly present in some isolates with a large mean depth, but be present due to sequencing error in others with a low mean

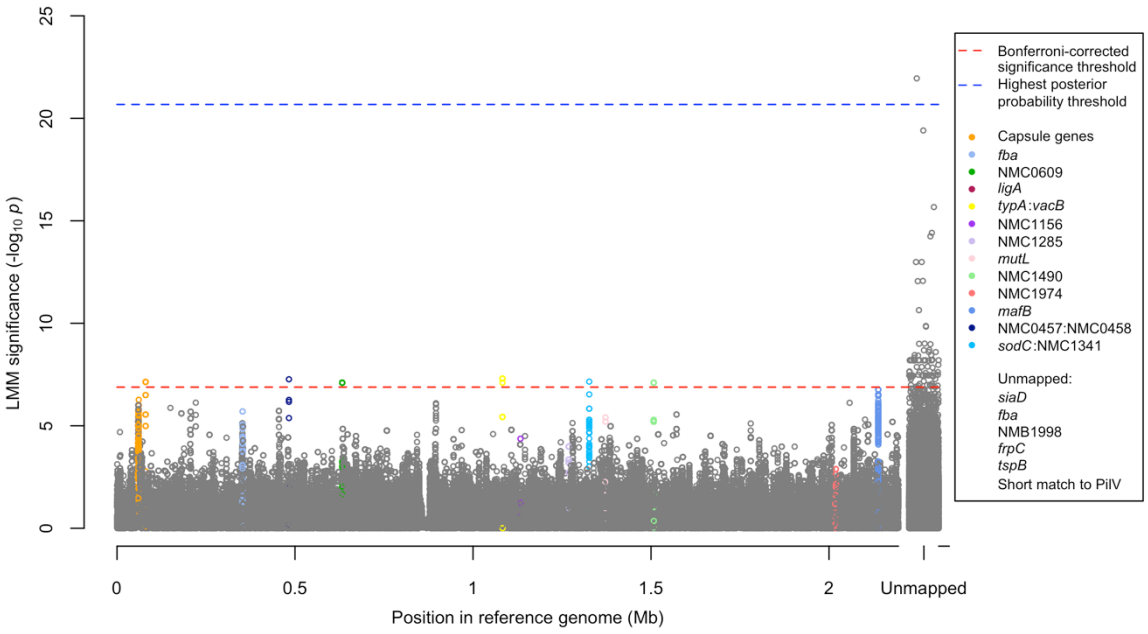

Figure 4.15 Kmer LMM significance against position in the reference genome. Manhattan plot showing the kmer results after using a post-counting filter of a mean depth of 50 or greater. Unmapped kmers plus kmers which did not map uniquely are plotted to the right of the figure and genes and intergenic regions containing significant variants are coloured. The most significant variants either did not map or did not map uniquely to the reference genome.

depth. This could pass the filtering threshold, which would mean defining the kmer as present in both sets of isolates, despite only being truly present in a subset of them.

To test whether this was an issue within this dataset, the number of times each significant kmer was present per isolate was converted into a standardised count (as described in Section 4.3.5). This was depicted on the phylogeny for the most significant kmer in Figure 4.16 which revealed how the standardised counts were often very divergent to the raw counts. The third ring on the phylogeny, from the inside out, depicts kmer presence/absence. Grey represents kmer absence (raw count of less than five) and blue represents kmer presence (raw count of five or greater). Standardised counts are shown in the outer ring in purple. When the standardised count was less than one, this means that the kmer was present at a lower depth than the mean depth across the genome in that isolate. This demonstrated that although the kmer was defined as present in the majority of isolates using the raw count threshold of five, it would be defined as absent in many of these isolates if using a threshold based on the standardised count, as many had

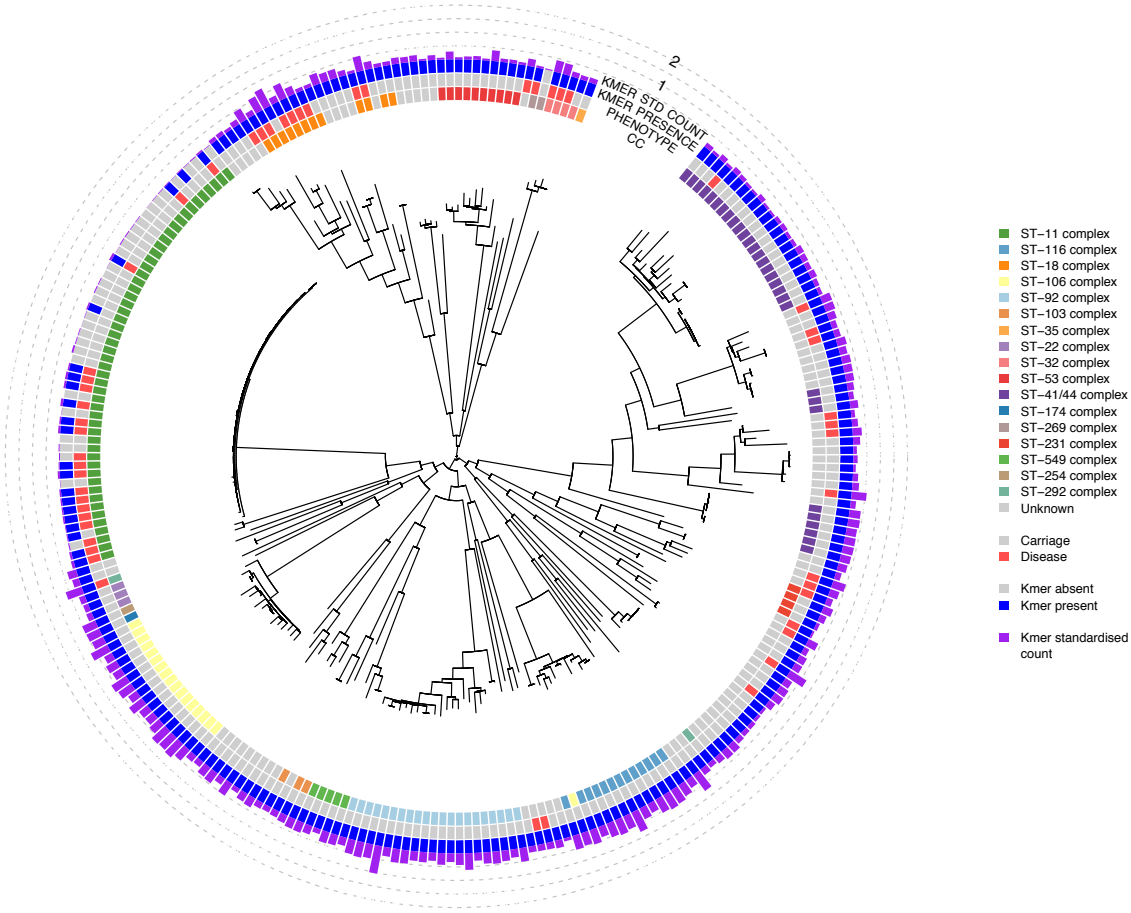

Figure 4.16 Standardised counts of the most significant kmer counted from sequencing reads on the phylogeny of the dataset. The third ring from the inside out is coloured grey if the kmer was originally determined as absent (raw count of <5) or blue if determined as present (>=5). The outer ring shows the standardised counts with dotted rings as a reference (Kmer std count = kmer standardised count).

standardised counts of close to zero and the majority of less than 0.5 (Figure 4.16).

It has previously been shown that errors accumulate along the read in Illumina sequencing, so errors are more often found at the end of a read than at the beginning (Schröder et al. 2010). Another measure taken to test the robustness of the kmers counted from reads was to identify the starting position of each of the 153 significant kmers (and their reverse complement) in the sequencing reads for each isolate. If the kmers were truly present, then we would expect to see their starting position evenly distributed throughout the reads. If present due to errors, their frequency may increase towards the end of the reads. Figure 4.17 shows examples of two significant kmers, with ST-11s and non ST-11s shown separately due to having different read lengths. Figure 4.17A depicts

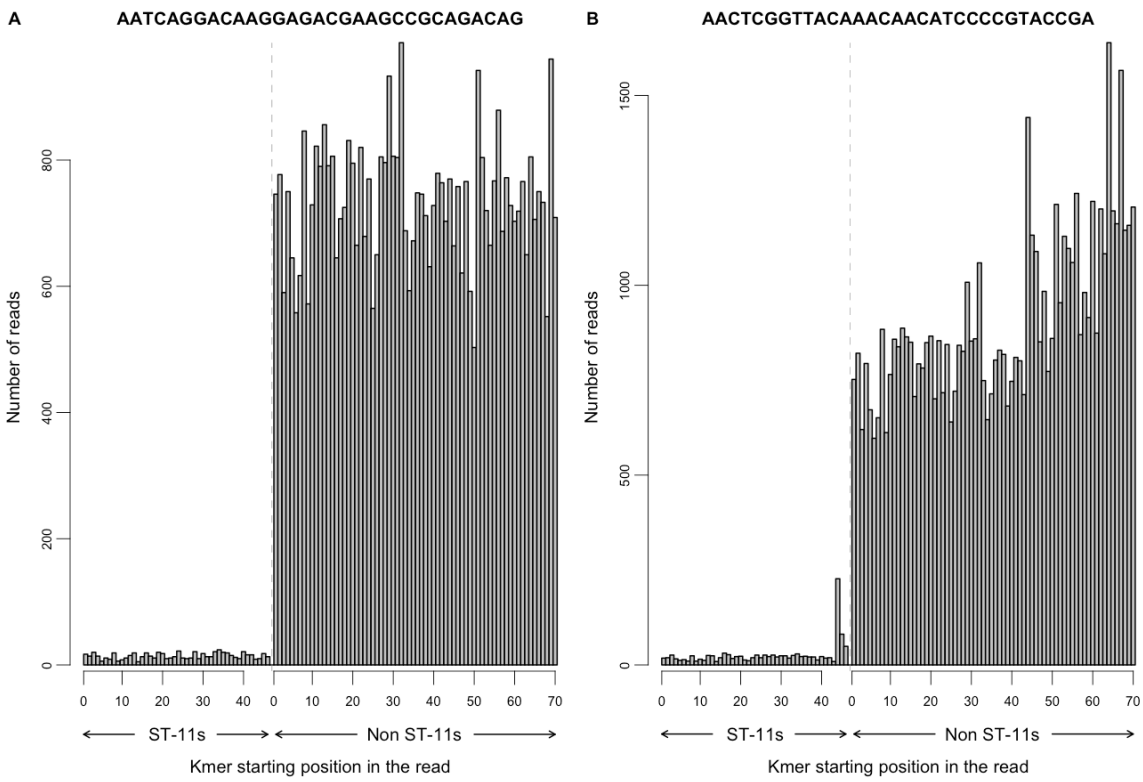

Figure 4.17 Starting position of two significant kmers in the sequencing reads of all isolates they were present in. A An example of what we would like to see, where the kmer was found to start evenly throughout the reads it was present in. **B** An example of a problematic kmer, where the kmer was increasingly likely to be seen towards the end of the read, which can occur due to errors.

what we would expect to see for a truly present kmer, where the starting position of the kmer was uniform across the read. Figure 4.17B however shows a kmer which was more frequently found at the end of the read. The presence of kmers with profiles such as that of Figure 4.17B suggested that the results of such kmers were not robust.

We have shown that there can be problems in choosing a reliable presence threshold when counting kmers from sequencing reads from data that has been put together somewhat opportunistically. We must conclude that although the results may be interesting, we cannot determine whether they are false positives. As standardised counts have revealed that counting kmers based on raw presence counts can be problematic, to be able to utilise counting kmers from sequencing reads the depth of each isolate would need to be taken into account. Therefore, we did not analyse this data further and instead defined kmers based on their presence in Velvet assemblies. Despite the potential

disadvantage of restricting the analysis to regions which assemble, we must accept the trade off in terms of robustness versus sensitivity.

4.4.5 Counting kmers from Velvet assemblies revealed additional associations with invasive disease

Here, we defined kmers based on their presence in Velvet assemblies in order to overcome the problems with determining an appropriate threshold when counting from sequencing reads.

4.4.5.1 The total length of assemblies and number of kmers per isolate was uniform

Assembly length was uniform across the isolates with total assembly length ranging between 2.07-2.30Mb and for contigs ≥ 1kb between 2.01-2.25Mb (Figure 4.18A). This resulted in the total number of kmers counted per isolate from the assemblies also being more uniform across isolates, ranging between 2.0-2.18×10^6 kmers, (Figure 4.18B) in comparison to counting kmers from sequencing reads (Figure 4.12). 7,806,583 variant kmers were counted across the dataset which were defined by 307,830 unique phylopatterns. The kmers were then tested for association with the phenotype and also assigned to the lineages identified in Section 4.4.2.2.

4.4.5.2 Identification of kmers counted from Velvet assemblies associated with invasive disease

465 kmers were significant after controlling for population structure using LMM and defining the kinship matrix and thus relatedness between isolates using SNPs. 63 of these kmers overlapped with the significant kmers counted from reads. We mapped the kmers to the reference genome FAM18 reporting just the best mapping per kmer. Kmers that mapped with a quality less than 10 were treated as unmapped. 29.7% (138/465) of the significant kmers did not map to the reference genome. This could reflect either different

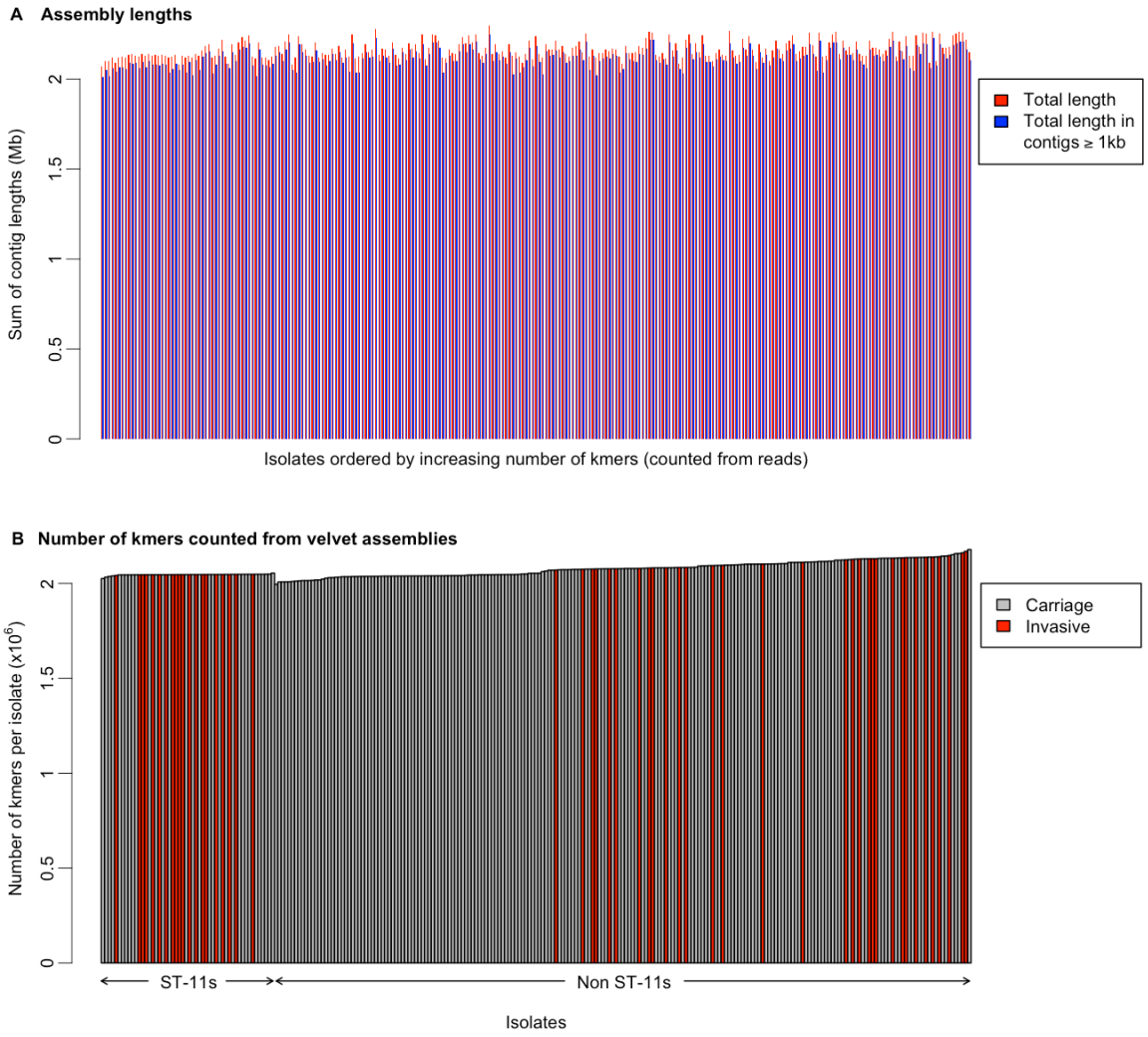

Figure 4.18 Velvet assembly lengths and number of kmers counted from each assembly. A The length of all Velvet assemblies in the dataset and the length of all contigs ≥ 1kb ordered by increasing number of kmers when counting kmers from sequencing reads. **B** The total number of kmers counted from each Velvet assembly split by whether the isolates were ST-11. Both the assembly lengths and number of kmers counted were relatively uniform across the dataset.

gene content between the isolates and the reference genome or sequence divergence from the reference genome. BLAST was used to identify the origin of the kmers which did not map to the FAM18 reference genome. The genes identified are shown in Figure 4.19. Significant kmers were found within the genes identified as significant in the SNP GWAS plus within *ctrG*, the intergenic region between *ctrE* and *ctrF*, *ligA*, NMC1156, NMC1974, *mafB*, the intergenic region between NMC0457:NMC0458, the intergenic region between *sodC* and NMC1341, *siaD*, NMB1998, *frpC*, *tspB* and a short match to PilV. All kmers represented a form of variation, rather than the presence/absence of a

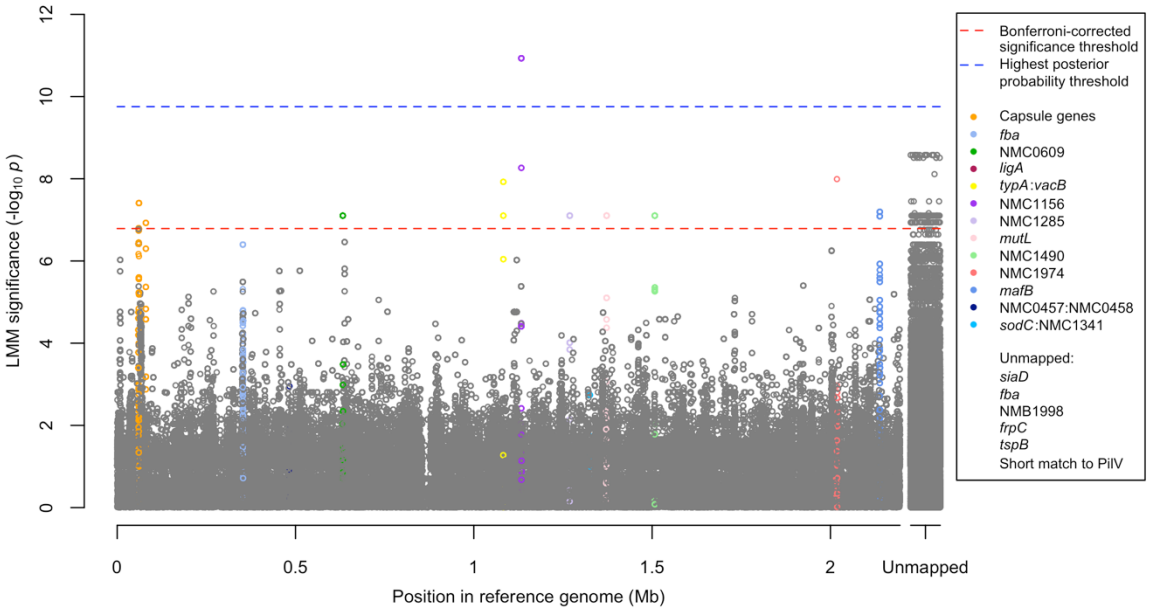

Figure 4.19 Manhattan plot of the kmer LMM results where kmers were counted from Velvet assemblies. Unmapped kmers are plotted to the right of the figure and genes and intergenic regions containing significant kmers are coloured.

gene, consistent with previous opinion that there is no virulence gene pool (Snyder & Saunders 2006; Schoen et al. 2008).

The kmers identified from Velvet assemblies were used to calculate the kinship matrix using GEMMA (Zhou & Stephens 2012) and kmers were tested for association and compared with the results using the kinship matrix built from SNPs. There is a concern that, particularly with the added vagaries of kmer counting, there might be a benefit to calculating the kinship matrix direct from the kmers rather than the SNPs, because the kinship matrix can additionally account for weak batch effects if present. The association results were very similar between using SNPs and kmers to build the kinship matrix (Figure 4.20). Just 14 of the 465 significant kmers lost significance when building the kinship matrix from kmers, no kmers gained significance and the heritability estimate was very similar at 35.35% with a standard error of 10.6%. It also made little difference to the QQ plot of the null versus expected distribution of *P* values for the kmers counted from Velvet assemblies (Figure 4.20), therefore we continued to investigate the results using SNPs to define the kinship matrix.

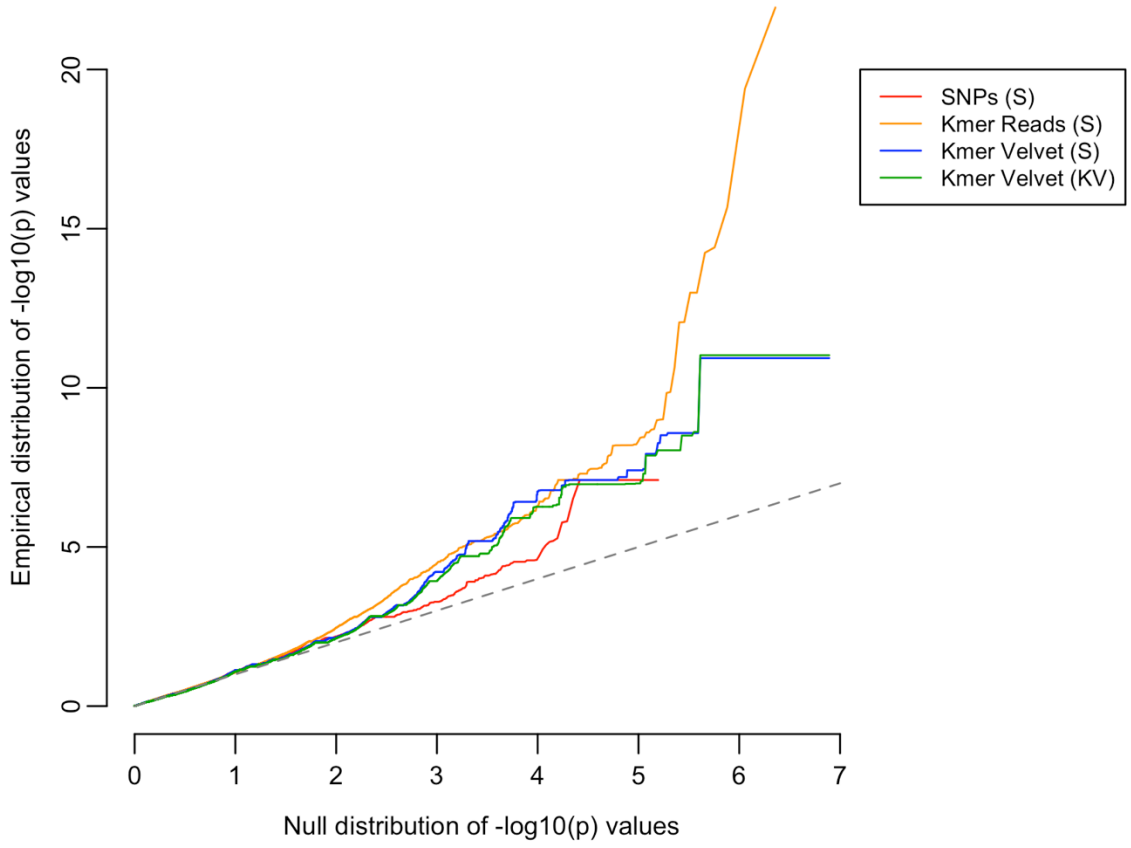

Figure 4.20 QQ plots showing the expected versus the empirical distribution of *P* values for all analyses. In brackets, the letters symbolise the type of genetic variation used to calculate the kinship matrix for the LMM analyses: S = biallelic SNPs, KV = variant kmers counted from Velvet assemblies. All studies showed a signature of overinflation of *P* values with respect to the null distribution, but the kmers counted from Velvet assemblies were less overinflated than the kmers counted from sequencing reads.

4.4.5.3 The majority of significant kmers counted from Velvet assemblies were associated with ST-11

By interpreting the locus effects in light of the lineage effects, 71% (331/465) of the significant kmers were most strongly correlated to lineage PC-1 (Figure 4.21B). The remainder of the significant kmers were most strongly correlated to either PC 2, 6, 7, 8 or 188. We can also see that overall, the SNPs and kmers associated with lineage PC-1 tended to lose significance after controlling for population structure (Figure 4.5A and Figure 4.21A) which we would expect to see if the control for population structure was sufficient.

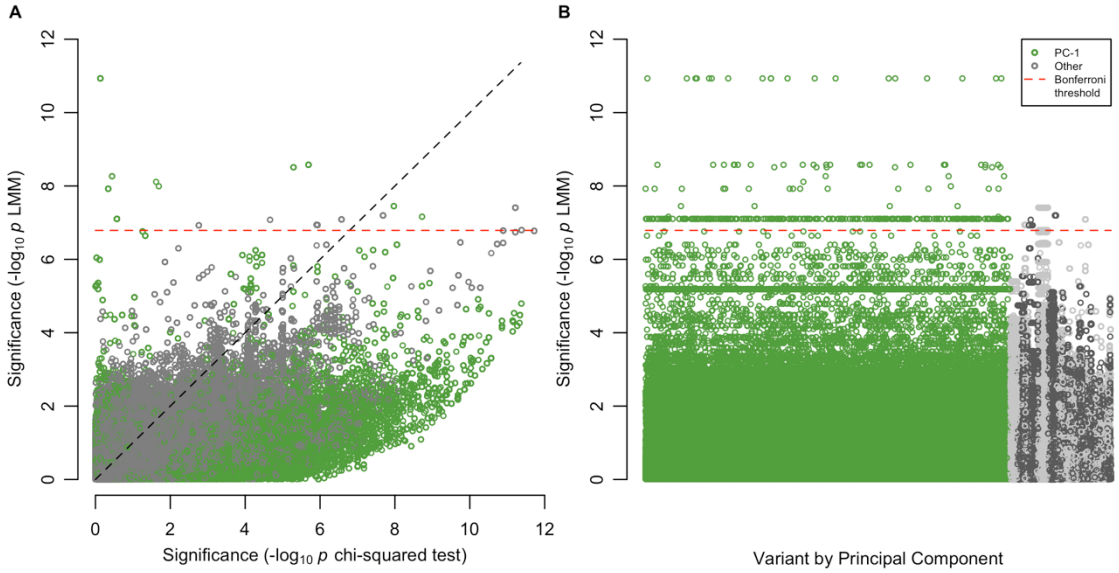

Figure 4.21 Interpreting the locus effects of the kmers counted from Velvet assemblies in light of the lineage effects. A $-\log_{10} P$ values of the uncorrected chi-squared test against the $-\log_{10} P$ values after controlling for population structure using LMM. **B** $-\log_{10} P$ values after controlling for population structure using LMM ordered on the x-axis by the principal component they were most strongly correlated with. This revealed that although the majority of lineage PC-1 associated variants were reduced in significance when controlling for population structure, most significant results post-LMM were still lineage PC-1 associated.

4.4.6 Identifying possible roles of significant variants in causing invasive disease

4.4.6.1 Variants in genes involved in the production of the capsule were associated with carriage versus invasive disease

Molecular work has revealed the importance of the capsule in invasive disease, and it has been shown to be necessary but not sufficient to cause invasive disease as discussed in Section 4.1.2. Multiple genes related to capsule production contained significant kmers: the intergenic region between *ctrE* and *ctrF*, *ctrG* and *siaD*. These will be discussed individually in this section.

ctrE and *ctrF* are capsule polysaccharide modification proteins part of capsule region B and conserved across all serogroups. The significant kmers in the intergenic region between the two genes all contained the major alleles of SNPs at FAM18 positions 80313, 80322 and 80327 based on mapped data, and all had a protective effect, with odds ratios of 0.26 (Figure 4.22).

σ^{70} -10 Pribnow box										
Consensus in *E. coli*	5'	-	**T**	**A**	T	A	A	**T**	-	3'
Significant kmers	5'	-	T	A	T	G	C	T	-	3'
-35 region										
Consensus in E. coli	5'		**T**	**T**	**G**	A	C	A	-	3'
Significant kmers	5'	-	T	T	G	C	A	T	-	3'

Table 4.2 Significant kmers in the intergenic region between *ctrE* and *ctrF* aligned with the consensus sequences for the *E. coli* σ^{70} -10 Pribnow box and -35 regions (Harley & Reynolds 1987) revealed sequence identity at the most conserved bases. The most conserved bases in *E. coli* for both regions are shown in bold and underlined for the consensus sequences.

It has previously been shown that *ctrE* and *ctrF* are likely transcribed independently (Tzeng 2005) but the promoter region is unknown. We hypothesise that the kmers captured the promoter region for *ctrF*. Examination of the sequence captured by the significant kmers revealed a match to four of the six nucleotides of the *E. coli* consensus sequence for the σ^{70} -10 Pribnow box and a match to three out of six nucleotides for the -35 sequence, separated by 17bp (Table 4.2) (Harley & Reynolds 1987). Analysis of *E. coli* promoter sequences has previously shown the three most conserved bases of the -10 and -35 sequences to be those in bold and underlined in Table 4.2, and that 92% of all sequences are optimally aligned with 17+/-1bp separating the -35 and -10 sequences (Harley & Reynolds 1987). Therefore, the significant kmers contained the most highly conserved residues of the -10 and -35 promoter sequences in *E. coli*, separated by 17bp, which is optimal for an actively transcribed promoter (Harley & Reynolds 1987).

The three SNPs covered by the kmers fell within the 17bp separating the putative -10 and -35 regions, so we hypothesise that divergence from the reference sequence in this region at those SNPs is associated with invasive disease, possibly due to an ablation of or increase in expression in *ctrF*. Although the kmers appeared to be capturing SNP variation, the SNPs covered by the significant kmers were not found to be significant in the SNP analysis. It appears that the kmers could be have captured an allelic effect, highlighting the power of the kmer analysis. Kmers covering just two of the three SNPs

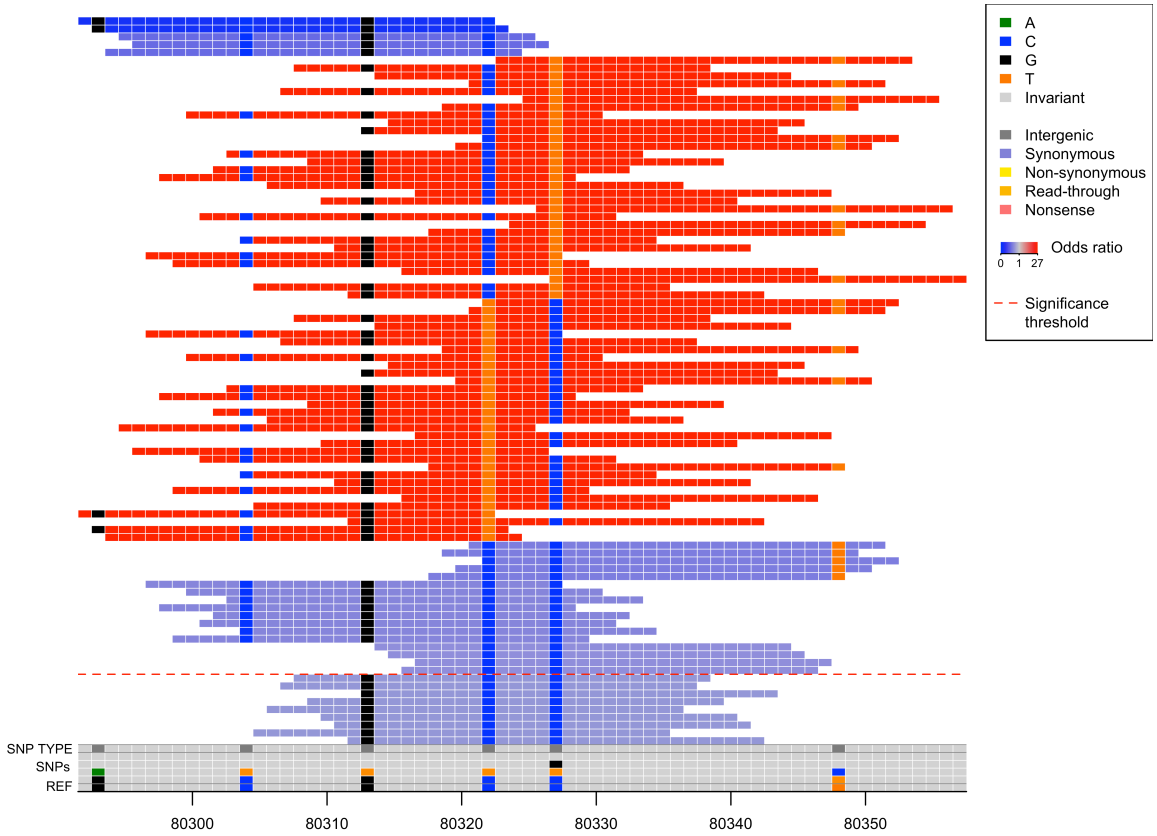

Figure 4.22 Close up of the significant kmers aligned to the intergenic region between *ctrE* and *ctrF*. From the bottom up, the reference is shown coloured grey if there were no variants between the kmers and the reference at that position, coloured by the reference base otherwise. The next four rows depict the SNPs based on the mapped data, from the most common allele at the bottom then decreasing in frequency, grey representing unseen alternative alleles. The sixth row depicts the type of SNP, here they are all dark grey and therefore intergenic. The kmers are then shown stacked from the most significant at the bottom decreasing in significance. Those below the red dashed line were above the Bonferroni-corrected significance threshold. Kmers are coloured by their odds ratio, from protective kmers coloured in blue through to risk kmers coloured in red. Kmers are also coloured by their allele at all variant sites (see key).

covered by the significant variants were not significant (Figure 4.22). Also, kmers which covered the three variants plus an additional variant at position 80304 were not significant, suggesting that the kmers were specifically capturing the combination of the major alleles of the three variants at positions 80313, 80322 and 80327 (Figure 4.22).

ctrG is part of capsule region A of serogroup B, C, W and Y and has been shown to have a role in the surface translocation of sialic acid capsules, essential in enabling the correct expression of sialic acid polysaccharides (Hobb et al. 2010). Significant kmers fell within two regions of the gene; in the first region the significant kmers covered a synonymous SNP at position 61015 with reference to FAM18 based on the mapped data,

Figure 4.23 Close up of the alignment of the significant kmers in *ctrG* to the reference genome FAM18. As before, variant sites are coloured by their allele, mapped SNP data is shown and the kmers are coloured by their odds ratio. Here, the significant kmers covered three SNPs: two non-synonymous and one synonymous, all containing the major allele based on the mapped SNP data.

and in the second, the significant kmers covered two non-synonymous SNPs at positions 60258 and 60275 (Figure 4.23). All significant kmers contained the major alleles (based

on the SNP data) for the three variants and were only present in invasive isolates.

ctrG was previously identified as a putative phase variable gene based on poly(T) repeats in the coding sequence (Saunders et al. 2000), but the significant kmers did not fall within this region. As the kmers appeared to be capturing SNP variation, this poses the question of why the SNPs were not significantly associated in the SNP analysis. As *ctrG* is only present in serogroups B, C, W and Y, it will therefore not have been present in all isolates in this dataset which also contained isolates of serogroups E, X and Z. When performing SNP imputation however, no differentiation was made between an uncalled base and a genuine deletion, therefore alleles will have been imputed at these variants in isolates not containing the gene. This masked the true signal and resulted in the SNPs not being significantly associated, and suggests that we should perhaps treat deletion as an additional state when imputing missing sites.

4.4.6.2 Putative phase variable regions were associated with invasive disease

Two genes containing putative phase variable regions contained significant kmers: NMC1156 and *siaD*. NMC1156 is a glycosyl transferase 2 family protein which has previously been predicted to be a strong candidate for phase variation due to poly(G) repeats in the coding sequence (Bentley et al. 2007). The region assembled with flanking sequence surrounding the poly(G) repeats in 233/261 isolates. In 26/261 isolates the contig ended in the poly(G) repeat region and in the remaining two isolates the region did not assemble. Poly(G) repeat length varied between 5-13bp, although we cannot exclude the possibility that the varying lengths were due to errors in the Illumina sequencing, as error rates are elevated in homopolymeric regions (Ross et al. 2013). However, a previous study found the range of repeat lengths to be between eight and twelve in a diverse collection of 20 genomes (Siena et al. 2016).

The significant kmers in NMC1156 were defined by two unique phylopatterns

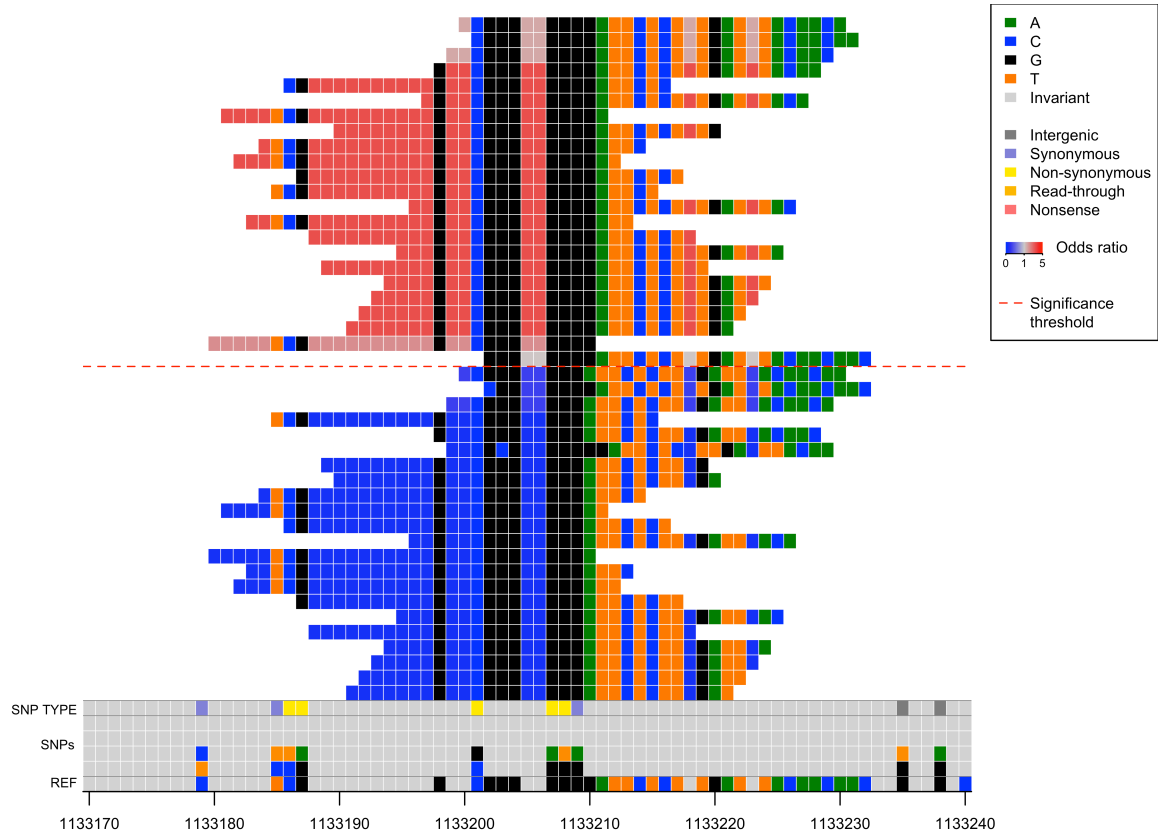

Figure 4.24 Close up of the alignment of the significant kmers which mapped to NMC1156. As before, variant sites are coloured by their allele, SNP data is shown and the kmers are coloured by their odds ratio. Here, the significant kmers covered seven SNPs: five non-synonymous and two synonymous. There were many differences between the reference and the downstream end of the kmers, indicating that they did not map well.

present in 39 and 52 individuals. The kmers covered the positions 1133180-1133232 with reference to FAM18 (Figure 4.24) and their presence was protective, with odds ratios of either 0.68 or 0.86. The alignment in Figure 4.24 revealed many sequence differences between the kmers and the reference genome FAM18, which suggested that the kmers were capturing variation other than SNP variation.

All significant kmers in NMC1156 covered the positions 1133202-1133210, and the poly(G) repeats in all isolates based on the Velvet *de novo* assemblies were found between positions 1133201-1133213. We bioinformatically predicted stop codon position for each isolate as described in Section 4.3.6. For the 52 isolates containing the significant kmers, all had a repeat length of 8 and a premature stop codon at either codon 338 (39/52) or codon 108 (13/52) (Table 4.3; Table 4.4). Of the remaining isolates for which the

	Repeat length								
	5	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Isolates not containing significant kmers	1	4	6	10	71	54	49	11	1
Isolates containing significant kmers	0	0	0	52	0	0	0	0	0

Table 4.3 The number of isolates with different NMC1156 poly(G) repeat lengths split by whether the isolate contained significant kmers in NMC1156.

	Stop codon prediction										
	NA	**33**	**108**	**109**	**110**	**119**	**120**	**121**	**337**	**338**	**339**
Isolates not containing significant kmers	47	5	0	45	1	3	43	25	2	33	3
Isolates containing significant kmers	0	0	13	0	0	0	0	0	0	39	0

Table 4.4 The number of isolates with different NMC1156 stop codon position predictions split by whether the isolate contained significant kmers in NMC1156.

significant kmers were absent, 10 had a repeat length of 8, the remainder between 5 and 13 but the majority were 9 or greater (Table 4.3). Where we could predict stop codon position, stop codons were found at codons 33, 109, 110, 119, 120, 121, 337, 338 and 339 (Table 4.4). Therefore, the kmers appeared to be representing particular length variants of NMC1156, and therefore divergence from these NMC1156 lengths was associated with invasive disease.

Significant kmers also captured a phase variable region in *siaD*, a capsule synthesis gene present in isolates with a sialic capsule (Harrison et al. 2013). Twenty-one significant kmers mapped to *siaD* and were all present in the same 68 individuals, 57 of which were serogroup B. The kmers capturing *siaD* did not map well to the FAM18 gene variant therefore a figure aligning the kmers with the SNP data is not shown, as we do not have SNP data to align to in this case. *siaD* phase variation occurs due to variation in homopolymeric tracts of As and Cs (Hammerschmidt et al. 1996; Weber et al. 2006) and a study of phase variable regions in a diverse collection of 20 meningococcal genomes found that the poly(A) length ranged between 8-10 bases (Siena et al. 2016).

BLAST was used to query the isolate Velvet contigs, which revealed that 82 isolates contained the gene. The 68 isolates which contained the significant kmers all contained poly(A) tract lengths of 9, and the remaining 14 *siaD* containing isolates had tract lengths of 8, 10 or 11. The significant kmers all overlapped one another and the 9 As. We bioinformatically predicted which state *siaD* was in, whether phase variable "on" or "off". Of the 68 isolates containing the significant kmers, in one isolate the gene was split over two contigs so it was not possible to determine stop codon position. For the remaining isolates we determined 63/68 to be "on" as they did not contain any frameshifts or premature stop codons. 4/68 contained an insertion at position 88 in the gene, at the position of a phase variable poly(C) region and therefore a frameshift and premature stop codon, so were predicted to be "off". All *siaD* positive isolates for which the significant kmers were absent contained premature stop codons. Therefore, presence of the significant kmers appeared to be representing the *siaD* "on" state within the poly(A) phase variable region, consistent with prior molecular work detailing the importance of *siaD* in capsule formation (Hammerschmidt et al. 1996) and therefore invasive disease (Jarvis & Vedros 1991; Spinosa et al. 2007).

4.4.6.3 Variants upstream of factor H binding protein were associated with invasive disease

A SNP variant within the gene *fba* was found to be significantly associated with carriage versus invasive disease (Figure 4.8). The SNP was a synonymous variant, found at position 900 within the gene at the third position of an Aspartic acid codon. Kmers capturing the SNP were also found to be significantly associated in the kmer GWAS (Figure 4.15). A second synonymous SNP, found at position 897 within *fba*, fell just below the genome-wide significance threshold (Figure 4.8). *fba* is a glycolytic enzyme with a cytosolic role in glycolysis and gluconeogenesis that can also be localised to the

	SNP 900 T	SNP 900 C
SNP 897 C	194 (4307)	1 (65)
SNP 897 T	0 (0)	66 (999)

Table 4.5 The number of isolates containing the four haplotypes of the two most strongly associated variants in the gene *fba* shown for the current dataset, plus the haplotypes for all isolates with complete *fba* alleles in pubMLST shown in brackets.

outer membrane (Tunio et al. 2010). It has also been shown to play a role in the adherence of *N. meningitidis* to human cells by binding human Glu-plasminogen via its C-terminal lysine residue (Tunio et al. 2010; Shams et al. 2016). *fba* is adjacent to the gene encoding factor H binding protein (fHbp), which enables the bacterium to evade complement mediated killing by binding the human complement factor H (CFH), the major negative regulator of the alternative complement pathway (Schneider et al. 2006). *fba* and *fhbp* can each be transcribed on monocistronic transcripts from their dedicated promoters, but can also be contranscribed on a bicistronic transcript from a promoter sequence upstream of *fba* (Oriente, Scarlato & Delany 2010). It has been suggested that the bicistronic transcript is likely due to inefficient termination resulting in read-through of the transcriptional terminator downstream of *fba* (Oriente, Scarlato & Delany 2010).

The minor alleles of the two SNPs were found together in all but one isolate in the current study, and in all but 65/5,271 isolates with a complete *fba* allele in the pubMLST database (http://pubmlst.org/neisseria/). In all cases where the minor alleles were not both present, it was the minor allele at position 900 (the significant variant) which was present (Table 4.5).

ST-11 can be categorised into three sublineages, defined as ET-37 and ET-15 (as determined by MLEE) which correspond with sublineage 1 and sublineage 2 (determined by rMLST) plus an additional sublineage 3 defined by rMLST (Lucidarme et al. 2015). All ST-11s in the Czech dataset analysed here were ET-15/sublineage 2, the lineage shown to be responsible for the elevated serogroup C, CC-11 disease in the 1990s/2000s (Lucidarme et al. 2015). Assessing the presence of the two alleles at the significant SNP

	ST-11 sublineage 1	ST-11 sublineage 2	ST-11 sublineage 3
SNP 900 T	640	18	3
SNP 900 C	10	266	0

Table 4.6 The presence of the minor allele of the significant SNP at position 900 in *fba* depends on ST-11 sublineage. 937 isolates within pubMLST were annotated with lineage and sublineage information, the table shows the number of isolates with the different alleles across sublineages.

within these lineages in pubMLST revealed that their presence was highly dependent on sublineage (Table 4.6). The major allele dominated in sublineage 1, but the minor allele dominanted in sublineage 2 (Table 4.6). The majority of the isolates assessed from pubMLST were invasive (93%, 876/937) as data deposited in pubMLST is biased towards invasive isolates, however this suggested that the presence of the minor allele at the significant SNP depends on ST-11 sublineage.

4.4.6.3.1 Variants in *fba* may create an additional putative FNR box

The FNR (fumerate and nitrate reduction regulator) protein is a global transcriptional regulator controlling gene expression in response to anaerobiosis in many bacteria. The FNR binding site consists of a highly conserved 5bp inverted repeat, separated by four non-specific bases (Eiglmeier et al. 1989). Previous studies have identified *fHbp* as part of the FNR regulon (Bartolini et al. 2006) and revealed a putative FNR box within the *fHbp* promoter sequence (TTGAC-N_4-ATCAA) just overlapping the -35 hexamer, differing from the *E. coli* FNR box consensus (TTGAT-N_4-ATCAA) by just three nucleotides (Table 4.7) (Eiglmeier et al. 1989; Spiro & Guest 1990; Oriente, Scarlato & Delany 2010). Levels of fHbp vary between strains, but this cannot be explained by mutations in the FNR box and cannot be correlated with the allele expressed, clonal complex, or geographic association (Masignani et al. 2003; McNeil et al. 2009; Oriente, Scarlato & Delany 2010). The significant SNP in *fba* was synonymous and the two codons were present at similar frequencies so the variant is unlikely to effect translational efficiency. Therefore, we hypothesise that the significant SNP in *fba* could have a regulatory effect on fHbp.

FNR box sequences													
E. coli consensus (Eiglmeier et al. 1989; Spiro & Guest 1990)	T	T	G	A	T	-	N_4	-	A	T	C	A	A
P_{fhbp} (Oriente, Scarlato & Delany 2010)	T	T	G	A	C	-	N_4	-	C	T	C	A	T
Major alleles (carriage associated)	T	C	G	A	T	-	N_4	-	G	C	A	A	A
Minor alleles (disease associated)	T	T	G	A	C	-	N_4	-	G	C	A	A	A

Table 4.7 SNPs from the current study in fba aligned with the predicted N. meningitidis fHbp promoter sequence and the E. coli FNR promoter consensus sequence.

Many FNR regulated promoters contain two DNA binding sites for FNR (Guest et al. 1996). FNR typically acts as a class II activator but it can also repress expression (Williams, Wing & Busby 1998). An example of this is the *E. coli yfiD* promoter dependent on FNR binding centred at position -40.5 in response to oxygen depletion (Marshall et al. 2001). When FNR is activated it occupies the site at -40.5 activating expression. As FNR becomes fully activated, the lower affinity upstream site at -93.5 is filled, downregulating expression and providing a mechanism for microaerobic induction of gene expression (Marshall et al. 2001). The spacing between and relative affinities of the two FNR sites are also crucial for FNR-mediated repression, with the inward-facing subunits being the most important (Marshall et al. 2001).

Examination of the sequence surrounding the significant SNP at position 900 plus the SNP at position 897 just below significance revealed the presence of a putative FNR box sequence. The major alleles combined produced the sequence TCGAT-N_4-GCAAA at positions 896-915, and the minor alleles the sequence TTGAC-N_4-GCAAA. This sequence defined by the minor alleles, which are risk alleles (OR = 6.3 and 5.8), differs from the *E. coli* FNR box consensus by just four nucleotides, with an upstream element identical to the upstream sequence of the putative P*fhbp* FNR box (Table 4.7) (Oriente, Scarlato & Delany 2010). It also has the same inverted repeat pattern match as the FNR

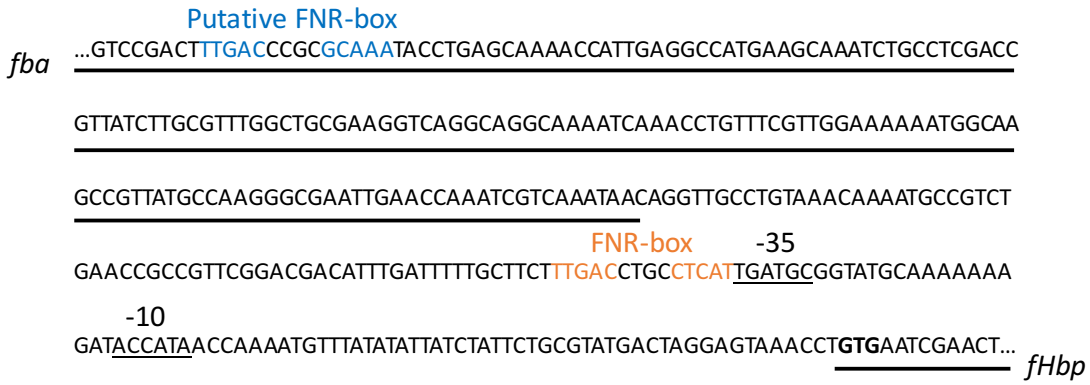

Figure 4.25 Sequence of the promoter region of *fhbp*. -35 and -10 elements for *fhbp* are underlined, and the *fhbp* start codon is shown in bold. The FNR-box previously predicted by Oriente, Scarlato & Delany (2010) is shown in orange, and the putative FNR box identified in the current study is shown in blue. Sequence shown is that of the reference genome MC58 (NC_003112.2) which contains the minor allele of the two SNPs of interest in *fba*.

box of the *fhbp* promoter.

This putative FNR box is separated from the FNR box of Pfhbp by 220bp. Studies investigating promoters with two FNR binding sites have explored the effect of distance between sites on FNR repression and found decreased repression with increased distance between FNR sites at distances shorter than 220bp, and it has been suggested that optimal centre-to-centre distances between two FNR sites are between 43-44bp (Barnard, Green & Busby 2003). However, FNR-mediated repression has been found to occur at greater distances. Marteyn et al. (2010) found that enhanced cell entry of *Shigella* in the absence of O_2 was dependent on FNR. A reduction of mRNA levels of two genes, *spa32* and *spa33*, was associated with decreased transcription in an FNR-dependent manner. *spa32* and *spa33* have two predicted FNR binding sites in their promoters, -156 and -67, and -205 and -116 upstream of the initiation codons, respectively, thus showing that FNR-mediated repression can occur at greater distances from the initiation codons and with greater spacing than 43-44bp. We therefore hypothesise that the minor alleles of the SNP in *fba* significantly associated with invasive disease, plus the SNP just below genome-wide significance, create an additional putative FNR box of potentially lower affinity affecting the transcription of fHbp (Figure 4.25).

4.5 Discussion

4.5.1 Summary

In this chapter, a genome-wide association study of carriage versus invasive disease of a natural population of 261 *Neisseria meningitidis* isolates from the Czech Republic in 1993 was undertaken. The principal findings were:

- We investigated whether we could identify loci significantly associated with carriage versus invasive disease and identified seven SNPs and 465 kmers counted from Velvet assemblies as significantly associated.
- All significant variants were of a reasonable frequency. The SNP MAFs were 16 and 26% and the kmer MAFs ranged from 3-43% with a mean of 18% and therefore could plausibly explain a reasonable proportion of the variance of the phenotype.
- We identified one lineage PC-1, representing ST-11, as significantly associated with the phenotype and invasive disease. This shows that the Wald test for lineage effects can identify known hyperinvasive lineages and therefore possibly identify unknown hypervirulent lineages using just bacterial genetic data.
- We identified significant variants in three genes involved in capsule production.
    - We hypothesise that significant kmers within the intergenic region between *ctrE* and *ctrF* captured a combination of three SNP alleles and the promoter of *ctrF*, and that divergence from the reference sequence in this region is associated with invasive disease, possibly due to an ablation of or increase in expression of *ctrF*.
    - Significant kmers in the gene *ctrG* appeared to be capturing SNP variation not identified by a SNP analysis, highlighting a potential need to account for deletions when imputing SNP data.

