Chapter 6

Conclusions

6.1 Summary of findings

S. pneumoniae is a human commensal, which in rare cases can invade a usually sterile niche. If the blood or CSF is invaded this usually leads to serious disease, called bacteremia and meningitis respectively. While virulence factors of the pathogen necessary for invasive disease have been identified from bottom-up lab based approaches (often relying on a mouse model of infection), the role of naturally occurring sequence variation in the pathogen genome in invasive disease is generally unknown.

I have used a large cohort of *S. pneumoniae* genomes isolated from invasive disease and asymptomatic carriage to determine the importance of sequence variation in disease susceptibility and severity, and to find the specific regions of the genome which contribute to these variations in phenotype. My main approach was to use GWAS, which is a hypothesis-free way of testing all genomic variants for association with a given phenotype. This approach does not require prior assumptions about which genes may affect the phenotype and does not rely on large-effect size gene knock-outs or animal models of disease.

In the context of bacterial populations, GWAS faces difficulties caused by strong population structure and highly plastic genomes. I developed a piece of software to help overcome these issues by finding an appropriate adjustment for population stratification, and using sequence elements (k-mers) to test for variation of the pan-genome. After testing this method using antibiotic resistance as a positive control, I then applied it to the phenotype of pneumococcal carriage duration, where I also developed a model to estimate carriage duration from longitudinal swab data. By adapting methods derived from human genetics, I was able to calculate the heritability caused by the pathogen genome, and identify which variants explained variation in this important epidemiological parameter.

Using a range of bioinformatic approaches I catalogued variation of the population of pneumococcal genomes sampled from the Netherlands, from both carriage and dis-

ease. I then performed associations between all of these variants and three phenotypes: invasive disease potential, severity and mortality. This analysis showed the importance of pneumococcal variation beyond serotype for invasive potential, but not in disease outcome, and identified many putative genes and regions associated with increased or decreased invasiveness. I also performed an analysis of within-host variation between blood and CSF isolates, and while I didn't find adaptation specific to either niche I did find evidence of selection on genes post-invasion.

Finally I performed a GWAS of host variation with susceptibility and severity of meningitis. I found these traits to be heritable, but despite attempts at meta-analysis with other studies the relatively low sample size and possible prototypic heterogeneity hasn't yet led to a confirmed association in either case. I also attempted a genome-to-genome analysis using both host and pathogen variation. I calculated the limit of detection given the small sample size, and using dimensionality reduction and biological hypotheses found possible interaction effects.

In summary, I have made the following advances. I have developed one of the first methods to overcome the challenges of bacterial GWAS, and showed that it works better than existing approaches. Using this technique, and others, I have quantified the effect of pneumococcal variation on variation in carriage duration beyond the resolution of serotype, and found some of the specific variants which affect it. I also used this top-down approach of assessing the genetics of pneumococcal meningitis, both in host and pathogen. This was not based around known required virulence factors, and used variation occurring in the natural population. Analysis of within-host diversity during meningitis found selection acting on additional genes. I calculated the heritability of host susceptibility to pneumococcal meningitis, and performed an association study using human genetic data. I also attempted the first genome-to-genome analysis with bacterial genomes and human genotypes.

6.1.1 Bacterial genome-wide association studies

Bacterial GWAS approaches have faced three main difficulties: lack of large sample collections, strong population structure confounding results and extensive pan-genomic variation. With the first restriction starting to be lifted, there is a need for scalable GWAS methods directly applicable to large populations of bacterial genomes. Such methods must account for population structure, and ideally assay variation in both the core and accessory genome without relying on a reference alignment.

The use of k-mers to assess pan-genomic variation had previously proven successful, so I wished to implement an approach which could efficiently perform associations using these as sequence variants. As the application of phylogeny based approaches are restricted due to their heavy computational burden and the need for an accurate recombination-

free tree, I opted to adapt regression-based methods used in human genetics to apply to bacterial GWAS. I wrote code to maximise the likelihoods of these regressions in C++, using efficient optimisation techniques as a first try, and more robust methods as a second pass.

