Chapter 4

Bacterial genetics contributing to invasive pneumococcal disease

Declaration of contributions

Stephen Bentley, Diederik van de Beek and Julian Parkhill supervised this work. Diederik and Arie van der Ende's groups designed and ran the MeninGene study on which this work is based, performed sample collection and DNA extraction. Philip Kremer performed the reciprocal best hits analysis. Ana Manso performed the long-range PCR to determine *ivr* allele type directly from clinical samples. I performed all other analyses.

Publication

The following has been published as:
Lees, J. A., Kremer, P. H. C., Manso, A. S., Croucher, N. J., Ferwerda, B., Serón, M. V., Oggioni M. R., Parkhill J., Brouwer M. C., van der Ende A., van de Beek D., Bentley, S. D. Large scale genomic analysis shows no evidence for pathogen adaptation between the blood and cerebrospinal fluid niches during bacterial meningitis. *Microbial Genomics*. 2017; 3. doi:10.1099/mgen.0.000103
Lees, J. A., Brouwer, M., van der Ende, A., Parkhill, J., van de Beek, D., Bentley, S. D. Within-Host Sampling of a Natural Population Shows Signs of Selection on Pde1 during Bacterial Meningitis. *Infection and Immunity*. 2017; 85(3). doi:10.1128/IAI.01061-16

4.1 Introduction

This chapter deals with the contribution of variation in the pathogen genome to bacterial meningitis. Using a GWAS framework, I first wished to test whether any variation in the pneumococcal genome is associated with susceptibility to meningitis or with a poor outcome of the infection. To study this, I used isolates collected as part of the MeninGene study (section 1.1.4). This part of the study consists of 3 089 pathogen isolates from the culture-proven cases of bacterial meningitis from the Netherlands. DNA was extracted from CSF and blood cultures and sequenced with 100bp paired end reads on the Illumina HiSeq platform. Of these sequences, 1 984 were *S. pneumoniae*. 751 carriage genomes from Dutch adults and children were also sequenced to use as controls for susceptibility analysis.

I first catalogued all forms of variation to use as the loci to test in a GWAS (section 4.3). While k-mers cover most of this variation, I also included tests of SNPs and genes due to their more straightforward interpretation. Some forms of variation such as inverting repeats, CNVs, recombinogenic antigens cannot be captured by these methods, so I developed new techniques to call variants at these loci. While this covered all forms of common variation detectable by short reads in the pneumococcal genome, rare variants may also play a role in disease pathogenesis. I annotated the predicted effect of rare coding variants to choose which to use in burden tests.

Using the SNP variation to tag other forms of variation in the genome, I was able to estimate the heritability of each of these traits (the proportion of variance in the phenotype is explained by variation within the genome). Finding evidence for pneumococcal genetic variants contributing to invasiveness other than serotype, I then used the methods presented in chapters 2 and 3 to test whether any of the specific variants that I called were associated with susceptibility to or severity of meningitis.

Section 4.5 concerns pathogen variation that occurs over the course of a single infection. Croucher, Mitchell et al. (2013) have previously shown that in a single patient bacteria appeared to adapt to the distinct conditions of blood and CSF. These are very different niches from that of nasopharyngeal carriage where this variation is well documented (Cremers et al., 2014), not least because the bacteria are exposed to different immune pressures (Habets et al., 2012) and have less time over which to accumulate mutations.

It is possible that bacteria inhabiting the nasopharynx are already well adapted for CSF invasion. However, genetic variants that enable invasion of the CSF are not expected to be under positive selection, since invasion is an evolutionary dead end for the bacterium. Studies of carriage alone will therefore be unable to detect selection during invasion. Current knowledge on within-host variation during invasive disease is mostly focused at the serotype and MLST level, and lacks the resolution and sample size to be able to address this question (Brueggemann et al., 2003; del Amo et al., 2015; D. A. Robinson et al., 2001).

Though the only whole genome based study suggests there is no difference between blood and CSF populations (at the gene level) in *S. pneumoniae* (Kulohoma et al., 2015), larger sample sizes are needed to better answer this question.

I therefore wished to expand the analysis of Croucher, Mitchell et al. (2013) by including more cases of disease, and used 938 pairs of genomes from the blood and the CSF of the same patient, and 54 pairs from the nasopharynx and CSF of the same patient sequenced as part of the MeninGene study described above. This sample set included both *N. meningitidis* isolates and *S. pneumoniae* isolates, each of which was analysed separately. As isolate pairs are matched they are closely related; the issue of population structure affecting bacterial GWAS is no longer a problem. Variants between pairs can be grouped by functional effect and tested for association with a niche straightforwardly.

4.2 Quality control and processing

In this section I discuss initial QC of isolates in the collection, and evaluations of both assembly and variant calling software to be used throughout the chapter.

Using a single *S. pneumoniae* isolate, I compared the quality of three assembly methods that have previously been shown to perform well on bacterial genomes (Magoc et al., 2013): Velvet (Zerbino & Birney, 2008), SPAdes (Bankevich et al., 2012) and SOAPdenovo2 & MaSuRCA (Zimin et al., 2013). Statistics from this comparison are shown in table 4.1. I decided that the SPAdes pipeline provided good quality assemblies while being easy to run, so assembled all isolates in the collection with v3.5 of the software using default settings. Additionally I ran velvet on all samples, which when k-mer length is optimised and scaffolds are improved, gave similar results to SPAdes. I corrected the resulting velvet assemblies with SSpace and GapFiller (Page et al., 2016). The assembly result used for each purpose will be stated throughout the rest of the thesis.

	Velvet	SPAdes	SOAPdenovo2 & MaSuRCA
# contigs	48	**7**	**7**
Total length	2 096 048	**2 205 585**	2 139 022
N50	77 648	429 779	**481 453**
# genes	2 073	**2 208**	2 166
CPU time	6 h	7.2 h	5.5 h
Maximum memory	3.7 GB	7.0 GB	4.3 GB
Disk space	0.1 GB	0.6 GB	4.2 GB

Table 4.1: Assembly and annotation of *S. pneumoniae* isolate 11822_8_30. N50 is the median contig length. For each performance metric the best scoring method is in bold.

I then analysed the quality of the SPAdes assemblies using quast (Gurevich et al., 2013) and kraken (Wood & Salzberg, 2014). I performed this analysis at the sample level, rather than at the contig level. As the primary aim is a GWAS I desired complete and comparable assemblies, so the number of included samples at each variant is the same. I found two assemblies which were predominantly another species, and discarded them. I also discarded five sequence runs with low yield, 17 with total lengths over 2.5Mb, two with total lengths under 1.8Mb and one with a GC content of 31.4%. This left 1 144 CSF isolates and 674 pairs of blood and CSF isolates for downstream analysis. For the carriage samples I removed 29 isolates contaminated with another species (determined by kraken, and the position on a preliminary core gene alignment phylogeny), and 8 isolates which showed evidence of being mixed samples (number of heterozygous SNPs in preliminary mapping was greater than two standard deviations above the mean). This left 693 carriage isolates for downstream analysis.

To compare variant calling methods I produced a set of true variant calls for 30 samples. I did this by simulating evolution of *S. pneumoniae* genomes along the branch of the tree between *S. pneumoniae* R6 (Hoskins et al., 2001) and the common ancestor with *S. mitis* B6 (Denapaite et al., 2010). The rates in the GTR matrix and insertion/deletion frequency distributions were estimated as in section 2.3.1. I created an average of 10 000 mutations with these rates, and Illumina paired end read data at 200x coverage simulated using pIRS (Hu et al., 2012).

Method	True positives	False negatives	False positives
bcftools	24922	900	244
freebayes	22253	3569	1465
GATK	25024	798	191

Table 4.2: Performance of variant calling algorithms on simulated data. True positives are SNPs or INDELs correctly called; false negatives are variant sites which were missed by the caller; false positives are sites without variation but called as a variant.

I mapped the reads with bwa-mem (H. Li, 2013), followed by samtools fixmate, sort and markdup. I then called variants using bcftools, freebayes (Garrison & Marth, 2012) and GATK (Van der Auwera et al., 2002). The results are shown in table 4.2. freebayes performed poorly due to its use on multiple nucleotide polymorphism (MNP)s, which were difficult to compare to the simulations. GATK performed the best on all measures, and in particular achieved much better power at calling indels. I used it for calling SNPs and indels throughout, unless otherwise stated.

4.3 Catalogue of all pneumococcal variation

In this section I detail how I catalogued population level variation in the pneumococcal genome. These variants are then used throughout the rest of this chapter as the predictor variable in GWAS with various phenotypes of interest, analysis of within-host variation and in chapter 5 as the phenotype in a genome to genome analysis. As discussed in section 2.2, variation in bacterial genomes is not well represented by short changes compared to a linear reference due to extensive variation of the accessory genome (Donati et al., 2010; McInerney et al., 2017), mosaic alleles created by recombination (Hanage et al., 2009), structural variation (Croucher, Coupland et al., 2014; Manso et al., 2014) and copy number variation (Howden et al., 2015). I used different techniques to determine the variation present in each sample from each of these sources to ensure maximum discovery power of the GWAS performed.

While short variants (i.e. SNPs and small indels) with respect to a single linear reference only partially covers the variation present in the pneumococcal population, it is still a useful dataset to produce. A genome alignment produced this way can be used to generate the phylogenetic relationship between all samples from the population and create discrete related clusters. Both of these are useful for QC, heritability analysis and evaluating population structure. Additionally, the effect of these variants on protein function can be straightforwardly predicted, making conclusions drawn from them more easily interpreted and also of use in indirect tests of association section 4.4.2.

I produced a whole genome alignment in two ways. Firstly I mapped reads to the ATCC 700669 reference using bwa mem with default settings

```
bwa mem reference.fa forward_reads.fastq reverse_reads.
    fastq | samtools fixmate -O bam - > output.bam
```

and finally marked duplicate reads in these binary sequence alignment/map (BAM) files using Picard. I then called variants from each of these BAM files separately using samtools mpileup and bcftools call, and as a population using GATK HaplotypeCaller. I then applied hard quality filters to each of these call sets to create initial calls. To select variants based on a correctly scaled sensitivity and specificity I used GATK VariantRecalibrator to scale the variant quality scores. This tool requires known true positive calls as a prior – I used the intersection of hard filtered variants from GATK and bcftools with 90% confidence (Q10), and filtered variants from the Maela and Massachusetts studies with 68% confidence (Q5) as recommended. After recalibration, I applied 99.9% power as the cut-off for variants to maximise sensitivity at this stage. Finally, I annotated the predicted consequence of all passing variants with variant effect predictor (VEP) (McLaren et al., 2010).

I also produced a core-genome alignment using roary (Page et al., 2015) with a 95% blast ID cut-off. Roary efficiently performs all by all alignment using every annotated protein in the dataset. Those matches with over 95% ID are assumed to be orthologs and are clustered and undergo multiple sequence alignment. Using a single cut-off will mean that some genes with orthologous function but without sequence homology (for example different alleles of a gene) will not be clustered together, and that some genes without orthologous function but with sequence homology will be incorrectly clustered. We chose the cut-off of 95% based on having the best balanced accuracy of these two error classes when using reciprocal best BLAST hits to define true orthologs (Ward & Moreno-Hagelsieb, 2014). As well as core genes (present in at least 99% of samples) roary also clusters accessory genes into COGs, which I later used as a variant in association. In this case the annotated function helped determine whether the cluster is showing presence or absence of gene or groups of different alleles of a gene that is being tested for association.

I counted k-mers using fsm-lite (section 2.2), which required 75Gb RAM and 14hrs CPU time to count all informative k-mers with a minor allele count (MAC) of ten or more. In this sample set there were 11.7M informative k-mers with 2.6M unique patterns. I called CNVs from the BAM files produced above using cn.mops (Klambauer et al., 2012) which fitted the coverage of mapped reads in 1kb windows with a mixture of Poisson distributions, and determined the most likely integer coverage value for each sample in each window. I extracted the inferred copy number from those windows which had support for a CNV from more than one sample.

4.3.1 Allelic variation of three pneumococcal antigens

I wished to determine whether sequence variation of pneumococcal antigens is associated with virulence and disease outcome. As well as being plausible GWAS hits, these antigens vary rapidly (Croucher, Vernikos et al., 2011), meaning sequence variation is not population stratified, which increases discovery power. Conversely, while the k-mer approach (section 2.2) either directly assays or indirectly tags most variation in the population, variation of these antigens may not be captured by this method. For example, *pspC* can be difficult to assemble due to repeats and copy number variation (Iannelli et al., 2002), and therefore k-mers from the gene sequence will not appear in the assembly, and not be counted or tested. In *pspA* and *zmpA*, mapping of k-mers may not be specific to the allele sequence due to sequence homology with orthologous and paralogous genes (Hollingshead et al., 2000; Bek-Thomsen et al., 2012).

Here I consider *pspC*/*cbpA*, *pspA* and *zmpA*, which have all been shown to have interactions with the host immune system (Croucher et al., 2017), but have variability that may not be assessed by the methods discussed above. I needed to develop a way first to classify possible alleles, then determine the allele of each sample from short read sequence data. For the latter issue, de novo assembly (followed by a BLAST with a set of reference alleles) is unreliable for completely reconstructing the gene sequences, but usually contained some information about the allele present. Alternatively mapping the sequence reads to a set of reference alleles is less affected by repeat sequences and may be more accurately used to find the allele of genes (Inouye et al., 2014), but determining the closest match is non-trivial. I decided to use a method which combines summary statistics from both of these techniques to determine the allele type. This has previously shown to be advantageous for antibiotic resistance typing from genomic data; Hunt et al. (2017) designed a method using combination of assembly and mapping which had improved type I and type II error rate over either technique alone.

I defer discussion of the variability and construction of a reference panel specific to each of these alleles until the sections below, and first discuss the typing method I applied to determine the allele of all three antigens given such a reference panel. I first generated statistics from the assemblies of all samples by running blastp between the annotated genes in both the velvet and SPAdes assemblies and the reference panel. From this, I extracted the % ID, number of mismatches, number of gaps, E-value and bitscore between the two assemblies of sample and every possible reference. For mapping I used srst2 (Inouye et al., 2014) in a mode which maps reads to all reference sequences, and reports information about coverage over every possible allele. I used the coverage, number of SNP mismatches, number of indel mismatches and number of truncated bases.

This led to a data frame with 16 predictors for every reference sequence, per sample (for example *pspC* had 48 references, so there were 768 predictors). When a match was not reported by blastp or srst2 I filled in value with the minimum reported value of the predictor (or maximum for the number of mismatch fields), and removed predictors without variation.