    - Significant kmers captured a phase variable region in *siaD*, with their presence appearing to represent the "on" state within the poly(A) phase variable region.
- We also identified significant genes in a further phase variable region, within the gene NMC1156, where the kmers appeared to be capturing particular length variants of the gene.
- We hypothesise that the combination of two SNPs in the gene *fba*, one significant and one just below genome-wide significance, create a putative FNR box affecting the transcription and therefore expression of fHbp.

Despite the small sampling frame, in many cases we were able to pinpoint statistical associations to individual loci, which is likely due to the high rate of recombination in meningococci (Feil et al. 1999). We previously demonstrated low power to detect locus effects for high MAF SNPs in *Escherichia coli* and *Klebsiella pneumoniae* simulations at similar sized sampling frames to the meningococcal dataset due to controlling for population structure and multiple testing (Chapter 3.4.4). However, in the current study of just 261 isolates, 7 SNPs and 465 kmers were significantly associated with the phenotype. This is despite the penalty for multiple testing correction; recombination breaks down linkage disequilibrium between sites increasing the number of unique phylopatterns, making the correction for multiple testing more stringent.

All significant variants identified represented types of genetic variation, rather than the presence and absence of genes, supporting and augmenting previous work positing that there is no virulence gene pool (Snyder & Saunders 2006; Schoen et al. 2008). Some results represented putative phase variable regions which enable the generation of genetic diversity at the point of replication. Kmers counted from Velvet assemblies representing putative phase variable regions were significant in *siaD* and

NMC1156. The phase variable regions have been experimentally validated in both *siaD* (Hammerschmidt et al. 1996; Weber et al. 2006; Siena et al. 2016) and NMC1156 (Siena et al. 2016). Given that these results were dependent on correctly identifying the length of the homopolymeric regions, they are susceptible to the occurrence of false positives by sequencing errors. Insertions and deletions have been shown to occur in up to 2% of bases within homopolymers by Illumina HiSeq data (Minoche, Dohm & Himmelbauer 2011). Genes associated with phase variation are typically surface-exposed or associated with surface structure biosynthesis or modification and therefore have been thought to be involved in virulence and host immune escape (Parkhill et al. 2000), however many phase variable regions have also been identified in non-pathogenic strains questioning this belief (Siena et al. 2016). Our results suggest that at least two phase variable regions are important in invasive disease in natural populations, and larger sample sizes may reveal the importance of additional phase variable regions. Further sequencing and functional validation would be required however given the potential for sequencing errors.

The heritability of the phenotype was identified and the phenotype was predicted from the genetic data. The heritability of the phenotype was estimated to be 36% with a standard error of 10%, higher than the R^2 between phenotype and clade of 23.7%, suggesting that the genetic data can explain more than just clade level differences in phenotype. This means that we should also be able to make phenotype predictions at a finer resolution than simply hyperinvasive lineage versus non-hyperinvasive lineage. We were in fact able to make reasonable predictions of the phenotype using the bacterial genetic data. All truly invasive isolates were predicted to be invasive, however some true carriage isolates were also predicted to be invasive. This was largely due to all ST-11 isolates being predicted as invasive isolates. The association of lineage PC-1 with invasive disease and its higher prevalence of invasive isolates partly explains, from a

statistical perspective, why all ST-11 isolates are predicted to be invasive. Epidemiologically, the tendency to over-predict invasiveness among ST-11 isolates suggests that the heightened virulence potential of this lineage is incompletely penetrant, and probably depends on a range of other factors, not least host health.

Previous studies have shown that lineages such as ST-11 are hypervirulent, and colonisation of hypervirulent strains is a known risk factor for invasive disease, but the underlying bacterial genetic variants remained undetermined. This work sheds new light on the variants contributing to ST-11 hyperinvasiveness through the identification of variants associated with invasion versus carriage that are also most correlated to the ST-11 lineage. In this case, these variants are also the most significant variants in the study.

4.5.2 Pitfalls with applying GWAS to bacteria

We identified some technical problems with conducting an association study in *Neisseria meningitidis* due to the large proportion of repetitive DNA. Counting kmers from sequencing reads requires a presence threshold which each kmer must pass to be defined as present. The presence of repetitive DNA makes this difficult, as errors in reads capturing repetitive regions could end up crossing thresholds which would be reasonable for the rest of the genome. Although the significant results representing repetitive regions when counting kmers from sequencing reads have the potential to be very interesting, we currently cannot distinguish them from false positives and therefore we chose not to interpret these results. Future work could be done to use individual count thresholds per genome, although this might not fully solve the issue. A superior approach to the elimination of WGS batch effects is to sequence the whole dataset at the same time, with experimental randomisation of cases and controls. Repetitive regions shown to be significant could be independently sequenced to verify whether they are genuine or not, although this would not solve the problem of false negatives. Counting kmers from

Velvet assemblies resulted in a more uniform number of kmers per isolate across the dataset, however this restricted the dataset to regions which could be assembled, meaning that we will inevitably have lost some information in the repetitive regions. We await low-error long-read sequencing to fully explore the bacterial genetic basis of virulence in meningococci and other important pathogens.

4.5.3 Validation

The results of this chapter led to a collaboration with Christoph Tang, who is leading a project aiming to functionally validate the results of the identified variants in *fba* by testing for altered expression of fHbp within an isogenic ST-11 background, differing only at the two identified SNP positions in *fba*.

Chapter 5

Genome-wide association study of *Campylobacter jejuni* wild bird versus chicken adaptation

5 Genome-wide association study of *Campylobacter jejuni* wild bird versus chicken adaptation

5.1 Introduction

Campylobacter, specifically *Campylobacter jejuni* and *Campylobacter coli* are major causes of human morbidity worldwide and are among the most common bacterial causes of gastroenteritis in humans in industrialised countries (Friedman et al. 2000). Around 90% of human disease is caused by *C. jejuni* and the majority of the remainder by *C. coli,* although many infections are likely not reported (Wheeler et al. 1999; Gillespie et al. 2002). Both species have multiple hosts and are also often found as harmless inhabitants of the gut microbiota of farmed animals (cattle, pigs, poultry and sheep), wild animals and birds (McCarthy et al. 2007; Griekspoor et al. 2013). *Campylobacter* can also be frequently isolated from water, sewage and the environment (Jones 2001).

5.1.1 Human disease

Human disease is sporadic, and routes of infection are typically through ingestion of contaminated food and from the environment. The main sources of human infection are farmed animals, in particular poultry (Sheppard et al. 2009; Strachan et al. 2009; Wilson et al. 2008) although more recently wild birds have been shown to be a consistent, albeit low, source of infection in Oxfordshire (Cody et al. 2015). The main risk factors for infection have been shown to be drinking non disinfected water, eating at barbecues and restaurants, consumption of undercooked chicken, raw seafood and oysters and unpasteurised milk and contact with farm animals (Kapperud et al. 2003; Friedman et al. 2004).

5.1.2 Host sources

Strains vary in their ability to colonise different hosts, and also in their ability to survive in the environment. Genotyping strains from multiple sources by multilocus sequence typing (MLST) has revealed that the population structure of *C. jejuni* segregates into specialist and generalist lineages (Sheppard et al. 2014). The two largest generalist lineages are the ST-45 and ST-21 complexes based on their phenotypic flexibility and large genetic diversity (Dingle et al. 2001; Gripp et al. 2011). These two generalist lineages are phylogenetically divergent despite their convergent ecological strategies (Sheppard et al. 2014). Although both complexes contain DNA estimated to have been obtained from various host specialist lineages, they are significantly isolated from one another based on an admixture analysis (Sheppard et al. 2014). This cannot be due to a lack of opportunity to recombine as strains are frequently isolated from the same hosts, and in vitro transformation experiments confirmed that recombination is functionally possible, suggesting that they potentially occupy different niches within the host (Sheppard et al. 2014).

Assessing strain prevalence across different sources enables the estimation of source attribution in cases of human disease (Wilson et al. 2008; Sheppard et al. 2009; Strachan et al. 2009; Pires et al. 2010). Dearlove et al. (2016) investigated using whole genome sequencing to improve the accuracy of attributing human cases to particular animal hosts. They focused on the common generalist strains, ST-45 complex, ST-21 complex and ST-828 complex, but found no phylogenetic association with host reservoir. This revealed at a finer-scale that there are no host associated sub-lineages within these complexes and that they appear to be genuine generalists (Dearlove et al. 2016). Other lineages have been shown to be specialists, such as the ST-61 complex which although frequently linked with human infection has been shown to be cattle and sheep associated (French et

al. 2005; Colles 2003). Both *C. jejuni* and *C. coli* show stronger evidence for host associated lineages than they do for geographic structure (Sheppard et al. 2010).

One hypothesis for the existence of generalist lineages is that they could have recently colonised multiple hosts and not had sufficient time to evolve host specialism (Sheppard et al. 2014). This however is not backed by the genetic diversity of generalist lineages, which is equivalent to that of specialist lineages, plus generalist lineages do not show evidence for recent clonal expansion (Sheppard et al. 2014). It appears that not all strains have equal potential to colonise different hosts, for example an *in vivo* infection experiment demonstrated that the European Robin could be colonised with a *C. jejuni* isolate retrieved from a closely related bird species but not from a human (Waldenström et al. 2010).

Strains also vary in their ability to survive in the environment outside of a host. *C. jejuni* seems to lack many genes involved with environmental stress resistance in other foodborne pathogens such as *E. coli* and *B. subtilis* (Park 2002). *Campylobacter* have been shown to be able to enter a viable but nonculturable (VBNC) state under unfavourable conditions such as low available nutrients. *Campylobacter* transform from a motile spiral form to a coccoid form with a reduction in size which could be an important contribution to survival in the environment (Rollins & Colwell 1986; Hazeleger et al. 1995). The ability to enter a VBNC state varies by strain, with some surviving for just days and others months (Talibart et al. 2000; Rollins & Colwell 1986). It is contentious whether *C. jejuni* can revert back from the VBNC state to a state of colonisation. The ability to revert appears to vary by strain, as some studies could not revert VBNC *C. jejuni* to colonise chicks (Beumer, de Vries & Rombouts 1992; Ziprin et al. 2003; Ziprin et al. 2004; Mederma et al. 1992) but others could (Baffone et al. 2006; Talibart et al. 2000; Saha, Saha & Sanyal 1991; Stern et al. 1994).

The generalist ST-45 strain has been shown to be able to survive in the environment. In a study of a structured sample of dairy farmland in the UK, wildlife and water isolates were most frequently ST-45 complex (French et al. 2005) and it was also shown to be the most prevalent ST within a two-year study sampling river water in the UK, where it was also reported to exhibit seasonality, as it was only found between April and August (Sopwith et al. 2008).

5.1.3 The *C. jejuni* genome

The *C. jejuni* genome is relatively small, typically around 1.6Mb in length (Parkhill et al. 2000; Pearson et al. 2007). *C. jejuni* has a weakly clonal population structure and segregates into clonal complexes (Dingle et al. 2001) despite being naturally competent for DNA uptake and diversity being generated by recombination at twice the rate of mutation (Wang & Taylor 1990; Wilson et al. 2009). Genetic exchange also occurs between *C. jejuni* and the closely related species *C. coli* and it has been suggested that this is causing the species to be converging (Sheppard et al. 2008). *C. jejuni* has a large pangenome; an estimate of the pangenome size of 130 *C. jejuni* isolates using a seven isolate pangeome reference panel was 3648 genes (Méric et al. 2014). The size of the pangenome does appear to plateau with the number of genomes included, as 92% of the total pangenome size was identified from just seven genomes and 99% from 75 of the genomes.

The genome of *C. jejuni* is unlike other Gram negative pathogenic bacteria, such as *Neisseria meningitidis*, as it has a very small number of repeat sequences and just one gene related to insertion sequences and phage associated sequences (Parkhill et al. 2000). A mechanism by which *C. jejuni* generates diversity at particular loci however is that of phase variation. Multiple genes have been identified to vary in length of polyG:C tracts (Parkhill et al. 2000). Phase variation enables adaptation to changing conditions due to

the switching of gene expression state as a result of the lengthening or shortening of repeat tracts. Homopolymeric regions in *C. jejuni* are typically found in genes related to production of three major surface structures, the capsule, lipooligosaccharide (LOS), and the flagella (Parkhill et al. 2000; Bacon et al. 2001; Caldwell et al. 1985; Guerry et al. 2002; Karlyshev et al. 2002).

5.1.4 Host adaptation

The focus of the current study is the colonisation of chickens and wild birds where they are colonised as a commensal organism with the possible exception of Ostriches (Verwoerd 2000). Chickens are a natural host for *Campylobacter*, and are frequently found to be colonised with *C. jejuni*. *C. jejuni* primarily localises in the mucus overlying the epithelial cells of the caeca, large intestine and cloaca, although systemic colonisation is sometimes detected outside of the gastrointestinal tract in the liver and spleen (Berndtson, Tivemo & Engvall 1992; Stern et al. 1988; Beery, Hugdahl & Doyle 1988; Meade et al. 2009). Colonisation ability appears to be able to evolve quickly; multiple studies have shown isolates which are poor at colonising chicks being able to consistently colonise chicks after several faecal-oral passages through chicks (Stern et al. 1988).

C. jejuni has been found to be common amongst many wild birds with prevalence varying by species, diet and season (Luechtefeld et al. 1980; Kapperud & Rosef 1983; Waldenström et al. 2002; Broman et al. 2002). Wild birds are thought to be an important source of infection to chickens by either gaining access to housed chickens or by contaminating accessible surface water or soil. Genotypes isolated from different wild bird species have been shown to be typically genetically distinct by MLST (Griekspoor et al. 2013). Wild birds have been shown to both be part of common generalist lineages and also specialist lineages such as a crow-only cluster identified by Weis et al. (2016). Colles et al. (2008) found that ST-21 and ST-45 complexes were the only lineages present in

geese, starlings and free-range chickens isolates sampled from the same location in Oxfordshire, therefore indicating little evidence for transmission between the hosts.

5.1.4.1 Host adaptation mechanisms

Although not fully understood, *C. jejuni* has clearly developed mechanisms which allow it to colonise and survive within its various hosts. Several genes and pathways allowing colonisation within chicken hosts have been identified to date. One such mechanism is the *Campylobacter* multidrug efflux pump (CME) encoded by the operon *cmeABC* which along with contributing to multidrug resistance also enables resistance to bile salts in the intestinal tract of the chicken, required for successful colonisation of *C. jejuni* in chickens (Lin et al. 2003).

Another is chemotaxis, a sensory system which senses stimuli to enable the directing of movement towards or away from favourable or unfavourable conditions mediated using its two flagella (Hugdahl, Beery & Doyle 1988). *C. jejuni* is attracted by the main component of mucus, the glycoprotein mucin, the mucin component L-fucose, bile, various amino acids, the salts of some organic acids, L-asparagine, formate and D-lactate (Hermans et al. 2011; Hugdahl, Beery & Doyle 1988; Vegge et al. 2009). Analysis of the genome of reference strain NCTC11168 revealed chemotaxis and aerotaxis gene orthologs (Marchant, Wren & Ketley 2002), and *C. jejuni* encodes for an extensive network of genes that are responsible for chemotaxis (Hendrixson & DiRita 2004). Transmembrane methyl-accepting chemotaxis proteins (MCP), also known as transducer-like proteins (Tlp), sense the chemotactic signals and then transmit the signals to the cytoplasmic core chemotaxis signal transduction (Che) proteins network (Marchant, Wren & Ketley 2002). The methylation status of the Tlp chemoreceptors are mediated by the enzymes CheB and CheR encoded by *cbeB* and *cheR* which encode a methylesterase and a methyltransferase respectively (Stephens, Loar & Alexandre 2006). Mutants of both

genes have been shown to have a reduced ability to colonise chick caeca (Kanungpean, Kakuda & Takai 2011). The Tlp signalling domains also form ternary complexes with CheA, a histidine kinase and CheW which couples CheA to the receptors (Gegner et al. 1992). Information sensed by the chemoreceptors is then transmitted to the flagellar motor switch proteins FliM and FliN via CheA and CheY, the response regulator (Sarkar, Paul & Blair 2010).

Motility and the flagella are also important for chicken colonisation (Hendrixson & DiRita 2004; Nachamkin et al. 1993). *Campylobacter* species are motile due to one or two polar flagella encoded by the adjacent genes *flaA* which encodes the major flagellin, and *flaB* which encodes the minor flagellin (Nuijten et al. 1990). *flaA* is important for optimal chicken colonisation (Wassenaar et al. 1993) but both are important for motility (Neal-Mckinney, Christensen & Konkel 2010). Motility is thought to be important so that *C. jejuni* can reach its niche, the mucus layer of the caecal and cloacal crypts (Beery, Hugdahl & Doyle 1988).

Other important factors behind chicken colonisation include surface accessible carbohydrate structures such as the previously mentioned flagella, lipooligosaccharides (LOS), the capsule and *O*- and *N*-linked glycans (Hermans et al. 2011). The *C. jejuni* surface polysaccharide LOS is an outer membrane glycolipid consisting of lipid A, a hydrophobic anchor, and an oligosaccharide (Karlyshev, Ketley & Wren 2005). Nineteen LOS classes have been identified in *C. jejuni* to date based on variable gene content (Parker et al. 2008) and LOS loci have been shown to be some of the most variable of the whole genome (Parkhill et al. 2000). Five of these classes contain loci necessary to synthesise sialylated LOS: classes A, B, C, M, and R (Parker et al. 2008).

LOS has been shown to be important in serum and bile resistance (Iwata et al. 2013), resistance to complement-mediated killing (Naito et al. 2010), chick colonisation,

adhesion, invasion and virulence (Javed et al. 2012; Perera et al. 2007; Kanipes et al. 2004; Louwen et al. 2008). *C. jejuni* LOS has also been identified as a TLR4 agonist, with LOS structure and sialylation status affecting TLR4 activation (Stephenson et al. 2013; Kuijf et al. 2010; Bax et al. 2011). Molecular mimicry between certain sialylated LOS structures and human gangliosides has been shown to cause Guillain-Barré syndrome (Yuki et al. 2004). Sialylated LOS structures have also been shown to confer bacteriophage resistance in *C. jejuni* (Louwen et al. 2013). Although isolates of unsialylated LOS class E also displayed bacteriophage resistance, they displayed inferior resistance in comparison to isolates containing a sialylated LOS (Louwen et al. 2013). It has also been observed that a reduced or absent CRISPR array, plus mutations in *cas* genes, is associated with containing *cst*-II loci and ganglioside-like LOS expression, hypothesising that the ganglioside-like LOS are a replacement for the CRISPR-Cas system for bacteriophage resistance (Louwen et al. 2013). LOS locus classes have previously been shown to be associated with *C. jejuni* population structure and particular MLSTs (Revez & Hänninen 2012).

Adhesion of *C. jejuni* to the chicken epithelial cells is also a key process in colonisation. Multiple studies have revealed the importance of adhesion, with reduced adhesion demonstrated in the absence of particular adhesins such as *capA*, *cadF* and *flpA* and sometimes also reduced colonisation (Ashgar et al. 2007; Flanagan et al. 2009). However, *capA* for example is not present universally in poultry isolates, so its contribution to chicken colonisation in natural populations is unclear (Flanagan et al. 2009).

5.1.4.2 Association studies

Various studies have therefore identified multiple genes and pathways to be associated with chicken colonisation, but these have typically been based on single knockout strains

and the presence or absence of these genes have not been validated across large natural populations. As *C. jejuni* contains host-associated lineages, this suggests that there is a genetic basis linking particular strains to particular hosts. Most studies investigating genetic host associations have done so at the lineage level using multi-locus sequence typing (MLST) or serotyping. Griekspoor et al. (2013) identified lineages associated with wild birds, observing strong nucleotide and allele population differentiation, but did not investigate which alleles were more frequent in wild birds. Lefébure et al. (2010) looked for evidence of host adaptation in 96 genomes of *C. jejuni* and *C. coli*, but found no link between gene content and host in *C. jejuni* using a Fisher's exact test. Although a link was found for *C. coli*, this could not be assigned to any particular gene. Lefébure et al. (2010) suggested that gene content is therefore not a major factor underlying host association and that instead gene regulation should be investigated. The study however was looking across isolates from multiple hosts, and did not control for population structure, which is necessary for association studies in bacteria.

In order to identify genomic variants associated with particular hosts, Hepworth et al. (2011) investigated 80 *C. jejuni* isolates obtained from humans, cattle, chicken, sheep, wild birds, rabbits, badgers and environmental water using comparative genomic hybridisation. A clade of water and wildlife isolates were defined by the absence of genes previously associated with chicken colonisation, such as genes in the *cdtABC* cluster encoding the cytolethal distending toxin, a human virulence factor. One representative strain was sequenced, revealing a deletion of the whole gene cluster, and a representative of a bank-vole restricted strain contained a deletion of *cdtA* (Hepworth et al. 2011).

Sheppard et al. (2013) applied a GWAS to investigate *Campylobacter* adaptation to cattle and chickens among 192 isolates. The kmer based study identified a seven-gene region which was cattle associated but often absent within isolates obtained from

chickens and wild birds. Three of the genes encoded proteins involved in vitamin B_5 synthesis, a vitamin that is rich in cereals and grains but low in grasses, the main diets of chicken and cattle respectively, suggesting that in order to survive within cattle, *Campylobacter* must produce the vitamin themselves.

Another genome-wide association study applied to *Campylobacter* investigated 600 genome sequences sampled from different stages of the poultry processing chain and human campylobacteriosis to investigate whether particular sequences were associated with human disease and therefore survival through the food chain (Yahara et al. 2017). Investigating the ST-21 and ST-45 complex isolates revealed genetic elements which increased in frequency through the poultry processing chain resulting in them being overrepresented in isolates obtained from human clinical cases. The elements included genes involved with formate metabolism, aerobic survival, oxidative respiration and nucleotide salvage, however associated variants were distinct between the ST-21 and ST-45 complexes indicating that they have evolved different solutions to survival through the food chain and colonising the human host and causing disease (Yahara et al. 2017).

Understanding how *C. jejuni* adapts to its various niches and hosts is key to understanding how it is maintained in the various hosts and therefore remains an infection source. More generally, over 60% of human infectious diseases are caused by pathogens that also infect animals (Karesh et al. 2012). Characterising host associated variants can assist in improving host attribution studies to gain a greater understanding of transmission and how *C. jejuni* is introduced and maintained in chicken flocks. For effective disease control, we need a greater understanding of the biology of the pathogen and how it adapts to its various hosts. Here we apply a genome-wide association study to investigate whether we can identify genetic variants within *C. jejuni* associated with adaptation to wild birds and chickens and also test whether we can identify host associated lineages.

5.2 Chapter Aims

- Investigate lineage-level differences in wild bird vs chicken host association and identify chicken or wild bird associated strains.
- Estimate heritability of the host association phenotype and how well we can predict the phenotype.
- Investigate associations between SNPs and kmers and chicken versus wild bird colonisation, and their implications on our understanding of host adaptation in *C. jejuni*.

5.3 Methods

5.3.1 Sampling frame

The dataset consisted of 489 *C. jejuni* isolates collated by Samuel Sheppard and Guillaume Méric who provided paired end Illumina reads. All isolates were sampled from either wild birds or chickens.

5.3.2 Variant calling

The sequencing reads were mapped to the reference genome NCTC11168 (Accession NC_002163), belonging to the ST-21 clonal complex, SNPs were conservatively called, and uncalled bases were imputed using ClonalFrameML (Didelot & Wilson 2015) as discussed in Chapter 2.3. 31bp kmers were counted using DSK from Velvet assemblies also as discussed in Chapter two. The numbers of all variants and unique phylopatterns can be found in Table 5.1. Bonferroni-corrected significance thresholds were calculated for the SNP and kmer analyses based on the number of phylopatterns with a MAF ≥ 1%, resulting in thresholds of 6.3 and 7.2 for SNPs and kmers, respectively. The Bonferroni-corrected significance threshold for lineages was 4.0.

Variant type	Biallelic SNPs	Tri-allelic SNPs	Tetra-allelic SNPs	Variant kmers counted from Velvet assemblies
Number of variants	220,484	12,986	827	16,620,479
Number of unique phylopatterns	98,289	12,553	826	780,266

Table 5.1 Total numbers of variants and unique variant patterns across individuals for all variant types.

5.3.3 Determining the proportion of reads assigned to *Campylobacter jejuni*

Before proceeding with the analysis, quality control measures were taken. Kraken (Wood & Salzberg 2014) was used to determine the proportion of kmers belonging to all of the species in its database. This revealed kmer matches of between 61.7-99.1% to *Campylobacter*, 51.7-94.2% to *Campylobacter jejuni*, 4.7-13.5% unclassified *Campylobacter*, 0.0002-8% other species and 0.8-37% unclassified (Figure 5.1). The lower limit of the range of kmer matches to *Campylobacter* (61.7%) and the upper limit on the range of unclassified kmers (37%) suggested that this could be due to poor sequencing for some isolates. Isolates with a percentage match of less than 70% to *Campylobacter* were removed from all analyses leaving 480 isolates for further analysis.

5.3.4 Quality control: Read length estimation

Read length was predicted for each isolate. Kmers were tested for association using estimated read length mode as a fixed effect in the LMM to assess whether this resulted in a batch effect.

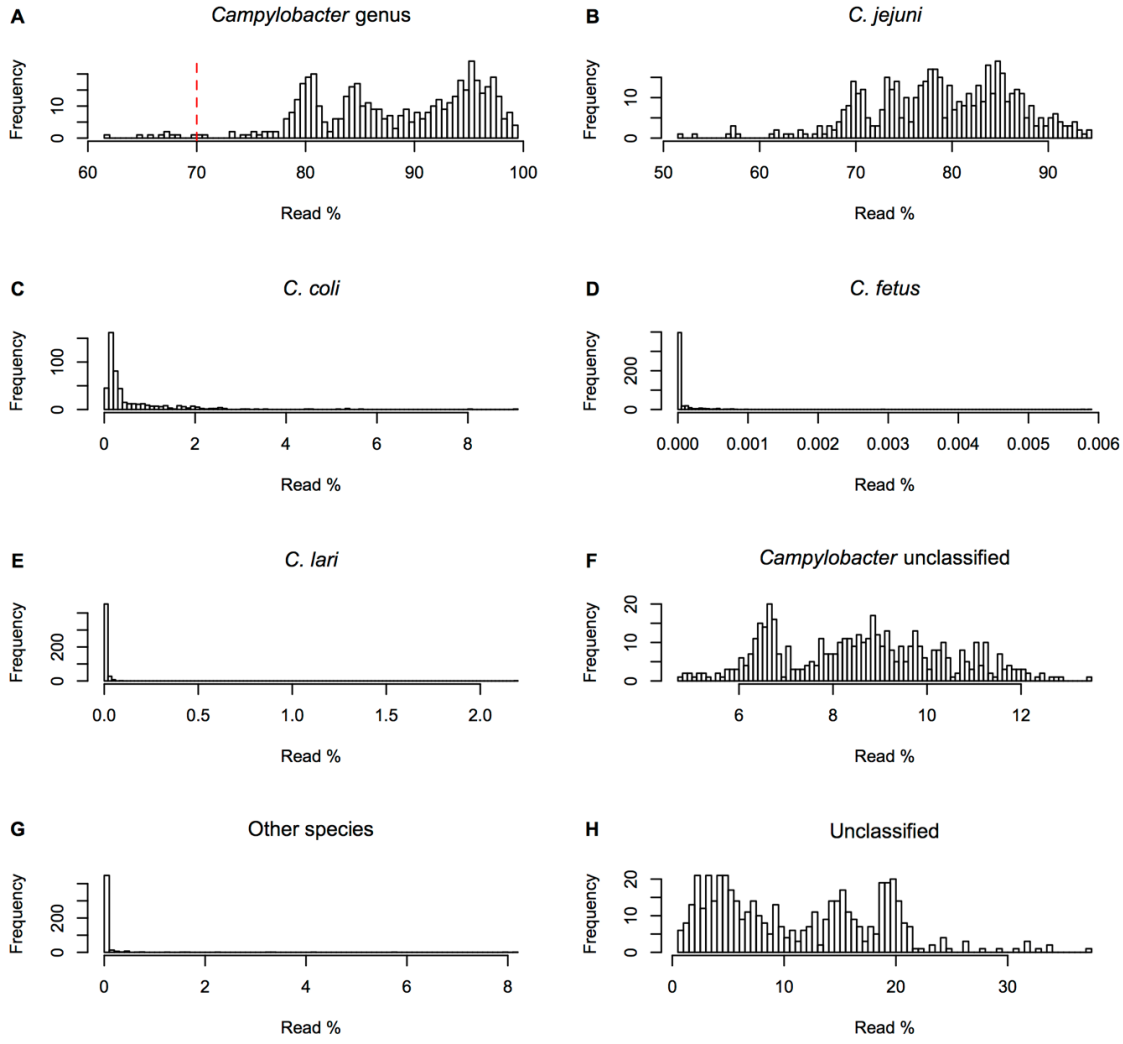

Figure 5.1 The proportion of reads assigned to *Campylobacter* species using Kraken. A *Campylobacter* genus; **B** *C. jejuni*; **C** *C. coli*; **D** *C. fetus*; **E** *C. lari*; **F** Kmers which could be assigned to the *Campylobacter* genus but could not be further classified to a species; **G** Non-*Campylobacter* species reads; **H** Unclassified kmers.

5.3.5 Calculating the kinship matrix and heritability using kmers

The kmers counted from Velvet assemblies were used to calculate the kinship matrix using GEMMA (Zhou & Stephens 2012) and kmers were tested for association using this kinship matrix and results compared with that of using the kinship matrix built from SNPs. The results of the LMM analysis using SNPs versus using kmers to build the kinship matrix were quite different (Figure 5.2). Using kmers to build the relatedness matrix may account for weak batch effects within the data, as these effects will be like

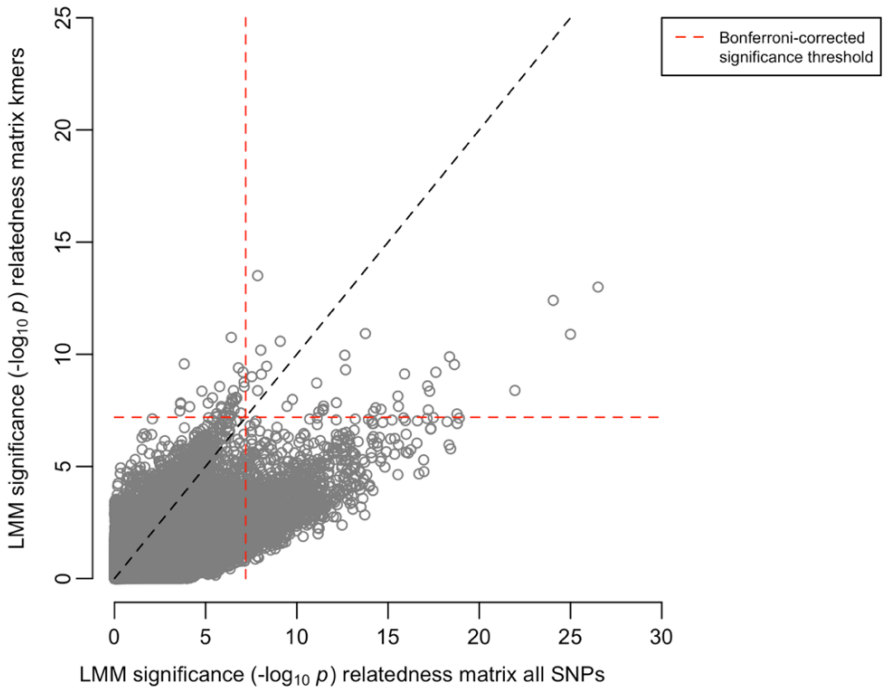

Figure 5.2 Kmer LMM kinship matrix comparison. Results of testing kmers for association with the phenotype using SNPs to control for population structure versus using kmers to control for population structure by building the kinship matrices using GEMMA for both types of variants. Significance is reduced when using kmers to build the kinship matrix, indicating superior control of unmeasured confounders such as weak batch effects, thus kmers were used to build the kinship matrix for all analyses.

axes of variation through the data which the kmers can then account for. All kmer LMM analyses therefore used kmers to build the kinship matrix.

5.3.6 Defining the presence of LOS alleles in Velvet assemblies

BLAST was used to identify LOS gene presence in the Velvet assembly contigs for all LOS classes identified to date using genomes from multiple reference genome as the queries (A, AF215659.1; B, AF401528.1; C, NC_002163.1; D, AF400669.1; E, NC_009839.1; F, AY434498.1; G, AY436358.1; H, EU404106.1; I, EU404107.1; J, EU404104.1; K, EU410350.1; L, EU404111.1; M, EF140720.1; N, AY816330.1; O, EF143352.1; P, AY943308.1; Q, EU404112.1; R, AY962325.1; S, EU404110.1) (Parker et al. 2008). A criterion of a match over at least 90% of the length with a minimum of 90% identity was used to classify genes across all LOS classes as present or absent. Presence vs absence of the genes was then tested for association with the phenotype,

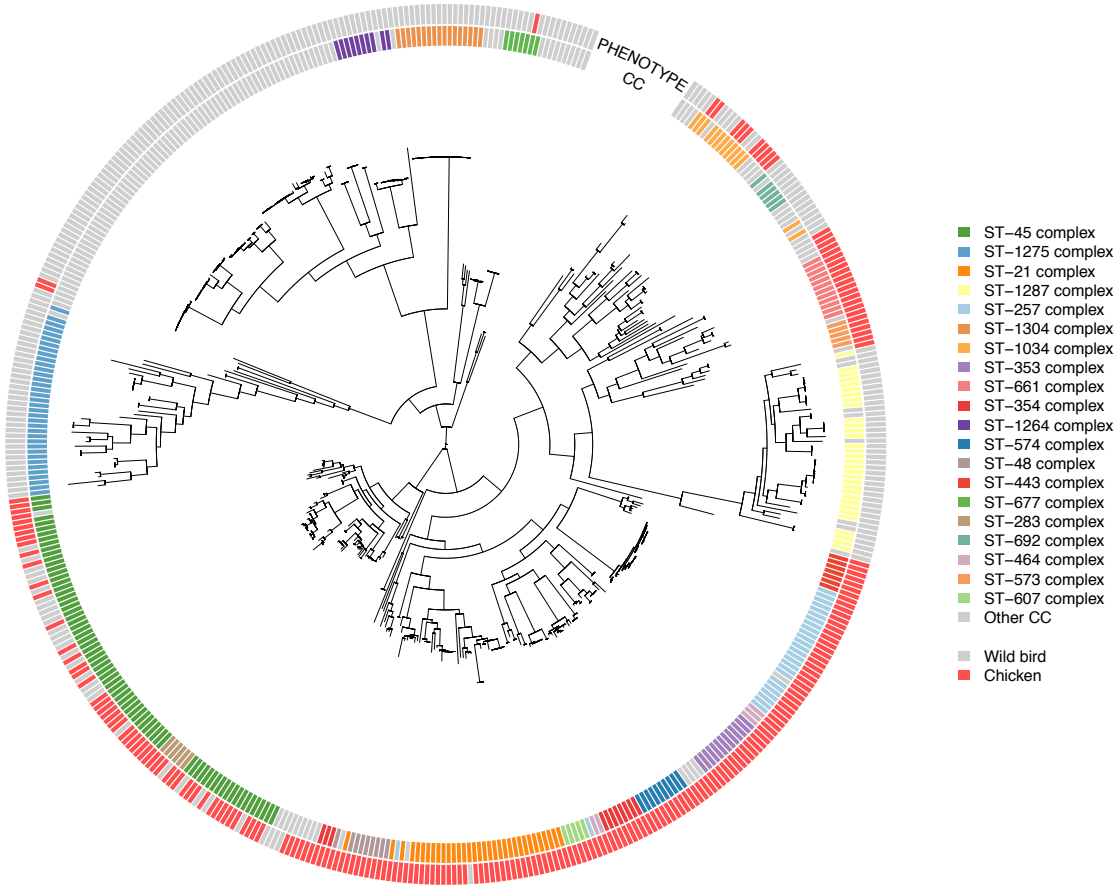

Figure 5.3 *Campylobacter jejuni* phylogeny. Maximum likelihood phylogeny built using RAxML based on biallelic SNPs and annotated with clonal complex (the 20 most common clonal complexes are coloured, all others are grey) and wild bird/chicken phenotype. Bootstrap supports are shown in Appendix B Figure B.3.

controlling for population structure using LMM (Zhou & Stephens 2012). A Bonferroni-corrected significance threshold based on the number of unique SNP phylopatterns with a minor allele frequency (MAF) of ≥1% (100,773) plus the number of unique LOS gene presence/absence phylopatterns with a MAF ≥1% (135) was applied, giving a Bonferroni-corrected significance threshold of $-\log10\ P = 6.3$.

5.4 Results

5.4.1 Population structure of the sampling frame

We built a maximum likelihood phylogeny of the data from the biallelic SNPs using RAxML (as described in Chapter 2.2). The midpoint rooted phylogeny is depicted in

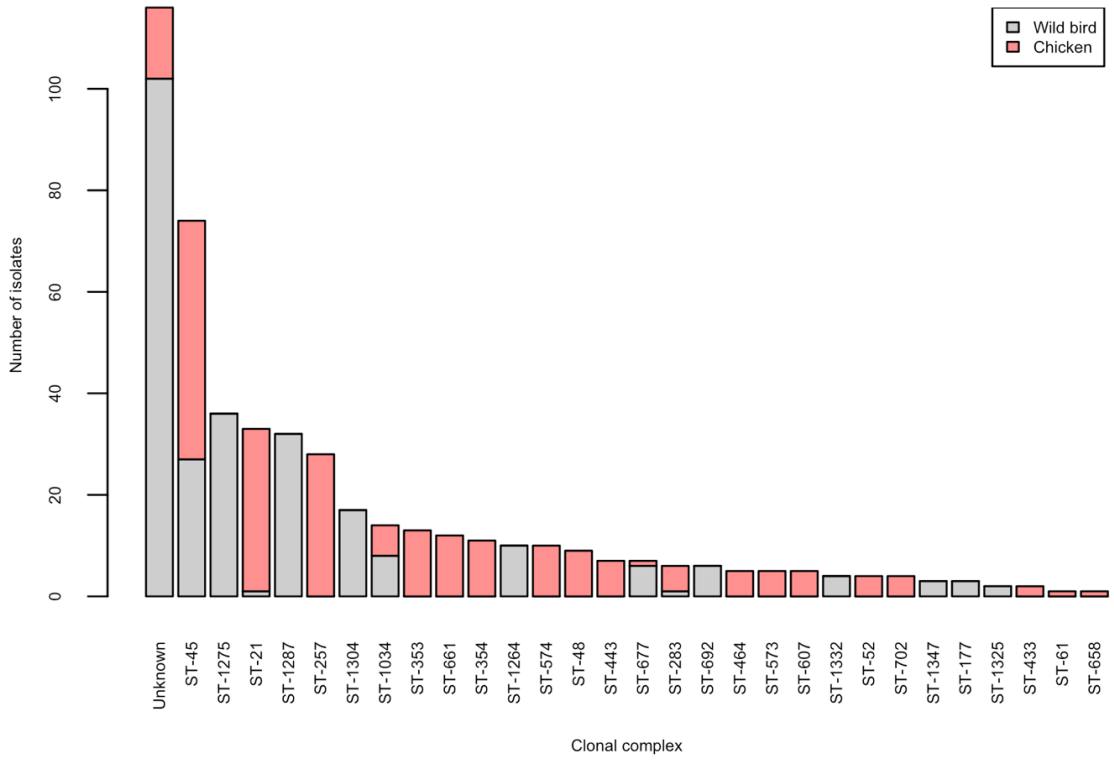

Figure 5.4 Distribution of wild bird and chicken isolates in each clonal complex. Total numbers of wild bird and chicken isolates per clonal complex are shown in grey and red respectively.

Figure 5.3 annotated by clonal complex (CC) designation and the phenotype, wild bird vs chicken colonisation. Bootstrap supports are shown in Appendix B Figure B.3. The population segregates into clonal lineages, as defined by clonal complexes, and the most common CC in the dataset was the generalist lineage ST-45 complex. The distribution of the phenotype across the major clonal complexes can be seen in Figure 5.4.

5.4.2 Read length estimation

Investigation of the read length of each isolate revealed that the dataset contained different read lengths (Figure 5.5). Most isolates had undergone adaptor trimming and therefore had variable read length. Eighty isolates had a constant read length, but 300 had the mode read length at 99% of the reads. All isolates with a mode read length of 301 were wild bird isolates, therefore a concern was that the different sequencing batches may have introduced batch effects which could have impacted the kmer analysis. Therefore,

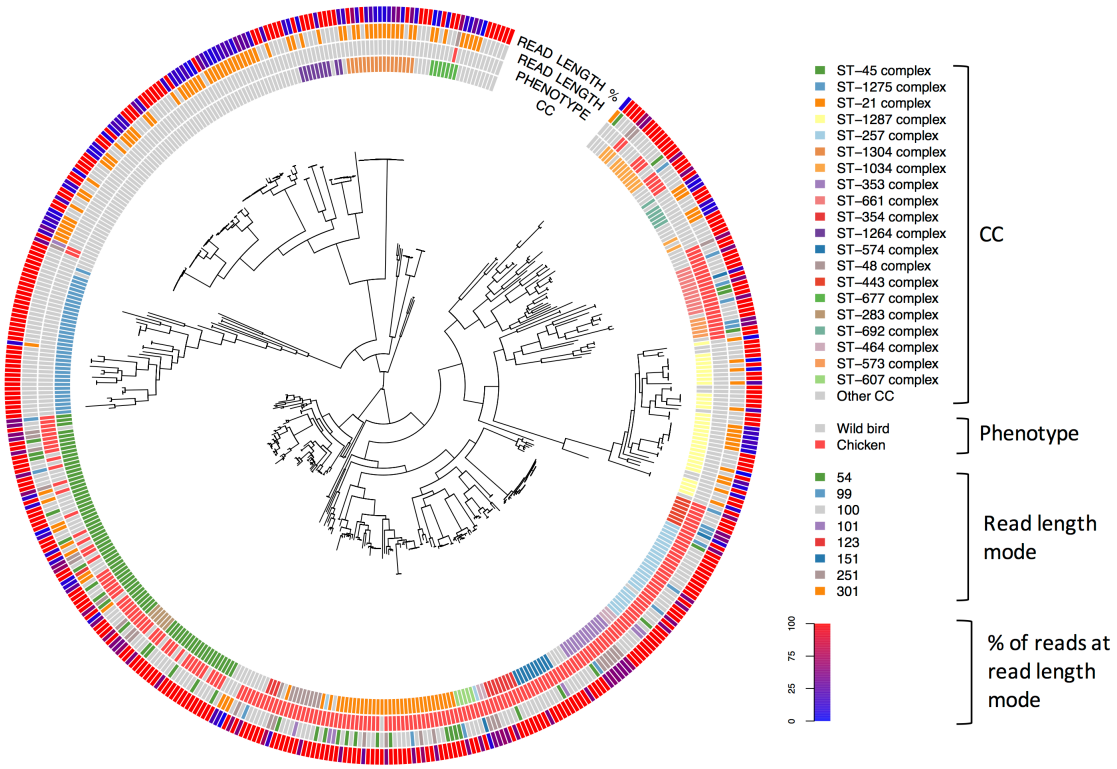

Figure 5.5 Read length mode annotated on the phylogeny. Phylogeny annotated by ST-complex, phenotype, read length mode and the percentage of reads at the read length mode per isolate. All isolates with a mode read length of 301 were wild bird isolates.

the effect of this was assessed in Section 5.4.4.2.

5.4.3 Four lineages were significantly host associated

Here, we investigated the proportion of phenotypic variation in host affinity explained by the bacterial genotype, and based on this identified how well the phenotype could be predicted from the bacterial genetic data. We also identified lineages defined by principal components and tested for their association with the phenotype.

5.4.3.1 Heritability of host association and predicting host association

Heritability, the proportion of the phenotypic variation that can be explained by the bacterial genotype, was estimated using the LMM null model in GEMMA (Zhou & Stephens 2012, Chapter 2.9.1). Heritability in the sample was estimated to be 83.6% with a standard error of 2.7%. This high estimate reflects the strong clustering of wild bird and

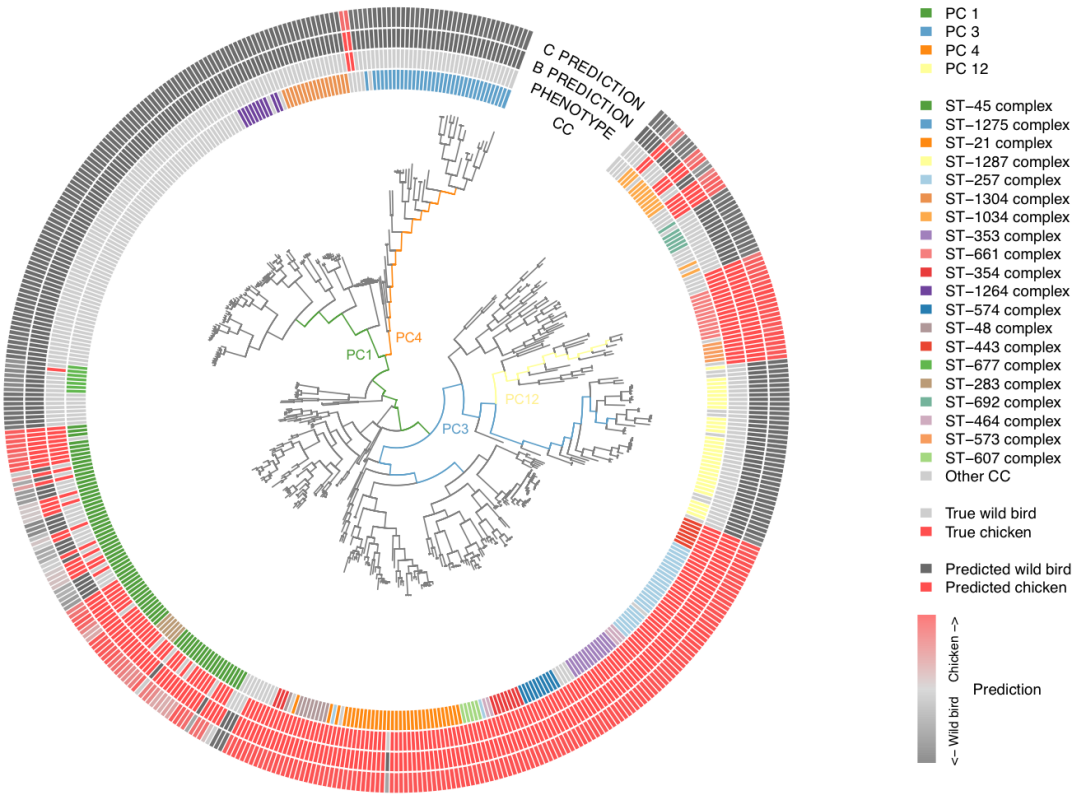

Figure 5.6 Predicting host association using SNP data. Phylogeny annotated by ST-complex, true phenotype (wild bird vs chicken colonisation), binary predicted phenotype ('B PREDICTION') and continuous predicted phenotype ('C PREDICTION'). The continuous prediction goes from dark grey (strongly predicted wild bird) through light grey to red (strongly predicted chicken).

chicken isolates on the phylogeny (Figure 5.3). The R^2 of the phenotype against clonal complex was 70%, meaning that WGS provided a greater explanation of the phenotype than clonal complex definitions (Figure 5.3).

The phenotype was predicted using the null Linear Mixed Model in which every SNP is used for prediction (described in Chapter 2.10.3), which is equivalent to a ridge regression (O'Hagan & Forster 2010). The phenotypes were coded as chicken colonisation (1) and wild bird colonisation (0) and mean centred, so positive values represented chicken colonisation. Binary wild bird vs chicken colonisation predictions resulted in correctly predicting 469/480 isolates (third ring from inside, Figure 5.6). The isolates wrongly predicted to be wild bird colonising were of the generalist ST-45 complex and the ST-677 complex. Of the isolates wrongly predicted to be chicken

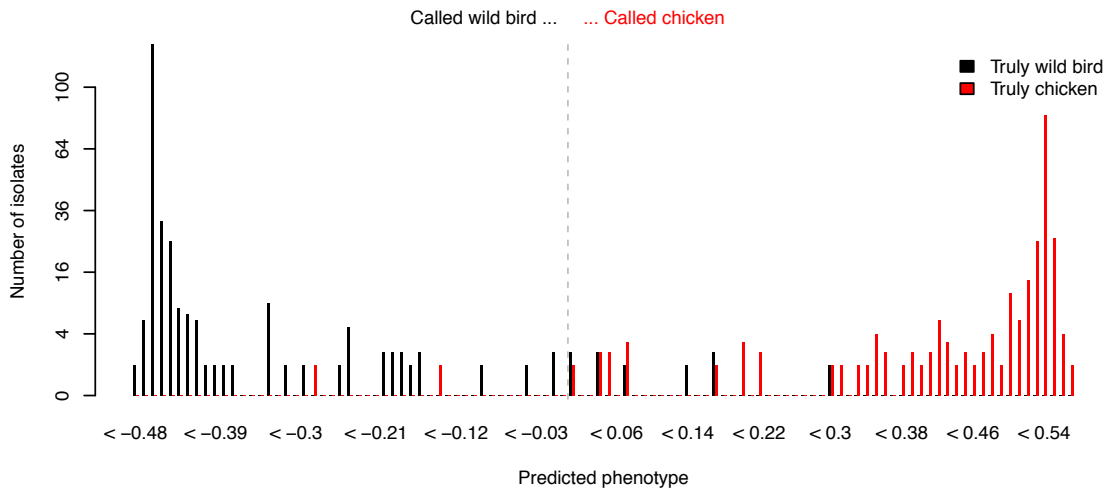

Figure 5.7 Phenotype prediction distribution. Histogram of the continuous phenotype predictions using SNP data. This shows that most predictions were strong, and that when a prediction was incorrect it was typically a weak prediction. Figure produced based on code provided by Daniel Wilson.

colonising, eight were ST-45 complex isolates and one was an ST-283 complex isolate. Visualising the predictions on a continuous scale on the phylogeny in Figure 5.6 (outer ring) and in Figure 5.7 revealed that the predictions made in the ST-45 complex, where the majority of incorrect predictions were made, were weaker predictions, i.e. closer to zero. Given that the ST-45 complex had the most phenotype switches and is a known generalist lineage, this was unsurprising (Dingle et al. 2001; Gripp et al. 2011).

5.4.3.2 Wald test for lineage-level differences in host association

Principal components analysis was performed on the biallelic imputed SNPs and a Wald test applied to assess the significance of associations between principal components (PCs) and wild bird colonisation versus chicken colonisation as described in Chapter 2.10.2. PCs 1-20 can be visualised on the phylogeny of the dataset in Figure 5.8 where the branch lengths have been square-rooted to allow for the visualisation of the branch colours at the tips of the phylogeny. The branches are coloured by assigning a pattern to each branch: 1 for all isolates one side of the branch, 0 for all isolates the other side of the branch. The colour represents the correlation of this pattern with the projection of the isolates onto

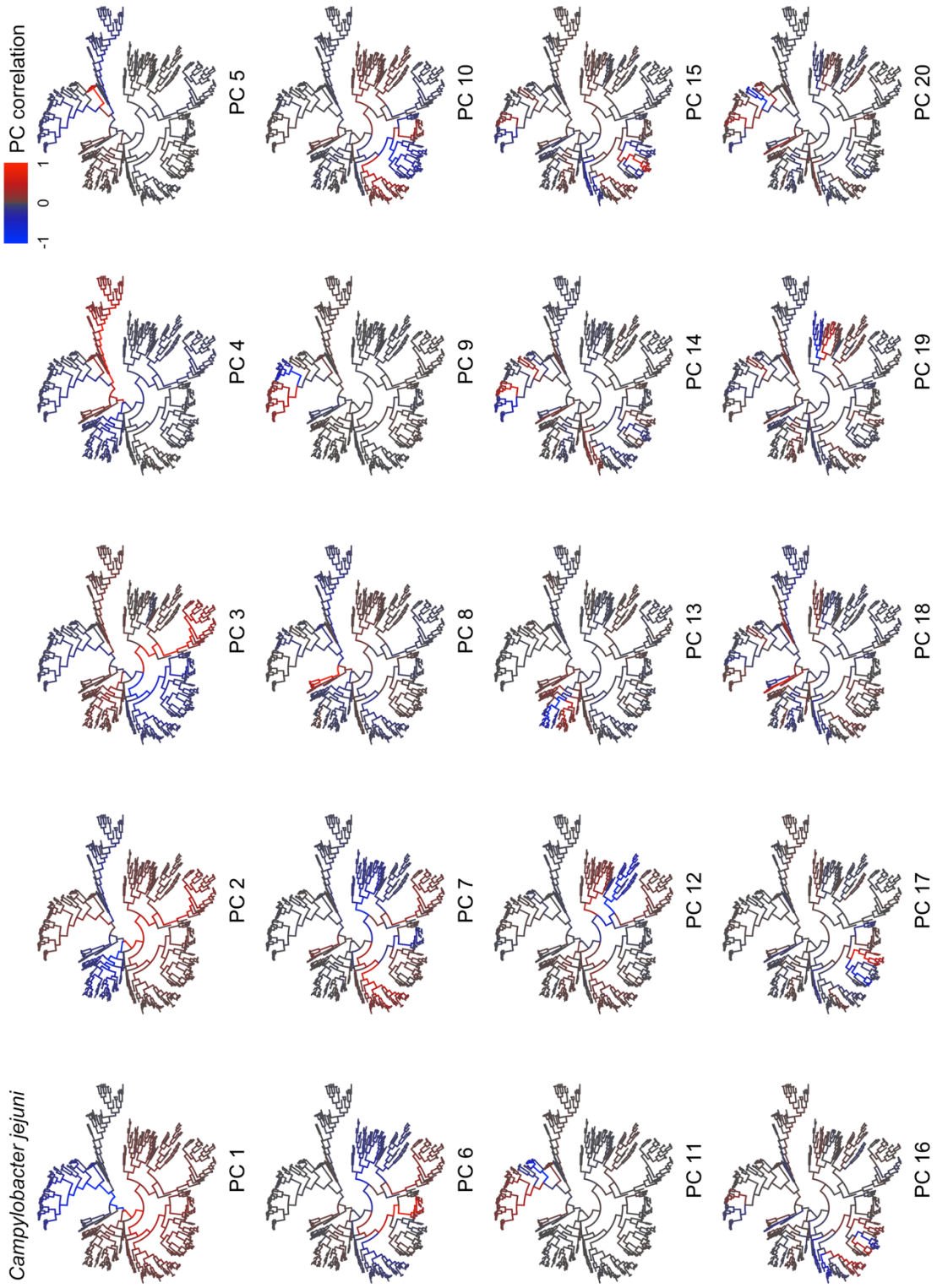

Figure 5.8 *C. jejuni* principal components annotated on the phylogeny. Branches are coloured by the correlation of the branch pattern with the projections of the individuals onto the principal components. Tracing from blue to red through the tree shows how the leading principal components trace paths through the tree and correspond to major lineages.

each PC. One can see that the leading PCs trace paths through the deepest branches of the tree, validating using PCs to define lineages in this dataset.

