Investigating the variability of virulence in natural populations of *Staphylococcus aureus* using whole genome sequencing

Bernadette Young
Wolfson College
University of Oxford
September 2017

A thesis submitted for the degree of Doctor of Philosophy, Trinity 2017

Abstract

Invasive *Staphylococcus aureus* disease is an important cause of morbidity and mortality, but is much rarer than asymptomatic carriage. The contribution of the bacterial genome to *S. aureus* infection is incompletely understood, and molecular epidemiology provides conflicting evidence. This thesis aims to improve our understanding of this contribution using the resolution afforded by whole-genome sequencing.

In a systematic study of *S. aureus* evolution in 105 hosts during invasive *S. aureus* disease, I demonstrate extensive within-host diversity, with evidence for varying selective pressures and within-host adaptation. Evidence for adaptation is strongest in genes under control of transcriptional regulatory systems, including Repressor of surface proteins (Rsp) ($p=10^{-6.4}$) – a recently discovered global virulence regulator – and the Accessory gene regulator (Agr) ($p=10^{-5.6}$), which are enriched for protein-altering variants 3.6- and 2.9-fold respectively. The development of invasive disease is associated with subtle changes in the transcriptional regulation of *Staphylococcus aureus* arising within hosts.

Applying recently developed tools for bacterial genome-wide association studies (GWAS), I present GWAS investigating for *S. aureus* genomic associations with two disease phenotypes: bacteraemia and pyomyositis. A study of *S. aureus* bacteraemia and carriage in 2001 isolates from the United Kingdom shows bacteraemia is not strongly bacterially determined: no lineages, genes or variants were significantly associated with bacteraemia.

In a study of 518 isolates from pyomyositis and carriage in Cambodian children, the presence of Panton-Valentine leukocidin (PVL) genes increases the odds of pyomyositis 130-fold ($p=10^{-18}$), and variation in these genes and an adjacent promoter region are sufficient to explain over 99.9% of the heritability of pyomyositis. These results establish staphylococcal pyomyositis, like tetanus and diphtheria, as a disease depending critically on expression of a toxin.

Microbial genomics offers unparalleled opportunities to understand infections, and here I demonstrate insights generated through pathogen evolution within hosts and bacterial GWAS.

Table of Contents

Abstract ii

1 Preface 1
1.1 Acknowledgements 1
1.2 Funding 2
1.3 Ethics statements 3
1.4. Declaration and attributions by chapter 4
1.4.1 Chapter 4 attributions 4
1.4.2 Chapter 5 attributions 5
1.4.3 Chapter 6 attributions 6
1.5 Publications 7
1.5.1 Publications arising from this thesis 7
1.5.2 Other publications during the period of this degree 7
1.5.3 Presentations 8
1.5.4 Posters 8
1.6 Abbreviations used in this thesis 9

2 Introduction 14
2.1 *Staphylococcus aureus* 14
2.1.1 *Staphylococcus aureus* microbiology 14
2.1.2 *Staphylococcus aureus* genome typing, population and diversity 14
2.2 *Staphylococcus aureus* carriage 17
2.3 *Staphylococcus aureus* disease 18
2.3.1 Spectrum of *S. aureus* infections in humans 18
2.3.2 Toxin mediated disease 18
2.3.3 *S. aureus* skin and soft tissue infection 19
2.3.4 *S. aureus* bacteraemia 19
2.3.5 Other invasive *S. aureus* infection 21
2.3.6 Hospital and community associated *S. aureus* bacteraemia 23
2.3.7 Role of *S. aureus* carriage in disease 24
2.4 *Staphylococcus aureus* virulence 27
2.4.1 Virulence factors in *S. aureus* 28
2.4.2 Virulence variation in *S. aureus* 30
2.5 *Staphylococcus aureus* and the host-pathogen relationship 36
2.5.1 Human immune system and host-pathogen interaction 36
2.5.2 Virulence may not be a "Red Queen's Race" 38
2.6 Within-host evolution as a source of altered virulence 39
2.7 Bacterial genome-wide association studies for investigation of virulence 44
2.7.1 GWAS for studying natural populations 44
2.7.2 Challenges for bacterial GWAS 45
2.7.3 Progress in bacterial GWAS 48
2.8 Thesis outline 51

3 Methods 68
3.1 Microbiological methods 68
3.1.1 Nasal swabs 68
3.1.2 Clinical specimens 70
3.1.3 DNA extraction 71
3.1.4 Toxin Elisa 72

3.2 Bacterial whole genome sequencing **73**
3.2.1 Short read sequencing 73
3.2.2 Mapping and *de novo* assembly of short read sequencing 74
3.2.3 Variant identification and annotation 76
3.2.4 Multi-locus Sequence Typing and Antimicrobial resistance prediction 78
3.3 Population genetics **78**
3.3.1 Phylogenetic tree-building 78
3.3.2 Ancestral state reconstruction 79
3.3.3 Pairwise genetic diversity 80
3.3.4 dN/dS within patient populations 80
3.4.5 omegaMap analysis for species wide d_N/d_S 81
3.4 Gene set enrichment analysis **81**
3.5 Genome-wide association study in bacteria **83**
3.5.1 Bacterial GWAS of SNPs and kmers using LMM to control for population structure 83
3.5.2 Testing for lineage associations 85
3.5.1 Quantifying association with Odds ratio 85
3.6 Model averaging with the use of harmonic mean *p*-values **86**
3.7 Accounting for multiple testing **86**

4 Within host evolution of *Staphylococcus aureus* in invasive disease reveals signatures of adaptation **91**
4.1 Introduction **91**
4.2 Materials and Methods **93**
4.3 Results **94**
4.3.1 Extensive diversity exists during *S. aureus* disease, with colonising and invasive isolates forming related but distinct clades 94
4.3.2 Variants arising within-host show relaxed purifying selection and an increase in stop codons occurring during infection 97
4.3.3. B variants are enriched for changes to transcriptional regulators 98
4.3.4 Variants separating colonising from invasive isolates are enriched for changes in gene regulators and in the cell surface proteins under their control 102
4.3.5 Adaptation occurs in multiple pathways in parallel across sites during infection 104
4.3.6 Adaptive signatures of Rsp, Agr and adhesins are unique to within-host evolution in infected patients 106
4.4 Discussion **112**

5 *Staphylococcus aureus* bacteraemia is not strongly determined by the bacterial genome at a population level **118**
5.1 Introduction **118**
5.2 Materials and methods **120**
5.2.1 Sampling frame for carriage controls 120
5.2.2 Sampling frame for bacteraemia cases 123
5.2.3 Epidemiological data 124
5.2.3 Ethics statements 126
5.2.4 Multiple testing 126
5.3 Results **127**
5.3.1 Global diversity of lineages found in *S. aureus* bacteraemia and carriage 127
5.3.2 No individual SNPs are found in significant association with bacteraemia 132
5.3.3 Kmer GWAS of SAB shows strong batch effects arising from refinements in sequencing technology 136
5.3.4 Antibiotic resistance-conferring mutations and genes are associated with SAB in a kmer GWAS 142

5.3.5 Cryptic population structure exists in *S. aureus* carriage, and specific variants are associated with healthcare-associated carriage. 147
5.3.6 SNPs in MRSA-associated mobile genetic elements and healthcare-associated foldase variants are associated with "community" associated SAB 154
5.3.7 Total heritability of *S. aureus* bacteraemia is low, and mostly accounted for by differences in MRSA rates. 158
5.3.8 Analysis of recent variants shows enrichment in the GraS regulon in *S. aureus* isolated from the bloodstream 159
5.4 Discussion 165

6 Genome-wide association study reveals the bacterial determinants of pyomyositis caused by *Staphylococcus aureus* 177
6.1 Pyomyositis 177
6.2 Materials and Methods 179
6.2.1 Pyomyositis strains 179
6.2.2 Carriage strains 179
6.2.3 Whole genome sequencing 180
6.2.4 Genome-wide association study 181
6.2.5 Toxin quantification 182
6.3 Results 182
6.3.1 Pyomyositis prevalence varies strongly by bacterial lineage 182
6.3.2 Kmers associated with disease reveal strong locus association with pyomyositis 185
6.3.3 Changes in PVL expression are associated with high risk kmers. 191
6.4 Discussion 194

7 Conclusions and future work 200
7.1 Within-host evolution of *S. aureus* is a promising avenue for understanding invasive disease 200
7.1.1 Summary of key findings 200
7.1.2 Implications and Future work 201
7.2 Two *S. aureus* virulence GWAS produce dramatically different results 204
7.2.1 Summary of key findings 204
7.2.2 Implications and future work 205
7.3 Going beyond the bacterium 208
7.4 Conclusion 209

Appendix 1 213

1 Preface

1.1 Acknowledgements

I would like to thank all the patients whose participation has enabled this research to take place. I'm deeply appreciative of their time and generosity.

I owe an enormous debt to my supervisors Derrick Crook and Daniel Wilson, as well as my sponsor Tim Peto. Derrick and Tim took a chance on a young Australian doctor a decade ago by giving me a clinical trial to run. They introduced me to both the exciting and challenging world of medical research, and a department full of great scientists, which has since evolved into the Modernising Medical Microbiology group. Both have continued to challenge my thinking about microbes and disease. Danny has been quite endlessly patient in teaching me about population biology and statistical genetics, and I've greatly enjoyed our long discussions about the whys and wherefores of these cunning bacteria.

I have been taught, aided and guided by an army of scientists and clinicians in the MMM. Sarah Walker has been an invaluable source of knowledge and wisdom in epidemiological methods. Martin Llewelyn has provided helpful insight in discussion of results. David Wyllie has been enthusiastic in moving the findings from genomic studies into the laboratory, and he has been a patient teacher in my forays into molecular microbiology.

Numerous collaborators have helped this project to succeed, and I've very grateful to all them. Kyle Knox, Ruth Miller and others through their work on the Oxfordshire *Staphylococcus aureus* carriage study have created an enormous resource for understanding *S. aureus*. Matt Scarborough and Rob Tilley, through their continuing enthusiasm helped construct large strain collections of *S. aureus*

bacteraemia. Catrin Moore worked over many years and despite obstacles to amass a unique collection for the study of pyomyositis.

I've greatly enjoyed my time as a DPhil student, not least because of the friends and colleagues who make such a wonderful workplace. My colleagues sharing "the registrar room" have supplied laughter and help in equal measure, and I owe much to Claire, Nick, Nicole, Ana, David, Tim and Leon. I'm grateful for the friendship and assistance of Sarah Earle and Jessie Wu – especially for their help when my R woes were overwhelming – and to Ali Vaughn and Betty Coles for absolutely everything.

Finally, none of this could have happened without the loving support of my family. To Toby, who has given endless support, love and patience through this D.Phil and my long student career, I can never give sufficient thanks. This makes, by my count, the eighth university degree between us over the last 18 years, and I'm grateful for a partner who values lifelong learning. I also thank Rose, who teaches me daily about the wide-eyed joy of discovery, and has shared me with this work for literally her entire life.

Perhaps fittingly, for a thesis on the power of heritability, I would like to dedicate this work to my parents. To Steve Young, whose unfailing love and support have been a lighthouse my whole life long, and to Kathleen Young, whose works were cut short: I think she'd be quite proud of this part of her legacy.

1.2 Funding

While undertaking my DPhil studies I have been supported by a Wellcome Trust Research Training Fellowship (Grant 101611/Z/13/Z).

This research was supported by the Oxford NIHR Biomedical Research Centre, a Mérieux Research Grant, the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at Oxford University in partnership with Public Health England (PHE) (grant HPRU-2012-10041), and the Health Innovation Challenge Fund (a parallel funding partnership between the Wellcome Trust (grant WT098615/Z/12/Z) and the Department of Health (grant HICF-T5-358)).

The study in chapter 6 was funded by the Wellcome Trust (MORU Grants 089275/H/09/Z and 089275/Z/09/Z) and the University of Oxford Medical Research Fund (MRF/MT2015/2180).

1.3 Ethics statements

Ethical approval for sequencing *S. aureus* isolates from routine clinical samples and linkage to patient data without individual patient consent in Oxford and Brighton in the U.K. was obtained from Berkshire Ethics Committee (10/H0505/83) and the U.K. National Information Governance Board [8-05(e)/2010].

Data about *S. aureus* bacteraemia in Oxfordshire, Brighton and Plymouth and bacterial isolates from these cases were collected for evaluations of clinical service provision. The UK National Research Ethics Service reviewed this data collection protocol, and deemed it was an evaluation of service, and therefore did not require ethics committee review.

The study in chapter 6 was approved by the Angkor Hospital for Children institutional review board and the Oxford Tropical Network Ethics committee [507-12]

1.4. Declaration and attributions by chapter

I, Bernadette Young, designed and conducted the analyses presented in this thesis, with the support of my supervisors and colleagues.

This work exists at an intersection between medical microbiology, epidemiology and statistical genetics, and as such arises from collaboration between numerous people with complementary expertise. A detailed description of this work, and assistance where it was received, follows below.

1.4.1 Chapter 4 attributions

I conceived of this study in discussion with Professor Derrick Crook and Professor Tim Peto, and designed a pilot study. The larger sampling frame with designed in conjunction with Dr John Paul, Prof Martin Llewelyn, Prof Crook, Prof Peto, Prof Sarah Walker and Prof Daniel Wilson.

I identified 50 paired samples of *S. aureus* from carriage and infection through the clinical microbiology laboratory at OUH, and further pairs were contributed by Dr James Price in from Royal Sussex Country Hospital, Brighton (n=9) and Dr Claire Gordon at OUH (n=46).

I performed the majority of sub-culture of isolates and preparation of DNA extracts for sequencing, with assistance from Dr Gordon, and MMM research assistants (Sanuki Perera, MMM Oxford; Elian Liu, MMM Oxford; Kevin Cole, MMM Brighton). DNA library preparation and whole genome sequencing was performed by WTCHG, Oxford. Processing of reads through a bioinformatics pipeline of read mapping and *de novo* assembly was performed by the MMM Bioinformatics team.

I performed *de novo* assembly on isolates in the study using Cortex, and identified variants using the Cortex variant caller, with technical advice from Dr Zam Iqbal. I identified mapping-based variants and built haplotrees using R scripts written by Prof Wilson. I undertook all variant annotation.

Dr Chieh-Hsi Wu built the phylogenetic tree of the study isolates along with a reference panel of *S. aureus* genomes, from which she constructed the MRCA sequence for each patient. I used this tree and MRCA data to determine ancestral states for all within-host variants, and identify nearest neighbour isolates.

Prof Wilson designed a gene set enrichment analysis method in discussion with Dr David Wyllie, and wrote an R script to execute this method. I applied this GSEA method to all variant sets described in the chapter and analysed the results. Prof Wilson designed the layout of Table 4.3. Prof Wilson undertook the OmegaMap analysis and performed an adjusted GSEA incorporating species wide d_N/d_S information and variant data I supplied. He drew figures 4.7 and 4.8.

I drafted a manuscript based on these findings in conjunction with Prof Wilson, and in the light of feedback received from co-authors and reviewers. This has been lodged in a pre-print server and is currently under peer review.

1.4.2 Chapter 5 attributions

I conceived of this study in discussion with Dr John Paul, Prof Martin Llewelyn, Prof Derrick Crook, Prof Tim Peto, Prof Sarah Walker and Prof Daniel Wilson.

Professor Sarah Walker developed the sampling frame for controls from available collections. I developed the sampling frame for cases under Prof Walker's supervision. I collected epidemiological data from a number of existing research databases as well as infection control reporting data.

DNA extraction was performed by me and MMM research staff Elian Liu, Laura Dunn, Kevin Cole and Ali Vaughn. Illumina sequencing was performed at the WTCHG.

Dr Chieh-Hsi Wu performed GWAS on provisional case control sets of cases and controls, in which she tested the performance of kmers generated directly from reads versus those from *de novo* assembly. Dr Wu also developed a method for filtering kmers arising from a phage which was introduced by the WTCHG to some samples.

I performed GWAS on a final set of 2001 cases and controls and two subsets of the collection and analysed the results.

1.4.3 Chapter 6 attributions

Dr Catrin Moore conceived of the study in discussion with Professor Nicholas Day and collected the clinical samples from pyomyositis and bacteraemia and associated epidemiological data, in collaboration with Prof Nick Day, director of the Mahidol-Oxford Tropical Medicine Research Unit (MORU). She selected carriage isolates from a study by Dr Emma Nickerson and supervised a further carriage cohort study with AHC staff.

Dr Moore extracted DNA and prepared it for whole genome sequencing. Illumina sequencing was carried out at the WTCHG. Read data was processed on the MMM sequencing pipeline by the MMM bioinformatics team.

The GWAS on this population was performed in collaboration between Dr Moore, Prof Wilson and I, and we formulated analysis plans in discussion together. I constructed phylogenies, performed association testing and interpreted the results. Sarah Earle, a doctoral student in Prof Wilson's group, provided technical support

in running the bacterial-GWAS scripts.

I developed an ELISA protocol and performed toxin quantification under the supervision of Dr David Wyllie.

I drafted a manuscript for publication based on these findings, which has been revised after extensive comments from Dr Moore and Prof Wilson.

1.5 Publications

1.5.1 Publications arising from this thesis

Young BC, Wu CH, Gordon NC, Cole K, Price JR, Liu E, Sheppard A, Perera S, Charlesworth J, Golubchik T, Iqbal Z, Bowden R, Massey R, Paul J, Crook DW, Peto TEA, Walker AS, Llewelyn M, Wylie DH, Wilson DJ. Severe infections emerge from commensal bacteria by adaptive evolution. Elife. 2017 Dec 19;6. pii: e30637. doi:10.7554/eLife.30637.

Das S, Lindemann C, Young BC, Muller J, Österreich B, Ternette N, Winkler AC, Paprotka K, Reinhardt R, Förstner KU, Allen E, et al. Natural mutations in a *Staphylococcus aureus* virulence regulator attenuate cytotoxicity but permit bacteremia and abscess formation Proc Natl Acad Sci U S A. 2016. 113(22):E3101-10.

1.5.2 Other publications during the period of this degree

Young BC, Votintseva AA, Foster D, Godwin H, Miller RR, Anson LW, Walker AS, Peto TE, Crook DW, Knox K. Multi-site and nasal swabbing for carriage of *Staphylococcus aureus*: what does a single nose swab predict? *J Hosp Infect*. 2017 S0195-6701(17)30059-2.

Laabei M, Uhlemann AC, Lowy FD, Austin ED, Yokoyama M, Ouadi K, Feil E, Thorpe HA, Williams B, Perkins M, Peacock SJ, Clarke SR, Dordel J, Holden M, Votintseva AA, Bowden R, Crook DW, Young BC, Wilson DJ, Recker M, Massey RC. Evolutionary Trade-Offs Underlie the Multi-faceted Virulence of *Staphylococcus aureus*. PLoS Biol. 2015 Sep 2;13(9).

1.5.3 Presentations

B. Young, E. Liu, C. Gordon, M. Llewelyn, J. Paul, T. Peto, D. Crook, S. Walker, D. Wilson "Invasive *Staphylococcus aureus* disease is associated with genomic changes arising within hosts" ECCMID 2016

B. Young, C. Gordon, D. Crook, T. Peto, J. Paul, M. Llewelyn, S. Walker, D. Wilson "Role of within-host evolution in *Staphylococcus aureus* infection" ECCMID 2016

B. Young, K. Cole, C. Wu, H. Seifert, S. Rieg, L. Lopez-Cortes, M. Gurguí, J. Lepe, H. Kim, W. Park, R. Tilley, M. Scarborough, J. Edgeworth, M. Llewelyn, D. Wilson, A. Kaasch "Polymorphisms in coagulase are associated with *Staphylococcus aureus* cardiac device infections" ECCMID 2017

1.5.4 Posters

Young B, Sona S, Poda S, Kumar V, Vuthy S, Songly H, Lyda L, Bousfield R, Emary K, Stoesser N, Parry CM, Day NPJ, Wilson D, Moore CE, "Panton-Valentine Leucocidin positive CC121 strain Methicillin-sensitive *Staphylococcus aureus* drives Pyomyositis in Cambodian children" ECCMID 2017

1.6 Abbreviations used in this thesis

Agr, accessory gene regulator

AHC, Angkor Hospital for Children

Bbp, bone sialoprotein binding protein

bp, base pairs

BSA, Bovine Serum Albumin

BSUH, Brighton and Sussex University Hospitals

CA, community associated

CAMP, cationic antimicrobial peptide

CC, Clonal Complex

CF, Cystic Fibrosis

ClfA and ClfB, Clumping factor A/B

CLSI, Clinical and Laboratory Standards Institute

CWA, cell-wall anchored proteins

DNA, deoxyribonucleic acid

ELISA, enzyme-linked immunosorbent assay

ECM, extracellular matrix

FnbA, FnbB, Fibronectin binding protein A/B

FWER, Family wide error rate

GBP, Great British Pounds

GWAS, genome-wide association study

HA, Healthcare associated

HGT, horizontal gene transfer

HLA, human leucocyte antigen

HMP, harmonic mean p-values

Ig, Immunoglobulin

Indel, insertion and deletion

kb, kilobase

LMM, linear mixed models

M, Molar

Mb, megabase

mcg, microgram

mg, milligram

MGE, mobile genetic elements

MHC, major histocompatibility complex

ml, millilitre

ML, maximum likelihood

MLST, Multi-locus sequence type

MMM, Modernising Medical Microbiology

MRSA, Methicillin resistant *Staphylococcus aureus*

MSCRAMM, Microbial surface component recognising adhesive matrix molecules

MSSA, Methicillin susceptible *Staphylococcus aureus*

MORU, Mahidol-Oxford Tropical Medicine Research Unit

NCBI, National Center for Biotechnology Information

NET, neutrophil extracellular traps

ng, nanogram

NHS, National Health Service

nm, nanometres

OR, odds ratio

OUH, Oxford University Hospitals

PBS, phosphate-buffered saline

PBP, penicillin-binding protein

PC, principle component

PCA, principle component analysis

PCR, polymerase chain reaction

PFGE, Pulsed-field gel electrophoresis

PWID, People who inject drugs

PVL, Panton-Valentine leucocidin

RNA, ribonucleic acid

Rsp, Repressor of surface proteins

SAB, *Staphylococcus aureus* bacteraemia

SAMMD, *Staphylococcus aureus* microarray meta-database

Sbi, staphylococcal binder of immunoglobulin

SCCmec, staphylococcal cassette chromosome *mec*

SCV, small colony variant

SdrC, SdrD, SdrE, Serine-aspartate repeat protein C/D/E

SNP, single nucleotide polymorphism

Spa, Staphylococcal protein A

SSSS, Staphylococcal scalded skin syndrome

SSTI, skin and soft tissue infection

ST, Sequence type

TSB, Tryptic soy broth

TSS, Toxic shock syndrome

µl, microlitre

µm, micrometre

UK, United Kingdom

USA, United States of America

WGS, whole-genome sequencing

WTCHG, Wellcome Trust Centre for Human Genetics

Chapter 2

2 Introduction

2.1 *Staphylococcus aureus*

2.1.1 *Staphylococcus aureus* microbiology

Staphylococcus aureus was first described as a micrococcus by Alexander Ogston in 1880, when he found the Gram-positive spherical bacteria while examining abcesses.[1] It was later named *S. aureus* in virtue of the characteristic golden colour of its colonies, by Frederich Rosenbach in 1884.[2] This facultative anaerobe is a frequent coloniser of humans, and other species: in humans, *S. aureus* is found most frequently in the anterior nares, as well as on the skin and the mucosa of the oropharynx.[3] It can be identified in the laboratory by growth in salt rich media, beta-haemolysis, the presence of deoxyribonuclease, and positive tests for catalase and coagulase.[4,5] Identification can be confirmed by the presence of protein A and clumping factor on the bacterial surface, evidenced by agglutination when mixed with latex beads coated with their respective ligands immunoglobulin G and fibrinogen.[4,5]

2.1.2 *Staphylococcus aureus* genome typing, population and diversity

Distinguishing between strains of *S. aureus* is crucial to understanding population dynamics, as well as epidemiological surveillance and outbreak investigation.[7] Prior to the advent of whole-genome sequencing (WGS), the method with the greatest

resolution between strains was pulsed field gel electrophoresis (PFGE) of DNA. DNA is fragmented using restriction enzymes, and these fragments separated by PFGE to produce a profile of fragments for comparison.[7] This method yields greater power to differentiate between strains than phage typing or typing based on capsular antibody binding.[8]

Chain termination sequencing technology, also known as Sanger sequencing, made comparison of DNA sequence possible. DNA typing schemes for *S. aureus* compared the DNA sequence of short regions of the genome. Allele typing fragments of seven relatively conserved housekeeping genes is used for Multi-Locus Sequence Typing (MLST)[2]. Strains with matching sequence type (ST) by MLST can be further clustered on the basis of partial matching into Clonal Complexes (CC), from which some inference can be made of ancestral relationships between strains.[10,11] Strain differentiation is also possible by sequencing a single gene, including the polymorphic repeat X region of the Staphylococcal Protein A (Spa) gene.[12] Sanger sequencing of this mini-satellite region (*spa*-typing) gives slightly reduced discrimination than PGFE,[13] but with greater speed and reproducibility. For methicillin resistant *S. aureus* (MRSA), typing can be performed based on PCR of the staphylococcal cassette chromosome *mec* (SCCmec).[14]

The population structure of *S. aureus* as revealed by these tools has been characterised as strongly clonal, with limited recombination: when closely related strains are examined in housekeeping genes, point mutation events are up to 15 times more common than recombination.[11] However comparison of housekeeping and adhesin genes between CCs reveals recombination occurs more frequently between lineages than within CCs.[15]

Whole genome sequencing of bacteria by Sanger sequencing is relatively expensive per base.[16] Parallel sequencing of very large numbers of short segments of DNA to

sequence whole genomes ("second generation" or "next-generation" sequencing) developed in 2005 has facilitated the generation of exponentially increasing volumes of sequence data for a given cost.[17] Since the first *S. aureus* genome was sequenced in 2001, an increasing number of *S. aureus* genomes from a range of human and animal sources have been characterised by WGS[18-32], yielding a wealth of new information about the *S. aureus* genome and population structure. The *S. aureus* genome is typically between 2.7 and 2.9 Mb in length, with up to 3000 open reading frames,[20] and a GC content of approximately 33%.[18]

Comparison of these whole genome sequences reveals extensive diversity in the genomic content of *S. aureus*.[33] Lindsay and Holden described the "core" genome – genes shared between strains – as well as an extensive array of variably present or "accessory" genes.[34] Accessory genes comprise around 25% of the genome on average and include determinants of antibiotic resistance, substrate use, metabolic functions and putative virulence determinants.[33] Study of the conservation of genes across the *S. aureus* species reveals an open pangenome: additional protein coding sequences continue to be found as more sequences are studied, as well as groups of genes dubbed "pseudocore" – present usually but not universally – breaking down the binary categories of core and accessory genes.[35]

Genes which are variably present are frequently found in association with mobile genetic elements (MGEs), and are therefore able to move via horizontal gene transfer (HGT).[34] MGEs in *S. aureus* include plasmids, phages, insertion sequences, transposons, pathogenicity islands and chromosome casettes.[33,36] *S. aureus* genomic diversity thus arises both by the accumulation of mutations and homologous recombination within core and accessory regions of the genome, and non-homologous movement of the accessory genome, where homologous and non-homologous HGT may be mediated via transformation, transduction and

conjugation of genetic material.[36] Studying across the species using WGS reveals extensive homologous recombination occurs within the core genome between lineages, particularly in regions adjacent to the integration of MGEs.[37]

2.2 *Staphylococcus aureus* carriage

In humans *S. aureus* is frequently found asymptomatically carried, forming part of the microbiome.[3,38] The most common site of carriage is the anterior nares,[3] though carriage in the throat, axilla and perineum are also common.[3,40] Point prevalence of *S. aureus* nasal carriage is estimated to be around 30% in cross-sectional studies.[3,41] Carriage is not static over time, and gain and loss of carriage are frequent events.[41,42] Strain typing of *S. aureus* identified in carriage also demonstrates that carried lineages can change during persistent and intermittent carriage, and the rates at which they change varies with carrier age and bacterial clonal complex.[41] Changing strain and carriage of multiple strains is common: of 166 individuals who carried *S. aureus* continuously throughout a 24 month period, only 109 (65.7%) carried a single *spa*-type throughout; 31 (18.7%) carried different *spa*-types in series, and 29 (17.5%) had a mixture of *spa*-type on at least one occasion.[43] Frequent sampling suggests that in young adults a new strain may be carried as often as every 8 weeks.[44]

Asymptomatic *S. aureus* carriage is a reservoir for transmission. *S. aureus* nasal carriage in both patients and healthcare workers can be a source of transmission in the healthcare setting – either directly between individuals[45] or through the environment[46] – and transmission between individuals in the community facilitates the spread of endemic clones.[47,48] Studies using the greater resolution of WGS have both confirmed both sporadic transmission and clonal outbreaks.[49,50]

2.3 *Staphylococcus aureus* disease

In addition to being a common commensal, *S. aureus* is a major human pathogen responsible for a broad range of human diseases.[51,52]

2.3.1 Spectrum of *S. aureus* infections in humans

Infection with *S. aureus* ranges from superficial and troublesome to invasive and life threatening (Figure 2.1).