To produce labelled training data I performed the same process on the reference panel itself, for each sequence using blastp against all the reference sequences and srst2 with simulated reads (these were error-free 100bp reads with 200x coverage and 350bp insert size with 80bp standard deviation (s.d.)). In all cases, on the test data simple variance analysis showed these statistics could be used to predict classification of alleles successfully (fig. A.13). I fitted a classifier to this training data (see section 4.3.1 for details), then finally used the trained model to predict the allele for all samples. The results are shown in fig. 4.1. As expected, all the antigens show some, but not total, concordance with background genotype. I used the above process for typing all antigens; I now discuss the specifics of constructing the reference panel for each antigen.

pspC/cbpA allele

The *pspC* gene, also known as *cbpA*, *hic*, *spsA* or *pbcA*, is paralogous to *pspA* and is known to have a number of immunogenic functions. These include binding host proteins C3, CFH and IgA, all of which are involved in the immune response to pneumococcal colonisation

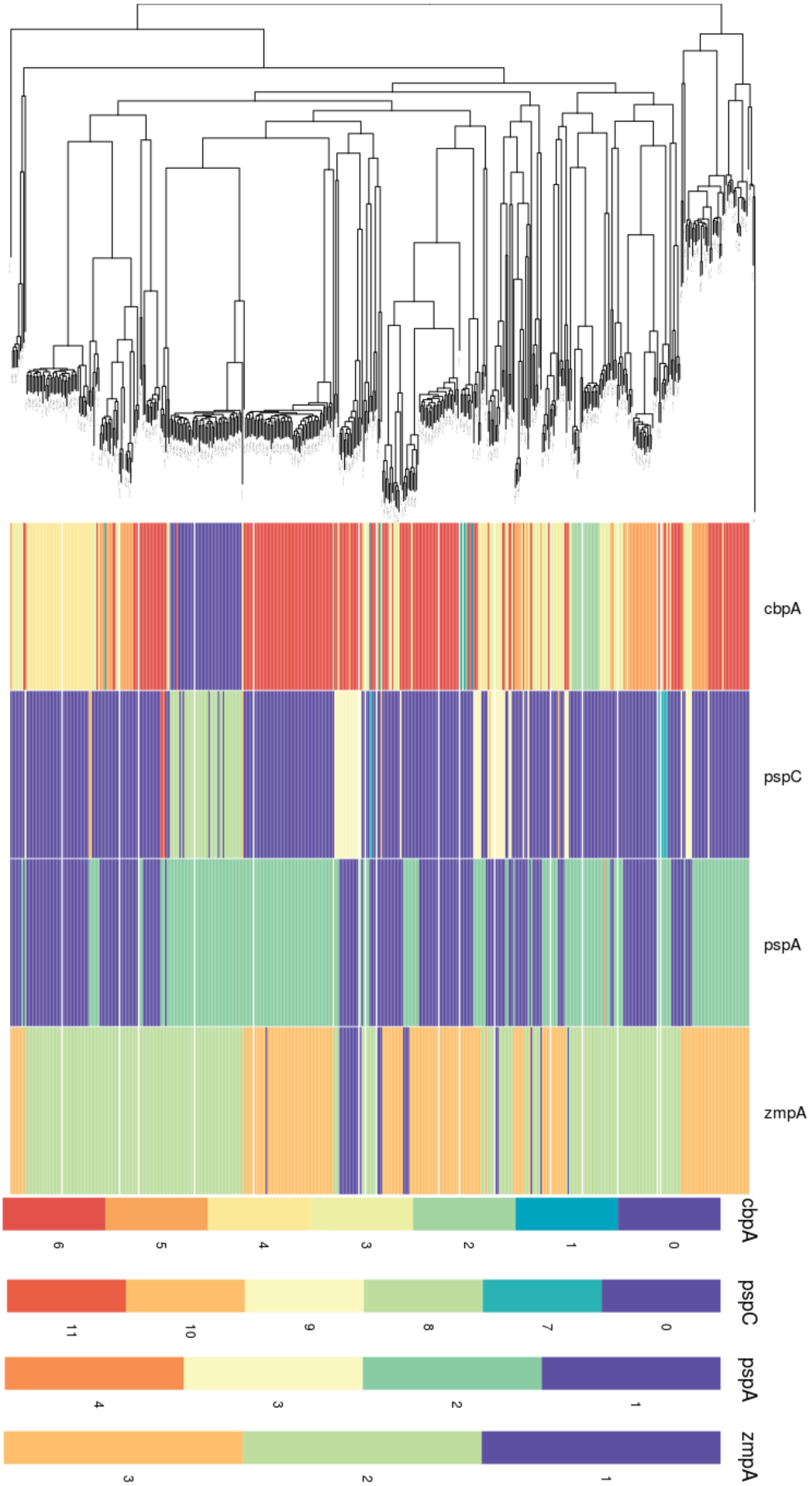

Figure 4.1: The inferred allele of pneumococcal antigens *zmpA*, *pspA* and *pspC*. Left: phylogenetic tree of CSF isolates. Right: tips coloured by the inferred allele for three antigens, and key. The first two columns are alleles 1–6 and 7–11 of pspC, which may have two copies present (Iannelli et al., 2002).

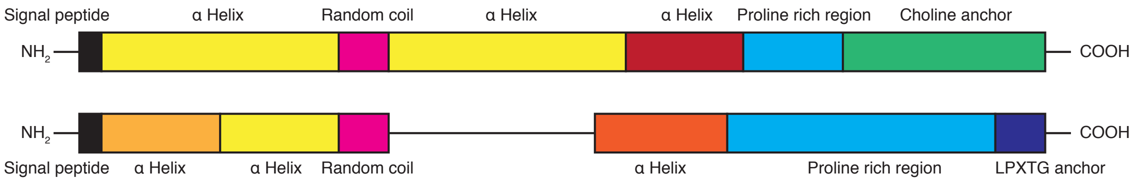

Figure 4.2: Pictographic alignment of the two forms of PspC, as in Iannelli et al. (2002). The top shows cbpA-5; all alleles *cbpA* 1-6 have a choline anchor, and otherwise vary in their α helix content. The bottom shows pspC-7; all alleles *pspC* 7-11 have an LPXTG anchor instead of a choline anchor.

(Brooks-Walter et al., 1999). The locus encoding PspC varies extensively, and two main forms exist (fig. 4.2) which are distinguished by having a choline anchor (alleles 1-6) or a LPXTG anchor (alleles 7-11) (Iannelli et al., 2002). Each genome may encode neither, one or both of these forms and they are normally found in tandem.

I used the existing classification of 11 alleles described by Iannelli et al. (2002), and the 48 sequences reported by these authors (fig. A.10). To allow for the fact that each of the two forms may be present or absent I trained two classifiers. The first, referred to as the *cbpA* allele, used alleles 1–6 and treated 7–11 as missing. The second, referred to as the *pspC* allele, used alleles 7–11 and treated 1–6 as missing. Though there was correlation between the two allele types (for example 4 and 10 were more likely to co-occur) I trained the two classifiers independently. I first checked whether the reference data could distinguish between the labels using PCA, and then predicted two different alleles for each sample.

I tried four different 'out of the box' classifiers: support vector machine (SVM) with a linear kernel, weighted k-nearest neighbours, random forests and DAPC (Jombart et al., 2010). I inspected the statistics and annotations to manually assign the allele pair for 25 genomes from across the tree, then using these truth values and compared the classification accuracy of each method. Table 4.3 shows that the SVM performed best; I used it for all four classifiers. Inspection of the feature importance showed the blastp bitscore, E-value, and number of mismatches as well as the srst2 number of truncated bases and number of mismatches were the most informative predictors.

Method	Balanced accuracy
SVM	0.86
kknn	0.73
Random forest	0.50
DAPC	0.14

Table 4.3: Comparison of classifiers of antigen alleles. The balanced accuracy is given by the average of $\frac{1}{2}(\text{sensitivity} + \text{specificity})$ for all alleles.

***pspA* and *zmpA* alleles**

PspA is a well studied pneumococcal antigen (Crain et al., 1990) which binds C3 (Tu et al., 1999) and lactoferrin (Shaper et al., 2004). Its locus is involved in both ancestral and recent recombination events which has created variation at the locus (Hollingshead et al., 2000; Croucher, Harris, Fraser et al., 2011). ZmpA, also known as Iga, is a zinc metalloprotease which cleaves IgA molecules (Wani et al., 1996). Similarly to *pspA*, the sequence is variable within the population and is under diversifying selection (Bek-Thomsen et al., 2012).

Croucher et al. (2017) have manually created clusters of sequences for both of these antigens using 616 carriage genomes (Croucher, Finkelstein et al., 2013). Sequences were combined into the same allele if their translated sequence was identical, giving 39 possible sequences for *pspA* and 18 possible sequences for *zmpA*. I used these sequences as the reference panel for each allele.

Unlike *pspC* where sequences had been further clustered based on functional domains by Iannelli et al. (2002), this reference panel contained very similar sequences with different allele labels. Using this directly for GWAS would lead to low power as the number of sequences with each allele would be very small, and the classification would also likely be poor due to the relative paucity of reference data for each allele. To avoid this I used the phylogentic relationship between sequences to clustered similar sequences into allele groups before training each classifier.

For both antigens I aligned the reference panel of amino acid sequences using MUSCLE (Edgar, 2004), and built a phylogeny with RAxML with a CAT+GAMMA model. To test the robustness of these phylogenies I ran 100 maximum-likelihood bootstrap replicates, and 10^6 mrbayes Markov-chain Monte Carlo (MCMC) iterations (discarding the first 25% as burn-in, sampling every 10^3 steps) to generate a sample of 750 trees from the posterior distribution. I compared the topology of these trees using treescape (Kendall & Colijn, 2015), and found the placement of ancestral branches of the topology were poorly resolved, though placement of sequences in main clades was well supported. I therefore took a cut through the deep branches of the two phylogenies, defining four alleles for *pspA* (fig. A.11) and three alleles for *zmpA* (fig. A.12). This phylogeny and classification is similar to three families previously defined for *pspA*, and three families previously seen for *zmpA*. Using these alleles I then fitted classifiers to the reference panels as in section 4.3.1, and predicted the allele for all samples in the study.

4.3.2 Phase variable type I R-M system allele (*ivr*)

Croucher, Coupland et al. (2014), J. Li et al. (2016), Manso et al. (2014) have highlighted a potential role in virulence for the *ivr* locus, a type I restriction-modification system with a phase-variable specificity gene allele of *hsdS* in the host specificity domain (fig. 4.3).

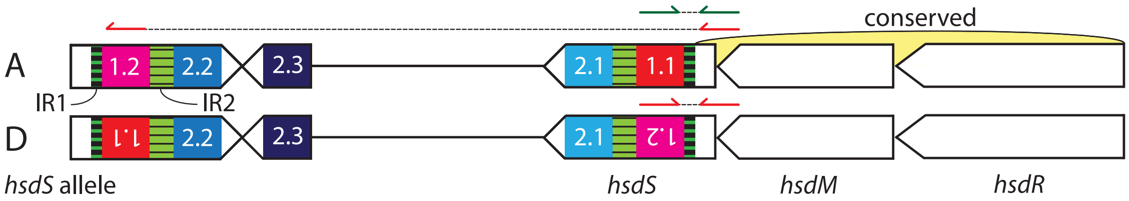

Figure 4.3: The structure of the *ivr* type I restriction-modification locus. The restriction (*hsdR*) and methylation (*hsdM*) subunits, and the 5' end of the specificity subunit (*hsdS*) are generally conserved. Inverted repeats IR1 (85bp) and IR2 (333bp) facilitate switching of downstream incomplete *hsdS* elements into the transcribed region. Top: The green read pair has the expected insert size, and suggests allele A (1.1, 2.1) is present. The red read pair is in the wrong orientation and has an anomalously large insert size. Bottom: The red read pair is consistent with the displayed inversion, suggesting allele D (1.2, 2.1) is present.

There are six possible different alleles A-F for *hsdS*, each corresponding to a different level of capsule expression. Some of these alleles are more successful in a murine model of invasion, whereas others are more successful in carriage.

Due to the high variation rate and structural rearrangement mediating the change the allele cannot reliably be determined using assembly and/or standard mapping of short read data. Instead, I extracted mates of reads mapping to the reverse strand of the conserved 5' region for each sample, and mapped with BLAT (Kent, 2002) to the possible alleles in position 1. This forms a vector r_i of length two for each sample i, with the number of reads mapped to 1.1 and 1.2. Similarly, to determine the 3' allele (position 2), I extracted pairs of reads mapping to each of the reverse strand of allele 1.1 and the forward strand of allele 1.2 and mapped to the three possible alleles in position 2. This forms a vector q_i of length six for each sample i, with the number of reads mapped to each allele A-F.

I performed this on all samples in the collection and found 677 of 693 carriage samples and 1 052 of 1 144 invasive CSF samples had at least one read mapping to an allele of the *ivr* locus *hsdS* gene. In the invasive samples, this corresponded to 621 CSF blood sample pairs. Those without any reads mapping had either a deletion of one component of the locus, or a large insertion mediated by the *ivr* recombinase.

4.4 GWAS of bacterial variants associated with meningitis

While it is well known that pneumococcal serotype contributes to invasive propensity (Hausdorff et al., 2000; Brueggemann et al., 2003), it is of great interest in the field of pneumococcal biology whether variation in other regions of the genome can independently affect invasiveness. Many virulence factors are known to be involved in and essential for pneumococcal colonisation and disease (Kadioglu et al., 2008), but whether natural variation in these regions affects clinical cases of disease has yet to be assessed. Indeed, the overall role of pneumococcal variation in invasive disease is as yet unknown, and therefore the proportion of variation in invasiveness which can be ascribed to the capsule and the proportion due to other factors cannot be determined. Additionally, the lack of large cohorts combining detailed clinical metadata with bacterial data means that little is known about about the effect of pneumococcal variation on disease outcome. Previous studies with small sample sizes have suggested a role for platelet binding (Tunjungputri et al., 2017) and arginine synthesis (Piet et al., 2014), with additional evidence from *in vitro* observations.

I first performed a heritability analysis to quantify the amount of variation due to the pneumococcal genome for each phenotype. As well as using the methods described in section 3.3 I also applied a phylogenetic mixed model assuming an Ornstein-Uhlenbeck (OU) process of trait evolution as implemented in the patherit package (Mitov & Stadler, 2016), which has previously been shown to be less biased than other techniques for estimating the heritability of pathogen traits (Blanquart et al., 2017). I performed 200 000 MCMC iterations, discarding the first half as burn-in and thinning the chain to every hundredth value. LDAK performs heritability estimation of this binary trait on the liability scale (Lynch & Walsh, 1998). I peformed this analysis within genomes collected from meningitis, stratified using GOS to define clinical outcome, and between genomes from carriage and genomes from meningitis (referred to as 'invasiveness').

Trait	Method		
	LDAK	OU	closest phylogenetic-pairs (CPP)
Invasiveness	0.983 ± 0.003	0.9936 (0.9928-0.9943)	0.995 (0.991-0.998)
Unfavourable outcome	0.006*	did not converge	0.05 (-0.04-0.16)
Death	0.0001*	0.02 (-0.07-0.11)	0.07 (-0.03-0.17)

Table 4.4: Estimated heritability of pneumococcal invasiveness and outcome due to variation of the pathogen genome. Values shown in brackets are the 95% CIs, where provided by the method, for LDAK the standard error is shown, unless the LRT p-value was > 0.05 so there is no support for a non-zero heritability (shown by an asterisk).

Table 4.4 shows the predicted heritability from each method. There is evidence that invasive propensity is highly heritable, but that disease outcome is not determined by natural variation of pathogen genetics. The latter is not surprising as invasive disease as an evolutionary dead end for the pathogen, adaptations affecting virulence over the short course of infection are unlikely to be selected for. The dependence on invasiveness is well known to depend on pneumococcal genetics, but not the degree. The high heritability estimated here, supported by three different techniques, suggests that in this population some bacteria are able to invade while others are not, with almost certainty depending on the genetic background. This is consistent with some serotypes not being found in invasive disease (Hausdorff et al., 2000), and their wide genetic separation from invasive serotypes. The complete heritability is likely an overestimate due to the binary nature of the trait, but does show that pathogen genetics are important in invasiveness and not likely to be important in severity.

I then wished to quantify the amount of this heritability which was due to serotype, which is the current focus of pneumococcal vaccination and the most well known invasiveness determinant, versus other factors. As in section 3.4.1 I used leave-one-out cross validation with lasso logistic regression to select the 36 serotypes (of 63 observed) which were informative of invasiveness. I then assessed the variance in invasiveness explained by these serotypes using Nagelkerke's pseudo R^2 from logistic regression (International Schizophrenia Consortium et al., 2009; Hosmer et al., 2013), which was 0.45. Caution should be used in directly interpreting this R^2 as variance explained, but it does show the model fit from serotype alone is not as good as using the pneumococcal kinships, suggesting there are factors other than serotype which affect invasiveness. I also checked whether invasiveness is well predicted by capsule charge, as has been previously suggested by Y. Li, Weinberger et al. (2013). Using the previously measured zeta potentials, and using the serogroup average when a serotype charge was not available, I performed the same logistic regression using charge as the predictor rather than serotype. Charge significantly affected invasiveness but was not as informative as the specific serotype ($p < 10^{-10}$; Nagelkerke's $R^2 = 0.08$), suggesting a role for finer structure of the capsule structure (Bentley et al., 2006).

