Chapter 6: Genome-wide association study identifies single nucleotide polymorphic changes associated with beta-lactam resistance

Introduction

Results

6.1 Identification of loci associated with beta-lactam non-susceptibility

- 6.1.1 Quality control and randomisation test
- 6.1.2 Population stratification
- 6.1.3 Case-control association analysis
- 6.1.4 Linkage disequilibrium
  - 6.1.4.1 Defining linkage blocks
  - 6.1.4.2 Larger linkage blocks observed in smaller dataset
- 6.1.5 Estimating the amount of the non-susceptible phenotype explained by co-detected loci in Maela and Massachusetts populations

6.2 Biological relevance of candidate loci

- 6.2.1 Candidate loci in genes participating in the peptidoglycan biosynthesis pathway
- 6.2.2 Candidate loci in genes outside the peptidoglycan biosynthesis pathway
- 6.2.3 Candidate loci in genes conferring resistance to other antibiotics

6.3 Beta-lactam specificity of resistance mutations

6.4 Distribution of candidate alleles in the Maela and Massachusetts populations

Conclusion

Declaration of work contributions:

Dr Nicholas Croucher kindly provided sequences for Massachusetts data.

6. Genome-wide association study identifies single nucleotide polymorphic changes associated with beta-lactam resistance

6.1 Introduction and aims

The previous chapter highlighted an increase in beta-lactam resistance and the mechanisms mediating the spread of resistance alleles in one local community. At the global level, a rise in beta-lactam resistant pneumococci over the recent decades has raised significant concerns (Potgieter, Carmichael *et al.* 1992, Carratalà, Alcaide *et al.* 1995, Donern amd Ferraro *et al.* 1996, and more recently WHO 2014). There have been tremendous efforts in identifying the genetic sources of resistance between 1980s-2000s. Comparative genomics between beta-lactam resistant and susceptible strains have identified highly variable regions in genes coding for penicillin-binding proteins (*pbp*) in resistant isolates; thereby highlighting *pbp* genes as key contributors to the resistance phenotype (Laible, Spratt *et al.* 1989, Dowson, Hutchison *et al.* 1989). Aside from determining nucleotide and amino acid alterations observed in beta-lactam resistant isolates, site-directed mutagenesis has been a useful tool in identifying mutations that give rise to resistant profiles (Laible and Hakenbeck 1991, Hakenbeck, Martin *et al.* 1994, Sifaoui, Kitzis *et al.* 1996, Smith and Klugman 1998 & 2003). Many of these studies were supported by further structural characterisations, revealing a change in the structure of penicillin binding proteins which lead to resistant phenotype (Gordon, Mouz *et al.* 2000, Dessen, Mouz *et al.* 2001, Job, Di Guilmi *et al.* 2003, Contreras-Martel, Job *et al.* 2006)

Identification of specific variants that might be associated with beta-lactam resistance can also be performed using a technique called genome-wide association study (GWAS). Although GWAS have been common in human genetics for many years (McCarthy, Abecasis *et al.* 2008), it was, at the time of this study, largely untried in bacteria because of the limited sample sizes and the confounding effects of bacterial clonal population structure. The clonal population structure may be less problematic in pneumococci as homologous recombination brings genetic admixture into pneumococcal populations in a manner akin to sexual reproduction in humans. Unlike human recombination, this does not occur every generation and only affects a small

part of the genome in each occurrence. On average, recombination in *S. pneumoniae* involves 2.3 kb of chromosomal DNA (Croucher, Harris *et al.* 2012), about twice the size of an average pneumococcal gene. This suggests that large numbers of recombination events must accumulate in order to disrupt the clonal structure, or break up haplotype blocks smaller than this size. Also, on average, one recombination event occurs every one to nine years, depending on the lineages. Therefore, to observe sufficient numbers of recombination events to disrupt clonal structure, a diverse and large population size would be required. The large size of the species-wide samples used in this study (3,071 pneumococcal genomes), and the highly recombinogenic nature of *S. pneumoniae* had the potential to allow sufficient resolution to be achieved and help refine the genetic determinants of resistance from large recombination fragments, previously described as mosaic genes, to discrete causative sites or smaller linkage blocks.

This gave an opportunity to investigate the molecular mechanisms of resistance down to single polymorphic changes. Ultimately, an increase in the resolution of detection and improved insights into the biology of resistance mechanisms for beta-lactams might contribute to the foundation for future application of genome sequencing in predicting antibiotic sensitivity in clinical settings and surveillance studies (Levine, O'Brien *et al.* 2012, Goldblatt, Ramakrishnan *et al.* 2013, GPS 2013).

This chapter aimed at:
i) Identifying genetic variants associated with beta-lactam non-susceptibility.
ii) Determining differential association of variants identified in i) to particular class of beta-lactams.
iii) Determining the distribution of resistant variants in different pneumococcal lineages including vaccine, and non-vaccine targeted lineages.

6.2 Methods

6.2.1 Subject populations

Two currently largest datasets for which whole genome sequences and beta-lactam susceptibility phenotype were available – Maela (3,085 isolates, also the subject of this thesis) and Massachusetts (616 isolates, Croucher, Finkelstein *et al.* 2013) - were employed in this analysis and the results were used to cross-validate each other.

6.2.2 Genotype callings and quality control

Bases were called from mapped sequences as discussed in 2.5, resulting in 392,524 and 198,248 SNP calls from the Maela and Massachusetts data respectively. This haploid bacterial information was handled as human mitochondrial sequence in PLINK v. 1.07 (Purcell, Neale *et al.* 2007). Since many thousands of genotypes are generated, a small genotyping error can lead to spurious GWAS results. Quality control thus is a critical step in performing GWAS. Here, minor allele frequency, which represents very low frequency alleles that likely reflect genotyping errors and proportion of missing genotypes per strain (genotype missingness rate) were estimated. Variants with minor allele frequency < 0.01, missingness by strain > 0.1 and missingness by variants > 0.1 were excluded from the analysis. For each site, the top two most common variants were parsed to the next analysis to reduce complexity in the test statistic.

6.2.3 Phenotype information

Beta-lactam susceptibilities were determined in both datasets by disk diffusion following the CLSI 2008 guidelines (CLSI 2008), generating 1,501 non-susceptible, 1,568 susceptible and 16 unknown phenotypes in the Maela data; 228 non-susceptible, 383 susceptible and 5 unknown phenotypes from the Massachusetts data. The minimum inhibitory concentrations (MIC) of non-susceptible isolates were confirmed by the E-test method as described in 2.1.1 and (Croucher, Finkelstein *et al.* 2013).

6.2.4 Determining the cut-off threshold

Randomisation tests were performed for both Maela and Massachusetts population to estimate the level of intrinsic noises generated by genetic variations alone as well as to determine a suitable cut-off for the analysis. 100 GWAS permutations were run with true genotypes but randomised binary phenotypes. With Bonferroni correction for multiple testing, there were no significant associations observed beyond the p-value 0.01. Therefore it was selected as a conservative threshold in the study reported here.

6.2.5 Case-control analysis

Variants associated with beta-lactam resistance were determined based on binary phenotypes: susceptible or non-susceptible conditioned on the pneumococcal population structure. BAPS clustering (generated in 2.6.2.2 and (Croucher, Finkelstein *et al.* 2013)) were used to represent population structure in the analysis. Based on known cluster information, the Cochran-Mantel-Haenszel (CMH) test for 2x2xK binary phenotype x variants | population cluster was employed with sites corrected for multiple testing using the Bonferroni correction at a p-value of 0.01.

6.2.6 Linkage analysis

Linkage disequilibrium was explicitly tested using Haploview (Barrett, Fry *et al.* 2005), which was devised for human genetics. However, bacterial recombination is not equivalent to human crossing over where linkage over long distance can be ignored. Therefore, the human genetics tool used (Barrett, Fry *et al.* 2005) would ignore any pairwise comparison over 500 kb. Here, the default setting was adjusted so that the tool considered all pairwise comparisons under 2,200 kb, which is equal to the size of the whole genome. The information was treated as male human X-chromosome to retain its haploidy, thereby incorporating all possible linkage predictions into our analysis. Using 95% confidence intervals as described in (Gabriel, Schaffner *et al.* 2002), a linkage block was identified as a region within a low recombination rate (here referred to as linkage loci). Physical linkage size detected in Maela and Massachusetts data were compared to illustrate the effect of population size on the power for separating causative SNPs from linked SNPs.

