Chapter 5

Human genetics contributing to invasive pneumococcal disease

Declaration of contributions

Jeff Barrett, Diederik van de Beek and Stephen Bentley supervised this work. Data collection: Diederik's group designed and ran the Dutch MeninGene study; Thomas Benfield designed and ran the Danish study; Matthjis Brouwer and Bart Ferwerda compiled clinical metadata for the Dutch cohort; Lars Henrik Ängquist, Thorkild Ingvor Arrild Sørensen and Ellen Aagaard Nøhr provided quality controlled genotype data for the GOYA study; Alexander Mentzer and Julian Knight provided summary statistics for the GenOSept study; Chao Tian and David Hinds provided summary statistics for the 23andme data. Philip Kremer performed re-impuation of the *CFH* region. I performed all other analyses.

5.1 Introduction

The previous chapter has considered variation present within the pneumococcal genome that is associated with colonisation and invasive disease, while mostly treating the infected hosts as identical, with the exception of section 3.6 where I showed infant age and previous colonisation were both associated with carriage duration. However, the hosts are in reality heterogeneous: as epidemiological parameters such as contact network (Dagan et al., 2002; P. C. Hill et al., 2010), vaccination status (Klugman, 2001), co-infections (McCullers, 2006; Siegel et al., 2014; Cohen et al., 2013) host age and immune response (Cobey & Lipsitch, 2012) have all been shown to affect invasive pneumococcal disease.

However, as well as varying in these ways, humans differ in the sequence content of their genomes. The contribution of human genetics to adult pneumococcal meningitis is presently unknown – both whether it affects the disease at all, and if so which specific regions of the genome contribute to the effect. Twin studies (Jepson, 1998; Burgner et al., 2006), linkage studies (Abel & Dessein, 1997) and then GWAS studies have all suggested a role for human variation for many bacterial diseases (Chapman & Hill, 2012). Association of HLA allele as well as other regions have been found. Despite likely being selected against over human history, variants pre-disposing to bacterial diseases as stable and enduring as tuberculosis have been found (Curtis et al., 2015; Sveinbjornsson et al., 2016).

I start this chapter by using genotype data from the MeninGene (section 1.1.4) cohort to calculate the heritability of susceptibility to and severity of meningitis (section 5.2). After I found that human genetics is expected to explain the variation in these traits, I performed a GWAS for each trait to find specific regions of the genome associated with bacterial meningitis and its progression. To obtain more evidence for the associations, and increase power, I then performed the same analysis in two additional cohorts, and finally meta-analysed the results of all of the studies with a further two previous cohorts for which we obtained summary statistics.

In section 5.3 I bring host and pathogen genetics together by performing a genome to genome analysis, using cases of pneumococcal meningitis from the MeninGene cohort where both the pathogen genome and corresponding host genotype was available. Rather than looking for human variants which affect meningitis susceptibility and severity regardless of the bacterial variation, this section attempts to find specific bacterial variation which correlates with specific host variation to contribute to disease. This can be considered an interaction, between the genomes. As interactions between host and pathogen proteins are known to be important in pathogenesis (Lambris et al., 2008; Serruto et al., 2010), this is a plausible avenue to explore and may further determine the genetic architecture contributing to infection in clinical cases of disease.

5.2 GWAS of human variation associated with meningitis

The MeninGene collection was built up in three batches over the course of this work: the final numbers along with each phenotype are shown in table 5.1. As the collection includes all consenting adults with culture-proven meningitis, all causative pathogen species are included in the collection. My analysis so far has mostly been restricted to pneumococcal meningitis, as being the most common cause of meningitis in adults it is the most well powered. However in this chapter I will also consider meningitis as a whole, which also includes cases caused by *N. meningitidis*, *L. monocytogenes* and *H. influenzae*. As well as microbiological data, clinical information has been collected for most cases, allowing an association of disease severity as in section 4.4. For the association I used genotype data from the ALS (van Es et al., 2009) and B-PROOF (van Wijngaarden et al., 2011) as population matched controls, all of whom were adults.

Cohort	Country	Age	Data	Samples	Phenotype
MeninGene	Netherlands	Adults	Illumina Omni array	1 149	Meningitis
				732	Pneumococcal meningitis
				277	Unfavourable outcome
ALS & BPROOF	Netherlands	Adults	Illumina Omni array	4 836	Controls
Benfield	Denmark	Children	Illumina Omni array	353	Pneumococcal meningitis
				873	Pneumococcal bacteremia
				473	Controls
GOYA	Denmark	Young adults	Illumina quad array	2 805	Controls
23andme	European	All	Summary statistics	842	Bacterial meningitis
				82 778	Controls
GenOSept	European	Adults	Summary statistics	220	Pneumococcal bacteremia
WTCCC	UK	Adults	Summary statistics	2 244	Controls

Table 5.1: Summary of cohorts with available human genotype data. The first section shows cohorts with full genotype data where I performed a GWAS; the second section is cohorts with the summary statistics from an existing GWAS used in meta-analysis only. Sample numbers are after the QC in section 5.2.1.

I also used data from Danish children with invasive pneumococcal disease (referred to here as the Benfield cohort). Using archived blood spots in the Danish national biobank, we extracted DNA for genotyping from cases of children with pneumococcal meningitis and bacteremia, as well as 473 population controls. As additional population matched controls I obtained the genotypes of controls from the GOYA study, which randomly sampled 2 805 healthy Danish young adults (Paternoster et al., 2011).

Finally, summary statistics were available from two existing studies. The first, performed by 23andme, gave participants a questionnaire on infectious diseases. Those responding yes to the question 'Have you ever had bacterial meningitis?' were classified as cases, and those responding no as controls ('I'm not sure' was also an option, and these responders were excluded from further analysis). The analysts performed a logistic

regression at all imputed SNPs using age, sex and the first four principal components as covariates (Tian et al., 2016). The second is the unpublished GenOSept study which included 220 adults with sepsis, who suffered shock in intensive care unit (ICU) and were either blood culture positive from pneumococcus, or were positive from pneumococcal antigen in their urine. The analysts used controls from WTCCC (Burton et al., 2007) and performed a regression at all imputed sites using a linear mixed model as implemented in `gemma` (Zhou & Stephens, 2012).

5.2.1 Genetic data processing

In this section I describe the set of steps I took to prepare genotyping intensity data for GWAS analysis. From the Dutch cohort there were initially 905 cases available from the collection since the Meningene study began, with a second batch of 94 new cases covering a subsequent winter, and a final third batch of 178 new cases covering a subsequent two winters. As controls, 1 981 samples from the ALS study, and 2 898 from the B-PROOF study were available from the start. From the Danish collections, 373 meningitis cases and 475 controls were available as called genotypes, and we genotyped 904 additional samples with pneumococcal bacteremia. I also applied for access to 2 817 samples from the GOYA study, which I received as quality controlled genotype calls.

The following analysis was completely repeated four times to arrive at the final SNP calls used in the association study. The processing steps and cut-offs used were the mostly same for all of these genotyping runs, however I do point out where steps differed based on cohort or run, and where cohorts or runs have been merged. Throughout, I have used a combination of plink v1.9 (Purcell et al., 2007; Chang et al., 2015) and my own perl scripts (https://github.com/johnlees/bioinformatics) to convert between different data formats.

Genotype calling

Genotyping arrays have hundreds of thousands of SNP probes, allowing for a relatively cheap assay of all common ($>$ 5% MAF) positions in the human genome. For each variant, there is a red florescently tagged probe which binds to the A allele, and a green probe which binds to the B allele. By comparing the relative intensities of these two colours across a large number of samples a genotype probability can be assigned to each sample in the run.

We processed raw genotyping data using Illumina's Beeline software to produce normalised intensity files. In these files, for each sample an x and y intensity is recorded at every SNP typed by the array, proportional to the amount of the A and B allele present. In the ideal case a sample homozygous for A would have high x and low to no y intensity, whereas a sample homozygous for B would have the opposite. Heterozygous samples would have half of each intensity. In practice the intensities are distributions (fig. 5.2),

and the best way to produce genotype calls is to plot x against y for many samples, and find three discrete clusters. As a final complication, at any given site some samples will act anomalously and either fail to produce an intensity, or worse produce an extra cluster which confounds the identification of the real genotype clusters. Such samples should be assigned a missing genotype at these sites.

As high quality genotyping is important for downstream imputation and any eventual fine-mapping (Spain & Barrett, 2015), I used optiCall (Shah et al., 2012) to deal with these issues and produce genotype calls for all the samples with genotype intensity data. This method has been shown to throw away fewer correctly typed variants than other methods, and produce more accurate calls overall. The algorithm first samples random intensities from across the genotyping run to generate priors of where the three genotype clusters are centred, then for each variant uses an EM algorithm to adjust class membership based on these priors and the observed data.

I ran optiCall using default settings on a per chromosome basis separately for each genotyping run, using the sample sex as a covariate. In the second and third rounds of Dutch case samples, each batch contained fewer than 200 samples. So at the rarer end of the SFS, less than one sample is expected to be in the homozygous rare category. While optiCall is robust to missing classes in rare variation, it needs reliable prior information to do so. To ensure high quality calling of these runs I therefore:

1. Combined the meningitis samples with intensity data from a run of 41 samples from a European population on the same platform, used by another study.
2. Treated the run ID as a covariate in optiCall.
3. Used chromosome 1 to generate priors for all other chromosomes, as it contains the most number of variants.

After calling, I discarded the samples from the other study. I will cover direct assessment of genotype call quality in section 5.2.1.

Quality control of genotype data

When performing QC of the called genotype data I followed the advice of C. A. Anderson et al. (2010), though using more modern and faster algorithms where appropriate. I first merged the first two runs of Dutch cases and controls, giving five sample sets to QC (Dutch combined, Dutch case batch three, Danish meningitis combined, Danish bacteremia and Danish controls).

For all these datasets, I performed the following basic QC steps using `plink`:

1. Predict sample sex using genotypic data (heterozygosity rate on X chromosome). Where discordant with recorded phenotypic sex, or the phenotypic sex was missing, I replaced it with the predicted value.

2. Remove samples with an overall heterozygosity rate above three standard deviations from the mean.

3. Remove samples with $> 3\%$ of genotypes missing.

4. Remove markers with $> 5\%$ of genotypes missing.

5. Remove markers with a significantly different call rate between cases and controls ($p < 10^{-5}$).

6. Remove markers with MAF $< 1\%$.

7. Remove markers out of Hardy-Weinberg equilibrium (HWE) ($p < 10^{-5}$).

Failing samples were removed before failing markers, to maximise the number of markers retained. Steps 2–5 remove those samples and markers which have not been genotyped well on the array, whereas step 6 removes those markers with insufficient power to inform imputation or association. Step 7 is useful in discarding genotype failures as almost all markers are close to being in HWE, so the number of samples in each genotype group can be related to the MAF. Departures from HWE are mainly due to genotyping failures, where clusters have been incorrectly merged or labelled. However, while a good first step, this step is not sufficient to remove all genotyping failures.

I then estimated sample ancestry and relatedness within each collection. To estimate degree of relatedness between samples I used KING with default settings (Manichaikul et al., 2010). For ancestry, I first removed palindromic SNPs (A/T or C/G) to minimise strand issues, and merged with the genotype data with 270 individuals from four different ancestries released as phase II of the HapMap project (International HapMap Consortium, 2005; International HapMap Consortium et al., 2007). I then used `eigenstrat` to perform PCA on the merge of samples and hapmap to identify and control for ancestry (A. L. Price et al., 2006).