In the rest of this section, using the variation defined for all samples as in section 4.3 and the GWAS methods developed in chapters 2 and 3, I tried to find the pneumococcal variants other than serotype which affect invasivness. Even though there is no evidence from the above heritability analysis that variation in the pneumococcal genome contributes to disease outcome I ran the same analysis on these phenotypes anyway – it may be that the common/core variation used to produce these estimates fails to tag variation in the accessory genome or phase variable regions which may contribute to outcome. In this case a lack of association will also provide further support for zero heritability due to the bacterial genome.

In the first section I consider association of common variants in the pan-genome (all of those described in section 4.3) with the phenotypes predominantly using the techniques already described. I then go on to asses the role of rare variation firstly using tests of selection, and more directly using an association combining variants with the same predicted effects. Finally I developed a model to test whether any particular *ivr* allele, or the amount of variation of the allele is associated with any of the phenotypes.

4.4.1 Role of common variation

Using the variants catalogued above, with previously described filtering thresholds, I performed a GWAS between the isolates from invasive disease and asymptomatic carriage, as well as unfavourable outcomes and/or death within the invasive isolates. I used SEER with the first ten MDS components to correct for population structure, as well as FaST-

LMM (Lippert et al., 2011) using the kinship matrix estimated from SNPs and INDELs as random effects.

Figure 4.4 shows the Q-Q plots of the resulting p-values from these methods on SNPs and k-mers with invasive versus carriage isolates. In both cases the test statistic from SEER is clearly highly overinflated for this population and phenotype, meaning a high significance threshold would be needed to remove population structure confounded associations. I have shown that invasiveness is highly heritable, so population structure being highly confounding is unsurprising. Increasing the number of fixed effect population structure covariates may help alleviate this issue, but as the LMM test statistic is better controlled, and as it was a successful method in chapter 3, I have used it for all associations of common variants with the three phenotypes. For significance thresholds I used the unique number of patterns as the number of tests in a Bonferroni correction, giving $p < 8.2 \times 10^{-7}$ for SNPs and $p < 1.9 \times 10^{-8}$ for k-mers. However, inspection of the Q-Q plots shows that for k-mers the LMM is still overinflated, so I have instead taken $p < 1 \times 10^{-16}$ to describe the top hits.

From all three of SNPs, COGs and k-mers by far the most highly associated variants are transposons. These mobile elements of DNA can insert into different places in the bacterial host genome through inverted repeat sequences, and coevolve with the bacterial population (Kleckner, 1981; Levin & Moran, 2011). In some cases transposons can carry cargo genes, such as antibiotic resistance conferring mechanisms, which increase host fitness (Croucher, Harris, Fraser et al., 2011). However, the transposons here appear to be simple elements lacking such cargo, and are therefore unlikely to explain a difference between carriage and invasive isolates directly. Most likely these transposons are present in some genetic backgrounds and not others, and are therefore a population structure confounded result. Their variability in position in the genome and specific sequence may mean they are less well controlled for against genetic background. Due to the lack of plausible functional link with the phenotype I do not consider them further here.

Other hits are shown in table 4.5, ordered by the variant type discovered. In some cases COGs were incorrectly clustered and actually represent two alleles of the a gene orthologs. For three of these alleles I found a positive association with invasive isolates from one allele, and a negative association from the alternative allele. To annotate the genes here I used the best blastp match to the core and accessory genome defined by Croucher, Finkelstein et al. (2013), and if not annotated already I used blastp with the `nt/nr` database to find annotated orthologs, and hmmscan and cd-hit to find functional domains to inform the annotation.

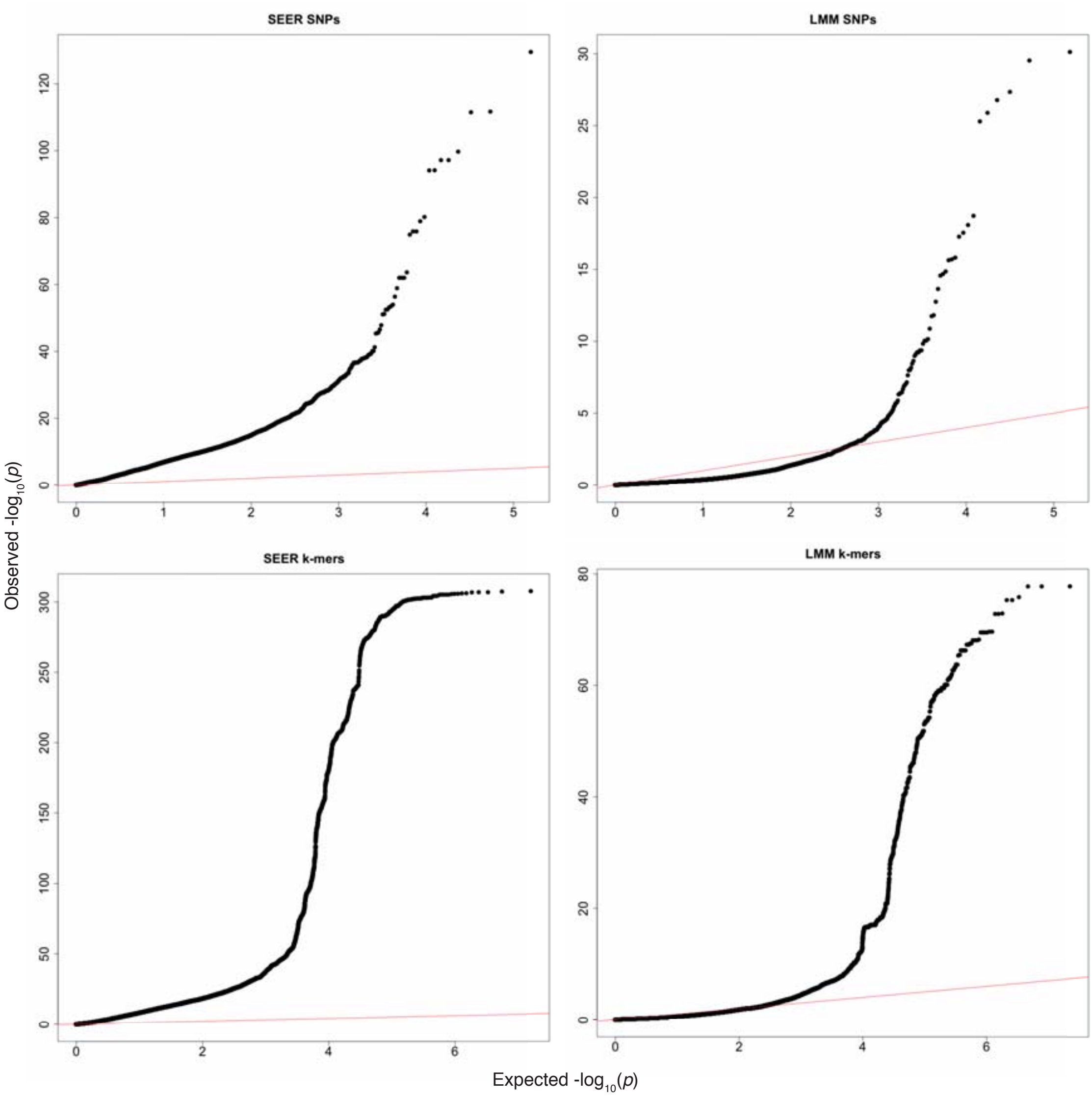

Figure 4.4: Quantile-Quantile plots for invasive *S. pneumoniae* GWAS methods. Red line is for observations following the null-hypothesis of no association, plotted points are observed p-values from each method. Top row: p-values from SNPs and INDELs from mapping; bottom row: p-values from k-mers. Left column: SEER run with the first ten population structure components. Right column: FaST-LMM run on the same input.

Gene ID	Annotation	Core	Method	p-value
FM211187.6011	*tlyC*; Membrane protein (upstream)	Yes	Mapped variants	7.7×10^{-31}
FM211187.977	*pbpX*; penicillin binding protein	Yes	Mapped variants	3.6×10^{-18}
FM211187.313	hypothetical protein (upstream)	Yes	Mapped variants	2×10^{-16}
FM211187.1802	*yhfE*; Aminopeptidase (upstream)	Yes	Mapped variants	1.0×10^{-9}
FM211187.1019	*wzh*; capsule synthesis	No	Mapped variants	3.6×10^{-9}
FM211187.150	*comA*; bacteriocin/competence (upstream)	Yes	Mapped variants	9.9×10^{-9}
FM211187.3083	*pbl3e*/*pldT*; bacteroicin	No	COG absent	4.0×10^{-10}
N/A	transcriptional regulator (pseudogene)	No	COG absent	1.4×10^{-8}
FM211187.3090	bacteriocin precursor	No	COG absent	1.7×10^{-8}
FM211187.6181	FtsX-family transport protein (ABC transporter permease)	No	COG alleles	4.7×10^{-9}
FM211187.6189	C4-dicarboxylate (citrate) ABC transporter	Yes	COG alleles	1.4×10^{-7}
FM211187.5843	23S rRNA (uracil-5-)-methyltransferase RumA2	Yes	COG alleles	5.5×10^{-7}
FM211187.939	galactose-6-phosphate isomerase	No	K-mers	3.0×10^{-60}
N/A	phage-related chromosomal island protein	No	K-mers	3.0×10^{-60}
FM211187.4259	Peptidase U32	Yes	K-mers	1.7×10^{-59}
FM211187.4090	*aroK*; Shikimate kinase	Yes	K-mers	1.7×10^{-59}
FM211187.1923	*yehU*; Sensor kinase	Yes	K-mers	3.1×10^{-59}
FM211187.6369	*patA*; efflux pump (upstream)	Yes	K-mers	2.0×10^{-54}

FM211187.6823	*tauA*; Nitrate/sulfonate/taurine ABC transporter solute-binding protein	Yes	K-mers	1.9×10^{-43}
FM211187.213	Galactose uptake PTS transporter, IIB subunit	Yes	K-mers	2.5×10^{-42}
FM211187.3677	*pyrB*; Aspartate carbamoyltransferase PyrB	Yes	K-mers	4.6×10^{-38}
FM211187.6594	*ulaA*; Pentose PTS transporter IIA	Yes	K-mers	3.7×10^{-25}

Table 4.5: Common variation associated with invasiveness using FaST-LMM. I have annotated the gene the significant locus overlaps, and intergenic variants are annotated with the nearest downstream genes as noted. Gene ID is the name in the ATCC 700669 reference if present; 'core' refers to whether this gene was in the core genome defined by Croucher, Finkelstein et al. (2013); method describes the type of variant that was found to be associated.

The *wzh* gene is involved in capsule synthesis and is part of the gene cassette which determines serotype (Bentley et al., 2006). As shown above and in previous studies, serotype has a large effect on invasiveness and hence this association serves as a positive control. The association of variants in *pbpX* is likely due to mosaic alleles which confer resistance to β-lactams being common in invasive serotypes, similar to what I found in section 3.5.2. The bacteriocins mediate intraspecies competition and determine strain fitness (Dawid et al., 2007), but a specific association with invasiveness independent of strain background has not previously been reported. *comA*, a core gene essential for competence, affects the expression of these bacteriocins so may represent an effect through the same pathway (Kjos et al., 2016).

The adhesin *yhfE* has previously been associated with virulence of *S. pneumoniae* (M. W. Robinson et al., 2013). This adhesin functions as a peptidase, hence the other peptidase may found to be associated also have similar role. Other genes found here previously associated with virulence in animal models include: *ulaA* which utilises ascorbic acid has been found to be upregulated in invasion (Afzal et al., 2015; Mahdi et al., 2015); *pyrB* is involved in cell wall biosynthesis and can affect virulence (Mohedano et al., 2005); *aroK* is involved in biofilm formation (Domenech et al., 2012); both *comA* and *tauA* were found to be essential for growth during meningitis using a genome-wide screen (Molzen, Burghout, Bootsma, Brandt, van der Gaast-de Jongh et al., 2011). For the other identified regions I couldn't find reference to a previous report relating them to a role in invasiveness or virulence of *S. pneumoniae*.

For unfavourable outcome and death, none of the above classes of variant reached genome-wide significance. This is consistent with the low heritability estimated for these phenotypes. No alleles of *pspC*, *pspA* or *zmpA* or any CNVs reached genome-wide significance for any of the phenotypes.

4.4.2 Role of rare variation

The availability of whole genome sequence data for these samples allows the identification of rare variants, here defined as those present in the population with MAF $< 1\%$, which are also plausible as having an effect on the phenotypes of interest. The amount of rare variation compared to common variation present in a population is informative of recent selection and population size changes (Ziheng Yang, 2006). An overall difference may therefore be informative of different selection on regions of the genome depending on the niche. In fig. 4.5a I have plotted the SFS by niche and predicted consequence to look for an overall difference. Across the range of common MAFs in both niches the proportion of synonymous/nonsynonymous/intergenic/LoF mutations is roughly constant and as expected (Ziheng Yang, 2006; Thorpe et al., 2017), though at low frequencies, there is an excess of potentially damaging variants.

Interestingly, there is a clear excess of rare variants in invasive samples compared to carriage samples. To quantify this difference and identify which regions of the genome are responsible for the excess of rare alleles I calculated Tajima's D for each coding sequence in the genome, and looked for differing signs of selection between cases and controls. Tajima (1989) developed the summary statistic D to look for differences between an observed population and an idealised population of a stable size evolving under neutral selection, where mutation frequency is dominated by drift rather than selection. By comparing the number of segregating sites with the average number of differences between pairs of sequences, a statistic D can be calculated. Deviations with $D < 0$ are indicative of selective sweeps and/or recent population expansion, whereas $D > 0$ is indicative of balancing selection and/or recent population contraction. In terms of differences between SFS, a negative D manifests as an excess of rare variants whereas a positive D manifests as a uniform distribution (Bamshad & Wooding, 2003).

For speed, I implemented code in C++ (https://github.com/johnlees/tajima-D) which uses the same optimised strain-wise distance calculation as SEER (section 2.3.2) to calculate the average number of pairwise strain differences $\hat{k}$. Unknown or gap sites are ignored in the calculation, and the codes produces the same value of D on standard test data. The code uses a variant call format (VCF) file as input, so is readily generalisable to other applications. Using this code, I calculated D for all coding sequences in the ATCC 700669 reference separately for carriage and invasive isolates, and the difference in D between niches.