After correcting for the number of lineages tested using a Bonferroni correction, we found PCs 1, 3, 4 and 12 to be significantly associated with wild bird vs chicken colonisation (Figure 5.9). Visualising the PCs on the phylogeny of the dataset in Figure 5.6 revealed that the most significant lineage, PC-1, represented the split of a wild bird lineage including the ST-1264 and ST-1304 complexes, from all other isolates. The ST-1264 and ST-1304 complexes have previously been shown to be host associated, ST-1264 was found only in Song Thrushes and ST-1304 has been found in Song Thrushes, European Blackbirds and Starlings (Waldenström et al. 2010; Griekspoor et al. 2013; Mohan et al. 2013).

Lineage PC-3 represented the split between the wild bird lineage ST-1287 complex previously identified in multiple wild bird species (Keller & Shriver 2014) and shown to be part of a wild bird only lineage (Griekspoor et al. 2013), and the chicken dominated lineage which contained ST complexes 21, 48, 257, 353, 354, 443, 574 and 607 amongst others. Lineage PC-4 represented the split of the wild bird lineage ST-1275 complex previously identified in wild birds (Keller & Shriver 2014), from the generalist ST-45 complex. Finally, lineage PC-12 represented the split between the chicken associated lineage containing ST-661 and ST-573 complexes previously shown to form a chicken associated lineage (Sheppard et al. 2014), and the phenotypically mixed lineage containing clonal complexes ST-1034 and ST-692 amongst others. The ST-1034 and ST-692 complexes have previously been shown to be isolated from both chickens and wild birds (Griekspoor et al. 2010). The test for lineage associations has revealed that both deep branches of the tree are host associated, such as the lineages represented by PCs 1, 3 and 4 but also slightly more recently evolved lineages such as that represented by PC-12

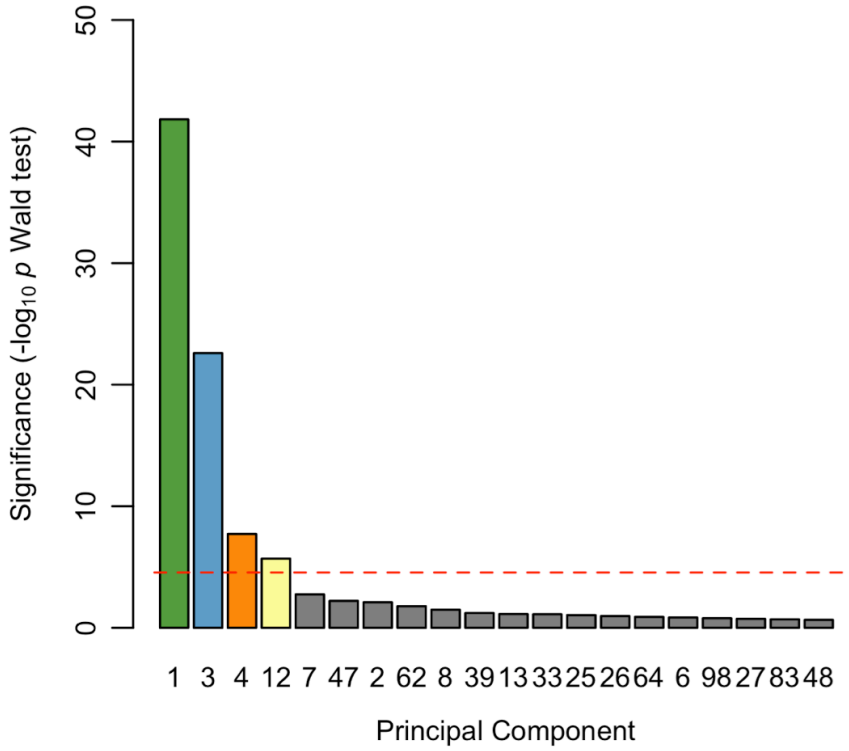

Figure 5.9 Results of the Wald test for lineage effects, testing for an association between PCs and the phenotype. PCs 1, 3, 4 and 12 were significantly associated with the phenotype. The red dotted line indicates the Bonferroni-corrected significance threshold.

which is chicken associated. Strong lineage associations have therefore resulted in high sample heritability and high host association prediction accuracy.

5.4.4 Thirteen SNPs and 1,164 kmers were significantly host associated

Here, we tested SNPs and kmers for association with wild bird versus chicken colonisation, performed a paired analysis to assess the independence of the signals of SNP association across loci and interpreted the locus results in light of the lineage associations identified in Section 5.4.3.2.

5.4.4.1 Identification of host associated variants

We conducted an association study to identify SNPs and kmers associated with wild bird or chicken hosts. Significance was assessed using Linear Mixed Models (LMM) using GEMMA (Zhou & Stephens 2012) to control for population structure. Bonferroni adjusted significance thresholds on the number of SNP phylopatterns and kmer

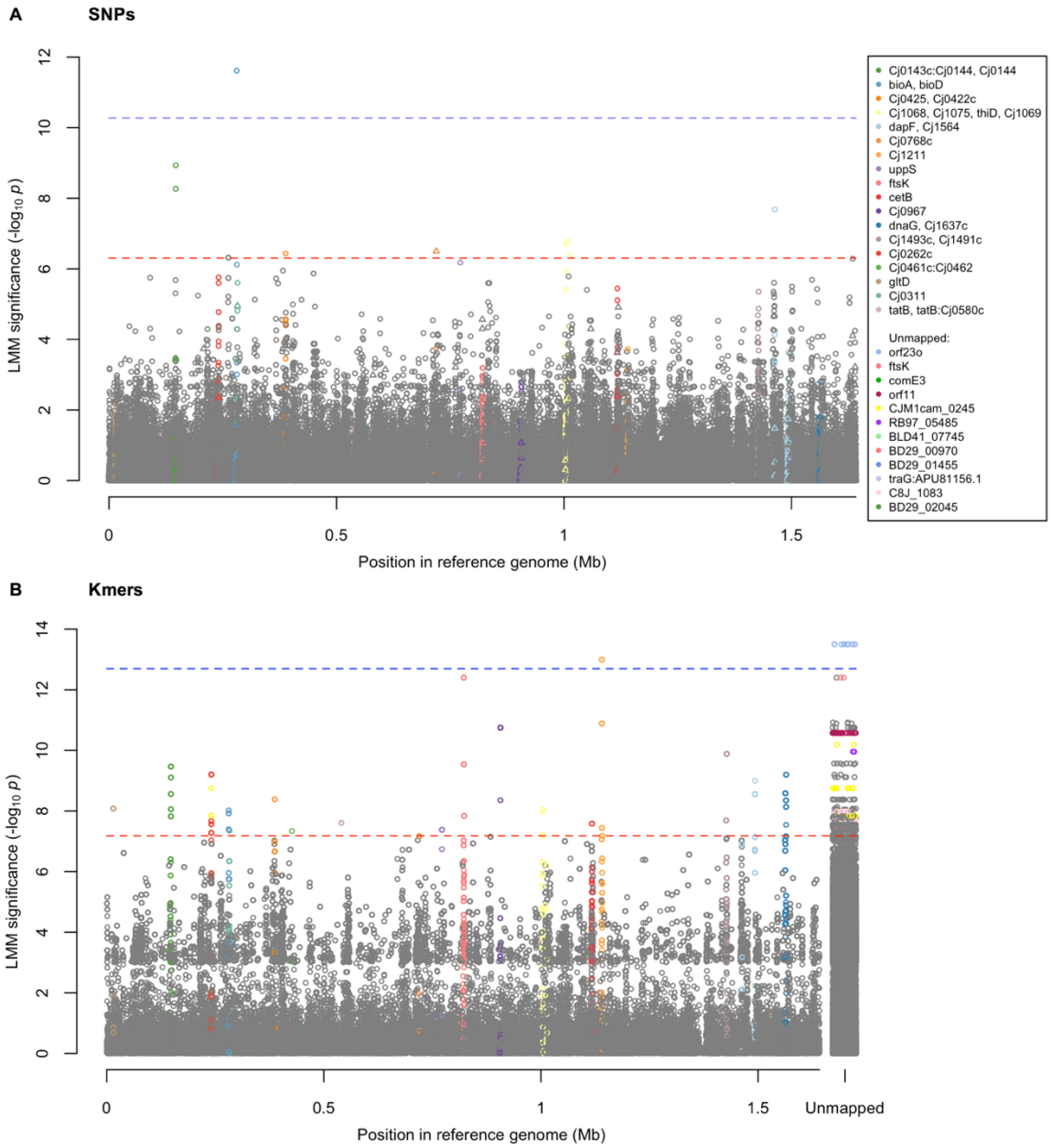

Figure 5.10 SNP and kmer significance after controlling for population structure using LMM. Only variants with a MAF≥1% are shown. SNP positions refer to their position in the reference genome FAM18. Kmer posiions refer to their position based on mapping to the same reference genome, plus some BLAST results to the reference genome. Genes and intergenic regions containing significant variants are coloured.

phylopatterns were applied to account for multiple testing for SNPs and kmers respectively. We found 13 SNPs and 1,164 kmers with minor allele frequencies (MAFs) of greater than 1% to be significantly associated with wild bird versus chicken colonisation (Figure 5.10). Significant SNPs are listed in Table 5.2 and genes containing significant kmers which map to the reference genome NCTC11168 are listed in Table 5.3.

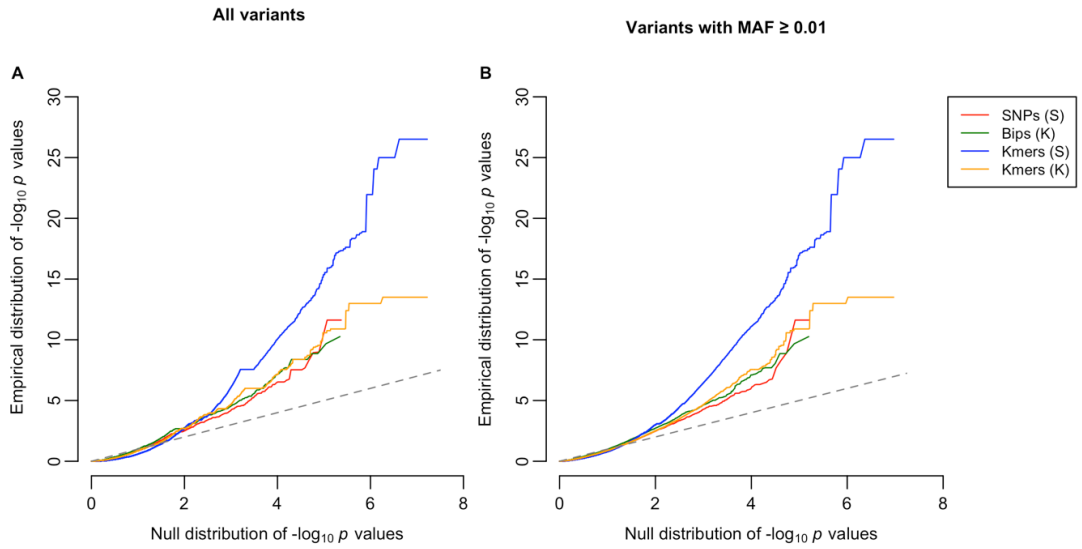

Figure 5.11 QQ plots showing the expected versus the empirical distribution of *P* values for all analyses. In brackets, the letters symbolise the type of genetic variation used to calculate the kinship matrix used to control for population structure; S = biallelic SNPs, K = variant kmers counted from Velvet assemblies. All studies show a signature of overinflation of *P* values with respect to the null distribution, but testing the kmers for association using the kmers to build the kinship matrix greatly reduces the overinflation in comparison to using the kinship matrix built from SNPs, and it then follows a similar path to that of the SNPs.

QQ plots of the null versus expected distributions of *P* values for SNPs and kmers are shown in Figure 5.11. This revealed that using kmers to build the kinship matrix for the kmer analysis greatly reduced the inflation of the *P* values against the null. As LMM can increase the significance of low frequency variants and insufficiently control for population structure for low frequency variants (Mathieson & McVean 2012), the frequency of all significant variants was assessed. The SNPs in *bioA* and *Cj1075* had a MAF of less than 5%, the SNPs in *Cj1068* and *thiD* had MAFs of less than 10% and the remainder had a MAF of greater than 10% (Table 5.2).

Of the 13 significant SNPs, some SNPs were heavily imputed (Table 5.2). 47% of isolates were imputed at the SNPs in *bioA*, 45% and 40% at the SNPs in the intergenic region between *Cj0143c* and *Cj0144* and 53% at the SNP in *Cj0425*. All other significant variants were imputed in 3% of isolates or less. The variants in the highly imputed regions were not found to be significant in the kmer analysis which assesses the true presence and absence of particular regions. There were however significant kmers in the

Position	A	C	G	T	Imputed %	-log$_{10}$ *P* value	OR	Type	Name	Product
280550	0	7 (4)	0	473 (250)	47.1	11.6	∞	S	*bioA*	adenosylmethionine--8-amino-7-oxononanoate aminotransferase BioA
280551	473 (249)	0	7 (4)	0	47.3	11.6	∞	NS	*bioA*	adenosylmethionine--8-amino-7-oxononanoate aminotransferase BioA
146630	318 (157)	0	162 (105)	0	45.4	8.9	83	I	*Cj0143c*: *Cj0144*	ABC transporter substrate-binding protein:methyl-accepting chemotaxis signal transduction protein
146577	208 (102)	0	272 (188)	0	39.6	8.3	0.21	I	*Cj0143c*: *Cj0144*	ABC transporter substrate-binding protein:methyl-accepting chemotaxis signal transduction protein
1463272	0	54 (52)	0	426 (425)	0.6	7.7	1.18	S	*dapF*	diaminopimelate epimerase
1009720	15 (15)	0	465 (459)	0	1.3	6.8	0.4	NS	*Cj1075*	flagellar assembly protein FliW
1003311	452 (451)	0	28 (27)	0	0.4	6.7	0.86	NS	*Cj1068*	zinc metalloprotease
719342	13 (13)	449 (438)	0	18 (17)	2.5	6.5		S/NS/NS	*Cj0768c*	3-polyprenyl-4-hydroxybenzoate decarboxylase
388255	0	265 (125)	0	215 (104)	52.3	6.4	0.004	S	*Cj0425*	periplasmic protein
1013501	0	437 (427)	0	43 (41)	2.5	6.4	0.7	NS	*thiD*	phosphomethylpyrimidine kinase
262332	0	389 (388)	0	91 (89)	0.6	6.3	0.017	S	*cheA*	Chemotaxis histidine kinase
262350	389 (388)	0	91 (90)	0	0.4	6.3	0.017	S	*cheA*	Chemotaxis histidine kinase
262358	389 (388)	0	91 (90)	0	0.4	6.3	0.017	S	*cheA*	Chemotaxis histidine kinase

Table 5.2 Significant SNP results after controlling for population structure using LMM. OR, odds ratio. SNP Type S, Synonymous; NS, Non-synonymous; I, Intergenic. Numbers per base represent base calls post-imputation first, and pre-imputation in brackets.

Name	# kmers	-log$_{10}$ *P* value	OR	MAF	Product
Cj1211	121	7.4-13.0	4.9, 2.9, 3.8	15.4	competence family protein
ftsK	42	7.8-12.4	4.5, 3.5, 4.8	13.8	DNA translocase FtsK
Cj0967	16	8.4-10.8	0, 0.46	1.5	periplasmic protein
Cj1493c	16	9.9	4.5	14.0	integral membrane protein
Cj0144/ *Cj0262c*/ *Cj1564*	47	7.8-9.5	122.9, ∞, 474.5, 119.6, ∞	30.0	methyl-accepting chemotaxis signal transduction proteins
Cj0262c	45	7.3-9.2	0.07, 0.1, 0.05, 0.04, 0.09, 0.06	7.7	methyl-accepting chemotaxis signal transduction protein
dnaG	26	7.5-9.2	3.06, 2.8, 2.7, 2.9	16.3	DNA primase
Cj1564	24	8.6-9.0	∞, ∞	33.0	methyl-accepting chemotaxis signal transduction protein
Cj1637c	20	8.6	3.2	15.8	periplasmic protein
Cj0422c	1	8.4	1.02	21.9	H-T-H containing protein
gltD	9	8.1	∞	2.3	glutamate synthase subunit beta
Cj1069	15	8.1	0.9	5.8	hypothetical protein
bioD	41	7.4-8.02	0.2, 0.09, 0.1	2.5	dethiobiotin synthetase
Cj1068	4	7.2-8.0	0.4, 0.8	6.3	zinc metalloprotease
Cj1491c	9	7.7	3.4	14.4	two-component regulator
tatB:Cj0580c	6	7.6	3.5	14.8	Sec-independent translocase TatB: coproporphyrinogen III oxidase
cetB	10	7.6	90.5	36.0	bipartate energy taxis response protein CetB
uppS	28	7.4	0.1	16.0	UDP diphosphate synthase
Cj0311	7	7.4	0.2	2.7	50S ribosomal protein L25/general stress protein Ctc

Table 5.3 Significant kmer results that map to the reference genome NCTC11168 after controlling for population structure using LMM. OR, odds ratio; MAF, minor allele frequency.

gene *Cj0144* (Figure 5.10).

We assessed the localised regions of the signals to look for signatures of local LD surrounding the SNPs and kmers to add support to the associations (Figure 5.12, described in Chapter 1.1.3.1) (The Wellcome Trust Case Control Consortium 2007). The synonymous SNP in *dapF* displayed the signature of inflated significance surrounding the SNP and decay in significance with increasing distance from the SNP (Figure 5.12C). Multiple other synonymous SNPs upstream of *dapF* were also just below the significance threshold. The signal in the *dapF* region was replicated in the kmer analysis, however due

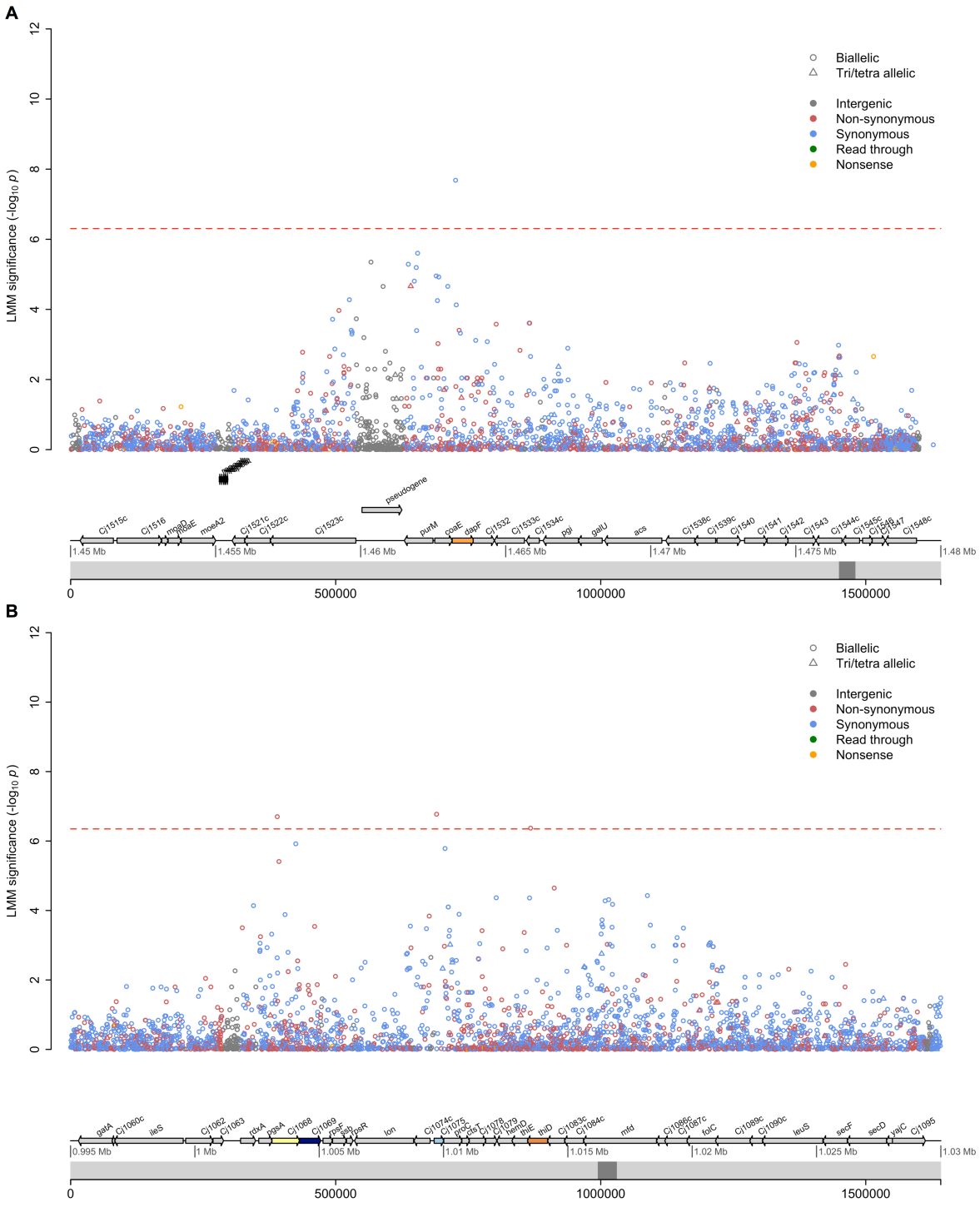

Figure 5.12 Close ups of the SNP manhattan plot focusing on the significant SNPs. SNPs are coloured: grey = intergenic; red = non-synonymous; blue = synonymous; green = read-through; yellow = nonsense. The coloured region of the bottom bars depict the region of the reference genome that is being shown. **A** *dapF*, *Cj1521c = cas2*; *Cj1522c = cas1*; *Cj1523c = cas9*; **B** *Cj1068*, *Cj1075* and *thiD*; **C** *Cj0768c*; **D** *cheA*. (*Continued on the next page*).

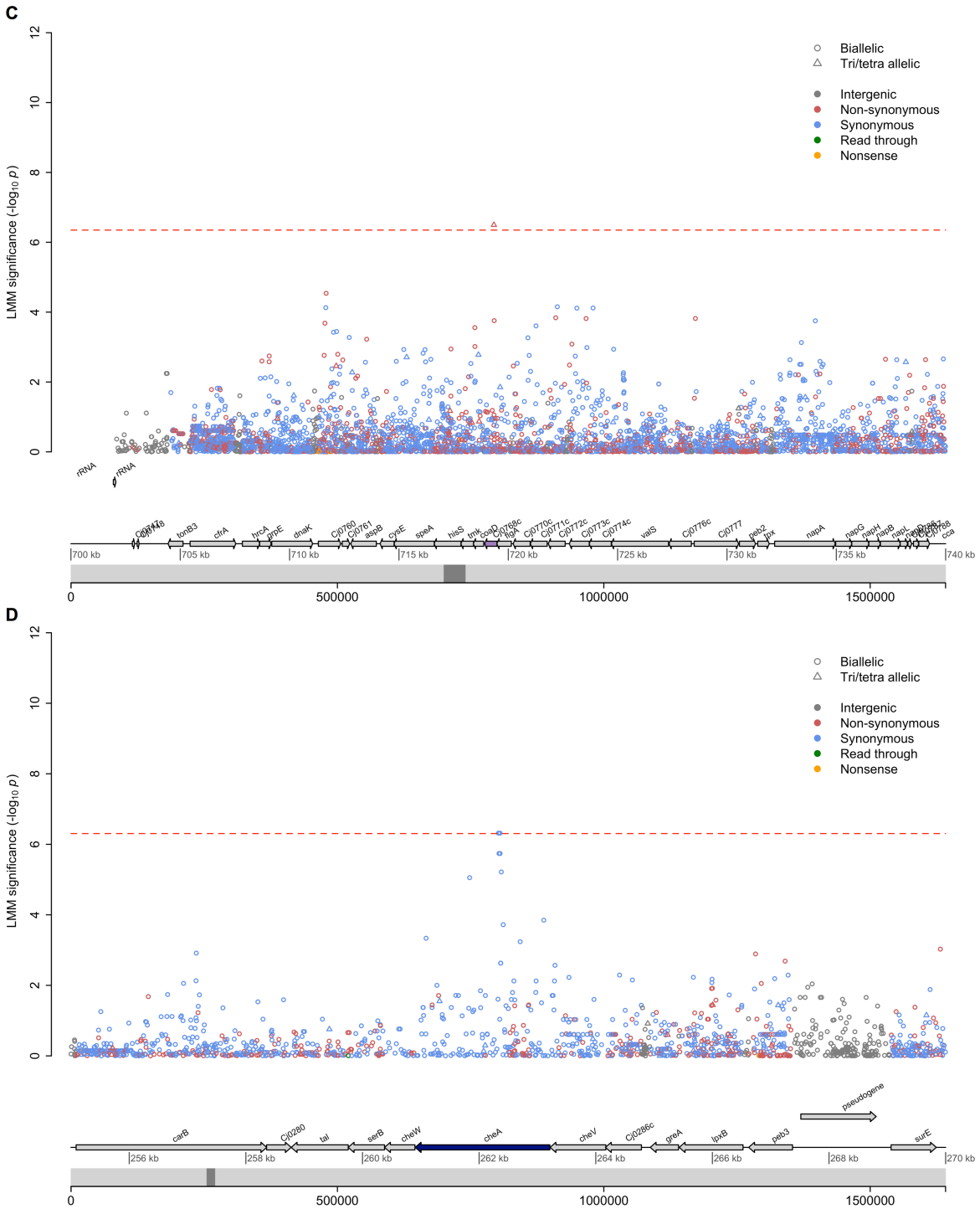

Figure 5.12 (contd.) Close ups of the SNP manhattan plot focusing on the significant SNPs. SNPs are coloured: grey = intergenic; red = non-synonymous; blue = synonymous; green = read-through; yellow = nonsense. The coloured region of the bottom bars depict the region of the reference genome that is being shown. **A** *dapF*; **B** *Cj1068*, *Cj1075* and *thiD*; **C** *Cj0768c*; **D** *cheA*.

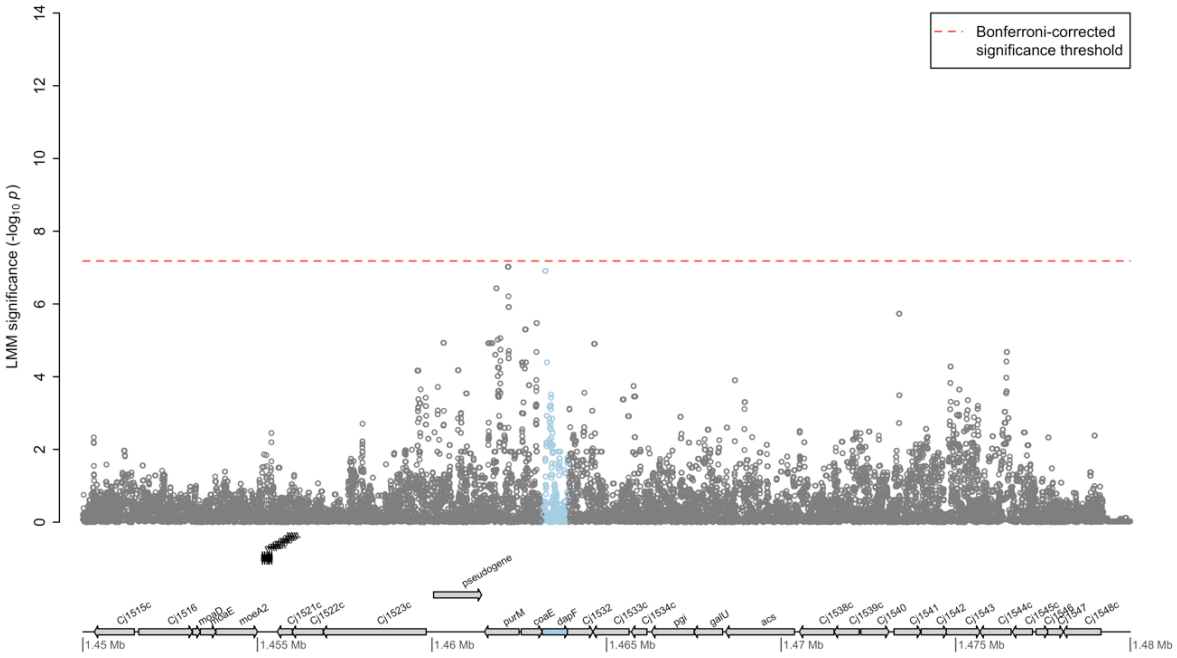

Figure 5.13 Close up of the kmers mapped to NCTC11168 focusing on the region containing *dapF*. The association in *dapF* was replicated in the kmer analysis, however due to the larger number of kmers tested and therefore the more stringest multiple testing correction, the variant was not significantly associated. All variants in *dapF* are coloured.

to the number of kmers tested and therefore a more stringent multiple testing correction threshold, the variants fell just below genome-wide significance (Figure 5.13). We did however see the signature of inflated significance in this region, and a decay of significance with increasing distance from the *purM* to *dapF* region (Figure 5.13).

We calculated approximate posterior probabilities for the biallelic SNPs, assuming one causal SNP, and subsequently the highest posterior probability threshold such that the sum of the posterior probabilities exceeding the threshold was ≥95%. Just two SNPs fell above this threshold, the two significant SNPs in *bioA*. However, these were low frequency SNPs, with a MAF of 1.5%, and they were imputed in 47% of isolates so could not explain a great proportion of the phenotypic variability.

With the exception of the two variants in the intergenic region between *Cj0143c*:*Cj0144*, when multiple variants were found within the same gene, the variants were in perfect LD and had the same phylopattern. All other variants had unique

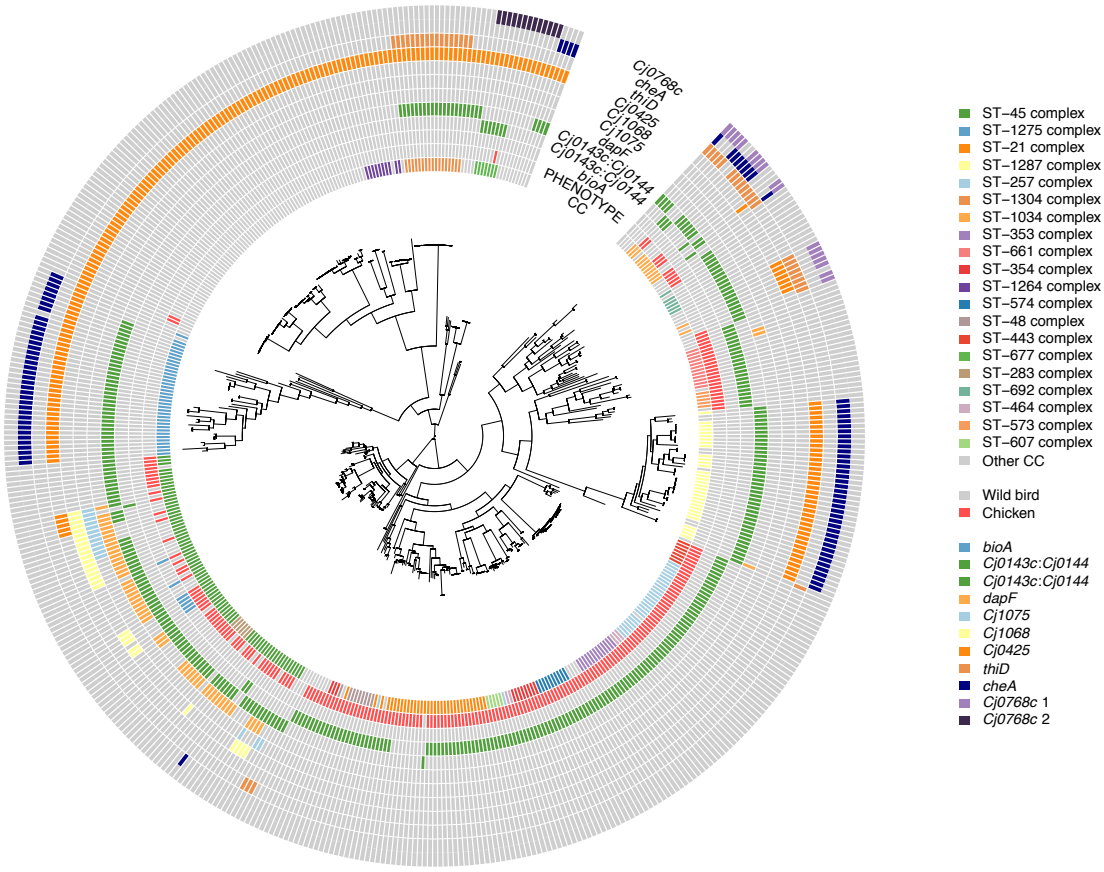

Figure 5.14 Phylogenetic distribution of the significant SNPs. The minor alleles of all significant SNPs are shown coloured on the phylogeny. The SNP in *dapF* was mostly present within the ST-45 complex but was also present in two isolates in the ST-661 complex, 6 in the ST-283 complex and one ST-1285 isolate with no clonal complex assignment.

phylopatterns, as visualised on the phylogeny of the dataset in Figure 5.14. For example, the minor allele of the significant SNP in *dapF*, although present at low frequencies in other regions of the phylogeny, was mostly present within the ST-45 complex where it was more prevalent within chicken isolates (Figure 5.14).

In the human setting, it has been shown that including PCs as additional fixed effects in the LMM can improve power and reduce false positives (Tucker, Price & Berger 2014; Widmer et al. 2014). So to ensure that population structure had been sufficiently controlled for and to investigate the robustness of the results, SNPs were tested for association, including PCs as additional fixed effects in the LMM. Results were compared with the original LMM results to test whether the SNPs retained significance.

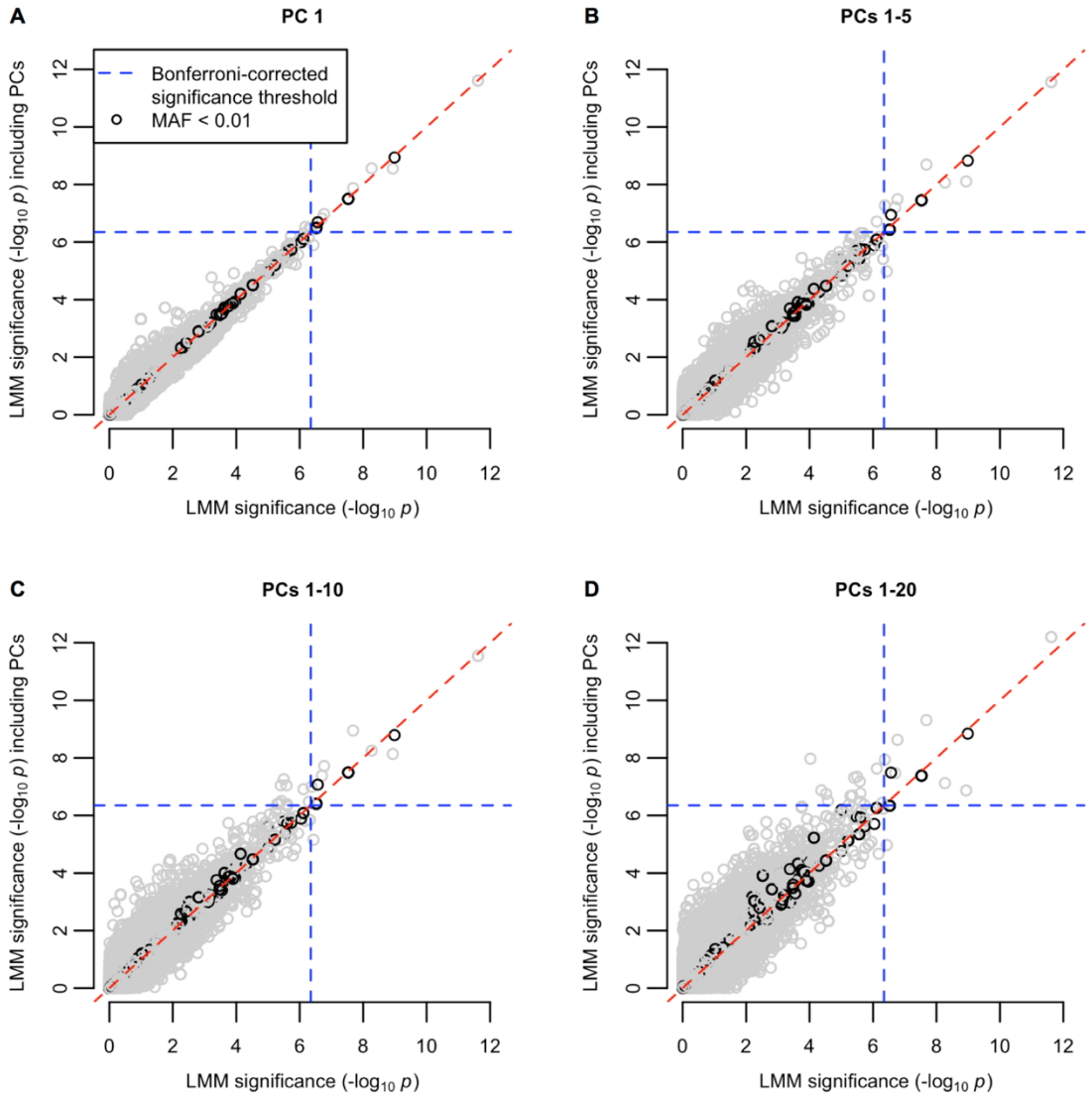

Figure 5.15 Biallelic SNP significance by LMM with and without including Principal Components as additional fixed effects. A PC 1 **B** PCs 1-5 **C** PCs 1-10 **D** PCs 1-20. Black circles represent SNPs with MAF < 0.01, grey circles represent SNPs with MAF ≥ 0.01. The SNP in *Cj0425* lost significance in all tests and the SNPs in *cheA* lost significance with the inclusion of 5 or more PCs but all other SNPs remained significant.

PCs 1, 1-5, 1-10 and 1-20 were all separately included. All originally significant SNPs remained significant in all tests except for the SNP in *Cj0425* and the SNPs in *cheA* (Figure 5.15). *Cj0425* lost significance in all tests and was not further analysed and the SNPs in *cheA* lost significance with the inclusion of five or more PCs (Figure 5.17B). The inclusion of PCs as fixed effects also did not significantly change the empirical distribution of *P* values against the null distribution, therefore all results which remained

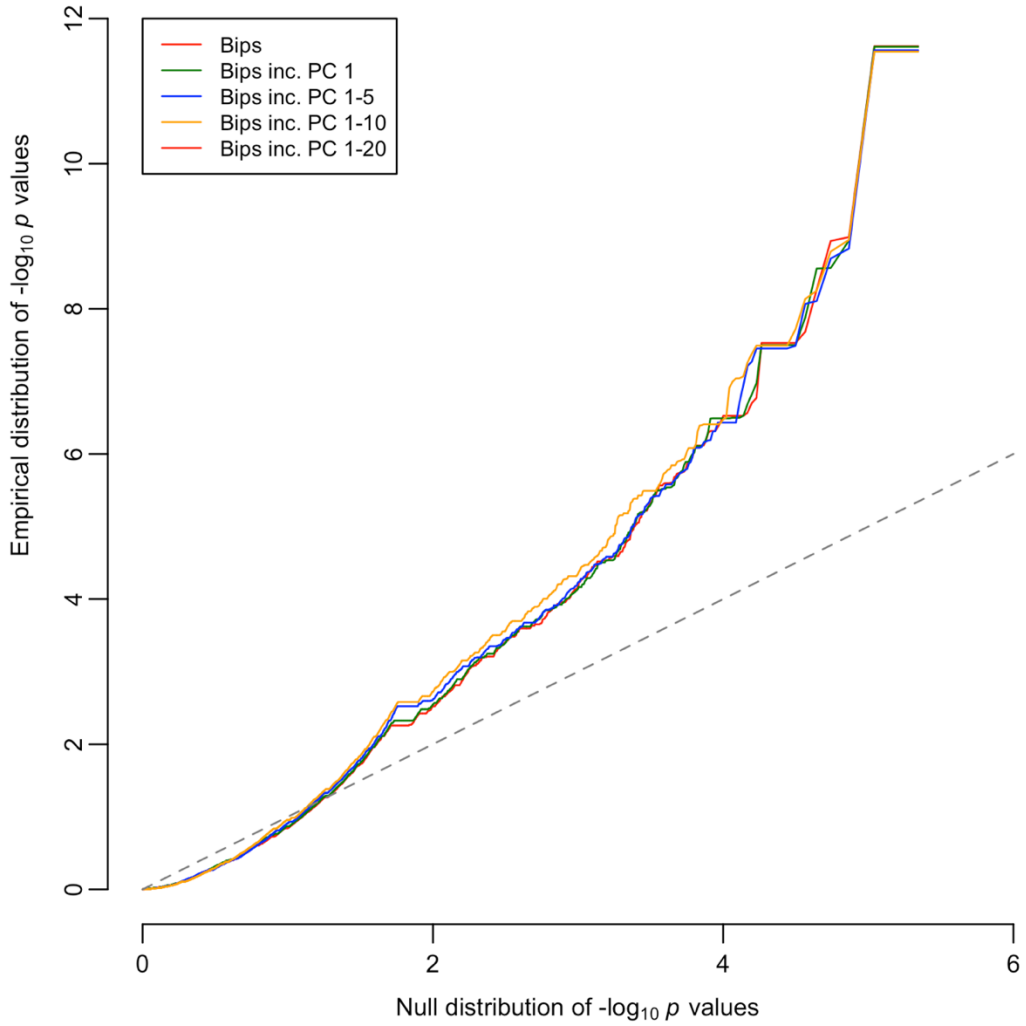

Figure 5.16 QQ plot showing the expected versus the empirical distribution of biallelic SNPs with and without including PCs as additional fixed effects. Including PCs as fixed effects does not significantly change the distribution suggesting that the original control for population structure is sufficient.

significant were analysed (Figure 5.16). The inclusion of the first 20 PCs resulted in an increase in significance in the region of the significant SNP in *dapF* (Figure 5.17A). SNPs in *purM* and in the intergenic region between *Cj1523c* and *purM* became significant giving confidence in the association signal in this region, as a more stringent control for population structure resulted in an increase of significance (Figure 5.17).

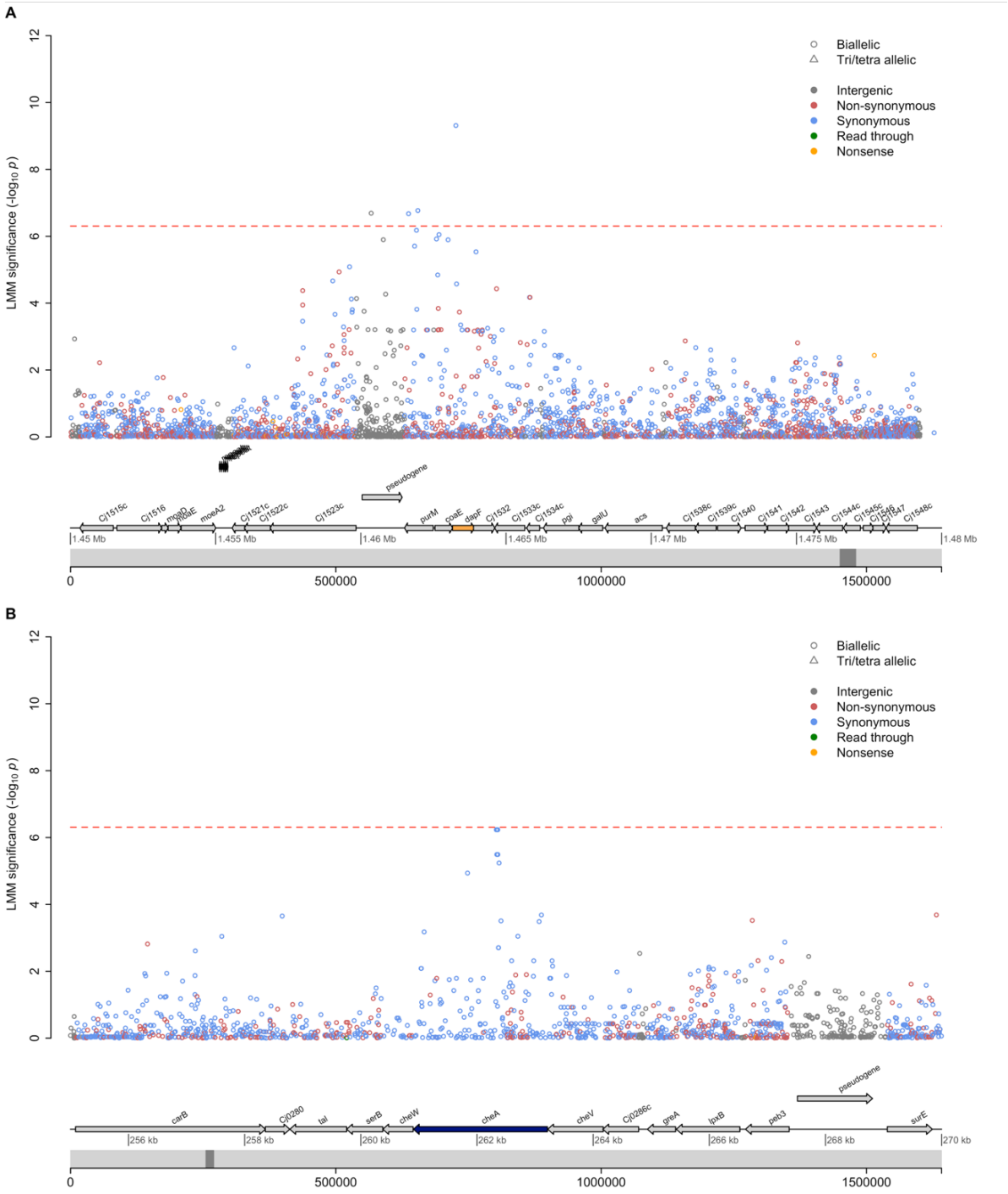

Figure 5.17 Close ups of the SNP Manhattan plot after including PCs 1-20 as additional fixed effects, focusing on the regions containing A *dapF* and B *cheA*. SNPs are coloured: grey = intergenic; red = non-synonymous; blue = synonymous; green = read-through; yellow = nonsense. The dark grey of the bottom bars depict the regions of the reference genome displayed. Including PCs 1-20 as additional fixed effects resulted in SNPs in *purM* and in the intergenic region/pseudogene between *Cj1523c* and *purM* becoming significant but also caused the SNPs in *cheA* to drop just below significance at $-\log_{10}p = 6.23$.

5.4.4.2 Kmer associations were robust to the inclusion of read length as a fixed effect

As the isolates were not all sequenced together and therefore had different read lengths, measures were taken to account for possible batch effects. The mode read length was estimated as discussed in Section 5.3.4 and these were coded as binary variables and included as additional fixed effects in the kmer analysis. Figure 5.18A shows the $-\log_{10} P$ values of the kmers from LMM analyses with and without including read length as a fixed effect. This shows that the significance of the top results changed by a significant amount for some of the kmers, some kmers lost significance and some also gained significance. However, viewing the results as a Manhattan plot revealed that the same genes were significant as when read length was not included as a fixed effect (Figure 5.18B). We believe that using kmers to build the kinship matrix should account for weak batch effects within the data, as they would in effect be an axis of variation through the data that could be controlled for. The inclusion of read length as a fixed effect would therefore be an even stricter control, but it is not clear which is the correct way to analyse the data. It was reassuring however that it was the same genes which were significant both with and without the inclusion of read length as fixed effects.

5.4.4.3 Many of the significant kmers contained homopolymers

We also investigated the presence of possible batch effects by identifying homopolymeric sequences within the kmers. Insertions and deletions have been shown to occur in up to 2% of bases within homopolymers by Illumina HiSeq data (Minoche, Dohm & Himmelbauer 2011). These errors could have been introduced non-randomly through the dataset due to the differing sequencing batches predicted by mode read length (Figure 5.5). The longest homopolymer was counted for each kmer, and the Manhattan plot was coloured by the longest homopolymer per kmer (Figure 5.19). This revealed that the most

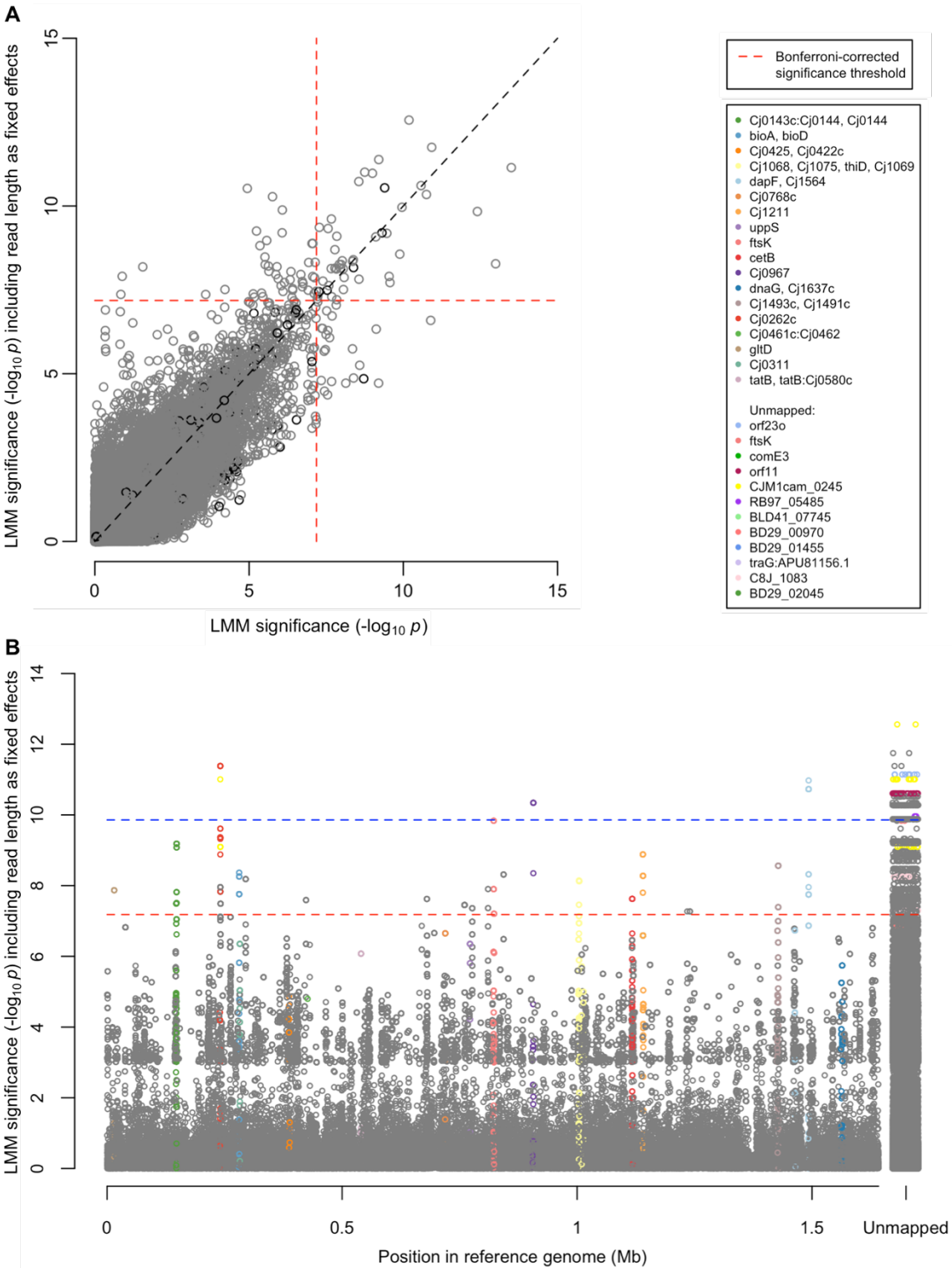

Figure 5.18 The effect of including read length as a fixed effect in the kmer LMM analysis. **A** Kmer LMM with and without including read length as fixed effects **B** Manhattan plot of the kmer LMM against NCTC11168 including read length as fixed effects. Although many results changed in significance, the same genes were significant after the extra correction for read length.

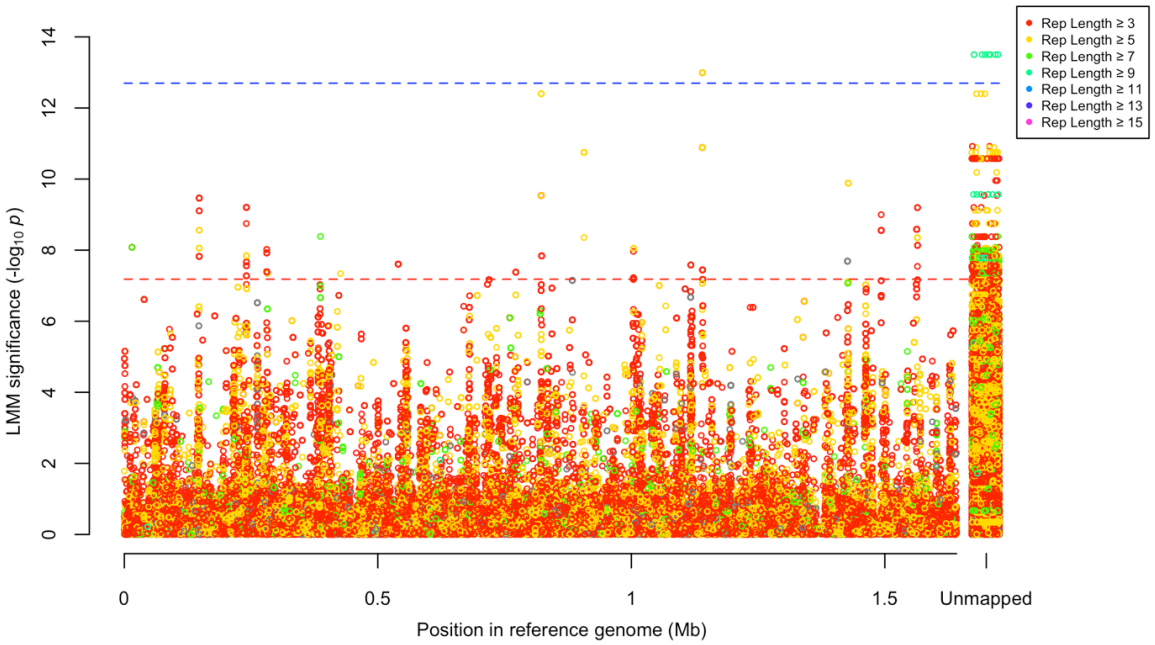

Figure 5.19 Kmer results mapped to NCTC11168 coloured by the length of the longest homopolymeric tract in the kmer. The most significant kmers contained at least one homopolymer of between 9 and 11 bases long.

significant kmers, which map to LOS *orf23o* (shown in green in the unmapped region in Figure 5.19) each contained at least one homopolymer of between 9 and 11 bases. Although these associations may be real, rather than batch effect artefacts, they should be treated with caution in the absence of experimentally confirming their existence.