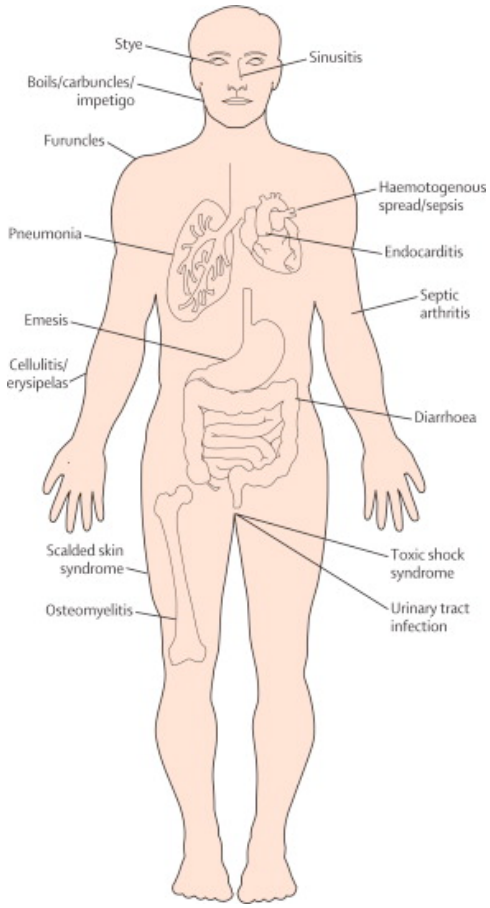


Figure 2.1: A summary of the diversity of *S. aureus* disease. Figure from Wertheim et al.[67]

2.3.2 Toxin mediated disease

S. aureus disease may result from the ingestion of pre-formed, heat stable toxin (gastroenteritis following *S. aureus* food-poisoning), or the production of toxins by *S. aureus in vivo* (Staphylococcal scalded skin syndrome (SSSS) and Toxic shock

syndrome (TSS)).[51] In TSS, an overwhelming cytokine release from by human T-cells occurs in response to stimulation by a toxin (e.g. toxic shock staphylococcal toxin 1 (TSST-1)) acting as a superantigen. This cytokine storm causes profound hypotension, fever and diffuse erythroderma.[53] In SSSS, desquamation occurs either locally or extensively, with blistering, skin loss, and dysregulation of fluid balance and temperature akin to severe burns.[53]

2.3.3 *S. aureus* skin and soft tissue infection

The commonest forms of *S. aureus* disease are superficial skin and soft tissue infections (SSTI). National survey data estimates that in the USA over 2001-3, more than 11 million ambulatory care visits were made annually with SSTI.[54] These include impetigo, cutaneous abscesses (e.g. boils, carbuncles and furunculosis), non-purulent cellulitis and surgical wound infection.[52,53] These infections may be self-limiting, but usually require treatment with incision and drainage of abscesses, antibiotics for non-purulent disease, or a combination of both in complicated infections.[52] Bacterial strains causing SSTI also appear to be efficiently transmitted: household members of children with MRSA SSTI have MRSA carriage rates 5-14 times higher than the general population.[55]

2.3.4 *S. aureus* bacteraemia

While superficial infections are most common, *S. aureus* disease also arises following bacterial ingress and proliferation in the bloodstream and deep tissues. *S. aureus* bacteraemia (SAB) is one of the commonest bloodstream infections.[56,57] In the UK, SAB causes a large and continually increasing burden of disease, with over 12,000 cases a year, 22.4 cases per 100,000 population per year, of which 6.7% are caused by MRSA (Figure 2.2).[58] In the UK, rates of MRSA bacteraemia rose at the end of the 20th century, but have declined dramatically over the last decade.[58] MRSA rates vary globally,[56] but this decline in MRSA bacteraemia rates has been reported

from multiple countries.[52]

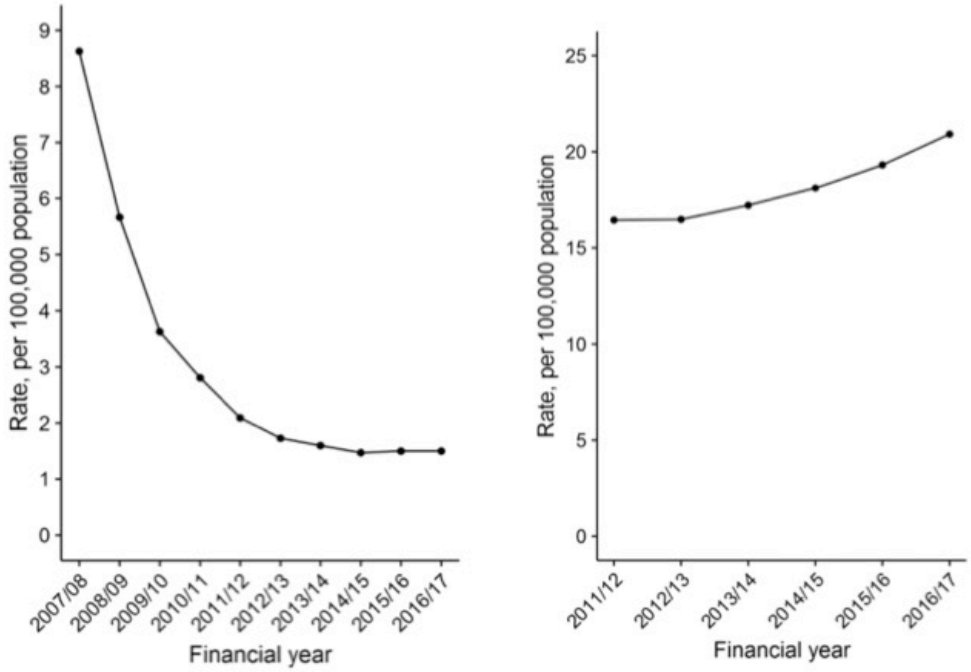


Figure 2.2: Rates of (A) methicillin-resistant and (B) methicillin-susceptible SAB per 100,000 population in England over periods of mandatory surveillance. Figures from Public Health England.[58]

Prospective studies of SAB report that the commonest portals for entry of *S. aureus* to the bloodstream are peripheral cannulas, and skin wounds or infections, but that in up to one third of cases, no route of entry for bacteria to the bloodstream is apparent on careful clinical assessment.[59] Sub-clinical skin wounds or occult infection may account for these cases, but this remains to be demonstrated.[60]

In addition to vulnerabilities which allow a portal of entry, a number of host risk factors are recognised for SAB, including age and gender. Population surveys show a bimodal distribution of risk with age: a peak occurs in those under 1 year old, then risk reduces to low levels in adolescence and young adulthood, before rising in middle age and continuing to increase with age thereafter.[58] SAB occurs in men with a rate reported 1.5-2 times that seen in women.[52,61,62] Rates vary with ethnicity, within country incidence being higher in the black population in the USA,

indigenous populations in Australia, and Maori and Pacific Island peoples in New Zealand.[52] Immunosuppression, including diabetes mellitus, is also a recognised risk factor.[52,61,62] Two populations are particularly at risk of SAB: people who inject drugs (PWID), who have a combination of regular vascular access, as well as higher rates of *S. aureus* carriage and soft tissue infection; and renal dialysis patients, in whom age, multifactorial immunosuppression and regular vascular access all combine to increase the risk of SAB.[52]

SAB continues to have a serious overall mortality rate. Rates are variably reported between 15-50%[52], with variation between country and centre.[61,62,63] Overall mortality has not decreased over recent decades.[52] Host factors impact on outcomes: mortality is consistently higher with age.[52,62,63,64] There is conflicting evidence between studies for the impact on mortality of healthcare-associated disease and MRSA infection.[52,61,62,63,64,65] This heterogeneity might be explained by the extensive variation in total mortality and clinical management observed between centres.[61,62]

Mortality also varies according to the focus of infection, and is lowest where a removable focus is identified.[52,62] Conversely, where a non-removable focus is identified the hazard ratio for inpatient mortality is double that seen when the only focus is a removable intra-vascular catheter,[61] while a respiratory focus has a hazard ratio 4.5 greater for mortality at 30 days.[62] Intriguingly, multiple prospective studies report significantly increased mortality in those with SAB in whom no focus can be identified.[61,62]

2.3.5 Other invasive *S. aureus* infection

S. aureus bacteraemia may occur alone or in association with other deep infections. SAB causes metastatic deep tissue invasion: circulating bacteria seed to deep

tissues in over 40% of cases, even with appropriate antimicrobial therapy.[60] Deep bacterial collections can also provide an on-going source of bacteraemia.[51]

Circulating bacteria may adhere to and infect heart valves, particularly prosthetic or damaged valves, and *S. aureus* has replaced viridans streptococci as the most common cause of infectious endocarditis (IE) in developed countries.[66] The relationship between SAB and IE is so strong that echocardiography to inspect the heart valves is routinely advised in SAB if a definite, removable source is not identified.[61]

The deep and uncommon soft tissue infections necrotizing fasciitis and pyomyositis can both be caused by *S. aureus*, with extensive tissue damage, inflammation and sepsis.[52,68] Osteo-articular infections of all forms are predominantly caused by *S. aureus*.[52] Staphylococcal osteomyelitis can be categorised into three groups using the Waldvogel classification: those arising from direct invasion of bacteria from superficial to deep tissues; infection in tissues with insufficient vascular supply; or infections arising from haematogenous seeding of bacteria to bones and joints.[69] Osteomyelitis in native bone is most commonly located in the vertebra of adults and long bones of children.[52] *S. aureus* is also a major cause of both native and prosthetic joint infections;[51,52] in the latter *S. aureus* biofilm formation makes treatment particularly challenging.[70]

Biofilms are thin layers of microbial cells, attached to a substrate or each other, and an extracellular matrix.[71] Bacteria existing in biofilms exhibit specific phenotypes of metabolism, growth rate and gene expression. *S. aureus* biofilms derive initially from expression of cell surface adhesins facilitating attachment, following which toxins and immune evasion proteins are expressed. *S. aureus* can grow in biofilms as small colony variants (SCV), small, slow-growing cells with reduced turnover of the targets of many antimicrobial agents. Biofilm formation thus allows *S. aureus* to

exist in relative protection from both host immunity and antimicrobial therapy.[71]

Biofilm formation is implicated in the frequency of *S. aureus* related medical device infections.[52] Cardiac devices (e.g. permanent pacemaker, implantable cardiac defibrillator), prosthetic heart valves, vascular grafts and urinary or intra-vascular catheters may all be infected with *S. aureus*. This infection may arise from direct spread from colonising flora or SSTI (e.g. intravascular catheter infection, pacemaker pocket infection), or seeding from bacteraemia (e.g. late prosthetic valve endocarditis, pacemaker wire infection).[51,52,66]

S. aureus is an important but uncommon cause of severe community acquired pneumonia, including secondary infection following influenza.[72] A severe, necrotizing, primary form of pneumonia is recognised, which carries high mortality.[73] *S. aureus* is also an important cause of pneumonia in individuals with cystic fibrosis (CF),[52] in hospital-associated pneumonia and ventilator associated pneumonia,[74] reflecting the ability of *S. aureus* to proliferate in abnormal airways.

2.3.6 Hospital and community associated *S. aureus* bacteraemia

The epidemiology, outcomes and microbiology of *S. aureus* disease, particularly bacteraemia, differ according to whether or not the disease is healthcare associated (HA).[62,75,76] Traditionally bloodstream infections have been deemed either community-acquired (CA) if developing before hospital admission or within 48 hours of admission to hospital, and nosocomial or "hospital acquired" if developing more than 48 hours after admission.[77] A third category is also recognised, of patients whose disease onset is prior to or early in admission, but who have significant recent hospital exposure. These infections share many of the features that distinguish hospital-acquired from CA bacteraemia: they occur in patients of older age and more co-morbidities, and are associated with longer length of stay

and mortality.[77] MRSA was previously a common cause of predominantly hospital-acquired disease, but the changing epidemiology of MRSA obscures this association. In the UK, MRSA bacteraemia rates have declined both overall and as a proportion of healthcare-acquired and healthcare-associated disease.[58] Globally, the spread of CA-MRSA lineages, including the USA300 strain (a successful ST8 MRSA clone responsible for a high burden of community acquired MRSA in the USA) as well as the implication of these strains in healthcare-associated disease have further blurred these distinctions.[78] In this thesis, I use the term "healthcare-associated" (HA) to include both hospital-acquired and healthcare-associated carriage and disease.

2.3.7 Role of *S. aureus* carriage in disease

Like many other pathogens, *S. aureus* is frequently a commensal organism: invasive disease is a relatively rare event compared with carriage. The prevalence of asymptomatic *S. aureus* nasal carriage is 30%[3] while invasive disease rates are between 22-31 cases per 100,000.[58,79] While disease, particularly superficial purulent infection, may facilitate transmission,[47,55] invasive disease is unnecessary[80] for transmission and may even be disadvantageous.[81]

Carriage and invasion are highly related: nasal carriage increases the risk of invasive disease,[67] and a prospective study of individuals colonized with *S. aureus* on admission to hospital demonstrated that individuals with carriage had a risk ratio for subsequent bacteraemia three times that of non-carriers.[82] Comparisons of nasal and bloodstream isolates in cross-sectional and prospective studies have demonstrated that carried and invasive bacteria are closely related: in 80-86% of infections blood and nasal isolates show identical PGFE profiles.[82,83] This relationship is the target of therapeutic interventions: nasal decolonisation successfully reduces both the incidence of post-surgical *S. aureus* wound

infections,[84] and of all bloodstream infections during and after intensive care unit admission.[85]

In addition to being an interventional target, this relationship also has the potential to provide insights into the origins and mechanisms of *S. aureus* infection. Increased availability of WGS has facilitated new insights into the bacterial population dynamics of this relationship. Sampling the transition from *S. aureus* carriage to bacteraemia in one individual, and comparing this with stable carriage over time in two other individuals, my collaborators and I previously reported dramatically narrowed diversity in bacteraemia compared with stable carriage (Figure 2.3).[86] The transition from carriage to disease was also associated with a burst of premature stop codons, in excess of rates observed in stable carriage in this individual and two others over time. These stop codons included one in a transcriptional regulator SAR2468, which was later characterised as Repressor of surface proteins.[87] These findings were significant for demonstrating that while *S. aureus* invasive and carriage populations in this individual were closely related, they were genetically distinct. Investigating the extent to which the findings of this case study generalise to *S. aureus* infections in general forms an important part of the motivation of this thesis.

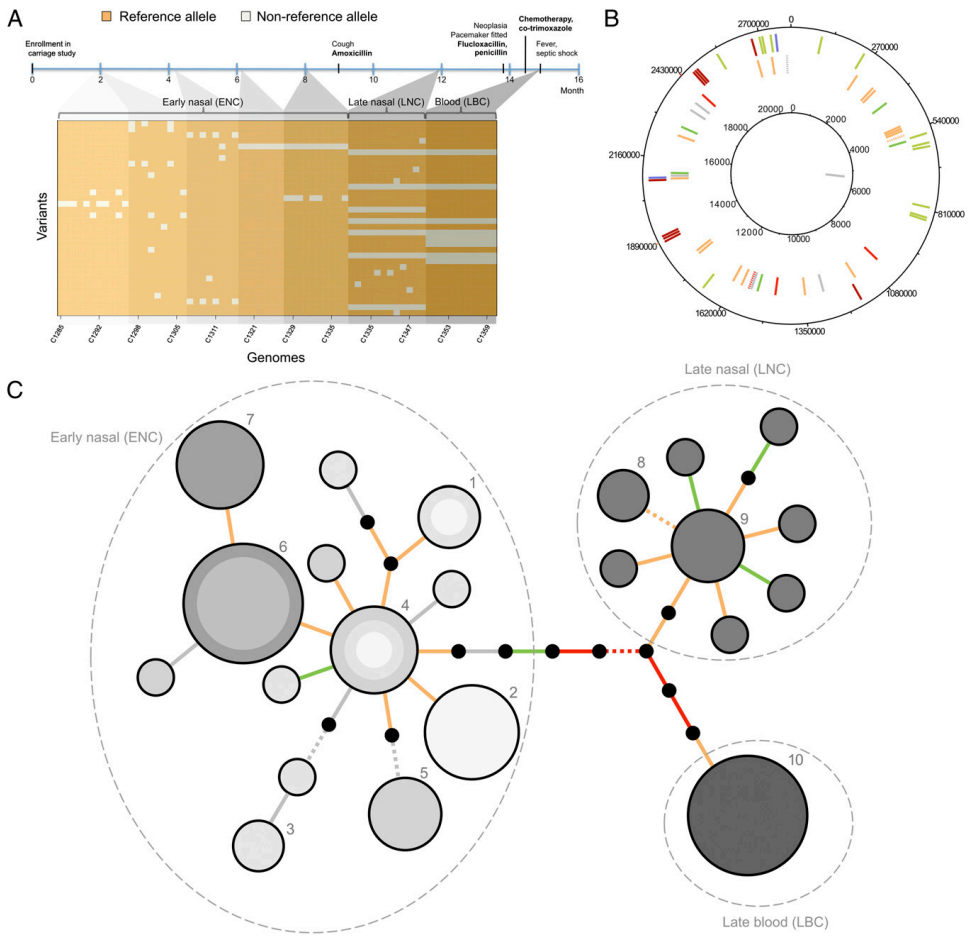


Figure 2.3: The genomic diversity during progression from carriage to disease in an individual (patient P). (A) Sampling frame, and temporal relationship of disease to carriage. **(B)** Distribution of variants on the chromosome of reference genome MSSA476, with MSSA476 putative virulence determinants marked. **(C)** Maximum-likelihood tree built from all sequences identified from Patient P. The size of nodes is proportional to the number of isolates with shared genotype, and the shading of nodes represents the time of isolation from earliest (white) to latest (dark grey). Black nodes are unobserved intermediate genotypes. Edges represent mutations, shaded by predicted on protein product: synonymous (green), non-synonymous (orange), premature stop codon (red), non-coding (grey). Solid edges are single point mutations, and dashed edges are indels. Figure from Young et al.[86]

Another study demonstrated the diversity of bacterial genomes during an outbreak of MRSA in a veterinary surgery, with multiple cases of nasal carriage and one case of invasive canine disease. Again, invasive isolates were closely clustered and harboured mutations distinguishing them from all other isolates found in carriage during the outbreak (Figure 2.4). In contrast carriage isolates from multiple sites in the infected individual and from asymptomatic individuals showed overlap of genome sequence found at multiple carriage sites and in multiple individuals.[88]

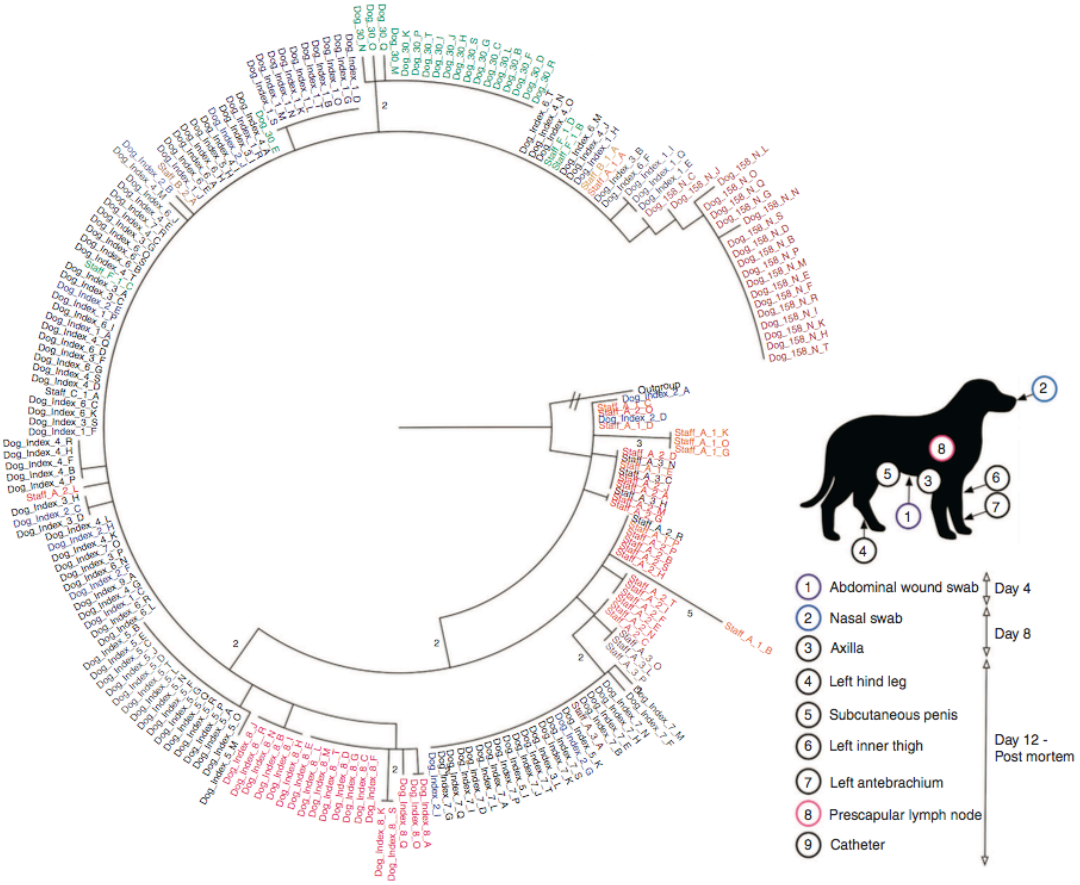


Figure 2.4: Maximum likelihood phylogeny generated from core genome SNPS in isolates from an MRSA outbreak in a veterinary setting. Numbers above branches represent the number of SNPs. Index case (with invasive disease sampled from a lymph node) represented graphically. Pink isolates are from invasive disease. (Figure from Paterson et al.[88])

These exploratory studies demonstrate the close relationship between *S. aureus* carriage and invasive disease, and suggest that fine-scale differences distinguish invasive isolates. Genomic studies of carriage and disease are thus a promising avenue for better understanding invasive *S. aureus* disease, but to date no systematic investigation has followed these preliminary analyses.

2.4 *Staphylococcus aureus* virulence

In the previous section I discussed how changes in virulent behaviour may be driven by variants arising within hosts. *S. aureus* virulence may also be altered by alleles occurring at high frequencies in the population.

2.4.1 Virulence factors in *S. aureus*

Staphylococcus aureus is a pathogen with an armory of factors facilitating tissue invasion, inflammation and evasion of host immune factors, including a thick peptidoglycan wall, polysaccharide capsule[51], toxins[89], complement control proteins[90] and bound adhesins[91]. The identification of bacterial virulence factors has traditionally relied upon comparing wild type and 'knock out' strains for their phenotype *in vitro* or in animal models of disease.[92] More recently WGS has allowed comparative genomics to identify homologs of established virulence factors, and contrast strains and species with different phenotypes.[20] These approaches have several limitations: they typically study genes in isolation, which may exist as part of complex processes, in systems that only simplistically model the host-pathogen interaction.[93]

Among the *S. aureus* toxins are those associated with specific toxinosis.[94] TSST-1 and staphylococcal enterotoxins B and C (SEB, SEC) are the causative agents in Toxic Shock Syndrome (TSS).[94,95] A range of enterotoxins cause gastroenteritis.[94] The exfoliative toxins A and B (ETA, ETB) cause Staphylococcal Scalded Skin Syndrome (SSSS).[96] Further secreted toxins include cytolytic toxins (haemolysins alpha-, beta- gamma- and delta-toxins and leucocidins LukDE, LukAB and Panton-Valentine leukocidin (PVL)) and a range of secreted enzymes including non-specific proteases, staphylokinase, coagulases and lipases.[97] These secreted toxins damage host tissues and aid in bacterial evasion and modulation of both the innate and adaptive human immune responses.[90,94,99]

Surface bound molecules have been identified with an array of virulence related functions. Microbial surface components recognising adhesive matrix molecules (MSCRAMMs) are bound to the *S. aureus* surface, contain immunoglobulin G (IgG)-like domains, and are capable of binding multiple host ligands, including fibrinogen,

fibronectin, keratin, epithelial cells and complement.[91] These proteins include collagen adhesin (Cna); fibronectin binding proteins A and B (FnbA, FnbB); clumping factors A and B; serine-apartate repeat (Sdr) proteins C, D and E; and the SdrE isoform bone sialoprotein-binding protein (Bbp).[91] Another cell wall anchored (CWA) protein Staphylococcal Protein A (Spa) binds the Fc portion of IgG, as does the lipoteichoic acid anchored staphylococcal binder of immunoglobulin (Sbi).[100] Collectively, CWA proteins have been implicated in biofilm formation, adhesion, nasal colonisation, epithelial invasion and immune evasion.[90,91,101] Additional immune evasion genes include surface expressed proteins such as Map (also known as Eap), which is an analogue of the human Major Histocompatibility Complex (MHC) class II protein,[102] and secreted proteins, including staphylococcal complement inhibitor (Scn).[103]

The expression of this multitude of virulence factors varies with regulatory control. *S. aureus* has over 100 conserved transcriptional regulators which regulate metabolic as well as virulence processes.[104] One of the best characterised is the quorum-sensing global transcriptional regulator Accessory Gene Regulator (Agr)(Figure 2.5).[105] Agr acts in response to increasing density of *S. aureus* bacteria by increasing expression of RNAIII, which mediates a switch in transcriptional activity away from surface proteins responsible for adhesion, towards toxins and other genes that result in tissue destruction.[106] This switch is critical during biofilm formation and maturation.[71] Variation of Agr function has been hypothesised to play a role in the emergence of the outbreak MRSA strain USA300 in the USA.[107] Negative regulation of virulence determinants is also observed: for example, CodY responds to a nutritionally constrained environment by negatively regulating expression of virulence genes.[108] Both of these regulators act within a complex network of regulatory systems.[104,109] *S. aureus* thus is able to vary expression through the course of an infection, as well as responding to environmental and

bacterial population changes by altering gene transcription and expression.[110] There is evidence that the effects of transcriptional regulators may differ between strains.[111] Four Agr groups have been defined, which vary across clonal complexes.[112]

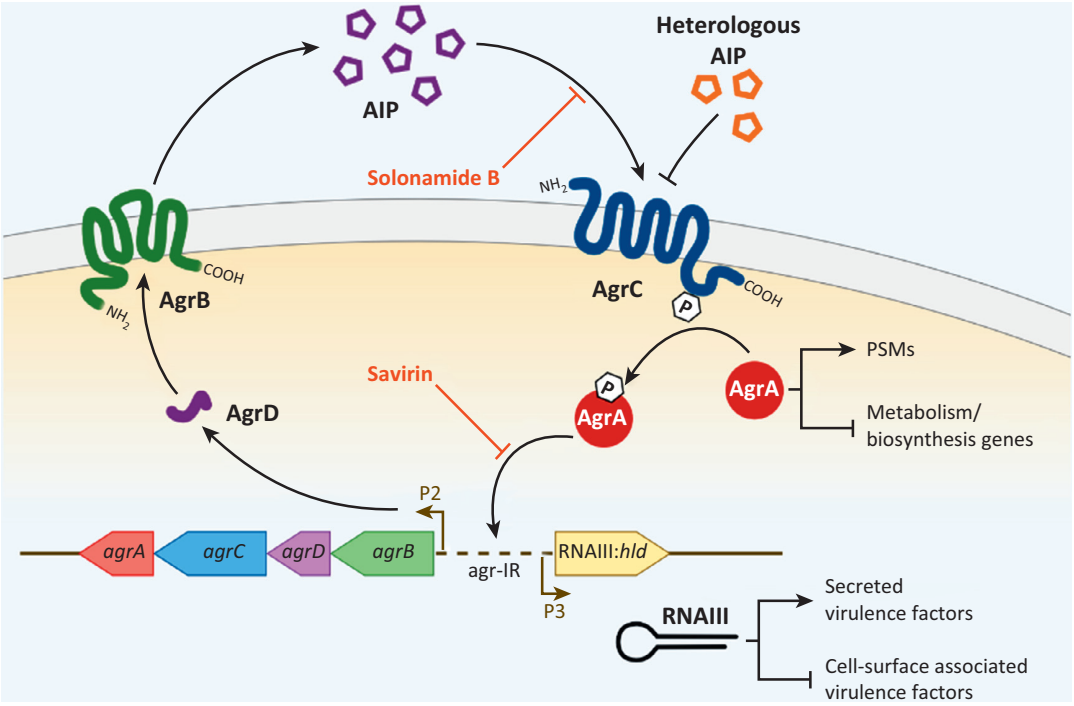


Figure 2.5: The Accessory Gene Regulator quorum-sensing system. Autoinducing peptide (AIP) is produced and secreted by *S. aureus* via AgrD and AgrB. This pheromone is sensed by the two-component system AgrC and AgrA. In response to AgrA phosphorylation, RNAIII is transcribed with downstream effects on gene transcription. AgrA also regulates production of phenol-soluble modulins (PSMs). Figure from Painter et al.[105]

2.4.2 Virulence variation in *S. aureus*

While many virulence factors which are thought to make *S. aureus* such a fearsome pathogen have been catalogued, the presence or absence of many of these in *S. aureus* infections has not been systematically established. The demonstrated requirement for specific toxin gene presence and toxin-mediated diseases (TSS[95], SSSS[96] and food-poisoning[94]) are the exception rather than the rule. A number of bacterial factors have been associated with *S. aureus* causing disease, though evidence from multiple studies is discordant.[113]

Virulence determinant	Reported effect (if any)	Reported association	Evidence for association	Evidence against association
I Genes				
1. Toxins TSST-1, SEB, SEC	Superantigens; bind non-specifically to T cells and stimulate cytokine response[94]	Causes TSS	Case-control studies, outbreak investigations[94,95]	
ETA, ETB	Cleave adherins of superficial skin layers[97]	Causes SSSS	Case-control studies and *in vivo* models[96]	
ETA		Increased risk of invasive *S. aureus* compared with carriage	Significantly associated with invasive disease in a case-control study of 334 UK isolates, after controlling for MLST[114]	No association in microarray study of 161 UK isolates[115]
ETB		Increased risk of clinical *S. aureus* disease compared with carriage	Significantly associated with clinical disease in a case-control study of 303 Brazilian isolates[116]	No association in case-control studies[114,115]
Staphylococcal enterotoxins (SE)	Superantigens[97], mechanism of emesis not established[98]	Causes food-poisoning	Case-control studies and outbreak investigations[94]	
SEP		Increased risk of MRSA bacteraemia after MRSA carriage	Case-control study of hospitalised patients with MRSA carriage in USA, examining presence of 30 genes[118]	
SEA, SED, SEE, SEI		Increased risk of infective endocarditis (IE)	Associated with IE in a multinational case-control study of 228 *S. aureus* isolates from IE compared with STI[140]	
SEJ		Increased risk of invasive *S. aureus* compared with carriage	Case-control studies as described above[114], and study of 91 isolates in India[119]	No association in case-control study [115]
Enterotoxin gene cluster (SEG, SI, SEM, SEN, SEO)		Decreased risk of invasive *S. aureus* compared with asymptomatic carriage	Case-control study of 391 Dutch strains comparing isolates from carriage and bacteraemia [120]	
delta-toxin	Binds to myeloid cell membranes resulting in haemolysis[97]	Increased risk of invasive *S. aureus* compared with carriage	Case-control study[114]	No association in case-control study [115]
PVL	Pore forming toxin, binds to and lyses myeloid cells, particularly neutrophils.[89]	Increased risk of skin and soft tissue infection	Significant association on meta-analysis of case-control studies[129]	
		Increased risk of pneumonia	Case series reports[126,127,128]	No association on meta-analysis [129]

Virulence determinant	Reported effect (if any)	Reported association	Evidence for association	Evidence against association
(PVL continued)		Increased risk of musculo-skeletal infection	Increased inflammation in USA case-control series of 85 PVL positive and negative osteomyelitis[130] Frequency reported in series of 136 isolates from infected sites, Benin[131]	No association on meta-analysis[129]
		Decreased risk of persistent MRSA bacteraemia	Case-control study of PVL positive and negative bacteria in 230 isolates from international trial of SAB treatment[133]	No effect in study of 222 isolates from MRSA bacteraemia[134]
LukDE	Pore forming toxin, lyses cells, recruits lymphocytes.[89]	Increased risk of invasive disease compared with carriage	Swedish case-control study of 134 isolates from carriage and invasive disease [121]	
2. Surface proteins FnbA	Binds fibrinogen and fibronectin, adheres to extracellular matrix (ECM); invasion of cells[91]	Increased risk of invasive *S. aureus* compared with carriage	Case-control study[114]	No association in case-control study [115]
Polymorphisms in FnbA	Enhanced binding to fibrinonectin[122]	Increased risk of cardiac device infection	Case-control studies from US[122] and Europe[123] comparing SAB with and without cardiac device infection (80 and 34 isolates respectively)	
FnbB	Binds fibrinogen and fibronectin, adheres to ECM; invasion of cells[91]	Increased risk of invasive disease compared with carriage	Case-control study [121]	No association in case-control study [115]
ClfB	Binds fibrinogen, keratin and loricrin facilitating adhesion to nasal epithelium and colonisation[91]	Increased risk of IE compared with soft tissue infection (STI)	Case-control study of 114 isolates comparing isolates from IE compared with *S. aureus* in STI[140]	
Cna	Binds collagen, adhering to collagen rich tissue and complement C1q, interferes with classical complement pathway[91]	Increased risk of invasive *S. aureus* compared with carriage	Case-control study[114]	No association in case-control study [115]
		Decreased risk of invasive *S. aureus* compared with carriage	Case-control study [121]	No association in case-control study [115]
SdrD and SdrE	SdrD: binds desquamated epithelial cells[91]	SdrD/E negative strains reduced ability to cause bone infections	Case control study of 401 isolates from nose, bone and blood; SdrD/E negative strains never found in bone infection.[157]	
SdrE	SdrE: binds complement factor H, degrades complement C3b[91]	Increased risk of invasive *S. aureus* compared with carriage	Case-control study[114]	No association in case-control study [115]
Map/Eap	Inhibits T cell function[102]	Increased risk of IE compared with soft tissue infection (STI)	Case-control study[140]	No association in case-control studies[114, 115]

Virulence determinant	Reported effect (if any)	Reported association	Evidence for association	Evidence against association
SasG	Biofilm formation [91]	Increased risk of invasive disease compared with carriage	Case-control study [121]	No association in case-control study [115]
3. Secreted proteins Scn	Complement inhibition[103]	Increased risk of invasive disease compared with carriage	Case-control study [121]	No association in case-control study [115]
Serine proteases A and B (SplA/B)		Increased risk of invasive disease compared with carriage	Case-control study [121]	No association in case-control study [115]
4**. Polysaccharide capsule type**	Resist phagocytosis[51]	Capsular type 5 associated with invasive disease, type 8 with carriage	Case-control study [121]	
5. Intercellular adhesin locus (ica)	Biofilm formation[71]	Increased risk of invasive *S. aureus* compared with carriage	Case-control study[114], case-control study of 151 bloodstream and carriage isolates in China[132]	No association in case-control study [115]
6. Agr group	Altered virulence expression[112]	Group II associated with invasive disease, Group III associated with carriage	Case-control study [121]	No association in case-control studies[114,115]
II Lineage				
No association		Association between CC and invasive *S. aureus* compared with carriage		No association in multiple case-control studies[11,115, 138]
CC-5 and CC-30		Association between CC-5 and CC-30 with SAB complicated by IE or osteo-articular infection	Case-control study of 379 isolates from carriage, bacteraemia and bacteraemia with haematogenous complications[137]	Case-control study of 100 sequential episodes of SAB shows no association[65]
CC-30	Hypothesised mechanisms for CC30: Premature stop in alpha-toxin, variant of agrC and high expression of Spa; variant of PSM α3[113].	Association between CC-30 with *S. aureus* bacteraemia complicated by IE	Case-control study[140]	No association in case-control study[121]
		Higher OR of being found in invasive disease (not statistically significant)	Case-control study of 178 cases and controls developing invasive *S. aureus* in hospital[139]	
CC-45		Lower OR of being found in invasive disease	Case-control study[139]	
Strains not in a CC		Lower mortality compared with *S. aureus* belonging to a defined CC	Case-control study[139]	

Table 2.1: Summary of published studies reporting association between *S. aureus* genome and virulence

Studies of variable gene presence designed to address this question have yielded conflicting results (Table 2.1). A comprehensive study comparing the detection of 33 virulence factors by PCR in a collection of *S. aureus* from 334 isolates (cultured from asymptomatic carriers and invasive disease) found that a small number were associated with invasive disease after controlling for population structure.[114] Three adhesins (FnbA, Cna, SdrE), three toxins (SEJ, ETA and delta-toxin) and the polysaccharide intercellular adhesin locus (ica) were all significantly more frequently found in invasive disease, and a linearly increasing odds ratio (OR) of disease was correlated with the presence of increasing number of these genes.[114] However when a subset of 161 isolates from this collection was studied using microarray technology with greater sensitivity to detecting sequence variation, no genes were found to be significantly associated with invasive disease.[115] A study comparing SAB with and without endocarditis found no single genes were different between the groups, but rather they report that a panel of 8 genetic markers correctly predicted 81% of cases to either SAB alone or SAB with endocarditis.[117]

The dispute over PVL, one hypothesised bacterial determinant, illustrates the conflicting and difficult nature of the evidence for bacterial genetic associations with disease. This pore-forming toxin of myeloid cells has been reported as a cause of severe *S. aureus* pneumonia based on its presence in severe disease and altered virulence *in vitro* of PVL positive strains, including over-expression of Spa.[124] Subsequent examination of the strains in which this association was reported found additional distant point mutations in the Agr locus producing the altered Spa expression previously attributed to PVL, underlining the importance of genetic background in studies of the impact of genes on virulence.[125] While extensive case series report co-occurrence of PVL and severe pneumonia,[126,127,128] meta-analysis finds no evidence for an increased rate of pneumonia from PVL positive strains compared to controls.[129] The role of PVL remains controversial, with opinion

divided as to whether this toxin is an important virulence factor, or merely a strain marker, linked to unidentified genetic risk factors.[135,136]

Studies of association between *S. aureus* lineage and propensity to cause disease have likewise yielded conflicting results (Table 2.1). Comparing UK carriage with invasive isolates in 334 isolates found no ST level association, except between a hospital outbreak clone (ST-36) and hospital-associated disease.[11] In contrast, a study of 371 isolates in the USA found that the two dominant lineages in the study (CC-5 and CC-30) were associated with SAB complicated by haematogenous bone or joint infection, or endocarditis[137] while CC30 was likewise associated with endocarditis risk in a multi-national study, in which CC30 isolates were also more likely to possess the ClfB, Cna and Map adhesins.[140]

Several explanations may account for these conflicting results. Technical differences in the performance of the methods used may explain the variability in findings: PCR-based studies depend upon conservation of the PCR primer-binding sequence to detect gene presence, and variability in these regions may be interpreted incorrectly as gene absence.[115] Micro-array studies can detect a greater range of variant genes, but are still incompletely sensitive and specific.[141] The sampling frames reported in these studies are from multiple continents, where the predominant lineages may vary substantially. These disparities may reflect true differences between lineages. Conversely they may represent the artefacts of *S. aureus* population structure. Studies cited here make variable attempts to account and control for population structure. Some do not account for this at all in association testing;[121,134] one study deliberately included only one isolate per PFGE profile in each group, thereby artificially skewing the populations found in cases and controls;[119] another used stratification by CC.[114] The crucial importance of proper control for population structure is illustrated in studies reporting likely

artefactual associations between both a dominant lineage and genes more frequently present in the dominant lineage.[121,140]

Evidence for a contribution of the bacterial genome to variation in virulent behaviour is currently mixed. Methods that have been applied to this question have faced technical limitations in their ability to account for small-scale genetic differences and the confounding effects of population structure.