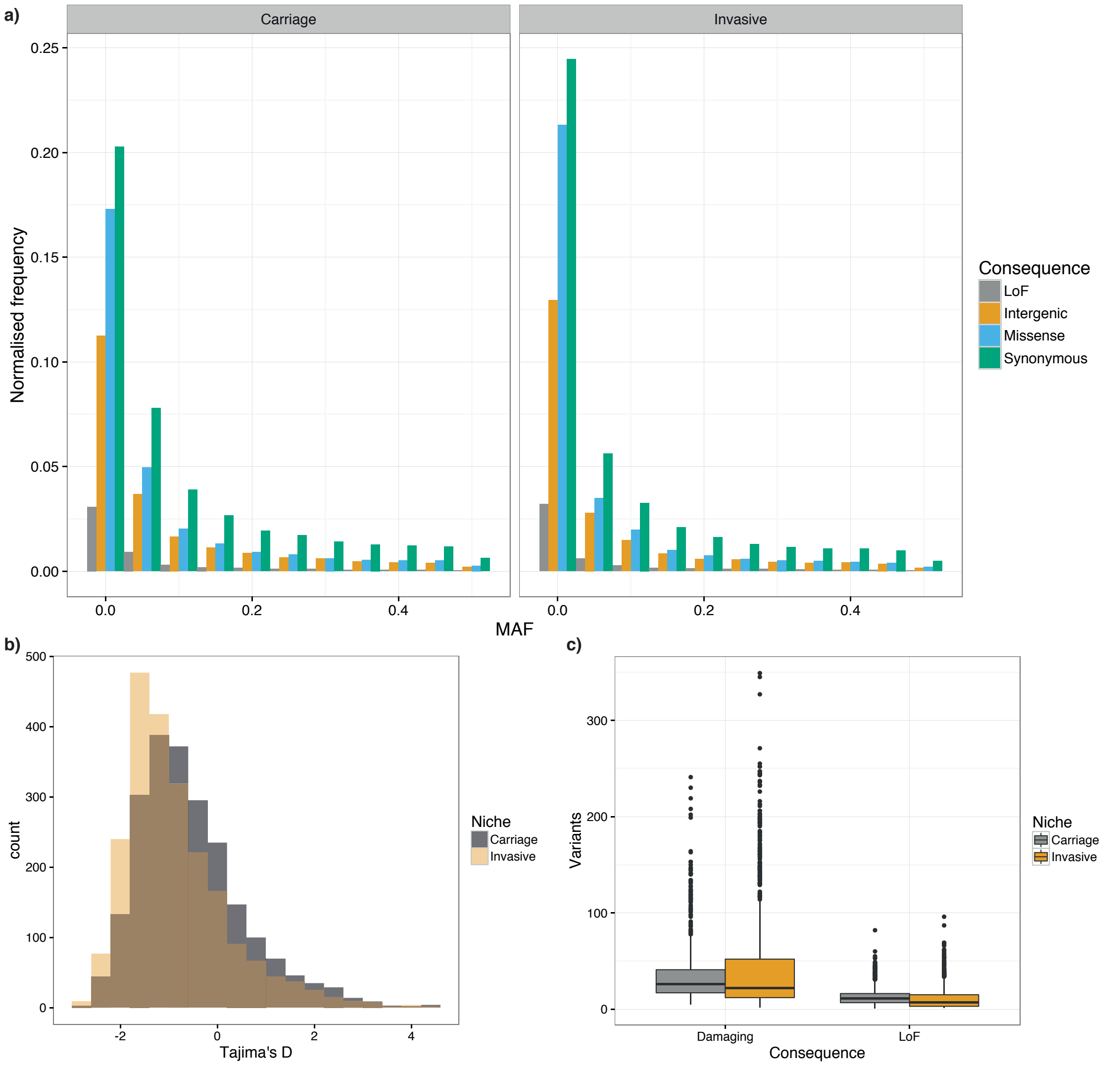

Figure 4.5: Differing burden and frequency of rare variation between invasive and carriage isolates, based on short variation called from mapping against the ATCC 700669 reference genome. LoF are frameshift or nonsense mutations. **a)** The SFS stratified by niche and by predicted consequence. Frequency has been normalised with respect to the number of samples in each population. **b)** Histogram of Tajima's D for all coding sequences in the genome, stratified by niche. **c)** Boxplot of number of rare variants per sample, stratified by niche and predicted consequence. Damaging mutations are LoF mutations and missense mutations predicted damaging by SIFT.

Comparison between *D* values to test for different selection between niches will only work within the same population, otherwise changing population size may cause an overall difference in *D*. The assumption that invasive and carriage populations are the same is potentially reasonable, as all invasive isolates must first have been carriage isolates, however the biased selection of case isolates used for GWAS and potential adaptation and population growth after invasion (described futher in section 4.5) may violate this assumption. In GWAS terms, although the calculation of Tajima's D uses rare variation, which is less prone to population structure confounding, common variation is also used which is affected by population structure.

To test for an overall difference I compared the distributions of *D* by gene in each phenotype shown in fig. 4.5b. Genes in invasive isolates had a lower average *D* (difference in medians -0.34; W = 1 996 100, $p < 10^{-10}$) and a more positively skewed *D* (difference in skewness 0.30; 95% bootstrapped CI 0.17-0.44). This difference in *D* may be representative of a difference in population dynamics or population structure between niches, or may show genuine differences in selection. To find individual genes which show a difference in selection between niches I then ran 44 000 permutations per gene with randomised phenotype labels to calculate a p-value on the difference in *D* between niches, to which I applied a Bonferroni correction to adjust for testing of all genes (Winantea et al., 2006). 156 genes had a significantly different *D* between niches; in table 4.6 I report 18 of these coding sequences which were outside of the 95% central mass of the *D* distribution for one niche but not the other. Due to potential population structure effects results should therefore be seen as suggestive, and potential for follow-up work.

Gene ID	Annotation	Invasive *D*	Carriage *D*	Direction
FM211187.1040	*wzx*; capsule synthesis	-2.53094	-1.79867	Negative in invasive
FM211187.5843	23S rRNA (uracil-5-)-methyltransferase RumA2	-2.4028	-1.63478	Negative in invasive
FM211187.2360	*ezrA*; septation ring formation regulator	-1.1051	-2.17726	Negative in carriage
FM211187.4024	replication initiator protein (on ICE)	-1.55767	-2.16733	Negative in carriage
FM211187.4026	hypothetical, contains FtsK gamma domain (on ICE)	-1.61993	-2.21525	Negative in carriage
FM211187.357	bacteriocin	4.19212	1.30796	Positive in invasive
FM211187.420	*tsaB*; tRNA threonylcarbamoyladenosine biosynthesis protein	3.49345	1.39805	Positive in invasive
FM211187.769	aceytltransferase	2.9055	1.80787	Positive in invasive
FM211187.1019	*wzh*; capsule synthesis	2.76882	1.63677	Positive in invasive
FM211187.1802	*yhfE*; Aminopeptidase	2.28654	1.19784	Positive in invasive

FM211187.1804	bacteroiocin	1.94491	-0.56384	Positive in invasive
FM211187.1806	*dacC*; D-alanyl-D-alanine carboxypeptidase	2.23028	0.722447	Positive in invasive
FM211187.5184	*dnaI*; primosomal protein	2.32212	0.197632	Positive in invasive
FM211187.3651	*tarI*; Ribitol-5-phosphate cytidylyltransferase	-0.146237	2.34171	Positive in carriage
FM211187.3804	*nanB*; neuraminidase	1.6805	3.19937	Positive in carriage
FM211187.5053	membrane protein	0.311619	2.46774	Positive in carriage
FM211187.5358	*secY*; accessory secretion system translocase	0.471641	2.36541	Positive in carriage

Table 4.6: Coding sequences with extreme values of Tajima's *D*, with a difference between carriage and invasive isolates as determined by permutation testing.

A positive *D* statistic implies common variants are being maintained in the population more than expected, suggesting that multiple alleles of the gene are common. The positive estimates of *D* in bacteriocins are consistent with their function, where having a different allele to competing strains is advantageous and increases fitness (Bogaardt et al., 2015; Miller et al., 2017). *nanB* is similarly involved in competition and in virulence (Shakhnovich et al., 2002; Brittan et al., 2012); the difference in *D* I found suggests that this selection may be more important in carriage where more common alleles appear to be maintained. A negative *D* suggests purifying selection acting on a gene. For example, *ezrA* is essential for growth in carriage (van Opijnen et al., 2009; Cleverley et al., 2014), so a negative *D* suggests that changes to the protein are not tolerated in this niche. As *wzx*, *wzh*, *yhfE*, RumA2 and bacteriocins were found to be associated with invasiveness above, this suggests that the difference in *D* I observed is less likely to be due to population stratification and more likely a real sign of selection. Genes found through these approach which may affect cell growth such as *ezrA*, *secY*, *dnaI* and *tarI* may make the population more or less immune stimulating, depending on their direction of effect.

Burden testing of coding sequences

I then wished to consider whether rare variants were associated with any of the three phenotypes. These variants will have occurred on terminal (or close to terminal) branches and therefore population structure is less of an issue than for common variants. Power to detect associations is proportional to MAF and OR, so and at low MAF, there is only power to detect those variants with a large effect size (Liu & Anderson, 2014). However, for rare alleles the statistical tests described so far lack the power to test for an association even for an infinite OR. In human genetics combining sets of variants with the same predicted effect on a more complex biological function (yet simpler than the whole phenotype), for example grouping rare LoF variants in the same gene, then testing the group for association with the phenotype of interest has been the most common approach (B. Li & Leal, 2008; Morris & Zeggini, 2010). This is known as a burden test – in bacterial genomes this technique has successfully found LoF variants associated with antibiotic resistance in *M. tuberculosis* (Desjardins et al., 2016).

In each test I used only variants with MAF $< 1\%$ from the variant calls derived from mapping. Using the annotations from VEP, I defined frameshift and stop gained mutations as LoF – 6 825 variants in total. I also analysed the effect of all predicted missense variants using Provean (Ng & Henikoff, 2003; Choi et al., 2012), and used the default threshold of -2.5 to select variants with a predicted effect on protein function – 26 206 of 50 383 missense variants passed this threshold. I combined these variants with LoF variants to define a damaging class. Figure 4.5c shows the overall burden of damaging rare variants between carriage and invasive samples; in both classes there was higher burden in carriage isolates (median LoF: invasive 7, carriage 11, W = 297 440, $p < 10^{-10}$; median damaging: invasive 22, carriage 26, W = 345 370, $p = 8 \times 10^{-4}$), so results showing a burden in carriage should be interpreted with caution.

I then used `plink/seq` to perform a burden test on all coding regions in the ATCC 700669 reference genome, which looked for an excess of rare damaging alleles in genes, and Bonferroni corrected all resulting p-values. I tested all six possible phenotypes: invasiveness, carriage, favourable outcome, unfavourable outcome, survival, death. For the latter four phenotypes based on clinical outcome no genes showed a significant burden of LoF or damaging variants. Table 4.7 shows the results for carriage and invasive isolates.

Gene ID	Annotation	p-value	Class	Direction
FM211187.1036	*wchV*; capsule synthesis	0.0022	LoF	Carriage
FM211187.1143	membrane protein	0.0022	LoF	Carriage
FM211187.1634	*bglG*; transcription anti-terminator	0.0022	LoF	Carriage
FM211187.3315	*zmpD*; zinc metalloprotease	0.0022	LoF	Carriage
FM211187.4588	*pclA*; collagen-like surface-anchored protein	0.0022	LoF	Carriage
FM211187.4679	platelet binding phage protein	0.0022	LoF	Carriage
FM211187.4714	prophage protein	0.0022	LoF	Carriage
FM211187.4939	membrane protein	0.0022	LoF	Carriage
FM211187.5113	*nanA*; neuraminidase	0.0022	LoF	Carriage
FM211187.5328	uncharacterised repeat protein	0.0022	LoF	Carriage
FM211187.5369	PsrP glycosyltransferase	0.0045	LoF	Carriage
FM211187.6773	*dusB*; tRNA-dihydrouridine synthase	0.0045	LoF	Carriage
FM211187.1025	*wze*; capsule synthesis	0.0067	LoF	Carriage
FM211187.4017	hypothetical protein (on ICE)	0.0067	LoF	Carriage
FM211187.1040	*wzx*; capsule synthesis	0.0089	LoF	Carriage
FM211187.92	cell wall-binding amidase/autolysin (pseudogene)	0.0089	LoF	Carriage
FM211187.6861	*comFC*; competence	0.011	LoF	Carriage
FM211187.6608	*pcpA*; choline binding protein	0.016	LoF	Carriage
FM211187.4717	prophage protein	0.018	LoF	Carriage
FM211187.2642	chlorohydrolase	0.029	LoF	Carriage
FM211187.5374	PsrP glycosyltransferase	0.038	LoF	Carriage
FM211187.1804	bacteriocin	0.039	LoF	Carriage

FM211187.3950	conjugal transfer protein (on ICE)	0.042	LoF	Carriage
FM211187.3204	*ybaB*; DNA-binding protein	0.0089	Damaging	Carriage
FM211187.4311	multidrug transporter	0.050	Damaging	Carriage
FM211187.4424	sortase-sorted surface anchored protein (pseudogene)	0.0067	LoF	Invasive
FM211187.2661	*bceA*; ABC exporter ATPase	0.0045	Damaging	Invasive
FM211187.3585	*smc*; Chromosome partition protein	0.0045	Damaging	Invasive
FM211187.5524	*trpD*; anthranilate phosphoribosyltransferase	0.0045	Damaging	Invasive
FM211187.2550	*fruA*; Fructose PTS ABC transporter	0.027	Damaging	Invasive
FM211187.3460	*ispA*; Farnesyl diphosphate synthase	0.038	Damaging	Invasive
FM211187.2615	*pfkA*; ATP-dependent 6-phosphofructokinase	0.042	Damaging	Invasive

Table 4.7: Burden testing of rare LoF and damaging variants in coding sequences associated with invasive or carriage isolates. P-values are Bonferroni corrected using the total number of genes.

Those regions found with a larger number of LoF variants in carriage than disease represent genes which are advantageous in invasion, and hence include a number of well-known virulence factors. Specifically, capsule related genes, *zmpD* and *nanA* have all been previously described as increasing virulence in animal models (Brueggemann et al., 2003; Bek-Thomsen et al., 2012; Brittan et al., 2012) and have some overlap with associations found through common variant association. The large effect size caused by these LoF mutations is similar to the gene knock-outs used in these experiments.

As well as these well-described virulence factors, I found four more genes which were more likely to be functional in invasive isolates which had been previously described as virulence related in a single or small number of studies. PsrP is an adhesin which has been shown to increase virulence in mice (Obert et al., 2006; Shivshankar et al., 2009), and found here were two genes which affect the protein's function. *pcpA* (Glover et al., 2008; Sánchez-Beato et al., 1998) and *pclA* (Paterson et al., 2008) are choline binding and surface anchored proteins respectively, both previously associated with virulence. Tunjungputri et al. (2017) have reported association with presence of the phage-derived platelet binding protein PblB with 30-day mortality of meningitis in humans – the platelet binding protein found here may have a similar role in invasiveness (though I did not find it to be associated with severity or mortality).

I could not find previous reports of association with virulence or invasive potential of the other hits in this category. Also, few of the genes found to be essential in a mouse model of meningitis (Molzen, Burghout, Bootsma, Brandt, van der Gaast-de Jongh et al., 2011) were found here, suggesting either that the induced variants do not occur in natural populations, that the mouse does not perfectly model human meningitis or that the sample size here was too low to discover these effects.

Only one gene was found to lose function more frequently in invasive disease, though as it is a pseudogene in the reference this is unlikely to be a real functional effect. For missense variants affecting protein function the direction of effect is less clear, as the variants may be fitness increasing or decreasing. This inconsistent direction may also make the burden test less powerful, and a test which does not rely on this assumption such as the SKAT test may be preferred (Wu et al., 2011; S. Lee et al., 2012). In carriage isolates, including missense variants also found *ybaB* and a multidrug transporter to be significantly altered in carriage but not in invasion. In invasive isolates a few more possible hits were found. *smc* is involved in cell division and growth, but also has epistatic links to much of the rest of the chromosome (Skwark et al., 2017). LoF in *trpD* has previously been associated with attenuated virulence (Hava & Camilli, 2002), and *fruA* as being associated with the switch in virulence between nasopharyngeal colonisation and bloodstream invasion (Trappetti et al., 2017).

4.4.3 Hierarchical Bayesian model for ivr allele prevalence

Manso et al. (2014), J. Li et al. (2016) have reported an association with *ivr* allele and invasive propensity in a murine model; this dataset offers the opportunity to test whether such as an association exists in clinical samples. As the *ivr* varies rapidly and independently from population structure (Croucher, Coupland et al., 2014) a simple association test can be performed for each allele. I first used the mapping approached described in section 4.3.2 to determine the *ivr* allele for each sample. However, as even a single colony contains heterogeneity at this locus, simply taking the allele with the most reads mapping to it in each sample gives a poor estimate of the overall presence of each allele in the invasive and carriage niches. To take into account the mix of alleles present in each sample, and to calculate confidence intervals, I developed a hierarchical Bayesian model for the allele in each niche (fig. 4.6). This simultaneously estimates the proportion of each colony pick with alleles A-F for both individual isolates (π), and summed over all the samples in each niche (μ). The model is applied this over i samples and c niches (in this case c can be blood, CSF or carriage).