I did not immediately discard these samples as they can still be included in a linear mixed model to increase discovery power (Lippert et al., 2011; Zhou & Stephens, 2012). Instead, I only removed identical samples, and recorded those which were related as third-degree or closer, and samples of non-European ancestry (PC1 < 0.07 in the hapmap projection; fig. 5.1). These were only removed in downstream analyses requiring unrelated samples from the same population.

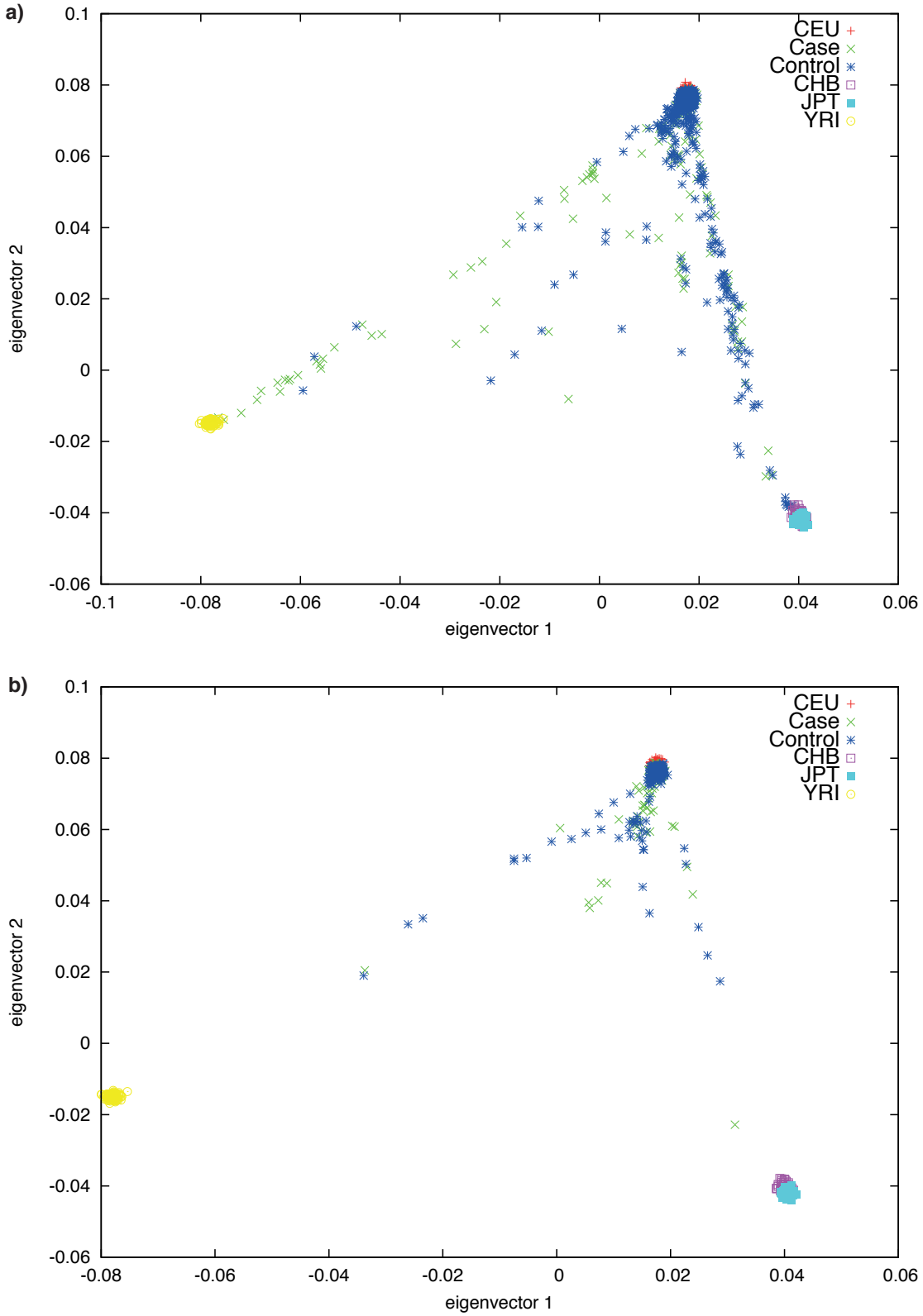


Figure 5.1: Projection of samples onto first two principal components of case (green crosses) and control (blue stars) samples from **a)** the Netherlands and **b)** Denmark with HapMap phase I populations. HapMap populations are 3 (red crosses) – CEU, European; 4 (pink squares) – CHB, Han Chinese; 5 (turquoise squares) – JPT, Japanese; 6 (yellow squares) – YRI, Yoruba Nigerians.

Using this first pass of QC, I performed an initial association test at all passing sites using a logistic regression. I removed all population divergent and third-degree or closer related samples, and fitted the basic model

$$\log\left(\frac{\boldsymbol{y}}{\boldsymbol{I}-\boldsymbol{y}}\right) = \boldsymbol{X}\boldsymbol{\beta} \tag{5.1}$$

at every marker, where $\boldsymbol{y}$ is the vector of binary phenotypes, $\boldsymbol{X}$ is the additive model matrix of genotypes (0 for homozygous common; 1 for heterozygous; 2 for homozygous rare) and β is the fitted slope. Using the Wald test p-values I found 226 sites suggestively associated with the susceptible phenotype $p < 10^{-4}$, and manually inspected the genotype cluster plots using Evoker (https://sourceforge.net/projects/evoker/files/). Many of these plots were miscalled in one or more cohorts, though in such a way that the HWE p-value managed to pass the filter set earlier. Some examples of faulty calling are shown in fig. 5.2 – all such identified variants were removed prior to downstream analysis and imputation. In addition, I performed an association within the control group, using the ALS study as cases and the B-PROOF study as controls. As there should be no overall phenotypic difference between these cohorts any significant results are likely artefacts from genotyping batch or incorrect calling (Burton et al., 2007). I therefore removed all markers with $p < 5 \times 10^{-8}$.

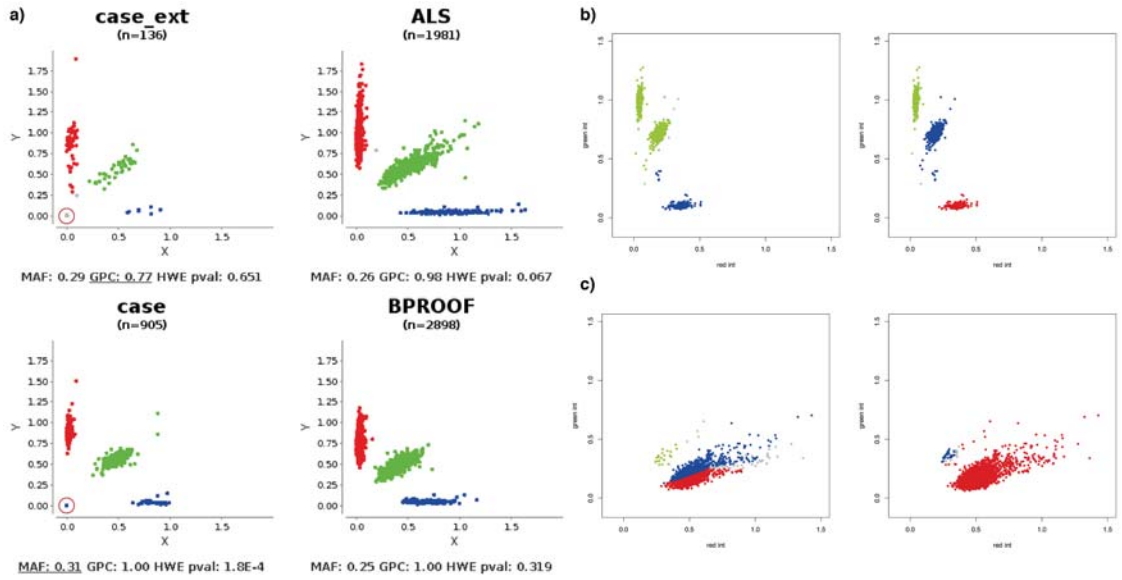


Figure 5.2: Examples of manual quality control of genotype cluster plots. All were removed rather than recalled. **a)** Evoker view of rs9516252. In cases missing genotypes have been mistakenly called as homozygous rare, whereas in cases-ext they were correct (red circles). **b)** A common mode of failure when cluster centres are not near the average. Left: incorrect identification of only two clusters at rs2717808. Right: corrected identification of three clusters. **c)** A common mode of failure when there are only two clusters at low MAF. Left: incorrect split into three clusters at rs17876189. Right: corrected identification of two clusters.

Imputation of untyped variants

To increase the power of GWAS it is common practice to impute the allele of untyped common variants in LD with those directly typed by the array, using haplotype information from whole genome sequenced population cohorts (Stranger et al., 2011). By finding overlap between genotyped alleles and haplotypes drawn from the population at these positions while taking into account population level LD it is possible to assign a probability of each genotype at all known variable positions. This increases the number of locations at which association can be tested for, mitigating the loss of low quality markers, and giving more information around signals of association. During genotype imputation all sites in the reference panel are assigned a most likely allele. At many common sites imputation accuracy is good ($R^2 = \sim 0.9$), and accuracy can be assessed through the INFO score which assesses how much information has been added at each position over the worst case of assigning the population MAF.

Humans are diploid organisms: they inherit one copy of a chromosome from their mother and the other from their father. However, as imputation works with haplotypes, a linear sequence along a single inherited chromosome, input genotypes must first be 'phased' into haplotypes. Phased data ensures that heterozygous SNPs are assigned to the chromosome they came from: for example if two alleles A/B were called as heterozygous and were next to each other possible haplotypes would be AA + BB or AB + BA (fig. 5.3). Data can be directly phased by barcoding which DNA molecules are being sequenced (Borgström et al., 2015), or by sequencing the sample's ancestors (mother and father). With genotype arrays used for GWAS direct phasing is not possible, but phased population reference panel data can be used to statistically estimate the most likely haplotypes of the input data (Delaneau et al., 2013; Loh et al., 2016).

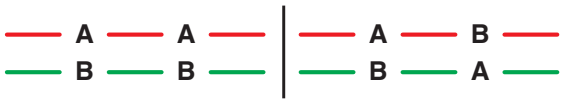


Figure 5.3: Demonstration of the effect of phasing. The subject is heterozygous for an A/B allele at two positions. The left panel shows one possibility, where the maternally inherited haplotype (red chromosome) is AA and the paternally inherited haplotype (green chromosome) is BB. The right panel shows the other possibility, of AB and BA haplotypes. Though there are another two possibilities gained from switching the parents, phasing does not distinguish these.

I performed phasing and imputation of variants using two methods. The first method, which I performed with the first batch of Dutch cases and controls, used the software `shapeit2` (Delaneau et al., 2013; O'Connell et al., 2014) and `impute2` (B. N. Howie et al., 2009; B. Howie et al., 2011) directly. I first merged the data, working with a file per chromosome across all case and control samples, then performed phasing with `shapeit2`. It is common to use the 1000 Genomes Project as the reference panel, as it contains a large collection of diverse haplotypes (1000 Genomes Project Consortium et al., 2015). It has

been shown that using a population specific reference panel can further increase imputation accuracy due to better matching, longer haplotypes being present between the reference panel and genotyped subjects (The Genome of the Netherlands Consortium, 2014). I therefore used `impute2` in reference panel merging mode, using both 1000 Genomes phase 3 (5 008 haplotypes) and The Genome of the Netherlands (GoNL) (998 haplotypes) as references to try and attain the best possible imputation accuracy for Dutch samples. I wrote a pipeline to automatically perform the imputation over a cluster system using this method by working in parallel on chunks of 2.5Mb at a time with a 250kb buffer to avoid loss of accuracy at the ends of each chunk, and automatically resubmitting failed jobs with more memory or wall-time as appropriate.