2.5 *Staphylococcus aureus* and the host-pathogen relationship

2.5.1 Human immune system and host-pathogen interaction

Crucial to consideration of bacterial determinants of virulence is the fact that any such determinants occur in the context of a relationship between the host and pathogen. Host vulnerability to infection almost certainly plays a role. As discussed in section 2.3, several host risk factors for development of *S. aureus* disease have been recognised, including advancing age, immunoparesis, injecting drug use and haemodialysis.[52] Host factors also affect outcomes: in particular, age is a recognised risk factor for mortality following SAB.[52,61,62,65]

There is limited evidence from human studies of a human genetic susceptibility to *S. aureus* carriage and infection. A twin study found no evidence of different rates of concordance between monozygotic and dizygotic twins, suggesting strong heritability of this phenotype is unlikely.[142] Two modestly powered genome-wide association studies (GWAS) of hospital acquired SAB found no SNPs at genome wide significance.[143,144] A large study of white individuals in the USA identified one genotyped SNP approaching genome-wide significance, and two imputed SNPs had OR >1.2 for *S. aureus* infection, significant at the genome wide level. All were intergenic SNPs in the region of human leucocyte antigen (HLA) loci.[145] HLA polymorphism has been associated with susceptibility to both bacterial and viral

infections.[146]

S. aureus carriage appears to arise partly as a "fit" between bacteria and host. In healthy volunteers, exposure to a mixture of *S. aureus* strains resulted most often in the previously carried strain being isolated from the nose.[147] The forces underlying this "fit" may be immunological, since carriage confers a degree of protection from mortality due to invasive disease for individuals carrying *S. aureus*. In a study of bacteraemia arising in individuals who were screened for *S. aureus* nasal carriage on admission to hospital, carriers and non-carriers demonstrated very different risk profiles. While non-carriers were at a lower risk of developing SAB, those who did develop SAB had a 4-fold higher mortality compared with those who had carried *S. aureus* prior to the onset of bacteraemia, a finding that has been replicated.[82,139]

These observations, along with findings from human GWAS, suggest the human immune system is an important site of interaction between host and pathogen. *S. aureus* has a host of factors that allow it to evade the human innate immune system.[90] These include: inhibition of neutrophil migration by CHIPs, and prevention of neutrophil migration by Map;[90] production of toxins and phenol-soluble modulins which destroy white cells;[97] resistance to opsonisation by capsule proteins and bacterial coating with fibrinogen bound to ClfA;[90] inactivation of complement by Extracellular fibrinogen binding protein and staphylokinase;[90] bacterial nuclease degradation of neutrophil extracellular traps (NET).[148]

Further, *S. aureus* can both subvert host cell autophagy and exist in a SCV form, both of which facilitate bacterial persistence within host cells, so that host immune cells may act as reservoir for infection, and seed metastatic infection.[60,149,150] Finally, *S. aureus* subverts adaptive immune responses to prevent the development of protective immunity.[99] The majority of *S. aureus* immune evasion genes are part of

the core genome,[99] which is evidence that these functions are crucial to *S. aureus* success.

There is evidence that the relationship with human disease shapes *S. aureus* population dynamics. Wertheim and colleagues found that, in a study of 94 individuals with invasive *S. aureus* infection, in-hospital mortality was significantly higher in those infected with strains belonging to a defined clonal complex (CC-1, 5, 9 12, 15, 25, 30, 45), compared with those infected with strains whose ST did not belong to any CC (29.0% vs 4.0%). They conclude this is evidence that common *S. aureus* lineages are successful at least in part because they have evolved greater virulence.

2.5.2 Virulence may not be a "Red Queen's Race"

While the evidence presented here suggests the evolution of virulence arises from a zero-sum battle between host and pathogen, several observations suggest that *S. aureus* disease is not a simple arms race. *S. aureus* lyses human neutrophils through the Agr-dependent expression of toxins, and Agr-defective strains show impaired neutrophil lysis. However *S. aureus* strains unable to express the Agr effector PSMs, while showing reduced cell toxicity, also showed increased dissemination in a mouse model, suggesting that the host responses to bacterially-mediated cytolysis are essential in clearing *S. aureus* from the bloodstream, and failure of this clearance may promote disseminated disease.[151]

Agr-defective bacteria are widely reported in bloodstream infections, are seen to arise during the course of infection and have been associated with persistent bloodstream infection.[152,153] Moreover, reduced cytotoxicity and Agr expression were also independent predictors of mortality in a study of nosocomial MRSA pneumonia.[154] This may be an artefact of altered Agr function being found in association with MRSA with Type II SCCmec, and hospital-acquired strains.[105,154]

However this association of reduced toxicity and more severe disease has also been documented outside of hospital associated lineages and Agr expression.[155]

A study of *S. aureus in vitro* toxicity, comparing isolates from carriage and superficial infections with isolates from bacteraemia, likewise found that bacterial strains from bloodstream infection showed significantly less toxicity to lymphocytes than strains from carriage and SSTI.[156] This finding was demonstrated both when comparing isolates found in carriage and disease from the single case study my colleagues and I published, and in a collection of USA300 isolates of the same lineage. These same bacteria showed no difference in their capacity *in vitro* to invade epithelial cells, provoke NET release, produce proteases, form biofilm or resist antimicrobial peptides. However, bacteria exhibiting low cytotoxicity were significantly better able to survive in the presence of human serum. Experimentally derived Agr-defective strains were likewise more fit than their isogenic mutants in the presence of 5% serum.[156]

One explanation for these findings is that reduced production of toxins, particularly those acting on host immune cells, may provide an advantage for *S. aureus* in some settings, particularly in the ability to survive within and disseminate via the human bloodstream. Thus, when considering the variation in disease caused by *S. aureus*, in addition to accounting for variability between hosts, we must also consider that traditional measures of virulence may fail to capture subtle changes in the host-pathogen balance underlying the infective process.

2.6 Within-host evolution as a source of altered virulence

S. aureus evolution has been shaped by this relationship with the human host over centuries or more,[157] but evolution can also be observed over much shorter time scales. Bacterial evolution over the short term has been found to be up to 2-5 orders

of magnitude higher than the evolutionary rate over longer time scales, and in *S. aureus* carriage has been estimated at approximately 8 SNPs per year.[158] Studies of *S. aureus* carriage have demonstrated that micro-diversity is the norm.[50,88,159] This variation arises from *de novo* mutation within the host and is shaped by forces including random drift, purifying and diversifying selection.[158]

Within-host evolution has been clearly demonstrated in response to the powerful selective pressure of antibiotic use. Evolution of resistance in response to antibiotic exposure is an example of a selective sweep: strains with newly-arising variants conferring antimicrobial resistance will have a comparative fitness advantage, and so are expected to increase in frequency under this selective pressure.[158] WGS of isolates taken during the course of a persistent *S. aureus* bloodstream infection (and failed treatment with multiple antibiotics) revealed that modest genomic change – just two SNPs – conferred significant phenotypic changes.[160] These SNPs reduced antibiotic susceptibility and permanently activated the "stringent response", resulting in SCV formation, with attenuated growth and reduced antibiotic suspectibility.[160] There are many such examples of antimicrobial resistance emerging in *S. aureus* within-host during treatment.[24,161]

Adaptive changes have also been demonstrated during host adaptation in opportunistic pathogens. Lieberman et al reported on evolution during an outbreak of *Burkholderia dolosa* among a cohort of CF patients: by demonstrating parallel evolution in multiple hosts, they showed which mutations were likely to be adaptive.[162] They observed a strong signal of fluoroquinolone resistance evolving in multiple hosts after transmission, with predictable mutations and drug resistance occurring through a limited number of paths. They were also able to identify candidate virulence genes by demonstrating greater than expected rates of mutations within single genes (Figure 2.6). Intriguingly, a loss of function mutation

in glycosyltransferase (which impairs O-antigen presentation) was inherited in the strain transmitted to 9 patients, and each of these independently acquired gain of function mutations in this gene. This suggests there is a trade-off between mutations favouring transmission and those which favour survival or virulence within hosts.[162] Further investigation of this outbreak strain demonstrated diverse communities with competing adaptive mutations, a finding which may be explained by large population size, segregation of lineages within the airway, or niche-specific adaptation.[163] McAdam et al found evidence of similar adaptation in *S. aureus* occurring in the airways of patients with CF, with changes in fusidic acid resistance, haemolysis and global virulence regulators.[164] These findings demonstrate that for opportunistic pathogens in an anatomically abnormal setting, within-host evolution yields insights into bacterial adaptation.

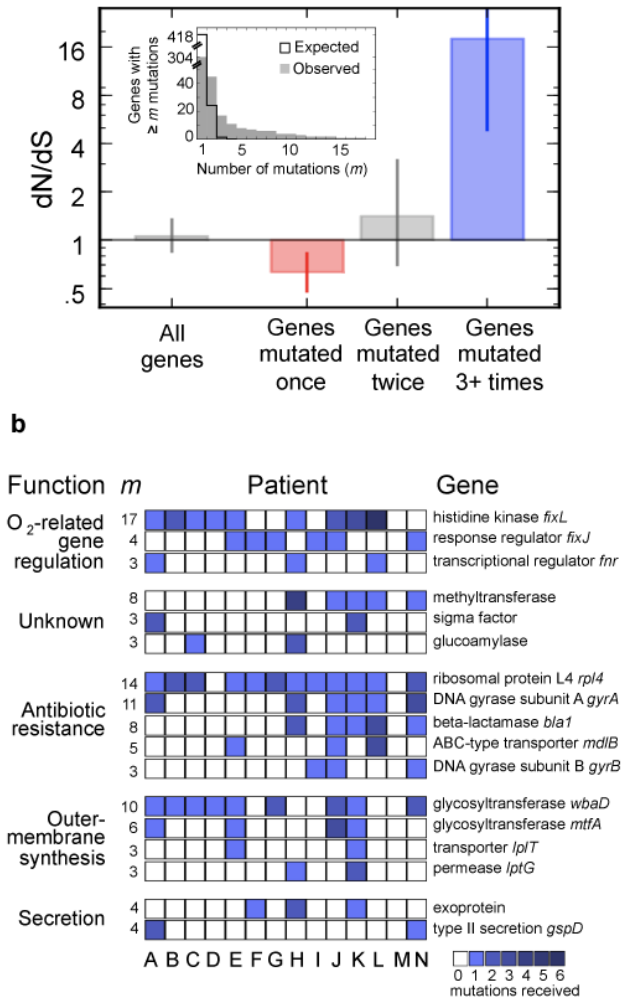


Figure 2.6: evidence of within-host selection from the observation of multiple mutations within genes. a) The distribution of mutations departs strongly from that expected under a neutral model and genes with multiple mutations exhibit higher dN/dS than those with 1 or 2 mutations. b) 17 genes with ≥3 mutations are grouped according by functional category. Rows represent each gene, columns represent the 14 patients, and boxes are shaded by the number of mutations per gene per patient. (Figure from Lieberman et al.[162])

Within-host changes within a diverse population may not represent straightforward adaption. Social dynamics in such populations make "cheating" possible within microbial communities. In *Pseudomonas aeruginosa*, this has been proposed to explain changes in iron metabolism during carriage over time in hosts with CF.[165] Within a population, there is a constant mutation pressure throughout the genome and some strains stop producing the iron-binding protein pyoverdine as a result. Normally, this would be deleterious and counter-selected, but if they co-exist with pyoverdine producers, these pyoverdine non-producers will continue to

benefit from uptake of pyoverdine-bound iron without the metabolic cost of pyoverdine production. Indeed, the rate at which pyoverdine production is lost by mutation far outweighs the rate at which the receptor which facilitates uptake of iron bound to pyoverdine is lost. Similar "cheating" has been demonstrated in *P. aeruginosa* quorum-sensing systems: strains with mutants leading to loss of the effector arm retained their signalling capacity and showed relative fitness advantages in culture.[166]

In understanding adaptation we must also consider trade-offs between different stages of an organism's life-history: changes that increase disease propensity in the short term may not be adaptive over longer time scales or across hosts. An analogous analysis has yielded insight into oncogenesis in human somatic cells. Here the search for causal or 'driver' mutations is complicated by the co-occurrence of many mutations that are merely associated or 'passenger' mutations.[167] Investigators compiled a list of mutations found in malignant cells, seeking genes with a higher than expected rate of non-synonymous mutations, or the accumulation of multiple non-synonymous mutations.[167] Searching for the accumulation of mutations in genes across many individuals – rather than a single driver mutation – uncovers both candidate genes and signatures of adaptation that are specific to many cancers.[167,168]

There are similarities between malignancy and microbial infection, including invasive bacterial infection.[169] Enhanced growth relative to other cells is an evolutionary advantage for the cell, but malignancy is likely to result in death of the organism. Likewise, bloodstream invasion is a temporary opportunity to extend growth for the bacteria, but is an evolutionary dead-end as bacteraemia will likely kill its host. Both might be considered 'accidents' in the long view, even though they represent the operation of selection pressures in the short term.

Bacterial whole-genome sequencing opens new possibilities to investigate the bacterial genomic changes in great detail over short time scales while interacting with the human host. It is now possible to systematically investigate within-host evolution of bacteria for clues to variability in *S. aureus* virulence.

2.7 Bacterial genome-wide association studies for investigation of virulence

2.7.1 GWAS for studying natural populations

As technological advances facilitated the investigation of the genetic basis of microbial virulence, Stanley Falkow proposed 3 "Molecular Koch's postulates", based on the Koch's original postulates for establishing the role of microbes in disease:[170]

1. The phenotype or property under investigation should be associated with pathogenic members of a genus of pathogenic strains of a species
2. Specific inactivation of the gene(s) associated with the suspected virulence trait should lead to a measurable loss in pathogenicity or virulence, or the gene(s) associated with the supposed virulence trait should be isolated by molecular methods. Specific inactivation or deletion of the gene(s) should lead to loss of function in the clone.
3. Reversion or allelic replacement of the mutated gene should lead to restoration of pathogenicity, or the replacement of the modified gene(s) for its allelic counterpart in the strain of origin should lead to loss of function and loss of pathogenicity or virulence. Restoration of pathogenicity should accompany the reintroduction of the wild-type gene(s).

Much investigation of the genetic basis for bacterial virulence has focussed on postulates two and three, and it is only recently that systematic surveys of bacterial genomes in natural populations have become possible.[171]

Genome-wide association studies (GWAS) are a powerful tool for identifying the genetic basis of phenotypes in natural populations. These identify genetic variants found in a group comprised of individuals exhibiting the phenotype (cases) and individuals who do not (controls). For any given variant, we can then test the null hypothesis that it is not found in different frequency between the two groups.[172,173,174] This testing is unbiased by prior hypotheses about what genes or variants may be causal in the phenotype of interest.[173] This approach was developed from the hypothesis that many variants have a small effect size in contributing to common diseases.[174]

In human studies, GWAS have yielded significant insights since the first human GWAS was published in 2005, revealing information about population genetics as well as the heritability and molecular basis of complex traits.[172] Around 10,000 associations have been demonstrated between SNPs and medically important phenotypes from genome-wide SNP surveys.[172]

2.7.2 Challenges for bacterial GWAS

GWAS in bacteria has lagged behind this success, despite the now widespread use of WGS in bacteria and resultant large datasets of bacterial genome sequences. One challenge is that GWAS depend on reliable phenotyping, and metadata collection and storage has not kept pace with the generation of sequencing data.[173] Bacteria also differ from humans in important respects, many of which mean that tools applied to human GWAS cannot be simply translated to bacteria.[173, 174]

Bacterial genomes are significantly smaller than humans: the *S. aureus* genome is approximately 3Mb long on a single chromosome, approximately 1000 times shorter than the human genome which is spread over 23 chromosomes.[18,175] Bacteria are haploid, meaning any allele is only ever present or absent rather than homozygous or heterozygous, so that the concept of dominant and recessive alleles

is not applicable. Bacterial genomes contain very little non-coding sequence compared with humans, as almost all of the bacterial genome is coding sequence.[174] This means that the number of genes is more similar than would be expected from total genome length: *S. aureus* contains approximately 3000 open reading frames, while the human genome has around ten times that. [18,175]

Sexual reproduction in humans involves genetic recombination with every generation. Recombination occurs most frequently at particular 'hot-spots', so that linkage disequilibrium (LD) occurs in block-like patterns across the genome.[176] This means that variants are strongly correlated with nearby variants, but this correlation rapidly decreases with increasing genetic distance. In human GWAS, LD is the basis of SNP imputation, so that "typing" of a single biallelic SNP can be used to infer the sequence for the surrounding region of the genome.[174] The nature of recombination in human chromosomes ensures genetic variants of interest are found across multiple different backgrounds.[173]

Bacteria reproduce asexually, and genomic variation will only be shared between individuals of different ancestry where there is horizontal gene transfer, or recurrent mutation (homoplasy).[177] The effects of homologous recombination lead to breakdown of LD in some regions.[174] LD in bacteria occurs in long haplotype blocks extending across the genome, meaning that variants across the full length of the bacterial genome can be in complete linkage.[173] Selection pressures exacerbate and maintain strong bacterial population structure.[174] Bacterial populations tend to consist of clusters of highly related individuals, separated by long branches, so that a large proportion of all genetic variation is population stratified. Earle and colleagues demonstrated that in four bacterial species (including *S. aureus*), the proportion of genetic variation captured by first 10 principal components (PCs) of variance was between 70 and 93%, compared to a mere 27% in human

chromosome 1.[178]

Population structure leads to two major difficulties in GWAS, both of which are typically more pronounced in bacteria. The first is that LD between variants creates difficulty in distinguishing causal from linked passenger mutations.[173] The second is that unmeasured confounders (e.g. environmental exposure) or sampling may differ systematically between sub-populations, and be misattributed to genetic differences between populations.

In human GWAS, population structure resulting from ancestry may be addressed by studying within an ethnic group, and using PC analysis to identify population outliers and control for more subtle population structure.[174] Another approach is the use of mixed models, which perform better in the presence of subtle relatedness, because they include all relatedness in the sample population, rather than that captured by a finite number of PCs.[179] Controlling for population structure in bacterial GWAS is essential to avoid false associations – variants which are not causally associated with the phenotype of interest, but are in LD with causal variants – as well as to prevent confounding by population-stratified differences in environmental exposure or sampling bias. However it means a substantial sacrifice of power, as a great proportion of bacterial genetic variance, including potentially causal variants, exists between lineages.

The nature of the genetic variation we wish to study in bacteria also differs from that in humans. Horizontal gene transfer does not occur in the human genome, where there is almost no "accessory" genome, while bacterial species can have extensive variation in the accessory genome.[173] Human studies rely largely on SNPs as markers of variation, usually through typing biallelic SNPs on genotyping chips, while bacterial studies are in a position to make use of WGS by virtue of their much smaller genomes.[174] Bacterial GWAS methods must be able to account for structural

genomic variation – including the variable presence of whole genes – as well as SNPs (including tri-and tetra-allelic SNPs) and indels.[173,174]

In addition to these complexities and challenges, bacterial GWAS have at least one advantage over human studies: our ability to manipulate bacterial genomes, and the short generation time mean that *in vitro* validation of effects and exploration of casual mechanisms can be undertaken in bacteria for many phenotypes.[173]

2.7.3 Progress in bacterial GWAS

Recent progress in bacterial GWAS has seen development and application of methods to address these challenges. A study by Sheppard and colleagues in 2013 was the first to demonstrate novel findings determined through bacterial GWAS, in a study of host species in *C. jejuni* and *C. coli*.[180] This study developed a "kmer" approach: for each assembled genome, all 30bp DNA "words" (or 30-mers) in the sequence were identified, and each of these was tested for association with the phenotype. This approach captures SNPs, indels and also variably present elements of the genome, and with it, Sheppard et al determined that differential vitamin B5 synthesis was associated with host species adaptation.[180] While SNP-based methods have yielded insights in bacterial GWAS of antibiotic resistance,[181] this kmer-based method allows for investigation of a wider range of genomic changes.

The adjustment of tools to control for population structure has also been essential to the success of bacterial GWAS. Laabei and colleagues performed the first bacterial GWAS for virulence, and used an approach of restricting their investigation to a single lineage, studying 90 isolates of the MRSA clone ST239.[182] They identified an *in vitro* toxicity phenotype that correlated with disease severity in a mouse model. They identified 122 variants associated with toxicity at $p<0.05$, and performed functional validation on a subset by producing transposon mutants, in which 4 of 13 candidate mutants demonstrated an effect on toxicity. Most

recently, these authors have extended this approach to use SNP GWAS, *in vitro* toxicity and biofilm formation and clinical data to identify loci that predict either i*n vitro* phenotype or mortality in SAB caused by CC-22 and CC-30.[183]

Studies within *S. aureus* clonal complexes are thus indicative that GWAS may lead to insights even into complicated phenotypes, but studies limited to a single ST will inevitably lack flexibility, and even within CC, population structure is evident. The use of linear mixed models (LMM) offers greater flexibility, and has been successfully applied in *S. aureus* to identify determinants of vancomycin-intermediate resistance.[184]

Bacterial specific GWAS tools incorporating both kmer-based testing and LMM or PCA-like approaches to control for bacterial population have subsequently been developed.[178,185] These have been validated by testing for bacterial determinants of antimicrobial resistance, and have demonstrated that locus effects associated with resistance can be identified.[178,185] Additionally, the relatedness that allows for control for population structure can be exploited to identify lineage-phenotype associations, and identify the loci associated phenotype within lineages, recovering the power to detect results that would otherwise be lost (Figure 2.7).[178] In addition to identifying known genetic associations, these tools have also been able to identify novel antimicrobial resistance determinants in *Escherichia coli* and candidates for invasive disease in *Streptococcus pyogenes*.[178,185]

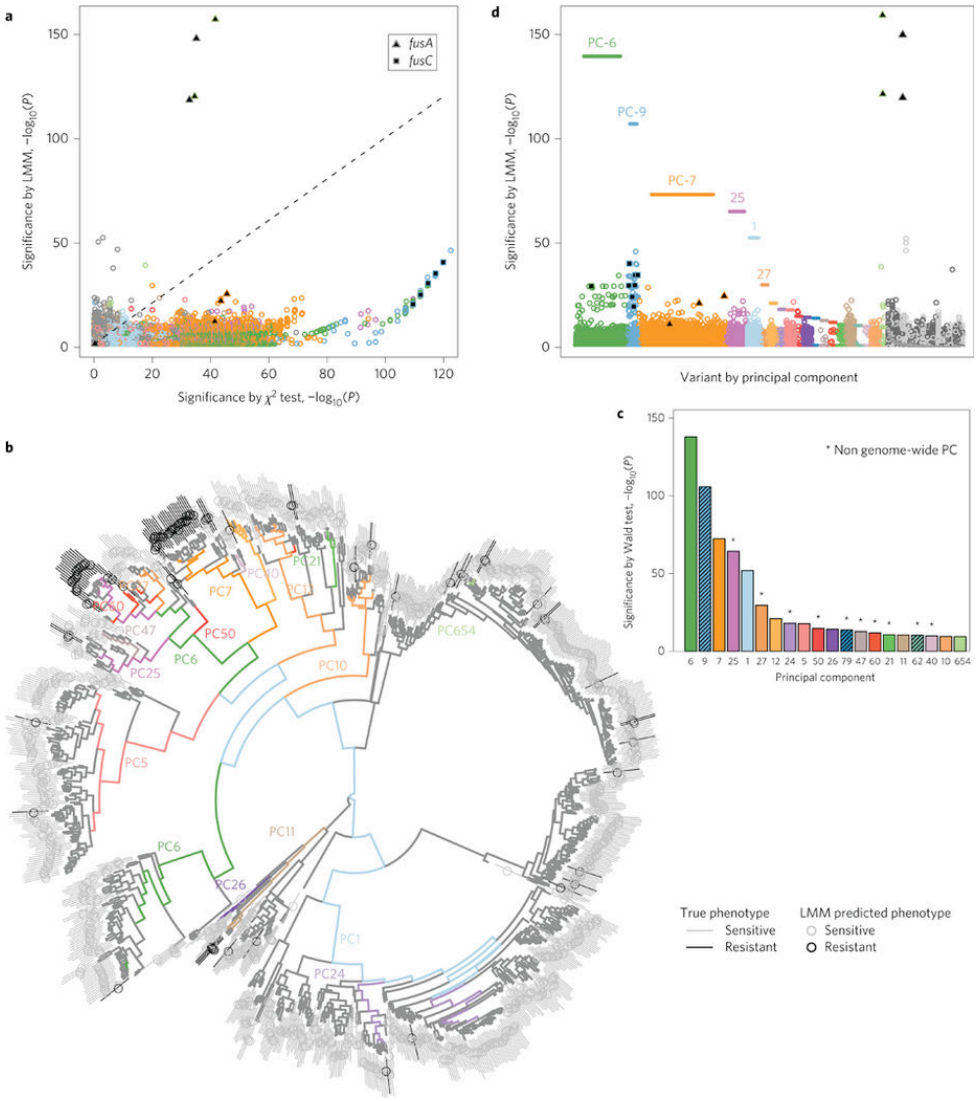


Figure 2.7: GWAS testing for fusidic acid resistance in *S. aureus*, controlling for population structure. **(A)** The effects of using LMM to control for population structure: the significance of the presence or absence of 31bp kmers using χ^2 test (x-axis), compared to testing in LMM (y-axis). Kmers most significantly correlated with one of the top 20 most significant PCs are coloured by that PC, otherwise grey. **(B)** The correspondence of the 20 most significant PCs with branches in the phylogeny; branches are coloured by the PC to which they correspond. **(C)** Results of testing PCs for association with the phenotype. **(D)** Manhattan plot showing significance of kmers after controlling for population structure, ordered by PC. This method is able to identify a locus effect that is independent of lineage (fusA), and a locus effect that underpins two lineages strongly correlated with fusidic acid resistance (fusC). (Figure from Earle et al. [178])

Microbial genome sequencing and the development of methods for bacterial genome-wide association studies thus present new opportunities to discover the bacterial genetic basis of serious infection in major pathogens. In particular, these methods allow study of natural populations and testing of the entire genome, and are not limited in discovery to the study of single, already characterised loci.

2.8 Thesis outline

Building on these tools and exploratory studies, in this thesis I harness the power of WGS to explore the genomic variation of *S. aureus* in natural populations. By understanding the nature of this variation, I aim to uncover associations to inform our understanding of the contribution of the bacterial genome to *S. aureus* virulence.

I first use paired isolates taken from individuals with concurrent nasal carriage and invasive disease. By comparing multiple isolates from both sites, I present a detailed picture of bacterial populations and within-host evolution during *S. aureus* infection. By cataloguing the variation that arises within and between these sites, I present a novel avenue for understanding *S. aureus* adaptation and virulence. The study of closely related isolates exhibiting different phenotypes is analogous to within family studies to identify the genetic basis of rare disease.

I then move to population-based studies of *S. aureus* disease, using collections of bacteria found in specific infections and comparing them to relevant controls by applying the tools of bacterial GWAS to the systematic search for virulence determinants. I present two case-control studies of important bacterial phenotypes.

The first is a GWAS of *S. aureus* bacteraemia and nasal carriage in patients over 13 years of age in the UK. This study uses a collection of 1017 cases and 984 controls, totalling 2001 isolates. Cases were identified from three centres in the southern UK: Oxford University Hospitals NHS Foundation Trust, Brighton and Sussex University Hospital NHS Trust, and Plymouth Hospitals NHS Trust. Controls were identified from studies of *S. aureus* carriage and transmission in Oxfordshire, UK.

Secondly I present a GWAS of pyomyositis. This is a characteristic invasive infection, most often seen in the tropics and usually caused by *S. aureus*. I compare a

collection of isolates from pyomyositis to *S. aureus* nasal carriage isolates, all of which were collected from children attending Angkor Hospital for Children in Siem Reap, Cambodia.

Together these studies uncover evidence of subtle adaption as well as strong bacterial genetic effects on pathogenesis, with implications for prevention and treatment of human disease. I also demonstrate several complications and limitations in the use of bacterial GWAS to understand infections.

References chapter 2

1. Classics in infectious diseases. "On abscesses". Alexander Ogston (1844-1929). Rev Infect Dis. 1984 Jan-Feb;6(1):122-8.

2. Licitra, G. Etymologia: Staphylococcus. Emerg Infect Dis. 2013 Sep; 19(9): 1553.

3. Kluytmans J, van Belkum A, Verbrugh H. Nasal carriage of *Staphylococcus aureus*: epidemiology, underlying mechanisms, and associated risks. Clin Microbiol Rev. 1997 Jul;10(3):505-20.

4. Public Health England. "Identification of *Staphylococcus* species, *Micrococcus* species and *Rothia* species." UK Standards for Microbiology Investigations ID 7: Issue no:3, Issue date 12.11.14

5. Bannerman TL and Peacock SJ "*Staphylococcus*, *Micrococcus* and other Catalase-Positive Cocci" in Murray PR, Baron EJ, Jorgensen JH, Landry ML, Pfaller MA, eds Manual of Clinical Microbiology, 9th ed. Washington USA: ASM Press, 2007: pp 390-411.

6. Li W, Raoult D, Fournier PE. Bacterial strain typing in the genomic era. FEMS Microbiol Rev. 2009 Sep;33(5):892-916.

7. Prevost G, Pottecher B, Dahlet M, Bientz M, Mantz J M, Piemont Y. Pulsed field gel electrophoresis as a new epidemiological tool for monitoring methicillin-resistant *Staphylococcus aureus* in an intensive care unit. J Hosp Infect. 1991;17:255–269.

8. Schlichting C, Branger C, Fournier J M, Witte W, Boutonnier A, Wolz C, Goullet P, Doring G. Typing of *Staphylococcus aureus* by pulsed-field gel electrophoresis, zymotyping, capsular typing, and phage typing: resolution of clonal relationships. J Clin Microbiol. 1993;31:227–232.

9. Enright MC, Day NP, Davies CE, Peacock SJ, Spratt BG. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus* J Clin Microbiol 2000 Mar;38(3):1008-15.

10. Enright MC, Robinson DA, Randle G, Feil EJ, Grundmann H, Spratt BG. The evolutionary history of methicillin-resistant *Staphylococcus aureus* (MRSA). Proc Natl Acad Sci U S A. 2002 May 28;99(11):7687-92.

11. Feil EJ, Cooper JE, Grundmann H, Robinson DA, Enright MC, Berendt T, Peacock SJ, Smith JM, Murphy M, Spratt BG, et al. How clonal is *Staphylococcus aureus*? J Bacteriol 2003;185(11):3307-3316.

12. Shopsin B, Gomez M, Montgomery SO, Smith DH, Waddington M, Dodge DE, Bost DA, Riehman M, Naidich S, Kreiswirth BN. Evaluation of protein A gene polymorphic region DNA sequencing for typing of *Staphylococcus aureus* strains. J Clin Microbiol. 1999 Nov;37(11):3556-63.

13. Tang YW, Waddington MG, Smith DH, Manahan JM, Kohner PC, Highsmith LM, Li H, Cockerill FR 3rd, Thompson RL, Montgomery SO, Persing DH. Comparison of protein A gene sequencing with pulsed-field gel electrophoresis and epidemiologic data for molecular typing of methicillin-resistant *Staphylococcus aureus*. J Clin Microbiol. 2000

Apr;38(4):1347-51.

14. International Working Group on the Classification of Staphylococcal Cassette Chromosome Elements (IWG-SCC). Classification of staphylococcal cassette chromosome mec (SCCmec): guidelines for reporting novel SCCmec elements. Antimicrob Agents Chemother. 2009 Dec;53(12):4961-7.

15. Basic-Hammer N, Vogel V, Basset P, Blanc DS. Impact of recombination on genetic variability within *Staphylococcus aureus* clonal complexes. Infect Genet Evol. 2010 Oct;10(7):1117-23.

16. Forde BM, O'Toole PW. Next-generation sequencing technologies and their impact on microbial genomics. Brief Funct Genomics. 2013 Sep;12(5):440-53.

17. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010 Jan;11(1):31-46.

18. Kuroda M, Ohta T, Uchiyama I, Baba T, Yuzawa H, Kobayashi I, Cui L, Oguchi A, Aoki K, Nagai Y, et al. Whole genome sequencing of meticillin-resistant *Staphylococcus aureus* Lancet 2001 Apr 21;357(9264):1225-40.

19. Baba T, Takeuchi F, Kuroda M, Yuzawa H, Aoki K, Oguchi A, Nagai Y, Iwama N, Asano K, Naimi T, Kuroda H, Cui L, Yamamoto K, Hiramatsu K. Genome and virulence determinants of high virulence community-acquired MRSA. Lancet. 2002 May 25;359(9320):1819-27.

