Host and pathogen genetics associated with pneumococcal meningitis

John Andrew Lees

Wellcome Trust Sanger Institute
Jesus College, University of Cambridge

July 2017

This dissertation is submitted for the degree of
Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed 60 000 words in length, as required by the School of Biological Sciences.

John Andrew Lees

July 2017

Acknowledgements

My main thanks have to go to Stephen Bentley, who has supervised me in exactly the way I would have wished. As well as being a steady hand on the tiller of my PhD, he has also shown me that it is possible to have a well-adjusted life, be kind to others and still be a great researcher. I will never forget all that he taught me (about Special Brew). Similar thanks must also go to my other supervisors. Jeff Barrett, who had no obligation to do so, fully accepted me into his research group. His group meetings and my conversations with him have really shaped this research and the way I think. Julian Parkhill has always been helpful and available for answering those tricky bacterial genomics questions no-one else could, and has given me many opportunities to present my work to other researchers (often in sunnier climes). Thanks too to Carl Anderson and John Welch, the other members of my thesis committee, whose freely contributed ideas have made this work better than it otherwise would have been.

Many collaborators have made this work possible. Nick Croucher is a pneumococcal master who I am lucky to know. Jukka Corander I am likewise lucky to know, and while explicitly involved with SEER I have felt his influence throughout this PhD. My Dutch friends Diederik, Matthijs, Philip, Arie and Bart have worked very hard on the unique dataset at the core of this thesis, and have been a pleasure to collaborate with. A special mention to Philip who even saw fit to invite me to his wedding. Paul Turner helped me throughout with the Maela dataset, and organised an excellent meeting in Cambodia without which chapter three wouldn’t exist.

To the people who I have talked to about my PhD, thanks for your contributions (tangible and intangible) to everything I’ve done here: Sumana, Katie, Tom, Jeremy, Liam, Marcia, Theresa, Simon, James, Sophia, Becca, Leo, Izzy, Claire, Darryl and Alison, and all the other members of teams 81 and 143 past and present. Thanks too to all of pathogen informatics and the graduate office, who have doubtless helped me in many ways. Finally, I am very grateful to the Wellcome Trust and MRC for funding this research.

In the knowledge that this is the only page most readers of this document will look at, the pressure to be witty or memorable is greatest here. I guess you’ll have to live with the Special Brew reference.

Summary

Host and pathogen genetics associated with pneumococcal meningitis

John Andrew Lees

Meningitis is an infection of the meninges, a layer of tissue surrounding the brain. In cases of pneumococcal meningitis (where the bacterium *Streptococcus pneumoniae* is the causative agent) this causes severe inflammation, requiring intensive care and rapid antibiotic treatment. The contribution of variation in host and pathogen genetics to pneumococcal meningitis is unknown. In this thesis I develop and apply statistical genetics techniques to identify genomic variation associated with the various stages of pneumococcal meningitis, including colonisation, invasion and severity.

I start by describing the development of a method to perform genome-wide association studies (GWAS) in bacteria, which can find variation in bacterial genomes associated with bacterial traits such as antibiotic resistance and virulence. I then applied this method to longitudinal samples from asymptomatic carriage, and found lineages and specific variants associated with altered duration of carriage. To assess meningitis versus carriage samples I applied similar analysis techniques, and found that the bacterial genome is crucial in determining invasive potential. As well as bacterial serotype, which I found to be the main effect, I discovered many independent sequence variants associated with disease. Separately, I analysed within host-diversity during the invasive phase of disease and found it to be of less relevance to disease progression.

Finally, I analysed host genotype data from four independent studies using GWAS and heritability estimates to determine the contribution of human sequence variation to pneumococcal meningitis. Host sequence accounted for some variation in susceptibility to and severity of meningitis. The work concludes with a combined analysis of pairs of bacterial and human sequences from meningitis cases, and finds variation correlated between the two.