6.2.7 Estimation of percentage of resistance in the population explained by candidate loci

High-stringency SNPs co-detected in both Maela and Massachusetts datasets were used to cross-predict resistance in each population separately. The proportion of resistance in the population that could be explained by co-predicted SNPs (here grouped into linkage loci) were plotted for each of the test populations. The order of loci added was permutated to accommodate all possible combinations.

6.2.8 Specificity to different classes of beta-lactams

To test whether or not there were SNPs conferring more specific resistance to certain classes of beta-lactam antibiotics, GWAS was repeated on the SNPs co-detected in both populations. The binary phenotypes were replaced with continuous phenotypes, penicillin MIC values and ceftriaxone MIC values. P-values calculated from penicillin MIC and ceftriaxone MIC for each SNP were grouped by the linkage structure computed as discussed above.

6.2.9 Prevalence of candidate loci in the population

For each BAPS cluster in both the Maela and Massachusetts data, the mean prevalence of candidate loci was calculated by averaging the frequency of linked SNPs detected in each locus per cluster size.

6.3 Results

6.3.1 Identification of loci associated with beta-lactam non-susceptibility

Genome-wide association study (GWAS) is a broadly used approach in human genetics to identify SNPs associated with complex diseases, ranging from cancer to mental health (Sullivan, Daly *et al.* 2012, Goldstein, Allen *et al.* 2013, Pharoah, Tsai *et al.* 2013). The test compares SNPs across a large population including individuals with and without the disease. GWAS reports SNPs enriched in the disease population (case) but absent in healthy population (control) as potential risk factors indicating that the individual that carries the risk alleles is more likely to develop the disease.

Similar to these studies conducted in humans, SNPs across beta-lactam susceptible and non-susceptible pneumococcal populations were compared here. GWAS was independently performed with genetic variants called from whole genome alignments of 3,085 pneumococcal strains collected from a carriage cohort in Maela and 616 strains from a carriage cohort in Massachusetts (Croucher, Finkelstein *et al.* 2013). The binary phenotypes, which are based on susceptibility and non-susceptibility to beta-lactams were determined using the Clinical and Laboratory Standard Institute guidelines (CLSI, 2008). Strains with penicillin minimum inhibitory concentration (MIC) $\leq$ 0.06 μg/ml are classified as susceptible; applying these cut-offs across the Maela and Massachusetts data gave 1,729 non-susceptible (case) and 1,951 susceptible (control) samples for performing GWAS (with 21 unknown). This section discusses how GWAS was performed, including essential corrections needed to minimise false positive rates. The result reports both discrete and linked genetic variants (here called loci) associated with beta-lactam non-susceptibility in the Maela and Massachusetts pneumococcal populations.

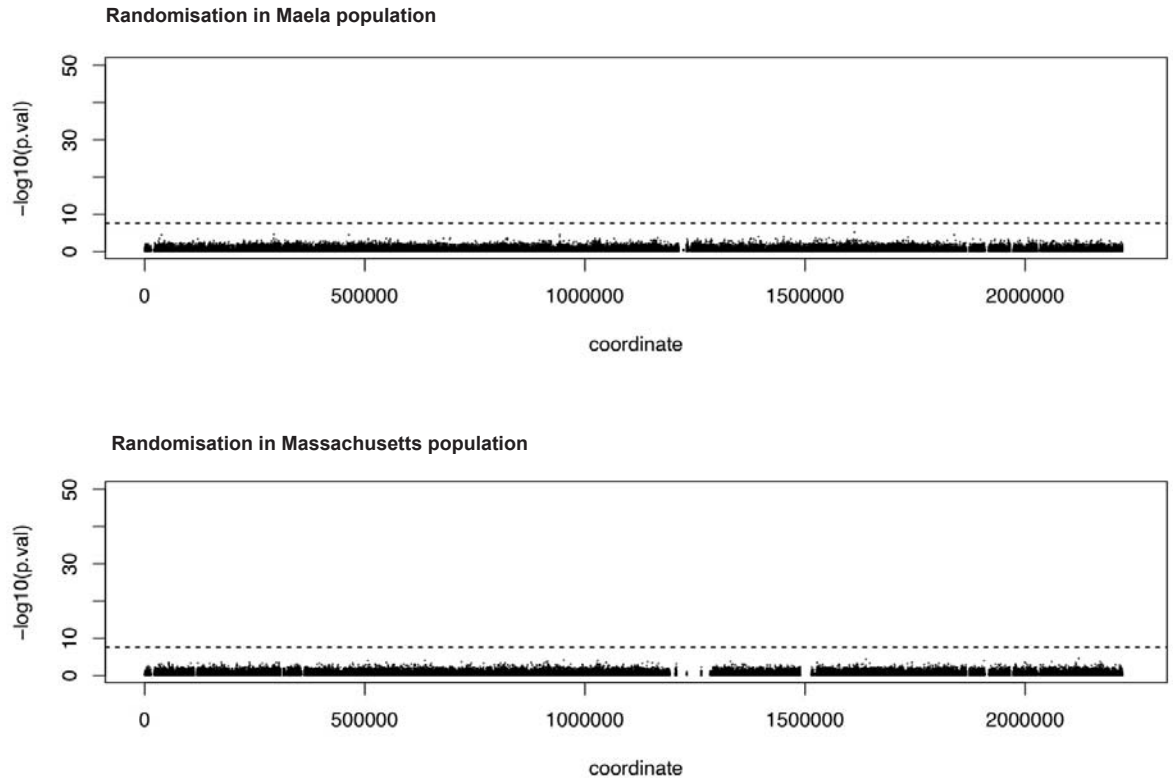
6.3.1.1 Randomisation test

Following quality control and data filtering, the basal intrinsic noise caused by genetic variation was estimated. As this noise can lead to false positives, it is essential to draw the result cut-off above the signals generated by noise. For each Maela and

Massachusetts data set, 100 GWAS permutations, with true genotypes but randomised binary phenotypes, were performed using PLINK (Purcell, Neale *et al.* 2007). Manhattan plots (**Figure 6.1**) summarise the randomised results from Maela (top) and Massachusetts data (bottom). None of the permutations of either datasets achieved significant association at p-value 0.01 with a Bonferroni correction for multiple testing, indicating a stringent threshold. Consequently, a significance p-value of 0.01 was applied throughout the study.

Figure 6.1 Randomised control for intrinsic noise based on genetic variation alone

Manhattan plots demonstrate no significant associations in either Maela or Massachusetts data using real genotype data and randomised resistance phenotype assignments. Horizontal dotted lines mark the cut-off with Bonferroni correction at p value = 0.01.



6.3.1.2 Population stratification

GWAS is sensitive to bias induced by population stratification. The test statistic is based on the assumption of independent observations. However, this is often violated as in humans; the cases may be overrepresented in a certain group of the population compared to the others (Wray, Yang *et al.* 2013), and in bacteria, the case population may be a part of the same clonal complex (Sheppard, Didelot *et al.* 2013). This leads to a true association locus being confounded by the underlying population structure, resulting in excessive false positive discoveries. As a consequence, population stratification is a necessary step in GWAS to make the studies consistent and replicable.

The Maela and Massachusetts populations consist of strains from a species-wide samplings; they respectively represent 277 and 154 known multilocus sequence types. Their population structures were predefined based on whole genome sequence similarity. The Bayesian based software BAPS was used to estimate the structure in both populations. This resulted in 33 and 16 initial clusters for the Maela and Massachusetts data, respectively. Due to the large sample size of the Maela data set, BAPS was additionally run in a hierarchical manner, generating secondary clusters within each primary cluster. These secondary clusters were used to represent the Maela population structure.