20. Holden MTG, Feil EJ, Lindsay JA, Peacock SJ, Day NPJ, Enright MC, Foster TJ, Moore CE, Hurst L, Atkin R, et al. Complete genomes of two clinical *Staphylococcus aureus* strains: Evidence for the rapid evolution of virulence and drug resistance Proc Natl Acad Sci U S A. 2004;101(26):9786 -9791.

21. Diep BA, Gill SR, Chang RF, Phan TH, Chen JH, Davidson MG, Lin F, Lin J, Carleton HA, Mongodin EF, et al. Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant *Staphylococcus aureus* Lancet. 2006 Mar 4;367(9512):731-9.

22. Gill SR, Fouts DE, Archer GL, Mongodin EF, DeBoy RT, Ravel J, Paulsen IT, Kolonay JF, Brinkac L, Beanan M, et al. Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant *Staphylococcus aureus* strain and a biofilm-producing methicillin-resistant *Staphylococcus epidermidis* strain J Bacteriol 2005;187(7):2426-2438.

23. Gillaspy AF, Worrell V, Orvis J, Roe BA, Dyer DW, Iandolo JJ. The *Staphylococcus aureus* NCTC8325 genome. In: V. Fischetti, R. Novick, J. Ferretti, D. Portnoy, J. Rood, editors. Gram positive pathogens. 1st ed. Washington, DC: ASM Press; 2006.

24. Mwangi MM, Wu SW, Zhou Y, Sieradzki K, de Lencastre H, Richardson P, Bruce D, Rubin E, Myers E, Siggia ED, Tomasz A. Tracking the in vivo evolution of multidrug resistance in *Staphylococcus aureus* by whole-genome sequencing. Proc Natl Acad Sci U S A. 2007 May 29;104(22):9451-6.

25. Herron-Olson L, Fitzgerald JR, Musser JM, Kapur V. Molecular correlates of host specialization in *Staphylococcus aureus* PLoS One 2007 Oct 31;2(10):e1120.

26. Baba T, Bae T, Schneewind O, Takeuchi F, Hiramatsu K. Genome sequence of *Staphylococcus aureus* strain Newman and comparative analysis of staphylococcal genomes: Polymorphism and evolution of two major pathogenicity islands J Bacteriol 2008 Jan;190(1):300-10.

27. Lowder BV, Guinane CM, Ben Zakour NL, Weinert LA, Conway-Morris A, Cartwright RA, Simpson AJ, Rambaut A, Nubel U, Fitzgerald JR. Recent human-to-poultry host jump, adaptation, and pandemic spread of *Staphylococcus aureus* Proc Natl Acad Sci U S A 2009 Nov 17;106(46):19545-50.

28. Holden MT, Lindsay JA, Corton C, Quail MA, Cockfield JD, Pathak S, Batra R, Parkhill J, Bentley SD, Edgeworth JD. Genome sequence of a recently emerged, highly transmissible, multi-antibiotic- and antiseptic-resistant variant of methicillin-resistant *Staphylococcus aureus*, sequence type 239 (TW) J Bacteriol 2010 Feb;192(3):888-92.

29. Schijffelen MJ, Boel CH, van Strijp JA, Fluit AC. Whole genome analysis of a livestock-associated methicillin-resistant *Staphylococcus aureus* ST398 isolate from a case of human endocarditis BMC Genomics 2010 Jun 14;11:376,2164-11-376.

30. Chua K, Seemann T, Harrison PF, Davies JK, Coutts SJ, Chen H, Haring V, Moore R, Howden BP, Stinear TP. Complete genome sequence of staphylococcus aureus strain JKD6159, a unique australian clone of ST93-IV community methicillin-resistant *Staphylococcus aureus* J Bacteriol 2010 Oct;192(20):5556-7.

31. Guinane CM, Ben Zakour NL, Tormo-Mas MA, Weinert LA, Lowder BV, Cartwright RA, Smyth DS, Smyth CJ, Lindsay JA, Gould KA, et al. Evolutionary genomics of *Staphylococcus aureus* reveals insights into the origin and molecular basis of ruminant host adaptation Genome Biol Evol 2010 Jul 12;2:454-66.

32. Holden MT, Hsu LY, Kurt K, Weinert LA, Mather AE, Harris SR, Strommenger B, Layer F, Witte W, de Lencastre H, et al. A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic Genome Res 2013 Apr;23(4):653-64.

33. Lindsay JA, Holden MT. Understanding the rise of the superbug: Investigation of the evolution and genomic variation of *Staphylococcus aureus* Funct Integr Genomics 2006 Jul;6(3):186-201

34. Lindsay JA, Holden MT. *Staphylococcus aureus*: Superbug, super genome? Trends Microbiol 2004 Aug;12(8):378-85.

35. Bosi E, Monk JM, Aziz RK, Fondi M, Nizet V, Palsson BØ. Comparative genome-scale modelling of *Staphylococcus aureus* strains identifies strain-specific metabolic capabilities linked to pathogenicity. Proc Natl Acad Sci U S A. 2016 Jun 28;113(26):E3801-9.

36. Malachowa N, DeLeo FR. Mobile genetic elements of *Staphylococcus aureus*. Cell Mol Life Sci. 2010 Sep;67(18):3057-71.

37. Everitt RG, Didelot X, Batty EM, H, et al. Mobile elements drive recombination hotspots in the core genome of *Staphylococcus aureus* Nat Commun 2014 May 23;5:3956.

38. Williams RE. Healthy carriage of *Staphylococcus aureus:* its prevalence and

importance. Bacteriol Rev. 1963;27:56-71.

39. Dancer SJ, Noble WC. Nasal, axillary, and perineal carriage of *Staphylococcus aureus* among women: identification of strains producing epidermolytic toxin. Journal of clinical pathology. 1991;44(8):681-4.

40. Mertz D, Frei R, Periat N, *et al*. Exclusive *Staphylococcus aureus* throat carriage: at-risk populations. Arch intern med. 2009;169(2):172-8.

41. Miller RR, Walker AS, Godwin H, Fung R, Votintseva A, Bowden R, Mant D, Peto TE, Crook DW, Knox K. Dynamics of acquisition and loss of carriage of *Staphylococcus aureus* strains in the community: the effect of clonal complex. J Infect. 2014 May;68(5):426-39.

42. VandenBergh MF, Yzerman EP, van Belkum A *et al*. Follow-up of *Staphylococcus aureus* nasal carriage after 8 years: redefining the persistent carrier state. J Clin Microbiol. 1999;37(10):3133-40.

43. Votintseva AA, Miller RR, Fung R, *et al*. Multiple-strain colonization in nasal carriers of *Staphylococcus aureus*. J Clin Microbiol. 2014;52(4):1192-200.

44. Ritchie SR, Isdale E, Priest P, Rainey PB, Thomas MG. The turnover of strains in intermittent and persistent nasal carriers of *Staphylococcus aureus*. J Infect. 2016 Mar;72(3):295-301.

45. McBryde ES, Bradley LC, Whitby M, McElwain DL. An investigation of contact transmission of methicillin-resistant *Staphylococcus aureus*. J Hosp Infect. 2004 Oct;58(2):104-8.

46. Otter JA, Yezli S, French GL. The role played by contaminated surfaces in the transmission of nosocomial pathogens. Infect Control Hosp Epidemiol. 2011 Jul;32(7):687-99.

47. Knox J, Uhlemann AC, Lowy FD. *Staphylococcus aureus* infections: transmission within households and the community. Trends Microbiol. 2015 Jul;23(7):437-44.

48. Tosas Auguet O, Betley JR, Stabler RA, Patel A, Ioannou A, Marbach H, Hearn P, Aryee A, Goldenberg SD, Otter JA, Desai N, Karadag T, Grundy C, Gaunt MW, Cooper BS, Edgeworth JD, Kypraios T. Evidence for Community Transmission of Community-Associated but Not Health-Care-Associated Methicillin-Resistant *Staphylococcus aureus* Strains Linked to Social and Material Deprivation: Spatial Analysis of Cross-sectional Data. PLoS Med. 2016 Jan 26;13(1):e1001944.

49. Gordon NC, Pichon B, Golubchik T, Wilson DJ, Paul J, Blanc DS, Cole K, Collins J, Cortes N, Cubbon M, Gould FK,et al. Whole-Genome Sequencing Reveals the Contribution of Long-Term Carriers in *Staphylococcus aureus* Outbreak Investigation. J Clin Microbiol. 2017 Jul;55(7):2188-2197.

50. Price JR, Golubchik T, Cole K, Wilson DJ, Crook DW, Thwaites GE, Bowden R, Walker AS, Peto TE, Paul J, Llewelyn MJ. Whole-genome sequencing shows that patient-to-patient transmission rarely accounts for acquisition of *Staphylococcus aureus* in an intensive care unit. Clin Infect Dis. 2014 Mar;58(5):609-18.

51. Lowy FD. *Staphylococcus aureus* infections N Engl J Med 1998 Aug 20;339(8):520-32.

52. Tong SY, Davis JS, Eichenberger E, Holland TL, Fowler VG Jr. *Staphylococcus aureus* infections: epidemiology, pathophysiology, clinical manifestations, and management. Clin Microbiol Rev. 2015 Jul;28(3):603-61.

53. Que YA and Moreillon P. *Staphylococcus aureus* (including Staphylococcal Toxic Shock Syndrome). In Bennett JE, Dolin R, Blaser MJ, eds. Mandell, Douglas and Bennett's principles and practice of infectious diseases. 8th ed. Philadelphia PA : Elsevier Saunders, 2015; 2237-2271

54. McCaig LF, McDonald LC, Mandal S, Jernigan DB. *Staphylococcus aureus*-associated skin and soft tissue infections in ambulatory care. Emerg Infect Dis. 2006 Nov;12(11):1715-23.

55. Fritz SA, Hogan PG, Hayek G, Eisenstein KA, Rodriguez M, Krauss M, Garbutt J, Fraser VJ. Staphylococcus aureus colonization in children with community-associated *Staphylococcus aureus* skin infections and their household contacts. Arch Pediatr Adolesc Med. 2012 Jun 1;166(6):551-7.

56. Anderson DJ, Moehring RW, Sloane R, Schmader KE, Weber DJ, Fowler VG,Jr, Smathers E, Sexton DJ. Bloodstream infections in community hospitals in the 21st century: A multicenter cohort study PLoS One 2014 Mar 18;9(3):e91713.

57. Shorr AF, Tabak YP, Killian AD, Gupta V, Liu LZ, Kollef MH. Healthcare-associated bloodstream infection: A distinct entity? insights from a large U.S. database Crit Care Med 2006 Oct;34(10):2588-95.

58. Annual Epidemiological Commentary: Mandatory MRSA, MSSA and *E. coli* bacteraemia and *C. difficile* infection data 2016/7, Public Health England 2017 [online].

59. Chang FY, MacDonald BB, Peacock JE,Jr, Musher DM, Triplett P, Mylotte JM, O'Donnell A, Wagener MM, Yu VL. A prospective multicenter study of *Staphylococcus aureus* bacteremia: Incidence of endocarditis, risk factors for mortality, and clinical impact of methicillin resistance Medicine (Baltimore) 2003 Sep;82(5):322-32.

60. Thwaites GE, Gant V. Are bloodstream leukocytes trojan horses for the metastasis of *Staphylococcus aureus*? Nat Rev Microbiol 2011 Mar;9(3):215-22.

61. Thwaites GE, United Kingdom Clinical Infection Research Group (UKCIRG). The management of *Staphylococcus aureus* bacteremia in the United Kingdom and Vietnam: A multi-centre evaluation PLoS One 2010 Dec 13;5(12):e14170.

62. Kaasch AJ, Barlow G, Edgeworth JD, Fowler VGJ, Hellmich M, Hopkins S, Kern WV, Llewelyn MJ, Rieg S, Rodriguez-Bano J, et al. *Staphylococcus aureus* bloodstream infection: A pooled analysis of five prospective, observational studies. The Journal of Infection 2014 Mar;68(3):242-51.

63. Laupland KB, Ross T, Gregson DB. *Staphylococcus aureus* bloodstream infections: risk factors, outcomes, and the influence of methicillin resistance in Calgary, Canada, 2000-2006. J Infect Dis. 2008 Aug 1;198(3):336-43.

64. van Hal SJ, Jensen SO, Vaska VL, Espedido BA, Paterson DL, Gosbell IB. Predictors of mortality in *Staphylococcus aureus* Bacteremia. Clin Microbiol Rev. 2012 Apr;25(2):362-86.

65. Price J, Baker G, Heath I, Walker-Bone K, Cubbon M, Curtis S, Enright MC, Lindsay J, Paul J, Llewelyn M. Clinical and Microbiological Determinants of Outcome in *Staphylococcus aureus* Bacteraemia. Int J Microbiol. 2010:65485.

66. Baddour LM, Wilson WR, Bayer AS, Fowler VG Jr, Tleyjeh IM, Rybak MJ, Barsic B, Lockhart PB, Gewitz MH, Levison ME, et al. Infective Endocarditis in Adults: Diagnosis, Antimicrobial Therapy, and Management of Complications: A Scientific Statement for Healthcare Professionals From the American Heart Association. Circulation. 2015 Oct 13;132(15):1435-86.

67. Wertheim HF, Melles DC, Vos MC, van Leeuwen W, van Belkum A, Verbrugh HA, Nouwen JL. The role of nasal carriage in *Staphylococcus aureus* infections Lancet Infect Dis 2005 Dec;5(12):751-62.

68. Verma S. Pyomyositis in children Curr Infect Dis Rep 2016 Mar;18(4):12,016-0520-2.

69. Waldvogel FA, Medoff G, Swartz MN. Osteomyelitis: a review of clinical features, therapeutic considerations and unusual aspects. N Engl J Med. 1970 Jan 22;282(4):198-206.

70. Barrett L, Atkins B. The clinical presentation of prosthetic joint infection. J Antimicrob Chemother. 2014 Sep;69 Suppl 1:i25-7.

71. Archer NK, Mazaitis MJ, Costerton JW, Leid JG, Powers ME, Shirtliff ME. *Staphylococcus aureus* biofilms: Properties, regulation, and roles in human disease. Virulence.2011 Sep-Oct;2(5):445-59.

72. Mandell LA. Community-acquired pneumonia: An overview. Postgrad Med. 2015 Aug;127(6):607-15.

73. Gillet Y, Issartel B, Vanhems P, Fournet JC, Lina G, Bes M, Vandenesch F, Piémont Y, Brousse N, Floret D, Etienne J. Association between *Staphylococcus aureus* strains carrying gene for Panton-Valentine leukocidin and highly lethal necrotising pneumonia in young immunocompetent patients. Lancet. 2002 Mar 2;359(9308):753-9.

74. American Thoracic Society; Infectious Diseases Society of America. Guidelines for the management of adults with hospital-acquired, ventilator-associated, and healthcare-associated pneumonia. Am J Respir Crit Care Med. 2005 Feb 15;171(4):388-416.

75. Forsblom E, Ruotsalainen E, Mölkänen T, Ollgren J, Lyytikäinen O, Järvinen A. Predisposing factors, disease progression and outcome in 430 prospectively followed patients of healthcare- and community-associated *Staphylococcus aureus* bacteraemia. J Hosp Infect. 2011 Jun;78(2):102-7

76. Morris AK, Russell CD. Enhanced surveillance of *Staphylococcus aureus* bacteraemia to identify targets for infection prevention. J Hosp Infect. 2016 Jun;93(2):169-74.

77. Lenz R, Leal JR, Church DL, Gregson DB, Ross T, Laupland KB. The distinct category of healthcare associated bloodstream infections. BMC Infect Dis. 2012 Apr 9;12:85.

78. Mediavilla JR, Chen L, Mathema B, Kreiswirth BN. Global epidemiology of community-associated methicillin resistant *Staphylococcus aureus* (CA-MRSA). Curr Opin Microbiol. 2012 Oct;15(5):588-95.

79. Klevens RM, Morrison MA, Nadle J, Petit S, Gershman K, Ray S, Harrison LH, Lynfield R, Dumyati G, Townes JM, Craig AS, Zell ER, Fosheim GE, McDougal LK, Carey RB, Fridkin SK; Active Bacterial Core surveillance (ABCs) MRSA Investigators. Invasive methicillin-resistant *Staphylococcus aureus* infections in the United States. JAMA. 2007 Oct 17;298(15):1763-71.

80. Casadevall A, Fang FC, Pirofski LA. Microbial virulence as an emergent property: Consequences and opportunities PLoS Pathog 2011 7(7):e1002136.

81. Brown SP, Cornforth DM, Mideo N. Evolution of virulence in opportunistic pathogens: Generalism, plasticity, and control Trends Microbiol. 2012. 20(7):336-42.

82. Wertheim HF, Vos MC, Ott A, van Belkum A, Voss A, Kluytmans JA, van Keulen PH, Vandenbroucke-Grauls CM, Meester MH, Verbrugh HA. Risk and outcome of nosocomial *Staphylococcus aureus* bacteraemia in nasal carriers versus non-carriers Lancet 2004 Aug 21-27;364(9435):703-5.

83. von Eiff C, Becker K, Machka K, Stammer H, Peters G. Nasal carriage as a source of *Staphylococcus aureus* bacteremia. Study Group. N Engl J Med. 2001 Jan 4;344(1):11-6.

84. Bode LG, Kluytmans JA, Wertheim HF, Bogaers D, Vandenbroucke-Grauls CM, Roosendaal R, Troelstra A, Box AT, Voss A, van der Tweel I, et al. Preventing surgical-site infections in nasal carriers of *Staphylococcus aureus* N Engl J Med 2010 Jan 7;362(1):9-17.

85. Huang SS, Septimus E, Kleinman K, Moody J, Hickok J, Avery TR, Lankiewicz J, Gombosev A, Terpstra L, Hartford F, et al. Targeted versus universal decolonization to prevent ICU infection N Engl J Med 2013 Jun 13;368(24):2255-65.

86. Young BC, Golubchik T, Batty EM, Fung R, Larner-Svensson H, Votintseva AA, Miller RR, Godwin H, Knox K, Everitt RG, et al. Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. Proc Natl Acad Sci U S A 2012;109(12):4550.

87. Lei MG, Cue D, Roux CM, Dunman PM, Lee CY. Rsp inhibits attachment and biofilm formation by repressing fnbA in *Staphylococcus aureus* MW2. J Bacteriol 2011;193(19):5231.

88. Paterson GK, Harrison EM, Gemma GRM, Welch JJ, Warland JH, Matthew TGH, Fiona JEM, Ba X, Koop G, Harris SR, et al. Capturing the cloud of diversity reveals complexity and heterogeneity of MRSA carriage, infection and transmission. Nature Communications 2015;6:6560.

89. Alonzo F,3rd, Torres VJ. The bicomponent pore-forming leucocidins of *Staphylococcus aureus* Microbiol Mol Biol Rev 2014 Jun;78(2):199-230.

90. Foster TJ. Immune evasion by staphylococci Nat Rev Microbiol 2005 Dec;3(12):948-58.

91. Foster TJ, Geoghegan JA, Ganesh VK, Höök M. Adhesion, invasion and evasion: The many functions of the surface proteins of *Staphylococcus aureus* Nature Reviews Microbiology 2013;12(1):49 - 62.

92. Wu HJ, Wang AH, Jennings MP. Discovery of virulence factors of pathogenic bacteria. Curr Opin Chem Biol. 2008 Feb;12(1):93-101.

93. Wassenaar TM, Gaastra W. Bacterial virulence: can we draw the line? FEMS Microbiol Lett. 2001 Jul 10;201(1):1-7.

94. Spaulding AR, Salgado-Pabón W, Kohler PL, Horswill AR, Leung DY, Schlievert PM. Staphylococcal and streptococcal superantigen exotoxins. Clin Microbiol Rev. 2013 Jul;26(3):422-47

95. Schlievert PM, Shands KN, Dan BB, Schmid GP, Nishimura RD. Identification and characterization of an exotoxin from *Staphylococcus aureus* associated with toxic-shock syndrome. J Infect Dis. 1981 Apr;143(4):509-16

96. Lina G, Gillet Y, Vandenesch F, Jones ME, Floret D, Etienne J. Toxin involvement in staphylococcal scalded skin syndrome. Clin Infect Dis. 1997 Dec;25(6):1369-73

97. Otto M. *Staphylococcus aureus* toxins. Curr Opin Microbiol. 2014 Feb;17:32-7

98. Argudín MÁ, Mendoza MC, Rodicio MR. Food poisoning and *Staphylococcus aureus* enterotoxins. Toxins(Basel). 2010 Jul;2(7):1751-73.

99. Thammavongsa V, Kim HK, Missiakas D, Schneewind O. Staphylococcal manipulation of host immune responses Nat Rev Microbiol 2015 Sep;13(9):529-43.

100. Kim HK, Thammavongsa V, Schneewind O, Missiakas D. Recurrent infections and immune evasion strategies of *Staphylococcus aureu*s. Curr Opin Microbiol. 2012 Feb;15(1):92-9

101. Becherelli M, Prachi P, Viciani E, et al. Protective Activity of the CnaBE3 Domain Conserved among *Staphylococcus aureus* Sdr Proteins. Diep BA, ed. *PLoS ONE*. 2013;8(9):e74718.

102. Lee LY, Miyamoto YJ, McIntyre BW, Höök M, McCrea KW, McDevitt D, Brown EL. The Staphylococcus aureus Map protein is an immunomodulator that interferes with T cell-mediated responses. J Clin Invest. 2002 Nov;110(10):1461-71.

103. Garcia BL, Zwarthoff SA, Rooijakkers SH, Geisbrecht BV. Novel Evasion Mechanisms of the Classical Complement Pathway. J Immunol. 2016 Sep 15;197(6):2051-6.

104. Ibarra JA, Pérez-Rueda E, Carroll RK, Shaw LN. Global analysis of transcriptional regulators in *Staphylococcus aureus*. BMC Genomics 2013;14:126.

105. Painter KL, Krishna A, Wigneshweraraj S, Edwards AM. What role does the quorum-sensing accessory gene regulator system play during *Staphylococcus aureus* bacteremia? Trends Microbiol 2014 Dec;22(12):676-85.

106. Novick RP. Autoinduction and signal transduction in the regulation of staphylococcal virulence. Mol Microbiol 2003;48(6):1429-49.

107. Cheung GY, Wang R, Khan BA, Sturdevant DE, Otto M. Role of the accessory gene regulator agr in community-associated methicillin-resistant *Staphylococcus aureus* pathogenesis. Infect Immun. 2011 May;79(5):1927-35.

108. Majerczyk CD, Sadykov MR, Luong TT, Lee C, Somerville GA, Sonenshein AL. *Staphylococcus aureus* CodY negatively regulates virulence gene expression. J Bacteriol. 2008 Apr;190(7):2257-65.

109. Priest NK, Rudkin JK, Feil EJ, van den Elsen JM, Cheung A, Peacock SJ, Laabei M, Lucks DA, Recker M, Massey RC. From genotype to phenotype: Can systems biology be used to predict *Staphylococcus aureu*s virulence? Nat Rev Microbiol 2012 Nov;10(11):791-7.

110. Tuchscherr L, Loffler B. *Staphylococcus aureus* dynamically adapts global regulators and virulence factor expression in the course from acute to chronic infection Curr Genet 2016 Feb;62(1):15-7.

111. Atwood DN, Loughran AJ, Courtney AP, Anthony AC, Meeker DG, Spencer HJ, Gupta RK, Lee CY, Beenken KE, Smeltzer MS. Comparative impact of diverse regulatory loci on *Staphylococcus aureus* biofilm formation. Microbiologyopen. 2015 Jun;4(3):436-51.

112. Jarraud S, Lyon GJ, Figueiredo AMS, Gerard L, Vandenesch F, Etienne J, Muir TW, Novick RP. Exfoliatin- producing strains define a fourth agr specificity group in *Staphylococcus aureus*. The Journal of Bacteriology 2000;182(22):6517.

113. Messina JA, Thaden JT, Sharma-Kuinkel BK, Fowler VG,Jr. Impact of bacterial and human genetic variation on *Staphylococcus aureus* infections PLoS Pathog 2016 Jan 14;12(1):e1005330.

114. Peacock SJ, Moore CE, Justice A, Kantzanou M, Story L, Mackie K, O'Neill G, Day NP. Virulent combinations of adhesin and toxin genes in natural populations of *Staphylococcus aureus*. Infect Immun. 2002 Sep;70(9):4987-96.

115. Lindsay JA, Moore CE, Day NP, Peacock SJ, Witney AA, Stabler RA, Husain SE, Butcher PD, Hinds J. Microarrays reveal that each of the ten dominant lineages of *Staphylococcus aureus* has a unique combination of surface-associated and regulatory genes J Bacteriol 2006 Jan;188(2):669-76.

116. Rodrigues MV, Branco Fortaleza CM, Martins Souza CS, Teixeira NB, Ribeiro de Souza da Cunha Mde L. Genetic Determinants of Methicillin Resistance and Virulence among *Staphylococcus aureus* Isolates Recovered from Clinical and Surveillance Cultures in a Brazilian Teaching Hospital. ISRN Microbiol. 2012 May 17;2012:975143.

117. Bouchiat C, Moreau K, Devillard S, Rasigade J, Mosnier A, Geissmann T, Bes M, Tristan A, Lina G, Laurent F, et al. Staphylococcus aureus infective endocarditis versus bacteremia strains: Subtle genetic differences at stake Infect Genet Evol. 2015 Dec;36:524-530.

118. Calderwood MS, Desjardins CA, Sakoulas G, Nicol R, Dubois A, Delaney ML, Kleinman K, Cosimi LA, Feldgarden M, Onderdonk AB, Birren BW, et al. Staphylococcal enterotoxin P predicts bacteremia in hospitalized patients colonized with methicillin-

resistant *Staphylococcus aureus*. J Infect Dis. 2014 Feb 15;209(4):571-7.

119. Bhatty M, Ray P, Singh R, Jain S, Sharma M. Presence of virulence determinants amongst *Staphylococcus aureus* isolates from nasal colonization, superficial & invasive infections. Indian J Med Res. 2013;138:143-6.

120. Van Belkum A, Melles DC, Snijders SV, van Leeuwen WB, Wertheim HF, Nouwen JL, Verbrugh HA, Etienne J. Clonal distribution and differential occurrence of the enterotoxin gene cluster, egc, in carriage- versus bacteremia-associated isolates of *Staphylococcus aureus*. J Clin Microbiol. 2006 Apr;44(4):1555-7.

121. Rasmussen G, Monecke S, Ehricht R, Söderquist B. Prevalence of clonal complexes and virulence genes among commensal and invasive *Staphylococcus aureus* isolates in Sweden. PLoS One. 2013 Oct 9;8(10):e77477.

122. Lower SK, Lamlertthon S, Casillas-Ituarte NN, Lins RD, Yongsunthon R, Taylor ES, DiBartola AC, Edmonson C, McIntyre LM, Reller LB, et al. Polymorphisms in fibronectin binding protein A of *Staphylococcus aureu*s are associated with infection of cardiovascular devices Proc Natl Acad Sci U S A 2011 Nov 8;108(45):18372-7.

123. Hos NJ, Rieg S, Kern WV, Jonas D, Fowler VG, Higgins PG, Seifert H, Kaasch AJ. Amino acid alterations in fibronectin binding protein A (FnBPA) and bacterial genotype are associated with cardiac device related infection in *Staphylococcus aureus* bacteraemia. J Infect. 2015 Feb;70(2):153-9.

124. Labandeira-Rey M, Couzon F, Boisset S, Brown EL, Bes M, Benito Y, Barbu EM, Vazquez V, Hook M, Etienne J, et al. *Staphylococcus aureus* Panton-Valentine Leukocidin causes necrotizing pneumonia Science. 2007 Feb 23;315(5815):1130-3.

125. Villaruz AE, Bubeck Wardenburg J, Khan BA, Whitney AR, Sturdevant DE, Gardner DJ, DeLeo FR, Otto M. A point mutation in the agr locus rather than expression of the Panton-Valentine leukocidin caused previously reported phenotypes in *Staphylococcus aureus* pneumonia and gene regulation J Infect Dis 2009 Sep 1;200(5):724-34.

126. Torrell E, Molin D, Tabno E, Ehrenborg C, Ryden C. Community- acquired pneumonia and bacteraemia in healthy young woman caused by methicillin-resistant *Staphylococcus aureus* (MRSA) carrying the genes encoding Panton-Valentine leukocidin (PVL). Scand J infect Dis 2005; 7: 902-4.

127. Micek T, Dunne M, Kollef MH. Pleuropulmonary complications and Panton-Valentine Leucocidin-positive community-acquired methicillin- resistant *Staphylococcus aureus*: importance of treatment with antimicrobials inhibiting exotoxin production. Chest 2005; 128: 2732-8.

128. Francis JS, Doherty MC, Lopatin U, et al. Severe community-onset pneumonia in healthy adults caused by methicillin-resistant *Staphylococcus aureus* carrying the Panton-Valentine leukocidin genes. Clin Infect Dis 2005; 40:100-7.

129. Shallcross LJ, Fragaszy E, Johnson AM, Hayward AC. The role of the Panton-Valentine leucocidin toxin in staphylococcal disease: A systematic review and meta-analysis Lancet Infect Dis 2013 Jan;13(1):43-54.

130. Bocchini CE, Hulten KG, Mason EO,Jr, Gonzalez BE, Hammerman WA, Kaplan SL. Panton-Valentine leukocidin genes are associated with enhanced inflammatory

response and local disease in acute hematogenous *Staphylococcus aureus* osteomyelitis in children Pediatrics. 2006 Feb;117(2):433-40.

131. Sina H, Ahoyo TA, Moussaoui W, Keller D, Bankole HS, Barogui Y, Stienstra Y, Kotchoni SO, Prevost G, Baba-Moussa L. Variability of antibiotic susceptibility and toxin production of *Staphylococcus aureus* strains isolated from skin, soft tissue, and bone related infections BMC Microbiol. 2013 Aug 8;13:188,2180-13-188.

132. Yu F, Yang L, Pan J, Chen C, Du J, Li Q, Huang J, Zhang X, Wang L. Prevalence of virulence genes among invasive and colonising *Staphylococcus aureus* isolates. J Hosp Infect. 2011 Jan;77(1):89-91.

133. Lalani T, Federspiel JJ, Boucher HW, Rude TH, Bae IG, Rybak MJ, Tonthat GT, Corey GR, Stryjewski ME, Sakoulas G, et al. Associations between the genotypes of *Staphylococcus aureus* bloodstream isolates and clinical characteristics and outcomes of bacteremic patients. J Clin Microbiol. 2008 Sep;46(9):2890-6.

134. Neuner EA, Casabar E, Reichley R, McKinnon PS. Clinical, microbiologic, and genetic determinants of persistent methicillin-resistant *Staphylococcus aureus* bacteremia. Diagn Microbiol Infect Dis. 2010 Jul;67(3):228-33.

135. Otto M. A MRSA-terious enemy among us: End of the PVL controversy? Nat Med 2011 Feb;17(2):169-70.

136. Day NP. Panton-Valentine leucocidin and staphylococcal disease Lancet Infect Dis 2013 Jan;13(1):5-6.

137. Fowler VG,Jr, Nelson CL, McIntyre LM, Kreiswirth BN, Monk A, Archer GL, Federspiel J, Naidich S, Remortel B, Rude T, et al. Potential associations between hematogenous complications and bacterial genotype in *Staphylococcus aureus* infection J Infect Dis 2007 Sep 1;196(5):738-47.

138. Melles DC, Gorkink RF, Boelens HA, Snijders SV, Peeters JK, Moorhouse MJ, van der Spek PJ, van Leeuwen WB, Simons G, Verbrugh HA, et al. Natural population dynamics and expansion of pathogenic clones of *Staphylococcus aureus*. J Clin Invest. 2004 Dec;114(12):1732-4.

139. Wertheim HF, van Leeuwen WB, Snijders S, Vos MC, Voss A, Vandenbroucke-Grauls CM, Kluytmans JA, Verbrugh HA, van Belkum A. Associations between *Staphylococcus aureus* Genotype, Infection, and In-Hospital Mortality: A Nested Case-Control Study. J Infect Dis. 2005 Oct 1;192(7):1196-2000.

140. Nienaber JJ, Sharma Kuinkel BK, Clarke-Pearson M, Lamlertthon S, Park L, Rude TH, Barriere S, Woods CW, Chu VH, Marín M, et al. Methicillin-susceptible *Staphylococcus aureus* endocarditis isolates are associated with clonal complex 30 genotype and a distinct repertoire of enterotoxins and adhesins. J Infect Dis. 2011 Sep 1;204(5):704-13.

141. Draghici S, Khatri P, Eklund AC, Szallasi Z. Reliability and reproducibility issues in DNA microarray measurements. Trends Genet. 2006 Feb;22(2):101-9.

142. Andersen PS, Pedersen JK, Fode P, Skov RL, Fowler VG Jr, Stegger M, Christensen K. Influence of host genetics and environment on nasal carriage of *Staphylococcus aureus* in danish middle-aged and elderly twins. J Infect Dis. 2012 Oct;206(8):1178-84.

143. Nelson CL, Pelak K, Podgoreanu MV, Ahn SH, Scott WK, Allen AS, Cowell LG, Rude TH, Zhang Y, Tong A, et al. A genome-wide association study of variants associated with acquisition of *Staphylococcus aureus* bacteremia in a healthcare setting BMC Infect Dis 2014 Feb 13;14:83,2334-14-83.

144. Ye Z, Vasco DA, Carter TC, Brilliant MH, Schrodi SJ, Shukla SK. Genome wide association study of SNP-, gene-, and pathway-based approaches to identify genes influencing susceptibility to *Staphylococcus aureus* infections Front Genet 2014 May 9;5:125.

145. DeLorenze GN, Nelson CL, Scott WK, Allen AS, Ray GT, Tsai A, Quesenberry CP, Fowler VG. Polymorphisms in HLA class II genes are associated with susceptibility to *Staphylococcus aureu*s Infection in a white population J Infect Dis 2016 Mar 1;213(5):816-23.