5.4.4.4 A paired SNP analysis indicated that the significant SNP associations across loci were not all independent signals

In order to determine whether the significant SNPs represented independent signals we conducted a paired analysis. All biallelic SNPs were tested using LMM including each of the unique significant SNP phylopatterns as an additional fixed effect individually, in addition to their inclusion as a random effect (Figure 5.20).

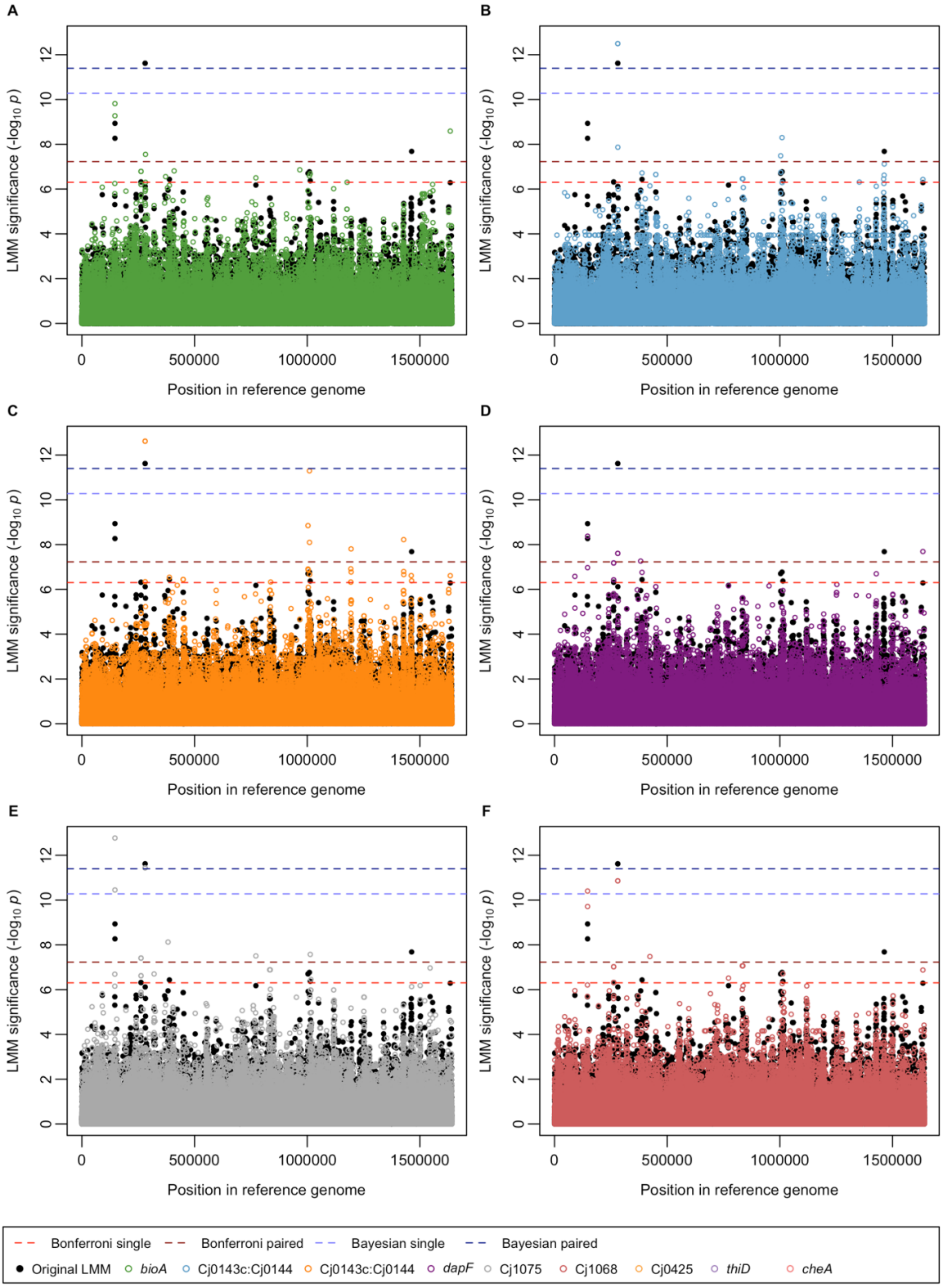

Figure 5.20 Results of the paired SNP analysis. The SNP included as an additional fixed effect was **A** *bioA*; **B** *Cj0143c*:*Cj0144*; **C** *Cj0143c*:*Cj0144*; **D** *dapF*; **E** *Cj1075*; **F** *Cj1068*; **G** *Cj0425*; **H** *thiD*; **I** *cheA*. Black dots depict the original LMM results. The coloured dots depict the results of including that SNP pattern as an additional fixed effect in the LMM. The original and new Bonferroni-corrected significance thresholds are shown along with the highest Bayesian posterior probability thresholds. (*Continued on the next page*).

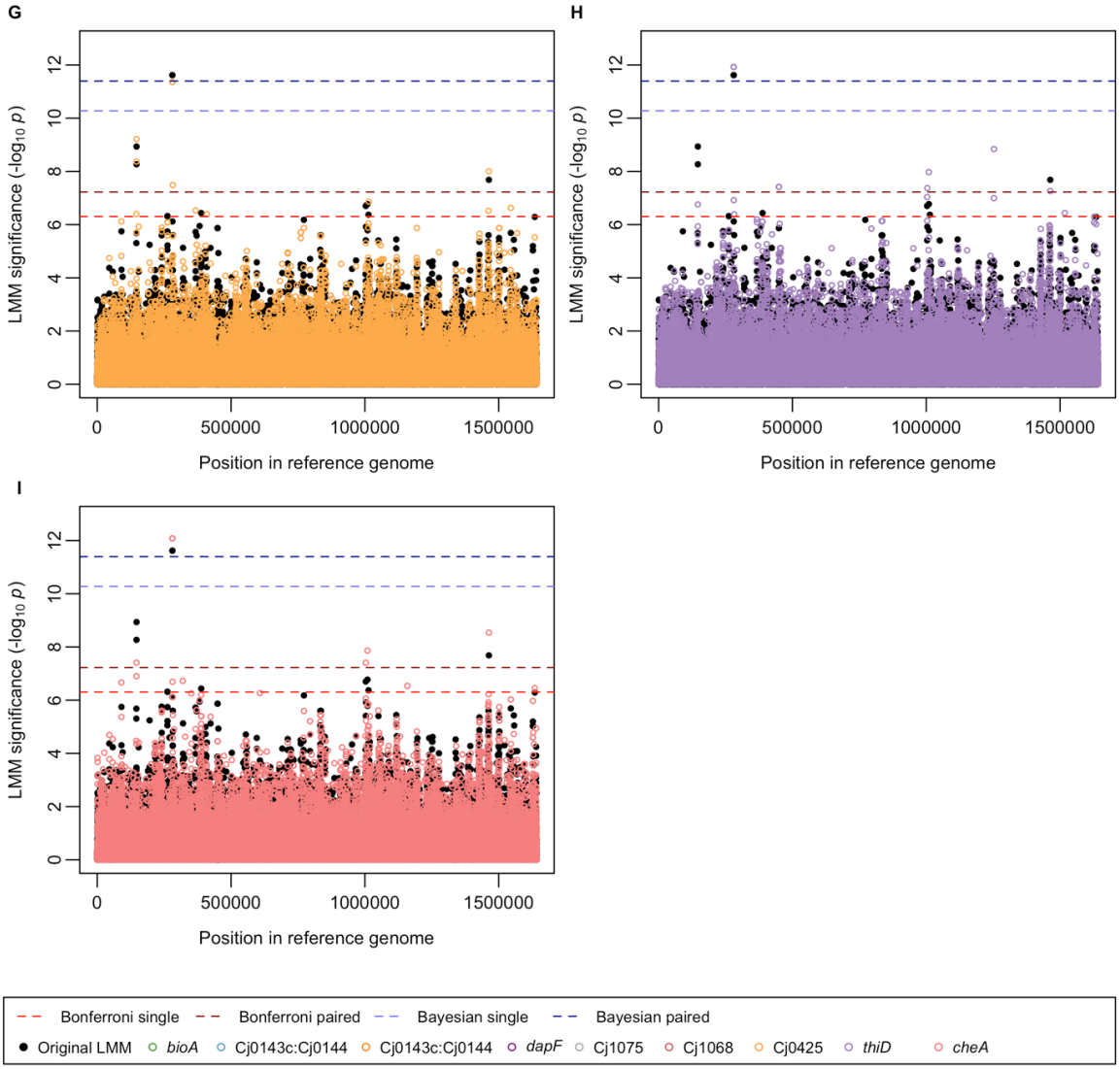

Figure 5.20 (Contd.) Results of the paired SNP analysis. The SNP included as an additional fixed effect was **A** *bioA*; **B** *Cj0143c*:*Cj0144*; **C** *Cj0143c*:*Cj0144*; **D** *dapF*; **E** *Cj1075*; **F** *Cj1068*; **G** *Cj0425*; **H** *thiD*; **I** *cheA*. Black dots depict the original LMM results. The coloured dots depict the results of including that SNP pattern as an additional fixed effect in the LMM. The original and new Bonferroni-corrected significance thresholds are shown along with the highest Bayesian posterior probability thresholds.

A new highest posterior probability threshold was calculated on the assumption that a single SNP pair was causally associated, which contained 11 candidate SNP pairs, consisting of SNPs in:

- The intergenic region between *Cj0143c* and *Cj0144* plus *Cj1075*
- *bioA* plus *Cj0143c*:*Cj0144*, *Cj1075*, *thiD* and *cheA*

A new Bonferroni-corrected significance threshold was also calculated to account for the number of pairs tested. Table 5.4 shows which SNPs retained significance within the

Gene	Original significance ($-\log_{10}p$)	Genes containing SNPs with signals independent to SNPs in the gene in column one
bioA	11.6	*Cj0143c*:*Cj0144*, *dapF*, *Cj1075*, *Cj1068*, *Cj0425*, *thiD*, *cheA*
Cj0143c:*Cj0144*	8.9	*bioA*, *dapF*, *Cj1075*, *Cj1068*, *Cj0425*, *cheA*
Cj0143c:*Cj0144*	8.3	*bioA*, *Cj1075*, *Cj1068*, *Cj0425*
dapF	7.7	*Cj0425*, *thiD*, *cheA*
Cj1075	6.8	*Cj0143c*:*Cj0144*(both), *thiD*, *cheA*
Cj1068	6.7	*Cj0143c*:*Cj0144*(both), *cheA*
Cj0425	6.4	-
thiD	6.4	*Cj1075*
cheA	6.3	*Cj1075*

Table 5.4 Results of the paired SNP analysis. For each gene in column one, column three lists the genes containing SNPs with independent signals.

inclusion of another SNP as a fixed effect.

This indicated that the SNP in *Cj0425* was not an independent signal from the other significant SNPs, as the inclusion of any other significant SNP phylopattern as an additional fixed effect resulted in loss of significance of the SNP. Most SNPs lost significance by the inclusion of the significant SNP in *dapF*, indicating that they were not independent signals. Given the region of inflated significance surrounding *dapF* (Figure 5.12C) this appears to be a promising association.

5.4.4.5 The majority of the significant variants were not associated with the significant lineage effects

We reassessed the locus effects in light of the lineage effects presented in Section 5.4.3.2 by assigning variants to lineages (Figure 5.21). This was achieved by taking the variant pattern across the isolates and determining the correlation of the variant pattern with the projections of the individuals onto each principal component. Variants were assigned to the lineage (PC) to which its absolute correlation was highest. This revealed that most SNP locus effects were not associated with any of the significant lineage effects. The exceptions to this were the SNPs in the intergenic region between *Cj0143c* and *Cj0144* associated with lineage PC-3 and the SNP in the gene *Cj0425*. The first of the SNPs in

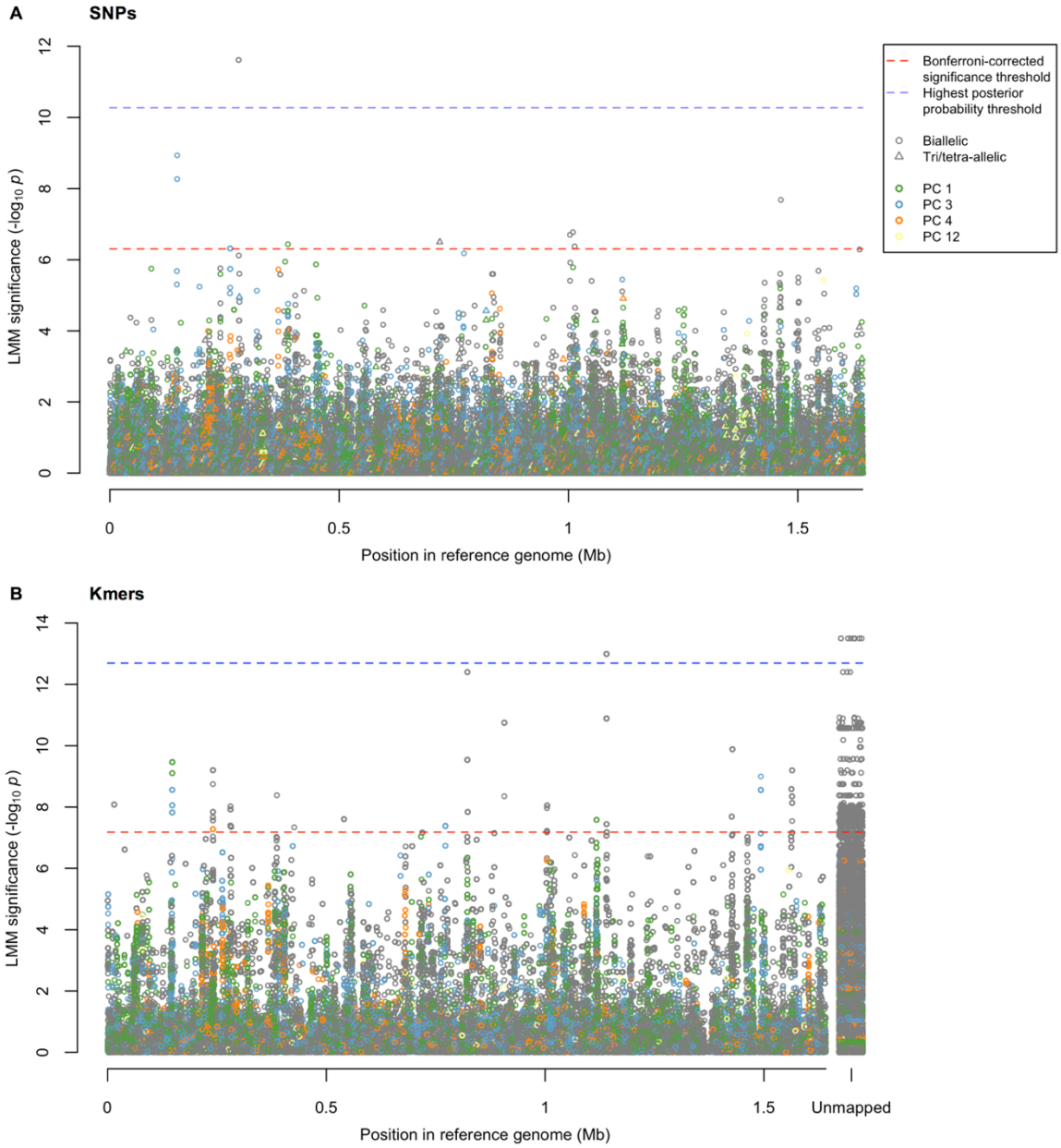

Figure 5.21 Interpreting the locus effects in light of the lineage effects for A SNPs and B kmers. SNP and kmer results are shown against reference genome NCTC11168 after controlling for population structure using LMM. Variants are coloured by the lineage effects, and are coloured by PC-1, PC-3, PC-4 or PC-12 if the variant pattern was most correlated with one of the four PCs. This revealed that most of the significant locus effects were not most correlated with any of the significant lineages.

the *Cj0143c*:*Cj0144* intergenic region was chicken associated, with an odds ratio (OR) of 83, and the other was wild-bird associated with an OR of 0.21 (Figure 5.14). The SNP in *Cj0425* was associated with lineage PC-1 and was wild-bird associated with an OR of 0.004. However, this SNP lost significance when including PCs as fixed effects, and 53% of the isolates were imputed at this site, so it has not been further analysed. By viewing

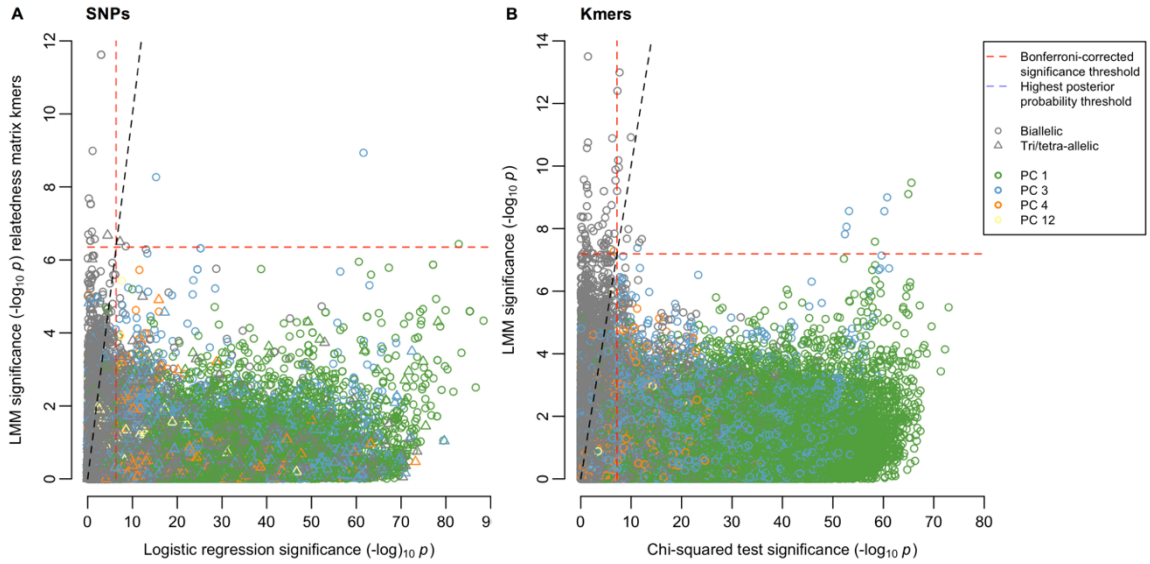

Figure 5.22 Interpreting the locus effects of the SNPs and kmers in light of the lineage effects - before and after controlling for population structure. A $-\log_{10} P$ values of the uncorrected logistic regression for SNPs against the $-\log_{10} P$ values after controlling for population structure using LMM. **B** $-\log_{10} P$ values of the uncorrected χ^2 test for kmers against the $-\log_{10} P$ values after controlling for population structure using LMM. For both analyses, variants most correlated to the significant lineage effects were greatly reduced in significance after controlling for population structure.

the effect of controlling for population structure using LMM on SNPs and kmers in comparison to logistic regression and χ^2 tests, respectively, we can see the differential effect on variants which are most correlated with the significant lineages versus those that are not. Variants most correlated with the significant lineage effect suffer loss of significance much more strongly than unstratified variants, as expected (Figure 5.22).

5.4.5 Identifying possible roles of significant host associated variants

5.4.5.1 A SNP variant downstream of cas1, cas2, cas9 and the CRISPR repeats was associated with chicken colonisation

The SNP GWAS revealed a synonymous biallelic SNP in the gene *dapF* to be significantly associated with wild bird vs chicken colonisation (Figure 5.12A). *dapF* is a diaminopimelate epimerase and is part of one of two lysine biosynthesis pathways (Velasco, Leguina & Lazcano 2002). *dapF* was previously classified as an essential gene under laboratory conditions by an *in silico* flux balance analysis aimed at classifying key

metabolic routes essential for energy and biomass generation and by a transposon mutagenesis method (Metris et al. 2011; Stahl & Stintzi 2011). It has also been shown to be essential in both *E. coli* and *B. subtilis* (Ashikaga et al. 2003; Gerdes et al. 2003). The region of the SNP association displayed the expected signature of an association of inflated significance decreasing with increasing distance from the significant variant (Figure 5.12A; The Wellcome Trust Case Control Consortium 2007). The centre of the region of elevated significance appeared to be not in *dapF* but in the gene *purM*, Inclusion of the first 20 PCs as additional fixed effects in the LMM lead to two SNPs in *purM* plus a SNP in the adjacent pseudogene *Cj1528* becoming significant (Figure 5.17A). *purM* encodes a phosphoribosylformylglycinamidine cyclo-ligase and is part of the purine biosynthesis pathway (Kappock, Ealick & Stubbe 2000). *purM* was also found to be essential by *in silico* flux balance analysis but not by transposon mutagenesis (Metris et al. 2011).

dapF and *purM* are also within close proximity of the CRISPR repeats, along with the CRISPR-associated genes *cas2*, *cas1*, and *cas9*. CRISPR (clustered regularly interspaced short palindromic repeats) is thought to be an adaptive microbial immune system, enabling immunity against plasmids and viruses (Horvath & Barrangou 2010). CRISPR arrays contain repeat sequences separated by spacer sequences, captured foreign DNA elements, and are typically found adjacent to *cas* genes (Barrangou et al. 2007). The spacer sequences recognise DNA targets and the Cas proteins enable their degradation (Hale et al. 2009; Gasiunas et al. 2012; Jinek et al. 2012). CRISPR-Cas variants have been classified into three main types, and CRISPR-containing *C. jejuni* strains are classified as type II (Dugar et al. 2013).

The significant SNP in *dapF* at the third base position in codon 39 was a synonymous SNP. The most significant kmers which mapped to the gene also covered the

SNP and are shown aligned to the SNP data in Figure 5.23A, however they fell just below genome-wide significance due to the more stringent multiple testing correction for kmers. *C. jejuni* is an AT-rich genome (Parkhill et al. 2000), and Fuglsang (2003) analysed the coding sequences of *C. jejuni* NCTC11168 (accession NC002163) to calculate relative synonymous codon usage values for each codon of each amino acid. Fuglsang found that codon usage is biased for all amino acids with synonymous codons, as AT-rich codons are typically used. The major allele at the significant SNP in *dapF* creates the codon GGT and the minor allele GGC, which both encode for Glycine. However, Fuglsang demonstrated that Glycine codons show a positive bias towards GGT and a negative bias towards GGC (Fuglsang 2003). Therefore, the alternative allele at the SNP in *dapF* creates a less common codon, so we hypothesise that it may reduce expression of *dapF*. This could also mean that the variants have been recently imported by horizontal gene transfer from another organism with a different codon usage bias. This is unlikely to be due to recombination with *C. coli* however as their codon usage and base composition are very similar (The Codon Usage Database http://www.kazusa.or.jp/codon/). Therefore, a possible link if any with the CRISPR system is not clear, and the role the variant might have in host adaptation remains to be understood.

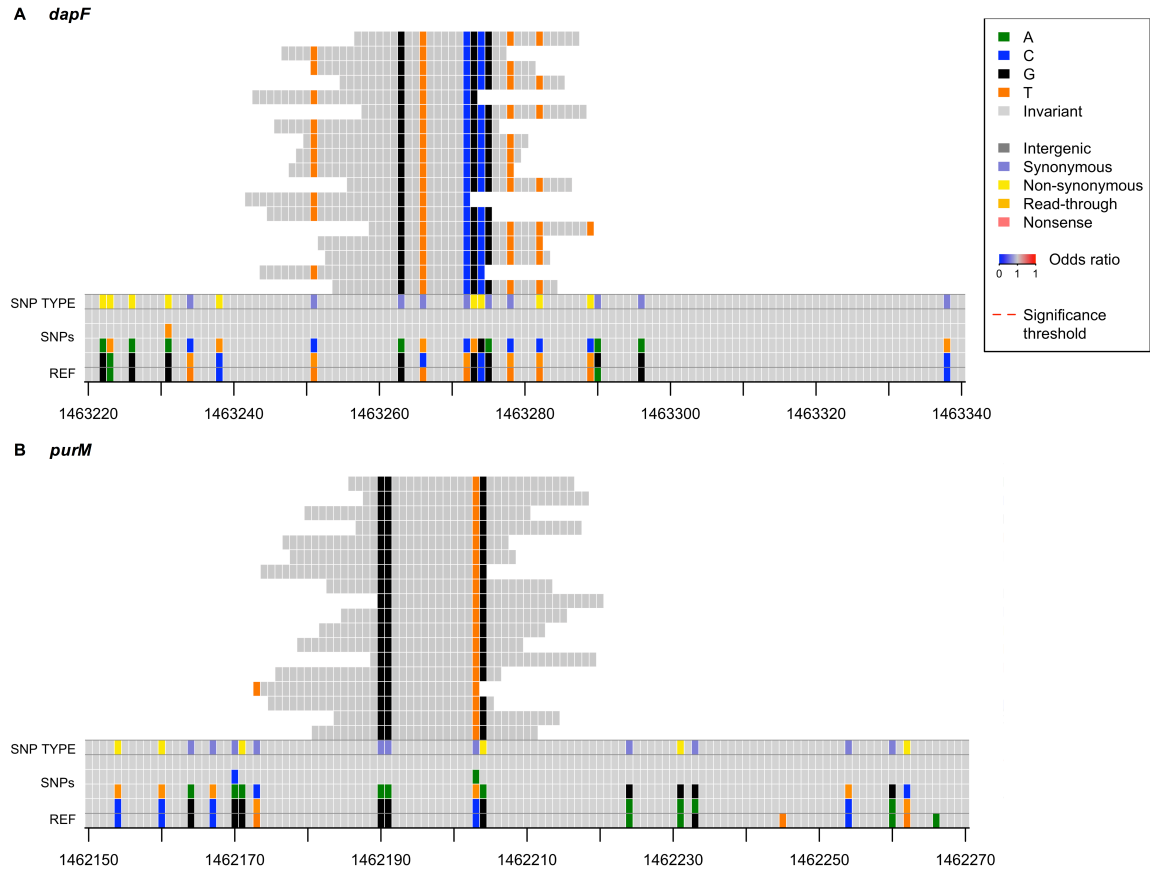

Figure 5.23 Close up of the most significant kmers within A *dapF* and B *purM*. From the bottom up, the reference is coloured grey if the site was invariant and there were no variants between the kmers and the reference at that position, coloured by the reference base otherwise. The next four rows show the SNPs based on the mapped data, from the most common allele at the bottom then decreasing in frequency. The sixth row shows SNP type, in this case blue for synonymous and yellow for non-synonymous. The kmers are shown where they mapped to and are coloured by their allele at all variant sites. The main kmer colours represent their odds ratios. All of the top kmers in *purM* contain three synonymous SNPs, and the top kmers in *dapF* cover the significant SNP at position 1463272.

The kmers in *purM* which fell just below significance, however, appear to have captured a conserved region containing the active site of the resultant protein. The most significant kmers are shown aligned to the SNP data in Figure 5.23B. They covered the positions 1461642-1461963, representing the first position of codon 92 to the last position of codon 107, respectively (Figure 5.23B). Li et al. (1999) determined the X-ray crystal structure of *purM* from *E. coli*, and also examined *purM* sequence alignments across multiple bacterial species which revealed that the most convincing region of homology were residues 94-102, a region containing residue D_{94} which is present in the putative

active-site cleft. Alignment of *C. jejuni* NCTC11168 *purM* with an *E. coli purM* (WP_053882113.1) revealed that the top kmers in *purM* correspond to the region of homology in *E. coli* containing the active site.

5.4.5.2 Variants in LOS genes were associated with chicken colonisation

Significant kmers mapped to multiple genes in the LOS biosynthesis pathway. Kmers mapped to LOS class O *orf23o* and LOS *orf11o* (Figure 5.10B, unmapped region), however these kmers were low frequency, at 1% and 2.9% respectively, and many contained long homopolymeric regions as discussed in Section 5.4.4.3 (Figure 5.19, unmapped region). Mapping the kmers to the LOS biosynthesis region of reference genome 81116 (NC_009839), which is LOS locus class E and has previously been shown to colonise chickens (Pearson et al. 2007), revealed a peak of significance in the region, and a decay of significance with increasing distance from the significant variants (Figure 5.24A). Four hundred and five significant kmers mapped to this region, with 81% containing a homopolymeric repeat length of five or fewer meaning that these LOS results appear more reliable (Figure 5.24A).

Most of the significant kmers mapped to C8J_RS05600/*orf26e* which codes for a hypothetical protein (Figure 5.24A). Observing the mapping of the kmers to *orf26e* revealed that they appeared to be tagging the presence vs absence of at least parts of the gene, rather than genomic variants. Some regions such as between 1.0892-1.0894Mb contained just one set of kmers defining presence of the region, and did not contain alternative alleles mapping to the same position (Figure 5.24B). Differing association signals along the gene could be because the LOS loci are hotspots for genetic exchange which can result in mosaicism (Parker et al. 2008). LOS classes O and P are thought to be potential intermediates between classes E and H. Gene content and synteny is highly similar between the classes, with the exception of class O containing a deletion of *orf28*

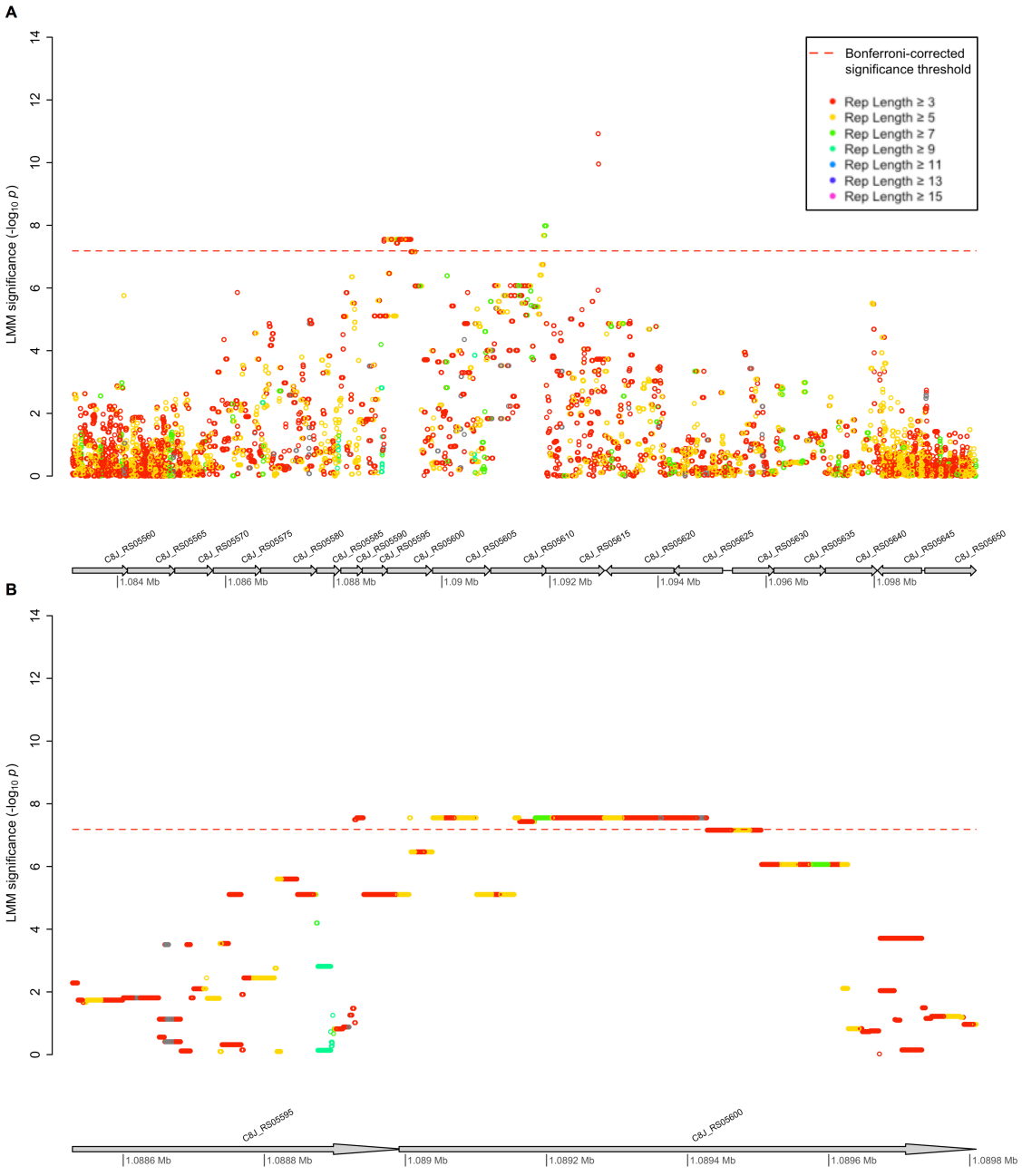

Figure 5.24 Kmer significance after controlling for population structure using LMM mapped to A the whole 81116 LOS class E and **B** LOS class E genes C8J_RS05595 and C8J_RS05600**.** Kmers are coloured by the length of the longest homopolymeric tract within the kmer. Most of the significant kmers which mapped to LOS class E only contained very short homopolymers.

and class P containing an insertion of *orf39* plus a deletion of the 5' region of *orf26* (Parker et al. 2008). These deletions and insertions all occur in class H, thus it has been hypothesised that class H arose as a result of a recombination between classes O and P (Parker et al. 2008). The variation in kmer significance across the gene

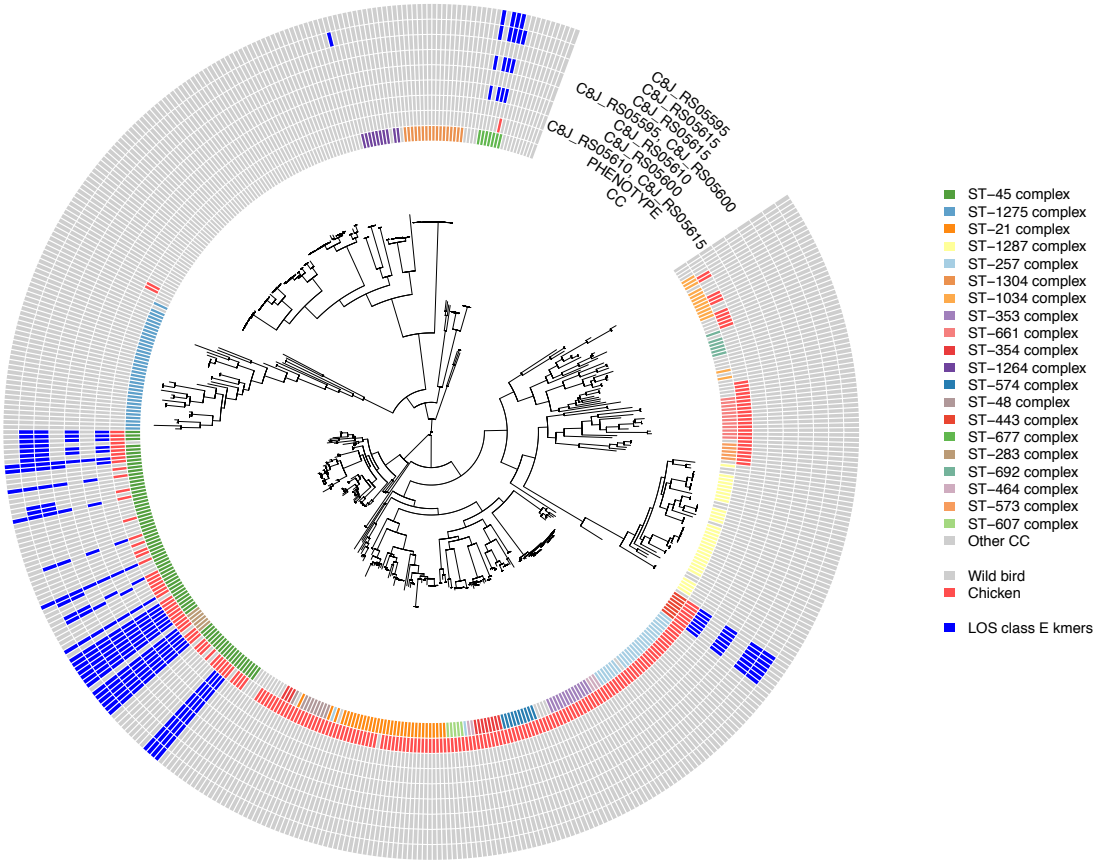

Figure 5.25 Distribution of the significant kmers which mapped to LOS biosynthesis locus class E on the phylogeny. All kmers are present in some ST-45 isolates.

C8J_RS05600/*orf26e* could therefore reflect this mosaicism. Kmers at the 3' end of the gene did not show a significant association with wild bird vs chicken colonisation which could be because the 3' end of the gene remains after the deletion of the 5' end in LOS locus classes P and H (Parker et al. 2008). Parker et al. (2005) use the absence of *orf26e* as a marker to designate isolates as LOS locus class H rather than E. Given that the significant kmers captured a gene used to designate isolates as LOS class E, and the kmers were more prevalent in chicken isolates (OR = 7.3-7.7) (Figure 5.25), this provides evidence for an association of LOS locus class E with chicken colonisation.

Observing the alignment of the kmers to C8J_RS05600 revealed that the significant kmers at the 5' end of the gene captured the conserved regions of the gene, i.e. when the gene was present, there were no variants in this region (Figure 5.26A). The

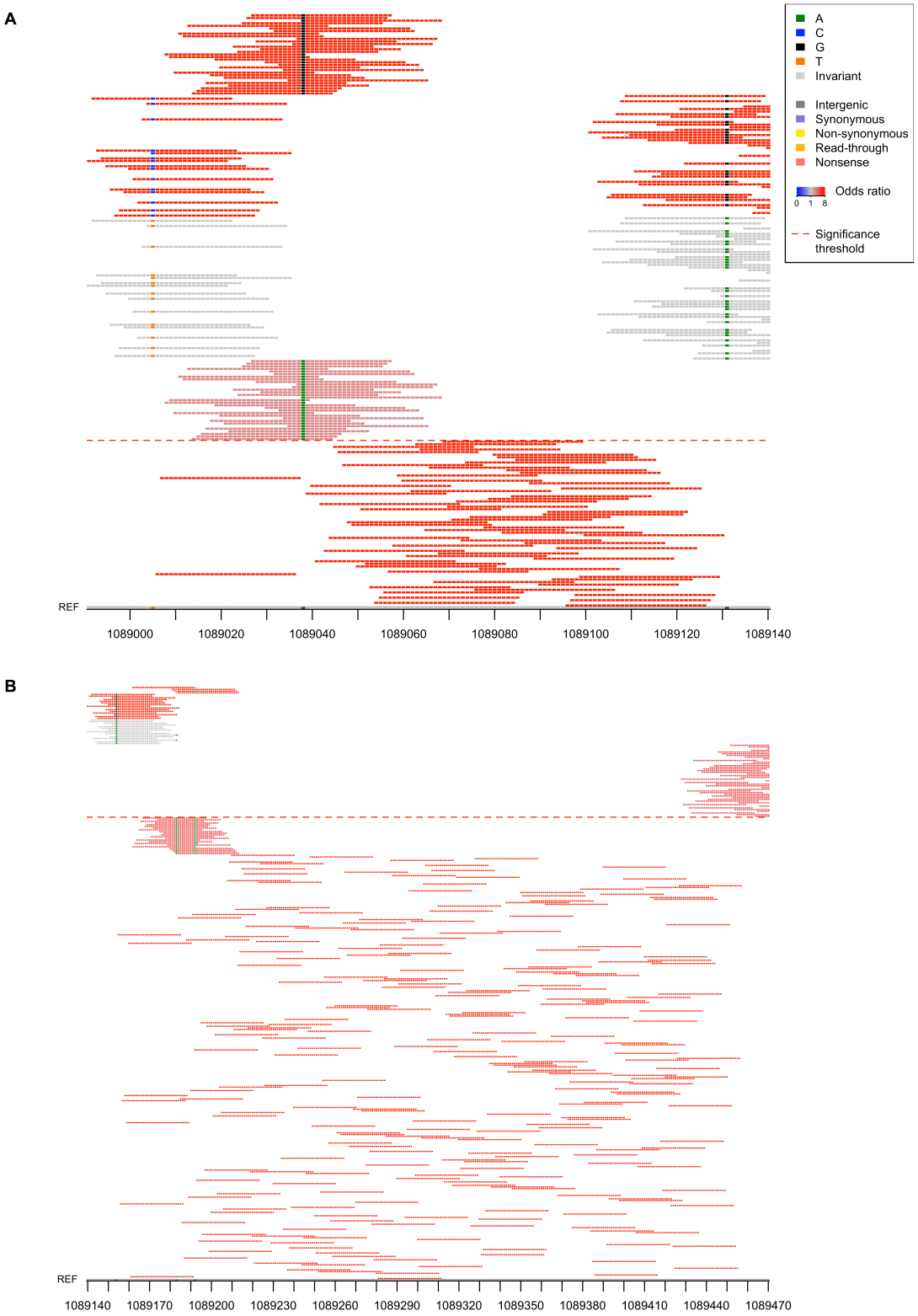

Figure 5.26 Close up of the most significant kmers mapping to two regions of the LOS locus class E gene C8J_RS05600/orf26e of reference 81116 NC_009839.1. The reference is coloured grey if the site was invariant and there were no variants between the kmers and the reference at that position, coloured by the reference base otherwise. The kmers are shown where they mapped to and are coloured by their allele at all variant sites. The main kmer colour represents their odds ratios, here all significant kmers are red, thus had odds ratios of greater than 1 and were chicken associated. The significant kmers were capturing conserved regions of the gene.

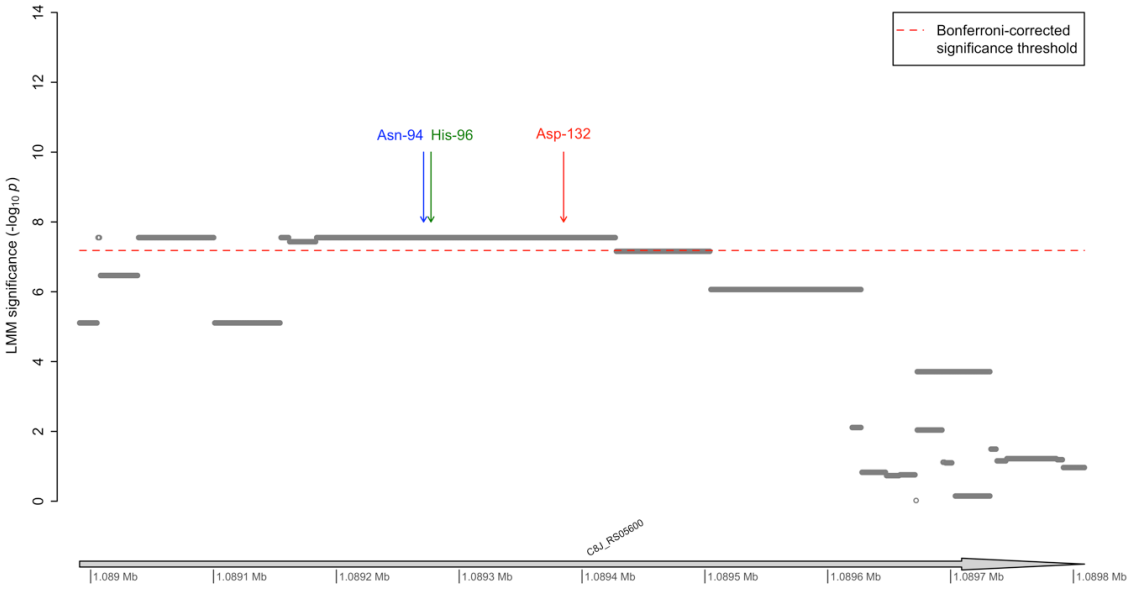

Figure 5.27 Kmer significance after controlling for population structure using LMM mapped to C8J_RS05600/orf26e of 81116. The three N-formyltransferase conserved residues are highlighted revealing that the significant kmers covered these residues and therefore the active site.

alignment of positions 1089000-1089140 revealed that the significant kmers covered all invariant sites, i.e. all kmers that did not cover sites 1089005, 1089038 and 1089131 where there is variation between the kmers that map there (Figure 5.26A).

Observing the alignment of the kmers slightly further downstream in C8J_RS05600 revealed that again, the kmers captured a region of the gene which did not vary across the dataset (Figure 5.26B). The only variation between significant kmers which mapped to the region was at positions 1089183 and 1089192, however only three and one kmer(s) contained the alternative allele at these sites, respectively, so they were low frequency variants. This indicated that the significant kmers represented presence of the gene.

C8J_RS05600/*orf26e* encodes an N-formyltransferase, also known as WlaRD, involved in the biosynthesis of Qui3NFo. A structural alignment of the protein with *E. coli* N-formyltranferase homologues by Thoden et al. (2013) previously revealed three conserved residues within the active site: Asn 94, His 96 and Asp 132. Thoden and

colleagues mutated these residues and that found mutation of any one of the residues resulted in a loss of enzymatic activity. One of the bands of significant kmers in C8J_RS05600/*orf26e* covered and contained all three conserved residues, therefore the significant kmers appeared to represent the conserved region of the gene encoding the active site of the protein (Figure 5.27).

The significant kmers mapping to LOS class E were all chicken associated, with odds ratios ranging from 7.1-46.5 at frequencies between 6-10%. The significant kmers were also associated with particular ST-complexes, which can be seen in Figure 5.25 which depicts the presence of the seven unique phylopatterns of the significant kmers. All kmers were present in some ST-45 and ST-283 complex isolates, four of the seven patterns were present in ST-443 complex isolates, three of the patterns were present in ST-667 complex isolates and one pattern was also present in all ST complexes referred to thus far plus ST-1325 (Figure 5.25). The kmers therefore potentially captured variation enabling chicken colonisation within the generalist lineage ST-45 complex.

BLAST was used to identify LOS gene presence for all LOS classes identified to date using genomes from multiple reference genome as the queries (detailed in 5.3.6). Results for the related classes E, H, O and P are shown in Figure 5.28, which revealed that this related group of LOS classes was associated with the ST-45 complex, which contained the greatest degree of mixing between chicken and wild-bird colonising isolates.

A stricter criterion of 90% length and 90% identity was used to classify genes across all LOS classes as present or absent, and genes were then tested for association using LMM, as detailed in Section 5.3.6 (Zhou & Stephens 2012). Just one gene was found to be significantly associated with the phenotype ($-\log_{10} P = 6.5$). The significant gene was *orf25p* of LOS class P, an acetyltransferase. The only genes with a $-\log_{10}p$-

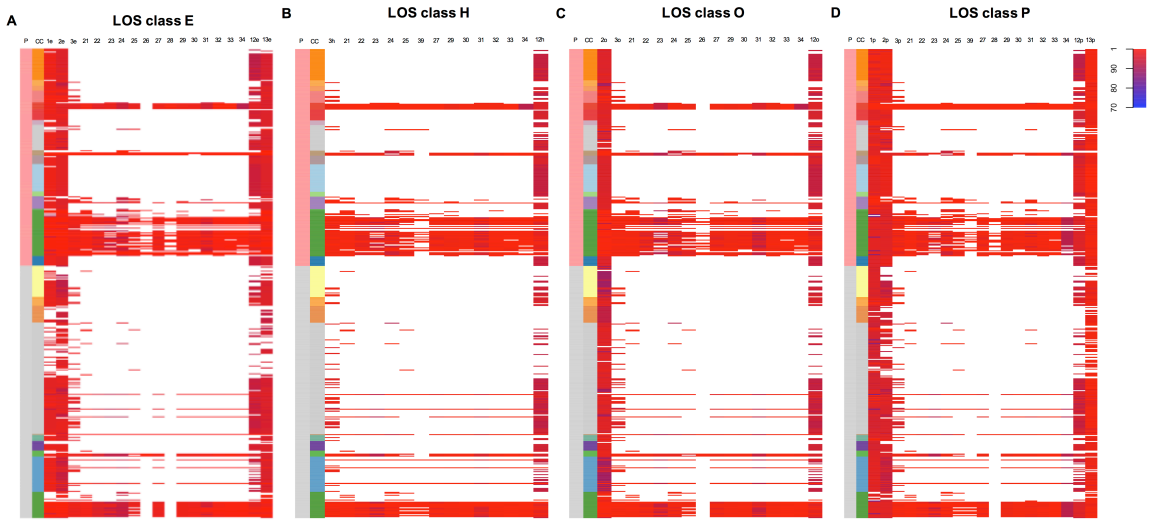

Figure 5.28 Results of scanning for each gene from LOS biosynthesis classes E, H, O and P against the assemblies using BLAST. A class E; **B** class H; **C** class O; **D** class P. Each row is an isolate, with phenotype shown in the first bar, red for chicken isolates and grey for wild bird isolates and clonal complex shown in the second bar. Each column is then a LOS biosynthesis locus. White means there was no BLAST results over at least 90% of the length with a minimum 70% identity. The colours represent the identity of matches over 90% of the length, from blue representing 70% to red representing 100% identity. This revealed these locus classes to be ST-45 complex associated (green).

value of greater than 4, and those closest to significance, were locus class E *orf26e* and *orf28e* ($-\log_{10} P = 5.8$ and 6.1, respectively), locus class O *orf26o* ($-\log_{10} P = 5.8$) and locus class P *orf28p* ($-\log_{10} P = 6.1$). The *orf26e* and *orf26o* genes both had the same phylopatterns, likewise for *orf28e* and *orf28p*. Therefore, assessing gene presence revealed a similar signal of association as the significant kmers within *orf26e* and *orf28e*.

In summary, there appear to be multiple interesting signals in the LOS biosynthesis loci. Significant kmers in the gene *orf26e* could simply be capturing the presence of the gene, but they may also be highlighting the importance of LOS class E more generally in chicken colonisation, but the nature of the associations remains to be elucidated.

5.4.5.3 Variants in multiple genes in the *C. jejuni* chemotaxis pathway were host associated

A synonymous SNP in *cheA*, a chemotaxis histidine kinase, and kmers mapping to the

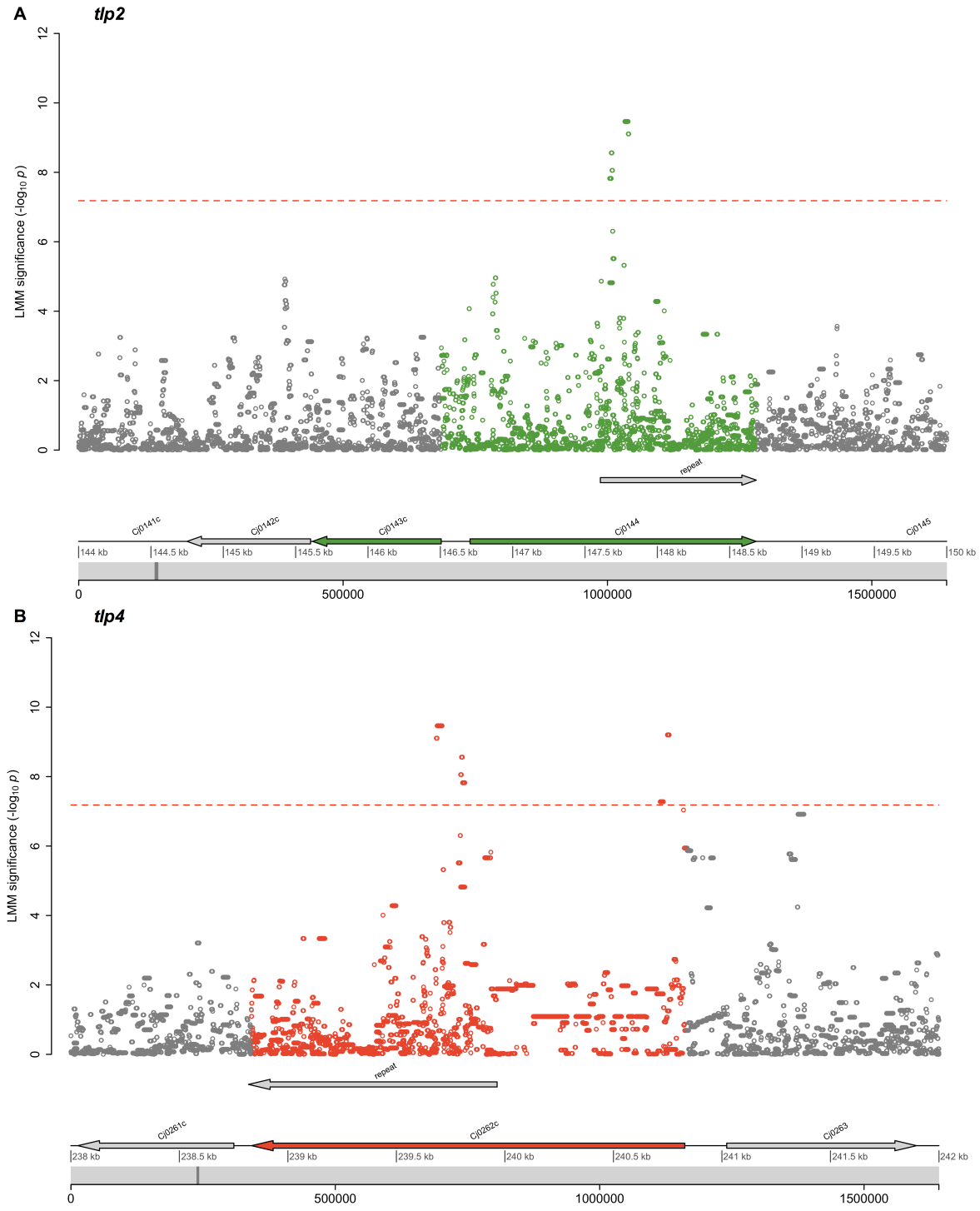

Figure 5.29 Close ups of kmer significance after controlling for population structure using LMM mapped to small regions of the genome. Kmers were mapped only to the regions shown to improve mapping in each region. **A** *Cj0144*/*tlp2*; **B** *Cj0262c*/*tlp4*; **C** *Cj1564*/*tlp3*; **D** *cetB*/*aer2*/*tlp9*; **E** *cheA*. Due to the stricter correction for multiple testing in the kmer analysis in comparison to the SNP analysis, kmers representing the significant SNP in *cheA* fell just below genome-wide significance. (*Continued on the next page*).