To work out how to deal with population structure I compared various approaches in terms of accuracy and computational burden for phylogeny reconstruction. Knowing that I would be using k-mers in the association, I found that a method using the Jaccard distance between subsets of overall k-mers was sufficient to control for population structure in my simulations, and for antibiotic resistance in *S. pneumoniae*. Since writing this code the minHash distance has been adapted for distances between genetic sequences (Ondov et al., 2016), and can now be used as a more efficient replacement for Jaccard distance. I used the eigenvectors with the three largest eigenvalues calculated from this pairwise distance matrices in a fixed effect logistic or linear regression, in analogy with the standard method used in human genetics. To deal with possible very large effect sizes in these regressions I used the LRT for significance, and Firth regression for when data was nearly separable (as in trimethoprim resistance).

This approach proved to be broadly successful for antibiotic resistance in *S. pneumoniae*, worked with simulated data, and found a potential virulence factor in *S. pyogenes*. However, in all of these cases the predicted effect size was very strong, and population structure was generally not strongly associated with the phenotypes tested. I did not test whether the population structure correction I applied here was more broadly applicable, and would be sufficient in other species or phenotypes where these conditions no longer hold. The use of more eigenvectors should improve the trade-off between false positive rate and power, but it may be the case that including them as random-effects under a linear mixed model may offer the best option. When used for carriage duration, I found that a LMM had slightly higher power for detecting homoplasic low frequency effects when compared to using fixed effects while controlling for false positive rate. However, it was not as useful as the fixed effects model for including possible lineage associated variants for follow-up elsewhere. For invasiveness of *S. pneumoniae*, the fixed effect model using ten population structure components appeared to have a high false positive rate, where the LMM offered better population structure control and was still powered to find associations.

I have therefore already observed situations in which different methods would be the best to use. A comparison of these possible methods based on a range of population structures, phenotype distribution, recombination rate/homoplasy, effect sizes, lineage and locus associations would be useful, and is not something I attempted here. It is difficult to simulate realistic bacterial phylogenies, and synthetic associations introduced as part of this kind of simulation may be easier to find than associations in real populations. To perform this comparison I would take observed sequence alignments from real populations, and introduce synthetic associations using eq. (2.11) over the range of parameters of interest.

A comparison of power and false positive rate of fixed and random effect regressions as well as phylogeny based methods would be useful for future applications.

The population structure correction I used in SEER is a reasonable start, and works well for strong effects such as antibiotic resistance. A comparison with other possibilities with positive and negative controls (either simulated or known associations) will help inform future development. I mostly tested methods on locus effects, and have ignored or controlled for lineage effects in the output. In the future, a ranking of lineage effects in the output would be useful in case lab-based follow-up of these sites is possible. Clear assignment of sites as either lineage or locus effects would be helpful too, and ancestral state reconstruction combined with a comparison between adjusted and unadjusted test statistics may help classify variants into one of these two classes.

A difficulty in both cases is picking a significance threshold. In my first attempts I reasoned that every possible site in the genome multiplied by all three possible mutations should be used as the number of tests, and backed this up with permutation testing. However, as samples are not independent and identically distributed due to their genetic relatedness then permuting phenotype labels may not be appropriate, as it assumes any switch of label has the same effect ignoring any covariance between samples. Permuting labels within population clusters may be better, but likely too conservative. Monte Carlo permutation using the a covariance structure calculated from the phylogeny is also possible, though with the usual caveats of computational burden and reliance on a high-quality tree. Inspection of Q-Q plots is useful and can visually allow for the identification of a breakpoint between population structure effects and a significant signal, and how much the former is affecting the association model overall. While this is not a consistent way of choosing a threshold, it can help with ranking the top hits. For the LMM, where population structure is well controlled at the lower end of the p-value spectrum, a conservative Bonferroni correction based on number of patterns seems appropriate based on the Q-Q plots tested here. For fixed effects models picking this threshold remains a challenge, though Q-Q plots can help.