I first modelled the state of the 5' allele (TRD1.j) only. For the two possible alleles 1.1 and 1.2, the number of reads mapping to each allele (a 2-vector r_i) was used as the number of successes in multinomial distribution z_c (c – index for niche). From these I inferred the proportion of each allele in each individual sample π_i, and in each niche overall μ_c. This was done by defining Dirichlet priors expressing the expected proportion of an allele in a given sample π_i to be drawn from a Dirichlet hyperprior representing the proportion of the allele that is found in each niche as a whole μ_c. The κ parameter sets the variance of all the individual sample allele distributions π_{ic} about the tissue average μ_c, with a higher κ corresponding to a smaller variance.

The hyperparameter A_μ, which encodes the total proportion of each allele we expected to see over all samples, was set to the average amount of the allele observed from the long range polymerase chain reaction (PCR) in a subset of 53 paired samples, as described in section 4.5.4.

The observed number of reads mapping to each allele, prior distributions defined above, and structure of the model in fig. 4.6 defines a likelihood which can be used to infer the most likely values of the parameters of interest π and μ. I used `Rjags` to perform MCMC sampling to simulate the posterior distribution of these parameters. I used 3 different starting points (i.e. three chains), and took and discarded 30 000 burn in steps, followed by 45 000 sampling steps. Noticeable auto-correlation was seen between consecutive samples, so only every third step in the chain was kept when sampling from the posterior. I manually inspected plots of each hyperparameter value and mean at each point in the chain, as well as the Gelman and Rubin convergence diagnostic, which showed that the chains had converged over the sampling interval.

To model both the 5' end (TRD 1.1 and 1.2) and the 3' end (TRD 2.1, 2.2 and 2.3) together, so each isolate i is represented by an allele A-F, for each isolate the total number of reads mapping n_i was drawn from the distribution in equation eq. (4.1)

$$n_i \sim \sum_j \pi_{ij} \cdot r_{ij} \tag{4.1}$$

where j is the index of the TRD region, r_{ij} is the number of reads in sample i that had a mate pair downstream from TRD1.j mapping to any TRD2 region, and π_i is the posterior for allele frequency in the sample.

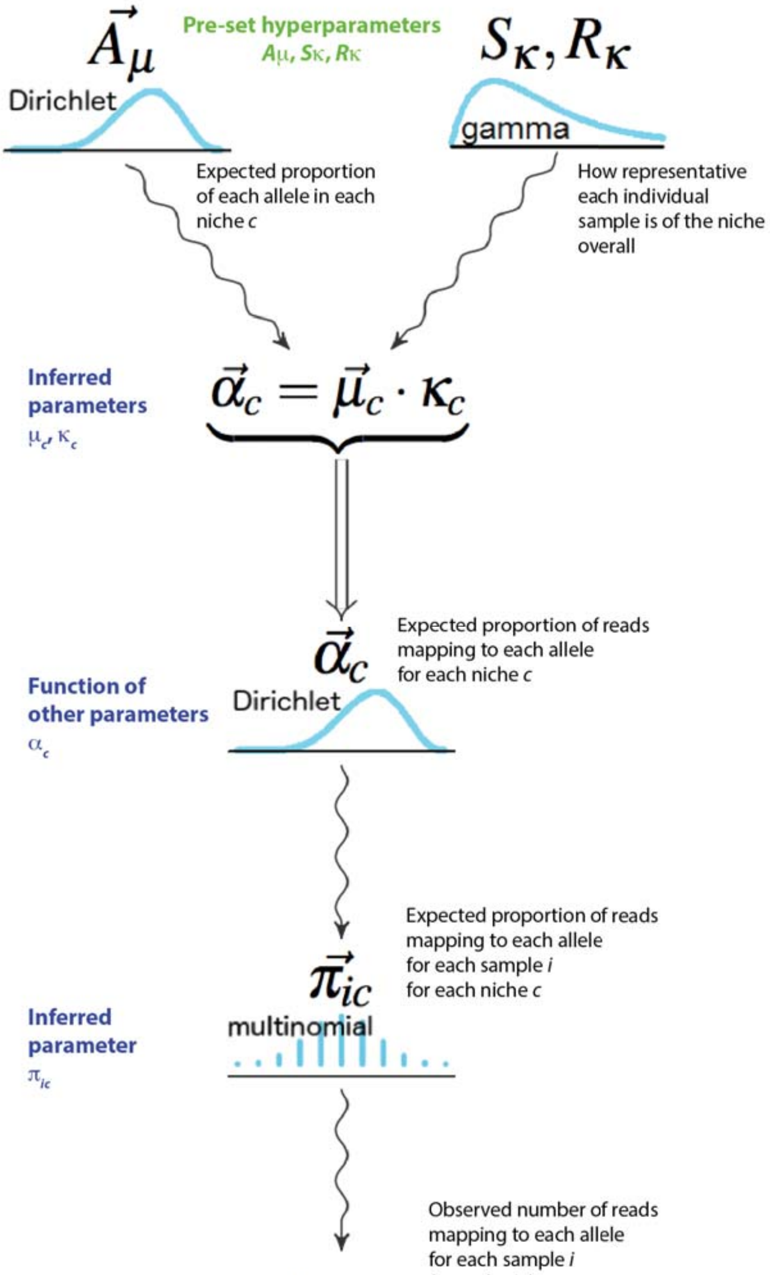
$\vec{A}_\mu$
Pre-set hyperparameters
Aμ, Sκ, Rκ
S_κ, R_κ
Dirichlet
gamma
Expected proportion of each allele in each niche c
How representative each individual sample is of the niche overall
Inferred parameters
μc, κc
$\underbrace{\vec{\alpha}_c = \vec{\mu}_c \cdot \kappa_c}$
$\vec{\alpha}_c$
Expected proportion of reads mapping to each allele for each niche c
Function of other parameters
αc
Dirichlet
Expected proportion of reads mapping to each allele for each sample i for each niche c
$\vec{\pi}_{ic}$
Inferred parameter
πic
multinomial
Observed number of reads mapping to each allele for each sample i for each niche c
Observed data
zic, nic
$\vec{z}_{ic}, n_{ic}$

Figure 4.6: Hierarchical model for *ivr* allele. Solid double arrows denote a deterministic relationship; wavy arrows represent a value drawn from a distribution. z is a vector of the number of reads mapping to each allele from a total of N reads mapping to the variable region; i is the sample number; c is an index for tissue type. μ_c, κ are hyperparameters for mean allele prevalence and how closely a sample is representative of a tissue type respectively. A_μ, B_μ are priors for allele prevalence in invasive disease. S_κ, R_κ are the shape and rate parameters for a gamma distribution, which were used to set a broad prior on κ.

The distribution for the number of reads mapping to each allele j, z_{ij} was then defined as in equation eq. (4.2)

$$z_{i,j} \sim \begin{cases} n_i \cdot \dfrac{q_{i,j}}{\|\vec{q}_i\|} \cdot \pi_{i,1.1}, & \text{if } j \in \text{A, B, E} \\ n_i \cdot \dfrac{q_{i,j}}{\|\vec{q}_i\|} \cdot \pi_{i,1.2}, & \text{if } j \in \text{C, D, F} \end{cases} \tag{4.2}$$

where q_i is a vector of length six which contains the number of reads mapped to each allele A-F as described above, and π, i and n are as previously. A single sample for z was taken for each isolate i. This 6-vector z_{ij} is then used as the observed data in the same model as above to infer π_i, and μ_c for the whole locus allele (A-F) rather than just the 5' end.

For the 5' allele (TRD1.j) a model using a single κ parameter rather than a κ indexed by tissue c was preferred (change in deviance information criterion $\Delta\text{DIC} = -0.523$ (Spiegelhalter et al., 2002)). For the 3' allele (TRD2.j), a model with a single κ parameter did not converge. A model with κ indexed by allele was used instead.

This simultaneously estimated the proportion of each colony pick with alleles A-F for both each individual isolate (π), and summed over all the samples in each niche (μ). I applied this over i samples and c niches (in this case c can be carriage/nasopharynx or CSF). The difference in mean of μ (corresponding to the mean allele frequency over all sample pairs for each allele) shows whether alleles are selected for in carriage or invasive disease, however as the confidence intervals overlapped for alleles, no particular allele was associated with invasive disease or carriage isolates. I also checked the diversity of alleles present in each sample by calculating the Shannon diversity index for each sample using the π vector. The median diversities were not significantly different (carriage 0.94; invasive 1.00).

The finding that *ivr* allele does not associate with invasive disease is at odds with the interpretation of Manso et al. (2014) that the capsule expression changes caused by each allele (through genome-wide methylation profile changes) are central to colonisation and disease. I found that, in clinical cases of meningitis, the allele of the *ivr* locus continues to be phase variable regardless of the niche the bacteria are in. Its purpose is likely to defend against phage (Croucher, Coupland et al., 2014), with little effect on disease course in natural human infection.

4.5 Genetic adaptation over the course of single infections

This section concerns whether the invasive pneumococcal population accumulates mutations as it moves from carriage, through blood to the CSF, and if it does whether this mutation represents adaptation to either of these niches. By sampling the same population longitudinally the issue of population structure is not an issue as for the convenience samples of cases and controls collected for GWAS, which will not be from the same population of bacteria. I called variation between pairs of samples (table 4.8), and looked for convergent evolution between different cases and/or signals of adaptation to a specific niche.

Organism	Number of pairs sequenced		Mean coverage
	blood/CSF	nasopharynx/CSF	
S. pneumoniae	674	6	91.7x
N. meningitidis	195	48	96.6x

Table 4.8: The number of paired samples analysed from the MeninGene study, and the average sequencing coverage.

I made assumptions about the evolution of bacteria within the host, under which I discuss the power of pairwise comparisons between single colonies taken from each niche to capture repeated evolution occurring post-invasion:

1. There is a bottleneck of a single bacterium upon invasion into the first sterile niche (usually blood), which then founds the post-invasion population (Gerlini et al., 2014; Moxon & Murphy, 1978).

2. A large invasive population is quickly established, as the population size approaches the carrying capacity of the blood/CSF. The population size is large enough for selection to operate efficiently.

3. As infection occurs in a mass transport system, populations are well mixed without any substructure. Therefore, the effective population size equals the census population size.

4. The bacterial growth rate within blood and CSF is similar.

Initially the population size is small, so selection is inefficient and the population-wide mutation rate is low. However, the eventual carrying capacity (the maximum number of cells) of the blood and CSF are large enough ($> 1.5 \times 10^5$ colony forming units (CFUs)) (Brown et al., 2004; La Scolea & Dryja, 1984) for beneficial mutations to fix rapidly. Due to the short generation time of around an hour (Allegrucci et al., 2006), this carrying capacity is reached early in the course of the disease (after 1-2 days) (Gang et al., 2015).

Crucially, population sizes where selection acts efficiently (Patwa & Wahl, 2008) are reached even earlier than this – a few hours after invasion. Therefore, mutations with a selective advantage occurring after the first stages of infection will eventually become fixed in the niche's population. So, sequence comparison between colony picks from each niche is likely to find adaptation that has occurred post invasion.

Similarity of the bacterial growth rate within blood and CSF is an important assumption because in 45% of the pneumococcal cases there was evidence that CSF invasion happened before blood invasion (patients had a documented prior CSF leak, otitis media or sinusitis (Brouwer, Heckenberg et al., 2010; Heckenberg et al., 2012)). This allowed me to search for post-adaptation invasion that happens in either direction in this species. I investigated the validity of this assumption using analysis of data on the *ivr* locus (section 4.5.4).

In carriage samples, although the population size is small (Y. Li, Thompson et al., 2013) carriage episodes can persist over many months (chapter 3), therefore allowing the potential for mutations conferring an advantage in an invasive niche to arise. Additionally, during carriage there is known to be population wide diversity (Cremers et al., 2014) and in some cases competition between strains (Cobey & Lipsitch, 2012). I only had access to the sequence of a single strain sampled from this diverse pool, which means I had less power to detect mutations either side of the bottleneck. Combined with the small sample size, this means only adaptive mutations with large selective advantages could be discovered in this part of the study.

Finally, I considered whether the culturing process will bias the results. In *S. pneumoniae* I found that two additional passages of the previous sample pair resulted in one additional insertion. In *N. meningitidis* a low rate of variation and no selection on phase-variable regions and no variation of other regions have been observed during the culture steps (Fransen et al., 2009; van der Ende et al., 1995; van der Ende et al., 2000). I therefore concluded that there will be minimal bias introduced during culturing, and that which is introduced will increase the frequency of mutations between pairs without bias towards either blood or CSF. Due to the higher power to detect variation between the blood and CSF, I present those results first in section 4.5.2, and the carriage/CSF results in section 4.5.5.

4.5.1 Reference free variant calling

As the amount of variation beween blood and CSF isolate pairs is very low, I needed to ensure I had sufficient power to call variants and did not suffer from an elevated false negative rate. I used the same simulation setup as in section 4.2, except generated an average of only 200 mutations between 100 simulated sample pairs.

To avoid reference bias, and missing variants in regions not present in an arbitrarily chosen reference genome, I then performed reference free variant calling between all sequence pairs of isolates using two methods: the 'hybrid' method (Uricaru et al., 2014) and Cortex (Iqbal et al., 2012). The former uses de novo assembly of the CSF sequence reads, mapping of reads from both the blood and CSF samples back to this sequence, then calling variants based on this mapping. Cortex uses an assembly method that keeps track of variation between samples as it traverses the de Bruijn graph.

In the hybrid method I used the SPAdes assembly of the CSF sample as the reference, then mapped reads from both members of the sample pair to this sequence using SNAP (Zaharia et al., 2011) followed by variant calling with bcftools v1.1 (H. Li, 2011) using the command:

```
samtools mpileup -C 50 -m 2 -F 0.0005 -d 1000 -t DP,SP -g -
   p -L 1000 -f assembly.fa mapping.bam | bcftools call -vm
    -P 1e-3 samples.txt
```

I filtered variants with QUAL $<$ 50, MQ $<$ 30, SP $>$ 30, MSQB $<$ 0.001, RPB $<$ 0.001 or DP $<$ 4 out.

For Cortex I first error corrected sample reads using quake (Kelley et al., 2010) to prevent false positive calls supported by very low coverage of reads. I then used the joint workflow of cortex with each set of corrected reads in its own path in the de Bruijn graph, and bubble calling was used to produce a second set of variants between samples. SNPs in the error corrected reads were also called using the graph-diff mode of SGA (Simpson & Durbin, 2012).

I then called variants between these sequences and a draft R6 assembly from simulated read data using both of the above methods; comparison with the mutations known to be introduced allowed power and false positive rate to be calculated – separately for SNPs and INDELs.

In addition to in silico simulation, I cultured blood/CSF paired strains 4038 and 4039 (Croucher, Mitchell et al., 2013) and resequenced them using the same 100bp Illumina paired end sequencing as the rest of the isolates in the study. The genomes of strains 4038 and 4039 have been exhaustively analysed using multiple sequencing technologies (Illumina, 454 and capillary sequencing), so represent high quality positive control data to assess the calling methods. I tested both methods on these data.

The highest power was achieved using hybrid mapping for SNPs and Cortex for INDELs: median power for calling SNPs was 90% using hybrid mapping, and 74% for INDELs using cortex (fig. 4.7a). SGA recovered few true variants. I therefore used this combination of methods, mapping for SNPs and cortex for INDELs, across all samples. When applied to the paired strains 4038/4039 the same mutations as originally reported are recovered, plus a 37bp insertion in *cysB* which was found to be introduced during culturing.