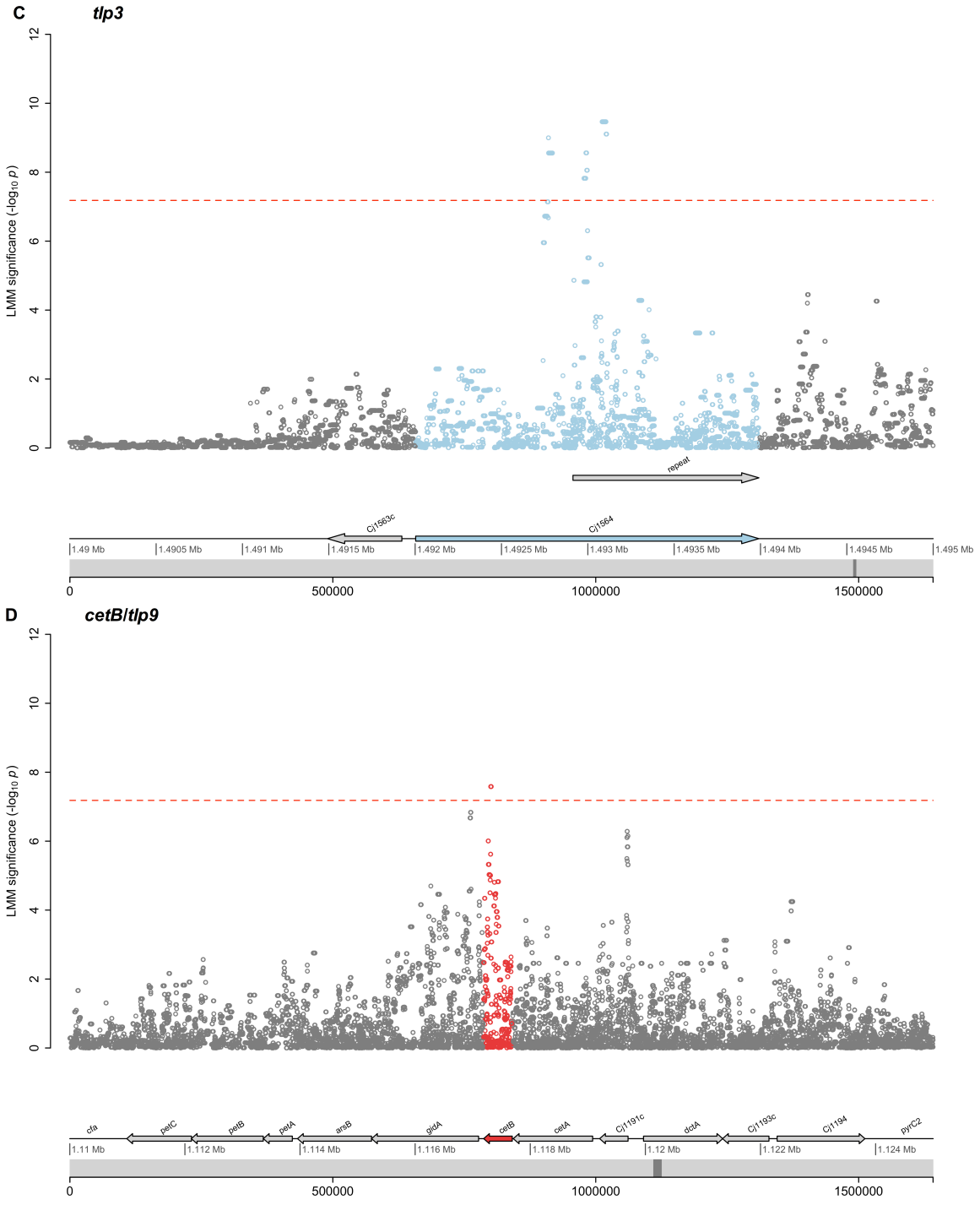

Figure 5.29 (Contd.) Close ups of kmer significance after controlling for population structure using LMM mapped to small regions of the genome. Kmers were mapped only to the regions shown to improve mapping in each region. **A** *Cj0144*/*tlp2*; **B** *Cj0262c*/*tlp4*; **C** *Cj1564*/*tlp3*; **D** *cetB*/*aer2*/*tlp9*; **E** *cheA*. Due to the stricter correction for multiple testing in the kmer analysis in comparison to the SNP analysis, kmers representing the significant SNP in *cheA* fell just below genome-wide significance. (*Continued on the next page*).

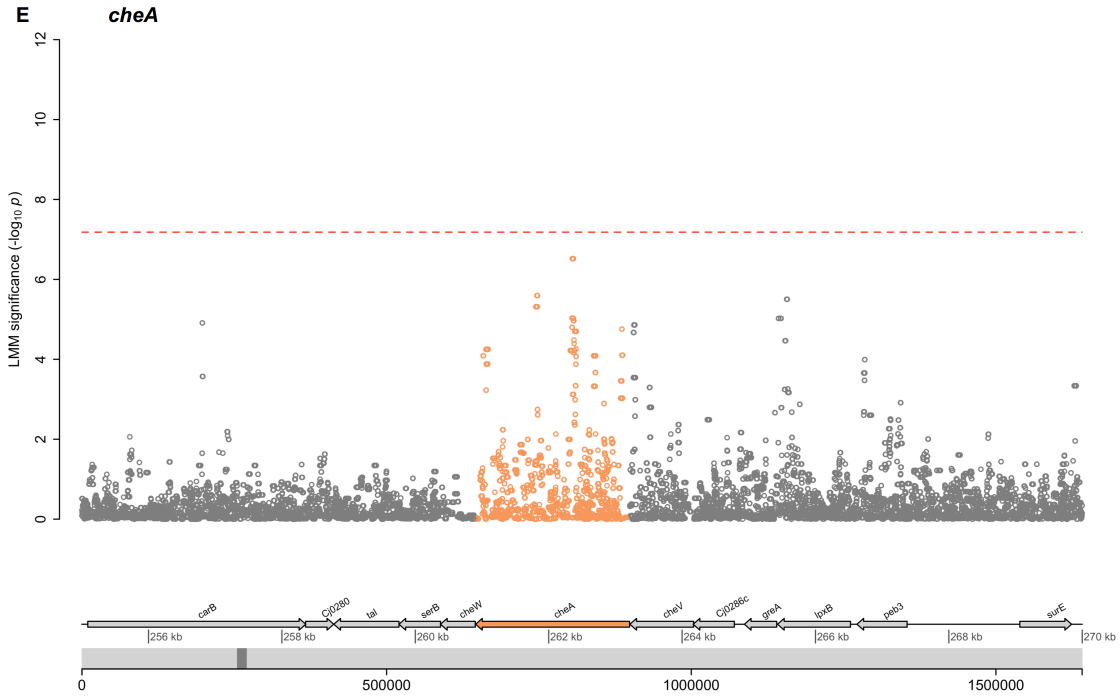

Figure 5.29 (Contd.) Close ups of kmer significance after controlling for population structure using LMM mapped to small regions of the genome. Kmers were mapped only to the regions shown to improve mapping in each region. **A** *Cj0144*/*tlp2*; **B** *Cj0262c*/*tlp4*; **C** *Cj1564*/*tlp3*; **D** *cetB*/*aer2*/*tlp9*; **E** *cheA*. Due to the stricter correction for multiple testing in the kmer analysis in comparison to the SNP analysis, kmers representing the significant SNP in *cheA* fell just below genome-wide significance.

genes *Cj0144*, *Cj0262c* and *Cj1564* encoding Tlps and *cetB*, an energy taxis response protein, were significantly host associated (Figure 5.10). We mapped kmers to the regions surrounding the genes individually to improve the mapping in each region (Figure 5.29).

Forty-seven kmers mapped to the repeat regions of *tlp2*, *tlp3* and *tlp4* (Figure 5.29A-C). The repeat regions are part of the signalling domains which are identical between Tlp2, Tlp3 and Tlp4 (Figure 5.36A). These kmers were chicken colonisation associated (OR of 119.6, 122.9, 474.5 or ∞; Table 5.3), which can be seen by the visualisation of the kmer phylopatterns in blue on the phylogeny in Figure 5.30. The most significant kmers in the region are shown aligned to the tlp4 repeat region in NCTC11168 in Figure 5.31; the mapped data from the current study does not contain any SNPs in this

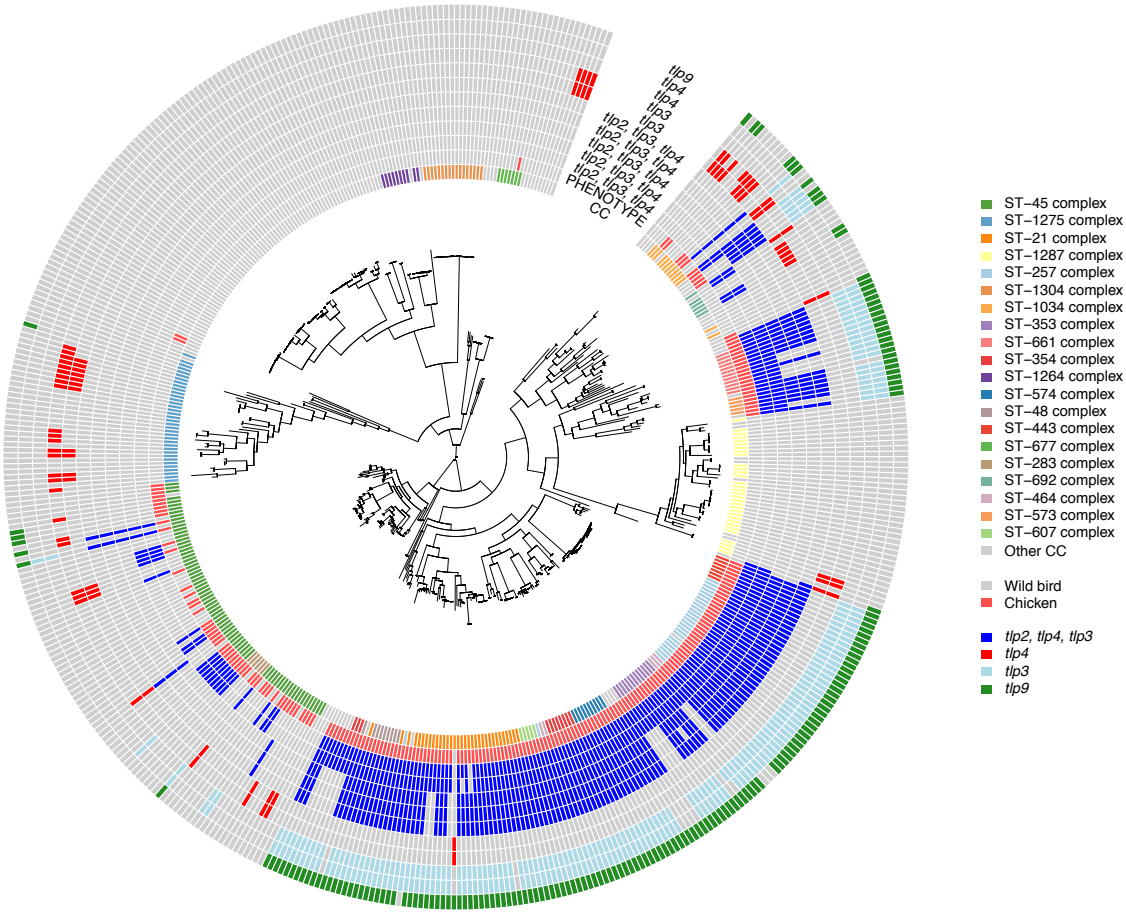

Figure 5.30 Distribution of the significant kmers which mapped to four Tlp genes on the phylogeny. Maximum likelihood phylogeny annotated by clonal complex, phenotype, and kmer presence/absence patterns for kmers which mapped to three genes, kmers which mapped equally well to the repeat regions of the genes *tlp2*, *tlp3* and *tlp4* (dark blue), kmers which mapped only to *tlp4* (red), kmers which mapped only to *tlp3* (light blue) and kmers which mapped only to *tlp9* (green).

region as repeat regions are deliberately masked in the SNP calling pipeline. As this repeat region is present in *tlp2*, *tlp3* and *tlp4*, and *tlp2* and *tlp3* have previously been found to be ubiquitous in a collection of 292 *C. jejuni* isolates (Mund et al. 2016), it is unlikely that the kmers were tagging gene presence/absence, but rather a form of genetic variation.

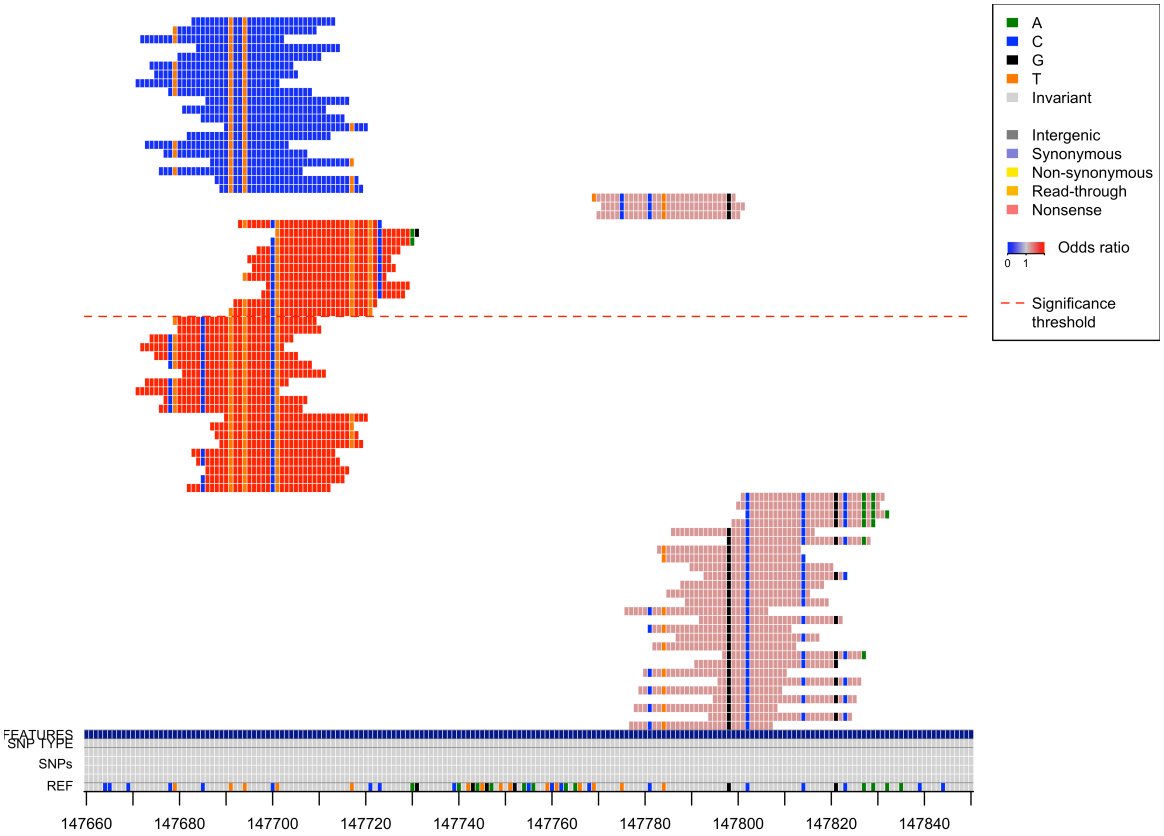

Figure 5.31 Close up of the alignment of the significant kmers which mapped to the signaling domain of *tlp2*, *tlp3* and *tlp4*, here shown mapped to *tlp4*. From the bottom up, the reference is shown coloured grey if there were no variants between the kmers and the reference at that position, coloured by the reference base otherwise. The next four rows show the SNPs based on the mapped data. In this region, there were no SNPs. The top row is coloured showing that this is a repetitive region of the genome. The kmers are then shown stacked from the most significant at the bottom decreasing in significance. Those below the red dashed line were above the bonferroni-corrected significance threshold. Kmers are coloured by their odds ratio, from protective kmers coloured in blue through to risk kmers coloured in red. Kmers are also coloured by their allele at all variant sites.

An additional 24 kmers mapped to *Cj1564*/*tlp3* outside of the repetitive region to the extracellular ligand binding domain (Figure 5.29C; Figure 5.36A). Examination of the alignment of the significant kmers in *Cj1564*/*tlp3* to the mapped SNP data revealed that all significant kmers covered variation at four biallelic non-synonymous SNPs (Figure 5.32). The significant kmers contained the minor allele of the SNP at position 1492795 and the major alleles of the SNPs at positions 1492796, 1492801 and 1492802. The kmers were only present in chicken isolates, as shown on the phylogeny in Figure 5.30.

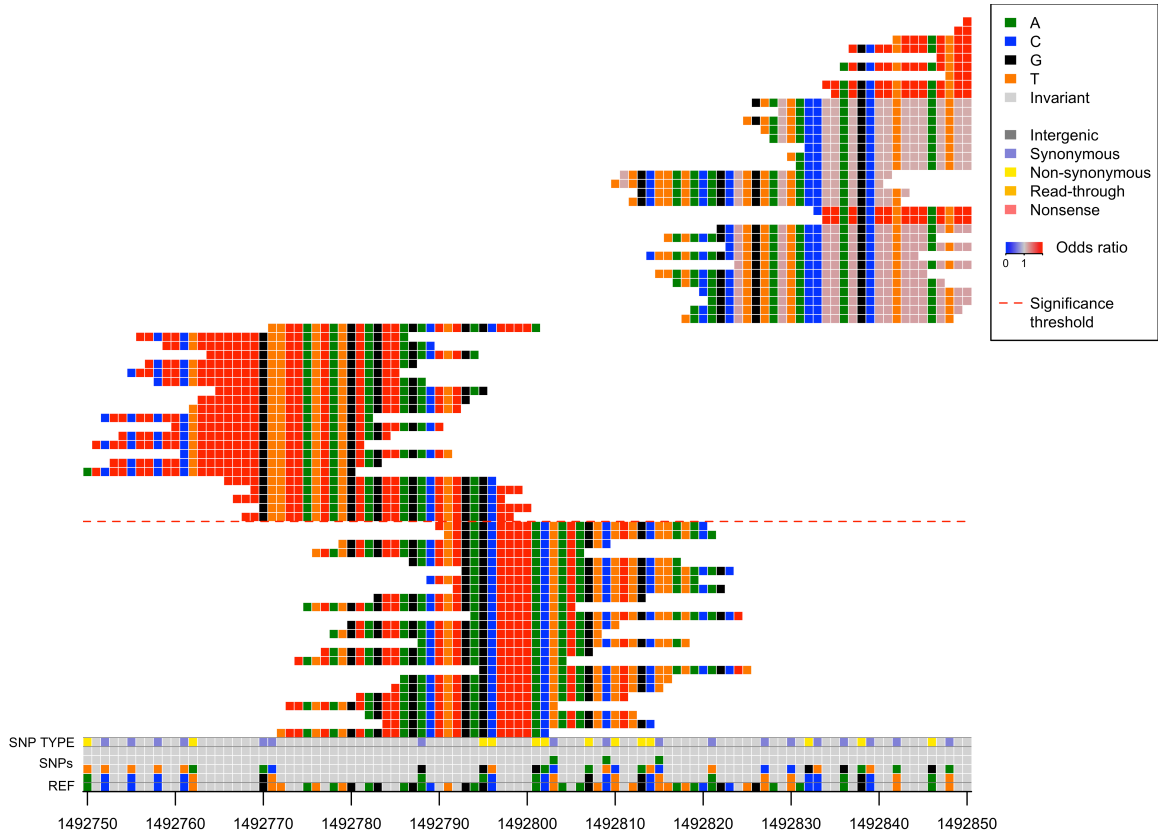

Figure 5.32 Close up of the alignment of the significant kmers which only mapped to *tlp3*. From the bottom up, the reference is shown coloured grey if there were no variants between the kmers and the reference at that position, coloured by the reference base otherwise. The next four rows show the SNPs based on the mapped data, from the most common allele at the bottom and then decreasing in frequency. The top row shows the SNPs types; here the kmers covered many non-synonymous SNPs. The kmers are shown stacked from the most significant at the bottom decreasing in significance. Those below the red dashed line were above the bonferroni-corrected significance threshold. Kmers are coloured by their odds ratio, from protective kmers coloured in blue through to risk kmers coloured in red. Kmers are also coloured by their allele at all variant sites.

The longest homopolymeric tract length in all of the kmers was three. As *tlp3* has been previously shown to be ubiquitous (Mund et al. 2016) the kmers were unlikely to have captured gene presence/absence but rather a type of genetic variation, most likely the non-synonymous SNPs.

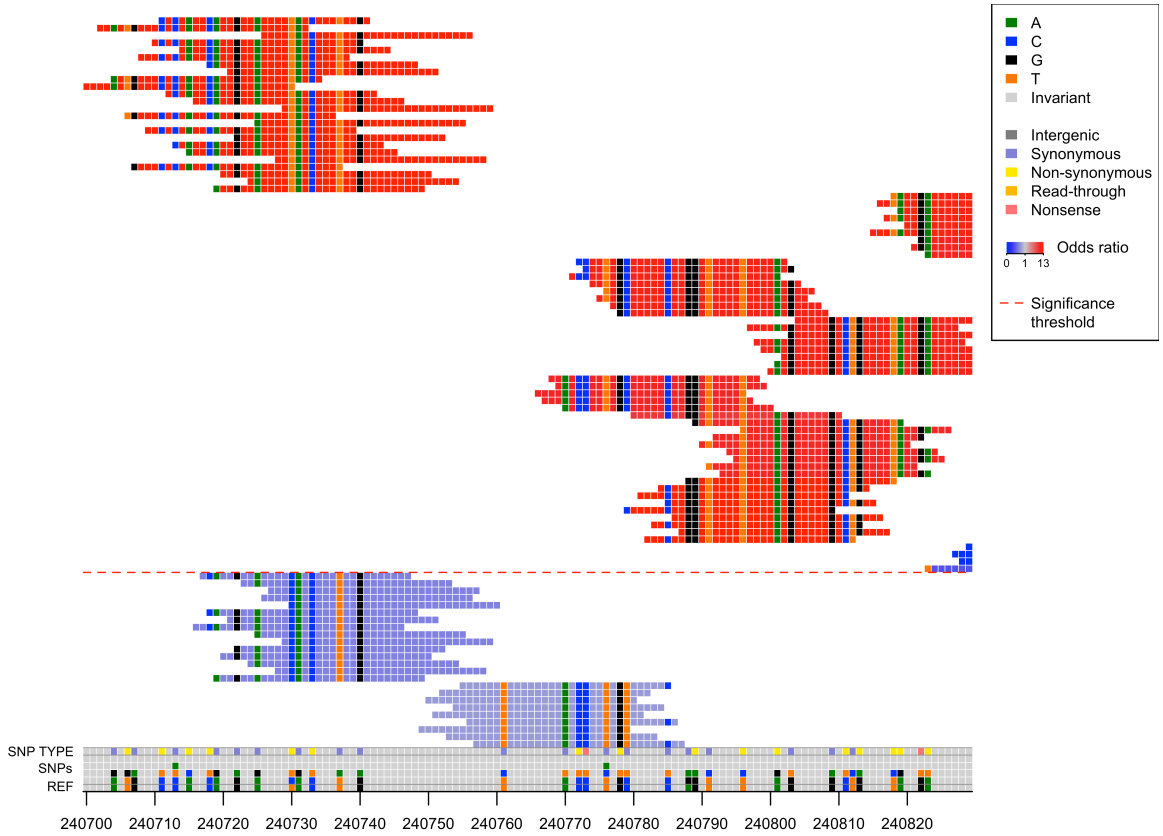

Figure 5.33 Close up of the alignment of the significant kmers which only mapped to *tlp4*. From the bottom up, the reference is shown coloured grey if there were no variants between the kmers and the reference at that position, coloured by the reference base otherwise. The next four rows show the SNPs based on the mapped data, from the most common allele at the bottom and then decreasing in frequency. The top row shows the SNP types; here the kmers covered non-synonymous SNPs, synonymous SNPs and a nonsense SNP. The kmers are then shown stacked from the most significant at the bottom decreasing in significance. Those below the red dashed line are above the bonferroni-corrected significance threshold. Kmers are coloured by their odds ratio, from protective kmers coloured in blue through to risk kmers coloured in red. Kmers are also coloured by their allele at all variant sites.

An additional 24 kmers mapped to *Cj0262c*/*tlp4* outside of the repeat region, to the N-terminal transmembrane region (Figure 5.29B; Figure 5.36A). Alignment of the significant kmers to the mapped SNP data revealed that the kmers cover many SNP variants (Figure 5.33). The nine most significant kmers covered seven SNPs, three synonymous, two non-synonymous and one nonsense, and contained the major allele for all SNPs except for the synonymous SNP at position 240779. The remainder of the significant kmers contained five SNPs, three synonymous and two non-synonymous, and contained the major allele at all sites. The kmers were wild bird associated (OR = 0.07,

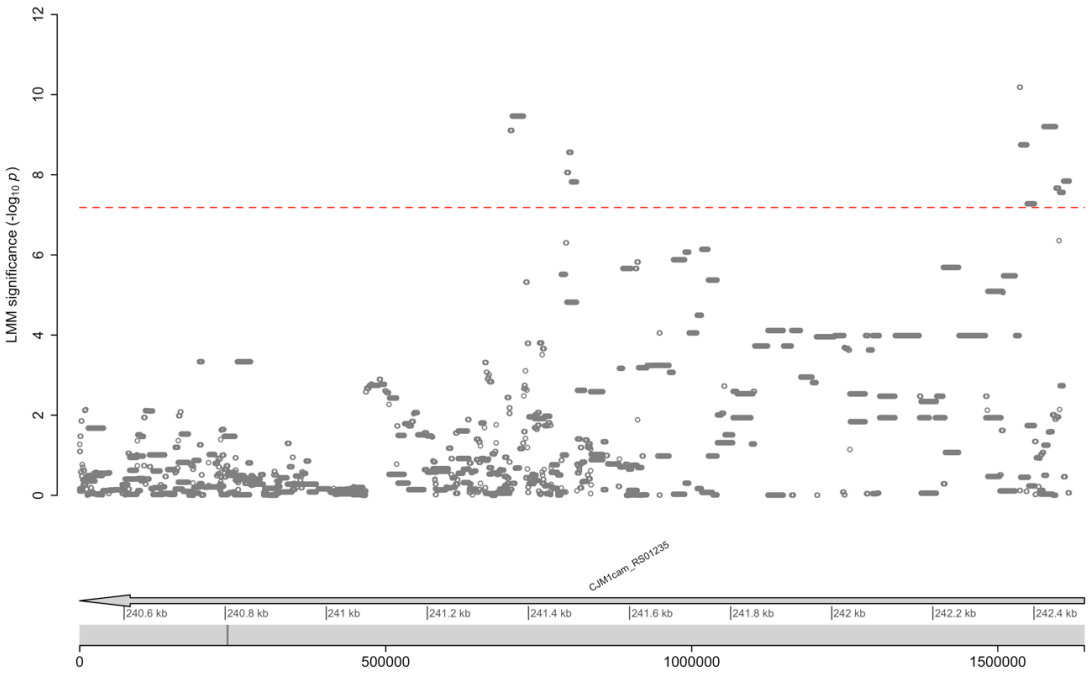

Figure 5.34 Kmer significance after controlling for population structure using LMM mapped to the *Cj0262c* homologue of the CJM1cam genome. The red dotted line indicates the Bonferroni-corrected significance threshold.

0.1, 0.05, 0.04, 0.09 and 0.06) and the presence of the significant kmers was visualised on the phylogeny in Figure 5.30. Of the 24 kmers which only mapped to *Cj0262c*/*tlp4*, 20 contained a homopolymeric region of length 4. Mund et al. (2016) previously showed that *tlp4* was not ubiquitous amongst 292 *C. jejuni* isolates, therefore it is possible that the kmers could have captured presence of the gene. However, as the kmers covered 12 SNPs it is likely that they were tagging the SNP variation, potentially an allelic combination of the SNPs, rather than presence of the gene.

Using BLAST, 37 of the kmers which did not map to the reference genome could be annotated as the gene *CJM1cam_0245* of the reference genome CJM1 (accession NZ_CP012149), another Tlp gene (Figure 5.34). Using BLAST to query NCTC11168 for the gene *CJM1cam_0245* revealed a match to bases 1-1134 and 1817-1989 of NCTC11168 *Cj0262c*/*tlp4* at 96% and 90% identity respectively. The kmers which mapped to the gene at around 241.4kb were the same kmers which mapped to the repetitive region of *tlp2*, *tlp3* and *tlp4*. *CJM1cam_0245* also has 99% identity to

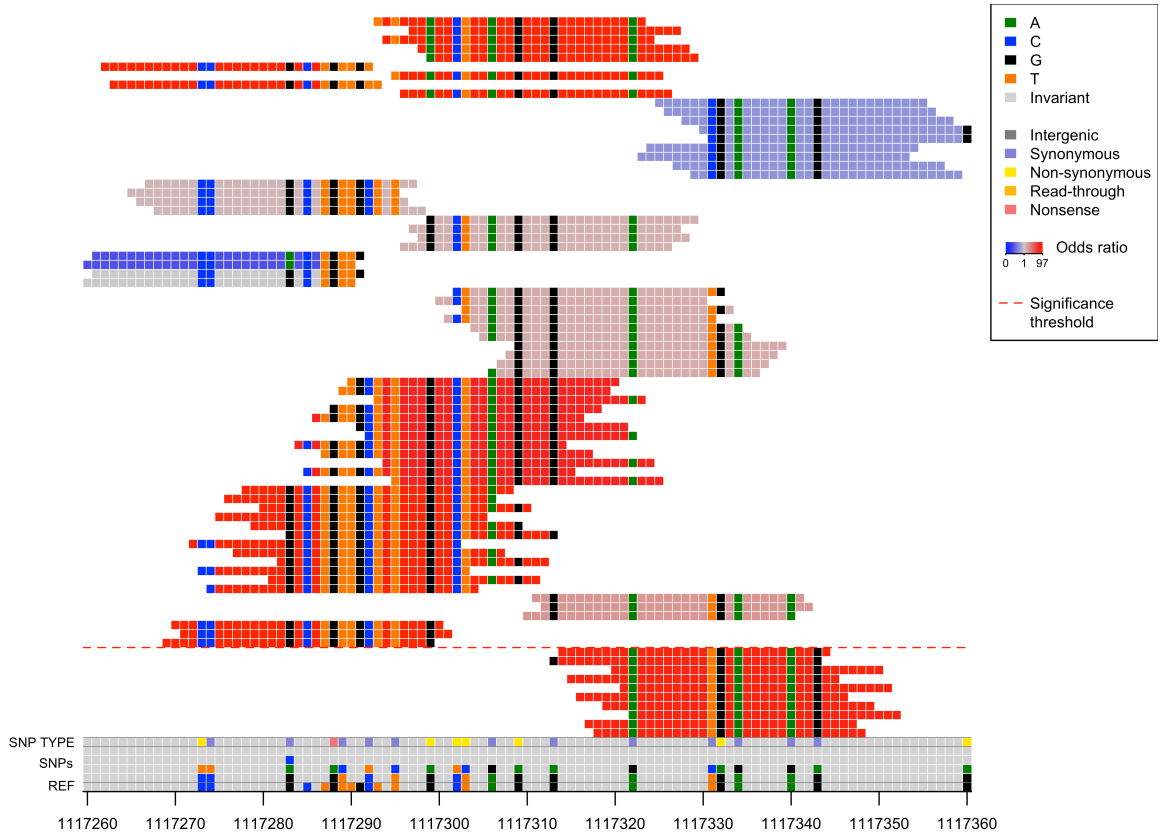

Figure 5.35 Close up of the alignment of the significant kmers which only mapped to *tlp9*/*cetB*. From the bottom up, the reference is shown coloured grey if there were no variants between the kmers and the reference at that position, coloured by the reference base otherwise. The next four rows show the SNPs based on the mapped data, from the most common allele at the bottom and then decreasing in frequency. The top row shows the SNP types; here the kmers covered non-synonymous, synonymous and nonsense SNPs. The kmers are shown stacked from the most significant at the bottom decreasing in significance. Those below the red dashed line were above the bonferroni-corrected significance threshold. Kmers are coloured by their odds ratio, from protective kmers coloured in blue through to risk kmers coloured in red. Kmers are also coloured by their allele at all variant sites.

H730_01610 of reference genome R14 (Accession CP005081). Mund et al. (2016) defined H730_01610 as a new receptor *tlp12*, where they described the sequence similarity between the gene and *Cj0262c*/*tlp4* at the 3' and 5' ends, and the the lack of significant sequence similarity between the two regions. Like the additional kmers which mapped to *Cj0262c*/*tlp4*, the additional kmers which mapped to *CJM1cam_0245* covered the same region, the N-terminal transmembrane domain.

Ten significant kmers mapped to *cetB*/*tlp9* outside of the PAS domain (Figure 5.29D; Figure 5.36B). Aligning the kmers to the mapped SNP data revealed that they

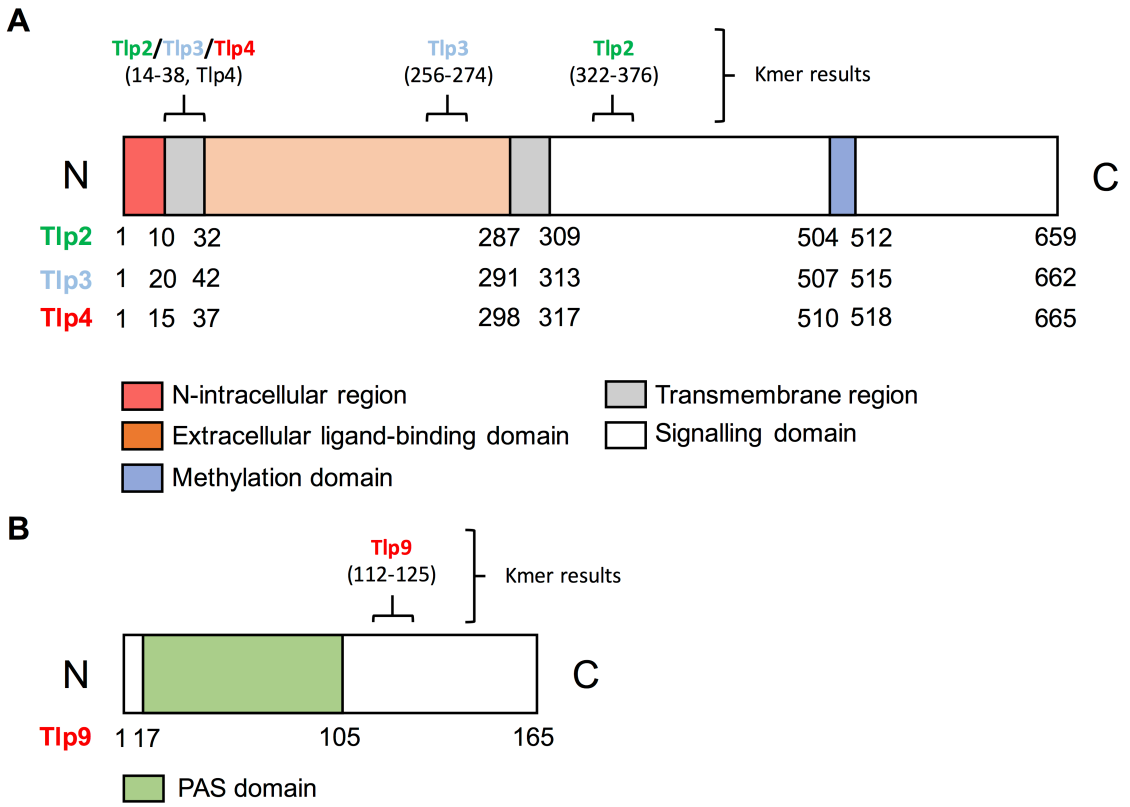

Figure 5.36 Functional domains of the *tlp* genes and where the significant kmers mapped. A Tlp2, Tlp3 and Tlp4 proteins of NCTC11168 adapted from Li et al. (2014); **B** Tlp9/*cetB* of NCTC11168 adapted from Reuter & van Vliet (2013).

covered six SNPs, five synonymous and one non-synonymous, and they contained the major allele for all SNPs (Figure 5.35). The kmers were chicken colonisation associated (OR = 90.5) and the presence of the kmers was visualised on the phylogeny in Figure 5.30. Five of the kmers had a homopolymeric tract length of 5, the others of three. Again, as *cetB* has previously been found to be ubiquitous (Mund et al. 2016), the kmers must have captured variation rather than the presence of the gene.

The signal seen in the SNP GWAS in the gene *cheA* replicated within the kmer analysis, but due to the number of kmers tested and therefore the more stringent multiple testing correction, the kmers in *cheA* were not significant (Figure 5.29E). We did still see the region of inflated significance and decay in significance with increasing distance from the top variant, however (Figure 5.29).

In summary, significant variants have been found in five genes involved in *C. jejuni* chemotaxis: *Cj0144*/*tlp2*, *Cj1564*/tlp3, *Cj0262c*/*tlp4*, *cetB*/*tlp9*, and *cheA*, however the nature of how these variants may influence host adaptation remains to be elucidated.

5.5 Discussion

5.5.1 Summary

In this chapter, a genome-wide association study of 480 *Campylobacter jejuni* isolates sampled from wild birds and chickens was undertaken.

- Sample heritability of wild bird versus chicken colonisation was estimated to be 83.6% (standard error = 2.7%) reflecting the strong clustering of wild bird and chicken isolates on the phylogeny. High sample heritability enabled accurate phenotypic predictions of 469/480 isolates.
- Twelve SNPs and 1,164 kmers were significantly associated with either wild bird or chicken colonisation.
- Four lineages were significantly associated with either wild bird or chicken colonisation in agreement with previously identified host associated lineages.
- We identified a SNP variant in the gene *dapF* downstream of *cas1*, *cas2*, *cas9* and the CRISPR repeats to be associated with chicken colonisation. Inclusion of the first 20 PCs as additional fixed effects in the LMM analysis resulted in two SNPs in the gene *purM* and one SNP in the pseudogene *Cj1528*, both adjacent to *dapF,* becoming significantly associated with chicken colonisation. The most significant kmers in *purM*, which fell just below the genome-wide significance threshold, captured the active site of the resultant protein. The SNP variant in *dapF* and the most significant kmers which captured the variant represented synonymous SNP variation, but we hypothesised that these SNPs have an effect on gene expression due to codon bias.

- We identified kmers capturing LOS class E to be associated with chicken colonisation. Four genes in the LOS class E biosynthesis locus contained kmers significantly associated with chicken colonisation, with most of the significant kmers concentrated in *orf26e*. The kmers appeared to be capturing the presence vs absence of the gene, which has previously been used as a marker to differentiate LOS locus class E and H (Parker et al. 2005). The signal varied along the gene, possibly reflecting the mosaicism seen in the region due to recombination. The significant kmers in *orf26e* captured the active site of the protein (Thoden et al. 2013).
- We identified significant kmers capturing variants in multiple genes in the chemotaxis pathway. Significant kmers were found in the genes *tlp2*, *tlp3*, *tlp4*, *tlp9* and *cheA*. Some of the variants were associated with wild bird colonisation and others chicken colonisation. This indicated that there are multiple strategies of adapting the chemotaxis pathway that are associated with host colonisation.

5.5.2 Accounting for possible batch effects in bacterial GWAS

We identified some potential issues with applying GWAS to opportunistically collected data. Experimental design is often not ideal in WGS studies, and in this study the sampling frame was sequenced in multiple batches, as evidenced by the range of mode read lengths detailed in Section 5.3.4. Of particular concern was that all isolates of the longest read length mode of 301 were wild bird isolates. Measures were therefore taken to account for potential batch effects. We applied a statistical control where the mode read length for each isolate was included as an additional fixed effect in the kmer LMM analysis. Although this had quite an effect on the significance of some of the top results, it was the same genes which were significant after accounting for read length.

We also applied a post-hoc sanity check to investigate possible batch effects, which was undertaken by identifying homopolymeric tracts within the kmers. Long homopolymers could potentially signal uncontrolled batch effects, as false insertions and deletions can occur within homopolymers by Illumina sequencing (Minoche, Dohm & Himmelbauer 2011) and the rate of errors may differ across the sequencing batches. Many of the significant results contained long homopolymers, for example the low frequency kmers in LOS *orf23o*. Although associations between the phenotype and kmers containing long homopolymers may be real, as opposed to an artefact due to a batch effect, caution must be taken with these results before experimentally validating their presence.

Of course, rather than attempting to control for possible batch effects and applying post-hoc sanity checks in order to catch any possible false positives, the best approach to combat batch effects in GWAS would be to sequence all isolates at the same time using experimental randomisation of the cases and controls. In terms of making the best that we can from available data, using kmer presence/absence to control for population structure when testing the presence of kmers for association should account for weak batch effects. Despite the opportunistically sampled data, however, we have been able to identify many locus effects which do not contain long homopolymers and are robust to using kmers to control for population structure and the inclusion of read lengths as additional fixed effects.

5.5.3 Surviving the extra-intestinal environment during host colonisation

When colonising a new host, *C. jejuni* may have to encounter novel bacteriophages, which are natural predators of the pathogen and have the ability to control the carriage of the bacterium (Carrillo et al. 2005; Wagenaar et al. 2005). Studies of UK broiler houses have shown bacteriophage isolation rates of between 42%-51% (Connerton et al. 2004;

El-Shibiny, Connerton & Connerton 2005). The presence of bacteriophages also influences the dominant strain in cases of mixed populations (Connerton et al. 2004; Scott et al. 2007).

The typical interaction between *C. jejuni* and bacteriophages is one of a predator/prey interaction where the phages are committed to lysis or lysogeny, however an alternative behaviour exists where the phage population is maintained at a stable equilibrium with the bacteria, known as the carrier state (Lwoff 1953). The bacteriophages remain associated with the bacteria, however they only propagate within a subset of the host population as a proportion are resistant (Siringan et al. 2014; Brathwaite et al. 2015). The carrier state of *C. jejuni* can be described by a non-motile phenotype and a reduction in the ability to adhere to and invade intestinal epithelial cells and colonise the chicken gut (Siringan et al. 2014; Brathwaite et al. 2015). A study by Brathwaite et al. (2015) investigated changes in gene expression in *C. jejuni* during the carrier state using RNA-seq in order to explain the accompanied phenotypic changes. They found that the major flagellin *flaA* was downregulated, explaining the impairment in mobility, along with downregulation of stress response genes *hrcA* and *hspR*, chemotactic response signalling genes *cheV*, *cheA* and *cheW* and three genes encoding the chemoreceptors Tlp1, Tlp6 and Tlp7. These changes were hypothesised to allow the bacteria and the bacteriophage to coexist, enabling *C. jejuni* to survive extra-intestinal passage.

A potential hypothesis is that the results highlighted in this chapter represent multiple pathways enabling survival during interaction with novel bacteriophages encountered during host switches. LOS class E isolates have previously shown effective resistance to tested bacteriophages, albeit lower resistance than the sialylated classes A, B and C, as the bacteriophages use the cell envelope as a receptor (Coward et al. 2006;

Louwen et al. 2013). Although we have not been able to hypothesise a function for the variants in *dapF* and *purM* downstream of the CRISPR-*cas* system, it is feasible that these are linked to the function of CRISPR-Cas and somehow work to enable the coexistence of *C. jejuni* and bacteriophages. Identification of phage-host interactions using CRISPR is likely to reflect recent host colonisation as opposed to historical, as CRISPR spacers are rapidly replaced (Edwards et al. 2016) and could therefore explain host adaptation within generalist lineages. Variants associated with wild birds and chickens at multiple stages in the chemotaxis pathway could also possibly enable the coexistence of *C. jejuni* and bacteriophages. The chemotaxis pathway enables the movement towards or away from favourable and unfavourable stimuli, with the movement caused by the flagella (Hugdahl, Beery & Doyle 1988). The study by Brathwaite and colleagues described above suggested the importance of the chemotactic response in surviving extra-intestinal environments.

5.5.4 Future directions

In order to tackle *C. jejuni* and remove it as an infection source, understanding how it adapts to its hosts and niches is key. Gaining a greater understanding of host associated variants will provide new targets for disease control. For example, insights into phage/host interactions will advance the application of phages as a biocontrol strategy (Zampara et al. 2017), as bacteriophage treatment is showing to be a promising approach for combating *C. jejuni* colonisation (Atterbury et al. 2003; Carrillo et al. 2005; El-Shibiny et al. 2009; Carvalho et al. 2010; Siringan et al. 2011).

This study investigated *C. jejuni* adaptation to wild bird and chicken colonisation. Previous studies have investigated *C. jejuni* adaptation to cattle and chickens (Sheppard et al. 2013) and to different stages of the poultry processing chain and human campylobacteriosis cases (Yahara et al. 2017). However, moving forward host adaptation

studies should look at the wider context of *C. jejuni* colonisation and investigate adaptation to multiple hosts within the same study in order to fully understand the epidemiology of *C. jejuni*. Methods will need to be adapted to investigate a categorical phenotype, but by doing so we will gain a much greater understanding of the factors enabling *C. jejuni* to adapt to its various hosts.

Chapter 6

Discussion

6 Discussion

The aim of this thesis was the development and application of genome-wide association studies to bacteria. Specifically, we wished to test the feasibility of GWAS in bacteria given the known genetic differences between bacteria and humans, where GWAS are well established. The feasibility of GWAS in bacteria was assessed by testing for associations with antimicrobial resistance phenotypes where the dominant underlying genetic mechanisms were well understood. The success of the proof of principle study provided the basis for two further GWAS, investigating phenotypes where the underlying bacterial genetic basis was less well understood. These were carriage vs invasive disease in *N. meningitidis* and wild bird vs chicken adaptation in *C. jejuni*. Below a summary of the thesis findings is presented, and future directions for bacterial association studies are discussed.

6.1 Summary of thesis findings

6.1.1 Chapter 3 – Proof of principle for bacterial genome-wide association studies

In Chapter 3, methods for the application of GWAS to bacteria were developed and applied. A GWAS of fusidic acid resistance in *S. aureus* was undertaken as a case study of the application of GWAS to bacteria. The genetic basis of resistance to eight further antimicrobials in *S. aureus* was investigated by testing SNPs and kmers. Results were also presented for antimicrobial resistance GWAS testing SNPs, genes and kmers in *E. coli*, *K. pneumoniae* and *M. tuberculosis* applied by colleagues. The key findings were:

- Failure to control for population structure resulted in a large number of false positives when testing for association between fusidic acid resistance and the presence or absence of 31bp kmers in *S. aureus*. Controlling for population structure alleviated this problem but resulted in a widespread reduction in

significance. This demonstrated that there is a substantial cost to adequately controlling for population structure.

- Testing for strain-level associations with phenotype, i.e. 'lineage effects' in bacterial GWAS identified important differences in resistance, and provided a novel way of prioritising variants for functional follow up by identifying the genetic variants with strongest leverage over the lineage effects identified.
- Comprehensively assessing the ability of GWAS to detect genuine causal variants revealed success in identifying known resistance-conferring mechanisms, and it also revealed a candidate novel association between the outer membrane porin *nmpC* and cefazolin resistance in *E. coli*.
- Testing kmers for association demonstrated some advantages over testing SNPs. For example, when multiple causal variants were within a kmer length, stronger significance was detected among kmer variants than SNPs because the presence of the wild type allele acted as a marker tagging the absence of multiple resistance-conferring mutant alleles, effectively acting to statistically pool low-frequency alternative alleles.

6.1.2 Chapter 4 – Genome-wide association study of *Neisseria meningitidis* carriage versus invasive disease

In Chapter 4, a GWAS of carriage versus invasive disease in *Neisseria meningitidis* was applied to investigate the genetic basis of virulence in a natural population of 261 isolates.

- Sample heritability was estimated to be 36%, with a standard error of 10%. This reflected reasonably accurate predictions of the phenotype using just the bacterial genetic data, demonstrating how virulence can be predicted using genetic data. Invasiveness was over-predicted, in particular all ST-11 isolates were predicted to be invasive, indicating that the heightened virulence potential of the hyperinvasive

ST-11 lineage dominant in the dataset is incompletely penetrant, probably reflecting the existence of other factors such as host health.

- The Wald test for lineage effects confirmed the known hyperinvasive ST 11 lineage as significantly associated with carriage versus invasive disease. The most significant variants associated with this lineage were also genome-wide significant, helping to explain the genetic basis behind ST 11 hypervirulence.
- Variants were identified in genes involved in capsule production and in phase variable regions as being significantly associated with carriage vs invasive disease, suggesting their importance in the virulence potential of meningococci in natural populations.
- A significantly associated SNP was identified in the fructose-bisphosphate aldolase *fba*, adjacent to known virulence factor factor H binding protein (*fhbp*), plus a SNP just short of genome-wide significance. I hypothesised that the combination of the minor alleles of the two SNPs produce a second FNR box for *fhbp* affecting the transcription and expression of fHbp. This result awaits functional validation.

6.1.3 Chapter 5 – Genome-wide association study of *Campylobacter jejuni* wild bird versus chicken adaptation

In Chapter 5 a GWAS was applied to investigate adaptation to wild bird and chicken hosts in a natural population of 480 *Campylobacter jejuni* isolates.

- Sample heritability was estimated to be 83.6% with a standard error of 2.7%, reflecting the strong genetic clustering of wild bird and chicken-colonising isolates in the population. This enabled the accurate prediction of 469/480 phenotypes using the bacterial genetic data, where isolates with incorrect predictions tended to be part of generalist lineages.

- Applying the Wald test for lineage effects to the data identified four lineages as being significantly associated with wild bird vs chicken colonisation. The four lineages identified were in agreement with studies which had previously identified host associated lineages, including wild bird colonising associated lineages ST-1264, ST-1304 and ST-1287 and chicken colonising associated lineages ST-661 and ST-573.
- Variants downstream of the CRISPR-Cas region, in genes involved in LOS biosynthesis and in genes at multiple stages of the chemotaxis pathway were identified as significantly associated with wild bird vs chicken colonisation. Although the nature of the associations is yet to be elucidated, I hypothesised that they represent multiple pathways which enable *C. jejuni* to survive encountering novel bacteriophages upon colonising a new host.

6.2 Challenges for bacterial GWAS

In the Introduction, Section 1.2.4, I identified a number of challenges facing the application of GWAS methods developed primarily in the eukaryotic setting to understand bacterial traits:

- Fundamental differences in reproduction between bacteria and eukaryotes, in particular the facultative nature of recombination in bacteria
- The strong structuring of bacteria into genetically clustered 'strains', so that the majority of genetic variation is often strain-stratified
- The existence of large, mobile, accessory genomes and diverse forms of genetic variation in bacteria

Here I take these challenges in turn and consider lessons learned.

The facultative recombination of bacteria leads to major differences in the patterns of LD between human and bacterial genomes, manifesting as genome-wide LD in

bacterial populations. I found that this was indeed a problem when testing for locus effects even for the highly recombining *N. meningitidis* where the most significant SNPs in the study of carriage vs invasive disease could not be distinguished statistically (Figure 4.6). The demonstration of this being a problem in a relatively highly recombining species suggests that this will be a widespread issue in bacterial GWAS. However fine-mapping was not as great an issue in the studies of antimicrobial resistance in Chapter 3, demonstrating the advantage given when causal variants are homoplastic. Resistance determinants were highly homoplastic as the selection pressure underlying antimicrobial resistance is extremely high, and this breaks down LD, assisting with fine-mapping. The caveat to this however, is that not all traits of interest may be strongly selected for. Bacterial GWAS will therefore need to learn from human approaches developed to take into account the functional consequence and biological plausibility of variants in order to prioritise statistically indistinguishable variants for functional follow-up studies.

Sample size and higher rates of recombination are important, however there is a trade off; the higher the rate of recombination and break down of LD, the greater the number of unique phylopatterns, and therefore the more stringent the multiple testing correction threshold. Surprisingly, simulations of *E. coli* and *K. pneumonia*e displayed the worst power to detect locus effects, despite higher recombination rates and therefore one would have presumed homoplasy. This is because combined with a small sample size, the more stringent multiple testing correction came at a cost, demonstrating the importance of both sample size and recombination rate (Section 3.4.4).

The second challenge with applying GWAS to bacteria is that the majority of variants are strain-stratified due to highly structured populations, and controlling the false positive rate by accounting for population structure risks loss of power to detect true causal associations. I found that kmers capturing the known *S. aureus* fusidic acid

resistance-determinant *fusC* greatly reduced in significance after controlling for population structure due to the variant only being present in strains ST-1 and ST-8 (Figure 3.6). The effect of controlling for population structure was also drastic in the studies of carriage vs invasive disease in *N. meningitidis* and in chicken vs wild bird host association in *C. jejuni* (e.g. Figure 4.5A; Figure 5.22). The large reduction in significance for these variants could result in missing causal variants which happen to be population stratified.

However, power lost when controlling for population structure can be regained by testing for lineage effects**,** as shown in Section 3.4.6. Defining lineages as PCs can not only capture important strain level differences which can be interpreted, but loci can also be assigned to the lineages, enabling a new way of interpreting the locus-level effects. Interpreting the most significant variants of those assigned to the significant lineage-level effects provides a way to prioritise variants for experimental follow up. In the case of fusidic acid resistance in *S. aureus* I found that the top variants within those assigned to the significant lineages were also genome-wide significant, as were the top variants within the hyperinvasive lineage effect identified in *N. meningitides*. However, it is possible that for other datasets and phenotypes, there will not be sufficient power to identify locus effects, and this would provide an alternative way of prioritising variants. Assigning variants to lineages can also help to understand the mechanisms behind the lineage effects, such as hyperinvasiveness, by identifying the variants with the greatest leverage over the lineage effects.

A final challenge with applying GWAS to bacteria is that of capturing the accessory genome. SNP based studies as employed in the human setting are only able to test core variation, and non-homologous genetic exchange in bacteria can result in extensive accessory genomes which often encode for important resistance and virulence

factors. Through this thesis I came to find that using kmers to capture genetic variation was a flexible way of addressing this problem. I found that kmers could define variation representing SNPs in the core genome (e.g. fusidic acid resistance determined by variants in *fusA* in the core genome of *S. aureus*), presence vs absence of accessory genes (e.g. variable presence of LOS class E *orf26e* in *C. jejuni*) and also small insertions and deletions (e.g. phase variable homopolymers in *siaD* in *N. meningitidis*).