Based on this clustering information, the Cochran-Mantel-Haenszel (CMH) association statistic was employed to test for association between beta-lactam non-susceptibility and specific variants, conditioned on the population clusters. The Bonferroni correction for multiple testing at p-value of 0.01, as discussed in 6.1.1, was used as the cut-off. The reduction in false positive rates after correction for underlying population structure was estimated by a parameter called genomic inflation factor. The inflation factor is defined as the ratio of observed distribution of the test statistic to the expected mean, thereby allowing the extent of inflation and false positive rate to be quantified (Devlin and Roeder 1999). A high inflation factor typically indicates a high rate of false positives where associations are influenced by population structure. The application of CMH test with clustering condition reduced

the inflation factors from 80.16 (mean chi-squared statistic = 68.99) to 2.56 (mean chi-square statistic = 3.05) in the Maela data, and 13.18 (mean chi-square statistic = 14.17) to 3.76 (mean chi-square statistic = 4.73) in the Massachusetts data. The decreases in genomic inflation factors in both populations indicate lower false positive rates due to underlying population structure. However, these observed inflation factors are still relatively high compared to GWAS conducted on the human nuclear genome, suggesting that intrinsic clonal population structure remains an issue for bacterial association studies.

6.3.1.3 Case-control association analysis

The CMH test described in 6.1.2 was performed, giving 858 and 1,721 SNPs associated with beta-lactam non-susceptibility in the Maela and Massachusetts populations, respectively (**Figure 6.2**, SNPs tabulated in **Appendix C - D**). Among these, 301 SNPs were found to be associated with non-susceptibility in both populations (**Figure 6.3**, tabulated in **Appendix E**). Given that the two data sets have different population structures, which have evolved independently, these co-detected SNPs represent a set of candidates in which one can have more confidence. The 301 co-detected SNPs consist of three SNPs localised in intergenic regions, and 298 SNPs found in coding sequences. The latter can be further divided into 71 non-synonymous and 227 synonymous SNPs. The detection of non-synonymous SNPs implies a functional effect which might contribute to beta-lactam non-susceptibility. Synonymous SNPs, on the other hand, might not play a causative role but could be tightly linked to causative SNPs with insufficient recombination in the data set to separate the link (here called “hitchhiking” SNPs), and therefore form part of the same haplotype block. These linkage structures and the limitation they imposed on the predictions will be explored in the next section.

Figure 6.2 Summary of the genome-wide association study conducted in two separate datasets

Manhattan plots summarise the association of whole-genome SNP variant with beta-lactam susceptibility in the Maela and Massachusetts data as well as particular gene regions which show strong associations. Top panel represents the statistical significance of association (y-axis) for each variant arranged in order on the genome (x-axis) in the Maela (red) and Massachusetts (blue) data. Horizontal dotted lines in both top and bottom panels indicate a significance cut-off after Bonferroni correction of p value = 0.01. Genes with significant associations are annotated on top. Genes coding for penicillin binding proteins: *pbp2x, pbp1a,* and *pbp2b*, whose roles in beta-lactam resistance are well characterised, are highlighted in grey. Bottom panel expands the view of penicillin binding protein genes where most of the significant associations are detected: from left to right *pbp2x, pbp1a,* and *pbp2b*. Protein domains identified within these genes are shaded in pale grey and labelled. The vertical dotted lines represent the active sites of the transpeptidase domain. Plus signs denote synonymous SNPs and dots denote non-synonymous SNPs.

(**Figure is shown on the next page**)

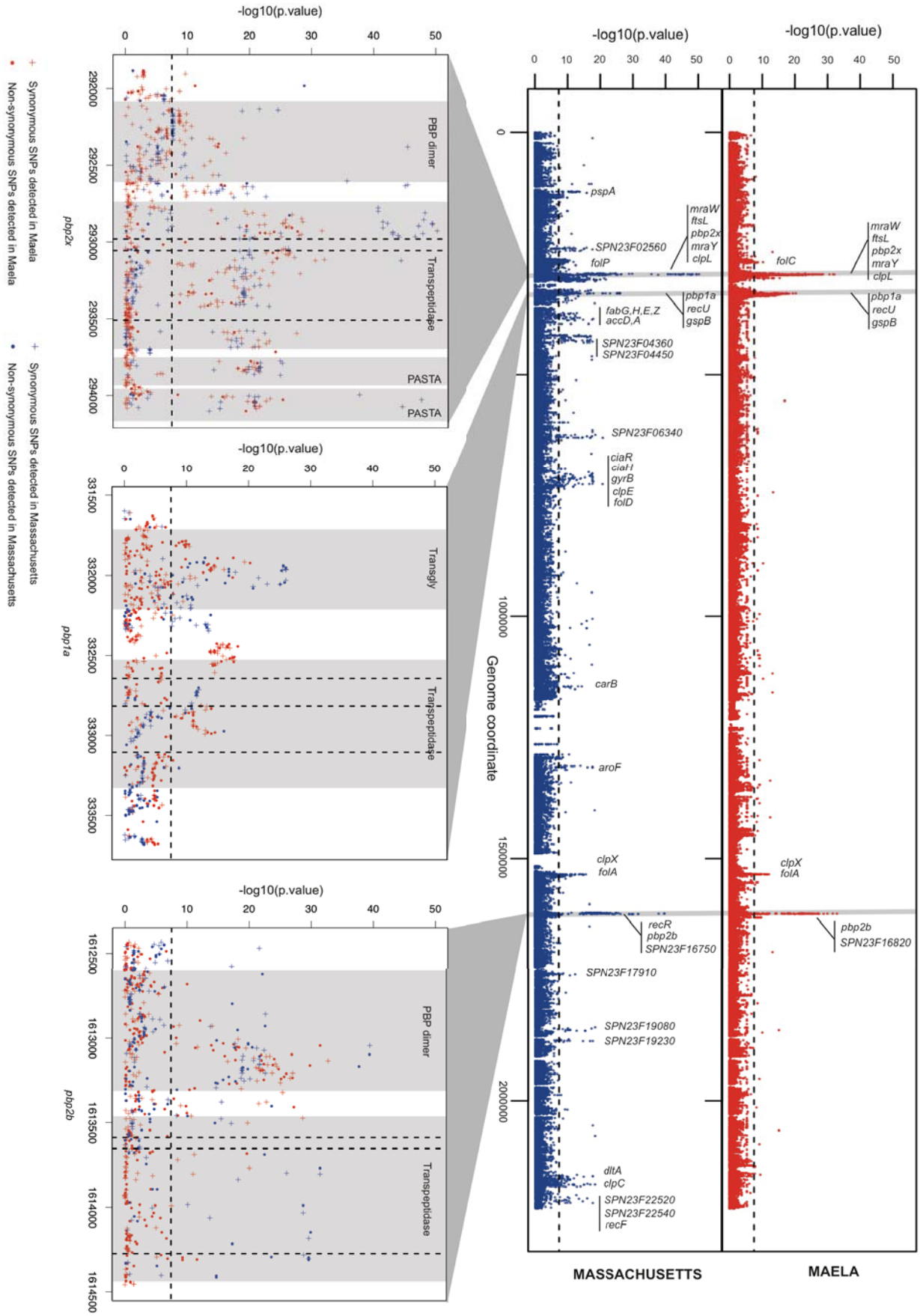
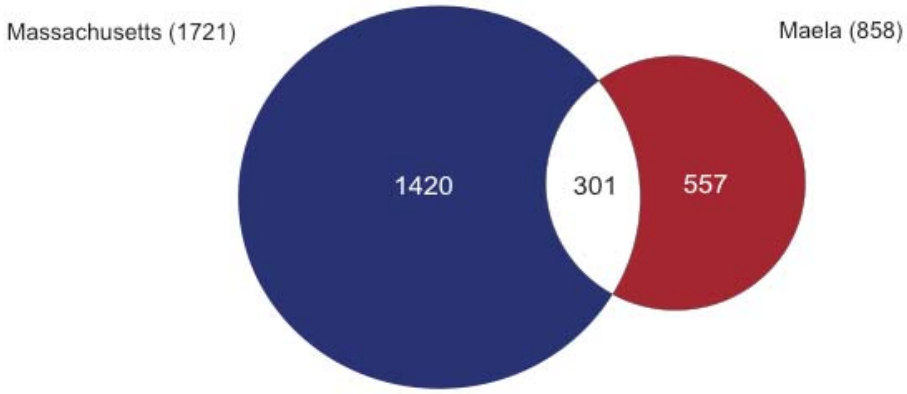