146. Tian C, Hinds DA, Hromatka BS, Kiefer AK, Eriksson N, Tung JY. Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. 2016, Sep 9; bioRxiv preprint.

147. Nouwen J, Boelens H, van Belkum A, Verbrugh H. Human factor in *Staphylococcus aureus* nasal carriage. Infect Immun. 2004 Nov; 72: 6685-8

148. Thammavongsa V, Missiakas DM, Schneewind O. *Staphylococcus aureus* degrades neutrophil extracellular traps to promote immune cell death Science 2013 Nov 15;342(6160):863-6.

149. O'Keeffe KM, Wilk MM, Leech JM, Murphy AG, Laabei M, Monk IR, Massey RC, Lindsay JA, Foster TJ, Geoghegan JA, et al. Manipulation of autophagy in phagocytes facilitates *Staphylococcus aureus* bloodstream infection Infect Immun 2015 Sep;83(9):3445-57.

150. Tuchscherr L, Medina E, Hussain M, Volker W, Heitmann V, Niemann S, Holzinger D, Roth J, Proctor RA, Becker K, et al. *Staphylococcus aureus* phenotype switching: An effective bacterial strategy to escape host immune response and establish a chronic infection EMBO Mol Med 2011 Mar;3(3):129-41

151. Cheung GY, Kretschmer D, Duong AC, Yeh AJ, Ho TV, Chen Y, Joo HS, Kreiswirth BN, Peschel A, Otto M. Production of an attenuated phenol-soluble modulin variant unique to the MRSA clonal complex 30 increases severity of bloodstream infection PLoS Pathog 2014 Aug 21;10(8):e1004298

152. Traber KE, Lee E, Benson S, Corrigan R, Cantera M, Shopsin B, Novick RP. Agr function in clinical *Staphylococcus aureus* isolates. Microbiology 2008;154(8):2265-74.

153. Fowler VG, Sakoulas G, Mcintyre LM, Meka VG, Arbeit RD, Cabell CH, Stryjewski ME, Eliopoulos GM, Reller LB, Corey GR, et al. Persistent bacteremia due to methicillin-resistant *Staphylococcus aureus* infection is associated with agr dysfunction and low-level in vitro resistance to thrombin- induced platelet microbicidal protein. J Infect Dis 2004;190(6):1140.

154. Rose HR, Holzman RS, Altman DR, Smyth DS, Wasserman GA, Kafer JM, Wible M, Mendes RE, Torres VJ, Shopsin B. Cytotoxic virulence predicts mortality in nosocomial pneumonia due to methicillin-resistant *Staphylococcus aureus* J Infect Dis 2015 Jun 15;211(12):1862-74.

155. Smeltzer MS. Staphylococcus aureus Pathogenesis: The Importance of Reduced Cytotoxicity. Trends Microbiol. 2016 Sep;24(9):681-2.

156. Laabei M, Uhlemann AC, Lowy FD, Austin ED, Yokoyama M, Ouadi K, Feil E, Thorpe HA, Williams B, Perkins M, et al. Evolutionary trade-offs underlie the multi-faceted virulence of *Staphylococcus aureus* PLoS Biol 2015 Sep 2;13(9):e1002229.

157. van Belkum A, Melles DC, Nouwen J, van Leeuwen WB, van Wamel W, Vos MC, Wertheim HF, Verbrugh HA. Co-evolutionary aspects of human colonisation and infection by *Staphylococcus aureus*. Infect Genet Evol. 2009 Jan;9(1):32-47.

158. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. Within-host evolution of bacterial pathogens Nat Rev Microbiol. 2016. 14(3):150-62.

159. Golubchik T, Batty EM, Miller RR, Farr H, Young BC, Larner-Svensson H, Fung R, Godwin H, Knox K, Votintseva A, et al. Within- host evolution of *Staphylococcus aureus* during asymptomatic carriage. PloS One 2013;8(5):e61319.

160. Gao W, Chua K, Davies JK, Newton HJ, Seemann T, Harrison PF, Holmes NE, Rhee HW, Hong JI, et al. Two novel point mutations in clinical *Staphylococcus aureus* reduce linezolid susceptibility and switch on the stringent response to promote persistent infection. PLoS Pathog. 2010 Jun 10;6(6):e1000944.

161. Howden BP, McEvoy CR, Allen DL, Chua K, Gao W, Harrison PF, Bell J, Coombs G, Bennett-Wood V, Porter JL, et al. Evolution of multidrug resistance during *Staphylococcus aureus* infection involves mutation of the essential two component regulator WalKR. PLoS Pathog. 2011 Nov;7(11):e1002359

162. Lieberman TD, Michel JB, Aingaran M, Potter-Bynoe G, Roux D, Davis MR,Jr, Skurnik D, Leiby N, LiPuma JJ, Goldberg JB, et al. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes Nat Genet 2011 Nov 13;43(12):1275-80.

163. Lieberman TD, Flett KB, Yelin I, Martin TR, McAdam AJ, Priebe GP, Kishony R. Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures Nat Genet 2014 Jan;46(1):82-7.

164. McAdam PR, Holmes A, Templeton KE, Fitzgerald JR. Adaptive evolution of *Staphylococcus aureus* during chronic endobronchial infection of a cystic fibrosis patient PLoS One 2011;6(9):e24301.

165. Andersen SB, Marvig RL, Molin S, Krogh Johansen H, Griffin AS. Long-term social dynamics drive loss of function in pathogenic bacteria Proc Natl Acad Sci U S A. 2015. Aug 25;112(34):10756-61.

166. Sandoz KM, Mitzimberg SM, Schuster M. Social cheating in *Pseudomonas aeruginosa* quorum sensing Proc Natl Acad Sci U S A 2007 Oct 2;104(40):15876-81

167. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, et al. Patterns of somatic mutation in human cancer genomes Nature 2007 Mar 8;446(7132):153-8.

168. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell

GR, Bolli N, Borg A, Borresen-Dale AL, et al. Signatures of mutational processes in human cancer Nature 2013 Aug 22;500(7463):415-21.

169. Muehlenbachs A, Bhatnagar J, Agudelo CA, Hidron A, Eberhard ML, Mathison BA, Frace MA, Ito A, Metcalfe MG, et al. Malignant Transformation of *Hymenolepis nana* in a Human Host. N Engl J Med. 2015 Nov 5;373(19):1845-52

170. Falkow S. Molecular Koch's postulates applied to microbial pathogenicity. Rev Infect Dis. 1988 Jul-Aug;10 Suppl 2:S274-6.

171. Falush D. Bacterial genomics: Microbial GWAS coming of age Nature Microbiology 2016;1(5):16059.

172. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017 Jul 6;101(1):5-22

173. Read TD, Massey RC. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. Genome Med. 2014 Nov 22;6(11):109

174. Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: Lessons from human GWAS Nat Rev Genet 2017 Jan;18(1):41-50.

175. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860-921.

176. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. Science. 2004 Apr 23;304(5670):581-4.

177. Dale JW and Park SF. Molecular Genetics of Bacteria. 4th ed. Chichester UK: John Wiley & Sons Ltd, 2004.

178. Earle SG, Wu CH, Charlesworth J, Stoesser N, Gordon NC, Walker TM, Spencer CC, Iqbal Z, Clifton DA, Hopkins KL, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies Nat Microbiol 2016 Apr 4;1:16041.

179. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010 Jul;11(7):459-63.

180. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, Bentley SD, Maiden MC, Parkhill J, Falush D. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter* Proc Natl Acad Sci U S A 2013 Jul 16;110(29):11923-7.

181. Chewapreecha C, Marttinen P, Croucher NJ, Salter SJ, Harris SR, Mather AE, Hanage WP, Goldblatt D, Nosten FH, Turner C, et al. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes PLoS Genet 2014 Aug 7;10(8):e1004547.

182. Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, Sloan TJ, Williams P, Endres JL, Bayles KW, Fey PD, et al. Predicting the virulence of MRSA from its genome sequence Genome Res 2014 May;24(5):839-49.

183. Recker M, Laabei M, Toleman MS, et al. Clonal differences in *Staphylococcus aureus* bacteraemia-associated mortality. Nat Microbiol. 2017 Aug 7. [Epub ahead of print].

184. Alam MT, Petit RA 3rd, Crispell EK, Thornton TA, Conneely KN, Jiang Y, Satola SW, Read TD. Dissecting vancomycin-intermediate resistance in *Staphylococcus aureus* using genome-wide association. Genome Biol Evol. 2014 Apr 30;6(5):1174-85.

185. Lees JA, Vehkala M, Valimaki N, Harris SR, Chewapreecha C, Croucher NJ, Marttinen P, Davies MR, Steer AC, Tong SY, et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes Nat Commun 2016 Sep 16;7:12797.

Chapter 3

3 Methods

3.1 Microbiological methods

3.1.1 Nasal swabs

For chapters 4 and 5, nasal swabs from carriage were performed as part of a study led by Dr Kyle Knox and Dr Ruth Miller of carriage in the community in Oxfordshire, UK. Further samples were collected in hospital inpatients either for routine screening of MRSA, or studies of in-hospital *S. aureus* carriage and transmission.

Community swabs were self-collected by study subjects. A study nurse demonstrated the method for swabbing the anterior nares, and a written information sheet was also provided. Subjects were instructed to place a dry, sterile cotton-tipped swab inside the nostril and rotate 3 times, then repeat on the next nostril with the same swab. Swabs were placed in charcoal and returned to the John Radcliffe Hospital by Royal Mail second-class post.[1]

Swabs collected for the in-hospital study were collected by hospital staff and infection control nurses. A dry sterile cotton-tipped swab was inserted into each nostril in turn, and rotated approximately 3 times. Swabs were placed in charcoal and returned to the clinical microbiology laboratory.

In all cases, MMM research assistants performed laboratory processing of swabs using a common standard operating procedure. Nasal swabs were incubated in 5%

NaCl broth overnight at 37°C, then streaked onto SASelect agar (BioRad) and incubated overnight at 37°C. *S. aureus* growth was identified by pink colony formation on SASelect, catalase testing, DNAase testing and slide agglutination testing (Staphaurex Plus; Oxoid, Basingstoke, United Kingdom). In cases where the results of these tests were unclear, a tube coagulase test was performed. Methicillin resistance was screened for by the zone of inhibition around an oxacillin antimicrobial susceptibility test disc (Becton Dickinson, Oxford, UK), and confirmed by Etest if unclear (Biomerieux, Marcy l'Etoile, France). If not proceeding immediately to DNA collection, a mixture of bacterial colonies (up to 50) were picked and stored in 0.9% saline with 15% glycerol at -80°C.

For DNA extraction of isolates previously confirmed to be *S. aureus*, a loopful of glycerol stock was streaked onto SASelect and incubated overnight at 37°C. *S. aureus* colonies were picked (multiple colonies for the study in chapter 4 a single colony for the studies in chapters 5 and 6), streaked each onto Columbia blood agar (CBA; Oxoid, Basingstoke, United Kingdom) and incubated overnight at 37°C for DNA extraction.

For chapter 6, nasal swabs were taken from two collections overseen by Prof Nick Day at the Angkor Hospital for Children (Siem Reap, Cambodia). The first was a collection taken from consecutive children attending the AHC over a one-month period September-October 2008.[2] Children attending the outpatient department had a swab taken of both nostrils on one occasion, and children admitted to the ward had twice weekly swabbing of nose, throat and axilla throughout their admission. Swabs were performed by study personnel swabbing both nares with a single swab, which was placed in sterile phosphate-buffered saline (PBS) and retuned to the clinical laboratory within 2 hours.

A second cohort of children were screened for *S. aureus* nasal carriage between the 2-7th of July 2012, including all children (≤16 years) attending as an outpatient at Angkor Hospital for Children whose parents consented for them to participate. Nasal swabs were performed on participants, using a sterile cotton-tipped swab (pre-moistened with PBS). The swab was rotated 3 times in each nostril in turn, then placed in a bottle containing sterile PBS. Samples were kept cool until plated in the laboratory within one hour.

In both studies, swabs were plated onto selective Mannitol Salt agar. *S. aureus* isolates were identified by golden colony colour, catalase and coagulase testing, and slide agglutination test (Staphaurex Plus; Oxoid, Basingstoke, United Kingdom). The Clinical Laboratory Standards Institute (CLSI) standards were followed for susceptibility testing and bacteria stored in tryptone soya broth (TSB) Glycerol at -80°C.

3.1.2 Clinical specimens

For chapter 4, clinical samples were handled by OUH and BSUH clinical microbiology laboratory staff according to standard laboratory operating procedures for pus, sterile site and blood cultures. *S. aureus* growth was confirmed according to standard operating procedures. When bacterial growth was confirmed as *S. aureus*, the primary culture plate was retrieved by me, Claire Gordon or James Price, and multiple colonies streaked onto CBA and incubated overnight at 37°C for DNA extraction. If not proceeding immediately to DNA collection, a mixture of bacterial colonies (up to 50) were picked and stored in 0.9% saline with 15% glycerol at -80°C.

In chapter 5, blood cultures were handled by OUH, BSUH and Plymouth Hospitals NHS trust clinical microbiology laboratory staff according to standard laboratory operating procedures for blood cultures, including organism identification and

antibiotic susceptibility testing. When *S. aureus* was identified in blood culture growth, clinical laboratories routinely kept isolates in glycerol stock at -80°C. For samples identified for inclusion in this study, this stock was identified, and a loopful streaked onto CBA and incubated overnight at 37°C. A single colony was picked and incubated overnight on CBA at 37°C for DNA extraction.

In chapter 6, pus and blood specimens from individuals with suspected pyomyositis were handled by AHC microbiology staff according to local clinical laboratory standard operating procedures. When *S. aureus* was identified in sterile sites or blood culture growth, clinical laboratories routinely kept a stock at -80°C. For specimens identified for inclusion in this study, stocks were retrieved from the routine freezer, and a loopful was streaked onto CBA and incubated overnight at 37°C. A single colony was picked and incubated overnight on CBA at 37°C for DNA extraction.

3.1.3 DNA extraction

DNA extraction was performed by me, Catrin Moore and MMM research assistants (see 1.4 for more details) on overnight growth of bacteria on CBA. Colonies were suspended in 0.9% saline and added to a commercial buffer and proteinase K (EDT and LDT; QuickGene; Fujifilm, Tokyo, Japan) with silica beads (Lysing Matrix B; MPBiomedicals, Santa Ana, CA), before mechanical lysis (FastPrep; MPBiomedicals, Santa Ana, CA). The mix was heated at 70°C for 10 minutes and centrifuged. DNA was extracted from the supernatant using a commercial kit (QuickGene; Fujifilm, Tokyo, Japan)

Purity of nucleic acids in the extract was confirmed by checking the ratio of absorbance at 260 nm and 230 nm on spectrophotometer (NanoDrop; Thermo Fisher Scientific, Waltham, USA) and DNA flurometric quantification was performed using a commercial kit (Quant-iT PicoGreen dsDNA Assay; Thermo Fisher Scientific,

Waltham, USA).

3.1.4 Toxin Elisa

I quantified PVL production using an enzyme-linked immunosorbent assay (ELISA).

Capture antibody (0.25 ng per well of mouse derived monoclonal (ID9) antibody against lukS-PV (ab190656, Abcam, Cambridge UK)) was diluted in PBS and bound to a 96-well microtitre plate (Maxisorp Nunc-Immunoplate; Thermo Fisher Scientific, Waltham, USA) and left to bind at 4°C overnight. Wells were washed with a solution of PBS and 0.05% Tween (Sigma-Aldrich, St Louis, USA) (PBST). Wells were blocked with a solution of PBS plus 1% Bovine Serum Albumin (Sigma-Aldrich, St Louis, USA)(PBS-BSA) and incubated at room temperature for an hour, before washing again with PBST.

Bacterial strains were cultured on Blood Agar (Oxoid, Oxoid, Basingstoke, United Kingdom) overnight. 5 colonies from overnight growth were suspended in 10mL Tryptone Soya Broth (Oxoid, Oxoid, Basingstoke, United Kingdom) and incubated at 37°C overnight on a shaking incubator. Supernatant was collected and filtered through a 0.2 μm filter. A 250 μL aliquot was removed and 10 μL undiluted pooled healthy mouse serum (Sigma-Aldrich, St Louis, USA, M5905-5ML) was added to overcome non-specific IgG binding by Spa and Sbi. Supernatants were kept at 4°C prior to testing the same day.

50 μl of supernatant with mouse serum was added to wells, along with a standard curve of recombinant LukS-PV over a range of concentrations 10^{-5}-30 ng/ml (IBT Bioservices, Rockville, USA, 0530-001). Wells were washed with PBST, and a detection antibody was added (rabbit derived polyclonal antibody against lukS-PV (ab190473, Abcam, Cambridge UK)). Cells were washed with PBST, and secondary detection anbibody added (goat-derived anti-rabbit antibodies, conjugated with

horseradish peroxidase (Jackson Immunoresearch, West Grove USA, 111-035-144). After incubation wells were washed and enzyme substrate added (3, 3', 5, 5'-Tetramethylbenzidine (TMB); Sigma-Aldrich, St Louis, USA). The reaction was stopped with 2M sulphuric acid, and absorbance was measured at 450nm and 650nm on a CLARIOstar microplate reader (BMG Labtech, Offenburg, Germany).

3.2 Bacterial whole genome sequencing

3.2.1 Short read sequencing

Whole genome sequencing was performed on the Illumina platform for short read sequencing. Library preparation and sequencing were performed by staff at the Wellcome Trust Centre for Human Genetics, Oxford, on Illumina HiSeq 2000 and HiSeq 2500 platforms (San Diego, USA). Paired reads of 100bp (pilot work for Chapter 4, 394 earliest isolates for chapter 5), and 151 bp (all other sequences) were generated.

The protocols used by WTCHG for Illumina sequencing are in routine use. For Illumina HiSeq, DNA is fragmented into random segments less than 800bp, and adaptors are ligated to 5′ and 3′ ends of each fragment. The fragments are then amplified by PCR, using primers that are complimentary to the adaptor sequences. This process adds sequences to the end of the adaptors, which will later hybridise fragments to the flow cell, as well as index sequences, so that many DNA libraries can be pooled in a single flow cell. PCR also enriches for fragments with adaptors at both ends (ie those that will successfully form clusters on the flow cell), and relatively depletes adaptor dimers.[3]

This library is purified to remove unbound adaptors or adaptor pairs, then loaded onto a flow cell, which is covered by oligonucleotides complimentary to adaptor sequences. Sequences are amplified in situ by bridge amplification, which generates

"clusters" – hundreds of copies of identical DNA fragments – across the flow cell.[4]

DNA sequencing by synthesis then proceeds: fluorescently tagged nucleotides with a reversible terminator are added to the bound DNA fragments, one nucleotide per round. The signal of each bound nucleotide is recorded each round. The terminator and fluorescent signal are removed between rounds, before incorporating another single, tagged nucleotide. This process is termed 'cyclic reversible termination';[4] the number of cycles determines the length of the reads generated. Reads have adaptors and tag sequences removed before being returned for bioinformatics processing.

Samples for Illumina sequencing were, as far as possible, batched together so that samples for the same project were sequenced with the same sequencing chemistry and read length, and to avoid systematic differences in the handling of cases and controls. For sequences in chapter 4, carriage and disease isolates from the same individual were randomly distributed across a 96 well plate of template DNA for sequencing on single flow cell. In total 15 plates of DNA were sequenced: 3 with samples for a pilot study in 2011, and 12 further plates in 2013. For chapter 5, all controls and 289 cases were sequenced randomly distributed across 15 plates in 2014. 417 cases came from a collection of blood culture isolates that had already been sequenced for other projects. A further 321 cases were subsequently collected and sequenced in 2015. We used 5% technical replication throughout for quality control. For chapter 6, case and control sequences were randomly distributed across 7 plates sequenced in 2016.

3.2.2 Mapping and *de novo* assembly of short read sequencing

Sequence reads were processed by colleagues operating the Modernising Medical Microbiology (MMM) bioinformatics pipeline, with additional project-specific mapping and assembly where required.

For reference-based mapping of reads, the pipeline used Stampy[4] to align paired reads against a reference genome. Stampy takes Illumina sequence reads and 'maps' them to a given reference, inferring the location on the genome from which the read most likely originated. Stampy employs an algorithm designed to optimise speed without loss of sensitivity in the presence of sequence variation, and uses a statistical model to produce a mapping quality score, reflecting the likelihood a read has mapped correctly. We used Stampy v1.0.22 as part of the MMM pipeline with no pre-mapping, an expected substitution rate of 0.01 and otherwise default settings for Stampy. In the standard processing of *S. aureus* reads, the MMM pipeline maps *S. aureus* reads to the reference genome MRSA252[6], a 2.9 Mb methicillin-resistant *S. aureus* isolated from an individual with hospital-acquired MRSA infection and part of the EMRSA-16 clone and ST-36.

Reference based read-mapping has several limitations. Firstly, where there are repeated sequences in the genome, reads may map equally well to multiple locations.[4] It is therefore difficult to accurately detect differences between the repeated sequences. In order to avoid erroneous variant calls arising from incorrect mapping to repeated regions, we identified repeated regions in the reference genome (identified by comparing the genome to itself with the alignment tool BLAST[7]), and masked these regions prior to variant calling.

Reference-based methods depend upon homology between the reference genome and the sequenced genome. Mapping algorithms can align reads in the presence of some variation, but when there is significant difference (eg sequence divergence or accessory genome content), reads from regions not represented in the reference genome either map incorrectly or fail to align.[8]

In order to overcome this limitation, when maximal sensitivity for detection of variation between genomes was desired, we employed patient-specific references, built from the *de novo* assembly of reads from an example genome, with contigs ordered against a completed reference sequence using Mauve[9], and repeat regions masked. In chapter 4, patient-specific references were created for each patient. In chapter 6, a clonal complex specific reference was created for CC121.

Assembly of reads was performed with two different genome assemblers – Velvet[10] and Cortex[11]. This approach requires no reference genome, but rather assembles short reads into longer contiguous sequences (called 'contigs'). Reads are broken up into all possible 'words' of a given length *k* (also called kmers). The assembler creates a de Bruijn graph from these kmers, where each edge in the graph represents a kmer, and the sequence is described by the path that traverses the graph visiting each edge or kmer once.[12] Velvet was run as part of the MMM pipeline, using version v1.0.18 and running the VelvetOptimiser v2.1.7, which optimises kmer and cutoff values and generates the assembly. Cortex assemblies were built using Cortex v1.0.5.21 at two kmer lengths (31 and 61bp).

3.2.3 Variant identification and annotation

Variant calling was performed from both mapped and assembled sequences.

For mapping-based variant calling, we called single nucleotide variants using SAMtools[13] v1.19, executed as part of the MMM pipeline. A number of internal quality filters, which have previously been validated, were employed and base calls that passed these filters were used to call variants.[14] Quality filters assessed:

1. The type of variant (only single nucleotide polymorphisms were included, due to unreliability of calling indels from mapped reads)

2. Number of reads at the site (minimum of 5, with at least 1 in each direction)

3. Depth of high-quality mapped reads covering the site (between 2.5 and 97.5 percentiles for depth of coverage at all sites along the genome

4. Calls only in unique regions of the genome (not within the identified repeat regions (see 3.2.2))

5. Agreement of reads (minimum of 75% of reads supporting the call)

6. No heterozygosity (calls were homozygous under a diploid model)

I also applied the Cortex Variant caller (v1.0.5.20) to identify variation between assemblies of closely related isolates.[11] I employed the "runcalls" command that calls variants independently for each isolate, using a reference for variant calling and co-ordinate identification. This builds and cleans (ie checks for errors) graphs for specified kmer lengths (I used kmer lengths 31 and 61), and compares these to a draft reference (a *de novo* assembly built using Velvet of one isolate in the group to be compared).[15] Cortex caller compares the graphs of each isolate in the set being studied with the reference and identifies areas of divergence or 'bubbles', then genotypes all samples in the set. As regions of graph divergence can be of any length, and the reference and comparison sequence may differ in their length at the bubble, and Cortex has 90% power to detect indels of up to 100bp in length for haploid sites.[11] I removed any variants that had fewer than 5 reads to support the alternate call, calls with greater than 10% of reads supporting the reference, and calls that localized to repeat regions of the genome.

Variant annotation was performed to further describe all variants. For SNPs called by mapping against a complete and annotated reference strain, the effect of the SNP was determined by substituting the alternate call into the reading frame and determining the alternate codon (if the base was in a coding sequence). This was done using Python script written by Dr Tanya Golubchik (chapter 4) and R scripts

written by Prof Daniel Wilson (Chapters 5 and 6).

In chapter 4, when variation was found using a patient-specific reference, these variants were annotated by first aligning to a complete and annotated reference strain (MRSA252) using Mauve. If no aligned position in MRSA252 could be found, I used additional annotated references. Where variation was found using Cortex only, I annotated the variant by first locating it by comparing the flanking sequence to MRSA252 (and other annotated references) using BLAST. If a variant was found in a reference other than MRSA252, I identified any ortholog to that gene in MRSA252 using geneDB[16] and KEGG[17].

3.2.4 Multi-locus Sequence Typing and Antimicrobial resistance prediction

I performed Multi-locus sequence typing *in silico*. A BLAST database was built from the contigs assembled by Velvet. This database was interrogated using BLAST for each of the 7 MLST loci.[18] An allele was confirmed if any locus in the *S. aureus* MLST database (http://saureus.mlst.net) was found with 100% sequence identity. A sequence type was confirmed if a seven allele profile was found which matched a profile in the MLST database.

Anti-microbial resistance was predicted *in silico* as part of the MMM pipeline using a panel of genes and polymorphisms which underlie antibiotic resistance in *S. aureus*,[19] using Mykrobe[20], which predicts antimicrobial resistance from short read sequencing data by comparing a de Bruijn graph of sequencing reads to one comprising the catalogue of genotypic resistance determinants.

3.3 Population genetics

3.3.1 Phylogenetic tree-building

Phylogenetic trees depict the inferred evolutionary relationships between the

sequenced genomes, and hence record the structure and evolutionary history of the population. Maximum likelihood methods seek to find the tree that maximises the probability of the genetic data (known as the likelihood) for a given evolutionary model.

In chapter 4, maximum likelihood phylogenies for isolates sequenced from a single individual were constructed with the assumption of no homoplasy (i.e. no repeat or back mutation and no recombination, making all mutations unique events in the population history). This was done using R code written by Prof Daniel Wilson.

In chapters 5, and 6, phylogenies were constructed using the sequences from mapping of reads to a closed reference strain. I built maximum likelihood trees using RAxML[21], assuming a general time reversible (GTR) model. RAxML fixes a minimum branch length, and this may be longer than a single substitution event, so to overcome this, branch-length estimates were recalibrated using the no-recombination option in ClonalFrameML[22]. In chapter 4, a phylogeny of study isolates along with a reference panel of *S. aureus* genomes from bloodstream and carriage samples was built by Dr Chieh-Hsi Wu in a 2 part process for greater efficiency: the sequences were first clustered using a neighbour joining algorithm[23], then RAxML was used to resolve the relationships within clusters.

3.3.2 Ancestral state reconstruction

In chapter 4, the large phylogeny of study isolates and reference panel was used to identify closely-related "nearest neighbours" for the isolates sampled from each patient. I employed this nearest neighbour as an out-group, and used the tree to reconstruct the sequence of the MRCA of colonizing and infecting bacteria within each patient using a maximum likelihood method[24] in ClonalFrameML[22]. This allowed me to categorise the alternate alleles for each variant found on mapping to MRSA252 as either ancestral (ie wild type) or derived (ie mutant arising in host).

For all other variants, I repeated the Cortex variant calling process, this time including the nearest neighbour. I identified the ancestral allele as that found in the nearest neighbour. Variants which segregated carriage and disease populations were predicted to have arisen in the disease population if the derived allele was found in disease, and predicted to have arisen in the carriage population if the derived allele was found in carriage.

3.3.3 Pairwise genetic diversity

In chapter 4, I calculated mean pairwise diversity (π) for the carriage and disease sites of each individual separately: π was calculated as the mean number of variants differing between each pair of genomes. I compared distributions of π using a Mann-Whitney-Wilcoxon.

3.3.4 dN/dS within patient populations

A measure of selection pressures, the d_N/d_S ratio compares the rate of non-synonymous with synonymous mutations, adjusted for the preponderance of non-synonymous changes expected in the absence of selection pressures (ie under strictly neutral evolution) because of the low redundancy of the genetic code. When sequences are compared, a ratio significantly less than one implies purifying selection pressure, meaning that newly-arising mutations are disfavoured on average, and a ratio greater than 1 suggests diversifying selection pressure, meaning that newly-arising mutations are favoured on average.

For assessing the d_N/d_S ratio within patients, I adjusted the ratio of raw counts of total numbers of non-synonymous and synonymous SNPs by the ratio expected under strict neutrality. I employed an expected rate of non-synonymous mutation 4.9 times higher than that of synonymous mutation in *S. aureus* based on codon usage in MRSA252 and the observed transition:transversion ratio in non-coding SNPs that had been previously described.[29]

3.4.5 omegaMap analysis for species wide d_N/d_S

To compare signals of adaptation evident within patients with the species wide-patterns of selection, Prof Wilson estimated d_N/d_S ratios between unrelated *S. aureus* sequences. He identified homologous coding sequences by aligning each of 16 closed genome sequences against MRSA252 using Mauve. Coding sequences overlapping those in MRSA252 thus identified were multiply aligned using PAGAN[25]. These alignments were compared using omegaMap[26], a Bayesian method which estimates selection coefficients while accounting for the potentially confounding effects of recombination.[27] Variation of d_N/d_S within genes was estimated using Monte Carlo Markov chain run twice for 10,000 iterations each, assuming a mean of 30 codons sharing the same d_N/d_S. Other model assumptions were: equal codon frequencies, an exponential prior distribution on the population scaled mutation rate (θ) with mean 0.05, an exponential prior distribution on the transition:transversion ratio (κ) with mean 3, and an exponential prior distribution on the d_N/d_S ratio (ω) with mean 0.2. For each gene, we computed the posterior mean d_N/d_S ratio, averaging over variation in selection pressures across each gene.

3.4 Gene set enrichment analysis

In order to detect a higher than expected number of variants occurring in a single gene, I performed a gene enrichment analysis. I used a form of generalised linear model – a Poisson regression – because the observed data were counts of the number of mutations occurring in a given gene. This analysis requires the assumption that no recombination occurred within hosts, which appeared sound as evidence of within-host recombination was only rarely observed in a previous study,[29] and physically clustered variants (which might be due to recombination) had been de-duplicated to single events.

Under the null hypothesis, in which mutations arising within-host are equally likely to occur anywhere in the genome, the expected number of variants in any gene *j* is modeled by $\lambda_0 L_j$ (where λ_0 is the expected number of mutations per kb and L_j is the length of gene *j* in kb).

Under the alternate hypothesis, the number of mutations arising in a particular gene may be relatively depleted or enriched for mutations, so the number of mutations in a gene of interest *i* is modelled as $\lambda_i L_i$ while mutations in other genes is modelled as $\lambda_1 L_j$. Using an inbuilt function of the statistical package R[26], maximum likelihood estimation was performed for parameters λ_0, λ_i and λ_1. The goodness of fit of the null versus alternative hypothesis was tested using a likelihood ratio test with one degree of freedom, and the null hypothesis was rejected if the *p*-value was less than 0.05, after adjusting for multiple testing (see section 3.6).

In addition to testing each gene for enrichment, I performed a gene set enrichment analysis, testing for enrichment in groups of genes with shared ontological classification or with a shared response in expression to regulatory changes.[30] This idea was initially developed for analyzing RNA expression data in human cancer: by pooling variants in genes with a common function or shared regulation, investigators improved their power to detect effects and were able to identify signals in common across independent studies.[30]

I used data from two sources to identify gene sets. Firstly, ontological classification from the BioCyc database,[31] which classifies genes based on their cellular location, molecular functions and biological processes. Secondly Dr Wyllie suggested the SAMMD database,[32] which pools data from transcriptional studies of *S. aureus*, and characterizes each gene according to whether it is up-regulated, down-regulated or not changed in response to altered experimental conditions or genetic mutations. I identified all unique gene sets in these databases that grouped 2 or more genes

(552 Biocyc classifications and 248 SAMMD pathways). Biocyc grouping formed two sets (in the classification or not), while SAMMD groupings could have two or three sets (up- and/or down-regulated genes and undifferentially regulated genes).

Again I employed a Poisson regression with the same null hypothesis of no within gene enrichment for variants, so that in a gene *j*, the expected number of variants in a gene was modeled by $\lambda_0 L_j$ (where λ_0 is the expected number of mutations per kb and L_j is the length of gene *j* in kb). Under the alternative hypothesis, the number of variants seen in each group is modelled as $\lambda_1 L_j$, $\lambda_2 L_j$ or $\lambda_3 L_j$ (for sets with 2 or 3 groupings), where L_j is the total length of genes in the grouping. As before I used R for maximum likelihood estimation given the data for parameters λ_0, λ_1, λ_2 and λ_3. The goodness of fit of the null versus alternative hypothesis was tested using a likelihood ratio test with one or two degrees of freedom (depending on the number of groupings), and the null hypothesis was rejected if the *p*-value was less than 0.05, after adjusting for multiple testing (see section 3.6).

I performed Gene enrichment analysis and Gene set enrichment analysis using R scripts written by Prof Daniel Wilson.

3.5 Genome-wide association study in bacteria

3.5.1 Bacterial GWAS of SNPs and kmers using LMM to control for population structure

I performed bacterial genome-wide association testing using an R package bacterialGWAS (https://github.com/jessiewu/bacterialGWAS), which implements a published method for bacterial GWAS, and can test both SNP and non-SNP variation.[34] For SNP testing I used the sequence from read mapping. All sites with variability within the study population were identified as either biallelic (two alleles found in the population) or tri-or-tetra-allelic (three or four alleles found in

the population. Missing calls were imputed using ClonalFrameML.

