Functional and evolutionary analyses of pneumococcal genome variation

Nicholas Jason Croucher

Downing College

University of Cambridge

2011

This dissertation is submitted for the degree of Doctor of Philosophy

Declaration

I hereby declare that this dissertation is my own work and contains nothing that is the outcome of work done in collaboration with others, except where specifically indicated here or elsewhere in the text.

All sequence data were produced by the sequencing and research and development teams at the Wellcome Trust Sanger Institute. The annotation of complete genome sequences was performed in conjunction with Dr. Stephen Bentley (Wellcome Trust Sanger Institute, Cambridge). Phylogenetic analyses of bacterial genome sequences were performed in collaboration with Dr. Simon Harris (Wellcome Trust Sanger Institute, Cambridge). All microarray experiments and statistical analyses were performed by the BμG@S group (St. George's Hospital, London). All statistical analyses of Omnilog data were performed by Lars Barquist (Wellcome Trust Sanger Institute, Cambridge).

None of the work described herein has been previously submitted for the purpose of obtaining another degree. This dissertation does not exceed 60,000 words in length, as required by the School of Biological Sciences.

Nicholas Jason Croucher
August 2011

Acknowledgements

I thank Julian Parkhill for allowing me the opportunity to undertake this work and for all his help and advice, in particular his ability to patiently correct me when I have been stubbornly wrong. I also owe a huge debt of gratitude to Stephen Bentley, upon whom the burden of my supervision has fallen most heavily; despite this, he has remained unflinchingly positive, understanding and encouraging, and he has proved to be a fantastic, friendly and wise supervisor. Similarly, Nick Thomson has been a perennially enthusiastic and supportive manager; it has been a pleasure to work for all three. I also thank Duncan Maskell for his perceptiveness and advice as my external supervisor, and the other members of my advisory committee: Gordon Dougan, who has also invested much effort in helping with my laboratory work, and Alex Bateman, who has organised the graduate students at the Sanger Institute so effectively.

Much of my informatics work could not have been completed without the help of Simon Harris and Matthew Holden, who have been two of the most friendly, conscientious and helpful people I could possibly hoped to have shared an office with. Thomas Otto also deserves much credit for shouting and singing me into being a better systems user. I am also grateful to Theresa Feltwell and Sally Whitehead for their tolerance of me, and the disruption I inevitably cause, in running their respective laboratories. I thank Maria Fookes and Del Pickard for their vast microbiological expertise and general helpfulness, as well as their unendingly friendly and amusing ways, along with Trevor Lawley, who also taught me a very practical approach to laboratory etiquette. I am also thankful to all of teams 15 and 81 for their patience and understanding, with a special mention of congratulations to the ever entertaining and hospitable Alan ‘Aunt’ Walker. Of course, none of this work could have been completed without the help of the informatics, systems, sequencing and library making teams of the Sanger Institute, which has been a brilliant place to work.

Externally, I have enjoyed many productive discussions with Bill Hanage and Christophe Fraser of Imperial College, London. Gavin Patterson, of the Vetinary School, Cambridge, was very kind in initially providing me with strains and advice on laboratory techniques. Tim Mitchell, of the University of Glasgow, has been a friendly and invaluable collaborator. In addition, I am greatly appreciative of the kindness of all our other collaborators who have provided us with relevant strains, which have been essential to this project.

I also thank my family, in particular my mum, my dad and my grandparents who have been so supportive and encouraging throughout these seven years at Cambridge, and beyond. The friends I have made during this time have also been central in making the bad times survivable and the good times so enjoyable. My final thanks are to Alina, who has had to demonstrate more patience and understanding than could reasonably be expected of anyone, and has made me so happy.

p.s. Mum, to save you flicking through to the end: no, I still haven't cured anything.

Abstract

Functional and evolutionary analyses of pneumococcal genome variation

Nicholas Jason Croucher

Streptococcus pneumoniae (the pneumococcus) is a human nasopharyngeal commensal and respiratory pathogen responsible for a high burden of morbidity and mortality worldwide. The bacterium's primary virulence factor appears to be its polysaccharide capsule, of which there are more than 90 different serologically-distinguishable types (serotypes). Although this categorisation was originally used for tracing pneumococcal epidemiology, the bacterium is naturally transformable, and hence is able to switch serotypes through horizontal exchange of capsule biosynthesis (*cps*) gene clusters. Therefore, following the emergence of multidrug-resistant lineages in the late 1970s, superior, multilocus-based typing schemes were devised for following pneumococcal evolution. Increasing antibiotic resistance also motivated the development of a heptavalent conjugate polysaccharide vaccine, which targeted seven *S. pneumoniae* serotypes, leading to a decrease in pneumococcal disease. However, this impact has been ameliorated by an increase in disease resulting from replacement by non-vaccine serotypes and switching of *cps* loci by strains previously expressing vaccine serotypes. This thesis describes the application of second-generation sequencing technologies to investigating the mechanisms by which the pneumococcus evolves, especially in response to such clinical interventions.