The use of k-mers worked well in the applications tested, and managed to find associations SNPs would not. They enjoyed the expected advantages from not requiring an alignment or clustering of orthologous genes. In cases where nearby SNPs independently affect the phenotype, which occurs in some antibiotic resistance genes, k-mers may be split up into lower frequency sequence units, lowering their power. In later chapters I therefore assessed variation through k-mers, SNPs and COGs where possible. The interpretation of k-mers has proved more challenging, due to the difficulty of mapping to the correct place (particularly with smaller k-mers) in a well annotated genome. Ideally k-mers would further be annotated by labelling SNPs and their predicted functional change in the k-mers, using the ancestral state as reference. However, to map an associated region, especially mediated by gene presence/absence and not fine-map the function, k-mers have been

successful.

6.1.2 Epidemiological variation of *S. pneumoniae*

Duration of carriage of *S. pneumoniae* is an important measure of strain fitness in epidemiological models, and its variation has been proposed as a mechanism by which antibiotic sensitive and resistant strains can coexist. Previous analysis of the source of this variation have been limited to serotype resolution, so using genome sequences from a longitudinal study cohort offered the opportunity to refine the analysis of variance.

I first developed HMMs for longitudinal swabbing data per serotype, to allow carriage acquisition and clearance rates and false negative swabbing rates to be estimated from the whole data rather than from a set of assumptions. The only model that converged for the most common serotypes was the simplest: two states for carrying and not-carrying. These parameters could then be applied to individual carriage episodes to infer the most likely durations based on the observed data. Using these durations as a continuous phenotype, I used a LMM to investigate and quantify the variance components caused by serotype and resistance, and GWAS to identify possible specific genetic variation which further contributed to variation in carriage duration.

I found that bacterial genomic variation had a significant effect on carriage duration, and that serotype was the largest lineage effect. However, only serotype 19F appeared to have a contribution independent of the genetic background. I also identified prophage k-mers which were associated with a lowered carriage duration, and evidence that this may work through interruption of the competence mechanism (by inserting into the *comYC* gene). These findings support the existence of duration and fitness modifying alleles in the natural population, which can be used to explain coexistence of antibiotic resistant and sensitive strains despite strong fitness differences depending on whether treatment is currently being applied (Lehtinen et al., 2017). The increased precision of the carriage estimates, per carriage episode rather than per serotype, along with provision of useful covariates such as *comYC* status, host age and previous carriage will also be useful data for models of coexistence and transmission.

However, one of the main limitations of this analysis was the monthly swabbing resolution. While clearly a large and well-sampled collection, the design of swabs spaced linearly in time to probe carriage durations which appear to be exponentially distributed is suboptimal. A design that would be better for this purpose is exponentially distributed sampling of cases that remain positive (Abdullahi et al., 2012a). Given the swabbing design available here, the estimates of effect sizes of the explanatory variables on carriage duration were therefore positively skewed.

As with all GWAS studies from a single population, results may be affected by batch effects in this population. Therefore meta-analysis of the results from this section with

another similar study would be useful before being generalised to the entire pneumococcal species. However, as this amount of sequencing has previously been unfeasible and as children need to be followed for two years, these studies are difficult to set up and long-running from start to finish. There are no other studies currently combining carriage duration estimates with genomic data, so this meta-analysis is not presently possible. We are aware of a similar study starting collection in Cape Town, South Africa, so I have released our results to facilitate comparison when this study's sequencing has been completed.

The function of altered carriage duration through *comYC* is an association only, and does not prove causation through this mechanism. While it is possible to make evolutionary arguments to support this interpretation, isogenic strains (controlling perfectly for genetic background) in an *in vivo* model would be needed to bolster this claim.

6.1.3 Host and pathogen genetics of pneumococcal meningitis

In chapters 4 and 5 I have used genomic variation of infecting bacteria and human host respectively to determine the impact of genetics on susceptibility to and severity of pneumococcal meningitis. Heritability analysis showed that for susceptibility, host genetics played a role and the genome sequence of the infecting strain is very important in whether invasive disease can occur. For severity of disease a different picture emerged: pathogen variation is unimportant, and host genetics is likely to play a small role. Though the estimation of specific heritabilities with binary phenotypes can be problematic, the data and multiple models support this overall conclusion.