I used a kmer-based approach to capture non-SNP variation.[35] Using the *de novo* assembled genome, all unique 31 base haplotypes were counted using dsk[36]. If a kmer was found in the assembly it was counted present for that genome, otherwise it was treated as absent.

The association of each SNP and kmer with the phenotype was tested, controlling for population structure and genetic background using a linear mixed model (LMM) implemented in the Genome-wide Efficient Mixed Model Association tool (GEMMA).[33]

The LMM models the phenotype as the sum of fixed effects of covariates, the fixed 'foreground' effect of the locus under investigation, the random 'background' effects of all loci and the random effects of the environment. This is expressed formally by the following:

$$y_i = w_{i1}\alpha_1 + \ldots + w_{ic}\alpha_c + X_{il}\beta_l + X_{il}\gamma_1 + \ldots + X_{il}\gamma_L + \varepsilon_i$$

This models the effect of a locus l on the phenotype y_i in individual i: w_{ij} is covariate j in individual i, α_j is the effect of covariate j and there are c covariates in total; X_{ij} is the genotype of locus j in individual i; β_l is the foreground effect of locus l, γ_j is the background effect of locus j (with L loci in total); ε_i is the effect of the environment (or error) on individual i.

Biallelic SNPs are used to model background effects because variants with 3 or 4 alleles are less common. The coefficients of the effects of background loci are assumed to follow independent normal distributions with a mean 0 variance $\lambda\tau^{-1}$ to be estimated. This assumption means the overall contribution of any given background locus is constrained to be small because most variants are unlikely to affect the phenotype, a statistical effect known as shrinkage. The environmental

effects are also treated as random effects following independent normal distributions with mean 0 and variance τ^{-1}.

GEMMA estimates the parameters of the model under maximum likelihood and performs a likelihood ratio test against the null hypothesis (that each SNP or kmer has foreground coefficient zero). GEMMA was run with no lower limit on minor allele frequency in order to include all variants. GEMMA was modified by Prof Wilson to output the ML log-likelihood for each SNP or kmer variant under the null and alternative, and $-\log_{10}$ *p*-values were calculated using R scripts in the bacterialGWAS package. GEMMA also outputs an estimate of overall heritability, or the proportion of variance in sample phenotypes that is explained by total genetic diversity.

3.5.2 Testing for lineage associations

I tested for associations between lineage and phenotype using an R package *bugwas* (available at https://github.com/sgearle/bugwas), which implements a published method for identifying lineage effects in bacterial GWAS.[34] The *bugwas* method identifies principle components from the relatedness matrix (which is used to model background effects in GEMMA) and tests the effect of each component against the null hypothesis of no background effect of each principle component using a Wald test.

3.5.1 Quantifying association with Odds ratio

Where an association was demonstrated between a genotype (eg SNP or kmer presence) and the phenotype of interest, the strength of this association was quantified using the odds ratio (OR). The OR of a genotype is calculated by taking the odds of finding genotype among cases and dividing this by the odds of finding the genotype among controls.

3.6 Model averaging with the use of harmonic mean *p*-values

The GWAS methods described above test the null hypothesis that no variant (SNP or kmer) is associated with bacterial phenotype against the alternate hypothesis that a specific SNP or kmer is associated with the phenotype of interest. Alternatively, one can hypothesise that at least one variant out of a group of candidates in a region (eg a gene or groups of genes) is associated with the phenotype. One can test this compound hypothesis by averaging the evidence for all the individual alternate hypotheses from that region.

A method for pooling evidence in a maximum likelihood setting, based on the use of harmonic mean *p*-values (HMP), has been developed recently.[37] The harmonic mean is the reciprocal of the arithmetic mean of the reciprocals of a set of values. Unlike Fisher's method for combining *p*-values, the HMP method does not require the assumption that values are independent.[37]

The evidence for the alternate hypothesis in a region of interest is determined by computing the harmonic mean of the *p*-values for each variant in that region. This HMP, which is directly interpretable when small (e.g. below 0.05), is used to compute a combined *p*-value that is compared to a significance threshold αw, which controls the family-wide error rate (FWER) at level α, where w is the proportion of all variants that are found in the region of interest.[37] *p*-values based on the HMP were computed using the R package "HarmonicMeanP" (available at https://github.com/cran/harmonicmeanp).

3.7 Accounting for multiple testing

A nominal false positive rate of 5% was set as the threshold for statistical significance. When multiple testing was performed, this was adjusted by applying a Bonferroni correction to control the FWER at the nominal false positive rate.[38]

When testing for enrichment of variants within genes, ontological classifications or transcriptional pathways, the threshold for significance was adjusted by a Bonferroni correction, weighted by the proportion of tests in each category.[39]

When testing the significance of a SNP, kmer or PC, we considered the effect of that kmer or PC to be significant if its *p*-value was lower than a false positive rate of 0.05, adjusted for multiple testing using Bonferroni. For PCs the number of tests performed was taken to be the number of PCs identified. For SNPS and kmers, I followed Earle and colleagues, taking the number of tests to be the number of unique phylogenetic patterns or 'phylopatterns', that is, the total number of ways in which alleles at genetic variants partition individuals into groups.[34] This reduces the multiple testing burden by only testing once for any given phylopattern. Since variants with the same phylopattern are in complete LD with one another, their effects on phenotype could not be teased apart statistically, and in that sense do not constitute separate tests.

References chapter 3

1. Miller RR, Walker AS, Godwin H, Fung R, Votintseva A, Bowden R, Mant D, Peto TE, Crook DW, Knox K. Dynamics of acquisition and loss of carriage of *Staphylococcus aureus* strains in the community: the effect of clonal complex. J Infect. 2014 May;68(5):426-39.

2. Nickerson EK, Wuthiekanun V, Kumar V, Amornchai P, Wongdeethai N, Chheng K, Chantratita N, Putchhat H, Thaipadungpanit J, Day NP, Peacock SJ. Emergence of community-associated methicillin-resistant *Staphylococcus aureus* carriage in children in Cambodia. Am J Trop Med Hyg. 2011 Feb;84(2):313-7.

3. Nextera XT DNA Library Prep Kit, Illumina, 2017. URL https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_nextera/nextera-xt/nextera-xt-library-prep-reference-guide-15031942-02.pdf

4. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010 Jan;11(1):31-46.

5. Lunter G, Goodson M. Stampy: A statistical algorithm for sensitive and fast mapping of illumina sequence reads Genome Res 2011 Jun;21(6):936-9.

6. Holden MTG, Feil EJ, Lindsay JA, Peacock SJ, Day NPJ, Enright MC, Foster TJ, Moore CE, Hurst L, Atkin R, et al. Complete genomes of two clinical *Staphylococcus aureus* strains: Evidence for the rapid evolution of virulence and drug resistance Proc Natl Acad Sci U S A. 2004;101(26):9786 -9791.

7. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool J Mol Biol 1990 Oct 5;215(3):403-10.

8. Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, Keim P, Morrow JB, Salit ML, Zook JM. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. Front Genet. 2015 Jul 7;6:235.

9. Darling AC, Mau B, Blattner FR, Perna NT. Mauve: Multiple alignment of conserved genomic sequence with rearrangements Genome Res. 2004; 14(7):1394-403.

10. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de bruijn graphs Genome Res 2008 May;18(5):821-9.

11. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet. 2012 Jan 8;44(2):226-32.

12. Compeau PE, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. Nat Biotechnol. 2011 Nov 8;29(11):987-91.

13. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9.

14. Didelot X, Eyre DW, Cule M, Ip CL, Ansari MA, Griffiths D, Vaughan A, O'Connor L, Golubchik T, Batty EM, et al. Microevolutionary analysis of *Clostridium difficile* genomes to investigate transmission Genome Biol 2012 Dec 21;13(12):R118,2012-13-12-r118

15. Iqbal Z, Turner I, McVean G. High-throughput microbial population genomics using the Cortex variation assembler. Bioinformatics. 2013 Jan 15;29(2):275-6.

16. Logan-Klumpler FJ, De Silva N, Boehme U, Rogers MB, Velarde G, McQuillan JA, Carver T, Aslett M, Olsen C, Subramanian S, et al. GeneDB--an annotation database for pathogens. Nucleic Acids Res. 2012 Jan;40(Database issue):D98-108.

17. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation Nucleic Acids Res. 2016; 44(D1):D457-62.

18. Enright MC, Day NP, Davies CE, Peacock SJ, Spratt BG. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus* J Clin Microbiol 2000 Mar;38(3):1008-15.

19. Gordon NC, Price JR, Cole K, Everitt R, Morgan M, Finney J, Kearns AM, Pichon B, Young B, Wilson DJ, et al. Prediction of *Staphylococcus aureus* antimicrobial resistance by whole-genome sequencing J Clin Microbiol 2014 Apr;52(4):1182-91.

20. Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, Earle S, Pankhurst LJ, Anson L, de Cesare M, et al. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. Nat Commun. 2015 Dec 21;6:10063.

21. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies Bioinformatics 2014 May 1;30(9):1312-3.

22. Didelot X, Wilson DJ. ClonalFrameML: Efficient inference of recombination in whole bacterial genomes PLoS Comput Biol 2015 Feb 12;11(2):e1004041.

23. Saitou N and Nei M. The neighbor-joining method: A new method for reconstructing phylogenetic trees Mol Biol Evol. 1987; 4(4):406-25.

24. Pupko T, Pe'er I, Shamir R, Graur D. A fast algorithm for joint reconstruction of ancestral amino acid sequences. Mol Biol Evol. 2000; 17(6), 890-6.

25. Loytynoja A, Vilella AJ, Goldman N. Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. Bioinformatics. 2012; 28(13), 1684-91.

26. Wilson DJ, McVean G. Estimating diversifying selection and functional constraint in the presence of recombination. Genetics. 2006;172(3), 1411-25.

27. Wilson DJ, Hernandez RD, Andolfatto P, Przeworski M. A population genetics-phylogenetics approach to inferring natural selection in coding sequences. PLoS Genet. 2011; 7(12), e1002395.

28. R Core Team. R: A Language and Environment for Statistical Computing, Vienna, Austria: R Foundation for Statistical Computing. 2015. URL https://www.R-project.org/

29. Golubchik T, Batty EM, Miller RR, Farr H, Young BC, Larner-Svensson H, Fung R,

Godwin H, Knox K, Votintseva A, et al. Within- host evolution of *Staphylococcus aureus* during asymptomatic carriage. PloS One 2013;8(5):e61319.

30. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50.

31. Caspi R, Billington R, Ferrer L,et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases Nucleic Acids Res. 2016 Jan 4;44(D1):D471-80.

32. Nagarajan V and Elasri M. SAMMD: Staphylococcus aureus microarray meta-database. BMC Genomics. 2007; 8(1):351.

33. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies Nat Genet. 2012 Jun 17;44(7):821-4.

34. Earle SG, Wu CH, Charlesworth J, Stoesser N, Gordon NC, Walker TM, Spencer CC, Iqbal Z, Clifton DA, Hopkins KL, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies Nat Microbiol 2016 Apr 4;1:16041.

35. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, Bentley SD, Maiden MC, Parkhill J, Falush D. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter* Proc Natl Acad Sci U S A 2013 Jul 16;110(29):11923-7.

36. Rizk G, Lavenier D, Chikhi R. DSK: K-mer counting with very low memory usage Bioinformatics 2013 Mar 1;29(5):652-3.

37. Wilson DJ, The harmonic mean p-value and model averaging by mean maximum likelihood. (Preprint BioRxiv, August 2 2017, doi: https://doi.org/10.1101/171751).

38. OJ Dunn. Estimation of the medians for dependent variables. Ann. Math. Stat. 1959 30, 192–197.

39. Roeder K, Wasserman L. Genome-Wide Significance Levels and Weighted Hypothesis Testing. Stat Sci. 2009 Nov;24(4):398-413.

Chapter 4

4 Within host evolution of *Staphylococcus aureus* in invasive disease reveals signatures of adaptation

4.1 Introduction

Staphylococcus aureus is a pathogen[1,2] with an armory of factors facilitating tissue invasion and evasion of host immune factors, including toxins[3], complement control proteins[4] and bound adhesins[5]. Many such factors are part of the variably present accessory genome.[6] While able to cause severe invasive disease, *S. aureus* - like many pathogens - is much more frequently a commensal organism. The prevalence of asymptomatic *S. aureus* nasal carriage is 30%;[1] the annual incidence of severe infection is 2 in 10,000.[2] Invasive disease is a relatively rare event that is unnecessary[7] and possibly adverse[8] for onward transmission. Carriage and invasion are highly related: 80% of infections are derived from nasally carried strains.[9]

We hypothesise that within-host evolution may act in this transition from commensalism to infection, and evidence for this hypothesis accrues from multiple sources. Within-host microdiversity is common in *S. aureus*, and changes arise rapidly enough to affect the host-pathogen balance.[10,11,12,13] Adaptation to antimicrobial selective pressures and the emergence of resistance – a process well documented in viruses such as HIV[14] – has been revealed in bacteria following the antibiotic treatment,[15] under experimental conditions[16] and in natural populations[17]. Changes to regulatory control is observed within hosts, with wide

ranging effects on bacterial phenotype.[17,18]

Sampling the transition from *S. aureus* carriage to bacteraemia, we previously reported narrowed diversity in bacteraemia and a regulatory gene change associated with the transition.[19] This gene was later identified as Repressor of surface proteins (Rsp).[20] Other investigators have reported diversity in a canine infection, with the most invasive samples distinguished by variants arising in the well-characterised accessory gene regulatory system (Agr).[12]

The wide-ranging effects of transcriptional regulators on gene expression mean that genomic changes can cause dramatic phenotypic changes.[15] Short-term evolution in bacteria on the scale occurring within hosts can cause mutation rates between two to five orders of magnitude greater than those seen over long periods of time, as bacteria exhibit in-host adaption, responding to selective pressures of antibiotic exposure and host immune response, or acquiring a competitive advantage over other bacteria.[21,22]

Such effects were demonstrated when we found isolates with naturally occurring loss-of-function variants in Rsp to have significantly attenuated cytotoxicity, haemolysis and lethality in a mouse model of infection.[23] These strains exhibited enhanced transcription of immune evasion proteins, but retained the capacity to disseminate and form abscesses in animal models. Thus genomic changes arising within-host can dramatically alter the *S. aureus* host-pathogen interaction.

In this chapter I present a systematic study of within-host evolution in invasive *S. aureus* disease. Comparing whole genome sequences of multiple isolates from nasal colonisation and infection, I comprehensively catalogued the extent and types of genomic change arising within infected hosts, providing insights into the adaptation evident in *S. aureus* disease.

4.2 Materials and Methods

We compared bacteria from 105 individuals with both *S. aureus* carriage and disease. Individuals with *S. aureus* cultured from both nasal swab and clinical samples were identified from the clinical laboratory of two UK hospitals. Clinical samples were cultured from blood, pus, soft tissue, bone or joint samples. This sample collection comprised 55 cases with bloodstream infection, and 50 with *S. aureus* cultured from a pus or tissue sample. Most individuals had *S. aureus* grown from a single infection site, but 5 individuals with bloodstream infection also had *S. aureus* cultured from an additional site: these were all considered bacteraemia cases. Nose and infection site were usually sampled on the same day (IQR 1 day prior-2 days later). 13/105 individuals (12.5%) had MRSA infection, and the remainder were MSSA.

Multiple colonies were sequenced from each nose or infection sample (12 in pilot work, then five) using Illumina Hi-Seq. In total 1162 high quality sequences were available for comparison. Variant detection was performed comparing all isolates from an individual using both mapping and *de novo* assembly.

In order to compare the findings observed within infected hosts to *S. aureus* evolution in other contexts, we assembled 5 reference panels. Reference panel I is a collection of 131 *S. aureus* isolates from a cross-sectional study of 13 individuals with *S. aureus* nasal carriage without infection.[10] Reference panel II is 1149 sequences, comprising a compilation of 95 unrelated samples from a carriage study of *S. aureus* in Oxfordshire (BioProject accession number PRJEB255),[24] 145 sequences from a study of within-host evolution of *S. aureus* in 3 individuals (BioProject PRJEB2892) [19] and 909 sequences from nasal carriage and bloodstream infection used in a study of whole genome sequencing to predict antimicrobial resistance (BioProject PRJEB6251).[25] We used these samples to improve our

reconstruction of ancestral genotypes in each patient. Reference panel III is a collection of 2001 *S. aureus* isolates sequenced from carriage and bacteraemia for a case control study (see Chapter 5 of this thesis). Reference panel IV is a collection of 237 genomes from longitudinal samples from 10 individuals with asymptomatic carriage.[10,26] Reference panel V is a collection of 16 closed reference genomes from human and animal samples: MRSA252 (Genbank accession number BX571856.1), MSSA476 (BX571857.1), COL (CP000046.1), NCTC 8325 (CP000253.1), Mu50 (BA000017.4), N315 (BA000018.3), USA300_FPR3757 (CP000255.1), JH1 (CP000736.1), Newman (AP009351.1), TW20 (FN433596.1), S0385 (AM990992.1), JKD6159 (CP002114.2), RF122 (AJ938182.1), ED133 (CP001996.1), ED98 (CP001781.1), EMRSA15 (HE681097.1). We used this panel to examine rates of change within genes at a species level.

4.3 Results

4.3.1 Extensive diversity exists during *S. aureus* disease, with colonising and invasive isolates forming related but distinct clades

Bacterial populations determined by whole-genome sequencing demonstrated a close relationship between carriage and disease (Table 4.1, Figure 4.1); only 10 individuals had unrelated lineages found at each site. When related, infecting and colonising bacteria usually formed distinct non-overlapping clades (74/95) with a mean of 5.7 substitutions separating them. There was never more than one overlapping genotype in those where the populations did not separate into clades (21/95 patients), consistent with a unique migration event and a tight bottleneck.[27]

	Relatedness of colonizing and infecting bacteria		
	Unrelated (≥1104 variants)	Closely related (≤66 variants)	
		Zero shared genotypes	One shared genotype
Bloodstream infections	4	43	8
Soft tissue infections	4	23	10
Bone & joint infections	2	8	3
Total	10	74	21

Table 4.1: Distribution of infection types and relatedness of nose-colonizing and invading *S. aureus* among 105 patients revealed by genomic comparison

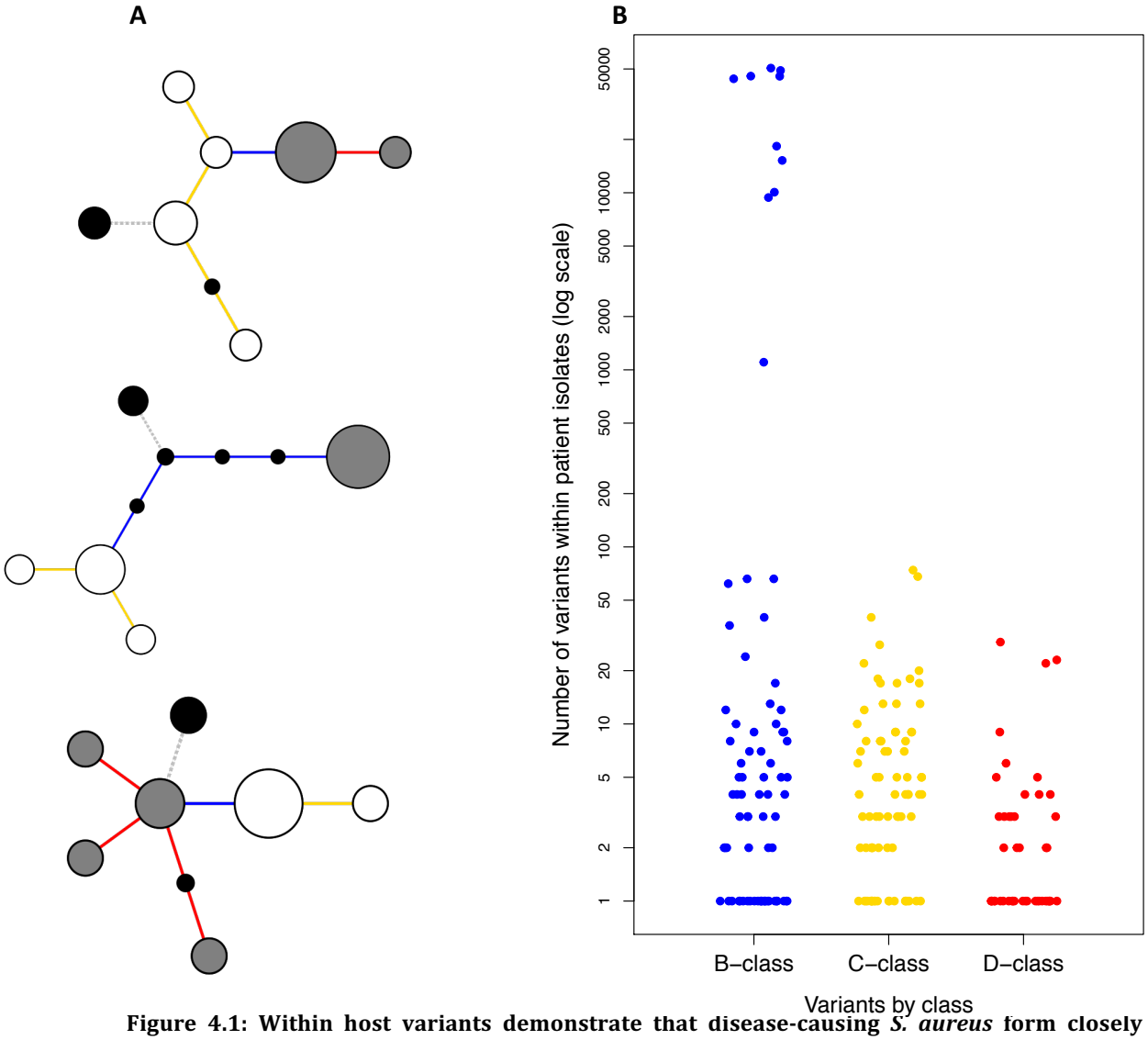


Figure 4.1: Within host variants demonstrate that disease-causing *S. aureus* form closely related but distinct populations from those found in carriage. (A). Example haplotrees. Nodes represent observed genotypes sampled from the nose (white) or infection site (grey), with area proportional to genotype frequency, or unobserved ancestral and intermediate genotypes (black). Edges represent mutations. Edge colour indicates that mutations occurring on those branches correspond to B-class substitutions between nose and infection site (blue), C-class variants among nose-colonizing bacteria (gold) or D-class variants among disease-causing bacteria (red). Branch to most-recent common ancestor represented by grey edge. **(B)**. Distribution of the number of variants found within patients according to the variant class

Within-host genomic diversity was found in 79 carriage populations, compared to 39 disease populations (Figure 4.2B). Up to 74 variants occurred within a single carriage population, and up to 29 in disease. Mean pairwise genetic diversity within carriage (π=2.8 variants) was similar to that reported in asymptomatic carriers (Reference Panel I[10], π = 4.1, p=0.13) but was significantly lower in disease (π=0.6, p=$10^{-10.0}$), and this reduction in diversity further supports a bottleneck during

invasion.

Using Reference panel II – a collection of sequences from other patients or carriers – to distinguish ancestral from derived alleles, we were able to determine the ancestral population in 49/95 individuals. In 80% (39/49) there was clear evidence that the nose population was ancestral, because all invading bacteria shared alleles that differed from both the carriage and ancestral allele. In the remaining 20% (10/50) the invading population appeared to be ancestral, indicating that nasal colonization is sometimes seeded from infecting bacteria. Confidence was high for just three of those patients, and they showed unusually high diversity, suggestive of persistent infections

4.3.2 Variants arising within-host show relaxed purifying selection and an increase in stop codons occurring during infection

To describe within-host variation, every variant was annotated for location within a coding sequence, and predicted effect on protein transcription (Table 4.2). Almost all variants arising within-host were unique: only two variants were observed to arise in two hosts each, both in non-coding regions.

Variants were annotated for their position on a Maximum Likelihood tree of each patient's isolates (Fig 4.1A): these were categorised according to whether they arose within the carriage isolates (C variants), within the disease isolates (D variants), or occurred on a branch separating the two populations (B variants) (Table 4.2). No significant differences were seen in the numbers of variants with different predicted effects on protein when comparing B, C and D variants. Protein truncating mutations occurred at a rate more than 3 times higher than that seen in a previous study of within-host evolution in asymptomatic *S. aureus* carriage.[10] This finding complements that of our previous report that an over-abundance of premature stop codons accompanied the transition from asymptomatic carriage to

invasive disease in a longitudinally sampled patient,[19] and gives further evidence that loss-of-function mutations are more common in infected patients. One possible explanation for this is a reduction in purifying selection, which would usually remove these mutations.

	B variants	C variants	D variants	Total		Asymptomatic carriers (Reference Panel I)[10]	
Non-coding	140	145	40	325	p=0.55[A]	45	p=1
Synonymous	93	93	26	213	p=0.50[A]	37	p=0.21[B]
Non-synonymous	265	325	82	672	p=0.57[A]	97	p=0.67[B]
Protein truncating	39	59	15	113	p=0.38[A]	5	p=0.01[B]**
Total	537	622	163	1322		184	
d_N/d_S	0.54	0.68	0.63			0.52	

Table 4.2: Variants by effect on protein and class arising within infected hosts. Number of each type of variant in each population, shown in comparison to that seen in stable carriage; Non-synonymous variants includes both amino-acid substitutions and short indels; Protein truncating variants include nonsense mutations and frameshift mutations with a premature stop codon; [A] Significance of difference in rate of each variant type occurring on different locations on the tree, Pearson's Chi-squared test, df = 2; [B] Significance of difference in rate observed in this study with rate of within-host variants in study of carriers with no *S. aureus* infections, Pearson's Chi-squared test with Yate's continuity correction, df =1.

Examining for varying selection pressures by comparing d_N/d_S, we found evidence of purifying selection on all branches of the tree with 0.54, 0.68 and 0.63 in B, C and D variants respectively, but this pressure was relatively relaxed, particularly in the nasal carriage population, compared with that seen in asymptomatic carriers (d_N/d_S =0.52). Along with the excess of protein truncating mutations, this provided evidence of reduced purifying selection on the bacterial genome during *S. aureus* infection.

4.3.3. B variants are enriched for changes to transcriptional regulators

We reasoned that B variants that change the predicted protein product should be enriched for any genomic changes that provide an advantage in infection. We had previously hypothesised these could be enriched for variants that affect transcriptional regulation,[19] and identified Rsp as a candidate gene for loss of

function mutations associated with invasion.

A single further Rsp variant was identified in this collection. One individual (P39N) had a SNP producing an alanine to proline substitution, predicted to disrupt the helix-turn-helix DNA binding domain. Transcriptional studies of these isolates undertaken in collaboration with Dr David Wyllie confirmed loss of expression of the key Rsp transcript SSR42 in the bloodstream isolate carrying this substitution.[23] Loss of function mutations to Rsp, though found within-host and associated with bloodstream infection, are therefore rare.

To identify loci under possible selection among the B variants we tested for a significant excess of variants predicted to affect the translated protein (ie either non-synonymous or protein truncating variants), combining variants by locus on MRSA252, adjusting for gene length (Table 4.3, Fig 4.2A). This aggregation by gene was required given than all but two variants in the study were unique to a single patient. Where the variant was described in a different reference genome, we used KEGG orthology of genes to identify orthologs within MRSA252. Orthologs in MRSA22 were found for 394/438 (90%) of protein altering variants with this method.

We found a significant excess of five protein-altering variants in *agrA*, the response regulator protein of the Agr system (p=$10^{-7.5}$), representing 58.3-fold enrichment for variants (Table 4.3, Fig 4.2). The *clfB* gene encoding clumping factor B, which binds human fibrinogen,[5] showed 15.9-fold enrichment with 5 protein-altering B variants, which approached significance after correction for multiple testing (p=$10^{-4.7}$). In contrast, C variants showed 19.0-fold enrichment with 6 protein altering variants in penicillin binding protein 2 (*pbp2*) – an enzyme in peptidoglycan synthesis and a site of action for beta-lactam antibiotics[28] (p=$10^{-6.0}$, Fig 4.3A). No loci were enriched for D variants (Figure 4.3), and no enrichment was found for

synonymous variants of any class (data not shown).

Gene group	No. protein-altering B-class variants		Cumulative length of genes (kb)		Enrichment		Significance ($-\log_{10} p$)
Locus							
agrA	5		0.7		58.27		**7.53**
clfB	5		2.6		15.87		4.70
Total	289		2363.8				
BioCyc Gene Ontology							
Cell wall	18		30.9		5.02		**7.03**
Cell adhesion	13		17.2		6.44		**6.47**
Pathogenesis	31		112.5		2.41		4.44
Total	288		2359.3				
SAMMD Expression Pathway	*Down-regulated*	*Up-regulated*	*Down-regulated*	*Up-regulated*	*Down-regulated*	*Up-regulated*	
Ovispirin-1	40	7	121.2	142.9	2.65	0.39	**7.80**
Temporin L	42	14	125.1	156.1	2.78	0.74	**6.86**
rsp	27	1	61.1	13.7	3.61	0.60	**6.35**
agrA (RN27)	9	30	41.0	85.0	1.83	2.94	**5.57**
VISA-vs-VSSA (Mu50 vs N315)	0	17	0	34.4	0	3.95	**5.27**
VISA-vs-VSSA (Mu50 vs Mu50-P)	0	17	0	36.7	0	3.70	**4.90**
VISA-vs-VSSA (isolate pair 2)	14	3	26.9	59.7	4.06	0.39	**4.71**
sarA (RN27)	6	23	49.9	57.7	0.97	3.22	**4.59**
agrA (UAMS-1 OD 1.0)	0	5	0	2.7	0	14.57	**4.52**
Pine-Oil Disinfectant-Reduced-Susceptibility	17	5	36.4	23.6	3.76	1.70	**4.44**
Total	275		2093.5				

Table 4.3: Enrichment found in genes, gene ontologies and expression pathways among B-class variants. Genes, gene ontologies and expression pathways exhibiting the most significant enrichments or depletions of protein-altering B-class variants separating nose microbiome and infection site bacteria. Enrichments below one represent depletions. The total number of variants and genes available for analysis differed by database. A $-\log_{10} p$-value above 5.2, 4.5 or 4.2 was considered genome-wide significant for loci, gene ontologies or expression pathways respectively (in bold).

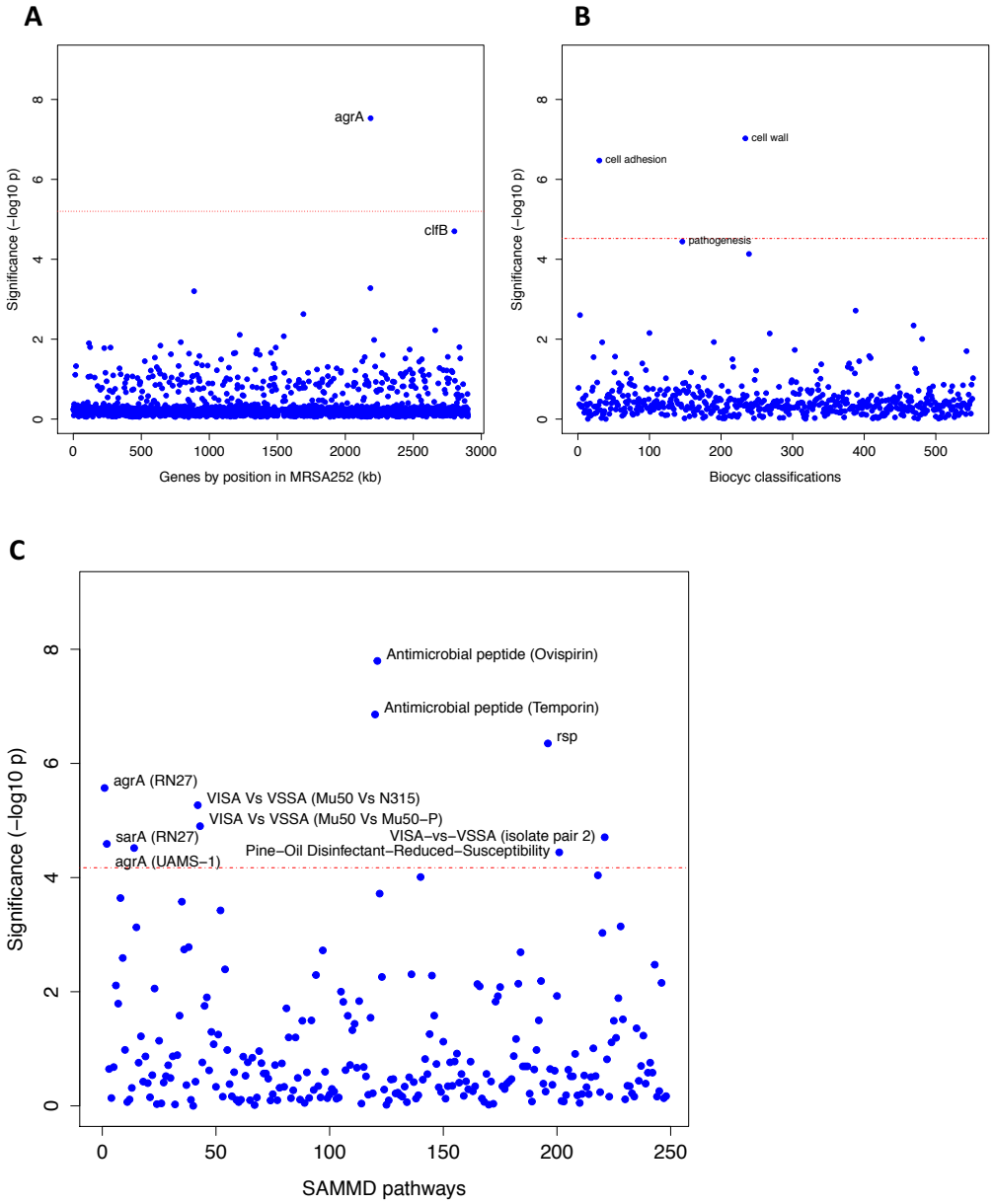


Figure 4.2: Genes, ontologies and pathways enriched for protein-altering substitutions between nose-colonizing and disease-causing bacteria within infected patients. (A) Significance of enrichment of 2650 individual genes. **(B)** Significance of enrichment of 552 gene sets defined by BioCyc gene ontologies. **(C)** Significance of enrichment of 248 gene sets defined by SAMMD expression pathways. Genes, pathways and ontologies that approach or exceed a Bonferroni-corrected significance threshold of $\alpha = 0.05$, weighted for the number of tests per category, (red lines) are named.