As more data became available later through the project, more efficient methods and sophisticated interfaces to phasing and imputation became available. Faster phasing became possible with `eagle2` (Loh et al., 2016) and faster imputation with `PBWT` (Durbin, 2014). This allowed the collection and use of the much larger and more diverse reference panel the haplotype reference consortium (HRC) (McCarthy et al., 2016). Though imputation accuracy is slightly lower than impute2, the efficient data structure and matching algorithm within `PBWT` allows rapid imputation even with the 63 000 haplotypes in release 1.1 of the HRC. The larger reference panel size overall gives good imputation accuracy, and includes both reference panels used in my previous imputation iteration. I therefore re-ran the phasing and imputation using this procedure, through the Sanger imputation server (https://imputation.sanger.ac.uk). Sex chromosomes were not included in this release, so all downstream analysis is of autosomes only.

To homogenise samples before imputation I used the HRC strand checking tool (http://www.well.ox.ac.uk/~wrayner/tools/#Checking). For each sample cohort, this checked whether alleles, strand of genotyping (which should all be on the positive strand, rather than the Illumina TOP strand), reference allele and MAF match with the reference panel. SNPs with MAF > 0.2 different from the reference panel are removed, which may assist with missed strand flips. I merged all samples with the same array version together (table 5.1) and then performed phasing and imputation.

Using the imputed data, I performed a final QC check on all the markers from the reference panel to remove low confidence sites. I re-applied the filters of MAF $> 1\%$ and HWE $p < 10^{-5}$, as well as removing any sites with an INFO score < 0.7 (suggesting poor imputation accuracy). After this step, 6.8M good quality SNPs were left for association. For phenotypes with lower numbers of cases (unfavourable outcome, genome to genome analysis) I applied a stricter MAF filter of $> 2\%$.

An initial association using eq. (5.1) revealed two quality issues not identified by the filters described. In both cases the issue was manifested by many highly significant p-values of markers, and non-significant values of those nearby and in LD with the lead variant. The first was a failure to match the strand between cases and controls, and in some

cases the imputation reference panel, at palindromic SNPs. At non-palindromic SNPs the reference strand is unambiguous and was correctly assigned by the strand checking tool, but at 1 722 (around 0.3% of genotyped positions) A/T or C/G SNPs with MAF $> 30\%$ neither allele or frequency mismatch could be used to disambiguate the genotype value. I used the Illumina genotype manifests data to ensure all genotypes were with respect to the positive reference strand rather than the Illumina TOP strand, and re-ran the imputation and subsequent QC on all affected cohorts.

The second issue was due to a mismatch of array design between cases and controls for the Danish bacteremia samples and GOYA controls. Despite performing separate QC and imputation of these cohorts to arrive at the same set of genotyped markers, a simple merge led to spurious association results. Although the imputation model in theory should allow for imputed sites to be merged when produced from different sets of calls, in practise subtle differences in genotyping quality and marker density for a large number of samples can easily lead to systematic differences between cases and controls. To match these two cohorts without introducing technical differences between them, I took the intersection of SNPs between the two panels and merged the genotype calls, then performed identical QC steps on the dataset as a whole. As this left only 291 830 markers ($\sim 50\%$ of that on a single array) I used `minimac3` via the Michigan imputation server (Sayantan Das et al., 2016) to perform imputation to the HRC, as this algorithm coped with the relative sparsity of markers better than `PBWT`.

Finally, as the *CFH* region was of particular interest given its previously reported association with meningococcal meningitis, we reimputed it for all the Dutch samples using `impute2`. In this mode we allowed `impute2` to infer the phasing during its MCMC which is far slower, but more accurate over this small region. This imputed data was used for meningococcal meningitis associations not reported here, and for the specific association with antigens in section 5.3.3.

5.2.2 Association results

Using the quality controlled genotype data I was able to perform three analyses on each cohort. The first was an estimation of heritability of each trait of interest, which represents the proportion of phenotypic variance explained by genetic variation. As in sections 3.3 and 4.4 I performed this calculation using different methods, as various technical limitations of each can bias estimates (Evans et al., 2017; Speed et al., 2017). All methods assume unrelated individuals with shared ancestry, so I filtered out these samples before performing heritability calculations.

I used the GCTA-GREML model, as implemented in `bolt-lmm` (Loh et al., 2015), which assumes normally distributed effect sizes with a variance equal to the genetic component of heritability σ_g^2 (J. Yang et al., 2010; J. Yang, Lee et al., 2011). Under this

assumption, restricted maximum likelihood optimisation of a LMM can be used to estimate h^2_{SNP}. This model does not adjust for LD, which in some cases may lead to underestimation of h^2_{SNP} (Speed et al., 2012). I therefore used LD-pruned SNPs as the input, and performed an additional heritability estimate with LDAK, which adjusts the weights of SNPs by their LD when calculating the kinship matrix used as the random effects in the linear mixed model.

After confirming that it is expected that a genetic contribution to the phenotype exists, I then ran an association scan. This performs a regression between variant and phenotype at every marker, though the use of an LMM allows ancestry and relatedness of samples to be included as random effects in the regression model. This means ancestrally divergent and related samples do not have to be completely removed, increasing the power to find associations without increasing type I error (A. L. Price, Zaitlen et al., 2010). It has previously been computationally prohibitive to fit this model to every imputed marker, but recent efficiency advances have allowed this technique to become commonplace (Lippert et al., 2011; Zhou & Stephens, 2012). I used `bolt-lmm` to perform the association (Loh et al., 2015), using LD-pruned genotyped markers to estimate the kinship matrix and random effects, and performing association at all genotyped and imputed sites. Where appropriate, I have included covariates such as immunocompromised status as fixed effects in the model.

The final question I wished to test using this data was whether there was evidence for difference of the genetic basis between similar sub-phenotypes of invasive disease. For example, is the association with *CFH* specific to meningococcal meningitis, or is it also shared by pneumococcal meningitis too? Overall, is there a difference in genetic susceptibility to different pathogens, or different manifestations of invasive disease? As the case numbers are low, these studies were underpowered to detect a difference through direct association of different sets of markers, or to calculate co-heritability. However, in such cases, performing an association between all cases and controls, and then between sub-phenotypes of cases may help test for an overall difference. Liley et al. (2017) have developed the subtest method which fits a mixture of Gaussians to the Z-scores from these two association tests, which compares the null model fit assuming no difference between subphenotypes and the alternative model when there is a difference. It can extract a p-value from the LRT which expresses the probability that the genetic basis for the subphenotypes are distinct. When running subtest I used the weights from LDAK to account for LD between associations, and performed 1 500 subsamples of 400 samples to generate the null-distributions of the test statistic.

Dutch cohort results

In the Meningene cohort I considered three different phenotypes: the susceptibility of adults to bacterial meningitis (using all cases), pneumococcal meningitis only, and severe (unfavourable clinical outcome) meningitis. In all of these associations I used immunocompromised status as a covariate (10% of cases) assuming that no controls were immunocompromised, as population prevalence is around 1% (van Veen et al., 2011; Harpaz et al., 2016).

The heritability analysis (table 5.2) showed that human genetic variation was expected to contribute to all of the phenotypes of interest. The size of the contribution varied, but was relatively high in comparison to other complex traits (Ge et al., 2017). In general LDAK estimated a higher heritability than GCTA-GREML, as expected from the structure of the models (Evans et al., 2017). Analysis using `subtest` as described above did not provide any evidence that pneumococcal meningitis was distinct from other bacterial meningitis (PLR = 0.25; $p = 0.75$) or that unfavourable outcome was distinct from overall meningitis susceptibility (PLR = 0.14; $p = 1.00$). However this may rely on relatively highly associated SNPs, which were not found with this few samples. Susceptibility to any meningitis has a significantly higher heritability than its sub-phenotypes, which also have heritability above zero. This is more consistent with some difference in genetic architecture between the phenotypes.

Phenotype	Method	Heritability	Error	Fit p-value
All meningitis	GCTA	0.418	0.064	2.4×10^{-6}
	LDAK	0.556	0.088	3.9×10^{-11}
Pneumococcal meningitis	GCTA	0.353	0.068	2.4×10^{-6}
	LDAK	0.416	0.096	3.9×10^{-6}
Unfavourable outcome	GCTA	0.192	0.067	2.8×10^{-5}
	LDAK	0.325	0.090	1.4×10^{-4}

Table 5.2: Human SNP heritability (h^2_{SNP}) of three meningitis phenotypes in Dutch adult cohort. Pneumococcal meningitis and unfavourable outcome are subsets from the 'all meningitis' phenotype. For each phenotype I estimated heritability using both GCTA-GREML and LDAK models, in every case there was evidence for heritability significantly above zero.

The Manhattan plots of the association results are shown in figs. 5.5 to 5.7. Across the three traits only one locus reached genome-wide significance: position 64680775 on chromosome 1, an intronic variant in *UBE2U*, was associated with unfavourable outcome (MAF = 0.43; OR = 1.62; p = 2.0×10^{-8}). *UBE2U* is part of the ubiquitin pathway (responsible for degrading proteins in the cell) (Gregory et al., 2006), but has not previously been associated with any other disease or trait. The signal also spanned *ROR1* (fig. 5.4), a protein of unknown function (Bainbridge et al., 2014) which has previously

been associated with cancers (Reddy et al., 1996) and pulmonary function (Lutz et al., 2015). Signals suggestive of significance for each trait are reported in table 5.3. Despite the lack of association from meningitis susceptibility, the heritability estimates above suggest that meta-analysis with more samples should be able to find associations with lower OR and MAF. I otherwise delay a detailed interpretation of results until they are replicated in an independent study and reach genome-wide significance in section 5.2.3.

Phenotype	Position	Effect allele	MAF	OR	p-value	Annotation
All meningitis	chr6:153582990	T	0.42	1.27	7.2×10^{-8}	Upstream of *RGS17*
Pneumococcal meningitis	chr6:117624549	G	0.46	0.77	8.8×10^{-7}	*ROS1* intron
	chr18:48403560	T	0.43	0.65	7.6×10^{-8}	*ME2/ELAC1/SMAD4*
	chr22:47506160	G	0.33	0.74	5.5×10^{-7}	*TBC1D22A* intron
Unfavourable meningitis	chr1:64680775	A	0.43	1.62	2.0×10^{-8}	*UBE2U/ROR1*
	chr4:182823804	A	0.33	1.58	4.1×10^{-7}	*AC108142.1* intron
	chr9:37382231	A	0.07	2.36	6.7×10^{-7}	*ZCCHC7/GRHPR*

Table 5.3: Signals of association in the Dutch cohort. I report the lead SNP at each associated locus with MAF $> 5\%$ and $p < 1 \times 10^{-6}$, and nearby annotated genes. The suggestive signal in all meningitis cases was also present when restricted to pneumococcal cases, albeit with a lower p-value of 3.9×10^{-7}.