I used simulations to compare against a simple method of mapping against an arbitrary reference, in this case TIGR4 (Tettelin et al., 2001). I found my reference free method has

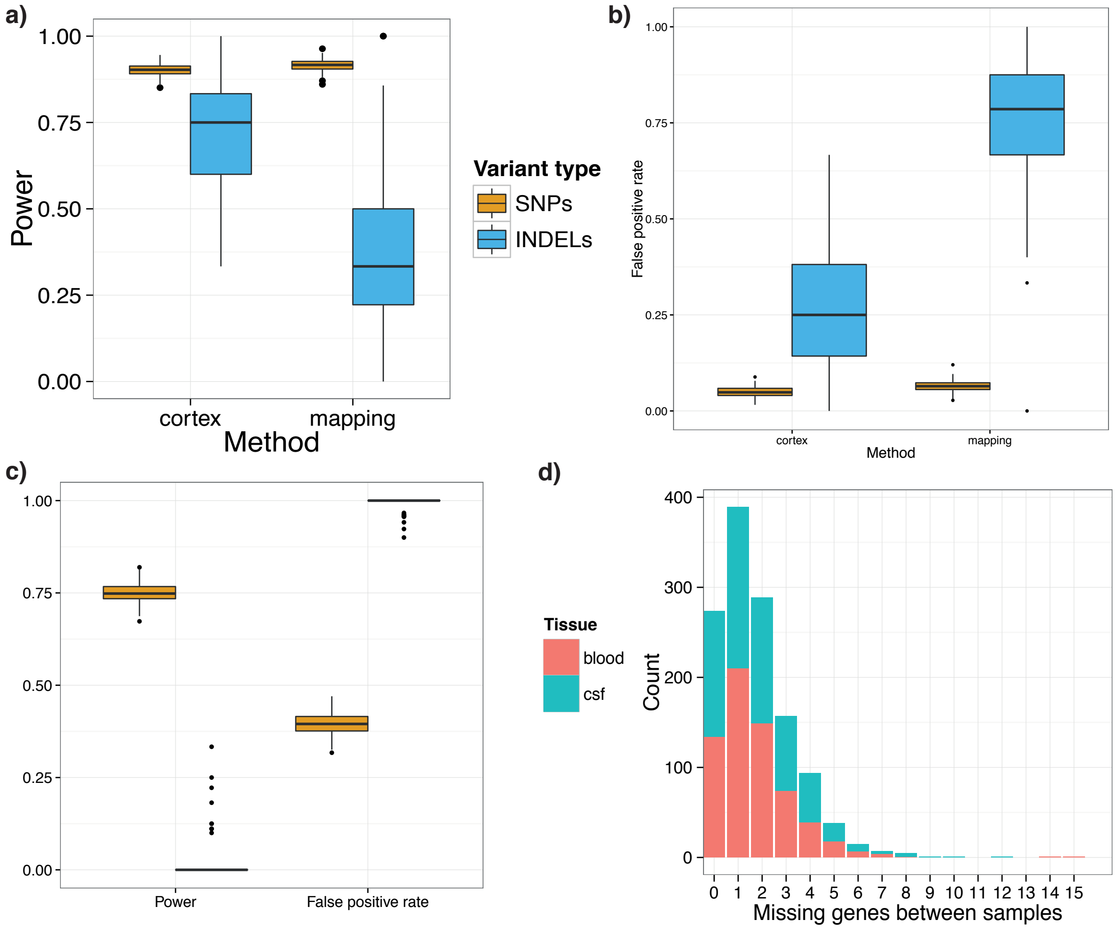

Figure 4.7: Performance of variant calling methods. SNPs (gold) and INDELs (blue) are shown separately. **a)** Boxplot of power (recall) for each method of variant calling for 100 simulated samples. **b)** shows the false discovery rate. **c)** Boxplot of power and false positive rate for reference based calling. Run on the same 100 simulated samples as a), calculated by number of false positives/number of true positives. **d)** Count of annotated genes present in blood but not CSF (red) or vice-versa (turquoise) between the 673 *S. pneumoniae* samples. The level of variation is inflated due to frequent misannotation of coding sequences (CDS)s.

greater power, especially for INDELs (fig. 4.7c), and a markedly reduced false positive rate. I also tested an assembly method alone to compare gene presence and absence, but this too suffered from a vastly elevated false positive rate (fig. 4.7d).

Variant direction and effect annotation

To be able to compare between samples using a consistent annotation, I mapped the called variants to the ATCC 700669 reference (Croucher et al., 2009) for *S. pneumoniae*, and MC58 reference (Tettelin et al., 2000) for *N. meningitidis*. This was done by taking a 300 base window around each variant and using blastn on these with the reference sequence. ‘Directionality’ was then relative to the reference used, and a binomial test with $\lambda = 0.5$ was used to test significance. I used VEP (McLaren et al., 2010) to annotate consequences of each SNP as synonymous, non-synonymous, or stop-gained and INDELs as frameshift or inframe.

4.5.2 No repeated post-invasion adaptation in coding regions across species

For each species I then counted the number of variants of any type between each blood/CSF isolate pair taken from a patient (fig. 4.8). In *S. pneumoniae* 452 of 674 paired samples (67%) were identical. The distribution of number of variants between isolate pairs is roughly Poisson (mean = 0.547), excluding outliers. Variation between *N. meningitidis* pairs also followed a roughly Poisson distribution (mean = 2.34), which when compared to *S. pneumoniae* showed a higher number of variants between blood and CSF isolates (Wilcoxon rank-sum test, W = 25 790, $p < 10^{-10}$) such that most pairs have at least one variant between the blood and CSF samples.

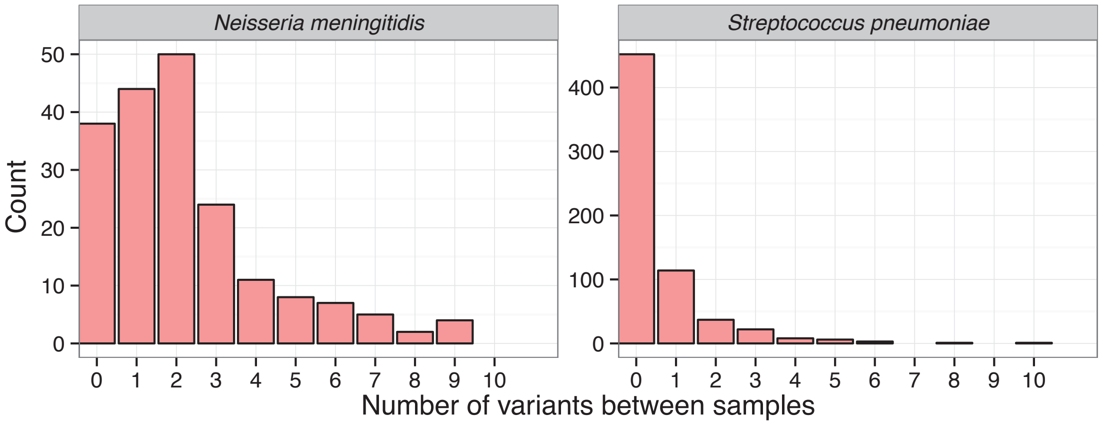

Figure 4.8: Histograms binned by number of variants between a blood/CSF sample pair, for both pathogens. Total pairs analysed in table 4.8. SNPs are from mapping, INDELs are from cortex. Three *S. pneumoniae* and one *N. meningitidis* sample with over 10 variants are not shown.

To test whether certain genotypic backgrounds were associated with a higher number of mutations that occurs post-invasion, I performed a linear fit of each MLST against number of mutations between blood and CSF isolates. I Bonferroni corrected the p-values of the slope for each MLST; at a significance level of 0.05 no MLST was associated with an increased number of mutations.

In both species, the mutations that do exist, if they cause the same functional change, could represent a signal of adaptation. To determine whether this is the case, the number of mutations in each CDS annotation was counted. I then performed a single-tailed Poisson test using the genome wide mutation rate per base pair multiplied by the gene length as the expected number of mutations. The resulting p-values were corrected for multiple testing using a Bonferroni correction with the total number of genes tested as the *m* tests; I have reported results with $p < 0.05$ in table 4.9.

Gene name	Gene length (bp)	Blood/CSF mutations	p-value
pde1 (SPD_2032)	1973	19	$< 10^{-10}$
dltD (SPD_2002)	1269	13	$< 10^{-10}$
dltB (SPD_2004)	1245	12	$< 10^{-10}$
dltA (SPD_2005)	1551	11	$< 10^{-10}$
clpX (SPD_1399)	1233	7	1.3×10^{-8}
wcaJ (SPD_1620)	693	6	3.4×10^{-8}
cysB (SPD_0513)	909	5	1.6×10^{-5}
cbpJ	1122	5	4.7×10^{-5}
amiC (SPD_1670)	1332	4	6.0×10^{-3}
marR	435	3	9.6×10^{-3}
fhuC	519	3	1.6×10^{-2}

Table 4.9: Genes containing significantly repeated mutations between blood and CSF isolate pairs in *S. pneumoniae*. Ordered by increasing p-value; locus tags refer to the D39 genome, if present.

The *dlt* operon, responsible for D-alanylation in teichoic acids in the cell wall (Deininger et al., 2007; Habets et al., 2012; Kovács et al., 2006), was the most frequently mutated locus: 36 mutations in 31 sample pairs (Poisson test $p < 10^{-10}$). This occurred in only 5% of samples, so adaptation to a niche due to variation in genes is not common. To investigate whether this represented adaptation to either blood or CSF, I annotated the effect of these variants, and determined whether they were specific to a niche. I mapped them to the R6 *S. pneumoniae* strain, which has a functional *dlt* operon and was therefore assumed to be the ancestral state. There was no directionality to the mutations: 19 occurred in the blood, and 11 in the CSF ($p = 0.2$). Only seven of the patients infected by these strains showed signs of blood invasion before CSF invasion (sinusitis or otitis); this also did not show directionality. I have plotted the position and nature of the mutations in fig. 4.9. Most of these mutations would be expected to cause LoF in the operon. Though this suggests this locus has a deleterious effect in invasive disease generally, the lack of directionality to the mutations means it does not show evidence of adaptation to either the blood or CSF specifically.

The next most significantly mutated gene was *pde1*. The *pde1* gene was first found to be essential for growth in an experimental meningitis model (Molzen, Burghout, Bootsma, Brandt, Der Gaast-De Jongh et al., 2011); further study by Cron et al. (2011) showed that *S. pneumoniae* mutants with *pde1* (SP2205 in TIGR4; SPD2032 in D39) and its paralogue *pde2* (SP1298 in TIGR4; SPD1153 in D39) knocked out exhibited reduced host cell adherence and attenuated virulence in a mouse model of meningitis. Following work confirmed that Pde1 acts as a phosphodiesterase, cleaving c-di-AMP into pApA (Bai et al., 2013; Kuipers et al., 2016). These signalling molecules are known to have broad effects

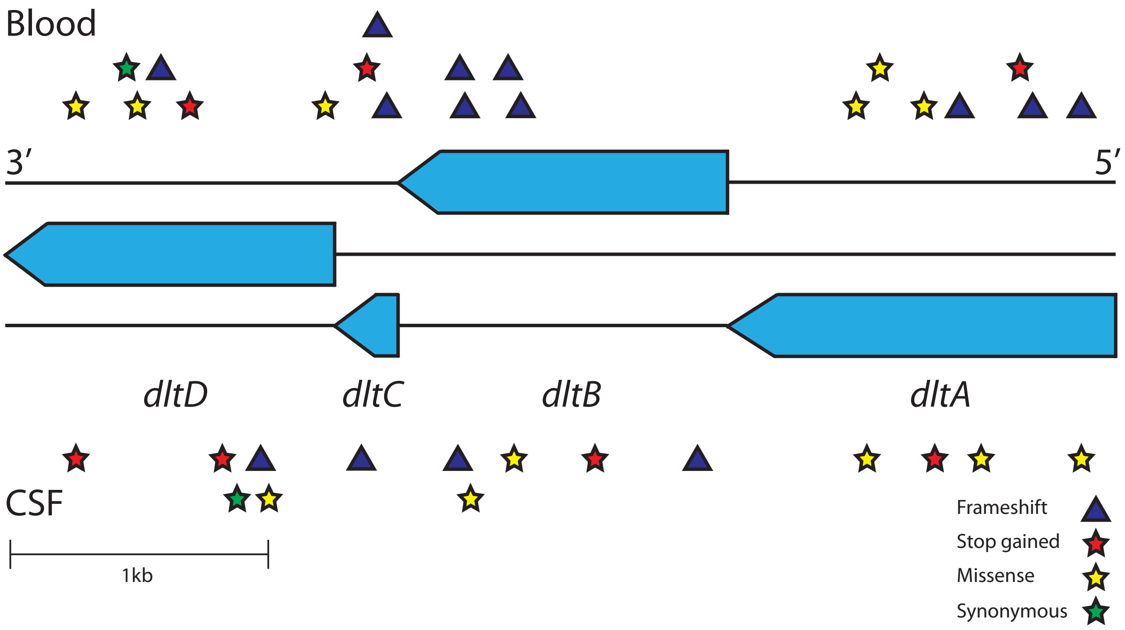

Figure 4.9: Mutations observed between all paired samples in the dlt operon. The operon consists of four genes in the three reading frames of the reverse strand. Mutations, displayed by type, in the blood strains are shown above the operon, and in the CSF strains below the operon.

on the cell (Tamayo et al., 2007) and were again shown to affect growth and virulence in a mouse model of pneumonia. In both studies, the authors suggested that these proteins are promising vaccine targets.

I therefore checked whether *pde1* appeared to be under selection in the sampled population. The ratio of nonsynonymous to synonymous mutations was neutral (dN/dS = 0.89) and contained variants with a SFS similar to that of other genes (fig. 4.10a and b; Tajima's D $= -1.44$; p $= 0.67$). However, as all the within-host mutations were nonsynonymous, this implied that selection may act on *pde1* during the course of invasive disease. I then computationally predicted the effect of the 19 mutations observed to occur in *pde1* using `SnpEff` and `PROVEAN` (Cingolani et al., 2012; Choi et al., 2012), and have plotted these along with the predicted functional domains in fig. 4.10c. Of these mutations, 14 are predicted to change protein function, without causing LoF. The mutations are not evenly distributed across the gene and are mostly clustered in the DHH family domain or just before it. While this does not allow a singular interpretation of the effect of these variants on gene function, this is consistent with selection acting on *pde1* during meningitis. This supports the conclusion of Cron et al. (2011) that *pde1* is essential for virulence, and lends credence to the idea it may be an effect component of a pneumococcal protein vaccine.

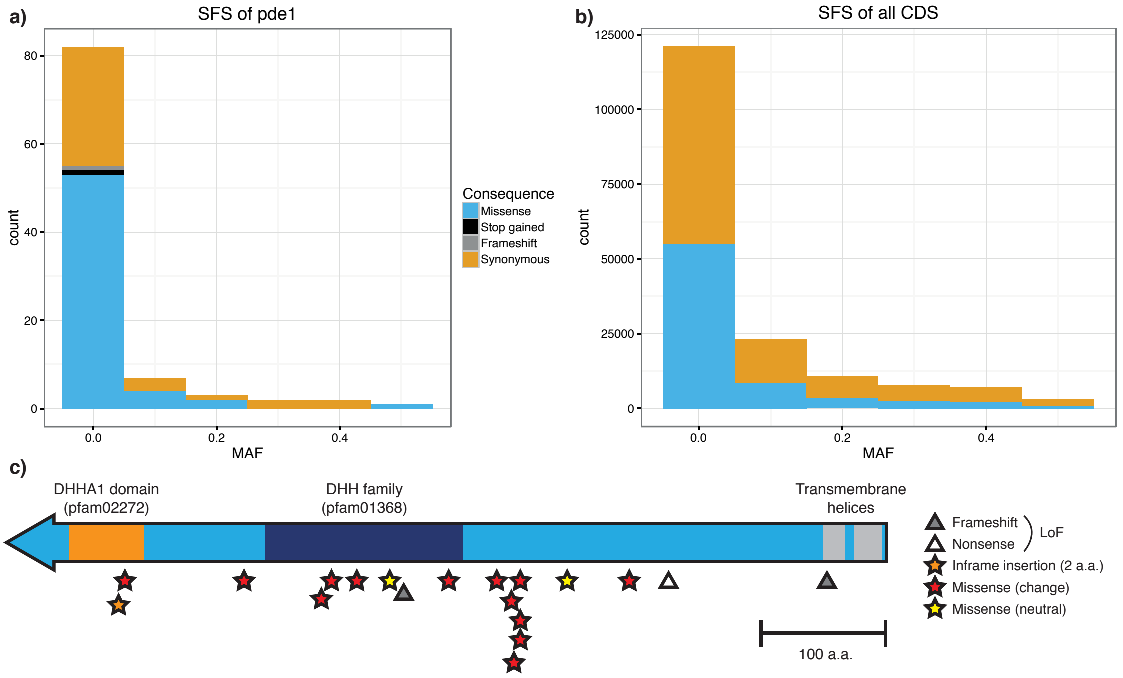

Figure 4.10: Evidence of selection on *pde1* during meningitis. Panels a and b show the SFS of mutations in just *pde1* and in all CDS, respectively. Variants are coloured according to the predicted effect. Panel c shows the positions and predicted effects of mutations observed in *pde1* during cases of meningitis and predicted pfam domains.