Synonymous SNPs detected in Maela
Non-synonymous SNPs detected in Maela
Synonymous SNPs detected in Massachusetts
Non-synonymous SNPs detected in Massachusetts
-log10(p.value)
pbp2x
PBP dimer
Transpeptidase
PASTA
PASTA
pbp1a
Transgly
Transpeptidase
pbp2b
PBP dimer
Transpeptidase
Genome coordinate
MASSACHUSETTS
MAELA
pspA
SPN23F02560
folP
mraW
ftsL
pbp2x
mraY
clpL
fabG,H,E,Z
accD,A
pbp1a
recU
gspB
SPN23F04360
SPN23F04450
SPN23F06340
ciaR
ciaH
gyrB
clpE
folD
carB
aroF
clpX
folA
recR
pbp2b
SPN23F16750
SPN23F17910
SPN23F19080
SPN23F19230
dltA
clpC
SPN23F22520
SPN23F22540
recF
folC
SPN23F16820

Figure 6.3 Summary of single nucleotide polymorphisms (SNPs) associated with beta-lactam non-susceptibility

A Venn diagram summarises the number of SNPs reaching significance in each of the Maela and Massachusetts datasets, and those that are co-detected in both.



6.3.1.4 Linkage disequilibrium

6.3.1.4.2 Defining linkage blocks

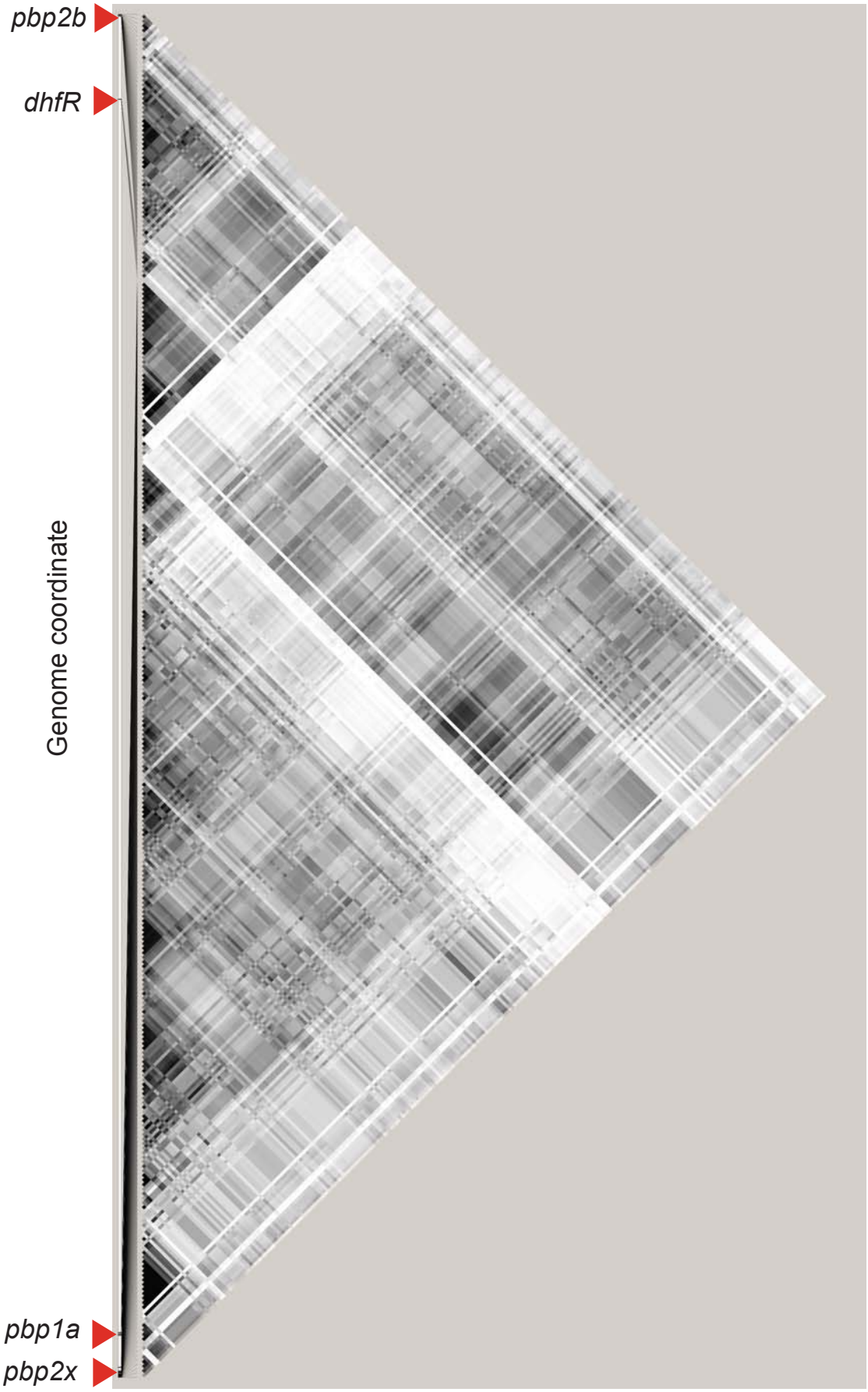
Linkage between candidate SNPs reported in 6.1.3 was explicitly tested using Haploview (Barrett, Fry *et al.* 2005). In humans, linkages between distant sites are disrupted by crossing-over, the recombination process where homologous chromosomes pair up and exchange different segments of their genetic material. However, bacterial recombination does not necessarily break long distance linkage. Therefore, Haploview was set to consider all possible pairwise comparisons over the entire size of *S. pneumoniae* genome (approximately 2,200 kb). This additional

setting allowed the application of Haploview, which was devised for human genetics, to be used in bacteria. Using 95% confidence bounds as described in (Gabriel, Schaffner *et al.* 2002), a haplotype blocks were identified as regions with a low recombination rate. Linkage information for SNPs detected in Maela and Massachusetts are listed in **Appendix C** and **D**, respectively. For SNPs co-detected in both Maela and Massachusetts populations, 51 linked loci were detected (**Figure 6.4)**. Among these, nine were single SNPs and 42 were in linkage blocks of between two and 19 SNPs, of which 12 contain only a single non-synonymous SNPs (**Appendix E**).

Figure 6.4 Linkage analysis for SNPs co-detected in two separate datasets

The Haploview plot illustrates linkage disequilibrium (r^2) between co-detected SNP candidates from the Maela and Massachusetts datasets. The bar of the left represents the genome position of the SNPs, connected by lines to the diamond plot on the right. A complete black diamond represents complete linkage disequilibrium between candidate SNPs (r^2=1), while a white diamond represents a perfect equilibrium (no linkage) (r^2 = 0)

(**Figure is shown on the next page**)