The first part concerns the Pneumococcal Molecular Epidemiology Network clone 1 (PMEN1) lineage, one of the first multidrug-resistant pneumococcal genotypes to become a worldwide problem. Complete sequencing of the *S. pneumoniae* ATCC 700669 type strain, combined with draft sequencing of a global collection of 240 isolates, quantified the impact of recombination across the chromosome, as well as revealing the diversity of conjugative elements and prophage in the population. The acquisition of antibiotic resistances and the evasion of the conjugate polysaccharide vaccine were both evident in among the strains. *In vitro* transformation experiments, in the same genetic background, were then used to perform a more detailed investigation of the types of homologous recombination events seen in the global population.

The second part of this dissertation describes the use of RNA sequencing to investigate the functional consequences of genomic variation. A novel method was developed and validated, and, when applied to *S. pneumoniae* ATCC 700669, revealed a family of expressed putative coding sequences that were formed by extended forms of the BOX interspersed repeat. This technique was also applied to two closely related strains of the PMEN31 lineage, both isolated from a single case of disease. This allowed the functional consequences of a small number of distinguishing polymorphisms on the global transcriptome to be ascertained, providing an insight into the level of pneumococcal evolution that can occur within an individual. Sequencing further members of this lineage showed that, although highly successful, this lineage has a much more static genotype than that of PMEN1.

The different mechanisms of pneumococcal genome variation are associated with evolution over different timescales, and in response to different selection pressures, but clearly interact in a number of ways. Hence the use of whole genome sequencing, surveying all the variation throughout the chromosome, will be crucial for greater understanding, and therefore improved control, of this important pathogen.