I found that the power of the kmer approach was demonstrated not only by its ability to investigate accessory gene variation and insertions and deletions not captured by core SNP analyses, but also by identifying SNP variation not found by testing SNPs in isolation. In the case of rifampicin resistance in *M. tuberculosis*, greater significance was found by testing kmers rather than SNPs individually, despite resistance being determined by individual SNPs (Figure 3.21). By pooling over multiple resistance conferring mutations, kmers capturing the wild-type sensitive haplotype were more significant than the individual SNPs. In this case, both the resistance-conferring SNPs and the kmers containing the wild-type alleles at these sites were genome-wide significant. However, in the case of carriage vs invasive disease in *N. meningitidis*, I found that kmers within the intergenic region between the genes *ctrE* and *ctrF* captured SNP variation not identified as significant by the SNP analysis (Figure 4.22). The significant kmers appeared to have captured an allelic effect of the combination of three alleles, at sites not found to be significant when tested individually. This revealed an improvement of testing kmers over SNPs and genes individually beyond that of capturing accessory and diverse variation.

The use of short read sequencing however means that the kmer approach is still somewhat limited in what it is able to capture; repetitive regions are difficult to assemble and therefore will not always be represented in kmer studies based on *de novo* assemblies, and counting kmers from sequencing reads suffers from the difficulty of determining an

appropriate presence threshold in the presence of repetitive regions. The future of long-read sequencing will therefore be important in investigating the contribution of these regions to phenotypes of interest as discussed below in Section 6.4.1.

6.3 Strengths and limitations of the GWAS approach

The approaches applied within this thesis consisted of testing genetic variants using LMM to control for population structure, and decomposing the background genetic effects of the LMM into the contribution of PCs which were also tested for association. There were three general limitations of the approach applied to testing for locus effects: correlated phenotypes, the assumption of additivity between variants and testing variants individually, and I address these in turn below.

False positives were encountered as a result of correlated phenotypes, due to multi-drug resistant isolates in the studies of antimicrobial resistance in Chapter 3. It will therefore be important to be aware of potential phenotype correlations, particularly for antimicrobial resistance where resistance-conferring mutations can be found to occur in predictable patterns (Manson et al. 2017). Multivariant analyses may be more appropriate in these settings where multiple phenotypes are known, as they have been shown to increase power to detect variants underlying not only pleiotropic effects, but also variants which affect just one of the known phenotypes (Morris et al. 2010, Stephens 2013).

Variants were also only tested for additive effects, whereas interactions between variants may be of importance in understanding the bacterial genetic basis behind important phenotypes. Testing for epistatic effects enables the investigation of the importance of interactions between loci, and the benefits and potential challenges of epistasis analyses are discussed in further detail below in Section 6.2.2.

LMMs have been shown to be the favoured approach in the human setting, however it is likely that many important traits of interest are polygenic, which has been

shown for resistance to many antimicrobials as discussed in Section 3.1.5.4, yet variants are typically tested individually. LMMs have been shown to increase power in the presence of polygenic effects (Yang et al. 2014), and testing for association using kmers can combine variants within a kmer length as discussed above. However future work could explore the possibility of testing variants for their joint effect on the phenotype, such as using a step-wise regression or computationally intensive MCMC methods, such as Gibbs samplers for the family of models known as scale mixtures of normals (SMN) (Rajaratnam et al. 2017). SMNs relax the highly constraining assumption of a normal prior distribution on effect sizes with other, more heavy-tailed distributions such as the exponential distribution (Bayesian LASSO) or Student's t distribution (Rajaratnam et al. 2017).

Testing variants individually is particularly underpowered when it comes to rare variants. Further development to the method could be achieved by combining rare variants and testing for a common direction of effect on the trait by treating them as a single covariate, whether that be pooling by gene, pathway, or other method of combining variants. This is known as a burden test, and has previously been applied to bacteria by Desjardins et al. (2016). This is appropriate for SNP based association testing but will be more difficult in the kmer setting, where it would rely on accurate assignment of kmers to each gene or pathway, and mapping short kmers can be problematic.

The second approach taken within this thesis to understand the bacterial genetic contribution to important phenotypes is that of testing lineages for their effect on the phenotype. The strength of the lineage test in defining lineages by PCs is that they are mutually uncorrelated, and therefore capture different axes of genetic variability. In comparison to defining lineages by phylogenetic clusters this should result in a better powered test for detecting lineage effects, and also does not require a phylogeny to define

lineages which would be inappropriate for recombining species. The limitation of defining lineages by PCs however is that they then need to be interpreted, however as shown in Figure 3.9, PCs can be interpreted phylogenetically.

6.4 Future directions

6.4.1 Long read sequencing to investigate the contribution of repetitive regions to traits of interest

The GWAS of carriage vs invasive disease in *N. meningitidis* in Chapter 4 highlighted how short read sequencing is limited in its ability to understand the contribution of repetitive DNA to important phenotypes. Illumina short-read sequencing produces reads which are often smaller than many repetitive sequences in the bacterial genome, with the largest repeat class being rDNA with an average repeat size between 3-4kb (Treangen et al. 2009; Delihas 2011). Repetitive elements are a great challenge to genome assembly algorithms as they often cannot be resolved, resulting in fragmented assemblies (Treangen & Salzberg 2012). The introduction of Pacific Biosciences and Oxford Nanopore Technologies platforms enabled the sequencing of fragments of over 10kb, however their per-base error rates are much higher than Illumina reads (Eid et al. 2009; Jain et al. 2015). Many tools are now available for long read assembly (Koren & Phillippy 2015), and long read assemblies have been shown to produce closed genomes despite the error rates (Chin et al. 2013; Koren et al. 2013; Loman, Quick & Simpson 2015). Tools are also emerging which make use of both short and long sequencing reads to produce hybrid assemblies, with recently developed tools including hybridSPAdes (Antipov et al. 2016), Canu (Koren et al. 2017) and Unicycler (Wick et al. 2017).

The utility of long read sequencing was displayed by Ashton et al. (2015) who demonstrated how nanopore sequencing technology could identify both the position and structure of an antibiotic resistance island. Ashton and colleagues applied a hybrid

approach, where nanopore sequencing reads were used to create a scaffold for an Illumina short-read assembly, which in all but two cases had been fragmented by the presence of insertion sequences. The insertion site of the resistance island was identified, and it was shown to be flanked by IS1 elements, which had not previously been characterised by other whole genome sequencing projects. When read lengths are longer than the rDNA operon, long-read sequencing can typically close bacterial genomes (Koren et al. 2013). The use of long-read sequencing for GWAS appears to be promising, and as error rates and costs reduce, this will enable us to access areas of the genome currently unexplored by GWAS.

6.4.2 Epistasis

Another important area of future research is that of epistasis. GWAS in bacteria has thus far typically tested for additive effects, however interactions between loci may be of importance in understanding the bacterial genetic basis to important traits. Identifying co-evolving loci could reveal important protein-protein interactions but also indirect interactions affecting traits of interest. However, distinguishing true epistatic interactions from linkage disequilibrium (LD) will be a challenge, as will controlling for the large multiple testing burden. Recent analyses include the study by Cui et al. (2015) who tested for associations between variants within 51 unrelated *Vibrio parahaemolyticus* isolates, identifying a single pair of distant loci which were strongly correlated. Skwark et al. (2017) also described a method developed to identify co-evolving loci under the strongest co-evolutionary pressures and assessed its ability to identify loci in 3,156 *Streptococcus pneumoniae* and 3,442 *Streptococcus pyogenes* isolates. The chromosomes were split into non-overlapping regions, and pairs of loci randomly selected from the regions were tested for association in order to reduce the effect of LD and improve computational efficiency. Analysis of *S. pneumoniae* identified 5,199 couplings affecting 1,936 sites as being part

of epistatic interactions, with most being links between genes which encode beta-lactam resistance-conferring mutations. Analysis of the more clonal *S. pyogenes* revealed 5,952 couplings, of which eight of the top 20 involved two RNA pseudouridine synthases, each linked to separate loci in the chromosome. The application of such methods will be a complementary approach to standard GWAS analyses and provide an alternative way of generating hypotheses which can be functionally investigated without requiring a phenotype.

6.4.3 Integration of human and bacterial data

A particularly exciting prospect is the potential to integrate bacterial GWAS with human GWAS data. Human infectious disease GWAS have been undertaken for many pathogens (Chapman & Hill 2012). For example, the study by Davila et al. (2010) discussed in Chapter 4 identified SNPs significantly associated with host susceptibility to meningococcal disease. Over 10 GWAS have also investigated susceptibility to tuberculosis (Uren et al. 2017) and multiple variants have been identified, such as a gene-poor region on chromosome 18q11.2 (Thye et al. 2010), a locus downstream of the gene *WT1* on chromosome 11p13 (Thye et al. 2012), variants in the gene *ASAP1* (Curtis et al. 2015) and in the *HLA* class II region (Sveinbjornsson et al. 2016).

Examples of the integration of host and pathogen data exists in the viral setting, but it has not yet been explored in bacteria. Bartha et al. (2013) investigated paired human and viral data from 1,071 individuals infected with HIV and tested for associations between host and viral DNA variants and plasma viral load. The strongest signals identified were between human variants capturing human leukocyte anitgen (HLA) class I alleles and viral variants capturing cytotoxic T lymphocyte (CTL) epitopes, and greater power to detect host factors was achieved when using viral sequence variation as an intermediate phenotype. A further study by Bartha et al. (2017) defined the proportion of

variation in HIV-1 set point viral load which could be assigned to viral and human genetic variants. Bartha and colleagues applied LMMs to estimate the sample heritability of set point viral load in 33 variants in HLA class I genes, in the viral phylogenetic tree, and in a combination of the two. The proportion of phenotypic variation attributed to the combination of the viral phylogenetic tree and HLA variants was 29.9% (standard errror = 12%), however 28.8% of phenotypic variation could be attributed to just the viral phylogenetic tree, demonstrating that the HLA variants do not explain additional heritability above the viral data.

A second paired viral and host GWAS study was performed by Ansari et al. (2017) investigating 542 individuals chronically infected with hepatitis C virus (HCV). The GWAS identified variants in human *IFNL4* and HLA genes which were associated with HCV variants, and also identified that an interaction between the host *IFNL4* variants and a particular substitution in the HCV NS5A protein determines HCV viral load. The application of combined bacterial and host association studies will be a powerful tool in understanding the genetic architecture of important diseases, and assist in accurately partitioning heritability between host and pathogen contributions. However, such studies share the problem of a larger multiple testing burden with studies of epistasis. Methodological advances are therefore also needed to facilitate these studies (see e.g. (Crawford et al. 2017; Wilson 2017).

6.5 Conclusion

In this thesis, methods for applying GWAS to bacteria were developed and applied, investigating the genetic basis of antimicrobial resistance as a proof of principle of the methods. This demonstrated the importance of controlling for population structure in bacterial analyses and the knowledge that can be gained from identifying lineage-level associations which are normally discarded in the process. The methods were then applied

to two novel phenotypes. The first, carriage vs invasive disease in *N. meningitidis*, revealed the importance of particular capsule variants and phase variable regions in natural variability in virulence in *N. meningitidis*. The study also identified a putative additional FNR box, a binding site for the global transcriptional regulator FNR protein, upstream of the gene encoding for factor H binding protein which may affect expression of the protein. The second, wild bird vs chicken colonisation in *C. jejuni*, revealed associations in loci within pathways which could plausibly be involved in bacteriophage resistance when encountering a new host. Future functional work will be required to understand the mechanisms behind these associations, but it is hoped that the methods developed and insights gained from real-life applications of GWAS to bacteria in this thesis can be used to further our knowledge of the bacterial genetic contribution to important phenotypes.

Chapter 7

Bibliography

7 Bibliography

Achtman, M 2008, 'Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens.', *Annual Review of Microbiology*, vol 62, pp. 53-70.

Alam, MT, Petit, RA, Crispell, EK, Thornton, TA, Conneely, KN, Jiang, Y, Satola, SW & Read, TD 2014, 'Dissecting vancomycin-intermediate resistance in staphylococcus aureus using genome-wide association.', *Genome Biology and Evolution*, vol 6, no. 5, pp. 1174-1185.

Alekshun, MN & Levy, SB 2007, 'Molecular Mechanisms of Antibacterial Multidrug Resistance', *Cell*, vol 128, no. 6, pp. 1037-1050.

Alexander, DH, Novembre, J & Lange, K 2009, 'Fast model-based estimation of ancestry in unrelated individuals', *Genome Research*, vol 19, no. 9, pp. 1655-1664.

Altshuler, D, Daly, MJ & Lander, ES 2008, 'Genetic Mapping in Human Disease', *Science*, vol 322, no. 5903, pp. 881-888.

Andre, E, Goeminne, L, Cabibbe, A, Beckert, P, Kabamba Mukadi, B, Mathys, V, Gagneux, S, Niemann, S, Van Ingen, J & Cambau, E 2017, 'Consensus numbering system for the rifampicin resistance-associated rpoB gene mutations in pathogenic mycobacteria', *Clinical Microbiology and Infection*, vol 23, no. 3, pp. 167-172.

Ansari, MA, Pedergnana, V, L C Ip, C, Magri, A, Von Delft, A, Bonsall, D, Chaturvedi, N, Bartha, I, Smith, D, Nicholson, G, McVean, G, Trebes, A, Piazza, P, Fellay, J, Cooke, G, Foster, GR, Consortium, S-H, Hudson, E, McLauchlan, J, Simmonds, P, et al. 2017, 'Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis C virus', *Nature Genetics*, vol 49, no. 5, pp. 666-673.

Antipov, D, Korobeynikov, A, McLean, JS & Pevzner, PA 2016, 'hybridSPAdes: an algorithm for hybrid assembly of short and long reads', *Bioinformatics*, vol 32, no. 7, pp. 1009-1015.

Ashgar, SSA, Oldfield, NJ, Wooldridge, KG, Jones, MA, Irving, GJ, Turner, DPJ & Ala'Aldeen, DAA 2007, 'CapA, an autotransporter protein of Campylobacter jejuni, mediates association with human epithelial cells and colonization of the chicken gut', *Journal of Bacteriology*, vol 189, no. 5, pp. 1856-1865.

Ashikaga, S, Aymerich, S, Bessieres, P, Boland, F, Brignell, SC, Bron, S, Bunai, K, Christiansen, LC, Danchin, A, Débarbouillé, M, Dervyn, E, Deuerling, E, Devine, K, Dreesen, O, Errington, J, Fillinger, S, Foster, SJ, Fujita, Y, Galizzi, A, Gardan, R, et al. 2003, 'Essential Bacillus subtilis genes.', *Proceedings of the National Academy of Sciences*, vol 100, no. 8, pp. 4678-4683.

Ashton, PM, Nair, S, Dallman, T, Rubino, S, Rabsch, W, Mwaigwisya, S, Wain, J & O'Grady, J 2015, 'MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island', *Nature Biotechnology*, vol 33, no. 3, pp.296-300.

Astle, W & Balding, D 2009, 'Population Structure and Cryptic Relatedness in Genetic Association Studies', *Statistical Science*, vol 24, no. 4, pp. 451-471.

Atterbury, RJ, Connerton, PL, Dodd, CER, Rees, CED & Connerton, IF 2003, 'Application of Host-Specific Bacteriophages to the Surface of Chicken Skin Leads to a Reduction in Recovery of Campylobacter jejuni', *Applied and Environmental Microbiology*, vol 69, no. 10, pp. 6302-6306.

Bacon, DJ, Szymanski, CM, Burr, DH, Silver, RP, Alm, RA & Guerry, P 2001, 'A phase-variable capsule is involved in virulence of Campylobacter jejuni 81-176', *Molecular Microbiology*, vol 40, no. 3, pp. 769-777.

Baffone, W, Casaroli, A, Citterio, B, Pierfelici, L, Campana, R, Vittoria, E, Guaglianone, E & Donelli, G 2006, 'Campylobacter jejuni loss of culturability in aqueous microcosms and ability to resuscitate in a mouse model', *International Journal of Food Microbiology*, vol 107, no. 1, pp. 83-91.

Baines, SL, Howden, BP, Heffernan, H, Stinear, TP, Carter, GP, Seemann, T, Kwong, JC, Ritchie, SR & Williamson, DA 2016, 'Rapid Emergence and Evolution of Staphylococcus aureus Clones Harboring fusC -Containing Staphylococcal Cassette Chromosome', *Antimicrobial Agents and Chemotherapy*, vol 60, no. 4, pp. 2359-2365.

Balding, DJ 2006, 'A tutorial on statistical methods for population association studies', *Nature Reviews Genetics*, vol 7, no. 10, pp. 781-791.

Barnard, AML, Green, J & Busby, SJW 2003, 'Transcription regulation by tandem-bound FNR at Escherichia coli promoters', *Journal of Bacteriology*, vol 185, no. 20, pp. 5993-6004.

Barrangou, R, Fremaux, C, Deveau, H, Richards, M, Boyaval, P, Moineau, S, Romero, DA & Horvath, P 2007, 'CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes', *Science*, vol 315, no. MARCH, pp. 1709-1712.

Bartha, I, Carlson, JM, Brumme, CJ, McLaren, PJ, Brumme, ZL, John, M, Haas, DW, Martinez-Picado, J, Dalmau, J, López-Galíndez, C, Casado, C, Rauch, A, Günthard, HF, Bernasconi, E, Vernazza, P, Klimkait, T, Yerly, S, O'Brien, SJ, Listgarten, J, Pfeifer, N, et al. 2013, 'A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control', *eLife*, vol 2, p. e01123.

Bartha, I, McLaren, PJ, Brumme, C, Harrigan, R, Telenti, A & Fellay, J 2017, 'Estimating the Respective Contributions of Human and Viral Genetic Variation to HIV Control', *PLOS Computational Biology*, vol 13, no. 2, p. e1005339.

Bartolini, E, Frigimelica, E, Giovinazzi, S, Galli, G, Shaik, Y, Genco, C, Welsch, JA, Granoff, DM, Grandi, G & Grifantini, R 2006, 'Role of FNR and FNR-regulated, sugar fermentation genes in Neisseria meningitidis infection', *Molecular Microbiology*, vol 60, no. 4, pp. 963-972.

Bax, M, Kuijf, ML, Heikema, AP, van Rijs, W, Bruijns, SCM, Garc´ıa-Vallejo, JJ, Crocker, PR, Jacobs, BC, van Vliet, SJ & van Kooyk, Y 2011, 'Campylobacter jejuni Lipooligosaccharides Modulate Dendritic Cell-Mediated T Cell Polarization in a Sialic Acid Linkage-Dependent Manner', *Infection and Immunity*, vol 79, no. 7, pp. 2681-2689.

Beery, JT, Hugdahl, MB & Doyle, MP 1988, 'Colonization of gastrointestinal tracts of chicks by Campylobacter jejuni', *Applied and Environmental Microbiology*, vol 54, no. 10, pp. 2365-2370.

Benjamini, Y & Hochberg, Y 1995, 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society. Series B (Methodological)*, vol 57, no. 1, pp. 289-300.

Bentley, SD & Parkhill, J 2015, 'Genomic perspectives on the evolution and spread of bacterial pathogens', *Proceedings of the Royal Society B: Biological Sciences*, vol 282, no. 1821.

Bentley, SD, Vernikos, GS, Snyder, LAS, Churcher, C, Arrowsmith, C, Chillingworth, T, Cronin, A, Davis, PH, Holroyd, NE, Jagels, K, Maddison, M, Moule, S, Rabbinowitsch, E, Sharp, S, Unwin, L, Whitehead, S, Quail, MA, Achtman, M, Barrell, B, Saunders, NJ, et al. 2007, 'Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18', *PLoS Genetics*, vol 3, no. 2, pp. 0230-0240.

Berndtson, E, Tivemo, M & Engvall, A 1992, 'Distribution and Numbers of Campylobacter-in Newly Slaughtered Broiler-Chickens and Hens', *International Journal of Food Microbiology*, vol 15, no. 1-2, pp. 45-50.

Besier, S, Ludwig, A, Brade, V & Wichelhaus, TA 2003, 'Molecular analysis of fusidic acid resistance in Staphylococcus aureus.', *Molecular microbiology*, vol 47, no. 2, pp. 463-469.

Beumer, RR, de Vries, J & Rombouts, FM 1992, 'Campylobacter jejuni non-culturable coccoid cells', *International Journal of Food Microbiology*, vol 15, no. 1-2, pp. 153-163.

Biagini, M, Spinsanti, M, De Angelis, G, Tomei, S, Ferlenghi, I, Scarselli, M, Rigat, F, Messuti, N, Biolchi, A, Muzzi, A, Anderloni, G, Brunelli, B, Cartocci, E, Buricchi, F, Tani, C, Stella, M, Moschioni, M, Del Tordello, E, Colaprico, A, Savino, S, et al. 2016, 'Expression of factor H binding protein in meningococcal strains can vary at least 15-fold and is genetically determined.', *Proceedings of the National Academy of Sciences*, vol 113, no. 10, pp. 2714-2719.

Biebl, A, Muendlein, A, Kinz, E, Drexel, H, Kabesch, M, Zenz, W, Elling, R, Müller, C, Keil, T, Lau, S & Simma, B 2015, 'Confirmation of Host Genetic Determinants in the CFH Region and Susceptibility to Meningococcal Disease in a Central European Study Sample.', *The Pediatric infectious disease journal*, vol 34, no. 10, pp. 1115-1117.

Bille, E, Ure, R, Gray, SJ, Kaczmarski, EB, McCarthy, ND, Nassif, X, Maiden, MCJ & Tinsley, CR 2008, 'Association of a bacteriophage with meningococcal disease in young adults.', *PloS one*, vol 3, no. 12, p. e3885.

Bille, E, Zahar, J-R, Perrin, A, Morelle, S, Kriz, P, Jolley, K, Maiden, MCJ, Dervin, C, Nassif, X & Tinsley, CR 2005, 'A chromosomally integrated bacteriophage in invasive meningococci.', *Journal of Experimental Medicine*, vol 201, no. 12, pp. 1905-1913.

Bolger, AM, Lohse, M & Usadel, B 2014, 'Trimmomatic: A flexible trimmer for Illumina sequence data', *Bioinformatics*, vol 30, no. 15, pp. 2114-2120.

Botstein, D, White, RL, Skolnick, M & Davis, RW 1980, 'Construction of a genetic linkage map in man using restriction fragment length polymorphisms.', *American Journal of Human Genetics*, vol 32, no. 3, pp. 314-331.

Bourgain, C, Hoffjan, S, Nicolae, R, Newman, D, Steiner, L, Walker, K, Reynolds, R, Ober, C & McPeek, MS 2003, 'Novel Case-Control Test in a Founder Population Identifies P-Selectin as an Atopy-Susceptibility Locus', *The American Journal of Human Genetics*, vol 73, no. 3, pp. 612-626.

Bradley, P, Gordon, NC, Walker, TM, Dunn, L, Heys, S, Huang, B, Earle, S, Pankhurst, LJ, Anson, L, de Cesare, M, Piazza, P, Votintseva, AA, Golubchik, T, Wilson, DJ, Wyllie, DH, Diel, R, Niemann, S, Feuerriegel, S, Kohl, TA, Ismail, N, et al. 2015, 'Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis', *Nature Communications*, vol 6, p. 10063.

Brathwaite, KJ, Siringan, P, Connerton, PL & Connerton, IF 2015, 'Host adaption to the bacteriophage carrier state of Campylobacter jejuni', *Research in Microbiology*, vol 166, no. 6, pp. 504-515.

Brochet, M, Rusniok, C, Couvé, E, Dramsi, S, Poyart, C, Trieu-Cuot, P, Kunst, F & Glaser, P 2008, 'Shaping a bacterial genome by large chromosomal replacements, the evolutionary history of Streptococcus agalactiae', *Proceedings of the National Academy of Sciences* , vol 105, no. 41, pp. 15961-15966.

Broman, T, Palmgren, H, Bergström, S, Sellin, M, Waldenström, J, Danielsson-Tham, ML & Olsen, B 2002, 'Campylobacter jejuni in Black-Headed Gulls (Larus ridibundus): Prevalence, Genotypes, and Influence on C. jejuni Epidemiology', *Journal of Clinical Microbiology*, vol 40, no. 12, pp. 4594-4602.

Brown, ED & Wright, GD 2016, 'Antibacterial drug discovery in the resistance era', *Nature*, vol 529, no. 7586, pp. 336-343.

Browning, BL & Browning, SR 2009, 'A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals', *The American Journal of Human Genetics*, vol 84, no. 2, pp. 210-223.

Browning, SR & Browning, BL 2011, 'Population structure can inflate SNP-based heritability estimates', *American Journal of Human Genetics*, vol 89, no. 1, pp. 191-193.

Brynildsrud, O, Bohlin, J, Scheffer, L & Eldholm, V 2016, 'Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary', *Genome Biology*, vol 17, no. 1, p. 238.

Budroni, S, Siena, E, Dunning Hotopp, JC, Seib, KL, Serruto, D, Nofroni, C, Comanducci, M, Riley, DR, Daugherty, SC, Angiuoli, SV, Covacci, A, Pizza, M, Rappuoli, R, Moxon, ER, Tettelin, H & Medini, D 2011, 'Neisseria meningitidis is structured in clades associated with restriction modification systems that modulate homologous recombination.', *Proceedings of the National Academy of Sciences*, vol 108, no. 11, pp. 4494-4499.

Caldwell, MB, Guerry, P, Lee, EC, Burans, JP & Walker, RI 1985, 'Reversible expression of Flagella in Campylobacter jejuni', *Infection and Immunity*, vol 50, no. 3, pp. 941-943.

Camacho, C, Coulouris, G, Avagyan, V, Ma, N, Papadopoulos, J, Bealer, K & Madden, TL 2009, 'BLAST+: architecture and applications', *BMC Bioinformatics*, vol 10, no. 1, p. 421.

Carrillo, CL, Atterbury, RJ, Connerton, PL, Dillon, E, Scott, A & Connerton, IF 2005, 'Bacteriophage Therapy To Reduce Campylobacter jejuni Colonization of Broiler Chickens', *Applied and Environmental Microbiology*, vol 71, no. 11, pp. 6554-6563.

Carroll, KC, Borek, AP, Burger, C, Glanz, B, Bhally, H, Henciak, S & Flayhart, DC 2006, 'Evaluation of the BD Phoenix Automated Microbiology System for Identification and Antimicrobial Susceptibility Testing of Staphylococci and Enterococci', *Journal of Clinical Microbiology*, vol 44, no. 6, pp. 2072-2077.

Carter, NJ 2013, 'Multicomponent Meningococcal Serogroup B Vaccine (4CMenB; Bexsero®): A Review of its Use in Primary and Booster Vaccination', *BioDrugs*, vol 27, no. 3, pp. 263-274.

Carvalho, CM, Gannon, BW, Halfhide, DE, Santos, SB, Hayes, CM, Roe, JM & Azeredo, J 2010, 'The in vivo efficacy of two administration routes of a phage cocktail to reduce numbers of Campylobacter coli and Campylobacter jejuni in chickens', *BMC Microbiology*, vol 10, no. 1, p. 232.

Caugant, DA & Maiden, MCJ 2009, 'Meningococcal carriage and disease-Population biology and evolution', *Vaccine*, vol 27, no. SUPPL. 2, pp. B64-B70.

Chapman, SJ & Hill, AVS 2012, 'Human genetic susceptibility to infectious disease', *Nature Reviews Genetics*, vol 13, no. 3, pp. 175-188.

Chen, L, Mathema, B, Pitout, JDD, Deleo, FR & Kreiswirth, N 2014, 'Epidemic Klebsiella pneumoniae ST258 Is a Hybrid Strain', *mBio*, vol 5, no. 3, pp. e01355-14.

Chen, PE & Shapiro, BJ 2015, 'The advent of genome-wide association studies for bacteria', *Current Opinion in Microbiology*, vol 25, pp. 17-24.

Cheng, L, Connor, TR, Sirén, J, Aanensen, DM & Corander, J 2013, 'Hierarchical and Spatially Explicit Clustering of DNA Sequences with BAPS Software', *Molecular Biology and Evolution*, vol 30, no. 5, pp. 1224-1228.

Chewapreecha, C, Marttinen, P, Croucher, NJ, Salter, SJ, Harris, SR, Mather, AE, Hanage, WP, Goldblatt, D, Nosten, FH, Turner, C, Turner, P, Bentley, SD & Parkhill, J 2014, 'Comprehensive Identification of Single Nucleotide Polymorphisms Associated with Beta-lactam Resistance within Pneumococcal Mosaic Genes', *PLoS Genetics*, vol 10, no. 8, p. e1004547.

Chin, C-S, Alexander, DH, Marks, P, Klammer, AA, Drake, J, Heiner, C, Clum, A, Copeland, A, Huddleston, J, Eichler, EE, Turner, SW & Korlach, J 2013, 'Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data', *Nature Methods*, vol 10, no. 6, pp. 563-569.

Christie, PJ & Dubnau, D 2005, 'The Ins and Outs of DNA Transfer in Bacteria', *Science*, vol 210, no. 5753, pp. 1456-1460.

Claus, H, Maiden, MCJ, Maag, R, Frosch, M & Vogel, U 2002, 'Many carried meningococci lack the genes required for capsule synthesis and transport', *Microbiology*, vol 148, no. 6, pp. 1813-1819.

Claus, H, Maiden, MCJ, Wilson, D, J McCarthy, ND, Jolley, K, a Urwin, R, Hessler, F, Frosch, M & Vogel, U 2005, 'Genetic analysis of meningococci carried by children and

young adults.', *The Journal of Infectious Diseases*, vol 191, no. 8, pp. 1263-1271.
Cody, AJ, Mccarthy, ND, Bray, JE, Wimalarathna, H, L, M, Colles, FM, Jansen van Rensburg, MJ, Dingle, KE, Waldenström, J & Maiden, MCJ 2015, 'Wild bird-associated Campylobacter jejuni isolates are a consistent source of human disease, in Oxfordshire, United Kingdom', *Environmental Microbiology Reports*, vol 7, no. 5, pp. 782-788.
Colles, FM, Dingle, KE, Cody, AJ & Maiden, MCJ 2008, 'Comparison of Campylobacter populations in wild geese with those in starlings and free-range poultry on the same farm', *Applied and Environmental Microbiology*, vol 74, no. 11, pp. 3583-3590.
Colles, FM, Jones, K, Harding, RM & Maiden, MCJ 2003, 'Genetic Diversity of Campylobacter jejuni Isolates from Farm Animals and the Farm Environment', *Applied and Environmental Microbiology*, vol 69, no. 12, pp. 7409-7413.
Collins, C & Didelot, X 2017, 'A Phylogenetic Method To Perform Genome-Wide Association Studies In Microbes That Accounts For Population Structure And Recombination', *bioRxiv*.
Connerton, PL, Carrillo, CML, Swift, C, Dillon, E, Rees, CED, Dodd, CER, Frost, J, Scott, A & Connerton, IF 2004, 'Longitudinal study of Campylobacter jejuni bacteriophages and their hosts from broiler chickens', *Applied and Environmental Microbiology*, vol 70, no. 7, pp. 3877-3883.
Corander, J, Marttinen, P, Sirén, J & Tang, J 2008, 'Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations', *BMC Bioinformatics*, vol 9, no. 1, p. 539.
Corander, J, Waldmann, P & Sillanpää, MJ 2003, 'Bayesian Analysis of Genetic Differentiation Between Populations', *Genetics*, vol 163, no. 1, pp. 367-374.
Corder, EH, Saunders, AM, Strittmatter, WJ, Schmechel, DE, Gaskell, PC, Small, GW, Roses, AD, Haines, JL & Pericak-Vance, MA 1993, 'Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families', *Science*, vol 261, no. 5123, pp. 921-923.
Correia, FF, Inouye, S & Inouye, M 1988, 'A family of small repeated elements with some transposon-like properties in the genome of Neisseria gonorrhoeae', *Journal of Biological Chemistry*, vol 263, no. 25, pp. 12194-12198.
Coward, C, Grant, AJ, Swift, C, Philp, J, Towler, R, Heydarian, M, Frost, JA & Maskell, DJ 2006, 'Phase-variable surface structures are required for infection of Campylobacter jejuni by bacteriophages', *Applied and Environmental Microbiology*, vol 72, no. 7, pp. 4638-4647.
Cox, A, Dunning, AM, Garcia-Closas, M, Balasubramanian, S, Reed, MWR, Pooley, KA, Scollen, S, Baynes, C, Ponder, BAJ, Chanock, S, Lissowska, J, Brinton, L, Peplonska, B, Southey, MC, Hopper, JL, McCredie, MRE, Giles, GG, Fletcher, O, Johnson, N, dos Santos Silva, I, et al. 2007, 'A common coding variant in CASP8 is associated with breast cancer risk', *Nature Genetics*, vol 39, no. 3, pp. 352-358.
Crawford, L, Zeng, P, Mukherjee, S & Zhou, X 2017, 'Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits', *PLOS Genetics*, vol 13, no. 7, p. e1006869.
Cui, Y, Yang, X, Didelot, X, Guo, C, Li, D, Yan, Y, Zhang, Y, Yuan, Y, Yang, H, Wang, J, Wang, J, Song, Y, Zhou, D, Falush, D & Yang, R 2015, 'Epidemic Clones, Oceanic Gene Pools, and Eco-LD in the Free Living Marine Pathogen Vibrio parahaemolyticus', *Molecular Biology and Evolution*, vol 32, no. 6, pp. 1396-1410.
Curtis, J, Luo, Y, Zenner, HL, Cuchet-Lourenco, D, Wu, C, Lo, K, Maes, M, Alisaac, A, Stebbings, E, Liu, JZ, Kopanitsa, L, Ignatyeva, O, Balabanova, Y, Nikolayevskyy, V, Baessmann, I, Thye, T, Meyer, CG, Nurnberg, P, Horstmann, RD, Drobniewski, F, et al. 2015, 'Susceptibility to tuberculosis is associated with variants in the ASAP1 gene encoding a regulator of dendritic cell migration', *Nature Genetics*, vol 47, no. 5, pp.

523-527.
Daly, MJ, Rioux, JD, Schaffner, SF, Hudson, TJ & Lander, ES 2001, 'High-resolution haplotype structure in the human genome', *Nature Genetics*, vol 29, no. 2, pp. 229-232.
Davies, J & Davies, D 2010, 'Origins and Evolution of Antibiotic Resistance', *Microbiology and Molecular Biology Reviews*, vol 74, no. 3, pp. 417-433.
Davila, S, Wright, VJ, Khor, CC, Sim, KS, Binder, A, Breunis, WB, Inwald, D, Nadel, S, Betts, H, Carrol, ED, de Groot, R, Hermans, PW, Hazelzet, J, Emonts, M, Lim, CC, Kuijpers, TW, Martinon-Torres, F, Salas, A, Zenz, W, Levin, M, et al. 2010, 'Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease', *Nature Genetics*, vol 42, no. 9, pp. 772-776.
Dawn Teare, M & Barrett, JH 2005, 'Genetic linkage studies', *The Lancet*, vol 366, no. 9490, pp. 1036-1044.
Dearlove, BL, Cody, AJ, Pascoe, B, Méric, G, Wilson, DJ & Sheppard, SK 2016, 'Rapid host switching in generalist Campylobacter strains erodes the signal for tracing human infections', *The ISME journal*, vol 10, no. 3, pp. 721-729.
Delihas, N 2011, 'Impact of Small Repeat Sequences on Bacterial Genome Evolution', *Genome Biology and Evolution*, vol 3, pp. 959-973.
Densen, P 1989, 'Interaction of complement with Neisseria meningitidis and Neisseria gonorrhoeae', *Clinical Microbiology Reviews*, vol 2, no. SUPPL., pp. 11-17.
Desjardins, CA, Cohen, KA, Munsamy, V, Abeel, T, Maharaj, K, Walker, BJ, Shea, TP, Almeida, DV, Manson, AL, Salazar, A, Padayatchi, N, O'Donnell, MR, Mlisana, KP, Wortman, J, Birren, BW, Grosset, J, Earl, AM & Pym, AS 2016, 'Genomic and functional analyses of Mycobacterium tuberculosis strains implicate ald in D-cycloserine resistance', *Nature Genetics*, vol 48, no. 5, pp. 544-551.
Devlin, B & Roeder, K 1999, 'Genomic Control for Association Studies', *Biometrics*, vol 55, no. 4, pp. 994-1004.
Diard, M & Hardt, W-D 2017, 'Evolution of bacterial virulence', *FEMS Microbiology Reviews*.
Didelot, X & Wilson, DJ 2015, 'ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes', *PLOS Computational Biology*, vol 11, p. e1004041.
Didelot, X, Bowden, R, Wilson, DJ, Peto, TEA & Crook, DW 2012, 'Transforming clinical microbiology with bacterial genome sequencing', *Nature Reviews Genetics*, vol 13, no. 9, pp. 601-612.
Didelot, X, Walker, AS, Peto, TE, Crook, DW & Wilson, DJ 2016, 'Within-host evolution of bacterial pathogens', *Nature Reviews Microbiology*, vol 14, no. 3, pp. 150-162.
Dingle, KE, Colles, FM, Wareing, DRAA, Ure, R, Fox, AJ, Bolton, FE, Bootsma, HJ, Willems, RJL, Urwin, R & Maiden, MCJ 2001, 'Multilocus Sequence Typing System for Campylobacter jejuni', *Journal of clinical microbiology*, vol 39, no. 1, pp. 14-23.
Dobie, D & Gray, J 2004, 'Fusidic acid resistance in Staphylococcus aureus', *Archives of Disease in Childhood*, vol 89, no. 1, pp. 74-77.
Donati, C, Hiller, NL, Tettelin, H, Muzzi, A, Croucher, NJ, Angiuoli, SV, Oggioni, M, Dunning Hotopp, JC, Hu, FZ, Riley, DR, Covacci, A, Mitchell, TJ, Bentley, SD, Kilian, M, Ehrlich, GD, Rappuoli, R, Moxon, ER & Masignani, V 2010, 'Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species', *Genome Biology*, vol 11, no. 10, p. R107.
Dong, L, Potter, J, White, E, Ulrich, C, Cardon, L & Peters, U 2008, 'Genetic susceptibility to cancer: The role of polymorphisms in candidate genes', *JAMA*, vol 299, no. 20, pp. 2423-2436.
Dorman, SE & Chaisson, RE 2007, 'From magic bullets back to the magic mountain: the rise of extensively drug-resistant tuberculosis.', *Nature Medicine*, vol 13, no. 3, pp. 295-298.
Dos Vultos, T, Mestre, O, Rauzier, J, Golec, M, Rastogi, N, Rasolofo, V, Tonjum, T, Sola, C, Matic, I & Gicquel, B 2008, 'Evolution and diversity of clonal bacteria: The paradigm

of Mycobacterium tuberculosis', *PLoS ONE*, vol 3, no. 2, pp. 1-10.
Doyle, MP & Erickson, MC 2006, 'Reducing the carriage of foodborne pathogens in livestock and poultry.', *Poultry Science*, vol 85, no. 6, pp. 960-973.
Draghici, S, Khatri, P, Eklund, AC & Szallasi, Z 2006, 'Reliability and reproducibility issues in DNA microarray measurements', *Trends in Genetics*, vol 22, no. 2, pp. 101-109.
Drobniewski, F, Nikolayevskyy, V, Maxeiner, H, Balabanova, Y, Casali, N, Kontsevaya, I & Ignatyeva, O 2013, 'Rapid diagnostics of tuberculosis and drug resistance in the industrialized world: clinical and public health benefits and barriers to implementation', *BMC Medicine*, vol 11, no. 1, p. 190.
Dugar, G, Herbig, A, Förstner, KU, Heidrich, N, Reinhardt, R, Nieselt, K & Sharma, CM 2013, 'High-Resolution Transcriptome Maps Reveal Strain-Specific Regulatory Features of Multiple Campylobacter jejuni Isolates', *PLOS Genetics*, vol 9, no. 5.
Dunn, OJ 1959, 'Estimation of the Medians for Dependent Variables', *The Annals of Mathematical Statistics*, vol 30, pp. 192-197.
Dunning Hotopp, JC, Grifantini, R, Kumar, N, Tzeng, YL, Fouts, D, Frigimelica, E, Draghi, M, Giuliani, MM, Rappuoli, R, Stephens, DS, Grandi, G & Tettelin, H 2006, 'Comparative genomics of Neisseria meningitidis: Core genome, islands of horizontal transfer and pathogen-specific genes', *Microbiology*, vol 152, no. 12, pp. 3733-3749.
Edwards, RA, McNair, K, Faust, K, Raes, J & Dutilh, BE 2016, 'Computational approaches to predict bacteriophage–host relationships', *FEMS Microbiology Reviews*, vol 40, no. 2, pp. 258-272.
Eid, J, Fehr, A, Gray, J, Luong, K, Lyle, J, Otto, G, Peluso, P, Rank, D, Baybayan, P, Bettman, B, Bibillo, A, Bjornson, K, Chaudhuri, B, Christians, F, Cicero, R, Clark, S, Dalal, R, DeWinter, A, Dixon, J, Foquet, M, et al. 2009, 'Real-Time DNA Sequencing from Single Polymerase Molecules', *Science*, vol 323, no. 5910, pp. 133-138.
Eiglmeier, K, Honoré, N, Iuchi, S, Lin, ECC & Cole, ST 1989, 'Molecular genetic analysis of FNR-dependent promoters', *Molecular Microbiology*, vol 3, no. 7, pp. 869-878.
El-Shibiny, A, Connerton, PL & Connerton, IF 2005, 'Enumeration and diversity of campylobacters and bacteriophages isolated during the rearing cycles of free-range and organic chickens', *Applied and Environmental Microbiology*, vol 71, no. 3, pp. 1259-1266.
El-Shibiny, A, Scott, A, Timms, A, Metawea, Y, Connerton, P & Connerton, I 2009, 'Application of a group II Campylobacter bacteriophage to reduce strains of Campylobacter jejuni and Campylobacter coli colonizing broiler chickens', *Journal of Food Protection*, vol 72, no. 4, pp. 733-740.
Everitt, RG, Didelot, X, Batty, EM, Miller, RR, Knox, K, Young, BC, Bowden, R, Auton, A, Votintseva, A, Larner-Svensson, H, Charlesworth, J, Golubchik, T, Ip, CLC, Godwin, H, Fung, R, Peto, TEA, Walker, AS, Crook, DW & Wilson, DJ 2014, 'Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus', *Nature Communications*, vol 5, p. 3956.
Eyre, DW, Golubchik, T, Gordon, NC, Bowden, R, Piazza, P, Batty, EM, Ip, CLC, Wilson, DJ, Didelot, X, O'Connor, L, Lay, R, Buck, D, Kearns, AM, Shaw, A, Paul, J, Wilcox, MH, Donnelly, PJ, Peto, TEA, Walker, AS & Crook, DW 2012, 'A pilot study of rapid benchtop sequencing of Staphylococcus aureus and Clostridium difficile for outbreak detection and surveillance.', *BMJ open*, vol 2, no. 3, p. e001124.
Fadista, J, Manning, AK, Florez, JC & Groop, L 2016, 'The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants', *European Journal of Human Genetics*, vol 24, no. 8, pp. 1202-1205.
Falkow, S 1988, 'Molecular Koch's Postulates Applied to Microbial Pathogenicity', *Reviews of Infectious Diseases*, vol 10, no. 2, pp. S274-S276.
Falush, D & Bowden, R 2006, 'Genome-wide association mapping in bacteria?', *Trends in Microbiology*, vol 14, no. 8, pp. 353-355.

Falush, D 2016, 'Bacterial genomics: Microbial GWAS coming of age', *Nature Microbiology*, vol 1, p. 16059.
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, et al. 2013, 'Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis', *Nature Genetics*, vol 45, no. 10, pp. 1183-1189.
Farhat, MR, Shapiro, BJ, Sheppard, SK, Colijn, C & Murray, M 2014, 'A phylogeny-based sampling strategy and power calculator informs genome-wide associations study design for microbial pathogens', *Genome Medicine*, vol 6, no. 11, p. 101.
Feil, EJ, Maiden, MC, Achtman, M & Spratt, BG 1999, 'The relative contributions of recombination and mutation to the divergence of clones of Neisseria meningitidis.', *Molecular Biology and Evolution*, vol 16, no. 11, pp. 1496-1502.
Flanagan, RC, Neal-McKinney, JM, Dhillon, AS, Miller, WG & Konkel, ME 2009, 'Examination of Campylobacter jejuni putative adhesins leads to the identification of a new protein, designated FlpA, required for chicken colonization', *Infection and Immunity*, vol 77, no. 6, pp. 2399-2407.
Fletcher, LD, Bernfield, L, Barniak, V, Farley, JE, Howell, A, Knauf, M, Ooi, P, Smith, RP, Weise, P, Wetherell, M, Xie, X, Zagursky, R, Zhang, Y & Zlotnick, GW 2004, 'Vaccine Potential of the Neisseria meningitidis 2086 Lipoprotein', *Infection and Immunity*, vol 72, no. 4, pp. 2088-2100.
French, N, Barrigas, M, Brown, P, Ribiero, P, Williams, N, Leatherbarrow, H, Birtles, R, Bolton, E, Fearnhead, P & Fox, A 2005, 'Spatial epidemiology and natural population structure of Campylobacter jejuni colonizing a farmland ecosystem', *Environmental Microbiology*, vol 7, no. 8, pp. 1116-1126.
Friedman, C, Neiman, J, Wegener, H & Tauxe, R 2000, 'Epidemiology of Campylobacter jejuni infections in the United States and other industrialised nations.', in *Campylobacter*, 2nd edn, ASM International, Washington, USA.
Friedman, CR, Hoekstra, RM, Samuel, M, Marcus, R, Bender, J, Shiferaw, B, Reddy, S, Ahuja, SD, Helfrick, DL, Hardnett, F, Carter, M, Anderson, B & Tauxe, RV 2004, 'Risk factors for sporadic Campylobacter infection in the United States: A case-control study in FoodNet sites.', *Clinical Infectious Diseases*, vol 38, no. 3, pp. S285-S296.
Fuglsang, A 2003, 'The genome of Campylobacter jejuni: codon and amino acid usage.', *APMIS: acta pathologica, microbiologica, et immunologica Scandinavica*, vol 111, no. 6, pp. 605-618.
Gabriel, SB, Schaffner, SF, Nguyen, H, Moore, JM, Roy, J, Blumenstiel, B, Higgins, J, Defelice, M, Lochner, A, Faggart, M, Liu-cordero, SN, Rotimi, C, Adeyemo, A, Cooper, R, Ward, R, Lander, ES, Daly, MJ & Altshuler, D 2002, 'The Structure of Haplotype Blocks in the Human Genome', *Science*, vol 296, no. 5576, pp. 2225-2229.
Ganesh, K, Allam, M, Wolter, N, Bratcher, HB, Harrison, OB, Lucidarme, J, Borrow, R, de Gouveia, L, Meiring, S, Birkhead, M, Maiden, MCJ, von Gottberg, A & du Plessis, M 2017, 'Molecular characterization of invasive capsule null Neisseria meningitidis in South Africa', *BMC Microbiology*, vol 17, no. 1, p. 40.
Gao, Y-G, Selmer, M, Dunham, CM, Weixlbaumer, A, Kelley, AC & Ramakrishnan, V 2009, 'The Structure of the Ribosome with Elongation Factor G Trapped in the Posttranslocational State', *Science*, vol 326, no. 5953, pp. 694-699.
Gardete, S & Tomasz, A 2014, 'Mechanisms of vancomycin resistance in Staphylococcus aureus', *The Journal of Clinical Investigation*, vol 124, no. 7, pp. 2836-2840.
Gasiunas, G, Barrangou, R, Horvath, P & Siksnys, V 2012, 'Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria.', *Proceedings of the National Academy of Sciences*, vol 109, no. 39, pp. E2579-E2586.

Gegner, JA, Graham, DR, Roth, AF & Dahlquist, FW 1992, 'Assembly of an MCP receptor, CheW, and kinase CheA complex in the bacterial chemotaxis signal transduction pathway', *Cell*, vol 70, no. 6, pp. 975-982.

Gerdes, S, Scholle, MD, Campbell, JW, Balázsi, G, Ravasz, E, Daugherty, MD, Somera, AL, Kyrpides, NC, Anderson, I, Gelfand, MS, Bhattacharya, A, Kapatral, V, D'Souza, M, Baev, MV, Grechkin, Y, Mseeh, F, Fonstein, M, Overbeek, R, Barabási, A-L, Oltvai, ZN, et al. 2003, 'Experimental determination and system level analysis of essential genes in Escherichia coli MG1655', *Journal of Bacteriology*, vol 185, no. 19, pp. 5673-5684.

Gillespie, IA, O'Brien, SJ, Frost, JA, Adak, GK, Horby, P, Swan, AV, Painter, MJ & Neal, KR 2002, 'A case-case comparison of Campylobacter coli and Campylobacter jejuni infection: A tool for generating hypotheses', *Emerging Infectious Diseases*, vol 8, no. 9, pp. 937-942.

Goodman, SD & Scocca, JJ 1988, 'Identification and arrangement of the DNA sequence recognized in specific transformation of Neisseria gonorrhoeae.', *Proceedings of the National Academy of Sciences*, vol 85, no. 18, pp. 6982-6986.

Gordon, NC, Price, JR, Cole, K, Everitt, R, Morgan, M, Finney, J, Kearns, AM, Pichon, B, Young, B, Wilson, DJ, Llewelyn, MJ, Paul, J, Peto, TEA, Crook, DW, Walker, AS & Golubchik, T 2014, 'Prediction of staphylococcus aureus antimicrobial resistance by whole-genome sequencing', *Journal of Clinical Microbiology*, vol 52, no. 4, pp. 1182-1191.

Griekspoor, P, Colles, FM, McCarthy, ND, Hansbro, PM, Ashhurst-Smith, C, Olsen, B, Hasselquist, D, Maiden, MCJ & Waldenström, J 2013, 'Marked host specificity and lack of phylogeographic population structure of Campylobacter jejuni in wild birds', *Molecular Ecology*, vol 22, no. 5, pp. 1463-1472.

Griekspoor, P, Engvall, EO, Olsen, B & Waldenström, J 2010, 'Multilocus sequence typing of Campylobacter jejuni from broilers', *Veterinary Microbiology*, vol 140, no. 1-2, pp. 180-185.

Gripp, E, Hlahla, D, Didelot, X, Kops, F, Maurischat, S, Tedin, K, Alter, T, Ellerbroek, L, Schreiber, K, Schomburg, D, Janssen, T, Bartholomäus, P, Hofreuter, D, Woltemate, S, Uhr, M, Brenneke, B, Grüning, P, Gerlach, G, Wieler, L, Suerbaum, S, et al. 2011, 'Closely related Campylobacter jejuni strains from different sources reveal a generalist rather than a specialist lifestyle', *BMC Genomics*, vol 12, no. 1, p. 584.

Guerry, P, Szymanski, CM, Prendergast, MM, Hickey, TE, Ewing, CP, Pattarini, DL & Moran, AP 2002, 'Phase Variation of Campylobacter jejuni 81-176 Lipooligosaccharide Affects Ganglioside Mimicry and Invasiveness In Vitro', *Infection and Immunity*, vol 70, no. 2, pp. 787-793.

Guest, JR, Green, J, Irvine, AS & Spiro, S 1996, 'The FNR modulon and FNR-regulated gene expression', in *Regulation of gene expression in Escherichia coli*, Springer US.

Gusella, JF, Wexler, NS, Conneally, PM, Naylor, SL, Anderson, MA, Tanzi, RE, Watkins, PC, Ottina, K, Wallace, MR, Sakaguchi, AY, Young, AB, Shoulson, I, Bonilla, E & Martin, JB 1983, 'A polymorphic DNA marker genetically linked to Huntington's disease', *Nature*, vol 306, no. 5940, pp. 234-238.

Hale, CR, Zhao, P, Olson, S, Duff, MO, Graveley, BR, Wells, L, Terns, RM & Terns, MP 2009, 'RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex', *Cell*, vol 139, no. 5, pp. 945-956.

Hall, BG 2014, 'SNP-Associations and Phenotype Predictions from Hundreds of Microbial Genomes without Genome Alignments', *PLOS ONE*, vol 9, no. 2, p. e90490.

Halperin, SA, Bettinger, JA, Greenwood, B, Harrison, LH, Jelfs, J, Ladhani, SN, McIntyre, P, Ramsay, ME & Sáfadi, MAP 2012, 'The changing and dynamic epidemiology of meningococcal disease', *Vaccine*, vol 30, no. SUPPL. 2, pp. B26-B36.

Hammerschmidt, S, Müller, A, Sillmann, H, Mühlenhoff, M, Borrow, R, Fox, A, Van Putten,

J, Zollinger, WD, Gerardy-Schahn, R & Frosch, M 1996, 'Capsule phase variation in Neisseria meningitidis serogroup B by slipped-strand mispairing in the polysialyltransferase gene (siaD): Correlation with bacterial invasion and the outbreak of meningococcal disease', *Molecular Microbiology*, vol 20, no. 6, pp. 1211-1220.

Harley, CB & Reynolds, RP 1987, 'Analysis of E. coli promoter sequences.', *Nucleic Acids Research*, vol 15, no. 5, pp. 2343-2361.

Harrison, OB, Claus, H, Jiang, Y, Bennett, JS, Bratcher, HB, Jolley, KA, Corton, C, Care, R, Poolman, JT, Zollinger, WD, Frasch, CE, Stephens, DS, Feavers, I, Frosch, M, Parkhill, J, Vogel, U, Quail, MA, Bentley, SD & Maiden, MCJ 2013, 'Description and nomenclature of Neisseria meningitidis capsule locus', *Emerging Infectious Diseases*, vol 19, no. 4, pp. 566-573.

Hazeleger, W, Janse, J, Koenraad, P, Beumer, R, Rombouts, F & Abee, T 1995, 'Temperature-dependent membrane fatty acid and cell physiology changes in coccoid forms of Campylobacter jejuni', *Applied and Environmental Microbiology*, vol 61, no. 7, pp. 2713-2719.

Hedge, J & Wilson, DJ 2014, 'Bacterial Phylogenetic Reconstruction from Whole Genomes Is Robust to Recombination but Demographic Inference Is Not', *mBio*, vol 5, no. 6, pp. e02158-14.

Henderson, IR, Owen, P & Nataro, JP 1999, 'Molecular switches - the ON and OFF of bacterial phase variation', *Molecular Microbiology*, vol 33, no. 5, pp. 919-932.

Hendrixson, DR & DiRita, VJ 2004, 'Identification of Campylobacter jejuni genes involved in commensal colonization of the chick gastrointestinal tract', *Molecular Microbiology*, vol 52, no. 2, pp. 471-484.

Hepworth, PJ, Ashelford, KE, Hinds, J, Gould, KA, Witney, AA, Williams, NJ, Leatherbarrow, H, French, NP, Birtles, RJ, Mendonca, C, Dorrell, N, Wren, BW, Wigley, P, Hall, N & Winstanley, C 2011, 'Genomic variations define divergence of water/wildlife-associated Campylobacter jejuni niche specialists from common clonal complexes', *Environmental Microbiology*, vol 13, no. 6, pp. 1549-1560.

Hermans, D, Van Deun, K, Martel, A, Van Immerseel, F, Messens, W, Heyndrickx, M, Haesebrouck, F & Pasmans, F 2011, 'Colonization factors of Campylobacter jejuni in the chicken gut', *Veterinary Research*, vol 42, no. 1, p. 82.