pbp2b
dhfR
Genome coordinate
pbp1a
pbp2x

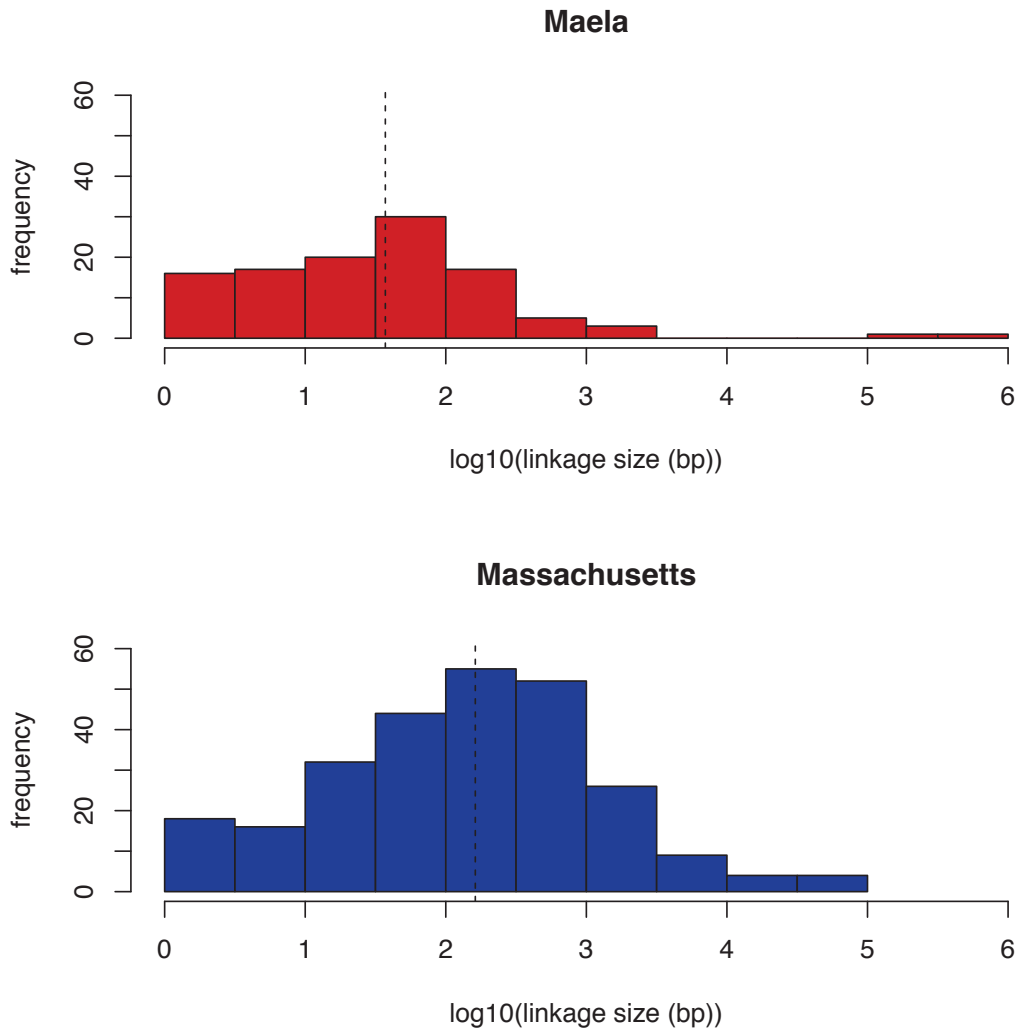
6.3.1.4.2 Larger linkage blocks observed in smaller dataset

As noted earlier, a total of 858 and 1,721 SNPs were detected in Maela and Massachusetts populations which comprise 3,085 and 616 strains respectively. The result was rather counter-intuitive, as one would expect higher number of candidate SNPs to be detected from the larger sample set of Maela than Massachusetts, which was not observed here. This could be due to the fact that population structures in the two settings were independently determined in previous works (discussed in 3.3.1.1 and Croucher, Finkelstein *et al.* 2013). Therefore, it is possible that the clustering information from the two data sets is not equivalent in their stringency. This subsequently leads to more strict control over population stratification in one population than the other.

Another potential explanation is that the Maela and Massachusetts data sets have different linkage structure. The sizes of the linkage blocks detected from the two populations in 6.1.4.1 were compared. Indeed, linkage block sizes detected in the Maela data were significantly smaller than the Massachusetts data (Mann Whitney test p-value 6.53×10^{-9}). Shown in **Figure 6.5**, the medians of Maela and Massachusetts linkage blocks were 37.5 bp and 165 bp respectively, presenting an inverse relationship between the size of the linkage structure and the sample size. This suggests that many of the candidate SNPs detected in the Massachusetts data are potentially hitchhikers, thereby resulting in greater false positives. Such observation possibly reflects the limitation of small data sets where there might not be enough recombination detection to sufficiently break the linkage structure. However, longer linkage blocks may also be expected from the Massachusetts population where carriage rates are lower, potentially reducing the frequency of opportunities for the strains to recombine.

Figure 6.5 Summary of physical linkage structure in two separate datasets

Size of linkage detected in the Maela (red) and Massachusetts (blue) association studies were plotted as histograms on log10 scale. Vertical dotted lines mark the median size of haplotype blocks that harbour candidate SNPs (37.5 bp in Maela data and 165 bp in Massachusetts data)



6.3.1.5 Estimating the amount of the non-susceptible phenotype explained by co-detected loci in the Maela and Massachusetts population

To estimate how much of the phenotypic resistance in the samples could be explained by the identified SNPs, cross-prediction tests were performed using only the SNPs co-

detected in both Maela and Massachusetts association studies. The co-predicted SNPs, grouped by their linkage structure, were tested back against each population. The results (**Figure 6.6**) show that close to 100% of the resistance in each population could be explained by all of the co-detected SNPs.

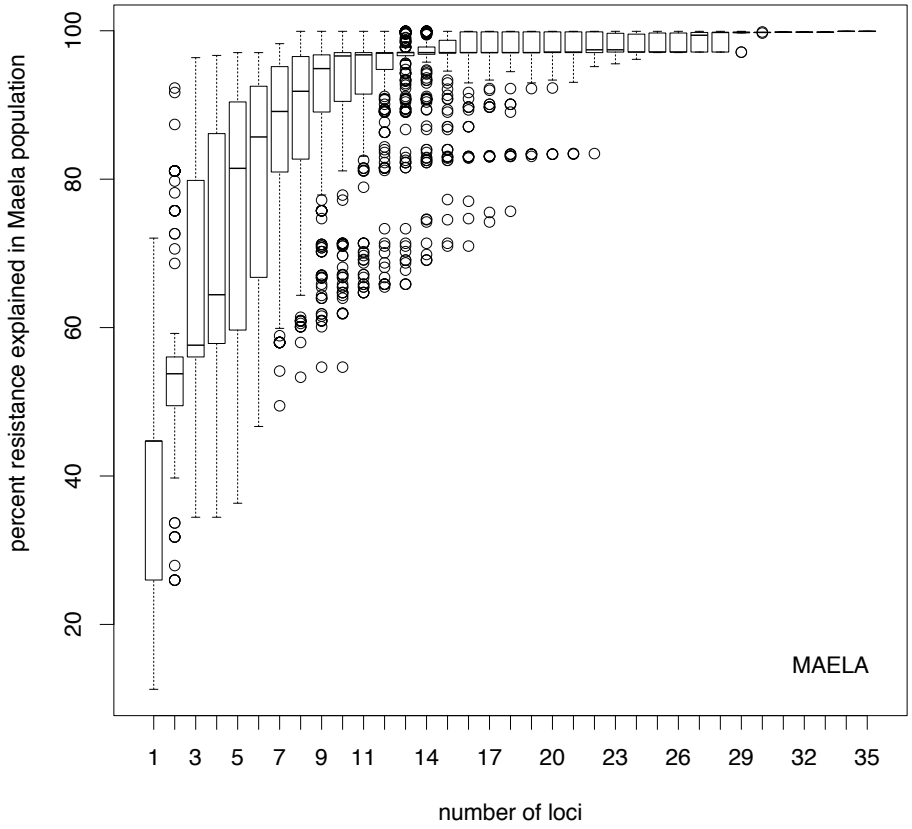
Unlike human polygenic traits where each locus contributes only a small effect on the phenotype, each of these bacterial loci appears to have much stronger effect. This is consistent with experimental characterisations where introductions of a single variant can sometimes lead to a change in pneumococcal beta-lactam susceptibility (Hakenbeck, Bruckner *et al.* 2012). This can be demonstrated using odds ratios, which indicate the size of the effect of each associated SNP. While human GWAS report a median odds ratio of 1.33 per SNP (Ku, Loy *et al.* 2010, Manolio 2010), this analysis gave a median odds ratio of 11.09 per SNP, indicating a stronger effect size.