Contents

1 Introduction **17**

1.1 Bacterial meningitis 17

1.1.1 Diagnosis, epidemiology and treatment 18

1.1.2 Causal organisms 20

1.1.3 Immune response to pneumococcal meningitis 21

1.1.4 A nationwide Dutch cohort 22

1.2 Pneumococcal biology 24

1.2.1 Importance of capsular serotype 24

1.2.2 Pneumococcal pathogenesis and immune evasion 25

1.2.3 Population studies of *S. pneumoniae* 29

1.2.4 Within-host variation of *S. pneumoniae* 32

1.3 Association mapping in humans 33

1.3.1 Genome-wide association studies 35

1.3.2 Heritability 40

1.3.3 Host susceptibility to infectious disease 42

1.4 Association mapping in bacteria 43

1.4.1 The effect of population structure 44

1.4.2 More variation and fewer samples 46

1.4.3 Early successes 47

1.4.4 Phylogenetic methods 49

1.4.5 Regression methods 50

1.5 Conclusions 51

2 Bacterial genome-wide association studies **53**

2.1 Introduction 54

2.2 K-mers as a generalised variant 55

2.2.1 Filtering k-mers 56

2.3 Accounting for population structure 57

2.3.1 Phylogenetic simulation of genomes 58

2.3.2 K-mer distance method producing covariates to control for population structure . . . 63

2.4 Association testing . . . 64

2.4.1 Significance cut-off . . . 67

2.4.2 Downstream interpretation of significant k-mers . . . 69

2.5 Development of SEER . . . 69

2.6 Benchmarking SEER . . . 70

2.6.1 Simulated data . . . 71

2.6.2 Antibiotic resistance in pneumococcal carriage . . . 74

2.6.3 Virulence of *Streptococcus pyogenes* . . . 77

2.7 Conclusions . . . 79

3 Variation in duration of asymptomatic pneumococcal carriage 80

3.1 Introduction . . . 81

3.2 Ascertainment of carriage episode duration using epidemiological modelling 82

3.2.1 Combining epidemiological data with genomic data . . . 86

3.3 Overall heritability of carriage duration is high . . . 87

3.4 Lineage effects on carriage duration . . . 88

3.4.1 Serotype and drug resistance explain part of the narrow-sense heritability . . . 90

3.4.2 Independent effects of serotype and genetic background . . . 92

3.4.3 Average carriage duration by serotype . . . 93

3.5 Additional loci identified by genome-wide association . . . 96

3.5.1 Prophage sequences associated with reduced carriage duration . . 98

3.5.2 Other loci associated with altered carriage duration . . . 101

3.6 Child age independently affects variance in carriage duration . . . 103

3.7 Conclusions . . . 105

4 Bacterial genetics contributing to invasive pneumococcal disease 107

4.1 Introduction . . . 108

4.2 Quality control and processing . . . 109

4.3 Catalogue of all pneumococcal variation . . . 111

4.3.1 Allelic variation of three pneumococcal antigens . . . 112

4.3.2 Phase variable type I R-M system allele (*ivr*) . . . 116

4.4 GWAS of bacterial variants associated with meningitis . . . 117

4.4.1 Role of common variation . . . 119

4.4.2 Role of rare variation . . . 124

4.4.3 Hierarchical Bayesian model for ivr allele prevalence . . . 131

4.5 Genetic adaptation over the course of single infections . . . 135

4.5.1 Reference free variant calling 136
4.5.2 No repeated post-invasion adaptation in coding regions across species 139
4.5.3 No evidence for repeated adaptation in intergenic regions in *S. pneumoniae* and *N. meningitidis* 144
4.5.4 No evidence for repeated adaptation in phase variable regions in *S. pneumoniae* and *N. meningitidis* 146
4.5.5 Carriage and invasive disease sample pairs show some evidence of repeated adaptation 149
4.6 Conclusions 151

5 Human genetics contributing to invasive pneumococcal disease 153
5.1 Introduction 154
5.2 GWAS of human variation associated with meningitis 155
5.2.1 Genetic data processing 156
5.2.2 Association results 163
5.2.3 Meta-analysis of four studies 170
5.3 Genome-to-genome analysis of host and pathogen variation 175
5.3.1 All by all variant association 176
5.3.2 Reduced representation of pathogen genome 177
5.3.3 Association of antigens 180
5.4 Conclusions 183