I was unable to find and validate specific host associations through meta-analysis with other studies given the current sample collection. This rules out the existence of common variants with large effect sizes, the fitness defect of which would be unlikely to exist evolutionarily. Whether the variation which contributes to this phenotype consists of low effect size common variants, or rarer large effect size variants is a question that will need to be answered by future studies with larger sample sizes and more sequencing covering the entire variant frequency spectrum.

I did not include the sex chromosomes in the present analysis due to difficulties with imputation, though I did perform an earlier analysis of the X chromosome when using `impute2` in the Dutch population that did not show any association with any of the phenotypes. Tools are being developed to deal with the sex chromosomes in the same way as the autosomes (Wise et al., 2013), and the imputation server and reference panel now allows the X chromosome to be included. Future analyses should therefore not ignore this variation.

Another issue was phenotype heterogeneity, as the cohorts differed in terms of participant age and the exact disease presentation. While these differences have not been

found to matter for many phenotypes, it is possible that differing effect sizes in the subtly different phenotypes here are making associations impossible to find given the model used. The sample size here may benefit from a specific model allowing for this heterogeneity and expected correlations between effect sizes (as evidenced by the lack of signal from subtest), though a simple first step would be to perform meta-analysis of only a subset of the available studies to test for this possibility.

For the bacterial genetic contribution to meningitis, using GWAS I found many regions of the genome to be associated with invasiveness. Reassuringly, positive controls such as capsule (which I separately estimated to account for half of the variation in invasiveness) and LoF mutations in virulence factors such as *zmpD* were found in this analysis. Some other genes had previously been reported to affect virulence in invasive disease models, and these results increase support for their importance in human disease too. The remaining regions were associated with virulence for the first time here, and may suggest new functions for these genes, or an impact on virulence through unknown interacting gene networks.

I used a simple burden test when testing the effect of rare variants, which would not be suitable if the variants included in the set had different directions of effect. While this is probably correct for LoF variants, a different test may increase power for rare missense variants affecting protein function. If there is still strong population structure at the tips of the tree the method I have used has not explicitly accounted for it. It would be possible to instead group variants manually, and perform the association using a LMM. A similar caveat exists with the Tajima's D analysis of differential selection, where permutation testing may be insufficient to correct for population structure. In this case, the confounding effect of different population histories or different effects of vaccine introduction may be impossible to disentangle from signatures of selection.

These GWAS results are particularly susceptible to batch effects, due to the difficulty of getting a perfectly matched sample of the population from carriage and invasive disease. When analysing binary traits, if a covariate (such as serotype) is perfectly correlated with the trait, then all the results will be confounded too. Therefore a crucial next step, before further interpretation, is replication and meta-analysis with another population where both carriage and disease have been sampled. Hits from both populations will then be much better supported as the confounders may cancel out if in random directions, and power will be raised for rarer and lower effect size variants. A project is underway in South Africa which has taken such a sample, so we intend to perform this meta-analysis using those sequences.

As mentioned in chapter 1 part of the power of GWAS over linkage studies comes from the simple study design, where as many samples as possible are used without necessarily worrying about matching for covariates or genetic background. These confounders can then be adjusted for in the downstream analysis instead, which maximises discovery power.

This is broadly true for bacterial genomes too, however the effect of population structure is a much stronger confounder, and for some phenotypes which are tightly correlated with genetic background (high heritability) this can make discovery of anything other than homoplasic variants impossible. An alternative study design is to instead compare variation from within the same bacterial population when it has divergent phenotypes. For example, sampling the diversity of the bacterial population within-host in the carriage niche and an invaded niche is not confounded by population structure (and also host covariates such as age and immune response) as the genetic background is the same. Performing a meta-analysis of the variation found to be associated with either niche across multiple samples will then find those variants which occur during infection which have allowed adaptation to the invaded niche.