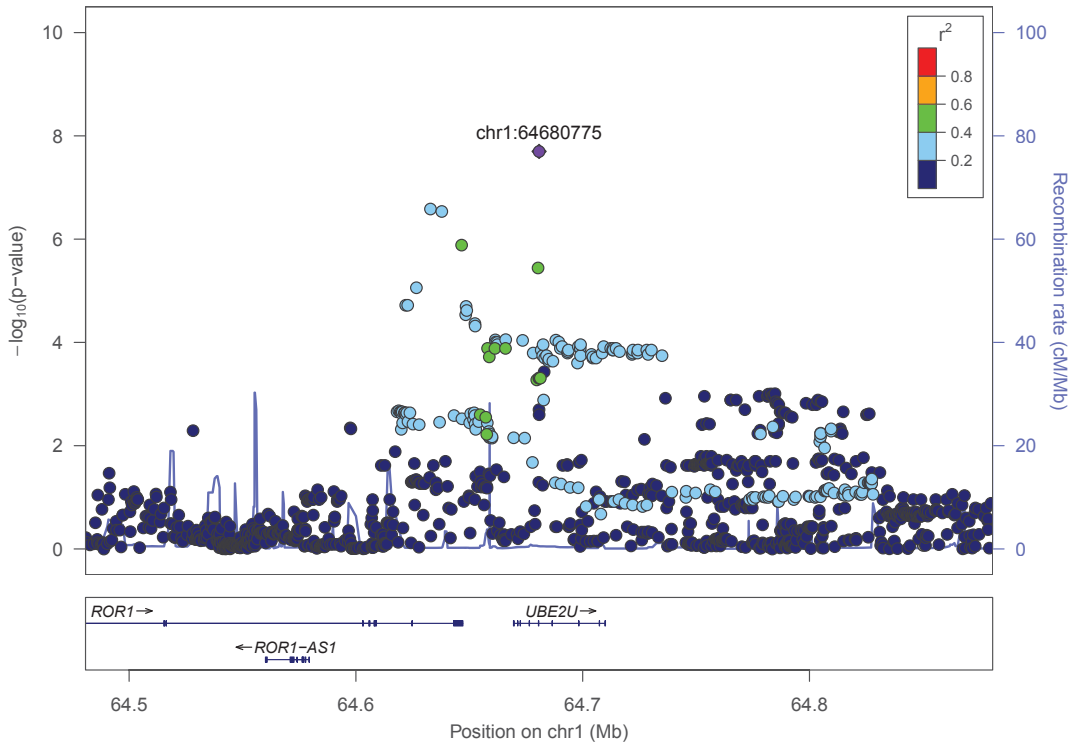


Figure 5.4: Locuszoom plot (Pruim et al., 2010) of association on chromosome 1 with unfavourable outcome, which is a zoom of the Manhattan plot on the locus. The lead SNP is a purple diamond, other markers are circles coloured by their r^2 with the lead SNP to show LD. The bottom panel shows annotated genes in the region, with exons as boxes and introns as lines. Recombination rate in cM/Mb is plotted as a pale blue line.

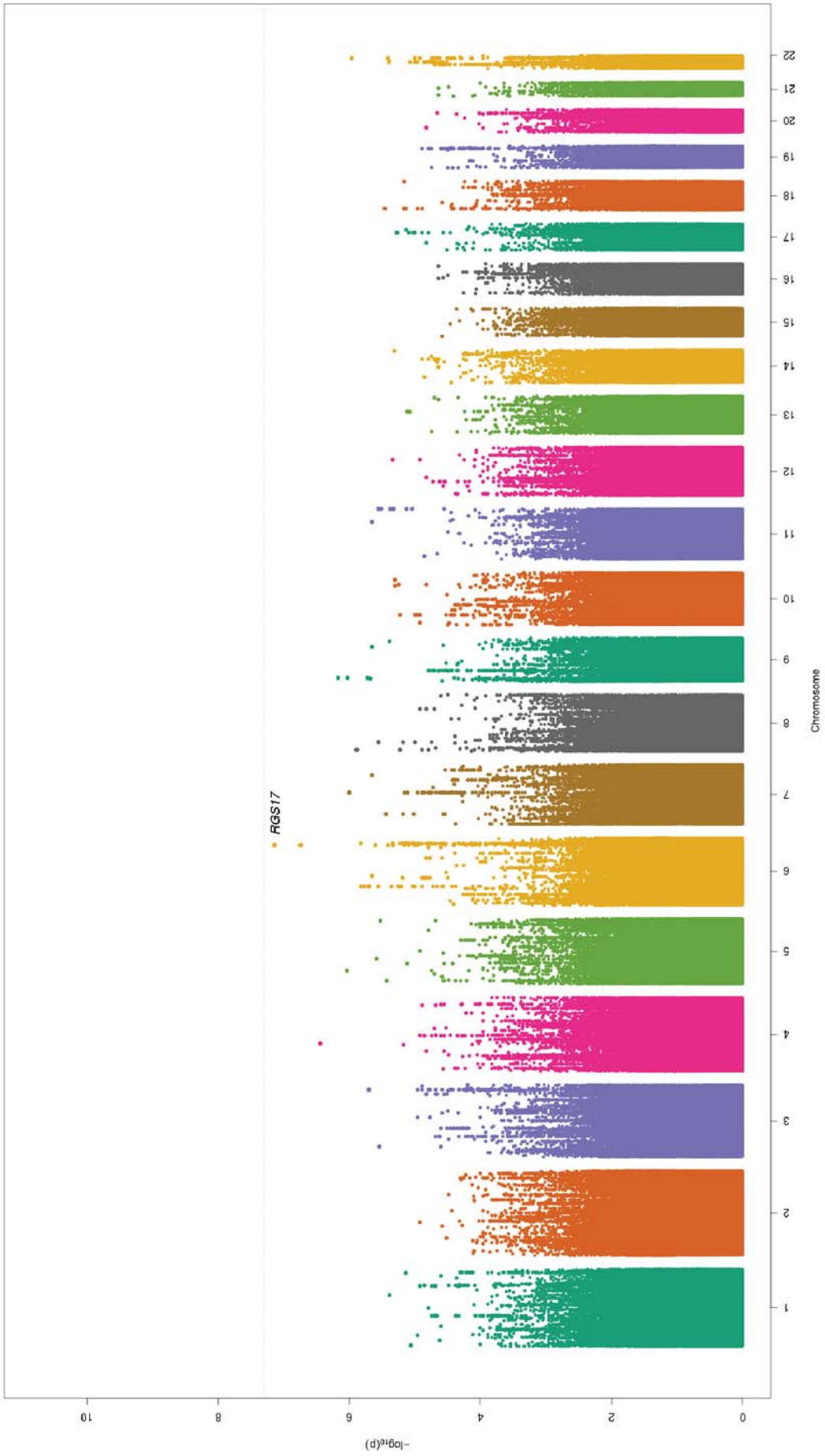


Figure 5.5: Manhattan plot from GWAS of all Dutch meningitis cases. Y-axis is $-\log_{10}(p)$ of the association values, x-axis is ordered by marker position on each autosome. Suggestive results from table 5.3 are annotated with nearby genes. Genome-wide significance is at 5×10^{-8}.

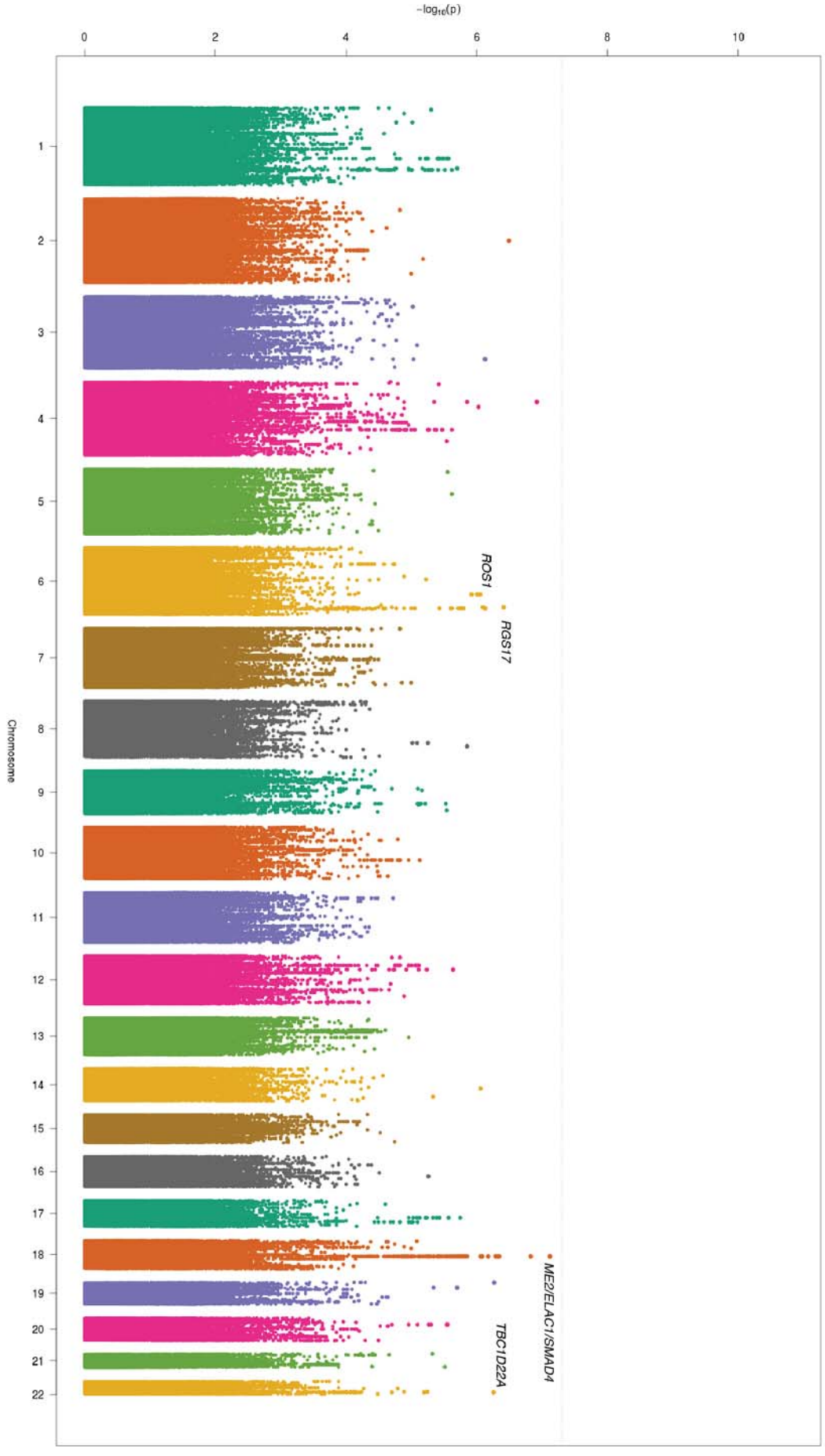


Figure 5.6: Manhattan plot from GWAS of Dutch pneumococcal meningitis cases. Y-axis is $-\log_{10}(p)$ of the association values, x-axis is ordered by marker position on each autosome. Suggestive results from table 5.3 are annotated with nearby genes. Genome-wide significance is at 5×10^{-8}.