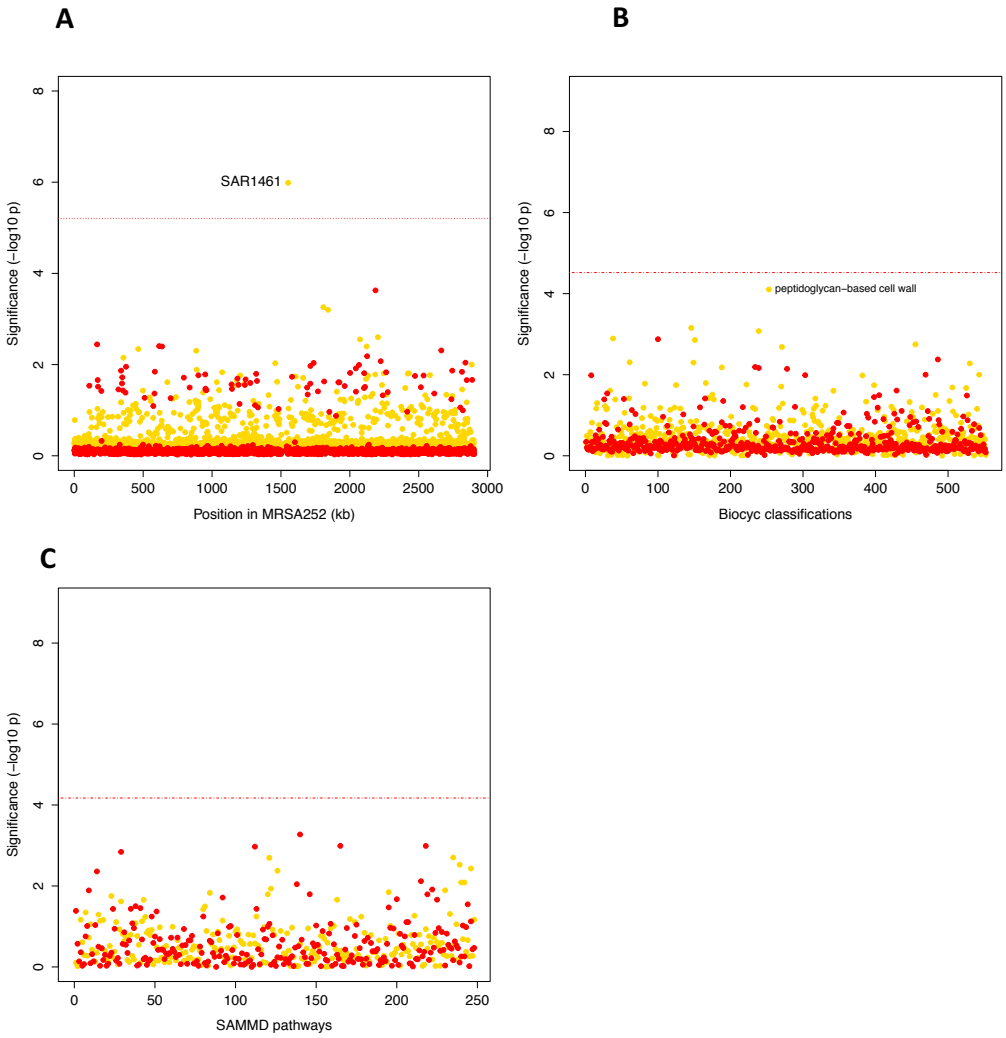


Figure 4.3: Genes, ontologies and pathways enriched for protein-altering transient variants within nose-colonizing and disease-causing bacteria. (A) Significance of enrichment of 2650 individual genes. **(B)** Significance of enrichment of 552 gene sets defined by BioCyc gene ontologies. **(C)** Significance of enrichment of 248 gene sets defined by SAMMD expression pathways. C-class variants among nose-colonizing bacteria are coloured gold, D-class variants among disease-causing bacteria are coloured red. Genes, pathways and ontologies that approach or exceed a Bonferroni-corrected significance threshold of $\alpha = 0.05$, weighted for the number of tests per category, (red lines) are named.

4.3.4 Variants separating colonising from invasive isolates are enriched for changes in gene regulators and in the cell surface proteins under their control

Only two genes showed evidence of enrichment when testing by locus, but this approach is likely underpowered to detect any but large effects. *S. aureus* contains over 2600 genes, more than twice the total number of variants identified in our study. To improve our power to detect areas of the genome systemically enriched

for variation, we performed a gene set enrichment analysis (GSEA) using genes in our reference genome (MRSA252).

We tested for an excess of variants using two independent methods of classifying gene groups. Firstly we used the ontological classifications in the Biocyc genome database[29] including all unique classifications of 2 or more loci. We tested for enrichment in genes belonging to each Biocyc ontology, compared with those that did not. We also used a database of transcription studies in *S. aureus* (SAMMD) to test for an excess of variants in the sets of genes with altered transcription in experiments comparing isogenic mutants or altered experimental conditions.[30] This has the advantage of being able to group more genes, including those not well characterised. We tested for enrichment in those genes whose expression was altered (either up-regulated or down-regulated) under the conditions studied, compared with those not differentially regulated.

Here we found the Rsp regulon was significantly enriched for variants (Table 4.3, Fig 4.2C), showing greater enrichment than that of any other transcriptional regulator ($p=10^{-6.4}$), with 3.6-fold enrichment for variants in genes down-regulated by Rsp, and 0.6-fold enrichment (ie a relative depletion) of variants in genes up-regulated by Rsp. In addition to the excess of variants found in the *agrA* locus, we found the genes up-regulated by *agrA* are enriched for variation in two strain backgrounds ($p=10^{-5.6}$, $p=10^{-4.5}$), as are those up-regulated by *sarA* ($p=10^{-4.6}$), which both directly up-regulates toxin expression and activates AgrA.[31] Significant enrichment also occurred within genes which are down-regulated by exposure to cationic antimicrobial peptides (CAMPs) involved in killing phagocytosed bacteria and mediating inflammatory responses[32] (ovispirin and temporin; $p=10^{-7.8}$ and $p=10^{-6.9}$ respectively) and after the acquisition of reduced susceptibility to antimicrobials (Vancomycin[33] and pine-oil disinfectant[34]).

Two Biocyc classifications showed enrichment for B variants (Fig 4.2B): genes located in the cell wall showed 5-fold enrichment ($p=10^{-7.0}$). These included variants in cell-wall anchored proteins (CWAs) implicated in bacterial adhesion, invasion and immune evasion,[5] including fibronectin binding protein A (*fnbA*), clumping factors A and B (*clfA, clfB*), serine rich adhesin for platelets (*sasA)* and staphylococcal protein A (*spa*) and bone sialic acid binding protein (*bbp*). Variants in four of these genes – *clfA*, *clfB*, *fnbA* and *bbp* – also contributed to 6.4-fold enrichment of variants in genes known to affect bacterial cell adhesion ($p=10^{-6.5}$).

Several genes contributed to multiple evolutionary signals. Variants in CWAs responsible for the enrichment in cell wall and cell adhesion ontologies also contributed to the enrichment in the Rsp regulon (along with urease genes *ureA* and *ureG* (a metabolic process important in biofilms[35]) as well as others).

The same enrichment was not observed in C or D variants, and no gene sets showed enrichment among these variants (Fig 4.3). These results showed that, unlike the variants arising within carriage or disease, B variants were enriched for changes to the regulatory loci and for the surface proteins whose expression they control.

4.3.5 Adaptation occurs in multiple pathways in parallel across sites during infection

We investigated whether the observed enrichment was limited to the site of infection. We classified variants according to whether the mutant allele was fund in the invasive or carriage population by identifying the allele shared with the closest related sequenced isolate in reference panel 2 ("nearest neighbour"). If the nearest neighbour allele was identical to the carriage allele, the variant was determined to be mutant in the disease population (B_D) and vice versa for carriage (B_C). 97% of variants could be thus classified. 521 (97%) B variants were typeable, with 281 of these being B_D (54%) variants.

Most significantly enriched pathways derived their signal from both B_D and B_C variants (Fig 4.4). Only the pathogenesis ontology showed differential enrichment, with B_D variants showing 3.1-fold enrichment ($p=10^{-4.6}$) and a statistically insignificant 1.7-fold enrichment in B_C-class variants ($p=0.13$). The variants driving this result include those in CWAs, as well as gamma haemolysin and regulatory loci implicated in toxin and virulence regulation (*rot*, *sarS*, and *saeR*).[36] With this exception, the processes driving evolution during infection appears common to carriage and infection sites, leading to convergent evolution across body sites. So while adaptation in pathogenesis genes appears specifically invasion-associated, other signals of adaptation in severely infected patients are driven by selection pressures that are as likely to favour mutants in nose-colonizing bacteria as in infecting bacteria. These changes may therefore be compensating for an altered within-host environment during infection.

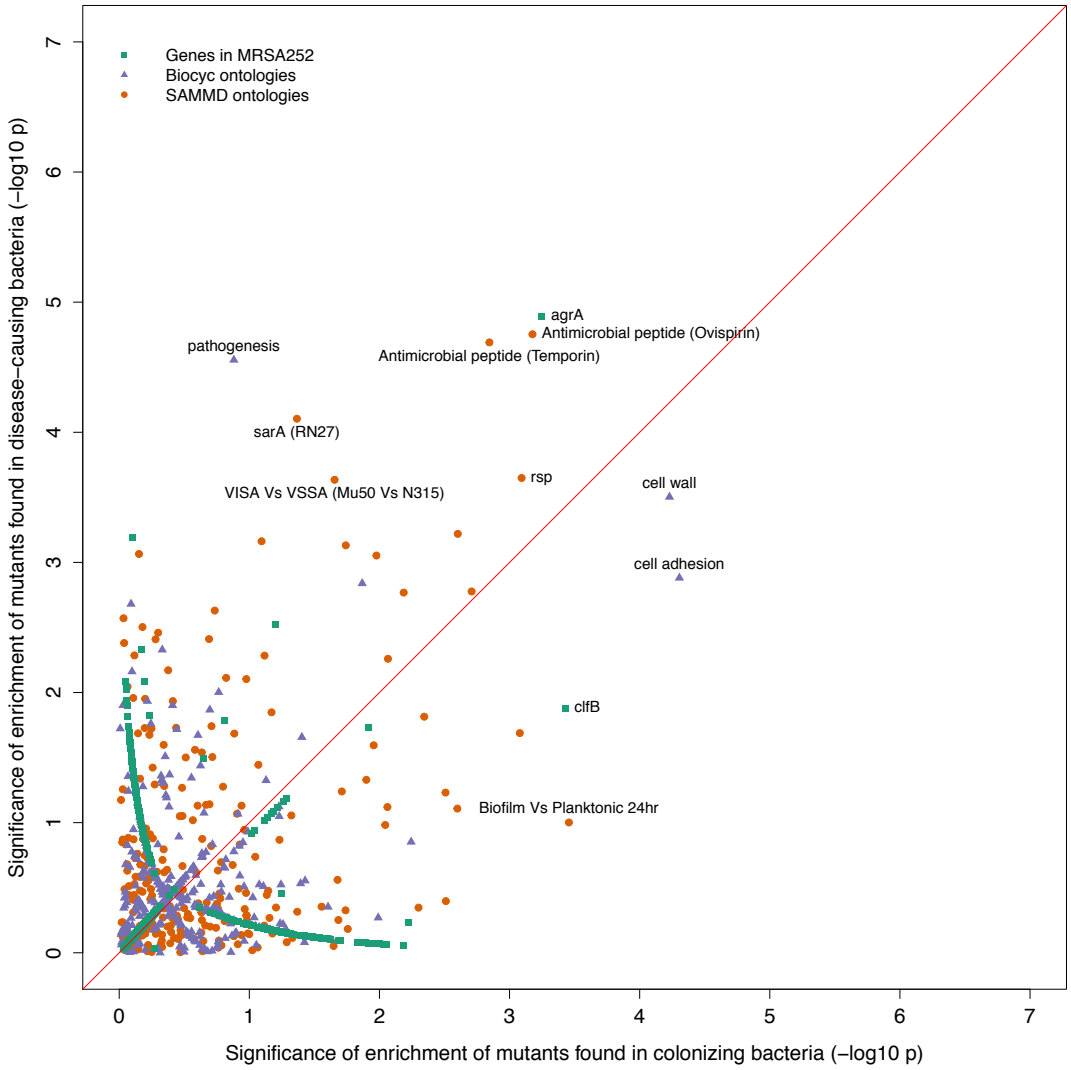


Figure 4.4: **Gene set enrichment analysis of B-class mutants occurring in the nose or the infection site**. Each point indicates the significance ($-\log_{10}$ *p*-values) of two tests for enrichment of protein-altering variants found among mutants in nose-colonizing bacteria vs disease-causing bacteria. The shape of each point represents the type of enrichment tested (squares: within 2650 genes in MRSA252, triangles: 552 Biocyc gene ontologies, circles: 248 SAMMD expression pathways). A line of 1:1 correspondence is plotted in red. A $-\log_{10}$ *p*-value above 5.2, 4.5 or 4.2 was considered genome-wide significant for loci, gene ontologies or expression pathways respectively.

4.3.6 Adaptive signatures of Rsp, Agr and adhesins are unique to within-host evolution in infected patients

Finding commonality between the signatures of adaptation in carriage and disease, we sought evidence of similar adaption in other settings.

First we examined recent variation in *S. aureus*. From a set of carriage and invasive isolates (Reference panel III), we identified 101 isolates whose most recent

common ancestor with the rest of the phylogeny was within 20 variants. By choosing sufficiently short branches we expect unique variants to be relatively recent in origin (*S. aureus* molecular clock has been estimated at 2.7 SNPs per megabase per year,[10] possibly higher in some lineages[37]), but are likely largely ancestral to within-host diversity. Analysing a total of 511 unique variants, we repeated GSEA and found near-significant enrichment only at the *agrC* locus (p=$10^{-4.7}$, Fig 4.5), and in no other genes or groupings.

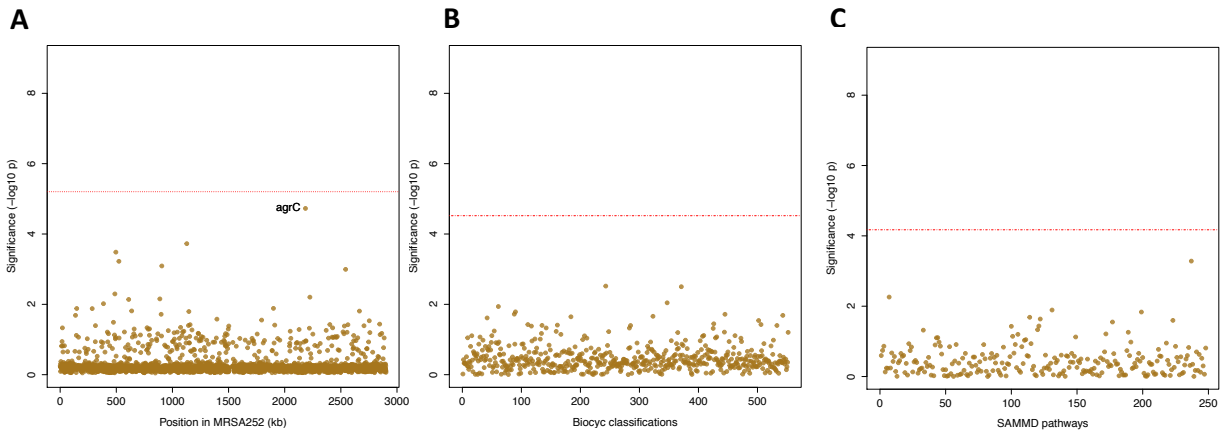


Figure 4.5: Genes, ontologies and pathways enriched for protein-altering variants among recent variants from 101 isolates from carriage and infection. (A) Significance of enrichment of 2650 individual genes. **(B)** Significance of enrichment of 552 gene sets defined by BioCyc gene ontologies. **(C)** Significance of enrichment of 248 gene sets defined by SAMMD expression pathways. Genes, pathways and ontologies that approach or exceed a Bonferroni-corrected significance threshold of $\alpha = 0.05$, weighted for the number of tests per category, (red lines) are named.

To address the modest sample size, we performed goodness-of-fit tests, focusing on the signals most significantly enriched in patients, which revealed significant depletions of substitutions during recent evolution relative to within patients in the pathogenesis ontology and AgrA regulon (p=$10^{-2.2}$, p=$10^{-3.0}$, Table 4.4).

Gene, classification or ontology	Number of variants on recent branches		Number of variants in B_D variants		Relative enrichment compared to B_D variants	
	n (total)	%	*n* (total)	%		
AgrA locus	0 (286)	0.0	3 (153)	2.0	0.0	p=0.04
AgrC locus	4 (285)	1.4	2 (154)	1.3	1.1	NS
Cell wall	6 (285)	2.1	9 (156)	5.8	0.4	p=0.05
Cell adhesion	5 (285)	1.8	6 (156)	3.8	0.5	NS
Pathogenesis	16 (285)	5.6	21 (156)	13.5	0.4	p=0.007
Rsp	13 (271)	4.8	16 (147)	10.9	0.4	p=0.03
AgrA (RN27)	13 (271)	4.8	21 (147)	14.3	0.3	p=0.001
SarA (RN27)	16 (271)	5.9	20 (147)	13.6	0.4	p=0.01
Antimicrobial peptide (Temporin)	51 (271)	18.8	35 (147)	23.8	0.8	NS
Antimicrobial peptide (Ovispirin)	49 (271)	18.1	27 (147)	18.4	1.0	NS
VISA-vs-VSSA (isolate pair 2)	10 (271)	3.7	9 (147)	6.1	0.6	NS
Pine-Oil Disinfectant-Reduced-Susceptibility	7 (271)	2.6	9 (147)	6.1	0.4	NS

Table 4.4: goodness of fit testing for ontologies showing enrichment in B-class variants compared with recent variants. For all ontologies showing enrichment in within-patient B-class variants, we identified the genes with variants contributing to the signal. We counted the number of protein-altering variants in these genes within patients, and compared to the enrichment in recent variants identified from clinical isolates in Reference Panel II. P values calculated using Fisher's exact test. *Variant totals are different for SAMMD pathways (*rsp, agrA, sarA*) and Biocyc ontologies (cell wall, cell adhesion, pathogenesis) because pathway information is available for a slightly different number of loci in each database.

We next examined the variation within-host during asymptomatic *S. aureus* carriage. These were identified from isolates from 10 subjects in a cohort study of *S. aureus* nasal carriage, sequenced at multiple time points during their carriage (Reference Panel IV[26]). The 235 variants found in stable carriage showed no variants in the *agrA* locus, the cell wall ontology or the cell adhesion ontology, and no loci or ontological classifications were enriched for variation (Figure 4.6). Again goodness-of-fit found significant depletions of protein-altering variants in carriers relative to patients in the *rsp*, *agr* and *sarA* regulons (p=$10^{-4.0}$) and the pathogenesis ontology (p=$10^{-3.2}$, Table 4.5).

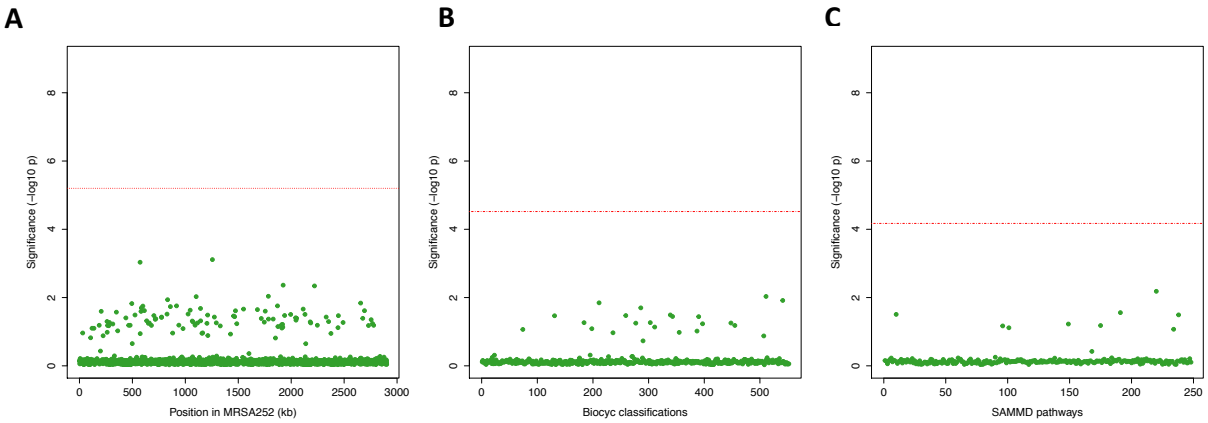


Figure 4.6: Genes, ontologies and pathways enriched for protein-altering variants among longitudinally sampled asymptomatic nasal carriers. (A) Significance of enrichment of 2650 individual genes. **(B)** Significance of enrichment of 552 gene sets defined by BioCyc gene ontologies. **(C)** Significance of enrichment of 248 gene sets defined by SAMMD expression pathways. Genes, pathways and ontologies that approach or exceed a Bonferroni-corrected significance threshold of $\alpha = 0.05$, weighted for the number of tests per category, (red lines) are named.

Gene Ontology or Expression Pathway (Loci with protein-altering B_D-class variants within patients)	Number of variants*		Relative depletion among carriers	
	Within patients	Within carriers		*p*-value
AgrA locus (SAR2126)	3/156	0/115	0.00	n.s.
Rsp transcriptional pathway (*spa*, SAR0143, *clfA*, SAR1014, SAR1745, *ureA*, *ureG*, SAR2427, *fnbA, clfB, sasA*, SAR2763)	16/147	0/109	0.00	0.0001 ***
SarA transcriptional pathway (SAR0109, *spa*, SAR0211, *pyrAA*, SAR1397, *agrC, agrA*, SAR2245, SAR2420, SAR2430, *hlgB, fnbA, arcC*, *sasA, lip*)	20/147	1/109 (agrC)	0.07	0.0001 ***
AgrA transcriptional pathway (*spa*, SAR0211, *pyrAA*, SAR1397, *sucA*, SAR1466, *hemL*, *agrC*, *agrA*, SAR2430, *hlgB*, *hlgC*, *clfB*, *arcC*, *sasA*, *lip*)	21/147	1/109 (agrC)	0.06	<0.0001 ***
Cell wall (*spa, clfA, fnbA, clfB, sasA*)	9/156	0/115	0.00	0.01 *
Cell adhesion *(clfA, fnbA, clfB*)	6/156	0/115	0.00	0.04 *
Pathogenesis (*spa*, SAR0115, SAR280, SAR0464, SAR0739, *saeR, clfA*, *ebh*, *rot*, SAR2035, SAR2448, *hlgA*, *hlgB*, *fnbA*, *clfB*, *sasA*)	21/156	2/115 (ebh)	0.13	0.0006**

Table 4.5: goodness of fit testing for variants found in severe infection compared to long-term carriers. For all ontologies showing enrichment in within-patient B_D-class variants, we identified the genes with variants contributing to the signal. We counted the number of protein-altering variants in these genes within patients, and compared to the number in long-term asymptomatic carriers. P values calculated using Fisher's exact test. *Variant totals are different for SAMMD pathways (*rsp, agrA, sarA*) and Biocyc ontologies (cell wall, cell adhesion, pathogenesis) because pathway information is available for a slightly different number of loci in each database.

Finally we examined relative rates of non-synonymous to synonymous substitution (d_N/d_S) at species-level evolution in *S. aureus* by examining completely sequenced unrelated bacteria. In the loci that contributed most to our within-patient signals – *agrA*, *agrC clfA*, *clfB*, *fnbA* and *sasA* – the d_N/d_S ratios showed no evidence for excess protein-altering change in these compared to other genes (Fig. 4.7). Accordingly, incorporating this locus-specific variability of d_N/d_S into the GSEA did not affect the results (Fig. 4.8).

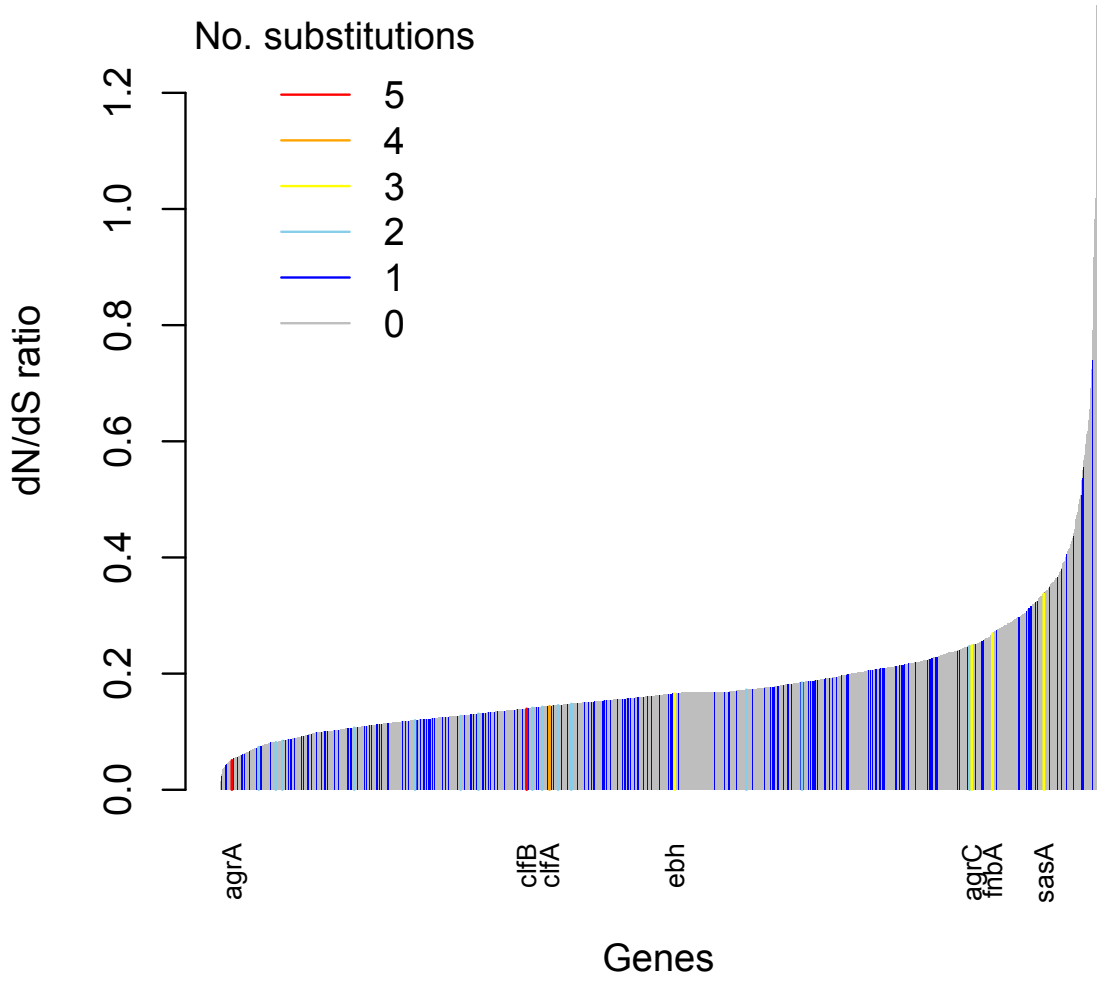


Figure. 4.7: Genes enriched for substitutions between nose-colonizing and disease-causing bacteria within patients are not the most rapidly evolving at the species level. An estimate of the d_N/d_S ratio between unrelated bacteria is shown for each gene, colour-coded by the number of protein-altering substitutions between nose-colonizing and disease-causing bacteria within patients. There was a negative Spearman rank correlation between d_N/d_S ratio and substitutions within patients (ρ=–0.04, p=0.02).

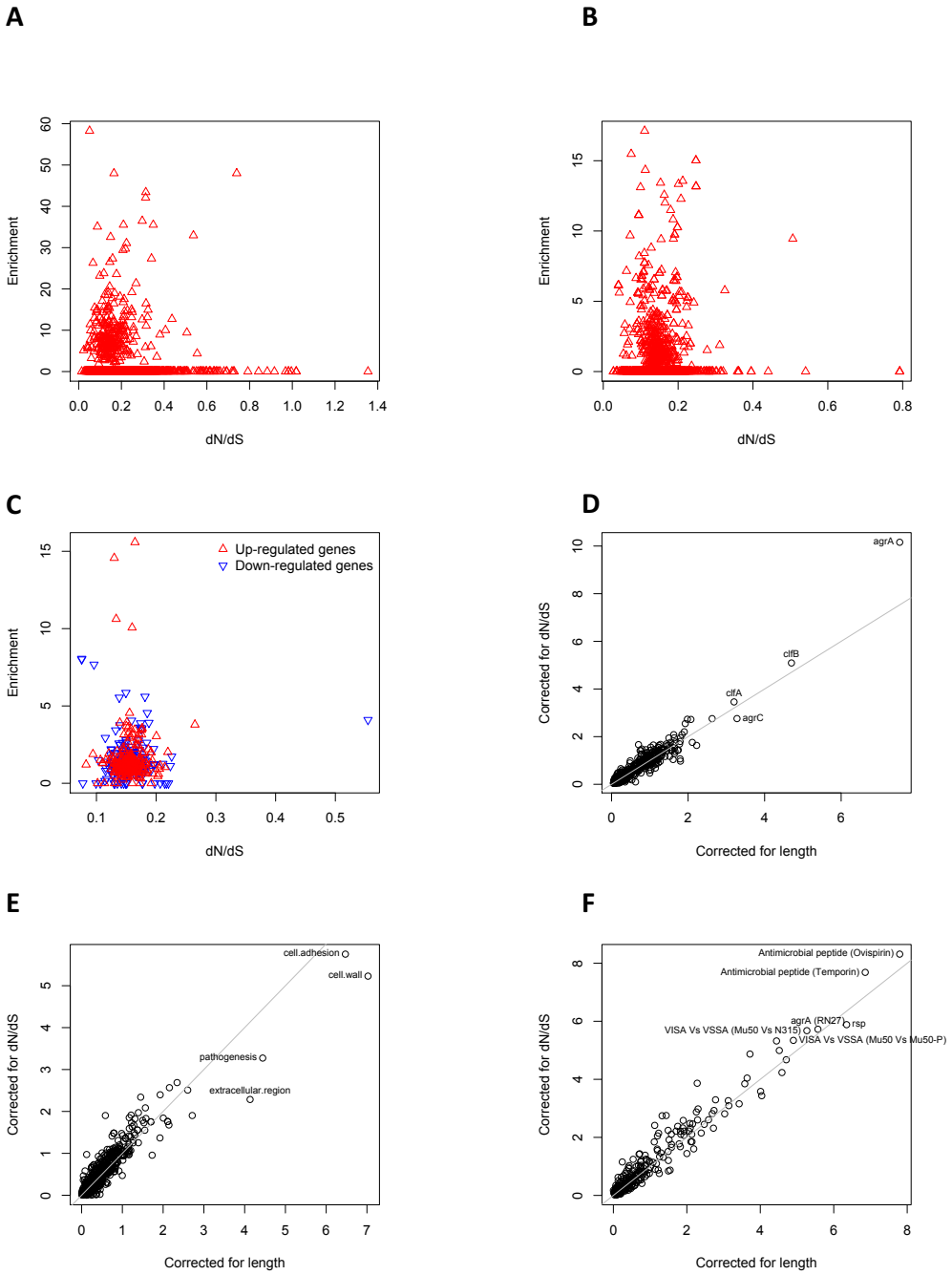


Figure 4.8: Gene set enrichment analysis is robust to species-level differences in d_N/d_S between genes. For every locus, expression pathway and gene ontology, we estimated d_N/d_S between unrelated *S. aureus*. There was no relationship between d_N/d_S and enrichment of protein-altering substitutions between nose-colonizing and disease-causing bacteria in a) loci, b) ontologies nor c) pathways (non-significant correlations, $p > 0.05$). When we incorporated variability in d_N/d_S between genes in the gene set enrichment analyses, the results were robust for d) loci, e) ontologies and f) pathways, showing only small differences in significance ($-\log_{10}$ p-value) between the analyses that correct for locus length only (horizontal axes) and those that correct for locus length and d_N/d_S (vertical axes).

Together, three lines of evidence supported our finding that adaptive signatures in the Rsp regulon, Agr and adhesins are specific to within-host evolution in infected patients, while the signature of adaptation in pathogenesis genes is specific to the site of infection in these patients.

4.4 Discussion

This study demonstrated that when serious, life-threatening infection with *S. aureus* occurs in patients who are also *S. aureus* carriers, the invasive bacteria commonly arise from commensal populations, but that detectable variation distinguishes colonising from invasive bacteria. We also discovered patterns of bacterial evolution specifically associated with invasive disease, suggesting that the changes arising in host are related to selective pressures common across sites of colonisation and invasion. Genes involved in *S. aureus* pathogenesis, including transcriptional regulators, showed enrichment specific to invasive bacteria, while both invasive and colonising bacteria showed adaptation across a range of regulators, cell wall adhesive proteins, and the genes responding to global regulators.