I performed this analysis between blood and CSF isolates, as previous work on a single case of pneumococcal meningitis had found convincing evidence for evolution occurring during invasive disease. When I expanded to hundreds of cases, I found no evidence of any variation causing adaptation to either the blood or CSF niche during disease. The sample size was large enough to conclusively state that variation occurring after invasion is rarely important for the progression of meningitis. However, when comparing the variation present in populations from invasion to carriage reference sequences I did find signs that *dlt* loses function in carriage more frequently than would be expected, and that *pde1* is under selection in invasion. To refine this analysis of variation occurring within-host between carriage and disease I would need to use more samples than analysed here, and also deeper sequencing of samples to assay the background of variation that exists within the founding population that is then selected.

6.2 Future directions

6.2.1 Bacterial GWAS methods

Since its release, I have received feedback about SEER which, if implemented, would make it into a more broadly usable and applicable piece of software for microbiologists. In terms of software development and installation, inclusion of SEER in a common 'container' would make installation automatic for those without C/C++ development experience, deal with differences between platforms and ensure all users are working with the same version of the code base.

I designed SEER with k-mers in mind, and therefore concentrated on making a scalable piece of software with a single input source. As mentioned, k-mers may not be the ideal variant when close SNPs are associated with a phenotype as the resulting k-mers will be split up into words of smaller frequency, and therefore power. For some purposes it may be useful to allow other forms of input such as VCF for short variants (SNPs and INDELs) with respect to a reference, and a general presence/absence matrix for COGs and aligned intergenic regions. The interpretation of k-mers can be challenging, both in finding a suitable reference (even from the entire `nr/nt`) to map to and annotate them with, and to determine whether they represent presence/absence of a region or variation within the region. It has been recently argued that population variation is best represented by a pan-genome graph, with shared haplotypes of any length being the natural variant (Marschall et al., 2016; Paten et al., 2017). Though the counting of informative k-mers goes some way toward testing longer variants, testing haplotypes may improve association power and make interpretation easier. A method has been proposed using unitigs (high confidence contigs not requiring repeat resolution), though this is not likely to scale beyond hundreds of samples (Jaillard et al., 2017). Integrating a scalable approach such as `vg` (variant graph – https://github.com/vgteam/vg) would be a promising way to include haplotype association.

Section 4.4.2 considered rare variation in GWAS assuming population structure was not an issue, due to low frequency variants occurring at the tips of the phylogeny. Including a way to input pathways of variants in SEER would relax this assumption, and also allow both gene-based burden tests (in either direction) to be extended to operons and functional pathways. Adding a model such as `SKAT` (Wu et al., 2011) would also improve power when rare variants in a functional pathway do not all act in the same direction on the phenotype of interest.

I picked a single method to adjust for population structure in SEER, but many others could be used. For example, as shown in chapters 3 and 4, the fixed effect model of SEER is in some cases a poor control for population structure. In the current implementation, BAPS clusters could be used as a categorical covariate in the regression giving a similar test to the CMH. A LMM has generally shown good control of population structure, likely thanks to

using all SNPs in the population structure correction rather than a proportion through picking the top principal components. The LMM normally has complexity $\mathcal{O}(MN^3)$, which is infeasible for the GWAS problems considered here and as sample sizes grow in future. The model of FaST-LMM rotates the design ($\boldsymbol{X}$) and relatedness ($\boldsymbol{G}$) matrices so the regression becomes linear along the eigenvectors of $\boldsymbol{G}$ (first using a singular value decomposition of $\boldsymbol{G}$), which with correct selection of $\boldsymbol{G}$ has complexity $\mathcal{O}(MN)$ (Lippert et al., 2011; Kadie & Heckerman, 2017). In this case, $\boldsymbol{G}$ is a SNP-wise distance between samples. This is similar to the $\mathcal{O}(MN^2)$ phylogenetic regression method of Pagel (1997) which transforms correlated error terms (due to relatedness between samples) into uncorrelated errors by diagonalising the variance-covariance matrix $\boldsymbol{G}$. In this case, $\boldsymbol{G}$ is the distance between the root and MRCA of each pair of samples. These methods could be included as new association models in SEER to allow for population structure correction when the current fixed effect model is not appropriate.