Hirschhorn, JN, Lohmueller, K, Byrne, E & Hirschhorn, K 2002, 'A comprehensive review of genetic association studies', *Genetics in Medicine*, vol 4, no. 2, pp. 45-61.

Hoang, LMN, Thomas, E, Tyler, S, Pollard, AJ, Stephens, G, Gustafson, L, McNabb, A, Pocock, I, Tsang, R & Tan, R 2005, 'Rapid and Fatal Meningococcal Disease Due to a Strain of Neisseria meningitidis Containing the Capsule Null Locus', *Clinical Infectious Diseases*, vol 40, no. 5, pp. e38-e42.

Hobb, RI, L., TY, Choudhury, BP, Carlson, RW & Stephens, DS 2010, 'Requirement of NMB0065 for connecting assembly and export of sialic acid capsular polysaccharides in Neisseria meningitidis', *Microbes and Infection*, vol 12, no. 6, pp. 476-487.

Holden, MTG, Feil, EJ, Lindsay, JA, Peacock, SJ, Day, NPJ, Enright, MC, Foster, TJ, Moore, CE, Hurst, L, Atkin, R, Barron, A, Bason, N, Bentley, SD, Chillingworth, C, Chillingworth, T, Churcher, C, Clark, L, Corton, C, Cronin, A, Doggett, J, et al. 2004, 'Complete genomes of two clinical Staphylococcus aureus strains: evidence for the rapid evolution of virulence and drug resistance.', *Proceedings of the National Academy of Sciences*, vol 101, no. 26, pp. 9786-9791.

Holt, KE, Wertheim, H, Zadoks, RN, Baker, S, Whitehouse, CA, Dance, D, Jenney, A, Connor, TR, Hsu, LY, Severin, J, Brisse, S, Cao, H, Wilksch, J, Gorrie, C, Schultz, MB, Edwards, DJ, Nguyen, KV, Nguyen, TV, Dao, TT, Mensink, M, et al. 2015, 'Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae , an urgent threat to public health', *Proceedings of*

the National Academy of Sciences, vol 112, no. 27, pp. E3574-E3581.
Horvath, P & Barrangou, R 2010, 'CRISPR/Cas, the immune system of bacteria and archaea.', *Science*, vol 327, no. 5962, pp. 167-170.
Howie, B, Fuchsberger, C, Stephens, M, Marchini, J & Abecasis, GR 2012, 'Fast and accurate genotype imputation in genome-wide association studies through pre-phasing', *Nature Genetics*, vol 44, no. 8, pp. 955-959.
Howie, BN, Donnelly, P & Marchini, J 2009, 'A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies', *PLOS Genetics*, vol 5, no. 6, p. e1000529.
Hugdahl, MB, Beery, JT & Doyle, MP 1988, 'Chemotactic behavior of Campylobacter jejuni', *Infection and Immunity*, vol 56, no. 6, pp. 1560-1566.
Hyatt, D, Chen, G-L, Locascio, PF, Land, ML, Larimer, FW & Hauser, LJ 2010, 'Prodigal: prokaryotic gene recognition and translation initiation site identification.', *BMC Bioinformatics*, vol 11, p. 119.
International Human Genome Sequencing Consortium 2001, 'Initial sequencing and analysis of the human genome', *Nature*, vol 409, no. 6822, pp. 860-921.
Ioannidis, JPA, Ntzani, EE, Trikalinos, TA & Contopoulos-ioannidis, DG 2001, 'Replication validity of genetic association studies', *Nature Genetics*, vol 29, no. 3, pp. 306-309.
Isenberg, HD 1988, 'Pathogenicity and Virulence : Another View', *Clinical Microbiology Reviews*, vol 1, no. 1, pp. 40-53.
Iwata, T, Chiku, K, Amano, KI, Kusumoto, M, Ohnishi-Kameyama, M, Ono, H & Akiba, M 2013, 'Effects of Lipooligosaccharide Inner Core Truncation on Bile Resistance and Chick Colonization by Campylobacter jejuni', *PLoS ONE*, vol 8, no. 2.
Jackson, RW, Vinatzer, B, Arnold, DL, Dorus, S & Murillo, J 2011, 'The influence of the accessory genome on bacterial pathogen evolution', *Mobile Genetic Elements*, vol 1, no. 1, pp. 55-65.
Jacoby, GA & Munoz-Price, LS 2005, 'The New β-Lactamases', *New England Journal of Medicine*, vol 352, no. 4, pp. 380-391.
Jain, M, Fiddes, IT, Miga, KH, Olsen, HE, Paten, B & Akeson, M 2015, 'Improved data analysis for the MinION nanopore sequencer', *Nature Methods*, vol 12, no. 4, pp. 351-356.
Jarvis, GA & Vedros, NA 1991, 'Sialic acid of group B Neisseria meningitidis regulates alternative complement pathway activation', *Infection and Immunity.*, vol 55, no. 1, pp. 174-180.
Javed, MA, Cawthraw, SA, Baig, A, Li, J, McNally, A, Oldfield, NJ, Newell, DG & Manning, G 2012, 'Cj1136 is required for lipooligosaccharide biosynthesis, hyperinvasion, and chick colonization by Campylobacter jejuni', *Infection and Immunity*, vol 80, no. 7, pp. 2361-2370.
Johnson, GCL, Esposito, L, Barratt, BJ, Smith, AN, Heward, J, Di, G, Ueda, H, Cordell, HJ, Eaves, IA, Dudbridge, F, Twells, RCJ, Hughes, W, Nutland, S, Stevens, H, Carr, P, Tuomilehto-wolf, E, Gough, SCL, Clayton, DG & Todd, JA 2001, 'Haplotype tagging for the identification of common disease genes', *Nature Genetics*, vol 29, no. 2, pp. 233-237.
Jolley, KA, Kalmusova, J, Feil, EJ, Gupta, S, Musilek, M, Kriz, P & Maiden, MCJ 2000, 'Carried meningococci in the Czech Republic: A diverse recombining population', *Journal of Clinical Microbiology*, vol 38, no. 12, pp. 4492-4498.
Jones, K 2001, 'Campylobacters in water, sewage and the environment', *Journal of Applied Microbiology*, vol 90, pp. 68S-79S.
Joseph, B, Schwarz, RF, Linke, B, Blom, J, Becker, A, Claus, H, Goesmann, A, Frosch, M, Müller, T, Vogel, U & Schoen, C 2011, 'Virulence evolution of the human pathogen neisseria meningitidis by recombination in the core and accessory genome', *PLoS ONE*, vol 6, no. 4.

Judge, K, Harris, SR, Reuter, S, Parkhill, J & Peacock, SJ 2015, 'Early insights into the potential of the Oxford Nanopore MinION for the detection of antimicrobial resistance genes', *Journal of Antimicrobial Chemotherapy*, vol 70, no. 10, pp. 2775-2778.
Kaatz, GW & Seo, SM 1997, 'Mechanisms of Fluoroquinolone Resistance in Genetically Related Strains of Staphylococcus aureus', *Antimicrobial Agents and Chemotherapy*, vol 41, no. 12, pp. 2733-2737.
Kang, HM, Sul, JH, Service, SK, Zaitlen, NA, Kong, SY, Freimer, NB, Sabatti, C & Eskin, E 2010, 'Variance component model to account for sample structure in genome-wide association studies', *Nature Genetics*, vol 42, no. 4, pp. 348-354.
Kang, HM, Zaitlen, NA, Wade, CM, Kirby, A, Heckerman, D, Daly, MJ & Eskin, E 2008, 'Efficient Control of Population Structure in Model Organism Association Mapping', *Genetics*, vol 178, no. 3, pp. 1709-1723.
Kanipes, MI, Holder, LC, Corcoran, AT, Moran, AP & Guerry, P 2004, 'A Deep-Rough Mutant of Campylobacter jejuni 81-176 Is Noninvasive for Intestinal Epithelial Cells', *Infection and Immunity*, vol 72, no. 4, pp. 2452-2455.
Kanungpean, D, Kakuda, T & Takai, S 2011, 'Participation of CheR and CheB in the chemosensory response of Campylobacter jejuni', *Microbiology*, vol 157, no. 5, pp. 1279-1289.
Kapperud, G & Rosef, O 1983, 'Avian wildlife reservoir of Campylobacter fetus subsp. jejuni, Yersinia spp., and Salmonella spp. in Norway', *Applied and Environmental Microbiology*, vol 45, no. 2, pp. 375-380.
Kapperud, G, Espeland, G, Wahl, E, Walde, A, Herikstad, H, Gustavsen, S, Tveits, I, Natås, O, Bevanger, L & Digranes, A 2003, 'Factors associated with increased and decreased risk of Campylobacter infection: A prospective case-control study in Norway', *American Journal of Epidemiology*, vol 158, no. 3, pp. 234-242.
Kappock, TJ, Ealick, SE & Stubbe, J 2000, 'Modular evolution of the purine biosynthetic pathway', *Current Opinion in Chemical Biology*, vol 4, no. 5, pp. 567-572.
Karesh, WB, Dobson, A, Lloyd-Smith, JO, Lubroth, J, Dixon, MA, Bennett, M, Aldrich, S, Harrington, T, Formenty, P, Loh, EH, MacHalaba, CC, Thomas, MJ & Heymann, DL 2012, 'Ecology of zoonoses: Natural and unnatural histories', *The Lancet*, vol 380, no. 9857, pp. 1936-1945.
Karlyshev, AV, Ketley, JM & Wren, BW 2005, 'The Campylobacter jejuni glycome', *FEMS Microbiology Reviews*, vol 29, no. 2, pp. 377-390.
Karlyshev, AV, Linton, D, Gregson, NA & Wren, BW 2002, 'A novel paralogous gene family involved in phase-variable flagella-mediated motility in Campylobacter jejuni', *Microbiology*, vol 148, no. 2, pp. 473-480.
Katz, LS, Sharma, NV, Harcourt, BH, Thomas, JD, Wang, X, Mayer, LW & Jordan, IK 2011, 'Using single-nucleotide polymorphisms to discriminate disease-associated from carried genomes of Neisseria meningitidis', *Journal of Bacteriology*, vol 193, no. 14, pp. 3633-3641.
Keenan, JD, Klugman, KP, Mcgee, L, Vidal, JE, Chochua, S, Hawkins, P, Cevallos, V, Gebre, T, Tadesse, Z, Emerson, PM, Jorgensen, JH, Gaynor, BD & Lietman, TM 2015, 'Evidence for Clonal Expansion After Antibiotic Selection Pressure: Pneumococcal Multilocus Sequence Types Before and After Mass Azithromycin Treatments', *The Journal of Infectious Diseases*, vol 211, no. 6, pp. 988-994.
Keller, JI & Shriver, WG 2014, 'Prevalence of three campylobacter species, C. jejuni, C. coli, and C. lari, using multilocus sequence typing in wild birds of the Mid-Atlantic region, USA.', *Journal of Wildlife Diseases*, vol 50, no. 1, pp. 31-41.
Kichaev, G, Yang, W-Y, Lindstrom, S, Hormozdiari, F, Eskin, E, Price, AL, Kraft, P & Pasaniuc, B 2014, 'Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies', *PLOS Genetics*, vol 10, no. 10, p. e1004722.
Klein, RJ, Zeiss, C, Chew, EY, Tsai, J-Y, Sackler, RS, Haynes, C, Henning, AK,

Sangiovanni, JP, Mane, SM, Mayne, ST, Bracken, MB, Ferris, FL, Ott, J, Barnstable, C & Hoh, J 2005, 'Complement Factor H Polymorphism in Age-Related Macular Degeneration', *Science*, vol 308, no. 5720, pp. 385-389.

Klughammer, J, Dittrich, M, Blom, J, Mitesser, V, Vogel, U, Frosch, M, Goesmann, A, Mu¨ller, T & Schoen, C 2017, 'Comparative genome sequencing reveals within-host genetic changes in neisseria meningitidis during invasive disease', *PLoS ONE*, vol 12, no. 1, pp. 1-29.

Knowlton, RG, Cohen-Haguenauer, O, Van Cong, N, Frezal, J, Brown, VA, Barker, D, Braman, JC, Schumm, JW, Tsui, L-C, Buchwald, M & Donis-Keller, H 1985, 'A polymorphic DNA marker linked to cystic fibrosis is located on chromosome 7', *Nature*, vol 318, no. 6044, pp. 380-382.

Koren, S & Phillippy, AM 2015, 'One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly', *Current Opinion in Microbiology*, vol 23, pp. 110-120.

Koren, S, Harhay, GP, Smith, TPL, Bono, JL, Harhay, DM, Mcvey, SD, Radune, D, Bergman, NH & Phillippy, AM 2013, 'Reducing assembly complexity of microbial genomes with single-molecule sequencing', *Genome Biology*, vol 14, no. 9, p. R101.

Koren, S, Walenz, BP, Berlin, K, Miller, JR, Bergman, NH & Phillippy, AM 2017, 'Canu : scalable and accurate long-read assembly via adaptive k -mer weighting and repeat separation', *Genome Research*, vol 27, no. 5, pp. 722-736.

Köser, CU, Bryant, JM, Becq, J, Török, ME, Ellington, MJ, Marti-Renom, MA, Carmichael, AJ, Parkhill, J, Smith, GP & Peacock, SJ 2013, 'Whole-Genome Sequencing for Rapid Susceptibility Testing of M. tuberculosis', *New England Journal of Medicine*, vol 369, no. 3, pp. 290-292.

Krízová, P, Musílek, M & Kalmusová, J 1997, 'Development of the epidemiological situation in invasive meningococcal disease in the Czech Republic caused by emerging Neisseria meningitidis clone ET-15/37', *Central European Journal of Public Health*, vol 5, no. 4, pp. 214-218.

Kuijf, ML, Samsom, JN, van Rijs, W, Bax, M, Huizinga, R, Heikema, AP, van Doorn, PA, van Belkum, A, van Kooyk, Y, Burgers, PC, Luider, TM, Endtz, HP, Nieuwenhuis, EES & Jacobs, BC 2010, 'TLR4-mediated sensing of Campylobacter jejuni by dendritic cells is determined by sialylation.', *Journal of Immunology*, vol 185, no. 1, pp. 748-755.

Laabei, M & Massey, R 2016, 'Using functional genomics to decipher the complexity of microbial pathogenicity', *Current Genetics*, vol 62, no. 3, pp. 523-525.

Laabei, M, Recker, M, Rudkin, JK, Aldeljawi, M, Gulay, Z, Sloan, TJ, Williams, P, Endres, JL, Bayles, KW, Fey, PD, Yajjala, VK, Widhelm, T, Hawkins, E, Lewis, K, Parfett, S, Scowen, L, Peacock, SJ, Holden, M, Wilson, D, Read, TD, et al. 2014, 'Predicting the virulence of MRSA from its genome sequence.', *Genome Research*, vol 24, no. 5, pp. 839-49.

Langmead, B & Salzberg, SL 2012, 'Fast gapped-read alignment with Bowtie 2', *Nature Methods*, vol 9, no. 4, pp. 357-359.

Langridge, G, C Fookes, M, Connor, TR, Feltwell, T, Feasey, N, Parsons, BN, Seth-Smith, HMB, Barquist, L, Stedman, A, Humphrey, T, Wigley, P, Peters, SE, Maskell, DJ, Corander, J, Chabalgoity, JA, Barrow, P, Parkhill, J, Dougan, G & Thomson, NR 2015, 'Patterns of genome evolution that have accompanied host adaptation in Salmonella', *Proceedings of the National Academy of Sciences* , vol 112, no. 3, pp. 863-968.

Lapierre, P & Gogarten, JP 2009, 'Estimating the size of the bacterial pan-genome', *Trends in Genetics*, vol 25, no. 3, pp. 107-110.

Lee, SH, Wray, NR, Goddard, ME & Visscher, PM 2011, 'Estimating missing heritability for disease from genome-wide association studies', *American Journal of Human Genetics*, vol 88, no. 3, pp. 294-305.

Lees, JA, Kremer, PHC, Manso, AS, Croucher, NJ, Ferwerda, B, Serón, M, Valls Oggioni,

MR, Parkhill, J, Brouwer, MC, Ende, AVD, Beek, DVD & Bentley, SD 2017, 'Large scale genomic analysis shows no evidence for pathogen adaptation between the blood and cerebrospinal fluid niches during bacterial meningitis', *Microbial Genomics*, vol 3, no. 1.

Lees, JA, Vehkala, M, Välimäki, N, Harris, SR, Chewapreecha, C, Croucher, NJ, Marttinen, P, Davies, MR, Steer, AC, Tong, SYC, Honkela, A, Parkhill, J, Bentley, SD & Corander, J 2016, 'Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes', *Nature Communications*, vol 7, p. 12797.

Lefébure, T, Bitar, PDP, Suzuki, H & Stanhope, MJ 2010, 'Evolutionary dynamics of complete Campylobacter pan-genomes and the bacterial species concept', *Genome Biology and Evolution*, vol 2, no. 1, pp. 646-655.

Levinson, G & Gutman, G 1987, 'Slipped-strand mispairing: a major mechanism for DNA sequence evolution.', *Molecular Biology and Evolution*, vol 4, no. 3, pp. 203-221.

Li, C, Kappock, TJ, Stubbe, J, Weaver, TM & Ealick, SE 1999, 'X-ray crystal structure of aminoimidazole ribonucleotide synthetase (PurM), from the Escherichia coli purine biosynthetic pathway at 2.5 Å resolution', *Structure*, vol 7, no. 9, pp. 1155-1166.

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*, vol 25, no. 16, pp. 2078-2079.

Li, W & Godzik, A 2006, 'Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences', *Bioinformatics*, vol 22, no. 13, pp. 1658-1659.

Li, Y, Willer, CJ, Ding, J, Scheet, P & Abecasis, GR 2010, 'MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes', *Genetic Epidemiology*, vol 34, no. 8, pp. 816-834.

Li, Z, Lou, H, Ojcius, DM, Sun, A, Sun, D, Zhao, J, Lin, X & Yan, J 2014, 'Methyl-accepting chemotaxis proteins 3 and 4 are responsible for Campylobacter jejuni chemotaxis and jejuna colonization in mice in response to sodium deoxycholate', *Journal of Medical Microbiology*, vol 63, no. PART 3, pp. 343-354.

Liao, R-Y, Mao, C, Qiu, L-X, Ding, H, Chen, Q & Pan, H-F 2010, 'TGFBR1*6A/9A polymorphism and cancer risk: a meta-analysis of 13,662 cases and 14,147 controls', *Molecular Biology Reports*, vol 37, no. 7, pp. 3227-3232.

Lin, J, Sahin, O, Michel, LO & Zhang, Q 2003, 'Critical Role of Multidrug Efflux Pump CmeABC in Bile Resistance and In Vivo Colonization of Campylobacter jejuni', *Infection and Immunity*, vol 71, no. 8, pp. 4250-4259.

Lindsay, JA, Moore, CE, Day, NP, Peacock, SJ, Witney, AA, Stabler, RA, Husain, SE, Butcher, PD, Hinds, J, Al, LET & Acteriol, JB 2006, 'Microarrays Reveal that Each of the Ten Dominant Lineages of Staphylococcus aureus Has a Unique Combination of Surface-Associated and Regulatory Genes', *Journal of Bacteriology*, vol 188, no. 2, pp. 669-676.

Lippert, C, Listgarten, J, Liu, Y, Kadie, CM, Davidson, RI & Heckerman, D 2011, 'FaST linear mixed models for genome-wide association studies', *Nature Methods*, vol 8, no. 10, pp. 833-835.

Listgarten, J, Lippert, C & Heckerman, D 2013, 'FaST-LMM-Select for addressing confounding from spatial structure and rare variants', *Nature Genetics*, vol 45, no. 5, pp. 470-471.

Listgarten, J, Lippert, C, Kadie, CM, Davidson, RI, Eskin, E & Heckerman, D 2012, 'Improved linear mixed models for genome-wide association studies.', *Nature Methods*, vol 9, no. 6, pp. 525-526.

Lo, H, Tang, CM & Exley, RM 2009, 'Mechanisms of avoidance of host immunity by Neisseria meningitidis and its effect on vaccine development', *The Lancet Infectious Diseases*, vol 9, no. 7, pp. 418-427.

Lohmueller, KE, Pearce, CL, Pike, M, Lander, ES & Hirschhorn, JN 2003, 'Meta-analysis of

genetic association studies supports a contribution of common variants to susceptibility to common disease', *Nature Genetics*, vol 33, no. 2, pp. 177-182.
Loman, NJ, Quick, J & Simpson, JT 2015, 'A complete bacterial genome assembled de novo using only nanopore sequencing data', *Nature Methods*, vol 12, no. 8, pp. 733-735.
Louwen, R, Heikema, A, Van Belkum, A, Ott, A, Gilbert, M, Ang, W, Endtz, HP, Bergman, MP & Nieuwenhuis, EE 2008, 'The sialylated lipooligosaccharide outer core in Campylobacter jejuni is an important determinant for epithelial cell invasion', *Infection and Immunity*, vol 76, no. 10, pp. 4431-4438.
Louwen, R, Horst-Kreft, D, De Boer, AG, Van Der Graaf, L, De Knegt, G, Hamersma, M, Heikema, AP, Timms, AR, Jacobs, BC, Wagenaar, JA, Endtz, HP, Van Der Oost, J, Wells, JM, Nieuwenhuis, EES, Van Vliet, AHM, Willemsen, PTJ, Van Baarlen, P & Van Belkum, A 2013, 'A novel link between Campylobacter jejuni bacteriophage defence, virulence and Guillain-Barré syndrome', *European Journal of Clinical Microbiology and Infectious Diseases*, vol 32, no. 2, pp. 207-226.
Lucidarme, J, Hill, DMC, Bratcher, HB, Gray, SJ, du Plessis, M, Tsang, RSW, Vazquez, JA, Taha, MK, Ceyhan, M, Efron, AM, Gorla, MC, Findlow, J, Jolley, KA, Maiden, MCJ & Borrow, R 2015, 'Genomic resolution of an aggressive, widespread, diverse and expanding meningococcal serogroup B, C and W lineage', *Journal of Infection*, vol 71, no. 5, pp. 544-552.
Luechtefeld, NAW, Blaser, MJ, Reller, LB & Wang, WLL 1980, 'Isolation of Campylobacter-Fetus Subsp Jejuni From Migratory Waterfowl', *Journal of Clinical Microbiology*, vol 12, no. 3, pp. 406-408.
Lunter, G & Goodson, M 2011, 'Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads.', *Genome Research*, vol 21, no. 6, pp. 936-939.
Lwoff, A 1953, 'Lysogeny', *Bacteriological Reviews*, vol 17, no. 4, pp. 269-337.
Maiden, MCJ, Bygraves, JA, Feil, E, Morelli, G, Russell, JE, Urwin, R, Zhang, Q, Zhou, J, Zurth, K, Caugant, DA, Feavers, IM, Achtman, M & Spratt, BG 1998, 'Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms', *Proceedings of the National Academy of Sciences*, vol 95, no. 6, pp. 3140-3145.
Manson, AL, Cohen, KA, Abeel, T, Desjardins, CA, Armstrong, DT, Barry III, CE, Brand, J, Consortium, TGG, Chapman, SB, Cho, S-N, Gabrielian, A, Gomez, J, Jodals, AM, Joloba, M, Jureen, P, Lee, JS, Malinga, L, Maiga, M, Nordenberg, D, Noroc, E, et al. 2017, 'Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance', *Nature Genetics*, vol 49, no. 3, pp. 395-402.
Mao, X, Ke, Z, Shi, X, Liu, S, Tang, B, Wang, J & Huang, H 2015, 'Diagnosis of drug resistance to fluoroquinolones, amikacin, capreomycin, kanamycin and ethambutol with genotype MTBDRsl assay: A meta-analysis', *Annals of Clinical and Laboratory Science*, vol 45, no. 5, pp. 533-544.
Marchant, J, Wren, B & Ketley, J 2002, 'Exploiting genome sequence: Predictions for mechanisms of Campylobacter chemotaxis', *Trends in Microbiology*, vol 10, no. 4, pp. 155-159.
Marchini, J & Howie, B 2010, 'Genotype imputation for genome-wide association studies.', *Nature Reviews Genetics*, vol 11, no. 7, pp. 499-511.
Marchini, J, Cardon, LR, Phillips, MS & Donnelly, P 2004, 'The effects of human population structure on large genetic association studies', *Nature Genetics*, vol 36, no. 5, pp. 512-517.
Marcusson, LL, Frimodt-Møller, N & Hughes, D 2009, 'Interplay in the selection of fluoroquinolone resistance and bacterial fitness', *PLoS Pathogens*, vol 5, no. 8, p. e1000541.
Marshall, FA, Messenger, SL, Wyborn, NR, Guest, JR, Wing, H, Busby, SJW & Green, J

2001, 'A novel promoter architecture for microaerobic activation by the anaerobic transcription factor FNR', *Molecular Microbiology*, vol 39, no. 3, pp. 747-753.
Marteyn, B, West, NP, Browning, DF, Cole, JA, Shaw, JG, Palm, F, Mounier, J, Prévost, M-C, Sansonetti, P & Tang, CM 2010, 'Modulation of Shigella virulence in response to available oxygen in vivo.', *Nature*, vol 465, no. 7296, pp. 355-358.
Martin, P, Van De Ven, T, Mouchel, N, Jeffries, AC, Hood, DW & Moxon, ER 2003, 'Experimentally revised repertoire of putative contingency loci in Neisseria meningitidis strain MC58: Evidence for a novel mechanism of phase variation', *Molecular Microbiology*, vol 50, no. 1, pp. 245-257.
Masignani, V, Comanducci, M, Giuliani, MM, Bambini, S, Adu-Bobie, J, Aricò, B, Brunelli, B, Pieri, A, Santini, L, Savino, S, Serruto, D, Litt, D, Kroll, S, Welsch, JA, Granoff, DM, Rappuoli, R & Pizza, M 2003, 'Vaccination against Neisseria meningitidis Using Three Variants of the Lipoprotein GNA1870', *The Journal of Experimental Medicine*, vol 197, no. 6, pp. 789-799.
Mathieson, I & McVean, G 2012, 'Differential confounding of rare and common variants in spatially structured populations', *Nature Genetics*, vol 44, no. 3, pp. 243-246.
Mathieson, I & McVean, G 2013, 'Reply to: "FaST-LMM-Select for addressing confounding from spatial structure and rare variants"', *Nature Genetics*, vol 45, no. 5, p. 471.
Maurelli, AT 2007, 'Black holes, antivirulence genes, and gene inactivation in the evolution of bacterial pathogens', *FEMS Microbiology Letters*, vol 267, no. 1, pp. 1-8.
Mccarthy, MI, Abecasis, GR, Cardon, LR, Goldstein, DB, Little, J & Ioannidis, JPA 2008, 'Genome-wide association studies for complex traits: consensus, uncertainty and challenges', *Nature Reviews Genetics*, vol 9, no. 5, pp. 356-369.
McCarthy, ND, Colles, FM, Dingle, KE, Bagnall, MC, Manning, G, Maiden, MCJ & Falush, D 2007, 'Host-associated genetic import in Campylobacter jejuni', *Emerging Infectious Diseases*, vol 13, no. 2, pp. 267-272.
McLaws, FB, Larsen, AR, Skov, RL, Chopra, I & O'Neill, AJ 2011, 'Distribution of fusidic acid resistance determinants in methicillin-resistant Staphylococcus aureus.', *Antimicrobial Agents and Chemotherapy*, vol 55, no. 3, pp. 1173-1176.
McNeil, LK, Murphy, E, Zhao, XJ, Guttmann, S, Harris, SL, Scott, AA, Tan, C, Mack, M, DaSilva, I, Alexander, K, Mason, K, Jiang, HQ, Zhu, D, Mininni, TL, Zlotnick, GW, Hoiseth, SK, Jones, TR, Pride, MW, Jansen, KU & Anderson, AS 2009, 'Detection of LP2086 on the cell surface of Neisseria meningitidis and its accessibility in the presence of serogroup B capsular polysaccharide', *Vaccine*, vol 27, no. 25-26, pp. 3417-3421.
McVean, GAT 2002, 'A Genealogical Interpretation of Linkage Disequilibrium', *Genetics*, vol 162, no. 2, pp. 987-991.
McVean, G 2009, 'A genealogical interpretation of principal components analysis', *PLoS Genetics*, vol 5, no. 10, p. e1000686.
McVean, GAT, Myers, SR, Hunt, S, Deloukas, P, Bentley, DR & Donnelly, P 2004, 'The Fine-Scale Structure of Recombination Rate Variation in the Human Genome', *Science*, vol 304, no. 5670, pp. 581-585.
Meade, KG, Narciandi, F, Cahalane, S, Reiman, C, Allan, B & O'Farrelly, C 2009, 'Comparative in vivo infection models yield insights on early host immune response to Campylobacter in chickens', *Immunogenetics*, vol 61, no. 2, pp. 101-110.
Mederma, GJ, Schets, FM, van de Giessen, AW & Havelaar, AH 1992, 'Lack of colonization of 1 day old chicks by viable, non-culturable Campylobacter jejuni', *Journal of Applied Bacteriology*, vol 72, no. 6, pp. 512-516.
Méric, G, Yahara, K, Mageiros, L, Pascoe, B, Maiden, MCJ, Jolley, KA & Sheppard, SK 2014, 'A reference pan-genome approach to comparative bacterial genomics: Identification of novel epidemiological markers in pathogenic Campylobacter', *PLoS ONE*, vol 9, no. 3.

Metris, A, Reuter, M, Gaskin, DJ, Baranyi, J & van Vliet, AH 2011, 'In vivo and in silico determination of essential genes of Campylobacter jejuni', *BMC Genomics*, vol 12, no. 1, p. 535.
Minoche, AE, Dohm, JC & Himmelbauer, H 2011, 'Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems', *Genome Biology*, vol 12, no. 11, p. R112.
Mohan, V, Stevenson, M, Marshall, J, Fearnhead, P, Holland, BR, Hotter, G & French, NP 2013, 'Campylobacter jejuni colonization and population structure in urban populations of ducks and starlings in New Zealand', *MicrobiologyOpen*, vol 2, no. 4, pp. 659-673.
Monecke, S, Slickers, P, Hotzel, H, Richter-Huhn, G, Pohle, M, Weber, S, Witte, W & Ehricht, R 2006, 'Microarray-based characterisation of a Panton-Valentine leukocidin-positive community-acquired strain of methicillin-resistant Staphylococcus aureus', *Clinical Microbiology and Infection*, vol 12, no. 8, pp. 718-728.
Moran, RA, Anantham, S, Holt, KE & Hall, RM 2017, 'Prediction of antibiotic resistance from antibiotic resistance genes detected in antibiotic-resistant commensal Escherichia coli using PCR or WGS', *Journal of Antimicrobial Chemotherapy*, vol 72, no. 3, pp. 700-704.
Morris, AP, Lindgren, CM, Zeggini, E, Timpson, NJ, Frayling, TM, Hattersley, AT & McCarthy, MI 2010, 'A powerful approach to sub-phenotype analysis in population-based genetic association studies', *Genetic Epidemiology*, vol 34, no. 4, pp. 335-343.
Müller, MG, Ing, JY, Cheng, MK-W, Flitter, BA & Moe, GR 2013, 'Identification of a phage-encoded Ig-binding protein from invasive Neisseria meningitidis.', *Journal of Immunology*, vol 191, no. 6, pp. 3287-3296.
Müller, MG, Moe, NE, Richards, PQ & Moe, GR 2015, 'Resistance of Neisseria meningitidis to human serum depends on T and B cell stimulating protein B', *Infection and Immunity*, vol 83, no. 4, pp. 1257-1264.
Mund, NL-A, Masanta, WO, Goldschmidt, A-M, Lugert, R, Groß, U & Zautner, AE 2016, 'Association of Campylobacter Jejuni ssp. Jejuni Chemotaxis Receptor Genes with Multilocus Sequence Types and Source of Isolation.', *European Journal of Microbiology & Immunology*, vol 6, no. 3, pp. 162-177.
Murphy, GL, Connell, TD, Barritt, DS, Koomey, M & Cannon, JG 1989, 'Phase variation of gonococcal protein II: Regulation of gene expression by slipped-strand mispairing of a repetitive DNA sequence', *Cell*, vol 56, no. 4, pp. 539-547.
Nachamkin, I, Yang, X-H, Stern, NJ & Russell, RB 1993, 'Role of Campylobacterjejuni Flagella as Colonization Factors for Three-Day-Old Chicks: Analysis with Flagellar Mutants', *Applied and Environmental Microbiology*, vol 59, no. 5, pp. 1269-1273.
Naito, M, Frirdich, E, Fields, JA, Pryjma, M, Li, J, Cameron, A, Gilbert, M, Thompson, SA & Gaynor, EC 2010, 'Effects of sequential Campylobacter jejuni 81-176 lipooligosaccharide core truncations on biofilm formation, stress survival, and pathogenesis', *Journal of Bacteriology*, vol 192, no. 8, pp. 2182-2192.
NCI-NHGRI Working Group 2007, 'Replicating genotype – phenotype associations', *Nature*, vol 447, no. 7175, pp. 655-660.
Neal-Mckinney, JM, Christensen, JE & Konkel, ME 2010, 'Amino-terminal residues dictate the export efficiency of the Campylobacter jejuni filament proteins via the flagellum', *Molecular Microbiology*, vol 76, no. 4, pp. 918-931.
Nepon, GT & Erlich, H 1991, 'MHC CLASS-II MOLECULES AND AUTOIMMUNITY', *Annual Review of Immunology*, vol 9, pp. 493-525.
Nonhoff, C, Rottiers, S & Struelens, MJ 2005, 'Evaluation of the Vitek 2 system for identification and antimicrobial susceptibility testing of Staphylococcus spp.', *Clinical Microbiology and Infection*, vol 11, no. 2, pp. 150-153.
Nuijten, PJM, van Aster, FJAM, Gaastra, W & van der Zeijst, BAM 1990, 'Structural and Functional Analysis of Two Campylobacter jejuni Flagellin Genes*', *The Journal of*

Biological Chemistry, vol 265, no. 29, pp. 17798-17804.
O'Brien, FG, Price, C, Grubb, WB & Gustafson, JE 2002, 'Genetic characterization of the fusidic acid and cadmium resistance determinants of Staphylococcus aureus plasmid pUB101', *Journal of Antimicrobial Chemotherapy*, vol 50, no. 0305-7453, pp. 313-321.
O'Hagan, A & Forster, J 2010, 'Bayesian Inference Ch. 11', in *Kendall's Advanced Theory of Statistics Volume 2B*, 2nd edn, Wiley-Blackwell.
O'Neill, A & John Chopra, I 2006, 'Molecular basis of fusB-mediated resistance to fusidic acid in Staphylococcus aureus.', *Molecular Microbiology*, vol 59, no. 2, pp. 664-676.
O'Neill, AJ, Larsen, AR, Henriksen, AS & Chopra, I 2004, 'A fusidic acid-resistant epidemic strain of Staphylococcus aureus carries the fusB determinant, whereas fusA mutations are prevalent in other resistant isolates', *Antimicrobial Agents and Chemotherapy*, vol 48, no. 9, pp. 3594-3597.
O'Neill, AJ, McLaws, F, Kahlmeter, G, Henriksen, AS & Chopra, I 2007, 'Genetic basis of resistance to fusidic acid in staphylococci.', *Antimicrobial Agents and Chemotherapy*, vol 51, no. 5, pp. 1737-1740.
Oriente, F, Scarlato, V & Delany, I 2010, 'Expression of factor H binding protein of meningococcus responds to oxygen limitation through a dedicated FNR-regulated promoter', *Journal of Bacteriology*, vol 192, no. 3, pp. 691-701.
Palmer, AC & Kishony, R 2013, 'Understanding, predicting and manipulating the genotypic evolution of antibiotic resistance', *Nature Reviews Genetics*, vol 14, no. 4, pp. 243-248.
Park, SF 2002, 'The physiology of Campylobacter species and its relevance to their role as foodborne pathogens', *International Journal of Food Microbiology*, vol 74, no. 3, pp. 177-188.
Parker, CT, Gilbert, M, Yuki, N, Endtz, HP & Mandrell, RE 2008, 'Characterization of lipooligosaccharide-biosynthetic loci of Campylobacter jejuni reveals new lipooligosaccharide classes: Evidence of mosaic organizations', *Journal of Bacteriology*, vol 190, no. 16, pp. 5681-5689.
Parker, CT, Horn, ST, Gilbert, M, Miller, WG, Woodward, DL & Mandrell, RE 2005, 'Comparison of Campylobacter jejuni lipooligosaccharide biosynthesis loci from a variety of sources', *Journal of Clinical Microbiology*, vol 43, no. 6, pp. 2771-2781.
Parkhill, J, Achtman, M, James, KD, Bentley, SD, Churcher, C, Klee, SR, Morelli, G, Basham, D, Brown, D, Chillingworth, T, Davies, RM, Davis, P, Devlin, K, Feltwell, T, Hamlin, N, Holroyd, S, Jagels, K, Leather, S, Moule, S, Mungall, K, et al. 2000, 'Complete DNA sequence of a serogroup A strain of Neisseria meningitidis Z2491.', *Nature*, vol 404, no. 6777, pp. 502-506.
Parkhill, J, Wren, BW, Mungall, K, Ketley, JM, Churcher, C, Basham, D, Chillingworth, T, Davies, RM, Feltwell, T, Holroyd, S, Jagels, K, Karlyshev, AV, Moule, S, Pallen, MJ, Penn, CW, Quail, MA, Rajandream, MA, Rutherford, KM, van Vliet, AH, Whitehead, S, et al. 2000, 'The genome sequence of the food-borne pathogen Campylobacter jejuni reveals hypervariable sequences.', *Nature*, vol 403, no. 6770, pp. 665-668.
Pascoe, B, Méric, G, Murray, S, Yahara, K, Mageiros, L, Bowen, R, Jones, NH, Jeeves, RE, Lappin-Scott, HM, Asakura, H & Sheppard, SK 2015, 'Enhanced biofilm formation and multi-host transmission evolve from divergent genetic backgrounds in Campylobacter jejuni', *Environmental Microbiology*, vol 17, no. 11, pp. 4779-4789.
Patterson, N, Price, AL & Reich, D 2006, 'Population Structure and Eigenanalysis', *PLOS Genetics*, vol 2, no. 12, p. e190.
Pe'er, I, Yelensky, R, Altshuler, D & Daly, MJ 2008, 'Estimation of the Multiple Testing Burden for Genomewide Association Studies of Nearly All Common Variants', *Genetic Epidemiology*, vol 32, no. 4, pp. 381-385.
Peacock, SJ, Moore, CE, Justice, A, Kantzanou, M, Story, L, Mackie, K, Neill, GO & Day, NPJ 2002, 'Virulent Combinations of Adhesin and Toxin Genes in Natural Populations of Staphylococcus aureus', *Infection and Immunity*, vol 70, no. 9, pp. 4987-4996.

Pearson, BM, Gaskin, DJH, Segers, RPAM, Wells, JM, Nuijten, PJM & Van Vliet, AHM 2007, 'The complete genome sequence of Campylobacter jejuni strain 81116 (NCTC11828)', *Journal of Bacteriology*, vol 189, no. 22, pp. 8402-8403.
Perera, VN, Nachamkin, I, Ung, H, Patterson, JH, McConville, MJ, Coloe, PJ & Fry, BN 2007, 'Molecular mimicry in Campylobacter jejuni: Role of the lipo-oligosaccharide core oligosaccharide in inducing anti-ganglioside antibodies', *FEMS Immunology and Medical Microbiology*, vol 50, no. 1, pp. 27-36.
Pérez-Losada, M, Browne, EB, Madsen, A, Wirth, T, Viscidi, RP & Crandall, KA 2006, 'Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data', *Infection, Genetics and Evolution*, vol 6, no. 2, pp. 97-112.
Pickrell, JK 2014, 'Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits', *The American Journal of Human Genetics*, vol 94, no. 4, pp. 559-573.
Pires, SM, Vigre, H, Makela, P & Hald, T 2010, 'Using Outbreak Data for Source Attribution of Human Salmonellosis and Campylobacteriosis in Europe', *Foodborne Pathogens and Disease*, vol 7, no. 11, pp. 1351-1361.
Porcu, E, Sanna, S, Fuchsberger, C & Fritsche, LG 2013, 'Genotype Imputation in Genome-Wide Association Studies', in *Current Protocols in Human Genetics*, John Wiley & Sons, Inc.
Power, RA, Parkhill, J & de Oliveira, T 2017, 'Microbial genome-wide association studies: lessons from human GWAS', *Nature Reviews Genetics*, vol 18, no. 1, pp. 41-50.
Price, AL, Helgason, A, Palsson, S, Stefansson, H, St. Clair, D, Andreassen, OA, Reich, D, Kong, A & Stefansson, K 2009, 'The Impact of Divergence Time on the Nature of Population Structure: An Example from Iceland', *PLOS Genetics*, vol 5, no. 6, p. e1000505.
Price, AL, Patterson, NJ, Plenge, RM, Weinblatt, ME, Shadick, NA & Reich, D 2006, 'Principal components analysis corrects for stratification in genome-wide association studies', *Nature Genetics*, vol 38, no. 8, pp. 904-909.
Price, AL, Spencer, CCA & Donnelly, P 2015, 'Progress and promise in understanding the genetic basis of common diseases', *Proceedings of the Royal Society B: Biological Sciences*, vol 282, no. 1821.
Price, AL, Zaitlen, NA, Reich, D & Patterson, N 2010, 'New approaches to population stratification in genome-wide association studies', *Nature Reviews Genetics*, vol 11, no. 7, pp. 459-463.
Pritchard, JK & Rosenberg, NA 1999, 'Use of Unlinked Genetic Markers to Detect Population Stratification in Association Studies', *The American Journal of Human Genetics*, vol 65, no. 1, pp. 220-228.
Pritchard, JK, Stephens, M & Donnelly, P 2000, 'Inference of Population Structure Using Multilocus Genotype Data', *Genetics*, vol 155, no. 2, pp. 945-959.
Pritchard, JK, Stephens, M, Rosenberg, NA & Donnelly, P 2000, 'Association Mapping in Structured Populations', *The American Journal of Human Genetics*, vol 67, no. 1, pp. 170-181.
Public Health England 2017, 'England world leaders in the use of whole genome sequencing to diagnose TB', *https://www.gov.uk/government/news/england-world-leaders-in-the-use-of-whole-genome-sequencing-to-diagnose-tb*.
Pupko, T, Pe, I, Shamir, R & Graur, D 2000, 'A Fast Algorithm for Joint Reconstruction of Ancestral Amino Acid Sequences', *Molecular Biology and Evolution*, vol 17, no. 6, pp. 890-896.
Purcell, S, Neale, B, Todd-brown, K, Thomas, L, Ferreira, MAR, Bender, D, Maller, J, Sklar, P, Bakker, PIWD, Daly, MJ & Sham, PC 2007, 'PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses', *The American Journal of Human Genetics*, vol 81, no. 3, pp. 559-575.

Quick, J, Ashton, P, Calus, S, Chatt, C, Gossain, S, Hawker, J, Nair, S, Neal, K, Nye, K, Peters, T, Pinna, E, De Robinson, E, Struthers, K, Webber, M, Catto, A, Dallman, TJ, Hawkey, P & Loman, NJ 2015, 'Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella', *Genome Biology*, vol 16, no. 1, p.114.

R Core Team 2015, 'R: A language and environment for statistical computing.', *R Foundation for Statistical Computing, Vienna, Austria.*, https://www.R-project.org/.

Ragoussis, J 2009, 'Genotyping Technologies for Genetic Research', *Annual Review of Genomics and Human Genetics*, vol 10, no. 1, pp. 117-133.

Rajaratnam, B, Sparks, D, Khare, K, Zhang L 2017, 'Scalable Bayesian shrinkage and uncertainty quantification in high-dimensional regression', *arXiv*:1703.09163

Read, TD & Massey, RC 2014, 'Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies : a new direction for bacteriology', *Genome Medicine*, vol 6, no. 11, p. 109.

Recker, M, Laabei, M, Toleman, MS, Reuter, S, Saunderson, RB, Blane, B, Török, ME, Ouadi, K, Stevens, E, Yokoyama, M, Steventon, J, Thompson, L, Milne, G, Bayliss, S, Bacon, L, Peacock, SJ & Massey, RC 2017, 'Clonal differences in Staphylococcus aureus bacteraemia-associated mortality', *Nature Microbiology*.

Reich, DE & Goldstein, DB 2001, 'Detecting association in a case-control study while correcting for population stratification', *Genetic Epidemiology*, vol 20, no. 1, pp. 4-16.

Reuter, M & van Vliet, AHM 2013, 'Signal Balancing by the CetABC and CetZ Chemoreceptors Controls Energy Taxis in Campylobacter jejuni', *PLoS ONE*, vol 8, no. 1, pp. 1-10.

Revez, J & Hänninen, ML 2012, 'Lipooligosaccharide locus classes are associated with certain Campylobacter jejuni multilocus sequence types', *European Journal of Clinical Microbiology and Infectious Diseases*, vol 31, no. 9, pp. 2203-2209.

Risch, N & Merikangas, K 1996, 'The Future of Genetic Studies of Complex Human Diseases', *Science*, vol 273, no. 5281, pp. 1516-1517.

Risch, NJ 2000, 'Searching for genetic determinants in the new millennium', *Nature*, vol 405, no. 6788, pp. 847-856.

Rizk, G, Lavenier, D & Chikhi, R 2013, 'DSK: k-mer counting with very low memory usage', *Bioinformatics*, vol 29, no. 5, pp. 652-653.

Robinson, DA & Enright, MC 2004, 'Evolution of Staphylococcus aureus by Large Chromosomal Replacements', *Journal of Bacteriology*, vol 186, no. 4, pp. 1060-1064.

Rollins, DM & Colwell, RR 1986, 'Viable but nonculturable stage of Campylobacter jejuni and its role in survival in the natural aquatic environment', *Applied and Environmental Microbiology*, vol 52, no. 3, pp. 531-538.

Rosenberg, NA, Pritchard, JK, Weber, JL, Cann, HM, Kidd, KK, Zhivotovsky, LA & Feldman, MW 2002, 'Genetic Structure of Human Populations', *Science*, vol 298, no. 5602, pp. 2381-2385.

Rosenstein, NE, Perkins, BA, Stephens, DS, Popovic, T & Hughes, JM 2001, 'Meningococcal Disease', *The New England Journal of Medicine*, vol 344, no. 18, pp. 1378-1388.

Ross, MG, Russ, C, Costello, M, Hollinger, A, Lennon, NJ, Hegarty, R, Nusbaum, C & Jaffe, DB 2013, 'Characterizing and measuring bias in sequence data', *Genome Biology*, vol 14, no. 5, p. R51.

Saha, SK, Saha, S & Sanyal, SC 1991, 'Recovery of injured Campylobacter jejuni cells after animal passage', *Applied and Environmental Microbiology*, vol 57, no. 11, pp. 3388-3389.

Salipante, SJ, Roach, DJ, Kitzman, JO, Snyder, MW, Stackhouse, B, Butler-wu, SM, Lee, C, Cookson, BT & Shendure, J 2015, 'Large-scale genomic sequencing of extraintestinal pathogenic Escherichia coli strains', *Genome Research*, vol 25, no. 1, pp. 119-128.

Sarkar, MK, Paul, K & Blair, DF 2010, 'Chemotaxis signaling protein CheY binds to the rotor protein FliN to control the direction of flagellar rotation in Escherichia coli.',

Proceedings of the National Academy of Sciences, vol 107, no. 20, pp. 9370-9375.
Saunders, NJ, Jeffries, AC, Peden, JF, Hood, DW, Tettelin, H, Rappuoli, R & Moxon, ER 2000, 'Repeat-associated phase variable genes in the complete genome sequence of Neisseria meningitidis strain MC58', *Molecular Microbiology*, vol 37, no. 1, pp. 207-215.
Scheet, P & Stephens, M 2006, 'A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data : Applications to Inferring Missing Genotypes and Haplotypic Phase', *The Americal Journal of Human Genetics*, vol 78, no. 4, pp. 629-644.
Schneider, MC, Exley, RM, Chan, H, Feavers, I, Kang, Y-H, Sim, RB & Tang, CM 2006, 'Functional Significance of Factor H Binding to Neisseria meningitidis', *The Journal of Immunology*, vol 176, no. 12, pp. 7566-7575.
Schoen, C, Blom, J, Claus, H, Schramm-Glück, A, Brandt, P, Müller, T, Goesmann, A, Joseph, B, Konietzny, S, Kurzai, O, Schmitt, C, Friedrich, T, Linke, B, Vogel, U & Frosch, M 2008, 'Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in Neisseria meningitidis.', *Proceedings of the National Academy of Sciences*, vol 105, no. 9, pp. 3473-3478.
Schröder, J, Bailey, J, Conway, T & Zobel, J 2010, 'Reference-free validation of short read data', *PLoS ONE*, vol 5, no. 9, pp. 1-11.
Scott, AE, Timms, AR, Connerton, PL, El-Shibiny, A & Connerton, IF 2007, 'Bacteriophage influence Campylobacter jejuni types populating broiler chickens', *Environmental Microbiology*, vol 9, no. 9, pp. 2341-2353.
Shams, F, Oldfield, NJ, Lai, SK, Tunio, SA, Wooldridge, KG & Turner, DPJ 2016, 'Fructose-1,6-bisphosphate aldolase of Neisseria meningitidis binds human plasminogen via its C-terminal lysine residue', *MicrobiologyOpen*, vol 5, no. 2, pp. 340-350.
Sheppard, SK, Cheng, L, Méric, G, De Haan, CPA, Llarena, AK, Marttinen, P, Vidal, A, Ridley, A, Clifton-Hadley, F, Connor, TR, Strachan, NJC, Forbes, K, Colles, FM, Jolley, KA, Bentley, SD, Maiden, MCJ, Hänninen, ML, Parkhill, J, Hanage, WP & Corander, J 2014, 'Cryptic ecology among host generalist Campylobacter jejuni in domestic animals', *Molecular Ecology*, vol 23, no. 10, pp. 2442-2451.
Sheppard, SK, Colles, F, Richardson, J, Cody, AJ, Elson, R, Lawson, A, Brick, G, Meldrum, R, Little, CL, Owen, RJ, Maiden, MCJ & McCarthy, ND 2010, 'Host association of campylobacter genotypes transcends geographic variations', *Applied and Environmental Microbiology*, vol 76, no. 15, pp. 5269-5277.
Sheppard, SK, Dallas, JF, Strachan, NJC, MacRae, M, McCarthy, ND, Wilson, DJ, Gormley, FJ, Falush, D, Ogden, ID, Maiden, MCJ & Forbes, KJ 2009, 'Campylobacter Genotyping to Determine the Source of Human Infection', *Clinical Infectious Diseases*, vol 48, no. 8, pp. 1072-1078.
Sheppard, SK, Didelot, X, Meric, G, Torralbo, A, Jolley, KA, Kelly, DJ, Bentley, SD, Maiden, MCJ, Parkhill, J & Falush, D 2013, 'Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter', *Proceedings of the National Academy of Sciences*, vol 110, no. 29, pp. 11923-11927.
Sheppard, SK, McCarthy, ND, Falush, D & Maiden, MCJ 2008, 'Convergence of Campylobacter species: Implications for bacterial evolution', *Science*, vol 320, no. 5873, pp. 237-239.
Shirley, M & Dhillon, S 2015, 'Bivalent rLP2086 Vaccine (Trumenba®): A Review in Active Immunization Against Invasive Meningococcal Group B Disease in Individuals Aged 10–25 Years', *BioDrugs*, vol 29, no. 5, pp. 353-361.
Siena, E, D'Aurizio, R, Riley, D, Tettelin, H, Guidotti, S, Torricelli, G, Moxon, ER & Medini, D 2016, 'In-silico prediction and deep-DNA sequencing validation indicate phase variation in 115 Neisseria meningitidis genes.', *BMC genomics*, vol 17, no. 1, p. 843.
Siringan, P, Connerton, PL, Cummings, NJ & Connerton, IF 2014, 'Alternative bacteriophage life cycles: the carrier state of Campylobacter jejuni.', *Open Biology*, vol 4, p. 130200.