6 Conclusions 185
6.1 Summary of findings 185
6.1.1 Bacterial genome-wide association studies 186
6.1.2 Epidemiological variation of *S. pneumoniae* 189
6.1.3 Host and pathogen genetics of pneumococcal meningitis 190
6.2 Future directions 193
6.2.1 Bacterial GWAS methods 193
6.2.2 Genetics affecting pneumococcal meningitis 195
6.2.3 Future of statistical genetics in bacterial diseases 197

Bibliography 198

A Supplementary information 240
A.1 Data access and code availability 240
A.2 Supplementary figures 241

List of Figures

1.1 Incentive for enrolling patients in the MeninGene study 23
1.2 Interactions between the immune system and *S. pneumoniae* 27
1.3 Overview of GWAS study design 37
1.4 Phylogenetic illustration of lineage and locus variants 45

2.1 Site frequency spectrum of different variants 57
2.2 Phylogeny used in simulations of population structure definition 59
2.3 Supertree of phylogenetic reconstruction methods 62
2.4 Number of dimensions needed in distance matrix projection 64
2.5 p-p plots showing the effect of k-mer filtering 68
2.6 Power of SEER on simulated data 72
2.7 Coverage of k-mers significantly associated with Trimethoprim resistance 73
2.8 Fine-mapping of trimethoprim resistance causal variant 76
2.9 Manhattan plots from GWAS on *Streptococcus pyogenes* invasiveness . . 78

3.1 Carriage duration swabbing and sequencing study design 83
3.2 HMMs of swab time series, and their goodness-of-fit 84
3.3 Distribution of carriage duration, and effect of monotonic transformation. 87
3.4 Mapping of carriage duration onto phylogeny 89
3.5 Change in carriage duration associated with capsule switching events . . . 94
3.6 Manhattan plot of SNPs associated with carriage duration 97
3.7 Manhattan plots of phage-associated SNPs associated with carriage duration100
3.8 Predicted mean carriage duration as a function of child age 103

4.1 Inferred allele of pneumococcal antigens 114
4.2 Alignment of the two forms of PspC 115
4.3 Structure of the inverting variable restriction locus 117
4.4 Q-Q plots for invasive *S. pneumoniae* GWAS methods 121
4.5 Differing burden and frequency of rare variation between invasive and carriage isolates 125
4.6 Hierarchical model for the inverting variable restriction locus 133

4.7 Performance of variant calling methods on paired samples 138
4.8 Frequency of variation between blood and CSF isolates 139
4.9 Loss of function mutations in *dlt* during meningitis 141
4.10 Evidence of selection on *pde1* during meningitis 142
4.11 Mutations observed between all *S. pneumoniae* pairs, overlaid onto a common reference 145
4.12 Mutations observed between all *N. meningitidis* pairs, overlaid onto a common reference 146
4.13 Prevalence of each ivr allele in blood and CSF samples 148
4.14 Frequency of variation between carriage and CSF isolates from the same patient 150

5.1 PCA of human samples from the Netherland and Denmark 159
5.2 Quality control of genotype cluster plots 160
5.3 Demonstration of the effect of phasing 161
5.4 Locuszoom plot of association on chromosome 1 with unfavourable outcome166
5.5 Manhattan plot from GWAS of Dutch meningitis cases 167
5.6 Manhattan plot from GWAS of Dutch pneumococcal meningitis cases . . 168
5.7 Manhattan plot from GWAS of Dutch severe meningitis cases 169
5.8 Manhattan plot from GWAS of Danish meningitis cases 171
5.9 Manhattan plot from GWAS of Danish bacteremia cases 172
5.10 Manhattan plot from meta-analysis of meningitis susceptibility 174
5.11 Power for detecting genome-to-genome interactions 178
5.12 PEER factor analysis in genome-to-genome strains 179
5.13 Locuszoom plot of association on chromosome 10 with sequence cluster 8 181
5.14 Antigen classification in genome-to-genome analysis 182
5.15 Locuszoom plot of association between imputed SNPs and *pspC* allele . . 183