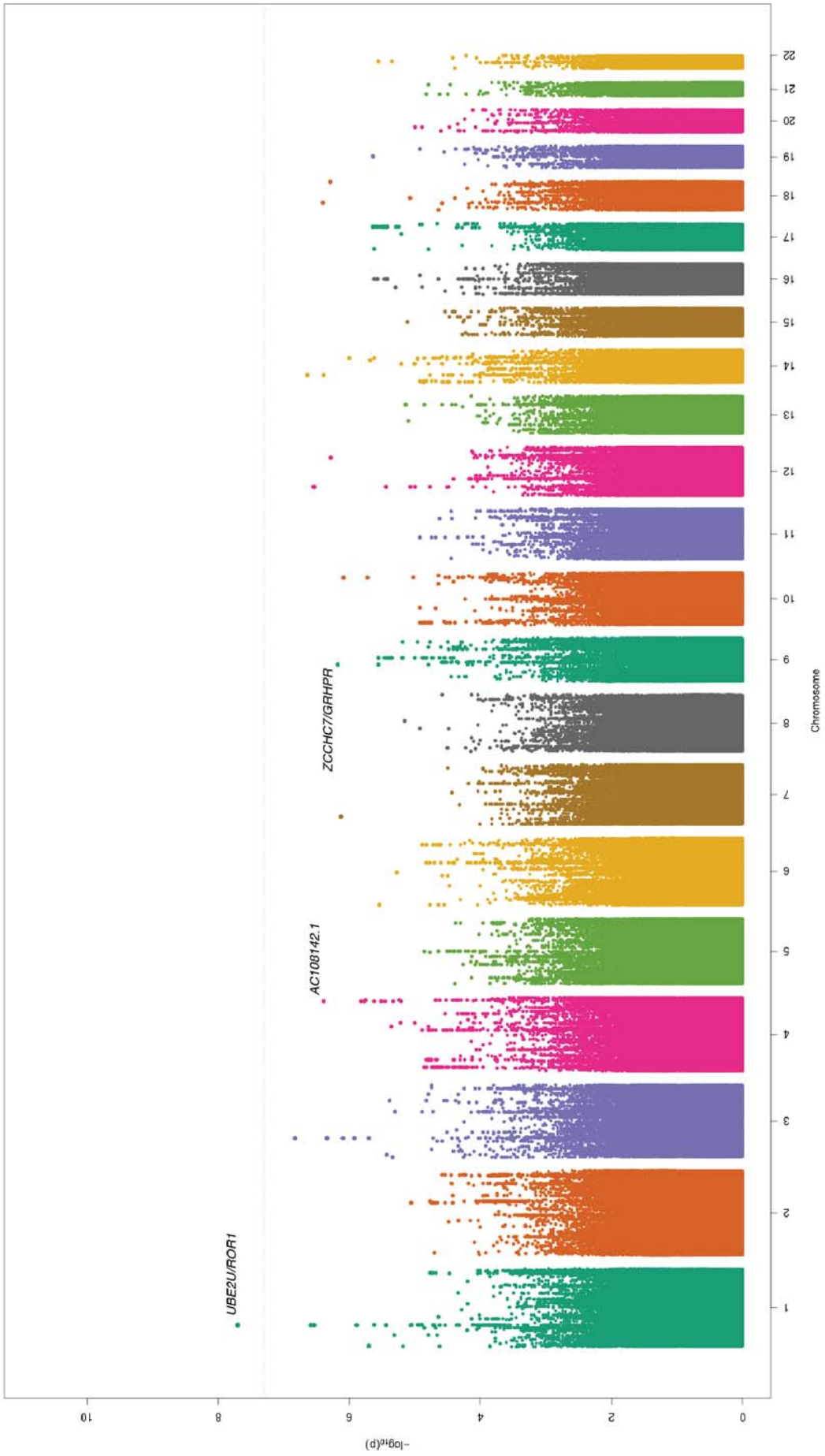


Figure 5.7: Manhattan plot from GWAS of all Dutch meningitis cases with an unfavourable outcome. Y-axis is $-\log_{10}(p)$ of the association values, x-axis is ordered by marker position on each autosome. Suggestive results from table 5.3 are annotated with nearby genes. Genome-wide significance is at 5×10^{-8}.

Danish cohort results

Once again analysis of all invasive pneumococcal disease, pneumococcal meningitis and pneumococcal bacteremia suggested a heritable component to each of these phenotypes (table 5.4), with estimates consistent with the Dutch study (although with wider confidence intervals, due to the smaller number of samples). Subtype did not provide any evidence that bacteremia and meningitis are genetically distinct phenotypes (PLR = 311; $p = 0.60$), as associations between the phenotypes followed a similar profile. No genome-wide significant associations were found for either pneumococcal meningitis or pneumococcal bacteremia (figs. 5.8 and 5.9). The only suggestive association (MAF $> 5\%$ and $p < 1 \times 10^{-6}$) was found on chromosome 14 at 67181537 (MAF = 0.14; OR = 0.45; p = 2.2×10^{-7}) in an intron of *GPHN*.

Phenotype	Method	Heritability	Error	Fit p-value
Invasive pneumococcal disease	GCTA	0.259	0.081	1.3×10^{-5}
	LDAK	0.285	0.092	8.5×10^{-4}
Pneumococcal meningitis	GCTA	0.727	0.451	5.1×10^{-7}
	LDAK	0.849	0.569	7.3×10^{-2}
Pneumococcal bacteremia	GCTA	0.371	0.098	1.4×10^{-5}
	LDAK	0.575	0.113	2.1×10^{-7}

Table 5.4: Human SNP heritability (h^2_{SNP}) of three pneumococcal phenotypes in Danish children cohort, as in table 5.2. Pneumococcal meningitis and bacteremia are subsets of the overall category of invasive disease.

5.2.3 Meta-analysis of four studies

An important step in GWAS is to confirm the results using an independent study population. As well as avoiding possible batch effects from a single cohort, this also increases sample size and power at true associations with an OR/MAF too low to find in the initial study. Here I did this analysis for meningitis susceptibility, which had the most total samples available. I used the summary statistics (p-value and β) that I generated from the Dutch and Danish cohorts, as well as summary statistics I received from 23andme and GenOSept (table 5.1).

I performed the meta-analysis between these studies using METAL (Willer et al., 2010). At each site the beta values (effect sizes and direction) and p-values from each study are converted into z-scores, which are then combined as a weighted sum with the weights given by the number of samples N in each study. This combined z-score gives the meta-analysis p-value. Before doing this I made sure all marker positions and alleles were given with respect to the same reference, as the direction of effect is crucial. For the association

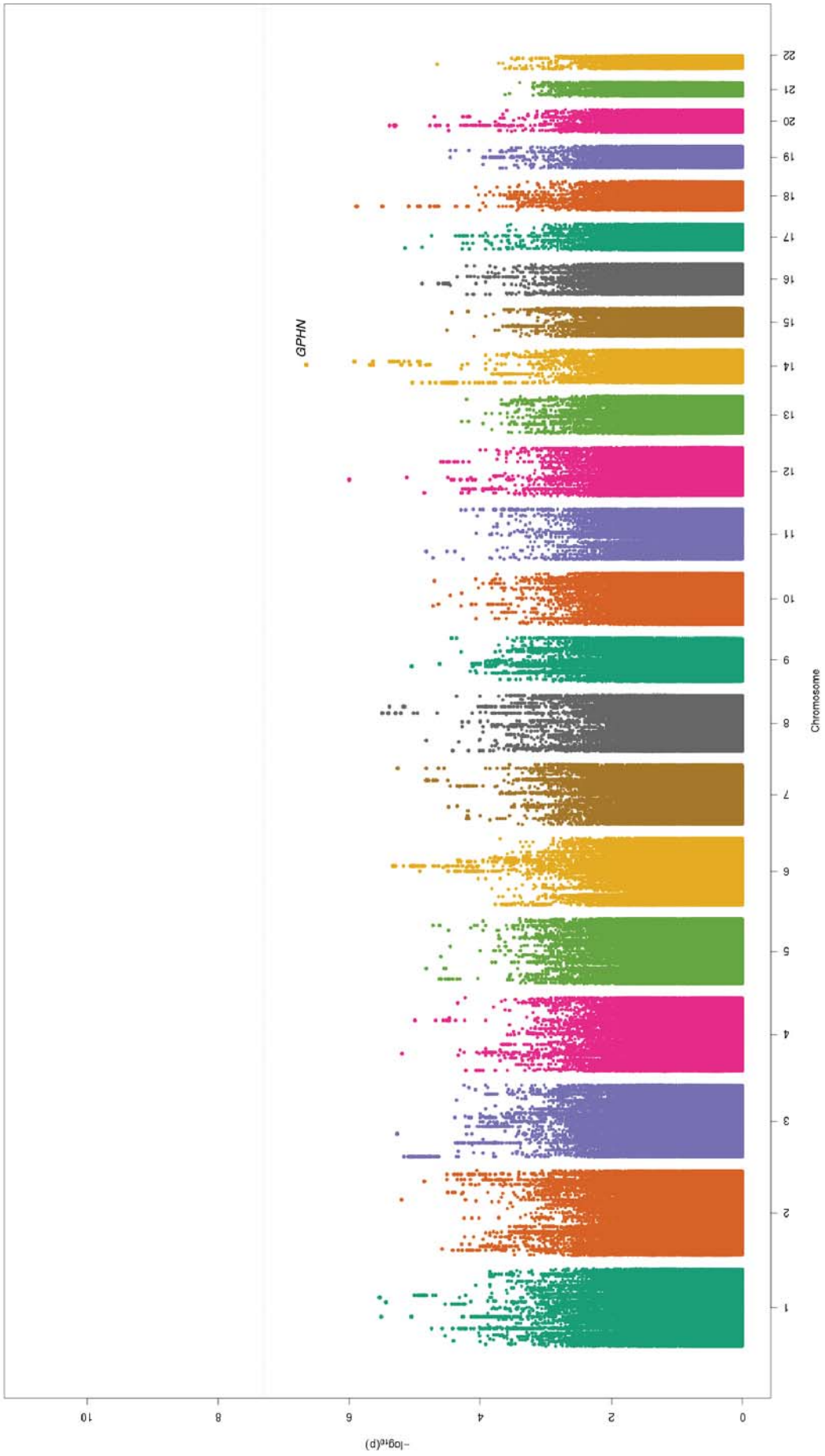


Figure 5.8: Manhattan plot from GWAS of Danish pneumococcal meningitis cases. Y-axis is $-\log_{10}(p)$ of the association values, x-axis is ordered by marker position on each autosome. The suggestive results is annotated with nearby genes. Genome-wide significance is at 5×10^{-8}.