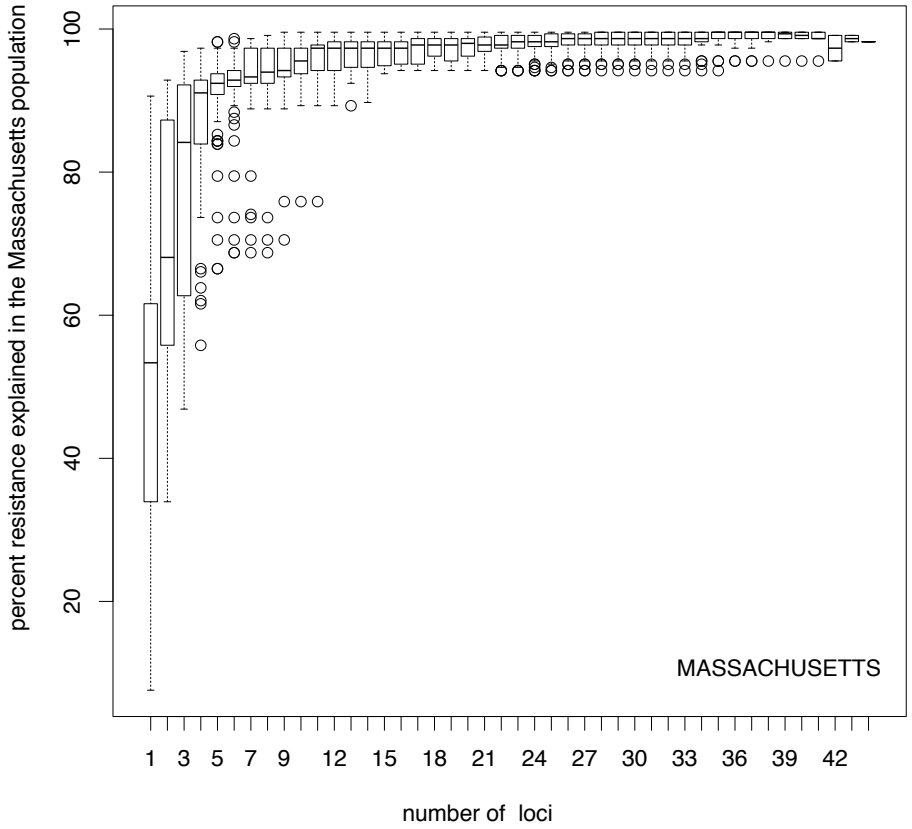
For both the Maela and Massachusetts populations, the percentage of resistance explained plateaued after the addition of approximately 10 loci in any order. This suggests that, at most, about 10 loci are required to make a susceptible strain non-susceptible and this can be done through multiple different combinations. However, in each resistant isolate, combinations of more than ten loci are commonly detected, perhaps indicating that not all loci are involved in conferring resistance. Some may play a compensatory role in reducing the fitness cost of resistance variants. In total, the co-detected variants are present in 100% and 98% of the Maela and Massachusetts resistant strains respectively, highlighting that a large proportion of possible resistance variants were captured in this analysis.

Figure 6.6 Percentage of the non-susceptible phenotype explained by co-detected loci in the Maela and Massachusetts populations

The plots represent proportions of resistance in the population (y-axis) explained by all combinations of increasing numbers of co-detected loci (x-axis), based on combinations of loci observed from both Maela and Massachusetts data.

(**Figure is shown on next page**)


percent resistance explained in Maela population
20
40
60
80
100
MAELA
1 3 5 7 9 11 14 17 20 23 26 29 32 35
number of loci


percent resistance explained in the Massachusetts population
20
40
60
80
100
MASSACHUSETTS
1 3 5 7 9 12 15 18 21 24 27 30 33 36 39 42
number of loci

6.3.2 Biological relevance of candidate loci

Following the identification of potential candidate loci associated with beta-lactam non-susceptibility, this section tries to explain the finding in 6.1 in biological context.

For both population settings, candidate loci show a higher enrichment in genes compared to intergenic regions than would be expected by chance (Fisher's Exact Test p-value > 0.0001). These loci are not randomly distributed across the whole genome, but clustered within certain genes (**Figure 6.2**). Loci co-detected in both populations are localised in genes participating in the peptidoglycan biosynthesis pathway, including penicillin binding proteins (*pbp2x*, *pbp1a*, *pbp2b*), two transferases required for cell wall biogenesis (*mraW*, *mraY*), the cell division pathway (*ftsL*, *gpsB*), heat shock protein and chaperones (*clpL*, *clpX*), the recombination pathway (*recU*) and a metabolic gene known to be involved in resistance to co-trimoxazole (*dhfR*). Some of these sites, particularly in the *pbp* genes, matched those previously reported to play an important role in beta-lactam resistance in the literature (**Appendix E**, providing independent validations to the methodology and some of the results. To my knowledge, out of 71 non-synonymous SNPs reported here, 43 SNPs are novel and potentially contribute to beta-lactam non-susceptibility in addition to those identified in previous studies.

6.3.2.1 Candidate loci in genes participating in the peptidoglycan biosynthesis pathway

Since beta-lactam antibiotics work by inhibiting cell wall biosynthesis, it is not surprising to observe significant associations between non-susceptible phenotypes and variants in genes participating in the peptidoglycan biosynthesis pathway, including *pbp2x*, *pbp1a*, *pbp2b*, *mraW* and *mraY*. Many single amino acid alterations in *pbp2x*, *pbp1a* and *pbp2b* have been previously demonstrated experimentally to increase pneumococcal resistance to beta-lactams (Hakenbeck, Bruckner *et al.* 2012). Mutations within or close to the active sites of the transpeptidase domain in penicillin binding proteins have been reported to be associated with penicillin resistance. By interfering with the formation of a covalent complex between the active site serine

and antibiotic molecules, these mutations help reduce the binding affinity of beta-lactam rings to the transpeptidase enzyme. This allows the pneumococci to form a functional cell wall, and thereby becoming non-susceptible. Many predicted loci co-localise with or surround the transpeptidase active sites. These are recognised as three conserved amino acid motifs, SXXK, SXN and KT(S)G and are highlighted as vertical dotted lines in the bottom panel of **Figure 6.2**. Amino acid alterations close to the active site often lead to conformational changes. An association at T338A in *pbp2x*, which is located next to the active site at position 337 was observed. The side chain of T338 is required for hydrogen-bonding, and the T338A substitution results in the distortion of the active site which lowers the binding affinity to beta-lactam (Hakenbeck, Bruckner *et al.* 2012). In *pbp1a*, an alteration from TSQF to NTGY at position 574-577 also matched previous reports. This alteration results in lower acylation efficiency *in vitro* (Job, Carapito *et al.* 2008). In addition to candidates known to confer structural changes, some candidates are consistent with compensatory mutations identified earlier. A substitution E285Q in *pbp1a* proposed to ameliorate the fitness cost caused by resistance in *pbp2b* (Albarracin Orio, Pinas *et al.* 2011) was also detected here. Other alterations that are consistent with previous literature were tabulated in **Appendix E**.

In addition to *pbp* genes, associations were observed in *mraY* and *mraW*, which encode transferases. Both function upstream of the *pbp* genes in the peptidoglycan biosynthesis pathway. They could potentially affect antibiotic susceptibility or represent compensatory mutations that interact epistatically with changes associated with resistance.

6.3.2.2 Candidate loci in genes outside the peptidoglycan biosynthesis pathway

The genome-wide screen provided an opportunity to identify associations outside the peptidoglycan biosynthesis pathway, which is the direct target of beta-lactams. The results obtained here highlighted nine loci found outside this pathway. The loci comprised 31 SNPs co-detected in both Maela and Massachusetts datasets; some of which were detected in the gene for heat shock protein *clpL*. Mutants lacking ClpL have been reported to be more susceptible to penicillin. The effect was attributed to

the ability of ClpL to interact with PBP2x and to stabilise *pbp2x* expression (Hakenbeck, Bruckner *et al.* 2012). In the Massachusetts data alone, amino acid alterations in a histidine kinase sensor, *ciaH* and its response regulator *ciaR*, also show an association with beta-lactam non-susceptibility. Previous reports showed that mutations in *ciaH* resulted in higher expression of *ciaR*. Hyperactivation of the regulon CiaR in turn leads to increased beta-lactam resistance (Muller, Marx *et al.* 2011, Hakenbeck, Bruckner *et al.* 2012). Associations between genes functioning in cell division, *ftsL* and *gpsB,* and beta-lactam resistance were observed in both Maela and Massachusetts datasets. Both proteins are required for complete cell wall formation. Depletion of GpsB leads to cell deformation (Land, Tsui *et al.* 2013). Based on known functions, these identified candidate loci potentially interact with *pbp* genes, either directly or indirectly through regulation or participating in cell wall formation. However, the hypotheses arising from these associations will require further experimental validation to explore the mechanisms of how these alterations might influence beta-lactam susceptibility.