A.1 Monotonic warping function from warped-lmm 241
A.2 Q-Q plots of carriage duration, and transformations 242
A.3 Carriage duration regression diagnostics and outlier removal 243
A.4 Histogram of pairwise patristic distances between carriage isolates 244
A.5 Lasso regression of lineage effects on carriage duration 245
A.6 Phage presence in assemblies by blastn hit length 246
A.7 Distribution of lengths of k-mers associated with carriage duration 246
A.8 Quantile-quantile plots of carriage duration association p-values 247
A.9 Possible SNPs associated with lineage and carriage duration. 248
A.10 Maximum likelihood tree of *pspC* protein alignment 249
A.11 Maximum likelihood tree of *pspA* protein alignment 250
A.12 Maximum likelihood tree of *zmpC* protein alignment 250

A.13 PCA plots of classifiers used on antigen training data 251
A.14 Difference in Shannon diversity index of the ivr allele within-host 252
A.15 Comparison of PEER factors and BAPS clusters 253

List of Tables

2.1 Accuracy and resource usage of phylogenetic reconstruction methods . . 61
2.2 Results from SEER for antibiotic resistance 74
2.3 Comparison of SEER with results from existing methods 76

3.1 Success of culturing unencapsulated *S. pneumoniae* 86
3.2 Coefficients from lasso regression model of carriage duration 92
3.3 Sojourn times for each serotype in carriage 95
3.4 Locus effects on carriage duration 98
3.5 Summary of variance in carriage duration explained 105

4.1 Comparison of assembly method performance 109
4.2 Comparison of variant calling method performance 110
4.3 Comparison of classifiers of antigen alleles 115
4.4 Heritability of pneumococcal virulence due to pathogen genome 118
4.5 Common variation associated with invasiveness 123
4.6 Genes with Tajima's *D* differences between carriage and invasive isolates 127
4.7 Burden testing of rare variants associated with invasiveness 130
4.8 Paired samples from the MeninGene study 135
4.9 Genes enriched for mutation between blood and CSF in *S. pneumoniae* . . 140
4.10 Genes enriched for mutation between blood and CSF in *N. meningitidis* . 143
4.11 Intergenic regions enriched for mutation between blood and CSF in *N. meningitidis* 145
4.12 CDS enriched for mutation between carriage and invasion in *N. meningitidis* 149

5.1 Summary of cohorts with human genotype data 155
5.2 Human SNP heritability of meningitis phenotypes in Dutch adults 165
5.3 Signals of association in the Dutch cohort 166
5.4 Human SNP heritability of pneumococcal phenotypes in Danish children 170
5.5 Clusters tested in genome-to-genome analysis 180

Acronyms

AF allele frequency. 56, 57

AIC Akaike information criterion. 84, 85

ALF artificial life framework. 58, 71

AMP anti-microbial peptide. 21, 26

BAM binary sequence alignment/map. 111, 112

BAPS Bayesian analysis of population structure. 48, 54, 59, 61–63, 79, 180, 193

BFGS Broyden–Fletcher–Goldfarb–Shanno. 65, 66

CDS coding sequences. 138, 139, 142, 145

CFU colony forming unit. 135

CI confidence interval. 48, 118, 126

CMH Cochran–Mantel–Haenszel. 48, 54, 75, 79, 193

CNV copy number variant. 39, 108, 112, 124, 143

COG cluster of orthologous genes. 30, 46, 49, 55–57, 112, 120, 122, 188, 193

CPP closest phylogenetic-pairs. 118

CSF cerebrospinal fluid. 17–21, 23, 77, 108–110, 114, 117, 132, 134–152, 185, 186, 192, 196

CSV comma separated values. 176

d.f. degrees of freedom. 36, 56, 66

DSM distributed string mining. 55, 56, 71, 73

FWER family-wise error rate. 36, 67

GoNL The Genome of the Netherlands. 162

GOS Glasgow outcome score. 20, 118

GTR generalised time reversible. 58, 60

GWAS genome wide association study. 17, 33, 36–39, 41–52, 54, 55, 57, 63, 75, 77, 79, 81, 82, 98, 106, 108–112, 116, 119, 121, 125, 135, 147, 151, 152, 154–156, 161, 167–172, 183, 185–187, 189, 191, 193–197