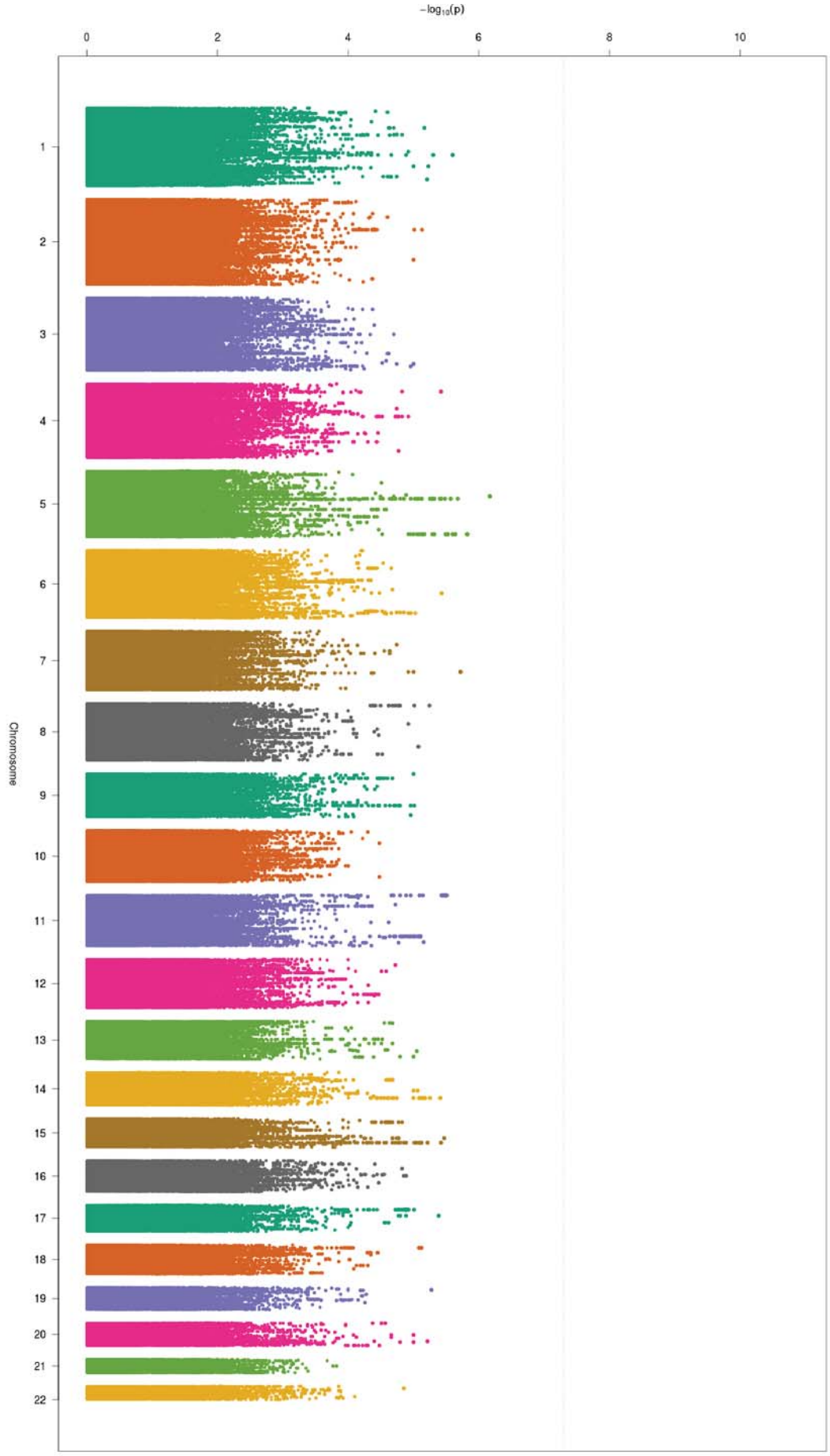


Figure 5.9: Manhattan plot from GWAS of all Danish pneumococcal bacteremia cases. Y-axis is $-\log_{10}(p)$ of the association values, x-axis is ordered by marker position on each autosome. Genome-wide significance is at 5×10^{-8}.

studies I performed using bolt-lmm I adjusted the beta values using the formula

$$\beta_{\text{adjusted}} = \beta \cdot \frac{1}{\pi \cdot (1-\pi)}$$
$$\text{where } \pi = \frac{N_{\text{cases}}}{N_{\text{cases}} + N_{\text{controls}}}$$

As the weight N for each study I used the effective sample size

$$N_{\text{eff}} = \frac{4}{\frac{1}{N_{cases}} + \frac{1}{N_{controls}}}$$

rather than the total number of samples, as some of the studies were highly biased to a larger number of controls (for example 23andme used 842 samples and 82 778 controls). I only included markers that had summary statistics from all studies in the meta-analysis (M = 5 627 710), to avoid effects of sample size heterogeneity in the final p-values.

Figure 5.10 shows the results of the meta-analysis genome-wide. No sites were significant in this analysis, and the additional data did not support the genome-wide significant hit in an intron of *CA10* reported by 23andme (Tian et al., 2016). A possible reason for these observations is due to heterogeneity of phenotype between the cohorts in the meta-analysis. The simple method used here assumes that sites must have the same direction of effect, and are independent observations of significance, and are on the same phenotype with no measurement error. However, the Dutch and Danish cohorts differ in that they analyse adult and childhood meningitis respectively, which differ in their immune system competence and their vaccination status (Imöhl et al., 2010; Rodrigo et al., 2014). GenOSept includes bacteremia cases, which may be different from meningitis specifically. Finally, 23andme uses self-reported status of bacterial meningitis. While self-reported data has generally been shown to be as good as hospital diagnoses for phenotype association, especially given the increased number of cases available, for difficult to diagnose infectious diseases such as lupus this has been shown not to be the case (Tian et al., 2016; Cortes et al., 2017). For bacterial meningitis cases have not been culture-proven, and may well be viral meningitis or not meningitis at all. If they are meningitis, most likely a wider range of pathogens compared to the other cohorts have been included.

A future analysis will include association statistics calculated from the UK biobank, which has a large collection of genotyped samples (N = 500 000) and hospital diagnoses for bacterial and pneumococcal meningitis. This may help to provide extra samples with a well-defined phenotype. Alternatively, modelling the heterogeneity in phenotype may help, though sample size is still likely to be a limiting factor.

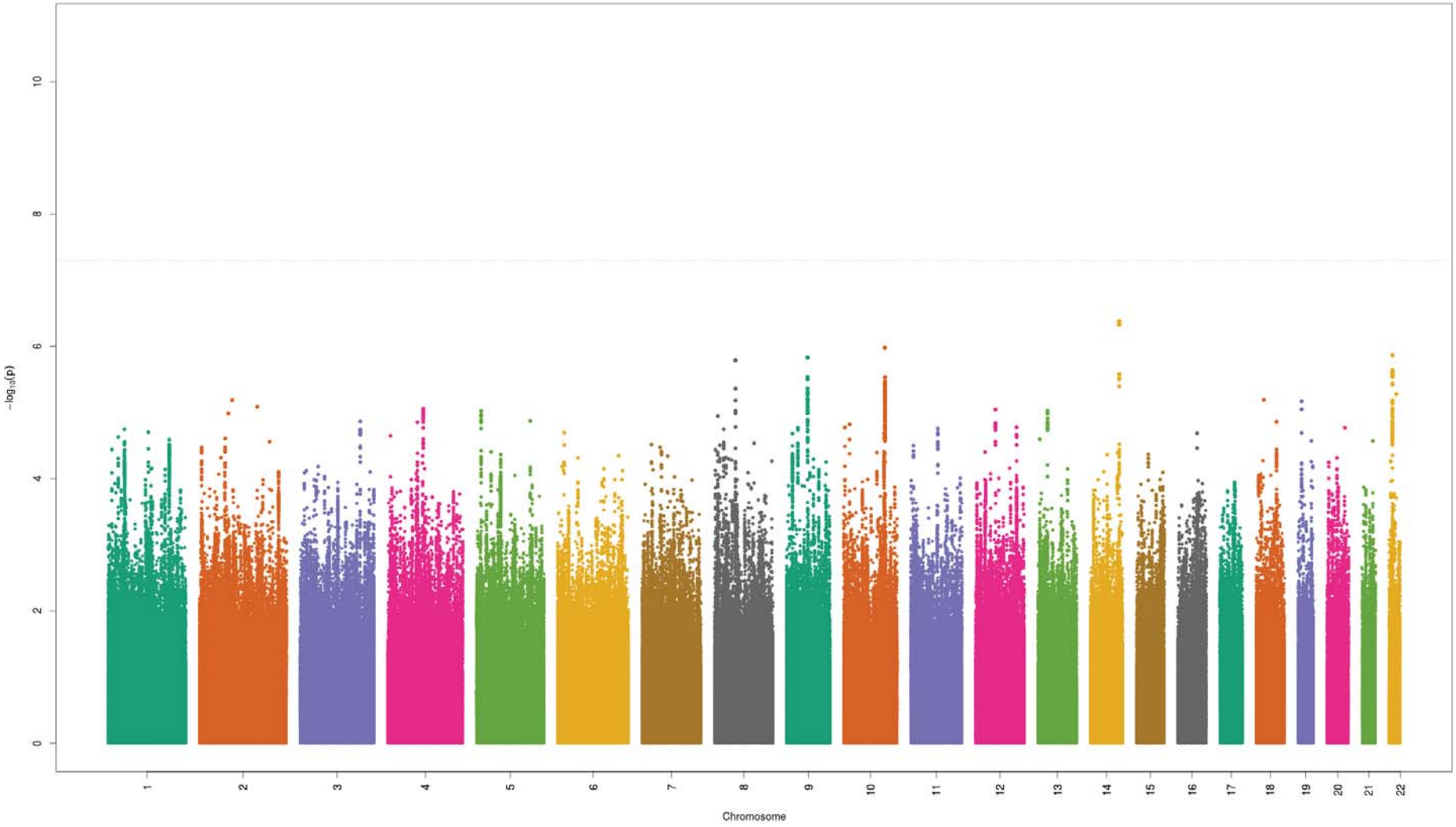


Figure 5.10: Manhattan plot from meta-analysis of meningitis susceptibility, including samples from the Dutch (adult meningitis cases) and Danish cohorts (child pneumococcal meningitis cases), 23andme (self-reported bacterial meningitis) and GenOSept (adult pneumococcal bacteremia). Y-axis is $-\log_{10}(p)$ of the association values, x-axis is ordered by marker position on each autosome. Genome-wide significance is at 5×10^{-8}.

5.3 Genome-to-genome analysis of host and pathogen variation

In this final section I aim to bring together data from chapter 4 and section 5.2 to search for genome-to-genome associations between the host and pathogen in cases of bacterial meningitis. By linking the datasets from the human and pathogen arm of the Meningene study and performing an association study between pairs of variants drawn from each genome over all these samples, I tested the hypothesis that certain bacteria are more likely to cause invasive disease in specific host genotypic backgrounds. This dataset is unique, and to the best of my knowledge the first time both host and pathogen have been sequenced for a bacterial infection. The present analysis does not require a phenotype, an advantage of such epistasis analyses (Skwark et al., 2017).

In viral infections, two previous analyses have been published attempting this analysis. Bartha et al. (2013) used host genotype and the infecting viral genome from 1 071 HIV patients to perform a logistic regression between every human SNP (of which there were ~ 7 million) and every viral amino acid (of which there were 3 000) while using the first two principal components to correct for viral population structure. The authors recapitulated the well known association with viral load and HLA allele, but were unable to find any new genome-to-genome links. They estimated having 80% power to detect a variant with MAF of 10% with an OR of 4.2 given their sample size and the number of tests being performed.

Azim Ansari et al. (2017) performed a similar analysis on 542 cases of hepatitis C infection. Again using imputed human genotypes and viral amino acids they performed a logistic regression between variants, using the first three principal components to control for human population structure, and the first ten to control for viral population structure. As well as finding expected associations with the HLA, they found a region of the viral genome associated with variability in *IFNL4*, though not quite reaching significance. However, the same human SNPs were found to be associated with viral load, for which the authors were able to conclude a link between the strength of selection acting on the viral population due to the *IFNL4* response, and the resulting fitness of circulating virions.

I wished to first remain agnostic to annotation or previous knowledge of host-pathogen interactions to attempt to uncover previously unknown genome to genome links in clinical cases of bacterial meningitis, following a similar design to the two viral studies. To do this, in section 5.3.1 I performed an association test between every genotyped human SNP and every bacterial mapped SNP/INDEL. However, given the small sample size and the large amount of variation between the two genomes, the power to overcome the multiple testing was very low for even moderate effect sizes. I therefore used unsupervised clustering techniques which use the correlation structure present in the bacterial population

to produce a lower dimensional representation of the bacterial genomic variation, lowering the multiple testing burden (section 5.3.2).

Finally I wished to test whether variation in host and pathogen protein which are well known to interact with each other is correlated in cases of disease (section 5.3.3). I used the detailed antigen calling already performed in section 4.3.1 as the bacterial variants, and tested for correlation with human variation. As these bacterial proteins are known to be broadly antigenic (Croucher et al., 2017), I tested not only the specific human gene involved in the interactions, but every imputed human variant to try and identify potential new interaction partners.

In the tests below I used the 460 samples which passed the QC filters from both sections 4.2 and 5.2.1. When performing associations on a sub-phenotype, as in splitting these samples into two based on cluster or antigen membership and testing human SNPs against this, I only tested those sub-phenotypes which contained at least 5% of samples. This avoided spurious results from testing rare (and underpowered) variants resulting from partitioning lower frequency variants into yet lower frequency phenotypes.