In all the other genes in table 4.9 the variants are non-synonymous SNPs distributed evenly between blood and CSF, therefore also showing no adaptation specific to either niche.

The most frequently mutated genes between pairs in *N. meningitidis* are shown in table 4.10. Top ranked are those relating to the pilus: *pilE* (19), *pilC* (6) and *pilQ* (4). Pilus genes are associated with immune interaction (Wörmann et al., 2014), and are therefore expected to be under diversifying selection; an excess of non-synonymous mutations (dN/dS = 1.39; p = 0.024) was consistent with this. The other notable gene with more mutations than expected in *N. meningitidis* was *porA*, encoding a variable protein which is a major determinant of immune reaction (Russell et al., 2004), in which 12 samples had frameshift mutations in one of two positions. Phase variation in the gene's promoter region, affecting its expression, is discussed in more detail below.

Gene name	Gene length (bp)	Blood/CSF mutations	p-value
pilE (NMB0018)	384	18	$< 10^{-10}$
lgtC	189	16	$< 10^{-10}$
hyaD	327	14	$< 10^{-10}$
oatA	1869	19	$< 10^{-10}$
hpuB (NMB1668)	2382	17	$< 10^{-10}$
porA (NMB1429)	1178	12	$< 10^{-10}$
lgtA (NMB1929)	1050	10	$< 10^{-10}$
kfoC	360	7	$< 10^{-10}$
cotSA	1134	7	9.2×10^{-9}
ssa1	3252	6	3.9×10^{-4}

Table 4.10: Genes containing significantly repeated mutations between blood and CSF isolate pairs in *N. meningitidis*. Ordered by increasing p-value; locus tags refer to the MC58 genome, if present.

The mutations in table 4.10 showed no association with blood or CSF specifically, so do not represent adaptation to either niche. Genetic variation in *pilE*, *hpuA*, *wbpC*, *porA* and *lgtB* within host has been observed previously in a single patient with a hypermutating *N. meningitidis* infection (Omer et al., 2011). These coding sequences overlap with those in table 4.10, which also suggests an elevated background mutation rate in these sequences, rather than strong selection between the blood and CSF niches.

Finally, I tested whether the increased mutation rate in the genes in tables 4.9 and 4.10 was associated with a particular genotype. I performed a logistic regression for each gene with over ten mutations reaching significance in the Poisson test, coding samples as one and zero based on whether they had a mutation in the gene being tested or not: no genes being mutated post invasion were associated with an MLST.

Copy number variation

I called CNVs between samples by first mapping each species to a single reference genome (ATCC 700669), then fitting the coverage of mapped reads with a mixture of Poisson distributions (Klambauer et al., 2012) as in section 4.3. Using windows of 1kb, I ranked regions by the number of sample pairs containing a discordant CNV call, as defined by the integer copy number being different between blood and CSF samples. I then inspected the top 5% of these regions.

In *S. pneumoniae* the most frequently varying region was due to poor quality mapping of a prophage region. The only other region with $p < 0.05$ was a change in copy number of 23S rRNA seen in a small number of sample pairs. In *N. meningitidis* mismapping in the *pilE*/*pilS* region accounts for the only CNV change.

4.5.3 No evidence for repeated adaptation in intergenic regions in *S. pneumoniae* and *N. meningitidis*

The previous result suggesting adaptation from blood to CSF was an intergenic change affecting the transcription of the *patAB* genes, encoding an efflux pump (Croucher, Mitchell et al., 2013). In general it is known that in pathogenic bacteria a common form of adaptation is mutation in intergenic regions, which may affect global transcription levels, causing a virulent phenotype (Gripenland et al., 2010; Johansson et al., 2002), antimicrobial resistance (Sreevatsan et al., 1997) and changing interaction with the host immune system (Magnusson et al., 2007). Changes in these regions have previously been shown to display signs of adaptation during single cases of bacterial disease (Marvig et al., 2014).

I therefore separately investigated the mutations in non-coding regions. Analysing the positions of these mutations required a consistent co-ordinate system across all sample pairs. To achieve this, I remapped the co-ordinates of each variant discovered in an intergenic region to the co-ordinates of the ATCC 700669 reference genome. I used the population matched carriage isolates as the ancestral state to determine whether these mutations occur in the blood or CSF isolate.

Figure 4.11 shows all mutations plotted genome-wide in *S. pneumoniae*. The peaks correspond to mutations in genes described in table 4.9. In the remaining 121 mutations in non-coding regions I observed no clustering by position. Over all pairs of samples, intergenic mutations were spread between blood and CSF isolates when compared to a carriage reference. This suggests none of the intergenic mutations are providing a selective advantage in either invasive niche.

The mutations in *N. meningitidis* are plotted in fig. 4.12, 110 of which were in non-coding regions. I observed enrichment (>1 mutation), but no niche specificity, in the upstream region of six genes. These mutations are listed in table 4.11. Some of the mutations upstream of *porA* and *opc* are in phase variable homopolymeric tracts, which are discussed more fully in section 4.5.4. The other mutations are upstream of the adhesins *hsf*/NMB0992 and NMB1994, which are involved in colonisation (Hung & Christodoulides, 2013) and immune interaction during invasion (Griffiths et al., 2011), and *frpB*/NMB1988 which is a surface antigen involved in iron uptake (Delany et al., 2006). Differential expression of these genes may be an important factor affecting invasion, but the mutations I observed that may affect this do not appear to be specific to blood or CSF.

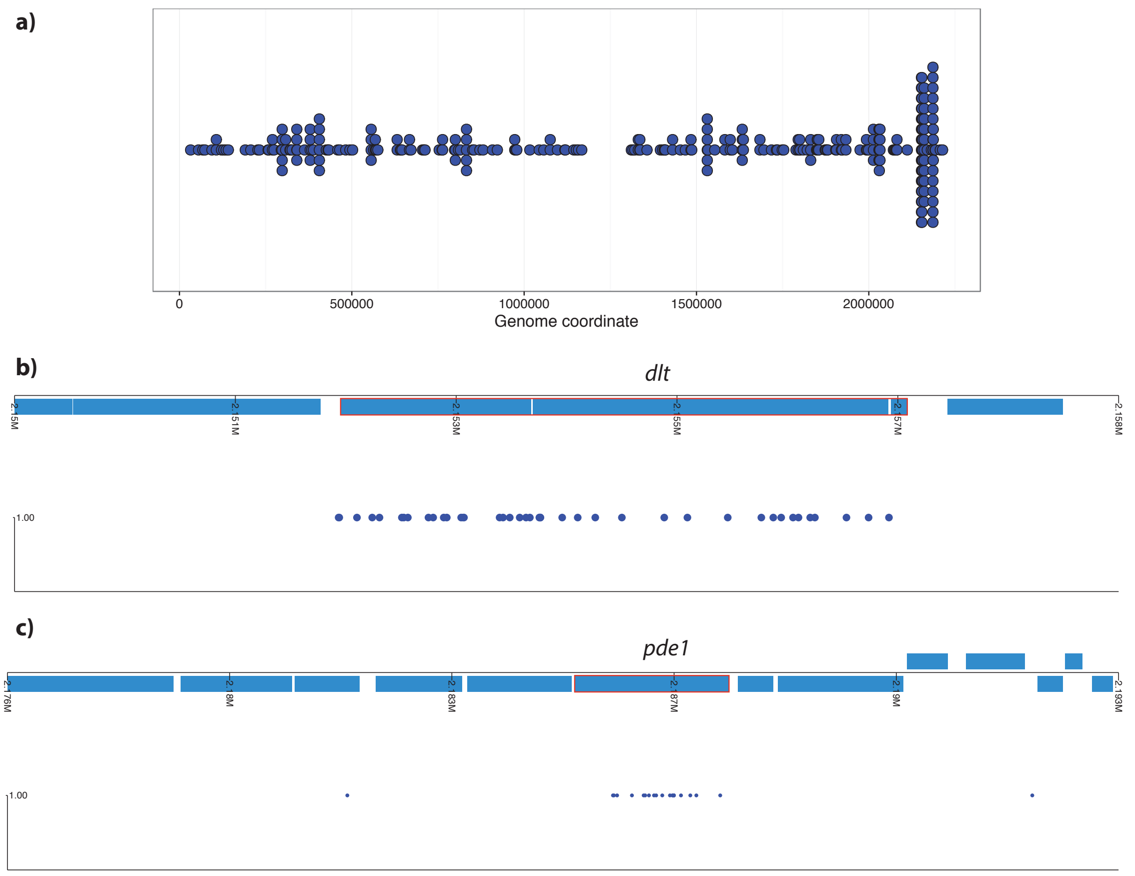

Figure 4.11: Mutations observed between all *S. pneumoniae* pairs, overlaid onto the Spn23F reference. Each blue point on the lower row corresponds to a SNP or INDEL variant observed between at least one sample pair. The blocks in the upper row represent CDSs, lying above or below the central line depending on whether they are on the forward or reverse strand respectively. The panels show **a)** whole genome (stacked, grouped by 1 000 bp windows); **b)** *dlt* operon (four genes in the centre, from 2 152 238 to 2 156 543 base pairs); **c)** *pde1* (gene in the centre from 2 185 398 to 2 187 371 base pairs).

Coordinates	Downstream gene	Blood/CSF mutations
1468329–1468331	*porA* (NMB1429)	7
1072215–1072328	*opc* (NMB1429)	7
1008872–1008985	*hsf* (NMB0992)	6
1315621–1315672	NMB1299	6
2092257–2092552	*frpB* (NMB1988)	5
2100124–2100258	NMB1994	4

Table 4.11: Intergenic regions containing significantly repeated mutations between CSF and blood isolate pairs in *N. meningitidis*. Ordered by increasing number of mutations; coordinates refer to the MC58 genome.

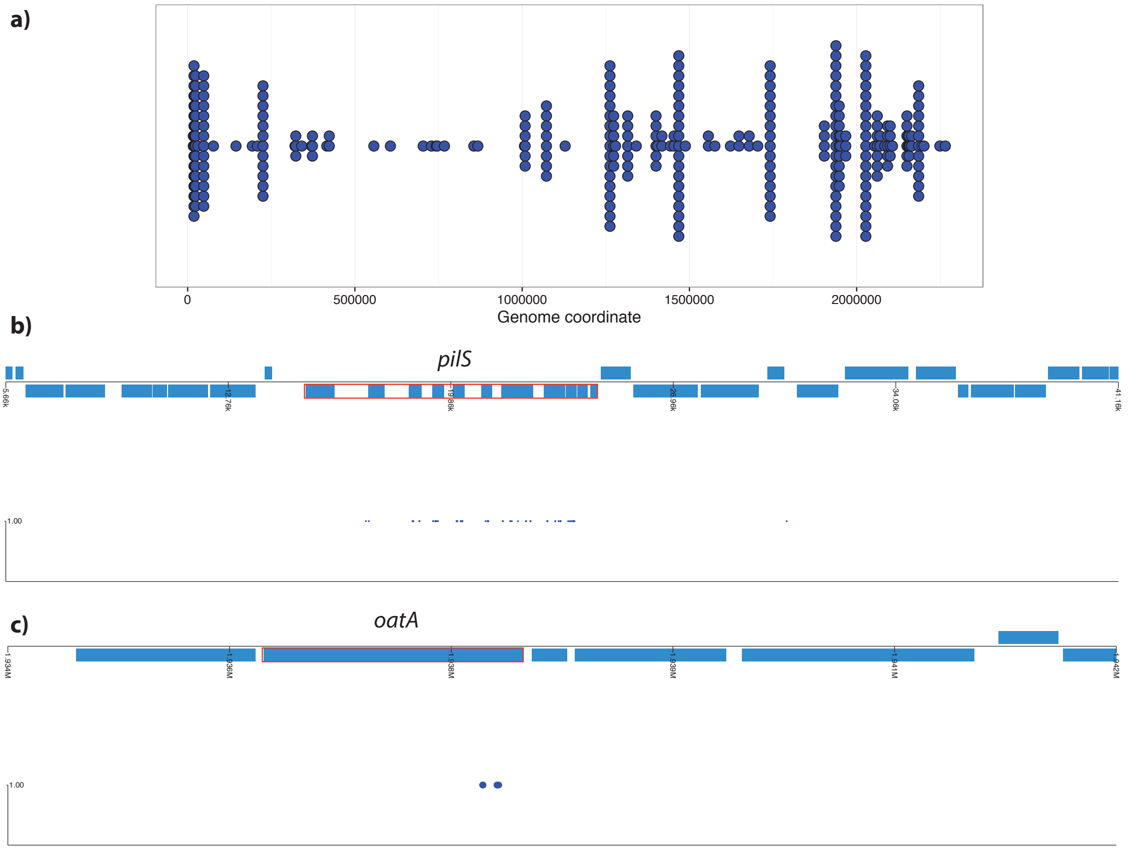

Figure 4.12: As fig. 4.11. **a)** whole genome. **b)** pilus encoding genes. Mapping to the MC58 reference places these incorrectly in the unexpressed *pilS* cassette; compared to the reference the isolates have recombined between *pilS* and the expressed *pilE*. **c)** *oatA*.

4.5.4 No evidence for repeated adaptation in phase variable regions in *S. pneumoniae* and *N. meningitidis*

Phase variable regions, which may also be intergenic, can mutate rapidly and are known to be a significant source of variation in pathogenic bacteria (Bucci et al., 1999). This mutation is an important mechanism of adaptation (Moxon et al., 1994), and meningococcal genomes in particular contain many of these elements (Snyder et al., 2001).

In *N. meningitidis* I observed six samples with single base changes in length of the phase-variable homopolymeric tract in the *porA* gene's promoter sequence, and five samples with the single base length changes in the analogous promoter sequence of *opc*. While changes in the length of these tracts will affect expression of the corresponding genes, both of which are major determinants of immune response (Sarkari et al., 1994; van der Ende et al., 2000), the tract length does not correlate with blood or CSF specifically. Consistent with this, *porA* expression has previously been found to be independent of whether isolates were taken from CSF, blood or throat (van der Ende et al., 2000).

In *S. pneumoniae* I was interested in whether the allele of the phase variable *ivr* locus discussed in section 4.3.2 was associated with either the blood or CSF niche specifically, as this could be a sign of adaptation. As the locus inversion is rapid and occurs within host, we first ensured that cultured samples are representative of the original clinical samples using PCR quantification of each allele. We therefore extracted DNA from a subset of 53 of 674 paired clinical CSF samples and the respective bacterial isolates.

Allele prevalence was quantified using a combined nested PCR protocol based on PCR amplification of the *ivr* locus (Manso et al., 2014). Allele prevalence was identical between the original clinical sample and cultured bacteria in 50 out of the 53 samples. The predictive power of the *in vitro* detected *ivr* allele prevalence in a pneumococcal culture for the original allele prevalence within the clinical sample was therefore sufficient to draw conclusions about adaptation from.