6.3.2.3 Candidate loci in genes conferring resistance to other antibiotics

In both Maela and Massachusetts data, association signals were unexpectedly detected in dihydrofolate reductase (*dhfR*) and dihydropteroate synthase (*folP*), genes whose allelic variants are known to confer resistance to another group of antibiotics, co-trimoxazole. The drug interferes with folate synthesis, which is essential for nucleotide biosynthesis, thereby inhibiting the bacterial DNA synthesis pathway (Maskell, Sefton *et al.* 2001). Since beta-lactams and co-trimoxazole target different pathways, and no known protein-protein interactions between the two pathways have been reported, it is unlikely that variants detected in *dhfR* and *folP* would contribute to a rise of beta-lactam resistance mechanistically. The linkage analysis discussed in 6.1.4.1 has shown that loci at *dhfR* and *folP* genes were not genetically linked to *pbp* genes or other beta-lactam targets, suggesting that they are not hitchhikers but were selected through advantages that their allelic variants had conferred to on the strains.

Based on clinical records from Thailand, co-trimoxazole has been listed as the second most frequently used antibiotic for upper respiratory treatment after beta-lactam

(Thamlikitkul and Apisitwittaya 2004). This contemporaneous use of both beta-lactam and co-trimoxazole antibiotics in the studied populations may have driven co-selection for resistance to the two unrelated groups of antibiotics. In both Maela and Massachusetts datasets, strains that are phenotypically resistant to beta-lactams are more likely to be phenotypically resistant to co-trimoxazole (**Table 6.1**, Fisher's exact test p-value < 2.2×10^{-16}). This suggests that the frequent use of two antibiotics might have created co-selection pressures, resulting in the detection of linked yet unrelated association signals.

Table 6.1 Co-occurrence of co-trimoxazole and beta-lactam resistance phenotypes

			Beta-lactam		**Fisher's exact test p-value and (odds ratio)**
			resistant	sensitive	
Maela	**Co-trimoxazole**	resistant	1,356	771	< 2.2×10^{-16} (10.36)
		intermediate	77	280	
		sensitive	68	517	
Massachusetts		resistant	102	38	< 2.2×10^{-16} (7.29)
		sensitive	125	341	

6.3.3 Beta-lactam specificity of resistance mutations

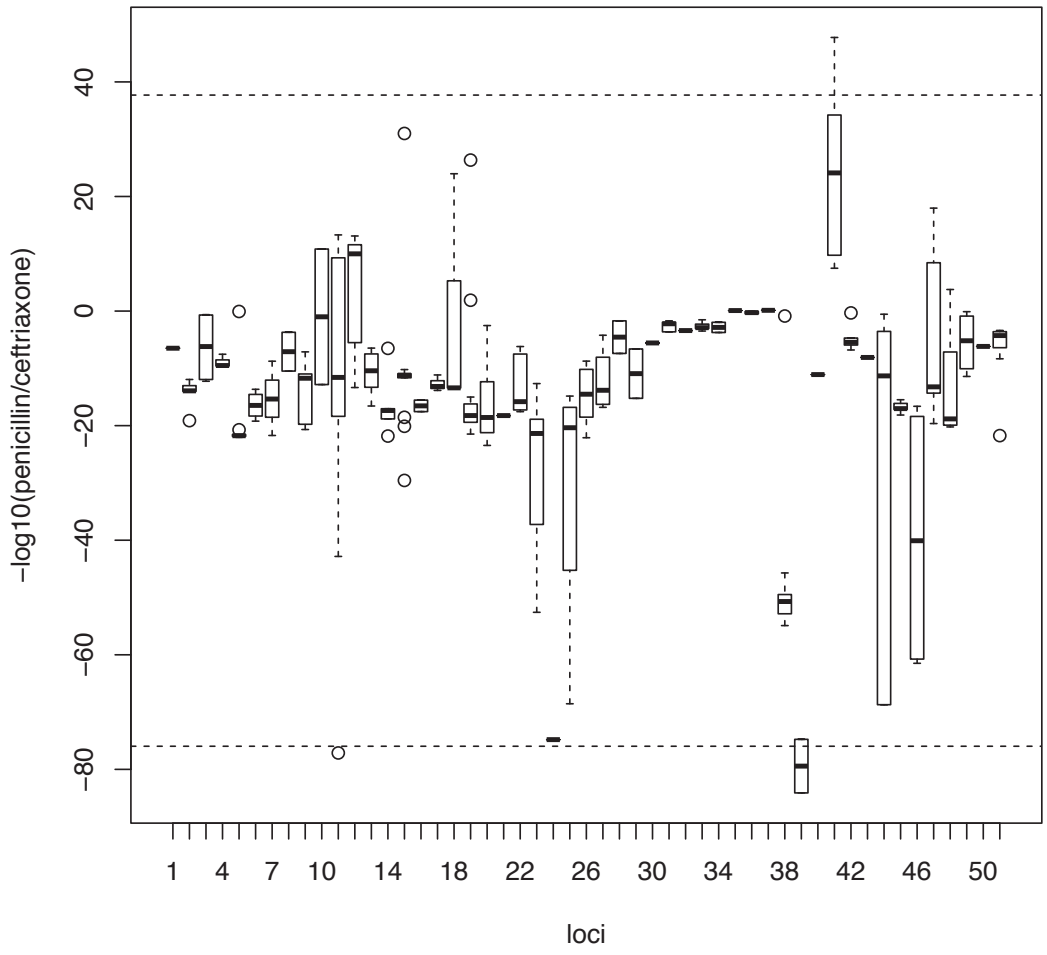
Although all beta-lactam antibiotics target peptidoglycan biosynthesis pathway, the group encompasses a range of drug molecules with different structures and chemical properties. Therefore, it is possible that each alteration detected earlier might confer greater resistance to certain classes of beta-lactam antibiotic than others. To test this hypothesis, the analyses were performed on the candidate SNPs identified in 6.1.3 using the continuous phenotypes recorded as the MIC value of two subclasses of beta-lactam antibiotics, penicillins and cephalosporins (here represented by ceftriaxone). Penicillins and cephalosporins display different structures. While the beta-lactam ring

is fused to a 5-membered thiaxolidine ring in penicillins, it is fused to a 6-membered dihydrothiazine ring in cephalosporins. Also, the side chains of the two drugs differ, leading to different kinetic properties (DePestel, Benninger *et al.* 2008).

The differential association of each locus with resistance to either penicillins or cephalosporins was summarised in **Figure 6.7**. Loci with stronger association with penicillin are distributed along the positive y-axis, while those showing a stronger association with cephalosporins are distributed along the negative y-axis. The result shows that some loci do not contribute equally to resistance to the different classes of beta-lactam antibiotic (Kruskal-Wallis rank sum test, p-value $< 2.2 \times 10^{-16}$). As can be seen in **Figure 6.7**, some loci show a strong preference towards penicillins or cephalosporins, suggesting specificity in resistance patterns.

Figure 6.7 Specificity of association signals for co-detected candidate loci with different classes of beta-lactam antibiotics

Bonferroni-adjusted p-values from associations with continuous phenotypes with each co-detected SNP were grouped into their linkage loci. Positive values on the y-axis show stronger association with penicillin resistance while negative values show stronger association with cephalosporin resistance. Horizontal dotted lines represent the 99th percentile.