Contents

Declaration ii

Acknowledgements iii

Abstract v

Contents vii

List of Figures xii

List of Tables xiv

Abbreviations xv

1 Introduction 1

1.1 The history and biology of the pneumococcus 1

1.1.1 "*Une maladie nouvelle*" 1

1.1.2 Taxonomic classification 2

1.1.3 Metabolic and microbiological characteristics 3

1.1.4 Microbiological identification 5

1.1.5 Pneumococcal serology 5

1.1.6 A transformable pathogen 8

1.2 Interactions with the host and microbiota 8

1.2.1 The pneumococcus as a human commensal 9

1.2.2 Competition within the nasopharynx 10

1.2.3 The pneumococcus as a respiratory pathogen 11

1.2.4 Risk factors for disease 12

1.2.5 Epidemiology of pneumococcal disease 13

1.2.6 Structures involved in pathogenesis 15

1.3 Pneumococcal genomics and epidemiology 18

1.3.1 The genome of *S. pneumoniae* TIGR4 18

1.3.2 Functional genomics experiments 19

1.3.3 Variation uncovered by dideoxy terminator sequencing 20

1.3.4 Epidemiological typing techniques 21

1.4 The emergence of antibiotic-resistant pneumococci 24

1.4.1 Early observations of resistance 24

1.4.2 Resistance to β lactams 25

1.4.3 Resistance to other antibiotics 27

1.4.4 The spread of the PMEN clones 29

1.5 Horizontal sequence exchange in the pneumococcus 30

1.5.1 The pneumococcal competence system 30

1.5.2 Transduction and pneumophage 32

1.5.3 Conjugative elements 33

1.6 Anti-pneumococcal vaccines 36

1.6.1 Early whole cell vaccines 36

1.6.2 Polysaccharide vaccines 36

1.6.3 Conjugate polysaccharide vaccines 37

1.7 The impact of second-generation sequencing technologies 40

1.7.1 Dideoxy terminator sequencing 41

1.7.2 Second-generation sequencing technologies 41

1.7.3 Bacterial population genomics 45

1.7.4 Bacterial RNA-seq 47

1.8 Summary 48

2 Materials and Methods 50

2.1 Culturing and transforming *S. pneumoniae* 50

2.1.1 Culturing of strains 50

2.1.2 Sampling of PMEN1 and ST180 isolates 50

2.1.3 Transformation experiments 51

2.1.4 Omnilog experiments 51

2.2 Extraction and analysis of nucleic acids 52

2.2.1 Genomic DNA extractions 52

2.2.2 RNA sample extractions 52

2.2.3 PCR and RT-PCR 53

2.3 Construction of mutant strains 54

2.3.1 Construction of *S. pneumoniae* TIGR4^{PUS} 54

2.3.2 Construction of *S. pneumoniae* ATCC 700669Δ*hexB* 55

2.4 DNA and RNA sequencing 56

2.4.1 Genome sequencing 56

2.4.2 Transcriptome sequencing 58

2.4.3 Oligonucleotide sequencing 58

2.5 Alignment and assembly of short sequence reads 59

2.5.1 Generation of whole genome alignments for phylogenetic analyses 59

2.5.2 Generation of a genome alignment for *in vitro* transformation analysis .60

2.5.3 Assembly of Illumina data 60

2.5.4 Detecting accessory genome components through mapping 61

2.5.5 Analysis of RNA-seq data..61

2.6 Bioinformatic analyses 61

2.6.1 Mathematical analyses 61

2.6.2 Annotation of sequences 61

2.6.3 Extraction of serotype and sequence type..62

2.6.4 Recombination and phylogenetic analyses 62

2.6.5 Analysis of gene disruption events..63

2.6.6 Bayesian phylogenetic analyses 63

2.6.7 Alignment of assembled sequences...64

2.6.8 Clustering of prophage elements 64

2.6.9 Analysis of repeat sequences...64

2.6.10 Prediction of RNA secondary structures..66

2.6.11 Analysis of *in vitro* transformation data...66

2.6.12 Differential gene expression and phenotype analyses 68

3 The genome of *S. pneumoniae* ATCC 700669 70

3.1 Introduction..70

3.1.1 The spread of the PMEN1 lineage...70

3.1.2 *S. pneumoniae* ATCC 700669 71

3.2 Description of the genome 71

3.2.1 Features of the chromosome...71

3.2.2 Genes implicated in pathogenesis...73

3.2.3 Prophage...74

3.2.4 ICE*Spn*23FST81 76

3.2.5 An ICE-derived genomic island 78

3.2.6 Pneumocidins...81

3.3 Discussion 83

4 The evolution of the PMEN1 lineage...86

4.1 Introduction..86

4.1.1 Global collection of PMEN1 strains..86

4.1.2 Detecting recombination in sequence data...86

4.2 Analysis of the PMEN1 population...89

4.2.1 Construction of the phylogeny 89
4.2.2 Recombination and antigenic variation...94
4.2.3 Population and serotype dynamics 97
4.2.4 Resistance to non–β-lactam antibiotics...102
4.2.5 Components of the accessory genome..107
4.3 Discussion 113
5 *In vitro* transformation of *S. pneumoniae* ATCC 700669 117
5.1 Introduction..117
5.2 Analysis of *in vitro* transformants 119
5.2.1 Genome-wide exchange of sequence between pneumococci...............119
5.2.2 Characterisation of 'capsule switching' recombinations 123
5.2.3 Mosaic recombinations in the *cps* locus...124
5.2.4 Analysis of secondary recombinations 126
5.2.5 Efficiency of polymorphism transfer 128
5.2.6 The role of mismatch repair 130
5.3 Discussion 131
6 A simple method for directional RNA-seq 136
6.1 Introduction..136
6.2 Description and validation of the RNA-seq methodology..........................136
6.2.1 Illumina sequencing libraries can be generated from ss-DNA.............136
6.2.2 ss-cDNA sequencing retains data directionality.................................141
6.2.3 ss-cDNA sequencing is quantitative 143
6.3 Discussion 144
7 Analysis of pneumococcal small interspersed repeats 145
7.1 Introduction..145
7.2 Analysis of small interspersed repeat sequences..147
7.2.1 Three families of pneumococcal repeats..147
7.2.2 Genomic distribution of pneumococcal repeats 150
7.2.3 Mobility of pneumococcal repeats...153
7.2.4 Repeat sequences in other streptococci..154
7.2.5 Genes affected by repeat element insertions.......................................155
7.2.6 Expressed open reading frames generated by large BOX elements......157
7.3 Discussion 160
8 The evolution of serotype 3 ST180 pneumococci..168