I then used the mapping method described in section 4.3.2 to determine the allele for all the paired samples from the read data. 621 sample pairs had reads mapping to *hsdS* from which an allele can be called. However, as even a single colony contains heterogeneity at this locus, simply taking the allele with the most reads mapping to it in each sample gave a poor estimate of the overall presence of each allele in the blood and CSF niches. To take into account the mix of alleles present in each sample, and to calculate confidence intervals, I used the same hierarchical Bayesian model for the *ivr* allele used for GWAS in section 4.4.3. This simultaneously estimated the proportion of each colony pick with alleles A-F for both each individual isolate (π), and summed over all the samples in each niche (μ). I applied this over i samples and c niches (in this case c can be blood or CSF).

For each pair of blood and CSF samples the difference in allele prevalence $\pi_{\mathrm{CSF}} - \pi_{\mathrm{blood}}$ was calculated. All *S. pneumoniae* samples had a difference in mean of at least one allele (as the highest posterior density (HPD) overlaps zero), highlighting the speed at which this locus inverts. While this means that between a single CSF and blood pair the allele at this locus usually changes, it is the mean of μ_c (corresponding to the mean allele frequency in each niche over all sample pairs) which tells us whether selection of an allele occurs in either the blood or CSF more generally. This is plotted in fig. 4.13. As the HPD overlap, no particular allele is associated with either blood or CSF *S. pneumoniae* isolates.

Manso et al. (2014) showed in a murine invasion model that an increase in proportion of alleles A and B occurs over the course of infection. I did not observe the same effect in these clinical samples, though the large confidence intervals from the mathematical model suggest that genomic data with a small insert size relative to the size of repeats in the locus is limited in resolving changes in this allele. A small selective effect of *ivr* allele between these niches would therefore not be detected, but strong selection for a particular allele (odds ratio > 2) can be ruled out.

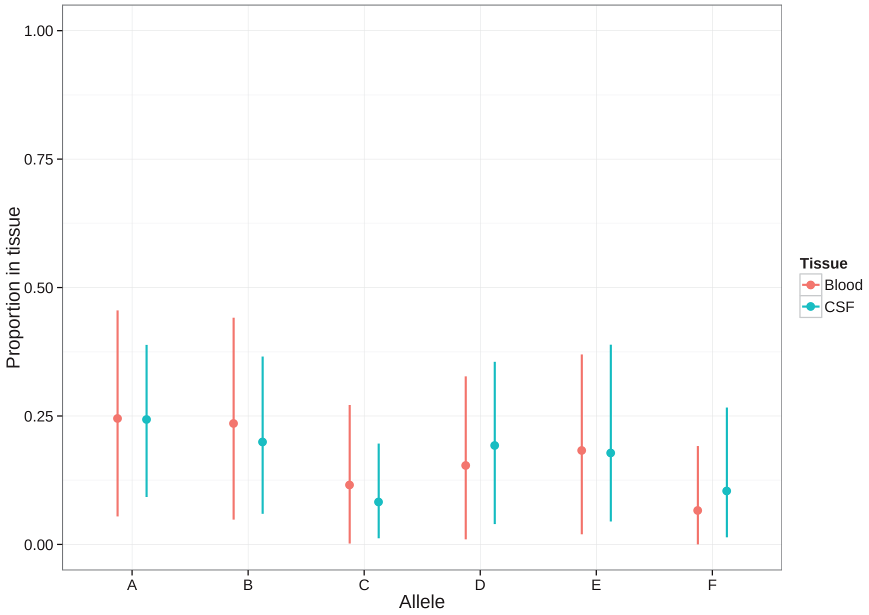

Figure 4.13: Mean and 95% HPD for μ_c. This shows the proportion of each allele present in each of blood (red) and CSF (turquoise) tissues pooling across all samples.

Diversity of *ivr* allele within samples

As the speed of inversion is rapid, I used the subsequent polymorphism of this locus to evaluate the assumptions about diversity of the bacterial population within each niche. I calculated the Shannon index of each sample's vectors μ_{CSF} and μ_{blood} to measure diversity of the sample in each niche. The mean Shannon index across CSF samples was 1.01 (95% HPD 0.39-1.51) and 0.98 (95% HPD 0.35-1.55) in the blood (fig. A.14). Looking at each sample pair individually, the difference between diversity in each niche appeared normally distributed with a mean of zero. Together, these observations suggested a similar rate of diversity generation in each niche. This is in line with the assumption that the two populations have similar mutation rates, and a similar number of generations between being founded and being sampled.

4.5.5 Carriage and invasive disease sample pairs show some evidence of repeated adaptation

Using the same methods, I also analysed pairs of genomes from 54 patients that were collected from the nasopharynx and CSF. Six of these were *S. pneumoniae*. In these samples, I detected only one sample with any variation (fig. 4.14), which was a two base insertion upstream of the *gph* gene. This is similar to the amount of mutation observed between the blood and CSF isolates, which is expected given the similar sampling timeframes. While I found that a functional *dlt* operon appears to have a deleterious effect in invasive disease, I did not observe mutation between the carriage and disease samples. However, this was expected given the small number of carriage samples relative to the effect size detected for this operon.

Between the remaining 48 *N. meningitidis* carriage and CSF isolate pairs small numbers of mutations were common. I went on to search for regions enriched for mutation, however in 8 samples I observed large numbers of mutations clustered close together (fig. 4.14). These represented single recombination events, so when analysing genes enriched for mutation I counted each recombination as a single event (Croucher, Page et al., 2015; Maiden et al., 1998).

Table 4.12 shows the results of this analysis. Similar genes are mutated as in the blood/CSF pairs, again with no specificity to either niche. In phase variable intergenic regions, I observed four sample pairs with an insertion or deletion in the *porA* promoter tract with no niche specificity. Otherwise, none of the regions above showed enrichment for mutation in either niche. These observations support the theory that these genes mutate at a higher rate but do not confer a selective advantage in any of the three niches studied.

Gene name	Gene length (bp)	Carriage/CSF mutations	p-value
lgtA (NMB1929)	1050	6	5.0×10^{-7}
oatA	1869	6	1.5×10^{-5}
hyaD	327	4	2.6×10^{-5}
pilE (NMB0018)	384	4	3.8×10^{-3}
pilT (NMB0052)	1131	4	3.5×10^{-3}
dca (NMB0415)	444	3	1.1×10^{-2}

Table 4.12: Genes containing significantly repeated mutations between nasopharyngeal and CSF isolate pairs in *N. meningitidis*. Ordered by increasing p-value; locus tags refer to the MC58 genome, if present.

A notable exception to this is the *dca* gene, a phase variable gene involved in competence in Neisseria gonorrhoea but of unknown function in N. meningitidis (Snyder et al., 2001; Snyder et al., 2003), in which all mutations are protein truncating variants in the invasive isolate. Similarly, though not reaching significance (due to the long length of the genes) were the *ggt* (NMB1057) and *czcD* (NMB1732) genes in which three recombina-

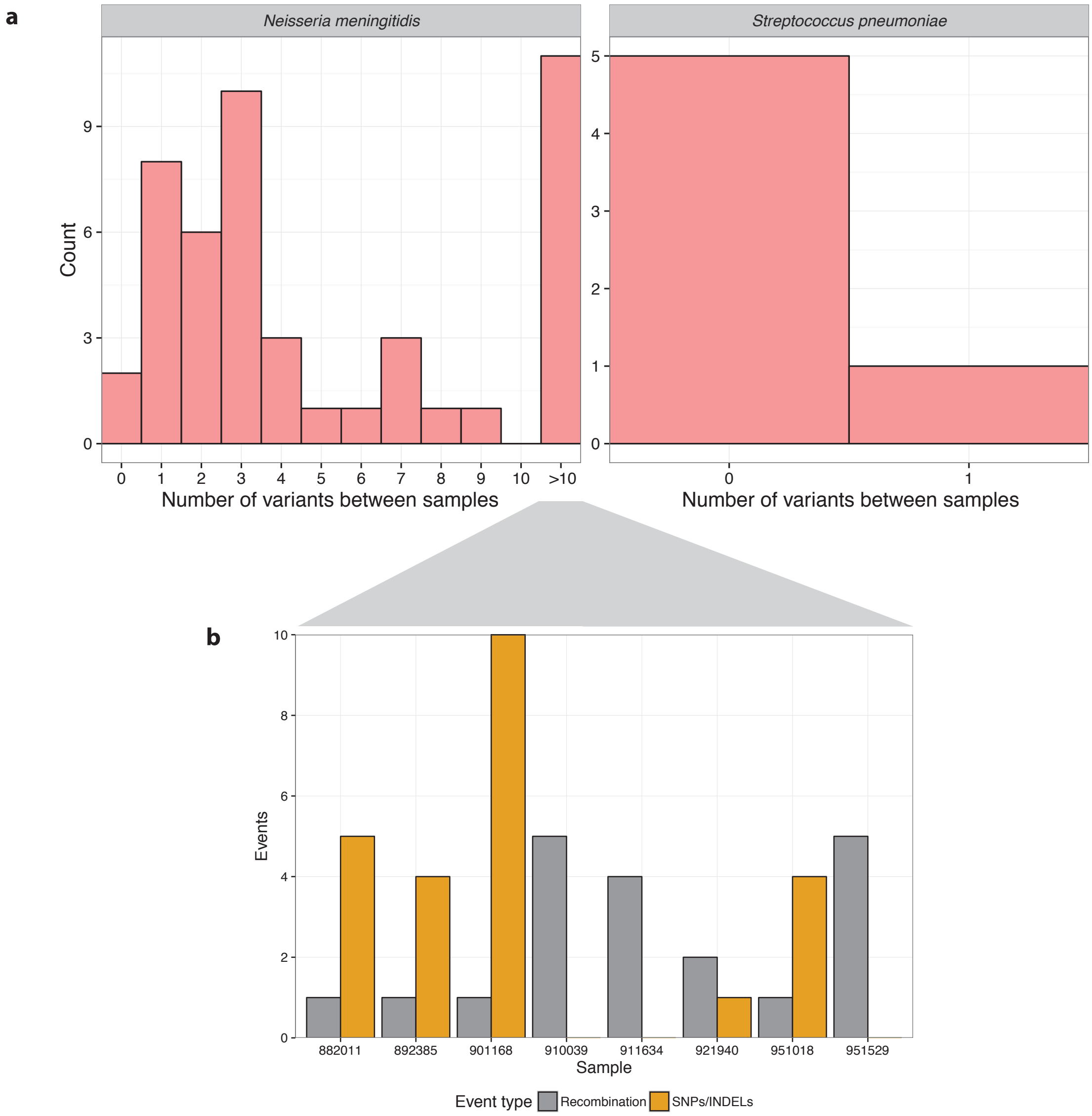

Figure 4.14: Histograms binned by number of variants between a carriage/CSF sample pair, for each bacterial species. **a)** As fig. 4.8. In *N. meningitidis* eleven samples with over ten variants between them due to recombination events are grouped. **b)** The number of recombination and SNP/INDEL events in samples in the group with over ten detected variants.

tions occurred, all of which were in the invasive isolate of the pair.

The mutations in these three genes therefore may confer a selective advantage in the invasive niche; the sequence at these loci in the invasive strains are the same as the MC58 reference, an invasive isolate itself. *ggt* has previously shown to be essential for *N. meningitidis* growth in CSF in rats (Takahashi et al., 2004), and metal exporters such as *czcD* have been shown to increase virulence in a mouse sepsis model (Veyrier et al., 2011). More such paired carriage and invasion samples would be needed to confirm if this is the case in human invasive disease.

4.6 Conclusions

In this chapter I have used a population of *S. pneumoniae* genomes to determine the contribution of naturally occurring bacterial variation to the progression of meningitis from asymptomatic carriage through blood invasion to CSF invasion. I first used a variety of bioinformatic methods to catalogue as wide a variety of variants as possible, particularly those which have previously been associated with virulence.

Using these variants and a matched collection of carriage and invasive isolates I found that the bacterial genome is crucial in determining invasive potential, with serotype likely to be the main factor. However, I did not find any evidence that the bacterial genome contributes to severity or outcome of disease. Using GWAS of both common and rare variants I found many regions and genes to be associated with invasive disease, independent of genetic background. Some of these have been previously described, whereas this is the first time others have been associated with invasive human disease. Genes involved in capsule synthesis, *yhfE*, RumA2, bacteriocins, *nanA* and *nanB* were associated with invasiveness using both common and rare variants, as well as analysis of selection.

The rare variant burden test found some well known virulence factors, showing that large effect size LoF mutations generated in lab mutants exist in the natural population, and further can affect disease in human infection. Common variants with smaller effect sizes may be the most interesting result of this approach in future, as the smaller effect sizes are harder to discover with bottom-up approaches, and their higher frequency in the population may make them more appealing vaccine targets.

I did not find evidence for association with invasiveness for some previously described variants. I did not find that the *ivr* allele was associated with invasive disease, suggesting that its function is to defend against highly variable prophage and that the variable capsule expression it can produce are not selected for in natural disease. The three antigen alleles were not associated with invasiveness, suggesting the allelic variants are a general form of diversifying selection without specific forms having a differing fitness in carriage or invasion.

These hits, as they rely on a single study population, are susceptible to batch effects specific to the Dutch setting or due to sampling bias of the collection. The association of positive controls such as capsule is reassuring, but replication in an independent population is necessary before further interpretation. The hits I have reported here will be useful for meta-analysis when further sampling and GWAS is performed.

As well as large scale population differences, previous studies have shown that substantial levels of genomic DNA sequence variation occur in bacteria colonising or infecting human hosts (Eyre et al., 2013; Kennemann et al., 2011; Morelli et al., 2010) and suggest that some of this variation may be due to selective adaptation (Croucher, Mitchell et al., 2013; Jorth et al., 2015; Marvig et al., 2014; L. Yang et al., 2011; Young et al., 2012). Such adaptations during invasive bacterial disease could lead to new insights into the processes of pathogenesis with the potential to inform therapies (Sudip Das et al., 2016; Didelot et al., 2016), which would be difficult to assess with GWAS due to the rapid disease progression. I have searched for variation in *S. pneumoniae* and *N. meningitidis*, by comparing the pan-genomes from bacteria isolated from both blood and CSF from the same individuals in 869 bacterial meningitis cases. The genetic background within-host is the same, so this comparison could be performed without population structure correction.

I found overall that blood and CSF isolates have very similar genetic sequences. The mutations observed are not randomly distributed throughout the genome, but are instead randomly distributed between blood and CSF isolates. These mutations are therefore an observation of a higher mutation rate in these regions during invasion (for example the pilus in *N. meningitidis*, which is known to be under diversifying selection) but not repeated adaptation to either niche. This study indicates that the previous observation of variation between blood and CSF isolates from a single case of meningitis (Croucher, Mitchell et al., 2013) was a rare event most likely driven by antibiotic selection pressure during treatment. The large sample size means that this eliminates the need to search for bacterial diversity between invaded host niches (blood and CSF) when trying to explain pathogenesis of meningitis, which is a tempting analysis for reference labs with both sets of samples available. However, my comparison between the genomes of carriage and invasive isolates did show some weak signals of adaptation. I found that *dlt* appeared to be deleterious in invasion, and that selection appeared to be acting on *pde1* during invasion. These genes were not associated with invasiveness in the GWAS, which may be due to insufficient power or population stratification.

I went on to analyse 54 samples comparing carriage and invasive isolates from the same patient. Though the sample size was lower, and fully sampled diversity within the nasopharynx was not available, I was able to get an insight into potential genetic differences between bacteria in these niches. I saw some of the same genes that mutate rapidly between blood and CSF isolates also do this between carriage and invasion. This supports the conclusion that these genes have a higher mutation rate, rather than giving a selective advantage to a niche. However the power in these comparisons was limited by sample size and single colony sequencing, so comparison with GWAS results is not possible.

In the next chapter I will perform a similar analysis on the effect of host genetics on bacterial meningitis, starting with the proportion of variability attributable to common host genetic variation for invasiveness and disease severity. Together, this will give an overall picture of host and pathogen genetics affecting pneumococcal meningitis.