The effect on GWAS power and false positive rate of these different population structure corrections is unknown, and will likely be different depending on variant penetrance, level of homoplasy and frequency. A simulation-based comparison between these methods over a range of situations would therefore be useful. Based on the simulations used in section 2.6.1, the best way to do this would be by adding in synthetic associations of different penetrance at various points of the phylogeny of a real population using eq. (2.11), which would allow varying homoplasy and frequency.

I used heritability and genomic partitioning to support the conclusions in chapters 3 and 4. While this is well-supported for continuous trait used in the former, the use of the liability scale for bacterial traits in the latter has not been properly explored. Extension to binary traits would be useful, and support of the applicability and robustness of the methods used from simulated data will be important for having faith in quantitative estimates. If this could be shown to work, the estimates of serotype importance may be better estimated in a framework where genetic background is separately accounted for.

The use of SEER has been exclusively to single traits, but with the increasing availability of high dimensional phenotypes as seen in genome-to-genome analysis (section 5.3), pheWAS (Bush et al., 2016) and eQTL studies (L. Franke & Jansen, 2009; Wang et al., 2009) the addition of a multitrait model could be considered. Transcriptomic data is now being produced for bacteria (Bruchmann et al., 2015), so improved association power of SEER for this purpose will be useful. Rather than associating every phenotype or transcript separately, necessitating a harsh multiple testing correction, the correlation structure of multiple traits can be exploited to find latent variables (biologically representing functional pathways) to test for association with genetic variation improving power (Marttinen & Corander, 2010; Marttinen et al., 2013; Marttinen et al., 2014). Recent implementations of non-negative matrix factorisation are fast, and a promising way to find latent variables in high dimensional phenotypes (Zhirong Yang et al., 2016) – so could be added as a further

module in SEER.

6.2.2 Genetics affecting pneumococcal meningitis

Further analysis using GWAS could further explain the biology of pneumococcal infection. A simple additional analysis would be adult versus child colonisation using the Dutch carriage population – I have already catalogued the variation, and host age is available for all samples. Any results may be informative of the differences in immune system evasion depending on host response, and could be important for vaccination which currently targets children.

In the carriage stage, bacteria will only persist in the population if they can be transmitted between hosts; 'transmissibility' of *S. pneumoniae* is therefore a measure of fitness. Alleles which affect transmissibility may also be a promising vaccine candidate, as compared to PCV they will reduce colonisation (and therefore disease) of all serotypes. Zafar et al. (2017) have shown that *ply* is necessary for transmission, as the host cell damage it causes increases shedding. A GWAS of *S. pneumoniae* transmissibility may be able to detect more subtle effects of alleles which occur in the natural population.

Nebenzahl-Guimaraes et al. (2016) performed a GWAS on transmissibility of *M. tuberculosis* by selecting low transmission strains from at-risk hosts with rare genotypes and high transmission strains from low-risk hosts with common genotypes. A similar way to perform this analysis would be to use the carriage durations I estimated in chapter 3 and assume equilibrium transmission in an susceptible-infected-recovered (SIR) model in the Maela population, which would then allow calculation of strain transmissibility from carriage duration divided by strain prevalence. However, evidence from infant mouse models suggests *S. pneumoniae* transmission may only occur shortly after colonisation, when inflammation is highest promoting increased shedding (Kono et al., 2016; Zafar et al., 2017). In this case a more complex transmission model using genetic similarity and infection times may be more appropriate, and model comparison between different functions of transmission intensity with respect to time would also be useful for inferring the biology of real-life transmission. Numminen et al. (2013) proposed a more flexible transmission model for the Maela population which was fitted with approximate Bayesian computation. Due to many proposals of the transmission tree being inconsistent with the observed infection times (and being assigned $\mathscr{L} = 0$) the fitting was computationally intensive; the use of the carriage durations estimated here rather than single time-points may ameliorate this problem. Inference of alleles affecting transmissiblity could then be jointly estimated in the process of inferring the transmission trees. Alternatively, if the dimension of genetic variation is too high, they could be inferred separately by first calculating strain-wise transmissibility from the transmission trees and then using these as a phenotype in GWAS. An alternative approach would be to sample within-host diversity

by deep sequencing of swabs, which allows finding the genotypes which make it through the transmission bottleneck in each case through ancestral state reconstruction (where the trait is the identity of the host). Averaged over many transmission chains, the variation shared by these genotypes would represent transmissibility alleles.