6.3.4 Distribution of candidate alleles in the Maela and Massachusetts populations

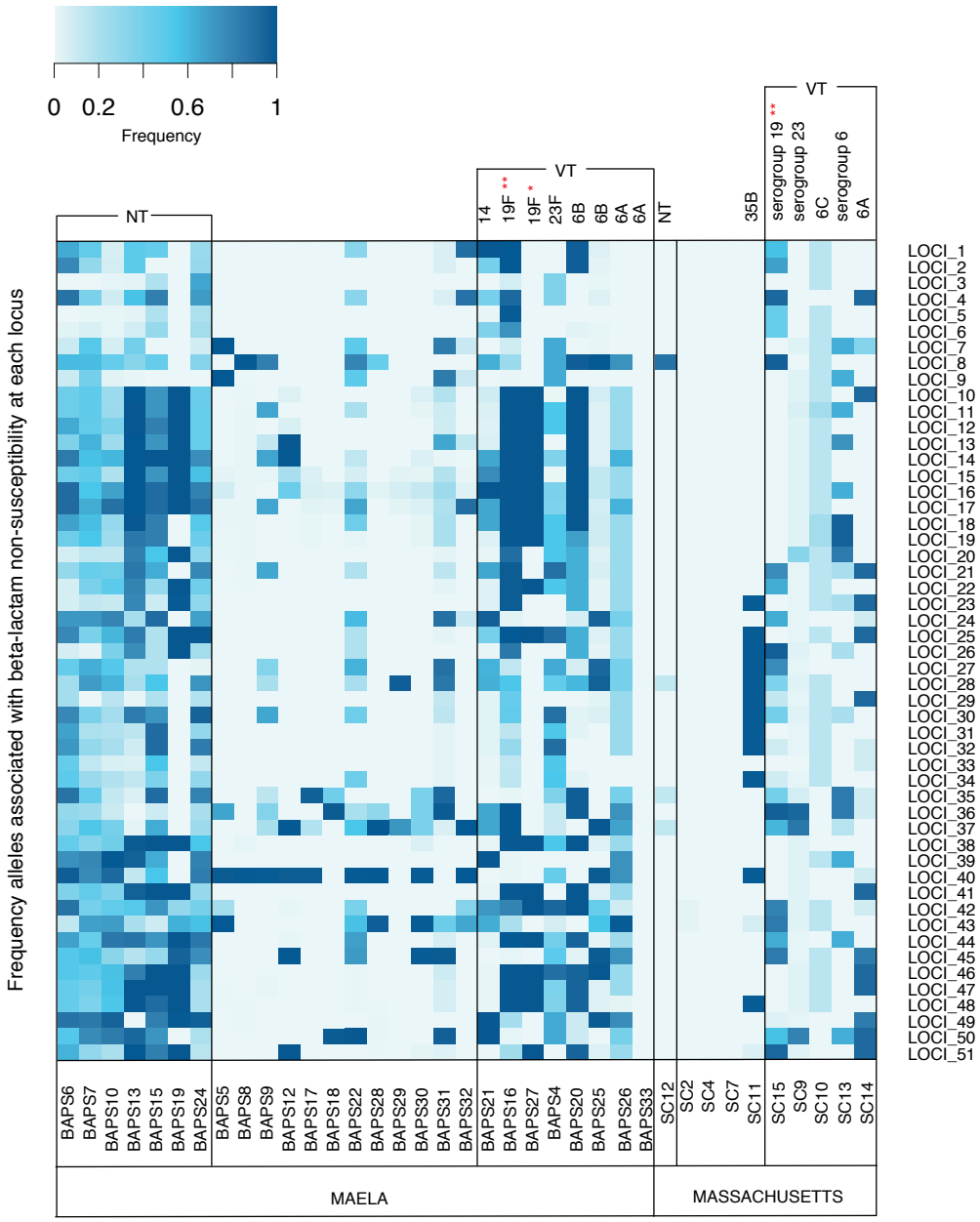
Next, the question posed in the previous chapter on the role of the nontypable strains in mediating the spread of beta-lactam resistance was investigated. Given that the pneumococcal population structure in both data sets is known (see 3.3.1.1 and Croucher, Finkelstein *et al* 2013), the prevalence of predicted beta-lactam resistance alleles in subpopulations was explored in each setting, particularly among nontypable clusters in the Maela population.

The result shows that candidate loci for beta-lactam resistance are heterogeneously distributed within subpopulations of both Maela and Massachusetts pneumococcal populations (**Figure 6.8**). The globally dispersed multidrug resistant lineages PMEN-14 and PMEN-1, along with other vaccine target lineages appear to carry candidate resistance alleles at a higher frequency. This reflects the vaccine's design to target serotypes associated with antibiotic resistance (Dagan 2009, Croucher, Finkelstein *et al.* 2013). Given this high frequency of resistance alleles, vaccine administration might be expected to help reduce beta-lactam resistance within the population as a whole. However, this has not been observed following the administration of pneumococcal conjugate vaccine (PCV7) (Huang, Platt *et al.* 2005, Huang, Hinrichsen *et al.* 2009), and this observation might provide a mechanism at the allelic level explaining why this may not occur. Nontypable lineages in Maela and serotype 35B in Massachusetts are non-vaccine lineages. Shown in **Figure 6.8**, each of these non-vaccine lineage harbours a high frequency of resistance alleles, allowing them to act as a reservoir for beta-lactam resistance following vaccination programmes.

For nontypable lineages, this observation provides an independent validation to their role as the source and sink of resistance alleles previously proposed in chapter 4 and chapter 5. Such a role might help generate more combinations of beta-lactam resistance alleles that are then seeded into the wider population. Such a trend observed in the NT lineages may thus have clinical consequences with respect to the spread of beta-lactam resistance in a community dominated by nontypable lineages like Maela.

Figure 6.8 Frequency of putative resistance alleles from candidate loci in the Maela and Massachusetts data

The resistance allele frequency at each locus (rows) in each population cluster (columns) is represented by different shades on a blue-grey scale. Vaccine and non-vaccine serotype status of each cluster are shown at the top of each column. Asterisks highlight two drug resistant global clones: * represents the PMEN1 lineage, ** represents the PMEN14 lineage. Note that mixed population clusters have been removed from the diagram.



6.4 Conclusion

Bacterial GWAS has allowed researchers to link individual elements of the genotype, including core genes, mobile genetic elements and SNPs, to specific phenotypes (Falush and Bowden 2006). This approach has been successfully performed in bacteria: *Campylobacter* (Sheppard, Didelot *et al.* 2013), *Staphylococcus aureus* (Laabei, Recker *et al.* 2014) and in this work, *Streptococcus pneumoniae*. Together, this holds a great promise for microbial functional genomics where one could design an experiment to systematically test the phenotypes followed by large-scale sequencing to draw the phenotype-genotype associations. Proposed by (Dutilh, Backus *et al.* 2013), a GWAS based approach may also be useful to link metagenomic entities including functions or taxa (operational taxonomic units) observed across metagenomic samples to clinical or environmental metadata. The term "metagenome-wide" association linking environmental parameters to metagenomic entities has been suggested.

Although these bacterial genome-wide and metagenome-wide associations are theoretically possible, their resolution will inherently be limited by the bacterial clonal population structure. This analysis used an *S. pneumoniae* dataset of large sampling size than previously, and the power to detect associated variants is therefore enhanced. Moreover, *pbp* genes, which are known targets of beta-lactams have been previously shown to be recombination hotspots, thereby significantly reducing the effect of long haplotype blocks. Together, this enabled a refinement of genetic determinants for beta-lactam non-susceptibility from "mosaic genes" to a single SNP or smaller linkage group. An increased resolution in the assignment of genetic variants will be useful for the prediction of antibiotic resistance/sensitivity from whole genome sequencing in surveillance and clinical studies.

A genome-wide screen allowed a search for loci associated with resistance beyond the known targets for beta-lactams in the peptidoglycan biosynthesis pathway, and reported associations in many previously characterised as well as many novel loci. The latter will require experimental validation to verify their contributions to resistance mechanisms. The results also show that loci can either contribute

universally to all beta-lactam resistance, or exert a stronger effect against certain classes of antibiotics. Moreover, these identified loci have a highly non-uniform distribution in the populations. They are enriched not only in vaccine-targeted but also non vaccine-targeted lineages, including the nontypable lineages detected in Maela. This provides further supportive evidence for the role of nontypables as the hub of genetic exchanges and a potential reservoir for antibiotic resistance genes in the Maela pneumococcal population, a concept that has been discussed throughout this thesis.