5.3.1 All by all variant association

To perform a correlation analysis between 7×10^6 imputed human variants and 1×10^5 requires around 10^{12} association tests, which even given the availability of a large number of CPU cores and the embarrassingly parallel nature of the problem is computationally challenging.

To approach this problem, I modified the SEER C++ code from chapter 2 to perform the association tests, as I had already optimised it for speed. I converted the VCF files with the human and pathogen variant calls to comma separated values (CSV) files, coding the human calls as 0, 1 or 2 based on the number of copies of the minor allele the genotype contained (the additive model). These CSV files then only contained the genotypes, and I stored site and sample level metadata in separate files – this separation allows much quicker processing of genotype data, especially when accessing specific chunks (Ganna et al., 2016). I extended the χ^2 test to a 3x2 table, and added efficient code for a 3x2 Fisher's exact test (https://github.com/chrchang/stats) which I applied when the assumptions of the χ^2 test were violated (by small expected values in the table counts, when MAF in either genome was low). I used a filter of $p < 5 \times 10^{-11}$ for this uncorrected test, which is equivalent to a Bonferroni correction with a significance level of $\alpha = 1$. I then tested the pairs of variants which passed this filter with a logistic regression, using the human SNP and the first three components of the bacterial MDS projection as the design matrix $\boldsymbol{X}$ and the bacterial variant as the response vector $\boldsymbol{y}$.

To parallelise the code I used 300 independent jobs. Each job first read in all the bacterial variants from the CSV file, and parsed these into a matrix stored in main memory.

The null log-likelihood for the logistic regression was calculated for each at this point, to avoid having to make this calculation multiple times when the same bacterial variant was tested against every human SNP. The chunk of human SNPs assigned to the job were then read in, and each one passing filtering was tested for association with every bacterial variant.

As the number of imputed human SNPs was still prohibitively large, I tested the genotyped human variants only. This is similar to testing an LD pruned subset of sites with the advantage that their genotype calls could be further investigated if they proved significant. Using this approach I tested 623 649 human SNPs for correlation with 113 059 associated bacterial variants (SNPs and INDELs from section 4.3). 1.8×10^{10} variant pairs passed filters of MAF $> 5\%$ in both human and bacterial population with $< 5\%$ of calls missing. Using 300 jobs the total computation time was 268 hrs, using 600Mb memory per job. 2 433 variant pairs passed the initial p-value filter for $p < 1$ when adjusting for multiple testing, but none of these were significantly associated at $p < 0.05$ when tested adjusting for bacterial population structure.

Due to the high multiple testing burden from the large number of variant pairs being tested, this number of samples would only detect strong correlations between genomic variants. This is plotted in fig. 5.11: assuming a MAF of 25% in each population, the sample size of 460 has 80% power to detect an epistatic effect with an odds ratio of 4. While bacterial population structure is less likely to be an issue for this analysis, it may still reduce the power to fine-map specific interactions. To find whether interactions exist at lower coupling strengths it would help to have more samples, as at sample sizes double this study the discovery power increases sharply. The number of samples is also currently too small to do a heritability analysis of the interaction effect.

While sample size fundamentally limits this analysis, there are some further steps to be taken. Firstly, the use of Direct Coupling Analysis has been shown to have greater power at detecting epistatic interactions in the *S. pneumoniae* genome than the simple χ^2 tests I have used (Skwark et al., 2017). However, an implementation of this which will scale to the size of the present problem does not exist. Instead, in subsequent sections I use a representation of the pathogen genome in a lower number of dimensions to attempt to reduce the multiple testing burden.

5.3.2 Reduced representation of pathogen genome

Given the difficulties encountered when testing every human variant against every bacterial variant, I wished to find a way to reduce the dimensionality of the problem. This problem is well known in eQTL studies, where both transcriptomic and genomic variation is measured, and an association is performed between the genetic variation and altered gene expression (Breitling et al., 2008; L. Franke & Jansen, 2009). One approach is to model the per-gene

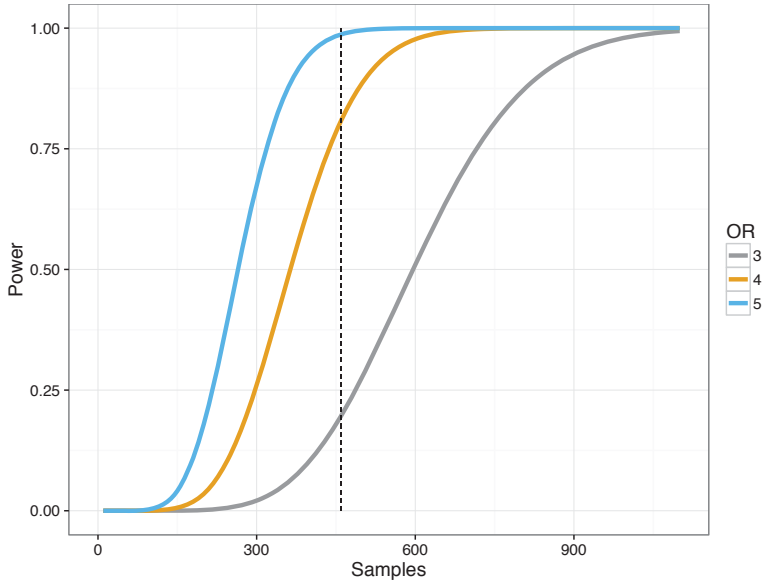


Figure 5.11: Power for detecting genome-to-genome interactions. Assuming no population structure effect, the power of detecting an correlation between genome positions at 25 % MAF at a range of ORs. The 460 samples I was able to use in this study is marked as a vertical dashed line.

levels of transcript variation as a smaller number of latent variables, each of which affect a number of transcripts. The simplest way to do this would be by PCA which would use the linear combinations of transcripts explaining most of the variation as the latent variables, though more sophisticated methods exist (Marttinen et al., 2013; Gillberg et al., 2016). In the present analogy, transcript variation corresponds to bacterial sequence variation, and the latent variables may combine these into features such as sequence type, serotype or antibiogram type.

A method which has been successfully used for this purpose is probabilistic estimation of expression residuals (PEER), which estimates latent factors and their per sample weighting from high dimensional input (Stegle et al., 2012). PEER's advantages over PCA are that: the latent factors estimated from the data do not have to be orthogonal, which may not always be biologically realistic; covariates can be included in the model fit such as batch effects or case/control status; the factors can be controlled to not be parallel with other known influences, for example serotype or sequence type.

I therefore ran PEER, learning 40 unobserved factors (though this is an unimportant setting, as automatic relevance determination is used to determine this from the data). The results are shown in fig. 5.12 – the first few factors can be seen to represent the large scale population structure, and some later factors represent finer scale population structure. I performed an association with all imputed variants against all the inferred factors, which gave uninflated results for the first twelve factors. Further factors gave spurious results at lower frequency variants.

While the PEER factors can be interpreted by the looking at the weights assigned to each input variants for the associated factor, I found this difficult to link directly to a biological

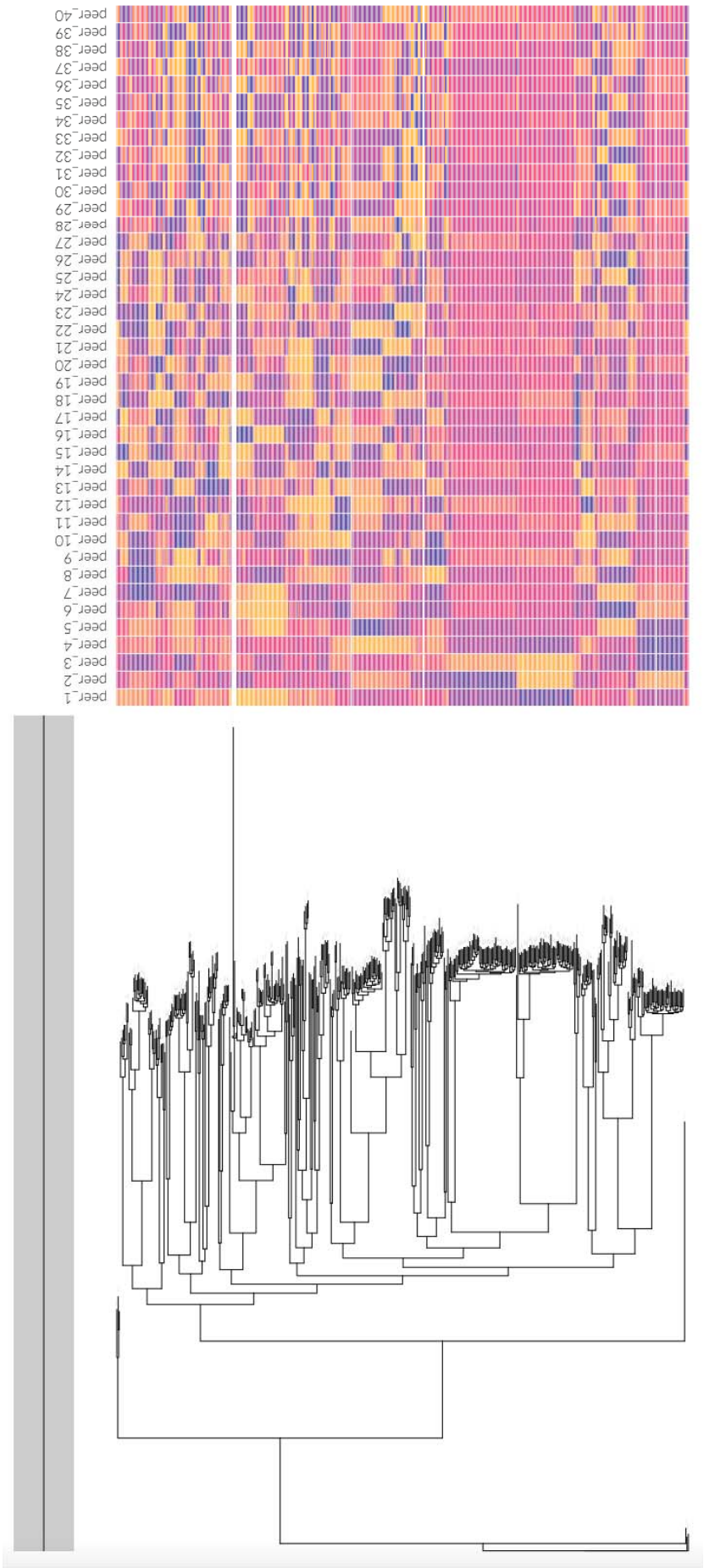


Figure 5.12: PEER factor analysis with $n = 40$ factors of the 460 *S. pneumoniae* genomes in the genome-to-genome analysis. The left panel shows the phylogeny drawn from the core-genome alignment with RAxML; the right panel shows the residual of each factor for the given sample on a continuous scale.

interpretation. Noting that the first components were describing population structure, I instead opted to instead test discrete population clusters for correlation with human variation as the interpretation of the bacterial variants was much more straightforward. This is essentially testing for lineage effects correlated with human variation, as the power to find locus effects is limited (as calculated above). I therefore created a core-genome alignment of these strains using `roary` as in section 4.3, and ran BAPS on this to generate population clusters. I found that the PEER components generally represented the same population structure as the BAPS clusters (fig. A.15).