In the analysis of host genetics affecting bacterial meningitis, a better model of the shared architecture between the subtypes of meningitis analysed may help find associations (Pickrell et al., 2016). Rather than using subtest with underpowered genotype data, it may be better to use LD-score regression between summary statistics from all the studies available, which would allow estimation of coheritabilities between the different sub-phenotypes (Bulik-Sullivan et al., 2015). To aid in increasing power for detecting host genetics we have applied to access the UK biobank (http://www.ukbiobank.ac.uk/), which is about to release 500 000 genotypes of a richly phenotyped UK adult population. These phenotypes include ICD-10 codes, which show hospital diagnoses for bacterial meningitis, split up by causal species. Additionally, date of death is available, allowing inference of clinical outcome. The large size and well-defined phenotype of these samples will allow us to perform another GWAS, and meta-analyse the results with those of chapter 5 for both susceptibility and severity increasing discovery power.

The genome-to-genome analysis was limited by the small sample size when testing massive numbers of combinations of possible interactions. In future, the ~1 200 samples from the Danish cohort will also have the causal *S. pneumoniae* sequenced, allowing this analysis to be expanded. It may also be possible to model the effect of genome-to-genome interactions on severity as well as bacterial and host factors, by analysing a combined model of the form:

$$\text{severity} \sim X_{\text{bacteria}} + X_{\text{host}} + X_{\text{interaction}}$$

where the interaction term is $X_{\text{bacteria}} \times X_{\text{host}}$.

Finally, I would propose the following extensions to assessing with-host diversity during bacterial meningitis. As I have shown that selection does not occur between blood and CSF samples, but that it probably does occur between carriage and CSF, a greater number of carriage and invasive samples from the same patient should be taken: greater both in terms of the number of patients enrolled and in the depth of coverage of the within-host diversity. This is a difficult study to set up: in the MeninGene cohort recent attempts to swab bacteria from the nasopharynx of bacterial meningitis patients before treatment started yielded no positive cultures, likely due to the small carriage population (Wyllie et al., 2014; Wyllie et al., 2016). Alternative culture-free methods such as DNA pull-down may be helpful, or alternative a study in an alternative population with high rates of carriage may be able to achieve sufficient sample size.

The analysis of this data would benefit from an improved null model of mutation. In section 4.5 I assumed a simple model of equal mutation rate per base and Poisson dispersion

of number of mutations, which led to regions with higher mutation rates being found, and may have suppressed the discovery of genes with lower mutation rates. Improving this through a more refined model of mutation rates depending on sequence context and using observed dispersion of the number of mutations would be a useful extension (Samocha et al., 2014; Aggarwala & Voight, 2016). If more mutations were observed, using the observed number of synonymous changes, which are assumed to be neutral, as a basis for the null would also help (Ding et al., 2008). Finally, experimental evolution without selection pressure may give the most accurate null model (Tenaillon et al., 2016), though an experiment recreating the bottlenecks encountered in pneumococcal meningitis has not yet been performed.

6.2.3 Future of statistical genetics in bacterial diseases

Statistical genetics, and specifically GWAS, of host and pathogen genetics contributing bacterial diseases is still in its infancy. Looking at the boom in human genetics and given the large sample sizes becoming available, it is reasonable to expect the field to continue to expand. The near future is likely to consist of further methodological improvements and analysis of new phenotypes, going on to functional validation and eventually integration with host data. I hope that I have presented some reasonable early steps in this field in this thesis, and that others find elements of what we've done useful for future research.

Thanks a lot for reading all the way to the end! (unless you skipped straight here)