Siringan, P, Connerton, PL, Payne, RJH & Connerton, IF 2011, 'Bacteriophage-mediated dispersal of Campylobacter jejuni biofilms', *Applied and Environmental Microbiology*, vol 77, no. 10, pp. 3320-3326.
Skwark, MJ, Croucher, NJ, Puranen, S, Chewapreecha, C, Pesonen, M, Xu, YY, Turner, P, Harris, SR, Beres, SB, Musser, JM, Parkhill, J, Bentley, SD, Aurell, E & Corander, J 2017, 'Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis', *PLOS Genetics*, vol 13, no. 2, p. e1006508.
Snyder, LAS & Saunders, NJ 2006, 'The majority of genes in the pathogenic Neisseria species are present in non-pathogenic Neisseria lactamica, including those designated as 'virulence genes'.', *BMC genomics*, vol 7, no. 1, p. 128.
Sopwith, W, Birtles, A, Matthews, M, Fox, A, Gee, S, Painter, M, Regan, M, Syed, Q & Bolton, E 2008, 'Identification of potential environmentally adapted Campylobacter jejuni strain, United Kingdom', *Emerging Infectious Diseases*, vol 14, no. 11, pp. 1769-1773.
Spinosa, MR, Progida, C, Talà, A, Cogli, L, Alifano, P & Bucci, C 2007, 'The Neisseria meningitidis capsule is important for intracellular survival in human cells', *Infection and Immunity*, vol 75, no. 7, pp. 3594-3603.
Spiro, S & Guest, JR 1990, 'FNR and its role in oxygen-regulated gene expression in Escherichia coli', *FEMS Microbiology Letters*, vol 75, no. 4, pp. 399-428.
Stahl, M & Stintzi, A 2011, 'Identification of essential genes in C. jejuni genome highlights hyper-variable plasticity regions', *Functional and Integrative Genomics*, vol 11, no. 2, pp. 241-257.
Stamatakis, A 2014, 'RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies', *Bioinformatics*, vol 30, no. 9, pp. 1312-1313.
Stegger, M, Price, LB, Larsen, AR, Gillece, JD, Waters, AE, Skov, R & Andersen, PS 2012, 'Genome sequence of staphylococcus aureus strain 11819-97, an ST80-IV european community-acquired methicillin-resistant isolate', *Journal of Bacteriology*, vol 194, no. 6, pp. 1625-1626.
Stephens, BB, Loar, SN & Alexandre, G 2006, 'Role of CheB and CheR in the complex chemotactic and aerotactic pathway of Azospirillum brasilense', *Journal of Bacteriology*, vol 188, no. 13, pp. 4759-4768.
Stephens, DS, Greenwood, B & Brandtzaeg, P 2007, 'Epidemic meningitis, meningococcaemia, and Neisseria meningitidis', *Lancet*, vol 369, no. 9580, pp. 2196-2210.
Stephens, M 2013, 'A Unified Framework for Association Analysis with Multiple Related Phenotypes', *PLOS ONE*, vol 8, no. 7, p. e65245.
Stephenson, HN, John, CM, Naz, N, Gundogdu, O, Dorrell, N, Wren, BW, Jarvis, GA & Bajaj-Elliott, M 2013, 'Campylobacter jejuni lipooligosaccharide sialylation, phosphorylation, and amide/ester linkage modifications fine-tune human toll-like receptor 4 activation', *Journal of Biological Chemistry*, vol 288, no. 27, pp. 19661-19672.
Stern, NJ, Bailey, JS, Blankenship, LC, Cox, NA & McHan, F 1988, 'Colonization Characteristics of Campylobacter jejuni in Chick Ceca', *Avian Diseases*, vol 32, no. 2, pp. 330-334.
Stern, NJ, Jones, DM, Wesley, IV & Rollins, DM 1994, 'Colonization of chicks by non-culturable Campylobacter spp.', *Letters in Applied Microbiology*, vol 18, pp. 333-336.
Stoesser, N, Batty, EM, Eyre, DW, Morgan, M, Wyllie, DH, Del Ojo Elias, C, Johnson, JR, Walker, AS, Peto, TEA & Crook, DW 2013, 'Predicting antimicrobial susceptibilities for Escherichia coli and Klebsiella pneumoniae isolates using whole genomic sequence data', *Journal of Antimicrobial Chemotherapy*, vol 68, no. 10, pp. 2234-2244.
Stoesser, N, Sheppard, AE, Peirano, G, Anson, LW, Pankhurst, L, Sebra, R, Phan, HTT, Kasarskis, A, Mathers, AJ, Peto, TEA, Bradford, P, Motyl, MR, Walker, AS, Crook,

DW & Pitout, JD 2017, 'Genomic epidemiology of global Klebsiella pneumoniae carbapenemase (KPC)-producing Escherichia coli', *Scientific Reports*, vol 7, no. 1, p. 5917.
Strachan, NJC, Gormley, FJ, Rotariu, O, Ogden, ID, Miller, G, Dunn, GM, Sheppard, SK, Dallas, JF, Reid, TMS, Howie, H, Maiden, MCJ & Forbes, KJ 2009, 'Attribution of Campylobacter infections in northeast Scotland to specific sources by use of multilocus sequence typing.', *The Journal of Infectious Diseases*, vol 199, no. 8, pp. 1205-1208.
Sturtevant, AH 1913, 'The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association', *Journal of Experimental Zoology*, vol 14, pp. 43-59.
Sun, S, Berg, OG, Roth, JR & Andersson, DI 2009, 'Contribution of gene amplification to evolution of increased antibiotic resistance in Salmonella typhimurium.', *Genetics*, vol 182, no. 4, pp. 1183-1195.
Sveinbjornsson, G, Gudbjartsson, DF, Halldorsson, BV, Kristinsson, KG, Gottfredsson, M, Barrett, JC, Gudmundsson, LJ, Blondal, K, Gylfason, A, Gudjonsson, SA, Helgadottir, HT, Jonasdottir, A, Jonasdottir, A, Karason, A, Kardum, LB, Knezevic, J, Kristjansson, H, Kristjansson, M, Love, A, Luo, Y, et al. 2016, 'HLA class II sequence variants influence tuberculosis risk in populations of European ancestry', *Nature Genetics*, vol 48, no. 3, pp. 318-322.
Svishcheva, GR, Axenovich, TI, Belonogova, NM, van Duijn, CM & Aulchenko, YS 2012, 'Rapid variance components-based method for whole-genome association analysis', *Nature Genetics*, vol 44, no. 10, pp. 1166-1170.
Talibart, R, Denis, M, Castillo, A, Cappelier, JM & Ermel, G 2000, 'Survival and recovery of viable but noncultivable forms of Campylobacter in aqueous microcosm', *International Journal of Food Microbiology*, vol 55, no. 1-3, pp. 263-267.
Tanaka, M, Wang, T, Onodera, Y, Uchida, Y & Sato, K 2000, 'Mechanism of quinolone resistance in Staphylococcus aureus', *Journal of Infection and Chemotherapy*, vol 6, no. 3, pp. 131-139.
Tanaka, N, Kinoshita, T & Masukawa, H 1968, 'Mechanism of protein synthesis inhibition by fusidic acid and related antibiotics', *Biochemical and Biophysical Research Communications*, vol 30, no. 3, pp. 278-283.
Telenti, A, Imboden, P, Marchesi, F, Matter, L, Schopfer, K, Bodmer, T, Lowrie, D, Colston, MJ & Cole, S 1993, 'Detection of rifampicin-resistance mutations in Mycobacterium tuberculosis', *The Lancet*, vol 341, no. 8846, pp. 647-651.
Tettelin, H, Saunders, NJ, Heidelberg, J, Jeffries, AC, Nelson, KE, Eisen, JA, Ketchum, KA, Hood, DW, Peden, JF, Dodson, RJ, Nelson, WC, Gwinn, ML, DeBoy, R, Peterson, JD, Hic, EK, Grandi, MPG, Sun, L, Smith, HO, Fraser, CM, Moxon, ER, et al. 2000, 'Complete Genome Sequence of Neisseria meningitidis Serogroup B Strain MC58', *Science*, vol 287, no. 5459, pp. 1809-1815.
The 1000 Genomes Project Consortium 2015, 'A global reference for human genetic variation', *Nature*, vol 526, no. 7571, pp. 68-74.
'The Codon Usage Database' http://www.kazusa.or.jp/codon/.
The International HapMap 3 Consortium 2010, 'Integrating common and rare genetic variation in diverse human populations', *Nature*, vol 467, no. 7311, pp. 52-58.
The International HapMap Consortium 2003, 'The International HapMap Project', *Nature*, vol 426, no. 6968, pp. 789-796.
The International HapMap Consortium 2005, 'A haplotype map of the human genome', *Nature*, vol 437, no. 7063, pp. 1299-1320.
The International HapMap Consortium 2007, 'A second generation human haplotype map of over 3. 1 million SNPs', *Nature*, vol 449, no. 7164, pp. 851-861.
The International SNP Map Working Group 2001, 'A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms', *Nature*, vol 409,

no. 6822, pp. 928-933.
The UniProt Consortium 2015, 'UniProt: a hub for protein information', *Nucleic Acids Research*, vol 43, no. D1, pp. D204-D212.
The Wellcome Trust Case Control Consortium 2007, 'Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls', *Nature*, vol 447, no. 7145, pp. 661-678.
Thoden, JB, Goneau, MF, Gilbert, M & Holden, HM 2013, 'Structure of a Sugar N-Formyltransferase from Campylobacter jejuni', *Biochemistry*, vol 52, no. 35, pp. 6114-6126.
Thomas, CM & Nielsen, KM 2005, 'Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria', *Nature Reviews Microbiology*, vol 3, no. 9, pp. 711-721.
Thornton, T & McPeek, MS 2007, 'Case-Control Association Testing with Related Individuals: A More Powerful Quasi-Likelihood Score Test', *The American Journal of Human Genetics*, vol 81, no. 2, pp. 321-337.
Thornton, T & Mcpeek, MS 2010, 'ROADTRIPS : Case-Control Association Testing with Partially or Completely Unknown Population and Pedigree Structure', *The American Journal of Human Genetics*, vol 86, no. 2, pp. 172-184.
Thye, T, Owusu-Dabo, E, Vannberg, FO, van Crevel, R, Curtis, J, Sahiratmadja, E, Balabanova, Y, Ehmen, C, Muntau, B, Ruge, G, Sievertsen, J, Gyapong, J, Nikolayevskyy, V, Hill, PC, Sirugo, G, Drobniewski, F, van de Vosse, E, Newport, M, Alisjahbana, B, Nejentsev, S, et al. 2012, 'Common variants at 11p13 are associated with susceptibility to tuberculosis', *Nature Genetics*, vol 44, no. 3, pp. 257-259.
Thye, T, Vannberg, FO, Wong, SH, Owusu-Dabo, E, Osei, I, Gyapong, J, Sirugo, G, Sisay-Joof, F, Enimil, A, Chinbuah, MA, Floyd, S, Warndorff, DK, Sichali, L, Malema, S, Crampin, AC, Ngwira, B, Teo, YY, Small, K, Rockett, K, Kwiatkowski, D, et al. 2010, 'Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2', *Nature Genetics*, vol 42, no. 9, pp. 739-741.
Treangen, TJ & Salzberg, SL 2012, 'Repetitive DNA and next-generation sequencing: computational challenges and solutions', *Nature Reviews Genetics*, vol 13, no. 1, pp. 36-46.
Treangen, TJ, Abraham, A-L, Touchon, M & Rocha, EPC 2009, 'Genesis, effects and fates of repeats in prokaryotic genomes', *FEMS Microbiology Reviews*, vol 33, no. 3, pp. 539-571.
Tucker, G, Price, AL & Berger, B 2014, 'Improving the power of GWAS and avoiding confounding from population stratification with PC-select', *Genetics*, vol 197, no. 3, pp. 1045-1049.
Tunio, SA, Oldfield, NJ, Berry, A, Ala'Aldeen, DAA, Wooldridge, KG & Turner, DPJ 2010, 'The moonlighting protein fructose-1, 6-bisphosphate aldolase of Neisseria meningitidis: Surface localization and role in host cell adhesion', *Molecular Microbiology*, vol 76, no. 3, pp. 605-615.
Tzeng, Y-L 2005, 'Translocation and Surface Expression of Lipidated Serogroup B Capsular Polysaccharide in Neisseria meningitidis', *Infection and Immunity*, vol 73, no. 3, pp. 1491-1505.
University of Oxford 2017, *Global team aim for faster, more effective TB diagnosis*, viewed August 2017, "http://www.ox.ac.uk/news/2016-03-24-global-team-aim-faster-more-effective-tb-diagnosis".
Uren, C, Henn, BM, Franke, A, Wittig, M, van Helden, PD, Hoal, EG & Möller, M 2017, 'A post-GWAS analysis of predicted regulatory variants and tuberculosis susceptibility', *PLOS ONE*, vol 12, no. 4, p. e0174738.
Vegge, CS, Brøndsted, L, Li, YP, Bang, DD & Ingmer, H 2009, 'Energy taxis drives Campylobacter jejuni toward the most favorable conditions for growth', *Applied and Environmental Microbiology*, vol 75, no. 16, pp. 5308-5314.
Velasco, AM, Leguina, JI & Lazcano, A 2002, 'Molecular evolution of the lysine biosynthetic

pathways', *Journal of Molecular Evolution*, vol 55, no. 4, pp. 445-459.
Venter, JC, Adams, MD, Myers, EW, Li, PW, Mural, RJ, Sutton, GG, Smith, HO, Yandell, M, Evans, CA, Holt, RA, Gocayne, JD, Amanatides, P, Ballew, RM, Huson, DH, Wortman, JR, Zhang, Q, Kodira, CD, Zheng, XH, Chen, L, Skupski, M, et al. 2001, 'The Sequence of the Human Genome', *Science*, vol 291, no. 5507, pp. 1304-1351.
Verwoerd, D 2000, 'Ostrich diseases', *Revue Scientifique et Technique*, vol 19, no. 2, pp. 638-661.
Viana, D, Comos, M, McAdam, PR, Ward, MJ, Selva, L, Guinane, CM, Gonzalez-Munoz, BM, Tristan, A, Foster, SJ, Fitzgerald, JR & Penades, JR 2015, 'A single natural nucleotide mutation alters bacterial pathogen host tropism', *Nature Genetics*, vol 47, no. 4, pp. 361-366.
Visscher, PM, Hill, WG & Wray, NR 2008, 'Heritability in the genomics era - concepts and misconceptions', *Nature Reviews Genetics*, vol 9, no. 4, pp. 255-266.
Vogel, U, Claus, H, Von Müller, L, Bunjes, D, Elias, J & Frosch, M 2004, 'Bacteremia in an immunocompromised patient caused by a commensal Neisseria meningitidis strain harboring the capsule null locus (cnl)', *Journal of Clinical Microbiology*, vol 42, no. 7, pp. 2898-2901.
Voight, BF & Pritchard, JK 2005, 'Confounding from Cryptic Relatedness in Case-Control Association Studies', *PLOS Genetics*, vol 1, no. 3, p. e32.
Vos, M & Didelot, X 2009, 'A comparison of homologous recombination rates in bacteria and archaea.', *The ISME Journal*, vol 3, no. 2, pp. 199-208.
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', *Journal of Clinical Microbiology*, vol 55, no. 5, pp. 1285-1298.
Wagenaar, JA, Bergen, MAPV, Mueller, MA, Wassenaar, TM & Carlton, RM 2005, 'Phage therapy reduces Campylobacter jejuni colonization in broilers', *Veterinary Microbiology*, vol 109, no. 3-4, pp. 275-283.
Wald, A 1943, 'Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large ', *Transactions of the American Mathematical Society*, vol 54, no. 3, pp. 426-482.
Waldenström, J, Axelsson-Olsson, D, Olsen, B, Hasselquist, D, Griekspoor, P, Jansson, L, Teneberg, S, Svensson, L & Ellström, P 2010, 'Campylobacter jejuni colonization in wild birds: Results from an infection experiment', *PLoS ONE*, vol 5, no. 2, pp. 1-8.
Waldenström, J, Broman, T, Carlsson, I, Hasselquist, D, Achterberg, RP, Wagenaar, JA & Olsen, B 2002, 'Prevalence of Campylobacter jejuni, Campylobacter lari, and Campylobacter coli in Different Ecological Guilds and Taxa of Migrating Birds', *Applied and Environmental Microbiology*, vol 68, no. 12, pp. 5911-5917.
Walker, TM, Kohl, TA, Omar, SV, Hedge, J, Del Ojo Elias, C, Bradley, P, Iqbal, Z, Feuerriegel, S, Niehaus, KE, Wilson, DJ, Clifton, DA, Kapatai, G, Ip, CLC, Bowden, R, Drobniewski, FA, Allix-Béguec, C, Gaudin, C, Parkhill, J, Diel, R, Supply, P, et al. 2015, 'Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: A retrospective cohort study', *The Lancet Infectious Diseases*, vol 15, no. 10, pp. 1193-1202.
Walport, MJ 2001, 'Complement: first of two parts', *New England Journal of Medicine*, vol 344, no. 14, pp. 1058-1066.
Wang, Y & Taylor, DE 1990, 'Natural transformation in Campylobacter species', *Journal of Bacteriology*, vol 172, no. 2, pp. 949-955.
Wassenaar, TM, Van Der Zeijst, BA, Ayling, R & Newell, DG 1993, 'Colonization of chicks by motility mutants of Campylobacter jejuni demonstrates the importance of flagellin A expression.', *Journal of General Microbiology*, vol 139, no. 6, pp. 1171-1175.
Webber, MA & Piddock, LJV 2003, 'The importance of efflux pumps in bacterial antibiotic

resistance', *Journal of Antimicrobial Chemotherapy*, vol 51, no. 1, pp. 9-11.
Weber, MVR, Claus, H, Maiden, MCJ, Frosch, M & Vogel, U 2006, 'Genetic mechanisms for loss of encapsulation in polysialyltransferase-gene-positive meningococci isolated from healthy carriers', *International Journal of Medical Microbiology*, vol 296, no. 7, pp. 475-484.
Weir, BS, Anderson, AD & Hepler, AB 2006, 'Genetic relatedness analysis: modern data and new challenges', *Nature Reviews Genetics*, vol 7, no. 10, pp. 771-780.
Weis, AM, Storey, DB, Taff, CC, Townsend, AK, Huang, BC, Kong, NT, Clothier, KA, Spinner, A, Byrne, BA & Weimer, BC 2016, 'Genomic Comparison of Campylobacter spp. and Their Potential for Zoonotic Transmission between Birds, Primates, and Livestock', *Applied & Environmental Microbiology*, vol 82, no. 24, pp. 7165-7175.
Wheeler, JG, Sethi, D, Cowden, JM, Wall, PG, Rodrigues, LC, Tompkins, DS, Hudson, MJ & Roderick, PJ 1999, 'Study of infectious intestinal disease in England: rates in the community, presenting to general practice, and reported to national surveillance. The Infectious Intestinal Disease Study Executive.', *BMJ*, vol 318, no. 7190, pp. 1046-1050.
Wick, RR, Judd, LM, Gorrie, CL & Holt, KE 2017, 'Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads', *PLOS Computational Biology*, vol 13, no. 6, p. e1005595.
Widmer, C, Lippert, C, Weissbrod, O, Fusi, N, Kadie, C, Davidson, R, Listgarten, J & Heckerman, D 2014, 'Further improvements to linear mixed models for genome-wide association studies.', *Scientific Reports*, vol 4, p. 6874.
Williams, SM, Wing, HJ & Busby, SJW 1998, 'Repression of transcription initiation by Escherichia coli FNR protein: Repression by FNR can be simple', *FEMS Microbiology Letters*, vol 163, no. 2, pp. 203-208.
Wilson, DJ 2017, 'The harmonic mean p-value and model averaging by mean maximum likelihood', *bioRxiv*, doi:10.1101/171751.
Wilson, DJ, Gabriel, E, Leatherbarrow, AJH, Cheesbrough, J, Gee, S, Bolton, E, Fox, A, Fearnhead, P, Hart, CA & Diggle, PJ 2008, 'Tracing the source of campylobacteriosis', *PLoS Genetics*, vol 4, no. 9.
Wilson, DJ, Gabriel, E, Leatherbarrow, AJH, Cheesbrough, J, Gee, S, Bolton, E, Fox, A, Hart, CA, Diggle, PJ & Fearnhead, P 2009, 'Rapid evolution and the importance of recombination to the gastroenteric pathogen Campylobacter jejuni', *Molecular Biology and Evolution*, vol 26, no. 2, pp. 385-397.
Wood, DE & Salzberg, SL 2014, 'Kraken: ultrafast metagenomic sequence classification using exact alignments', *Genome Biology*, vol 15, no. 3, p. R46.
World Health Organisation 2012, *The evolving threat of antimicrobial resistance: options for action*, viewed August 2017, "http://apps.who.int/iris/bitstream/10665/75389/1/WHO_IER_PSP_2012.2_eng.pdf"
Wu, C, DeWan, A, Hoh, J & Wang, Z 2011, 'A Comparison of Association Methods Correcting for Population Stratification in Case–Control Studies', *Annals of Human Genetics*, vol 75, no. 3, pp. 418-427.
Wyres, KL, Gorrie, C, Edwards, DJ, Wertheim, HFL, Hsu, LY, Van Kinh, N, Zadoks, R, Baker, S & Holt, KE 2015, 'Extensive Capsule Locus Variation and Large-Scale Genomic Recombination within the Klebsiella pneumoniae Clonal Group 258', *Genome Biology and Evolution*, vol 7, no. 5, pp. 1267-1279.
Yahara, K, Méric, G, Taylor, AJ, de Vries, SPW, Murray, S, Pascoe, B, Mageiros, L, Torralbo, A, Vidal, A, Ridley, A, Komukai, S, Wimalarathna, H, Cody, AJ, Colles, FM, McCarthy, N, Harris, D, Bray, JE, Jolley, KA, Maiden, MCJ, Bentley, SD, et al. 2017, 'Genome-wide association of functional traits linked with Campylobacter jejuni survival from farm to fork', *Environmental Microbiology*, vol 19, no. 1, pp. 361-380.
Yang, J, Lee, SH, Goddard, ME & Visscher, PM 2011, 'GCTA: A Tool for Genome-wide Complex Trait Analysis', *The American Journal of Human Genetics*, vol 88, no. 1, pp.76-82

Yang, J, Zaitlen, NA, Goddard, ME, Visscher, PM & Price, AL 2014, 'Advantages and pitfalls in the application of mixed-model association methods', *Nature Genetics*, vol 46, no. 2, pp. 100-106.

Yazdankhah, SP, Kriz, P, Tzanakaki, G, Kremastinou, J, Kalmusova, J, Musilek, M, Alvestad, T, Jolley, KA, Wilson, DJ, McCarthy, ND, Caugant, DA & Maiden, MCJ 2004, 'Distribution of serogroups and genotypes among disease-associated and carried isolates of Neisseria meningitidis from the Czech Republic, Greece, and Norway', *Journal of Clinical Microbiology*, vol 42, no. 11, pp. 5146-5153.

Yu, J, Pressoir, G, Briggs, WH, Vroh Bi, I, Yamasaki, M, Doebley, JF, McMullen, MD, Gaut, BS, Nielsen, DM, Holland, JB, Kresovich, S & Buckler, ES 2006, 'A unified mixed-model method for association mapping that accounts for multiple levels of relatedness', *Nature Genetics*, vol 38, no. 2, pp. 203-208.

Yuki, N, Susuki, K, Koga, M, Nishimoto, Y, Odaka, M, Hirata, K, Taguchi, K, Miyatake, T, Furukawa, K, Kobata, T & Yamada, M 2004, 'Carbohydrate mimicry between human ganglioside GM1 and Campylobacter jejuni lipooligosaccharide causes Guillain-Barre syndrome', *Proceedings of the National Academy of Sciences*, vol 101, no. 31, pp. 11404-11409.

Zaitlen, N & Kraft, P 2012, 'Heritability in the genome-wide association era', *Human Genetics*, vol 131, no. 10, pp. 1655-1664.

Zaitlen, N, Pas, B, Gur, T, Ziv, E & Halperin, E 2010, 'Leveraging Genetic Variability across Populations for the Identification of Causal Variants', *The American Journal of Human Genetics*, vol 86, no. 1, pp. 23-33.

Zampara, A, Sørensen, MCH, Elsser-Gravesen, A & Brøndsted, L 2017, 'Significance of phage-host interactions for biocontrol of Campylobacter jejuni in food', *Food Control*, vol 73, pp. 1169-1175.

Zankari, E, Hasman, H, Kaas, RS, Seyfarth, AM, Agersø, Y, Lund, O, Larsen, MV & Aarestrup, FM 2013, 'Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing', *Journal of Antimicrobial Chemotherapy*, vol 68, no. 4, pp. 771-777.

Zeggini, E, Scott, LJ, Saxena, R & Voight, BF 2008, 'Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes', *Nature Genetics*, vol 40, no. 5, pp. 638-645.

Zerbino, DR & Birney, E 2008, 'Velvet: algorithms for de novo short read assembly using de Bruijn graphs.', *Genome Research*, vol 18, no. 5, pp. 821-829.

Zhang, K, Deng, M, Chen, T, Waterman, MS & Sun, F 2002, 'A dynamic programming algorithm for haplotype block partitioning', *Proceedings of the National Academy of Sciences*, vol 99, no. 11, pp. 7335-7339.

Zhang, Z, Ersoz, E, Lai, C-Q, Todhunter, RJ, Tiwari, HK, Gore, MA, Bradbury, PJ, Yu, J, Arnett, DK, Ordovas, JM & Buckler, ES 2010, 'Mixed linear model approach adapted for genome-wide association studies', *Nature Genetics*, vol 42, no. 4, pp. 355-360.

Zhou, J 2003, 'Microarrays for bacterial detection and microbial community analysis', *Current Opinion in Microbiology*, vol 6, no. 3, pp. 288-294.

Zhou, X & Stephens, M 2012, 'Genome-wide efficient mixed-model analysis for association studies', *Nature Genetics*, vol 44, no. 7, pp. 821-824.

Zhu, X, Zhang, S, Zhao, H & Cooper, RS 2002, 'Association Mapping , Using a Mixture Model for Complex Traits', *Genetic Epidemiology*, vol 23, no. 2, pp. 181-196.

Ziprin, RL, Droleskey, RE, Hume, ME & Harvey, RB 2003, 'Failure of Viable Nonculturable Campylobacter jejuni to Colonize the Cecum of Newly Hatched Leghorn', *Avian Diseases*, vol 47, no. 3, pp. 753-758.

Ziprin, RL, Harvey, RB, Ziprin, RL & Harvey, RB 2004, 'Inability of Cecal Microflora to Promote Reversion of Viable Nonculturable Campylobacter jejuni', *Avian Diseases*, vol 48, no. 3, pp. 647-650.

Chapter 8

Appendices

8 Appendices

8.1 Appendix A

Drug	Gene	Study	Variant	Genome position (in reference genome) or BLAST accession LMM	Genome position (in reference genome) or BLAST accession Pre-LMM	Alleles	Type	Ctrl 1	Ctrl 2	Ctrl 3	Case 1	Case 2	Case 3	Odds ratio	-log$_{10}$ *P*	Rank	-log$_{10}$ *P* LMM	Rank LMM
AMP	*β-lactamase genes*	SNP / gene	$bla_{TEM-208}$	NC_017659.1				51	1	-	59	130	-	112.4	20.1	1	19.4	1
		Kmer	$bla_{TEM-208}$	NC_017654.1 (1455 - 1485)				52	0	-	53	136	-	Inf	23.6	6	19.7	6
	Tn3-like transposase	Kmer	Linked to $bla_{OXA-181}$	KP400525.1 (51445-51475)				50	2	-	43	146	-	84.9	25.9	1	21.2	1
CIP	*gyrA*	SNP / gene	D87N, D87W	2626015		C, T, A	NS	147	2	1	5	86	0	1264.2, 0, 0	55.7	2	18.5	2
		Kmer	-	2626026 - 2626056		-	-	136	14	-	1	90	-	874.3	41.5	45	43.4	1
	gyrB	SNP / gene		4380590		A, C, G	S	107	35	8	7	39	45	17.0, 86.0, 5.0	26	43	21.4	8
		Kmer																
	parC	SNP / gene	S80I	3595065		G, A	NS	136	14	-	1	90	-	874.3	59.6	1	55.8	1
		Kmer		3595065 - 3595095				4	146	-	86	5		0.002	45.6	1	38.7	28
	parE	SNP / gene	I529L	3610603		T, G	NS	142	8	-	46	45	-	17.4	15.2	6626	14.8	156
		Kmer		NC_013361.1 (627842 - 627812)														
CFZ	*β-lactamase genes*	SNP / gene	$bla_{CTX-M-15}$	U5SQ39				102	0	-	96	43	-	Inf	12.7	2	6	3
		Kmer	$bla_{CTX-M-15}$	DQ335219.1 (405 - 435)				102	0	-	107	32	-	Inf	6.71	121710	3.99	3690
	nmpC	SNP / gene		P21420				15	87	-	91	48	-	0.09	15.4	1	12.4	1
		Kmer	*nmpC*	1985557 – 1985587				16	86	-	91	48	-	0.1	13.8	1	9.6	1

Continued on the next page

Drug	Gene	Study	Variant	Genome position (in reference genome) or BLAST accession LMM / Pre-LMM	Alleles	Type	Ctrl 1	Ctrl 2	Ctrl 3	Case 1	Case 2	Case 3	Odds ratio	-log$_{10}$ *P*	Rank	-log$_{10}$ *P* LMM	Rank LMM
CXM	*β-lactamase genes*	SNP / gene	*bla*$_{CTX-M-14}$	U5SQ39			159	1	-	39	42	-	171.2	23.2	1	18.99	1
		Kmer	*bla*$_{CTX-M-15}$	KP268826.1 (7 - 37)			160	0	-	50	31	-	Inf	16.3	1598	15.4	470
	Intergenic	Kmer	Linked to bla$_{CMY-2}$ (31177 - 32322)	LC019731.1 (31015 - 31045)			160	0	-	38	43	-	Inf	25.6	1	20	1
CRO	*β-lactamase genes*	SNP / gene	*bla*$_{CTX-M-15}$	NC_022648.1			185	1	-	13	42	-	597.7	34.5	1	48.2	1
		Kmer	*bla*$_{CTX-M-15}$	KP268826.1 (7 - 37)			186	0	-	24	31	-	Inf	27.3	1403	34.9	470
	tnpA – ISECP1	Kmer	Linked to bla$_{CTX-M-132}$ (8362 - 9237; complement)	KM207012.2 (9298 – 9328)			186	0	-	12	43	-	Inf	39.7	1	56.4	1
GEN	aac	SNP / gene	*aac(3)-II*	ESD46483.1			192	1	-	9	39	-	832	35.5	1	68.4	1
		Kmer	*aac(3)-II*	CP008735.1 (7913-7943)	CP008735.1 (7913-7943)		193	0	-	9	39	-	Inf	41.9	1	74	1
TOB	aac	SNP / gene	*aac(3)-II*	ESD46483.1			174	0	-	27	40	-	Inf	28.6	1	30.5	1
		Kmer	*aac(3)-II*	KJ850481(1 34-164)	KJ850481(1 34-164)		174	0		27	40	-	Inf	28.2	1	30.5	1

Table A.1 ***Escherichia coli*** **GWAS results.** AMP = Ampicillin, CFZ = Cefazolin, CIP = Ciprofloxacin, CRO = Ceftriaxone, CXM = Cefuroxime, GEN = Gentamicin, TOB = Tobramycin. Case = phenotypically resistant, control = phenotypically sensitive, aac = Aminoglycoside N-acetyltransferase genes, ant = Aminoglycoside N-acetyltransferase genes, aph = Aminoglycoside O- phosphotransferase genes. Causal gene names are coloured according to their resistance causing mechanism, red if its presence determines resistance, blue if substitutions within the gene causes resistance.

Drug	Gene	Study	Variant	Genome position (in reference genome) or BLAST accession LMM Pre-LMM	Alleles	Type	Ctrl 1	Ctrl 2	Ctrl 3	Case 1	Case 2	Case 3	Odds ratio	-log10(*p*)	Rank	-log10(*p*) LMM	Rank LMM
CFZ	*β-lactamase genes*	SNP / gene	*bla*$_{CTX-M-15}$	A0A075VKM9			122	1	-	20	33	-	201.3	20.8	1	15.2	2
		Kmer	*bla*$_{CTX-M-15}$	DQ335219.1 (110-140)			122	1	-	20	33	-	201.3	20.6	762	15.2	837
	HP from ISEcp1	SNP / gene					122	1	-	20	33	-	201.3	20.8	1	15.2	2
	protein WbuC	SNP / gene		AIG86706.1			122	1	-	20	33	-	201.3	20.8	1	15.2	1
	ISEcp1 tnpA	Kmer	Linked to *bla*$_{CTX-M-15}$	EU418923.1 (10812 - 10842, reverse)			122	1	-	19	34	-	218.3	21.3	1	18.3	1
CXM	*β-lactamase genes*	SNP / gene	*bla*$_{CTX-M-15}$	A0A075VKM9			129	1	-	13	33	-	327.5	24.2	1	23.4	1
		Kmer	*bla*$_{CTX-M-24}$	NC_022078.1 (127606 - 127633, reverse)			129	1	-	13	33	-	327.5	25	772	23.4	1480
	ISEcp1 tnpA	Kmer	Linked to *bla*$_{CTX-M-15}$	EU418923.1 (10812 - 10842, reverse)			129	1	-	12	34	-	365.5	25.9	1	26.6	1
	HP from ISEcp1	SNP / gene					129	1	-	13	33	-	327.5	24.2	1	23.4	1
	protein wbuC	SNP / gene		AIG86706.1			129	1	-	13	33	-	327.5	24.2	1	23.4	1
CRO	*β-lactamase genes*	SNP / gene	*bla*$_{CTX-M-15}$	AIG86707.1			140	1	-	2	33	-	2310	32.8	1	60.5	1
		Kmer	*bla*$_{CTX-M-24}$	NC_022078.1 (127606 - 127633, reverse)			140	1	-	2	33	-	2310	35.4	762	60.5	803
	HP from ISEcp1	SNP / gene					140	1	-	2	33	-	2310	32.8	1	60.5	1
	protein wbuC	SNP / gene		AIG86706.1			140	1	-	2	33	-	2310	32.8	1	60.5	1
	ISEcp1 tnpA	Kmer	Linked to *bla*$_{CTX-M-15}$	EU418923.1 (10812 - 10842, reverse)			140	1	-	1	34	-	4760	36.7	1	76	1

Continued on the next page

Drug	Gene	Study	Variant	Genome position (in reference genome) or BLAST accession		Alleles	Type	Ctrl 1	Ctrl 2	Ctrl 3	Case 1	Case 2	Case 3	Odds ratio	-log$_{10}$ *P*	Rank	-log$_{10}$ *P* LMM	Rank LMM
				LMM	Pre-LMM													
CIP	Plasmid-mediated quinolone resistance genes	SNP / gene	*aac(6')-Ib-c*	ACV60575.1				138	4	-	8	26	-	112.1	20	5	16.7	4
		SNP / gene	*qnr-B1*	A0A075VJL2				140	2	-	9	25	-	194.4	20.7	2	19.5	2
		Kmer	*qnr-B19*	JX298080.1 (481 – 511)				130	2	-	9	25	-	180.6	25	1846	19.5	4423
	tnpA (truncated)	Kmer	Linked to *qnr-B19*	JX298080.1 (1520 – 1550)				135	7	-	3	31	-	199.3	27.3	1	28.5	1
	tnpA	SNP / gene	Linked to aac	BAD08693.1				131	11	-	2	32	-	190.5	23.5	1	19.9	1
GEN	aac	SNP / gene	*aac(3)-II*	AHI38985.1				145	0	-	0	31	-	NA	36.8	1	>100	1
		Kmer	*aac(3)-II*	AJD77170.1(383 – 413)				145	0	-	0	31	-	Inf	39.4	1	15.6	397987
	ant	SNP / gene	*aac(6)*	AIG86041.1				141	4	-	5	26	-	183.3	22.1	6	18.6	7
	tmrB_2 (Tunicamycin resistance protein)	Kmer	Linked to aacA4 (314520 – 315119, complement)	CP011314.1 (310752 – 310782)				145	0	-	2	29	-	Inf	36.5	519	146.4	1
TOB	aac	SNP / gene	*aac(3)-II*	AIG86707.1				140	0	-	5	31	-	Inf	30.4	1	43.3	1
		Kmer	*aac(3)-II*	LK391770.1 (21132 - 21162)				140	0	-	5	31	-	Inf	33	1	43.3	1
	ant	SNP / gene	*aac(6)*	AIG86041.1				139	1	-	7	29	-	575.6	25.6	3	27	3

Table A.2 ***Klebsiella pneumoniae*** **GWAS results.** CFZ = Cefazolin, CIP = Ciprofloxacin, CRO = Ceftriaxone, CXM = Cefuroxime, GEN = Gentamicin, TOB = Tobramycin. Case = phenotypically resistant, control = phenotypically sensitive. aac = Aminoglycoside N-acetyltransferase genes, ant = Aminoglycoside N- acetyltransferase genes, aph = Aminoglycoside O-phosphotransferase genes. Causal gene names are coloured according to their resistance causing mechanism, red if its presence determines resistance, blue if substitutions within the gene causes resistance.

Drug	Gene	Study	Variant	Genome position (in reference genome) or BLAST accession		Alleles	Type	Ctrl 1	Ctrl 2	Ctrl 3	Case 1	Case 2	Case 3	Odds ratio	-log$_{10}$ *P*	Rank	-log$_{10}$ *P* LMM	Rank LMM
				LMM	Pre-LMM													
EMB	*embB*	SNP	M306L, M306V	4247429		A, G, C	NS	1586	2	1	24	16	1	528.7, 66.1, 0.1	25.6	2	82.8	1
		Kmer		4247429 - 4247459				31	1558	-	31	10	-	0.006	130.2	1	107.5	1
	rpoB	SNP	S450L, S450W	761155		C, T, G	NS	1563	23	3	17	23	1	91.9,	27.5	1	45.9	2
														30.6				
														0.333				
INH	*katG*	SNP	S315T	2155168		C, G	NS	1468	2	-	86	153	-	2475.8	151.1	1	169.4	1
		Kmer		2155145 - 2155175	2155145 - 2155175			1468	2	-	87	152	-	1282.4	220.9	1	172.4	1
PZA	*pncA*	SNP	V125G	2288868	2288868	A, C	NS	1662	0	-	41	4	-	Inf	7.2	142	60	1
		Kmer		2288847-2288877	2288847-2288877			1662	0	-	41	4	-	Inf	33.3	7890	60	1
		Kmer						1	1661	-	7	38	-	0.003	50.2	174	25.7	653
	rpoB	SNP	S450L, S450W	761155	761155	C, T, G	NS	1632	28	2	23	21	1	53.2, 35.5, 0.7	22.3	1	54.4	2
RIF	*rpoB*	SNP	S450L, S450W	761155	761155	C, T, G	NS	1486	0	1	34	49	3	Inf, 131.1, 0	73.2	1	269.8	1
		Kmer		761136 - 761166	761136 - 761166			7	1480	-	70	16	-	0.001	250	1	0	1
		Kmer		761126 - 761156	761126 - 761156			6	1481	-	66	20	-	0.001	237.2	14	321.7	1

Table A.3 ***Mycobacterium tuberculosis*** **GWAS results.** EMB = Ethambutol, INH = Isoniazid, PZA = Pyrazinamide, RIF = Rifampicin. Case = phenotypically resistant, control = phenotypically sensitive. Causal gene names are coloured according to their resistance causing mechanism, red if its presence determines resistance, blue if substitutions within the gene causes resistance.

Drug	Gene	Study	Variant	Genome position (in reference genome) or BLAST accession		Alleles	Type	Ctrl 1	Ctrl 2	Ctrl 3	Case 1	Case 2	Case 3	Odds ratio	-log$_{10}$ *P*	Rank	-log$_{10}$ *P* LMM	Rank LMM
				LMM	Pre-LMM													
CIP	*grlA*	SNP / gene	S80F, S80Y, F80Y	1419998		C, T, A	NS, NS, NS	745	5	0	14	226	2	2405.3, Inf, Inf	198.6	1	138.2	1
		Kmer			1419968-1420098			745	5	-	18	224	-	1854.2	190.5	1	99.1	9
		Kmer		1419995-1420025				22	728	-	228	14	-	0.002	177.2	46	113.6	1
	gyrA	SNP / gene	S84L	7255		C, T	NS	745	5	-	20	222	-	1653.9	190.3	2	81.6	2
		Kmer		7726 - 7756				6	744	-	223	19	-	0.0007	188.3	23	82.3	47
ERY	*ermC*	SNP / gene	Q2FDD1					774	2	-	103	113	-	424.6	85	1	193.5	1
		Kmer		NC_022228.1 (279-309)	NC_022228.1 (743-781)			770	6	-	102	114	-	143.4	94.7	1	192.4	12
	ermA	SNP / gene	P0A0H3					775	1	-	121	95	-	608.5	70.7	12	75.7	3
	ermT	SNP / gene	clusters with ermA in pan-genome															
	Intergenic 30bp upstream to *ermC*	Kmer		AE002098.2 (75634-75664)				772	4	-	106	110	-	200.3	93.1	92	195.7	1
FUS	*fusA*	SNP / gene	L461-, L461S, -461S	601084		T, A, C	Nonsense, NS, read-through	908	0	-	83	1	-	Inf	11	9754	62.3	1
		kmer		601054-601084				1	907	-	18	66	-	0.004	41.6	115876	157.4	1
	fusC	SNP / gene	Q8GNY5					908	0	-	40	44	-	Inf	54.3	4	19.3	39
		kmer		NC_002953.3 (53216-53246)				908	0	-	36	48	-	Inf	119.9	75	39.7	58
	fusB	SNP / gene	Q8GNY5					908	0	-	76	8	-	Inf	9.6	18494	16.6	76
	SAS0040	Kmer						908	0	-	35	49	-	Inf	122.5	1	45.4	36
	SAS0037	SNP / gene						908	0	-	38	46	-	Inf	57	1	28.5	21
	SAS0040	SNP / gene						908	0	-	38	46	-	Inf	57	1	28.5	21

Continued on the next page

Drug	Gene	Study	Variant	Genome position (in reference genome) or BLAST accession		Alleles	Type	Ctrl 1	Ctrl 2	Ctrl 3	Case 1	Case 2	Case 3	Odds ratio	-log$_{10}$ *P*	Rank	-log$_{10}$ *P* LMM	Rank LMM
				LMM	Pre-LMM													
GEN	*aacA/aphD*	SNP/ gene	P0A0C1					981	0	-	2	9	-	Inf	21.1	1	380.8	1
		Kmer	-	AY971367.1 (727-2683)				981	0	-	2	9	-	Inf	177.4	1	380.8	1
	GNAT acetyltransferase	SNP/ gene	D2J631					981	0	-	2	9	-	Inf	21.1	1	380.8	1
MET	*mecA*	SNP/ gene	P60185					773	4	-	3	212	-	13656.3	209.6	1	374.9	1
		Kmer		NC_022604.1 (78438- 78468)				772	4	-	3	213	-	13703	208.2	1	375.6	1
	HP in SCC-mec	SNP/ gene						773	4	-	3	212	-	13656.3	209.6	1	374.9	1
PEN	*blaZ*	SNP/ gene	P00807					145	23	-	28	796	-	179.2	118.6	1	140.1	1
		Kmer		NC_022604.1 (2824752-2824782)				143	25	-	9	815	-	518	166.4	2	210.5	2
	blaI			NC_022604.1 (2822414-2822444)				142	26	-	7	817	-	637.4	167.8	1	216.2	1
RIF	*rpoB*	SNP/ gene	H481Y	592271		C, T	NS	983	1	-	3	5	-	1638.3	11.1	1	158.6	1
		kmer		592260-592290				2	98	-	6	2	-	0.0007	122	1	177.9	1
TET	*tetK*	SNP/ gene	B0FYM6					945	1	-	9	37	-	3885	58	2	315.4	2
		Kmer		KM281803.1 (1198-1228)				945	1	-	4	42	-	9922.5	192.6	1	464.1	1
	tetL	cluster with *tetK* in the pan-genome																
	tetM																	
TRI	*repC*	SNP/ gene	Q5701					944	2	-	6	40	-	3146.7	62.9	1	364.6	1
	dfrB	SNP/ gene	F99Y	1497290		A, T	NS	308	0	-	10	5	-	Inf	7.9	1	28.5	1
		Kmer		1497269- 1497299				308	0	-	10	5	-	Inf	23.8	1	28.5	1
	dfrG	SNP/ gene						308	0	-	12	3		Inf	4.9	2	16.1	1
	dfrA	clusters with dfrG in pan-genome																
	orfu1	SNP/ gene						308	0	-	12	3		Inf	4.9	2	16.1	1
	LPXTG surface protein	SNP/ gene						301	7	-	9	6		28.7	5.6	1	13.5	2

Table A.4 *Staphylococcus aureus* GWAS results. CIP = Ciprofloxacin, ERY = Erythromycin, FUS = Fusidic acid, GEN = Gentamicin, MET = Methicillin, PEN = Penicillin, RIF = Rifampicin, TET = Tetracycline, TRI = Trimethoprim. Case = phenotypically resistant, control = phenotypically sensitive. Causal gene names are coloured according to their resistance causing mechanism, red if its presence determines resistance, blue if substitutions within the gene causes resistance.

8.2 Appendix B

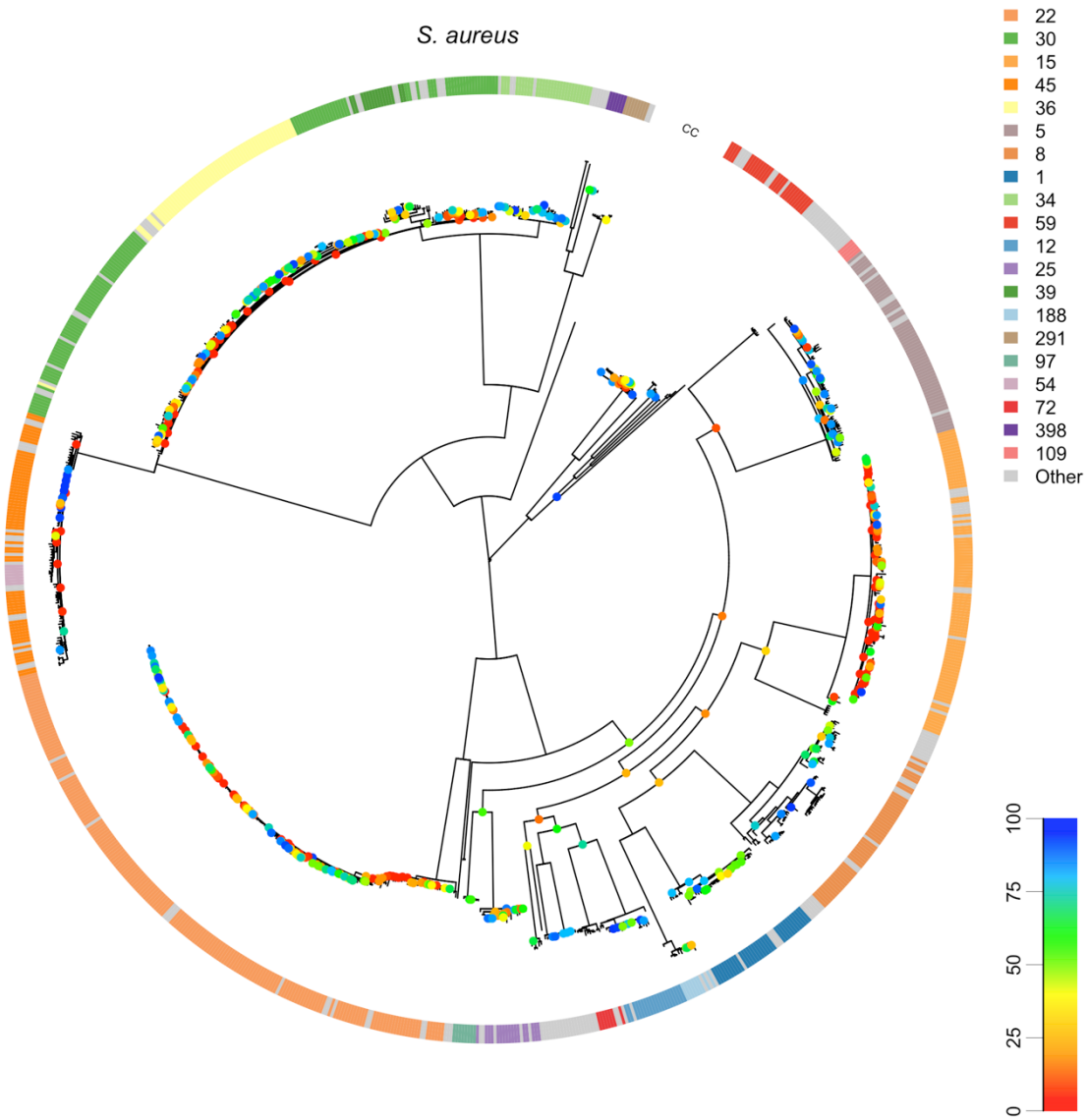

Figure B.1 Bootstrap support for the maximum likelihood estimated *S. aureus* RAxML phylogeny. One hundred rapid bootstraps were performed using RAxML.The outer ring displays clonal complex, and nodes with <95% support are coloured according to the rainbow colour legend to depict their node support.

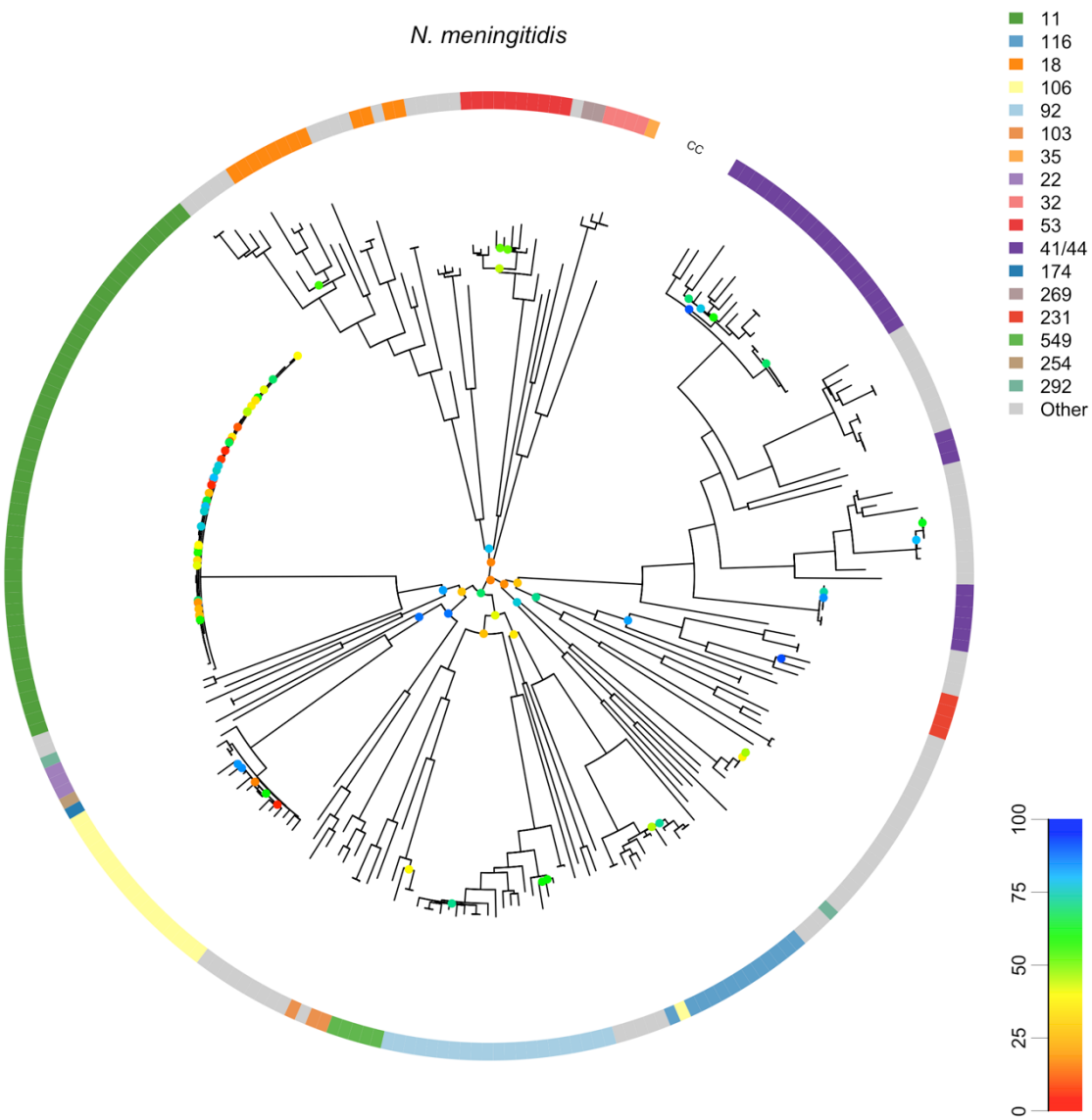

Figure B.2 Bootstrap support for the maximum likelihood estimated *N. meningitidis* RAxML phylogeny. One hundred rapid bootstraps were performed using RAxML. The outer ring displays clonal complex, and nodes with <95% support are coloured according to the rainbow colour legend to depict their node support.

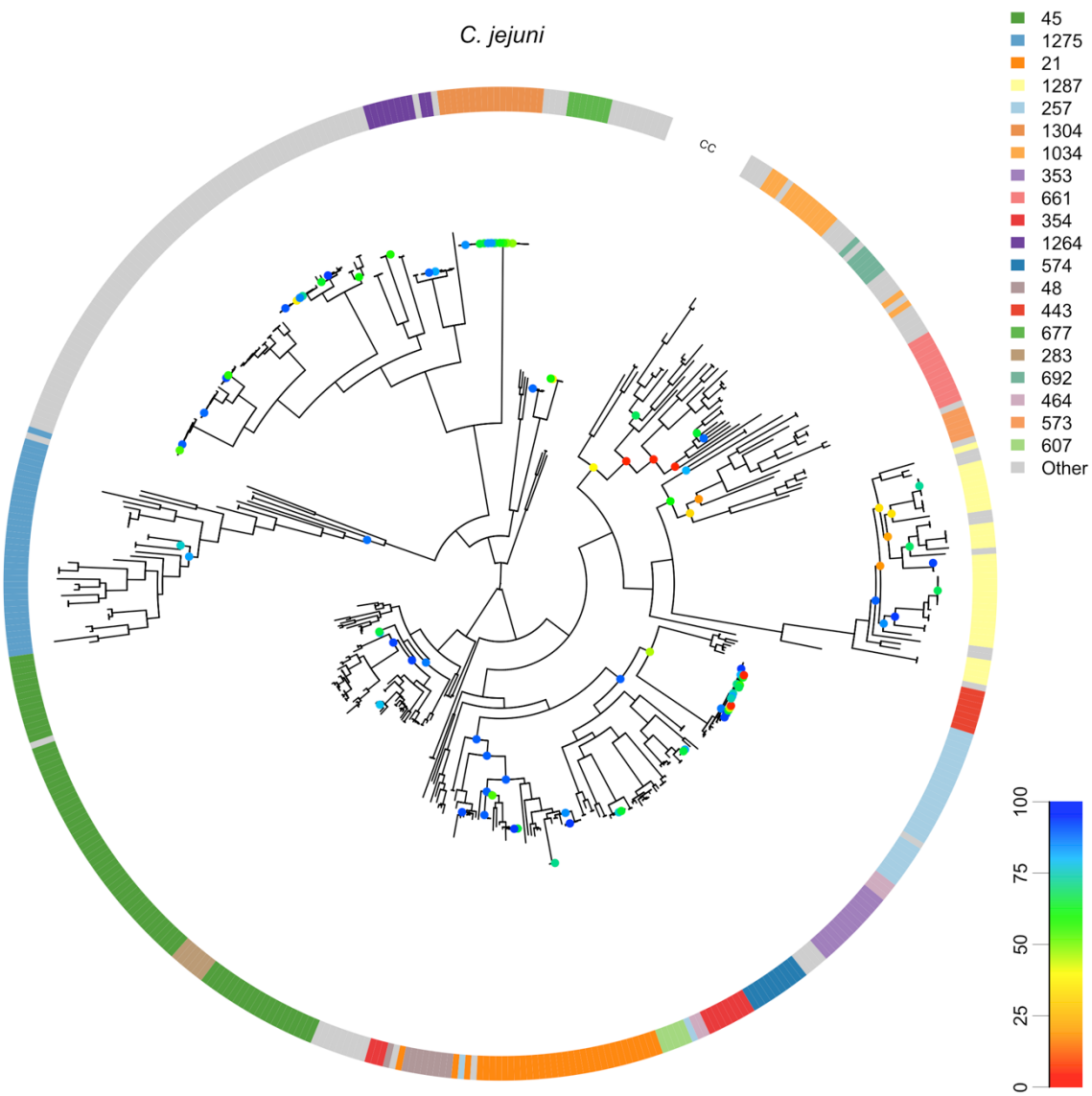

Figure B.3 Bootstrap support for the maximum likelihood estimated *C. jejuni* RAxML phylogeny. One hundred rapid bootstraps were performed using RAxML.The outer ring displays clonal complex, and nodes with <95% support are coloured according to the rainbow colour legend to depict their node support.