8.1 Introduction....168
8.2 Genetics of the serotype 3 ST180 population....169
8.2.1 The genome of *S. pneumoniae* OXC141....169
8.2.2 The phylogeny of ST180....170
8.3 Comparison of *S. pneumoniae* 99-4038 and 99-4039....177
8.3.1 Transcriptional profiles of *S. pneumoniae* 99-4038 and 99-4039....177
8.3.2 Phenotypic differences between *S. pneumoniae* 99-4038 and 99-4039.178
8.3.3 The mechanism of *patAB* upregulation....181
8.4 Discussion....182
9 Discussion....190
9.1 Pneumococcal transformation....190
9.1.1 Biases in the detection of recombination....190
9.1.2 Impact of biases on *r/m* estimates....192
9.1.3 The advantage of being transformable....193
9.1.4 Inferences from PMEN1 and ST180....195
9.2 Site-specific and homologous recombination....196
9.2.1 The characteristics of horizontal sequence transfer....196
9.2.2 Transformation as bacterial 'gene therapy'....197
9.2.3 The interaction of mobile elements and transformation....199
9.2.4 Criticism of this model....201
9.2.5 Comparing the evolutionary dynamics of PMEN1 and ST180....202
9.3 Concluding remarks....203
Appendix I: Primer sequences....204
Appendix II: PMEN1 strains....207
Appendix III: EMBL accession codes....213
Appendix IV: Serotype 3 strains....214
References....217

List of Figures

Figure 1.1 A timeline of pneumococcal research 35
Figure 1.2 A summary of the SGST methods..44
Figure 2.1 Construction of *S. pneumoniae* ATCC 700669Δ*hexB* 55
Figure 3.1 Circular diagram representing the *S. pneumoniae* ATCC 700669 chromosome. 72
Figure 3.2 Representation of ICE*Sp*23FST81 75
Figure 3.3 Comparison of streptococcal integrative and conjugative elements 77
Figure 3.4 Comparison of PPI-1 of ATCC 700669 with ICE*Sp*23FST81 78
Figure 3.5 Alignment of PPI-1 sequences from pneumococcal genome data............80
Figure 3.6 Lantibiotic synthesis gene clusters and predicted structures 82
Figure 3.7 Comparison between *blp* loci...84
Figure 4.1 Phylogeography and sequence variation of PMEN1................................94
Figure 4.2 Construction of the PMEN1 phylogeny 96
Figure 4.3 Robustness of the PMEN1 phylogeny..97
Figure 4.4 Histogram showing the variation of recombination event lengths............98
Figure 4.5 Population dynamics of PMEN1 in the USA 99
Figure 4.6 Recombinations causing changes in serotype.......................................100
Figure 4.7 Variation in the serogroup 19 *cps* loci...101
Figure 4.8 Presence of macrolide-resistance cassettes in Tn*916*.............................103
Figure 4.9 Distribution of macrolide resistance cassettes among PMEN1 strains. ...104
Figure 4.10 Deletions affecting ICE*Sp*23FST81 107
Figure 4.11 The putative ICE ICE*Sp*11876...108
Figure 4.12 The putative ICE ICE*Sp*11930...109
Figure 4.13 The putative ICE*St*1-like element ICE*Sp*8140....................................110
Figure 4.14 Prophage sequences in PMEN1 111
Figure 4.15 Distribution of prophage sequences between PMEN1 isolates 112
Figure 4.16 Competence of isolates carrying prophage inserted in *comYC*.............114
Figure 5.1 Distribution of transformation events across the genome 119
Figure 5.2 Structure of transformation events within the primary locus 122
Figure 5.3 Distribution of transformation event sizes...125
Figure 5.4 Frequencies of SNPs within transformation events 127

Figure 5.5 Knock out of *hexB* gene in *S. pneumoniae* ATCC 700669$^{lab}\Delta$*hexB* 129
Figure 5.6 Distribution of transformation events in the absence of MMR 130
Figure 5.7 Recombination sizes in PMEN1 isolates and *in vitro* transformants.......132
Figure 6.1 Example of the data produced by sequencing ss-cDNA and ds-cDNA...137
Figure 6.2 Potential mechanisms for the attachment of adapters to ss-cDNA..........139
Figure 6.3 Studying the attachment of adapters to an oligonucleotide.....................141
Figure 6.4 Comparison of RPKMs from ss-cDNA and ds-cDNA sequencing 143
Figure 7.1 HMM logos representing pneumococcal interspersed repeats 147
Figure 7.2 Predicted repeat sequence secondary structures 149
Figure 7.3 Distribution of repeat sequences within the pneumococcal genome. 150
Figure 7.4 Distribution of repeat sequences relative to CDS functional classes.......151
Figure 7.5 Repeat sequence variation between pneumococcal genomes.................153
Figure 7.6 Repeat sequence expression congruent with genome annotation............157
Figure 7.7 Potential misannotation of the *S. pneumoniae* ATCC 700669 genome...159
Figure 8.1 Circular diagram representing the *S. pneumoniae* OXC141 genome 170
Figure 8.2 Phylogenetic analysis of serotype 3 isolates..171
Figure 8.3 Distribution of detected recombination lengths in serotype 3 172
Figure 8.4 Heatmap showing the distribution of prophage sequences......................174
Figure 8.5 Activity of prophage ϕOXC141..176
Figure 8.6 Distribution of accessory genome loci 179
Figure 8.7 Expression of the *patAB* locus 180
Figure 8.8 Leader sequence upstream of *patAB* 181
Figure 8.9 The effect of the PUS. 183
Figure 8.10 Model to explain the differences caused by the PUS 184
Figure 9.1 Modelling the ability to detect recombinations and mutations................191
Figure 9.2 Model for the role of transformation in pneumococci 199