Cluster	Serotype	Samples	Tested
1	4	17	-
2	-	145	✓
3	8/11A/33F	49	✓
4	10A/35F	22	-
5	23A/B/F	32	✓
6	6B	14	-
7	22F	39	✓
8	9N/15B/19A	47	✓
9	3	47	✓
10	7F	55	✓

Table 5.5: Number of samples in each population cluster. Cluster two is a polyphyletic 'bin' cluster. The dominant serotypes for each cluster, where they account for $> 50\%$ of the isolates, are listed.

Table 5.5 lists the ten clusters found in the data, and the dominant serotypes for each cluster. I ran an association with the BAPS clusters with at least 10% of samples in the subphenotype. The only result reaching genome-wide significance was an association between cluster eight (serotypes 9N/15B/19A) and variants on chromosome 10 (fig. 5.13). The lead variant is at position 134046136 on chromosome 10 (MAF = 0.27; OR = 4.28; p = 1.2×10^{-8}) located in an intron of *STK32C*, a serine/threonine kinase highly expressed in the brain. The high effect size estimated for the interaction is consistent with the power predicted in fig. 5.11.

5.3.3 Association of antigens

This section considers known interactions between host and pathogen proteins, and whether variation in the coding sequence or surrounding regions of each gene is correlated in cases of bacterial meningitis. *S. pneumoniae* has many virulence factors, some of which are known to interact with specific human proteins (Kadioglu et al., 2008). However, I was mostly interested in the interactions where the pneumococcal protein contains sequence variation, ideally somewhat independent of population structure. These regions have

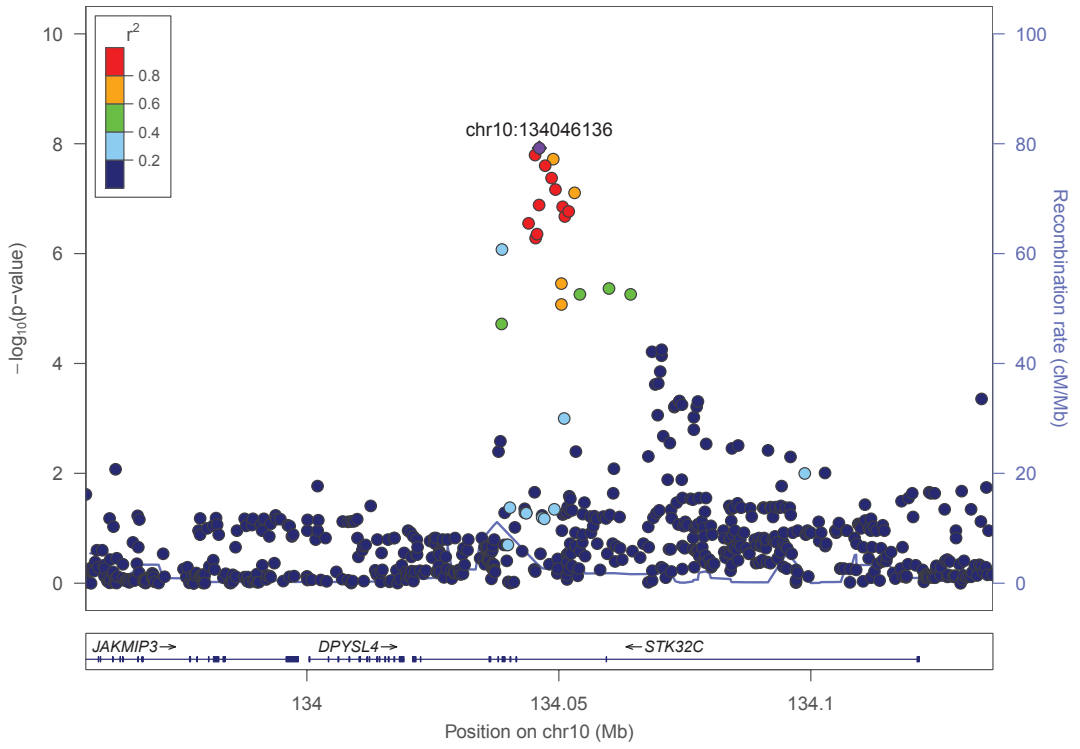


Figure 5.13: Locuszoom plot of association on chromosome 10 with sequence cluster 8, as in fig. 5.4.

higher power to be detected in an association analysis, and the higher rate of variation is potentially a sign of diversifying selection, which may mean the variation is more likely to be associated with specific interactions with the human immune system. Using a combined mapping and assembly approach, followed by a supervised machine learning, in section 4.3.1 I have classified the *pspA*, *pspC* and *zmpA* allele of every sample in the Meningene collection.

pspA is known to bind to C3b, preventing decomposition on the pneumococcal surface and blocking the complement pathway response to infection (Tu et al., 1999). The *LTF* gene encodes lactoferrin, an iron-binding protein found in the granules of neutrophils. This protein is bacteriocidal, and forms part of the innate immune response against pneumococci. It has been found that *pspA* binds lactoferrin to the surface of the pneumococcus, thus reducing their killing by this protein (Shaper et al., 2004).

Like *pspA*, *pspC* has been shown to bind C3 and prevent opsonic decomposition on the pneumococcal surface (Q. Cheng et al., 2000). In addition, some forms of *pspC* have been shown to bind factor H (Janulczyk et al., 2000; Dave et al., 2001). Factor H inhibits complement activation by preventing C3b degrading and activating the next step in the complement pathway. By binding this protein to the surface, the pneumococcus further prevents activation of C3. This locus in the human genome is also known to be involved in susceptibility to invasive meningococcal disease (Davila et al., 2010).

Finally I tested allelic variation of *zmpA*, which is a protease known to bind IgA

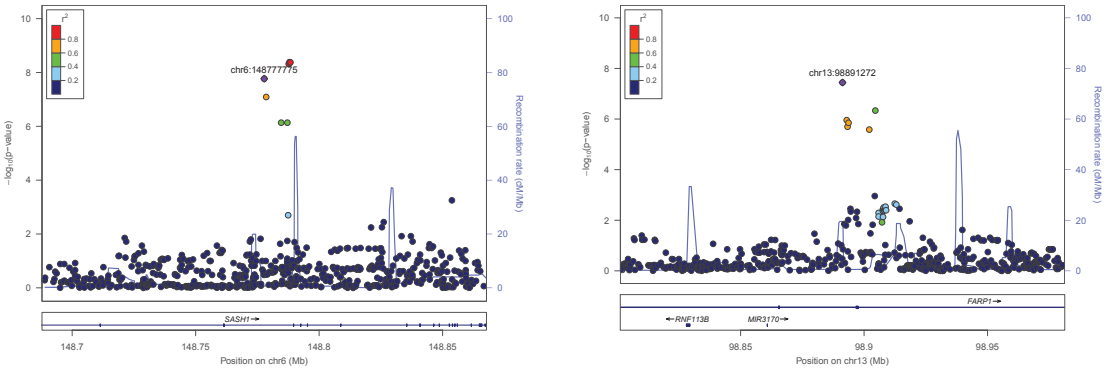
(Wani et al., 1996). This is the most abundant antibody in the nasopharynx, and is an important part of the immune response to pneumococcal infection (Cerutti & Rescigno, 2008). However, it is not produced by simple translation from a single gene and instead involves a pathway covering the HLA along with other regions of the genome (Fagarasan & Honjo, 2003; Ferreira et al., 2010).

For all of the antigen alleles with enough observations (fig. 5.14) I performed an association against all imputed human variants as in section 5.2.2. I used a more accurate imputation of the *CFH* region due to its potential relevance in these interactions. For each test I produced a genome-wide Manhattan plot, and a locuszoom plot for the known interaction partner.

Antigen	Allele	Samples	Tested
pspA	1	214	✓
	2	231	✓
	3	1	-
	4	1	-
cbpA	0	44	✓
	1	6	-
	2	17	-
	3	84	✓
	4	45	✓
	5	60	✓
	6	191	✓
pspC	0	347	✓
	7	7	-
	8	39	✓
	9	45	✓
	10	6	-
	11	3	-
zmpA	1	26	-
	2	236	✓
	3	185	✓

Figure 5.14: Antigen classification of *pspA*, *pspC* and *zmpA*. The total number of samples in the genome-to-genome analysis with each allele is shown, and those where an association test performed are noted.

None of the bacterial antigen alleles were significantly correlated with variants in their human interacting counterparts at the suggestive level ($p < 10^{-5}$). However, there were two associations of *pspC* allele reaching genome-wide significance elsewhere in the genome. Figure 5.15 shows a locuszoom plot of each of these associations. The first is between *pspC*-8 and position 148788006 on chromosome 6 (MAF = 0.08; OR = 9.20; p = 4.1×10^{-9}). This is in *SASH1*, which has previously been found to have decreased expression during meningococcal meningitis (https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-11755/) . The second is between *pspC*-9 and position 98891272 on chromosome 13 (MAF = 0.16; OR = 6.30; p = 3.6×10^{-8}), in *FARP1*.



(a) *pspC* allele 8, and chromosome 6 (*SASH1*) **(b)** *pspC* allele 9, and chromosome 13 (*FARP1*)

Figure 5.15: Locuszoom plot of association between *pspC* allele and imputed human SNPs, as in fig. 5.4.

5.4 Conclusions

This chapter has considered the effect of host variation on susceptibility to and severity of pneumococcal meningitis. By using two relatively large well-phenotyped cohorts from the Netherlands and Denmark, I have estimated h^2_{SNP} to be around 30-40% for susceptibility, and around 25% for severity. This suggests that human genetics plays a role in determining how likely invasive disease is, given that a bacteria which is capable of invasion has colonised the individual (chapter 4). Additionally, I have shown that host genetics explains some of the variability in disease outcome after invasion has happened, which may occur by variation in immune response.

I then attempted to use GWAS to find specific variants which contribute to these traits, and while I found signals reaching significance in the Dutch population, none of them have replicated when meta-analysed with summary statistics from other similar studies. No data from other studies is currently available associating human variation with disease outcome, so any planned future confirmation of the association with *UBE2U* may have to use an *in vivo* model of pneumococcal meningitis.

It is difficult to collect bacterial meningitis cases due to: 1) their rarity, and 2) the difficulty to confirm the causative organism by culture. It is even more difficult to determine which of those cases resulted in a poor clinical outcome, as this requires a study design with patient follow-up potentially months after discharge from hospital. The number of cases collected by the collaborators for this analysis is impressive, and this has allowed the first heritability estimates of these traits to be made. These estimates suggest that continuation of the Meningene cohort is warranted, as is the meta-analysis with other well phenotyped studies. With enough cases, specific associations replicating in multiple cohorts will be found. The attempt at meta-analysis I performed here did not find any hits, perhaps due to heterogeneity of phenotype between cohorts. Additionally, a previously reported association in an intron of *CA10* could not be confirmed.

The only previously known genetic association with meningitis is the *CFH* region, which the minor allele is protective for susceptibility to invasive meningococcal disease in children (Davila et al., 2010). I did not find this association with pneumococcal meningitis, though when I restricted analysis to adult meningococcal cases, meta-analysis with the Dutch cohort did not refute its existence. This may suggest a difference in the host response based on invading pathogen, with *CFH* binding being less important for pneumococcal infection.

In the genome-to-genome analysis I was able to put a limit on the strength of interactions that could be detected. Despite being underpowered given the large combined complexity of the host and pathogen populations, I was able to find possible correlations between lineage and host variants. Additionally, some antigen alleles showed possible correlation with variants in the host, though not in regions they are known to interact with. The lack of association may point at the variability of antigen binding of host-proteins being uninvolved in disease course, or may just be limited by a combination of small sample size and high antigenic diversity. The other possible hits from this single study will need replication before biological meaning can confidently be inferred, but the method here shows how such analysis might be done for a bacterial infection, and the results can be used in any future meta-analysis with similar studies.