H. influenzae *Haemophilus influenzae*. 20, 24, 155

HLA human leukocyte antigen. 43, 154, 175, 182

HMM hidden Markov model. 82, 84, 85, 95, 189

HPD highest posterior density. 147, 148

HRC haplotype reference consortium. 162, 163

HWE Hardy-Weinberg equilibrium. 158, 160, 162

ICE integrative conjugative element. 29, 31, 74, 91, 126, 129, 130

ICU intensive care unit. 156

IPD invasive pneumococcal disease. 18, 24

ivr inverting variable restriction. 32, 116, 117, 119, 131, 136, 146, 147, 252

JC Jukes-Cantor. 60

KC Kendall-Colijn. 60–62

L. monocytogenes *Listeria monocytogenes*. 20, 47, 58, 155

LD linkage disequilibrium. 30, 34–38, 42, 44–46, 49, 57, 73–75, 79, 88, 96, 98–100, 102, 161, 162, 164, 166, 177, 196

LMM linear mixed model. 39, 50, 86, 88–90, 93, 99, 102, 105, 120, 164, 187–189, 191, 193, 194, 246

LOD logarithm of odds. 33

LoF loss of function. 28, 39, 51, 124, 125, 128–131, 140, 141, 151, 191

LRT likelihood ratio test. 62, 66, 67, 90, 118, 164, 187

M. tuberculosis *Mycobacterium tuberculosis*. 43, 46, 47, 50, 128, 195

MAC membrane attack complex. 22, 27, 112

MAF minor allele frequency. 34, 36, 38, 39, 42, 55, 68, 71, 77, 96, 98, 99, 124, 128, 156, 158, 160–163, 165, 166, 170, 175–178, 180, 183

MCMC Markov-chain Monte Carlo. 116, 118, 132, 163

MDS multidimensional scaling. 63–65, 67, 68, 119, 176

MIC minimum inhibitory concentration. 93

MLST multi-locus sequence typing. 30, 47, 59, 61, 62, 108, 139, 143

MNP multiple nucleotide polymorphism. 110

MRCA most recent common ancestor. 58, 194

N. gonorrhoeae *Neisseria gonorrhoeae*. 66

N. meningitidis *Neisseria meningitidis*. 20, 21, 43, 46, 47, 99, 109, 135, 136, 138, 139, 142–146, 149, 150, 152, 155

NCD normalised compression distance. 60–62

NJ neighbour joining. 60–62

NT non-typable. 25, 31, 82, 85, 86, 90, 95

OR odds-ratio. 19, 45, 48, 49, 71, 72, 128, 165, 166, 170, 175, 178, 180, 183

OU Ornstein-Uhlenbeck. 118

pbp penicillin binding protein. 29, 49

PCA principal component analysis. 39, 115, 158, 178, 251

PCR polymerase chain reaction. 132, 147

PCV pneumococcal conjugate vaccine. 21, 31, 82, 195

PEER probabilistic estimation of expression residuals. 178–180

ply pneumolysin. 26, 195

QC quality control. 36, 109, 111, 155, 157, 160, 162, 163, 176

S. aureus *Staphlyococcus aureus*. 24

S. mitis *Streptococcus mitis*. 24, 58, 110

S. pneumoniae *Streptococcus pneumoniae*. 20–22, 24–28, 30–33, 43, 46, 48, 49, 56–58, 64, 67, 70, 71, 74, 75, 81, 86, 88, 95, 99, 105, 108–110, 121, 123, 135, 136, 138–140, 143–147, 149, 151, 152, 177, 179, 180, 185, 187, 189, 195, 196

S. pyogenes *Streptococcus pyogenes*. 9, 53, 56, 64, 70, 77–79, 187, 240

s.d. standard deviation. 113

SEER sequence element enrichment analysis. 53, 55, 59, 61–67, 69, 70, 74–77, 79, 81, 89, 102, 119, 120, 124, 176, 188, 193–195, 240

SFS site frequency spectrum. 56, 57, 124, 125, 141, 142, 157

SIR susceptible-infected-recovered. 195

SNP single nucleotide polymorphism. 31, 35–39, 46–49, 54–58, 61, 63, 70, 73, 75–79, 86–88, 91, 96–102, 108, 110, 111, 113, 120, 121, 137–139, 142, 145, 150, 156, 158, 161–166, 170, 175–177, 183, 188, 193, 194, 247, 248

SVM support vector machine. 115

VCF variant call format. 124, 176, 193

VEP variant effect predictor. 111, 128, 138

WHO World Health Organisation. 82