List of Tables

Table 4.1 Convergence of the PMEN1 phylogeny 115

Table 4.2 CDSs frequently disrupted by mutations in the PMEN1 phylogeny.........116

Table 5.1 Association of indels with *in vitro* recombination events........................135

Table 7.1 Distribution of repeat elements relative to CDSs...................................164

Table 7.2 Association of repeat sequences with *in vitro* recombination events........165

Table 7.3 Frequency of repeats in streptococcal genome sequences. 166

Table 8.1 Microarray analysis of differential expression between *S. pneumoniae* 99-4038 and 99-4039..185

Table 8.2 CDSs found to be significantly more highly expressed in *S. pneumoniae* 99-4039 than in *S. pneumoniae* 99-4038 using RNA-seq. 186

Table 8.3 CDSs found to be significantly less highly expressed in *S. pneumoniae* 99-4039 than in *S. pneumoniae* 99-4038 using RNA-seq. 187

Table 8.4 Comparison of *S. pneumoniae* 99-4038 and 99-4039 using phenotype microarrays 188

Table 8.5 Microarray analysis of differential expression between *S. pneumoniae* TIGR4 and TIGR4^{PUS}. 189

Abbreviations

aa	Amino acid
ABC	ATP-binding cassette
ATCC	American type culture collection
ATP	Adenosine triphosphate
BAC	Bacterial artificial chromosome
BHI	Brain heart infusion
BLAST	Basic local alignment search tool
bp	Base pair
CBP	Choline-binding protein
CC	Clonal complex
cDNA	Complementary DNA
CDS	Coding sequences
CGH	Comparative genome hybridisation
CSF	Cerebrospinal fluid
CSP	Competence stimulating peptide
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleoside triphosphate
ds	Double stranded
DUS	DNA uptake sequence
EDTA	Ethylenediaminetetraacetic acid
EMBL	European Molecular Biology Laboratories
FR	Flanking region
GC	Guanine and cytosine
GMP	Guanosine monophosphate
HIV	Human immunodeficiency virus
HMM	Hidden Markov model
ICE	Integrative and conjugative element
IPD	Invasive pneumococcal disease
IS	Insertion sequence
LB	Luria broth
MCS	Multiple cloning site

MEPS	Minimum efficiently processed segment
MGE	Mobile genetic element
MITE	Miniature inverted repeat transposable element
MLEE	Multilocus enzyme electrophoresis
MLST	Multilocus sequence typing
MMR	Mismatch repair
MR	Mosaic recombination
mRNA	Messenger RNA
nt	Nucleotide
NTP	Nucleoside triphosphate
ORF	Open reading frame
PBP	Penicillin binding protein
PCR	Polymerase chain reaction
PCV	Polysaccharide conjugate vaccine
PEP	Phosphoenolpyruvate
PFGE	Pulsed field gel electrophoresis
PMEN	Pneumococcal molecular epidemiology network
PPI-1	Pneumococcal pathogenicity island 1
PTS	Phosphotransferase system
PUS	*patAB* upregulatory SNP
rDNA	Ribosomal DNA
RNA	Ribonucleic acid
RPKM	Reads per kilobase per million mapped reads
rRNA	Ribosomal RNA
RSS	Recombined sequence segment
RT-PCR	Reverse transcription PCR
RUP	Repeat unit of pneumococcus
SGST	Second generation sequencing technology
SNP	Single nucleotide polymorphism
ss	Single stranded
ST	Sequence type
TCA	Tricarboxylic acid
TIR	Terminal inverted repeats
tRNA	Transfer RNA

TSD	Target sequence duplication
USS	Uptake signal sequence
UTR	Untranslated region
UV	Ultraviolet
VNTR	Variable number tandem repeat