Among these, we showed striking enrichment affecting the targets of the recently described regulator Rsp. Having previous reported naturally occurring loss of function variants within Rsp itself,[23] here we observed the effector arm of Rsp was enriched for variants that arose in individuals with invasive *S. aureus* disease, while no similar enrichment was seen in long-term carriage.

In addition to this novel finding, we observed changes to the AgrA gene and Agr regulon. Agr mutants have been observed to arise *in vivo*, with mixed Agr function in populations from infected individuals,[18] but these relatively common mutations tend not to be transmitted.[38] This observation may be due to relative decrease in

toxin-mediated virulence in Agr mutants, with their observation in disease explained by host susceptibility,[39] but mere accident does not seem sufficient to explain this repeated finding of *agrA* mutation within hosts. Alternatively, Agr dysfunction may have enhanced fitness during infection but not for transmission,[39] for example via altered expression of Agr-dependent PSMs, which facilitate immune-mediated clearance of bacteria from the bloodstream.[40]

A growing body of evidence suggests that *S. aureus* bloodstream infection may be inversely correlated with traditional virulence assays. Studies of laboratory-created Rsp knock-outs and animal models of disease concluded that Rsp is essential to virulence,[20,35] but we found it inactivated in human bloodstream infection, and used these naturally-occurring mutants to demonstrate that Rsp mutants remain able to disseminate in a mouse model of disease despite reduced lethality.[23] With collaborators at the University of Bath we found that Rsp mutant *S. aureus* isolates causing bloodstream infection exhibited lower cytotoxicity but enhanced survival in the presence of human serum.[41]

This highlights the relevance of natural populations to studying determinants of bacterial disease. Future studies of *S. aureus* virulence can leverage these findings to focus *in vitro* studies on genes and pathways known to have clinical virulence, taking into account the trade-offs between fitness for ecological niches.

This study demonstrates the value of studying within-host evolution to elucidate the complex phenotype of virulence. Unlike relatively deterministic genotype-phenotype relationships underlying antibiotic resistance, bacterial virulence is likely to involve multiple pathways with functional redundancy and variable penetration.[13] Additionally, host factors are likely to be important. In this study we controlled for many host factors as well as bacterial strain by studying strains within-hosts, effectively matching cases and controls for host environment and

bacterial genetic background.

The majority of cases of infection studied here were found to arise from a colonising population. We might consider this as an evolutionary analogy to malignancy: severe infection such as bloodstream invasion is a temporary growth opportunity for bacteria, but an evolutionary dead-end, as it is highly likely to result in host death. Both might be considered 'accidents' in the long view, though they represent the operation of selection pressures in the short term. Thus, rather than a long evolutionary arms race towards a single goal of "virulence", *S. aureus* faces trade-offs between different periods of its life history, with some adaptations enabling colonisation of epithelial surfaces, while others favour dissemination through infection. This hypothesis has some support from the observation that signals of enrichment that were strong over the short term were absent at a species-wide level, so that variants arising within infected hosts are not transmitted.

References Chapter 4

1. Lowy FD. *Staphylococcus aureus* infections N Engl J Med 1998 Aug 20;339(8):520-32.

2. Public Health England *Staphylococcus aureus* (MRSA and MSSA) bacteraemia mandatory reports,2014/15, Public Health England 2015 [online]

3. Alonzo F,3rd, Torres VJ. The bicomponent pore-forming leucocidins of *Staphylococcus aureus* Microbiol Mol Biol Rev 2014 Jun;78(2):199-230.

4. Foster TJ. Immune evasion by staphylococci Nat Rev Microbiol 2005 Dec;3(12):948-58.

5. Foster TJ, Geoghegan JA, Ganesh VK, Höök M. Adhesion, invasion and evasion: The many functions of the surface proteins of *Staphylococcus aureus* Nature Reviews Microbiology 2013;12(1):49 - 62.

6. Thammavongsa V, Kim HK, Missiakas D, Schneewind O. Staphylococcal manipulation of host immune responses Nat Rev Microbiol 2015 Sep;13(9):529-43.

7. Casadevall A, Fang FC, Pirofski LA. Microbial virulence as an emergent property: Consequences and opportunities PLoS Pathog 2011 7(7):e1002136.

8. Brown SP, Cornforth DM, Mideo N. Evolution of virulence in opportunistic pathogens: Generalism, plasticity, and control Trends Microbiol. 2012. 20(7):336-42

9. Wertheim HF, Vos MC, Ott A, van Belkum A, Voss A, Kluytmans JA, van Keulen PH, Vandenbroucke-Grauls CM, Meester MH, Verbrugh HA. Risk and outcome of nosocomial *Staphylococcus aureus* bacteraemia in nasal carriers versus non-carriers Lancet 2004 Aug 21-27;364(9435):703-5.

10. Golubchik T, Batty EM, Miller RR, Farr H, Young BC, Larner-Svensson H, Fung R, Godwin H, Knox K, Votintseva A, et al. Within- host evolution of *Staphylococcus aureus* during asymptomatic carriage. PloS One 2013;8(5):e61319.

11. Price JR, Golubchik T, Cole K, Wilson DJ, Crook DW, Thwaites GE, Bowden R, Walker AS, Peto TE, Paul J, et al. Whole-genome sequencing shows that patient-to-patient transmission rarely accounts for acquisition of *Staphylococcus aureus* in an intensive care unit Clin Infect Dis 2014 Mar;58(5):609-18.

12. Paterson GK, Harrison EM, Gemma GRM, Welch JJ, Warland JH, Matthew TGH, Fiona JEM, Ba X, Koop G, Harris SR, et al. Capturing the cloud of diversity reveals complexity and heterogeneity of MRSA carriage, infection and transmission. Nature Communications 2015;6.

13. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. Within-host evolution of bacterial pathogens Nat Rev Microbiol. 2016. 14(3):150-62.

14. Fraser C, et al. Virulence and pathogenesis of HIV-1 infection: An evolutionary perspective Science. 2014. 343(6177):1243727.

15. Gao W, Chua K, Davies JK, Newton HJ, Seemann T, Harrison PF, Holmes NE, Rhee HW, Hong JI, Hartland EL, et al. Two novel point mutations in clinical *Staphylococcus aureus* reduce linezolid susceptibility and switch on the stringent response to promote persistent infection PLoS Pathog 2010 Jun 10;6(6):e1000944.

16. Palmer AC and Kishony R. Understanding, predicting and manipulating the genotypic evolution of antibiotic resistance Nat Rev Genet. 2013. 14(4):243-8.

17. Howden BP, et al. Evolution of multidrug resistance during *Staphylococcus aureus* infection involves mutation of the essential two component regulator WalKR PLoS Pathog. 2011. 7(11):e1002359.

18. Traber KE, Lee E, Benson S, Corrigan R, Cantera M, Shopsin B, Novick RP. Agr function in clinical *Staphylococcus aureus* isolates. Microbiology 2008;154(8):2265-74.

19. Young BC, Golubchik T, Batty EM, Fung R, Larner-Svensson H, Votintseva AA, Miller RR, Godwin H, Knox K, Everitt RG, et al. Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. Proc Natl Acad Sci U S A 2012;109(12):4550.

20. Lei MG, Cue D, Roux CM, Dunman PM, Lee CY. Rsp inhibits attachment and biofilm formation by repressing fnbA in *Staphylococcus aureus* MW2. J Bacteriol 2011;193(19):5231.

21. Tuchscherr L, Loffler B. *Staphylococcus aureus* dynamically adapts global regulators and virulence factor expression in the course from acute to chronic infection Curr Genet 2016 Feb;62(1):15-7.

22. Andersen SB, Marvig RL, Molin S, Krogh Johansen H, Griffin AS. Long-term social dynamics drive loss of function in pathogenic bacteria Proc Natl Acad Sci U S A. 2015. Aug 25;112(34):10756-61.

23. Das S, Lindemann C, Young BC, Muller J, Österreich B, Ternette N, Winkler AC, Paprotka K, Reinhardt R, Förstner KU, Allen E, et al. Natural mutations in a *Staphylococcus aureus* virulence regulator attenuate cytotoxicity but permit bacteremia and abscess formation Proc Natl Acad Sci U S A. 2016. 113(22):E3101-10.

24. Everitt RG, et al. 2014. Mobile elements drive recombination hotspots in the core genome of *Staphylococcus aureus* Nat Commun 5:3956.

25. Gordon NC, et al. 2014. Prediction of *Staphylococcus aureus* antimicrobial resistance by whole-genome sequencing J Clin Microbiol 52(4):1182-91.

26. Gordon NC, Pichon B, Golubchik T, Wilson DJ, Paul J, Blanc DS, Cole K, Collins J, Cortes N, Cubbon M, et al. Whole-Genome Sequencing Reveals the Contribution of Long-Term Carriers in *Staphylococcus aureus* Outbreak Investigation. J Clin Microbiol. 2017.Jul;55(7):2188-2197.

27. Margolis E and Levin BR. Within-host evolution for the invasiveness of commensal bacteria: An experimental study of bacteremias resulting from *Haemophilus influenzae* nasal carriage J Infect Dis. 2007. 196(7):1068-1075.

28. Leski TA and Tomasz A. Role of penicillin-binding protein 2 (PBP2) in the antibiotic susceptibility and cell wall cross-linking of *Staphylococcus aureus*: Evidence for the cooperative functioning of PBP2, PBP4, and PBP2A J Bacteriol. 2005. 187(5):1815-24.

29. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases Nucleic Acids Res. 2016. 44(D1):D471-80.

30. Nagarajan V and Elasri M. SAMMD: *Staphylococcus aureus* microarray meta- database. BMC Genomics. 2007. 8(1):351.

31. Cheung AL, Nishina KA, Trotonda MP, Tamber S. The SarA protein family of *Staphylococcus aureus*. Int J Biochem Cell Biol. 2008;40(3):355-61.

32. Pietiäinen M, François P, Hyyryläinen HL, Tangomo M, Sass V, Sahl HG, Schrenzel J, Kontinen VP. Transcriptome analysis of the responses of *Staphylococcus aureus* to antimicrobial peptides and characterization of the roles of vraDE and vraSR in antimicrobial resistance. BMC Genomics. 2009 Sep 14;10:429.

33. Howden BP Smith DJ, Mansell A, Johnson PD, Ward PB, Stinear TP, Davies JK. Different bacterial gene expression patterns and attenuated host immune responses are associated with the evolution of low-level vancomycin resistance during persistent methicillin-resistant *Staphylococcus aureus* bacteraemia BMC Microbiol. 2008 8:39,2180-8-39.

34. Lamichhane-Khadka R, Riordan JT, Delgado A, Muthaiyan A, Reynolds TD, Wilkinson BJ, Gustafson JE. Genetic changes that correlate with the pine-oil disinfectant-reduced susceptibility mechanism of *Staphylococcus aureus*. J Appl Microbiol. 2008 Dec;105(6):1973-81.

35. Li T, He L, Song Y, Villaruz AE, Joo HS, Liu Q, Zhu Y, Wang Y, Qin J, Otto M, et al. AraC-type regulator rsp adapts *Staphylococcus aureus* gene expression to acute infection Infect Immun 2015 Dec 28.

36. Ibarra JA, Pérez-Rueda E, Carroll RK, Shaw LN. Global analysis of transcriptional regulators in *Staphylococcus aureus*. BMC Genomics 2013;14:126.

37. Long SW, Beres SB, Olsen RJ, Musser JM. Absence of patient-to-patient intrahospital transmission of Staphylococcus aureus as determined by whole-genome sequencing. MBio. 2014 Oct 7;5(5):e01692-14.

38. Shopsin B, Eaton C, Wasserman GA, Mathema B, Adhikari RP, Agolory S, Altman DR, Holzman RS, Kreiswirth BN, Novick RP. Mutations in agr do not persist in natural populations of methicillin- resistant *Staphylococcus aureus*. J Infect Dis 2010;202(10):1593.

39. Smyth DS, Kafer JM, Wasserman GA, Velickovic L, Mathema B, Holzman RS, Knipe TA, Becker K, Von Eiff C, Peters G, et al. Nasal carriage as a source of agr- defective *Staphylococcus aureus* bacteremia. J Infect Dis 2012;206(8):1168.

40. Cheung GY, Kretschmer D, Duong AC, Yeh AJ, Ho TV, Chen Y, Joo HS, Kreiswirth BN, Peschel A, Otto M. Production of an attenuated phenol-soluble modulin variant unique to the MRSA clonal complex 30 increases severity of bloodstream infection PLoS Pathog 2014 Aug 21;10(8):e1004298.

41. Laabei M, Uhlemann AC, Lowy FD, Austin ED, Yokoyama M, Ouadi K, Feil E, Thorpe HA, Williams B, Perkins M, et al. Evolutionary trade-offs underlie the multi-faceted virulence of *Staphylococcus aureus* PLoS Biol 2015 Sep 2;13(9):e1002229.

Chapter 5

5 *Staphylococcus aureus* bacteraemia is not strongly determined by the bacterial genome at a population level

5.1 Introduction

Staphylococcus aureus is one of the leading causes of bloodstream infection worldwide.[1,2,3,4] 12,000 cases occur each year in the UK.[5] The mortality of *S. aureus* bacteraemia (SAB) has not declined over recent years, and mortality at 30 days remains generally over 20% even with appropriate therapy.[6,7] *S. aureus* bacteraemia can be both community-associated (CA) or healthcare-associated (HA), including health-care associated disease with an onset in the community.[8]

There is some evidence for bacterial genotype contributing to the likelihood of bacteraemia. The presence of putative virulence factors has been associated with invasive *S. aureus* disease: secreted enterotoxins and haemolysins, surface proteins which mediate tissue attachment, invasion and immune evasion, and variation in the Agr regulatory system have all shown association with an increased odds ratio (OR) of being found in invasive *S. aureus* disease in case control studies.[9,10,11] However, the evidence for these associations is inconsistent, and conflicting results are reported between studies (see Chapter 2, Table 2.1). Likewise, strains belonging to major Clonal Complexes (CCs) have shown conflicting evidence of association with invasive disease.[12,13]

It has been argued that the different selection pressures at play in HA and CA disease are responsible for the differences in bacteria commonly observed causing HA and CA SAB. The healthcare setting features greater selection pressure in the form of antibiotic use, but in this setting patients may also have a lower barrier to infection, as skin breaks and vascular access devices provide physical opportunities for bacterial entry to the bloodstream, and reduced immune function increases susceptibility to infection.[14] Painter and colleagues reason that the finding of loss of function mutations in the Accessory Gene Regulator (Agr) in HA bloodstream infections is driven by the selective pressure of antibiotic use.[14] MRSA strains express altered Penicillin Binding Proteins (PBPs), either constitutively or in response to β-lactam exposure. These PBPs directly reduce Agr expression and therefore toxin production and neutrophil evasion, which these authors argue is a cost or "price" of antibiotic resistance. They contend this disadvantage can more often be overcome in hospitalised patients, and this explains the finding of Agr mutants in this setting.

Other evidence suggests that altered bacterial toxicity is not just a feature of lowered barrier to bloodstream entry in a compromised host, but rather that reduced bacterial toxicity is associated with enhanced bacterial survival in the bloodstream. In two patients studied as part of a comprehensive study of within-host variation in *S. aureus* (see Chapter 4), we identified loss of function mutations in Repressor of Surface Proteins (Rsp) arising in host and distinguishing carriage from bloodstream isolates. Strains carrying these mutations were associated with attenuated mortality in a mouse model of disease when compared with isogenic strains carrying the wild type allele, but nevertheless retained their ability to disseminate, and caused the same number of abscesses as Rs*p*-producing strains.[15] A collection of USA300 isolates showed significantly reduced toxicity to T cells, compared with USA300 isolates found in nasal carriage or soft tissue infection.[16]

These observations support the hypothesis that subtle genetic variation could increase the likelihood of SAB with certain isolates of *S. aureus*, and make this an attractive subject for GWAS. Bacteraemia has advantages as a phenotype for GWAS because the finding of *S. aureus* on blood culture is an objectively defined phenotype, which could plausibly be bacterially determined, since it differs from the organism's usual ecological niche in the anterior nares.

To date no GWAS have reported bacterial determinants of *S. aureus* bloodstream infection, though there is growing interest in this area. A recent study investigated associations between bacterial genetic variants and predicted severity of disease based on toxin production and severity of disease in a mouse model within a single lineage (ST239).[17] Here we present the first GWAS of *S. aureus* bacteraemia across the entire bacterial population, studying 2001 isolates from nasal carriage and bacteraemia and applying validated tools for identifying bacterial lineages, variants and genes in GWAS.

5.2 Materials and methods

Bacteria were processed and GWAS analysis methods were followed as described in Chapter 3. Special considerations of sampling and collection of epidemiological data that apply to this study are detailed below.

5.2.1 Sampling frame for carriage controls

S. aureus isolated from nasal carriage in individuals without *S. aureus* bacteraemia were identified from two studies of *S. aureus* carriage in Oxfordshire UK. The first was a study of carriage in adults in the community,[18] and the second was a study of *S. aureus* carriage and transmission in 3 wards of the John Radcliffe Hospital, Oxford.

In the first study, individuals ≥16 years of age were invited to participate in the

study when attending one of 5 General Practices in the catchment of OUH NHS Foundation Trust, and during orientation to Oxford University. 1123 individuals enrolled, of which 360 carried *S. aureus* at recruitment. All individuals carrying *S. aureus* at recruitment and 211 swab-negative individuals were invited to supply nasal swabs at 2 monthly intervals between July 2009 and April 2013.[18] Carried *S. aureus* strains were typed using *spa*-typing. A study of 100 isolates from this study demonstrated that the global diversity of *S. aureus* was well represented in this population.[19]

We identified individuals with community-associated *S. aureus* carriage from this cohort. Individuals with *S. aureus* clinical infections over the course of the study were excluded. Where co-habiting individuals carried the same *spa*-type, only 1 individual was considered for inclusion to avoid over-sampling of strains being transmitted between individuals in the study population. Finally, individuals with an inpatient hospital stay in the year prior to first carriage observation were excluded, leaving 396 individuals with community-associated *S. aureus* carriage. (Figure 5.1)

In the second study, individuals admitted to three study wards had nasal swabs for *S. aureus* carriage performed on admission and at fortnightly intervals until discharge between September 2009 and August 2011. 1146 individuals were found to have *S. aureus* carriage on any swab during the study (enrolled from Intensive Treatment Unit (ITU) (n=729), Trauma unit (n=352) or one of two Geratology wards (n=65)). Carried *S. aureus* strains were typed using *spa*-typing.

Individuals with *S. aureus* isolated from a clinical sample in the previous 12 months were excluded. Individuals who had first *S. aureus* carriage isolation on day of admission or the following day, and met the definition for community carriage used above (no overnight stay in the preceding 12 months) were included as community-

associated controls. Individuals who had *S. aureus* first isolated more than 48 hours after admission or who had been admitted for ≥3 nights in the preceding 28 days were included as healthcare-associated controls. (Figure 5.2)

For each included individual, we selected the last instance of the most frequently carried *spa*-type, which represented the isolate with the greatest opportunity to cause disease without having done so.

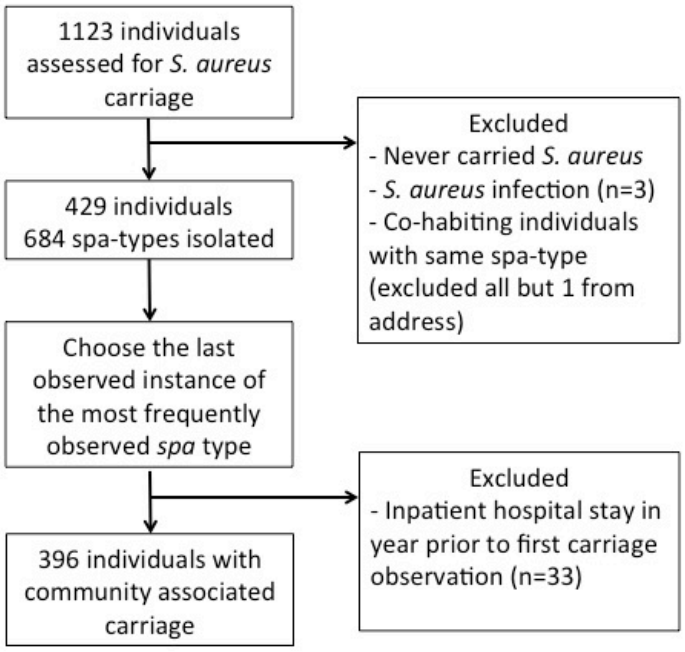


Figure 5.1: Flow chart of control selection from study of *S. aureus* carriage in community

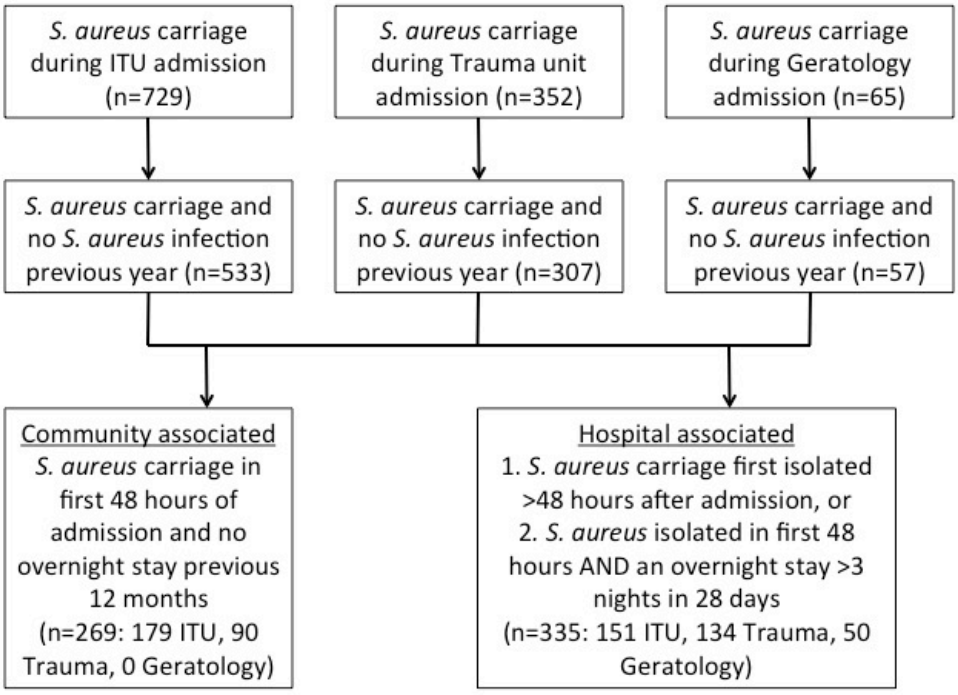


Figure 5.2: Flow chart of control selection from study of *S. aureus* carriage in hospital inpatients

5.2.2 Sampling frame for bacteraemia cases

For a given sample size, our power to detect effects is greatest with a matched number of cases and controls. Having identified nearly 1000 population controls, we aimed for a sample size of approximately 1000 cases. Because bacteraemia is less frequent than carriage, *S. aureus* bacteraemia cases were drawn over a longer period of time, and comprised three broad collections of sequenced isolates.

Cases were eligible for inclusion if they met the following criteria:

1. *S. aureus* isolated from bloodstream
2. Blood culture isolate was available for sequencing
3. A minimum dataset was present in available clinical database (See 5.2.3)
4. Patient >13 years of age (as this was minimum age for inclusion in clinical

databases)

5. Isolation of *S. aureus* was not deemed to be a contaminant on clinical assessment

Sequences were drawn firstly from a historic collection of SAB isolates from OUH NHS trust (Oxford UK) and BSUH NHS trust (Brighton, UK) identified between 2008 and 2012, which were sequenced in 2011 and 2012 for studies of *S. aureus* resistance.[20] From 612 strains isolated between 2008 and 2012, we identified 417 isolates that represented the first sequenced SAB episode in unique individuals.

Next was a collection was 119 cases occurring in Oxford 2012-2013, along with a collection of 160 SAB from Plymouth collected over 2008-2012. These were sequenced in 2014, along with all the identified controls. Finally, to increase our power, subsequent cases occurring in Oxford and Brighton were sequenced in 2015, adding 199 cases from Oxford (2013-14) and 122 from Brighton (2012-14).

5.2.3 Epidemiological data

Epidemiological data about individuals with SAB accrued from several sources. Firstly, the Infections in Oxfordshire Research Database (IORD) which links information about patient attendances with results from pathology services in an anonymised manner for research purposes.[21]

Secondly, information on episodes of SAB was collected as part of on-going service evaluation studies, as part of multi-centre collaborations with the UK Clinical Infection Research Group (UKCIRG) and the International *Staphylococcus aureus* Collaboration (ISAC).

Finally, information was collected on infections that may be healthcare acquired as part of infection control surveillance and reporting, which has been mandatory for MRSA bacteraemia since 2007, and MSSA bacteraemia since 2011.

To avoid sampling bias driven by individuals with recurrent disease, only the first episode of SAB in an individual was included, and the earliest available isolate was selected for inclusion. Within UKCIRG and ISAC databases, patient identifiers were included that allowed identification of individuals. The IORD database generates a unique, anonymous "cluster id" for each individual, and by using this identifier we were able to remove repeat episodes in an individual.

For each episode of SAB we recorded patient gender and age at the time of infection. We recorded the date of admission, the number of days between admission and first blood culture from which *S. aureus* was cultured, and the number of days since the most recent discharge from hospital. These data constituted the minimum dataset for inclusion. If available, the clinically determined focus of infection and patient mortality data 90 days after infection were also recorded.

Cases were deemed healthcare-associated (HA) if the first blood culture positive for *S. aureus* was collected on the third day or later of a hospital admission, or if the patient had an inpatient admission in the previous 90 days. Cases were deemed community-associated if the first blood culture positive for *S. aureus* was collected on the first or second calendar day of admission, and there was no inpatient admission in the previous 90 days.

Our definition of hospital exposure was more stringent for cases than controls. For cases we considered admissions in the previous 90 days, a time frame used widely in the literature[22], and for mandatory reporting. However controls were only deemed community-associated if there was no admission in the previous 12 months. This stricter definition was chosen in order to sample controls that most truly reflected strains transmitted and carried in the community without causing disease.

Epidemiological data for controls, including admission and any clinical *S. aureus* infection was accessed from IORD for individuals sampled on ITU, Geratology and Trauma wards. For other community carriage controls the Oxfordshire carriage study database included patient and GP reports of infection, and through identifying patient information collected for study purposes, we were able to record the isolation of *S. aureus* from clinical samples as well as any admissions to OUH NHS trust using IORD.

5.2.3 Ethics statements

Ethical approval for sequencing *S. aureus* isolates from routine clinical samples and linkage to patient data without individual patient consent in Oxford and Brighton in the U.K. was obtained from Berkshire Ethics Committee (10/H0505/83) and the U.K. National Information Governance Board [8-05(e)/2010].

Data about *S. aureus* bacteraemia in Oxfordshire, Brighton and Plymouth were collected for evaluations of clinical service provision. The UK National Research Ethics Service reviewed this data collection protocol, and deemed it was an evaluation of service, and therefore did not require ethics committee review.

5.2.4 Multiple testing

When testing PCs, SNPs and kmers for association with phenotype, a Bonferroni correction was applied to a nominal false discovery rate of 5%, as detailed in Methods 3.7. In a set of 2001 isolates we identified 2001 PCs, 907,487 phylopatterns of kmers and 112,292 phylopatterns of SNPs (after imputation of missing calls). This gave adjusted significance thresholds of $10^{-4.6}$, $10^{-7.3}$ and $10^{-6.4}$ respectively. A kmer GWAS on subset of these (1610 isolates) identified 763,051 phylopatterns of kmers, giving an adjusted significance threshold of $10^{-7.2}$.

Two sub-analyses were conducted. A GWAS to identify bacterial lineages and genetic variants associated with healthcare acquisition of carriage on 984 isolates

from CA and HA carriage identified 984 PCs, 668,232 kmer phylopatterns and 76,048 phylopatterns of SNPs. This gave adjusted significance thresholds of $10^{-4.3}$, $10^{-7.2}$ and $10^{-6.2}$ respectively. A GWAS to identify bacterial lineages and genetic variants associated with bacteraemia comparing 1285 isolates from CA carriage and CA SAB identified 1285 PCs, 743,840 kmer phylopatterns and 89,462 phylopatterns of SNPs. This gave adjusted significance thresholds of $10^{-4.4}$, $10^{-7.2}$ and $10^{-6.3}$ respectively.

5.3 Results

5.3.1 Global diversity of lineages found in *S. aureus* bacteraemia and carriage

We sequenced 984 isolates from unique individuals with carriage and 1017 isolates from unique individuals with *S. aureus* bacteraemia. Table 5.1 summarises the characteristics of patients in with SAB compared to those with carriage.

	Bacteraemia	**Carriage**	
Number of sequences	1017	984	
Community-associated (CA)	631	654	
Healthcare-associated (HA)	386 (38.0%)	330 (33.5%)	
Age (median (IQR))	68 (53-79)	59 (38-76)	$P < 10^{-5}$ (Mann-Whitney U test)
Male sex (n (%))	659 (68.4%)	511 (51.9%)	$P < 10^{-5}$ (χ^2 test)
MRSA (n (%))	138 (13.6%)	54 (5.5%)	$P < 10^{-5}$ (χ^2 test)

Table 5.1: Cases and controls included in study. HA includes hospital-onset disease and community-onset-HCA disease.

As has been reported in previous studies[2] the median age of individuals with bacteraemia was significantly higher than the median age of carriers. Also consistent with known risk factors for *S. aureus* bacteraemia,[2] we found a

significantly higher proportion of cases occurred in men. The ratio of men to women in SAB was 2.1, while in carriage the ratio was 1.1. Cases were much more likely than controls to be methicillin resistant, and this higher rate was also found when CA and HA isolates were compared (Table 5.2). A lesser degree of hospital exposure in CA controls may partly explain the excess of MRSA isolates causing CA disease compared with carriage. However we also found MRSA 1.5 times more frequently in HA cases than controls (p=0.04), despite identical inclusion criteria regarding hospital exposure in these groups (Table 5.2).

	Cases	Controls	
CA (n)	75 (11.9%)	19 (2.9%)	$p < 10^{-5}$ (χ^2 test)
HA (n)	63 (15.3%)	35 (10.6%)	p=0.035 (χ^2 test)
Total (n)	138 (13.6%)	54 (5.5%)	

Table 5.2: MRSA in CA and HA *S. aureus* bacteraemia and carriage. Rates of methicillin resistance in cases (bloodstream infection) are compared with controls (asymptomatic nasal carriage) found in community acquired (CA) and heathcare-acquired (HA) *Staphylococcus aureus*.

There is evidence from multiple studies that the different rates observed in this study population are in keeping with wider prevalence. Studies of MRSA in Oxfordshire found the total prevalence of MRSA in community carriage to be below 1%,[18] and demonstrated that CA MRSA bacteraemia can occur individuals with no hospital exposure in the preceding 12 months.[23] These low rates of MRSA in CA carriage are consistent with several observational studies reporting prevalence of MRSA nasal carriage of 1.9% in community[24] and 2.0% on admission to hospital,[25] as well as a UK national audit which reported the prevalence of MRSA carriage on admission to hospital was 1.4% in 2013.[26]

Mortality at 7 days, 30 days and 90 days was recorded for 927/1017 (91.1%) cases (Table 5.3). Mortality data were missing for 6/672 Oxford cases (0.9%), 6/160

(3.7%) of Plymouth cases, and 70/187 (37.4%) of Brighton cases. Mortality rates differed between centres at 7 days, but these differences were not statistically significant (p=0.25, χ^2 test with 2 degrees of freedom). There was no significant difference between centres in mortality at 30 days and 90 days (p=0.6 and p=0.9). Overall mortality was 13.3% at 7 days, 26.4% at 30 days, and 34.5% at 90 days. Not all this mortality is likely to be attributed directly to SAB, and this high mortality probably reflects both an elderly population with co-morbidities as well as a significant mortality burden of SAB.

Centre	All cause mortality by 7 days	All cause mortality by 30 days	All cause mortality by 90 days
Oxford	94/666 (14.1%)	180/666 (27.0%)	228/666 (34.2%)
Brighton	15/117 (12.8%)	27/117 (23.1%)	38/117 (32.5%)
Plymouth	14/144 (9.7%)	38/144 (26.4%)	54/144 (37.5%)
All	123/927 (13.3%)	245/927 (26.4%)	320/927 (34.5%)

Table 5.3: Mortality at 7, 30 and 90 days in SAB cases.

We collected data on the focus of infection, when it could be identified, where available (Table 5.4). There was 9.2% missing data for this item. Soft tissue infections were the most commonly established focus, but cases where no focus was identifiable by clinical examination were the second largest group.

Focus	*n*	%	UKCIRG (%)
Central venous catheter	101	10.9	18.2
Peripheral venous catheter	47	5.1	5.8
Endocarditis	42	4.7	5.8
Soft tissue	244	26.4	19.5
Respiratory	52	5.6	3.7
Urinary	35	3.8	(NA)
Thrombus	11	1.2	(NA)
Vascular implant	17	1.8	(NA)
Other	134	14.5	13.6
Not established	235	25.5	18.8
(Missing data)	94	9.2%	

Table 5.4: Focus of infection for 1017 cases. Proportions (of available data) are compared with those reported in a prospective cohort of cases in the UK (UKCIRG).[6]

A phylogeny of 2001 cases and controls demonstrated that the global diversity of *S. aureus* was represented in this study (Figure 5.3). Cases and controls did not obviously cluster in the tree, with the exception of a cluster of HA cases of ST-36, a hospital-associated clone, which previously caused an outbreak in Brighton.[27] Two lineages dominated MRSA isolates – ST-22 and ST-36 – consistent with other reports of the epidemiology of MRSA in Oxfordshire and elsewhere in the UK.[28,29] No isolates belonged to the third most common UK HA-MRSA clone ST-239,[29] and only sporadic isolates of MRSA CC-8 were observed.

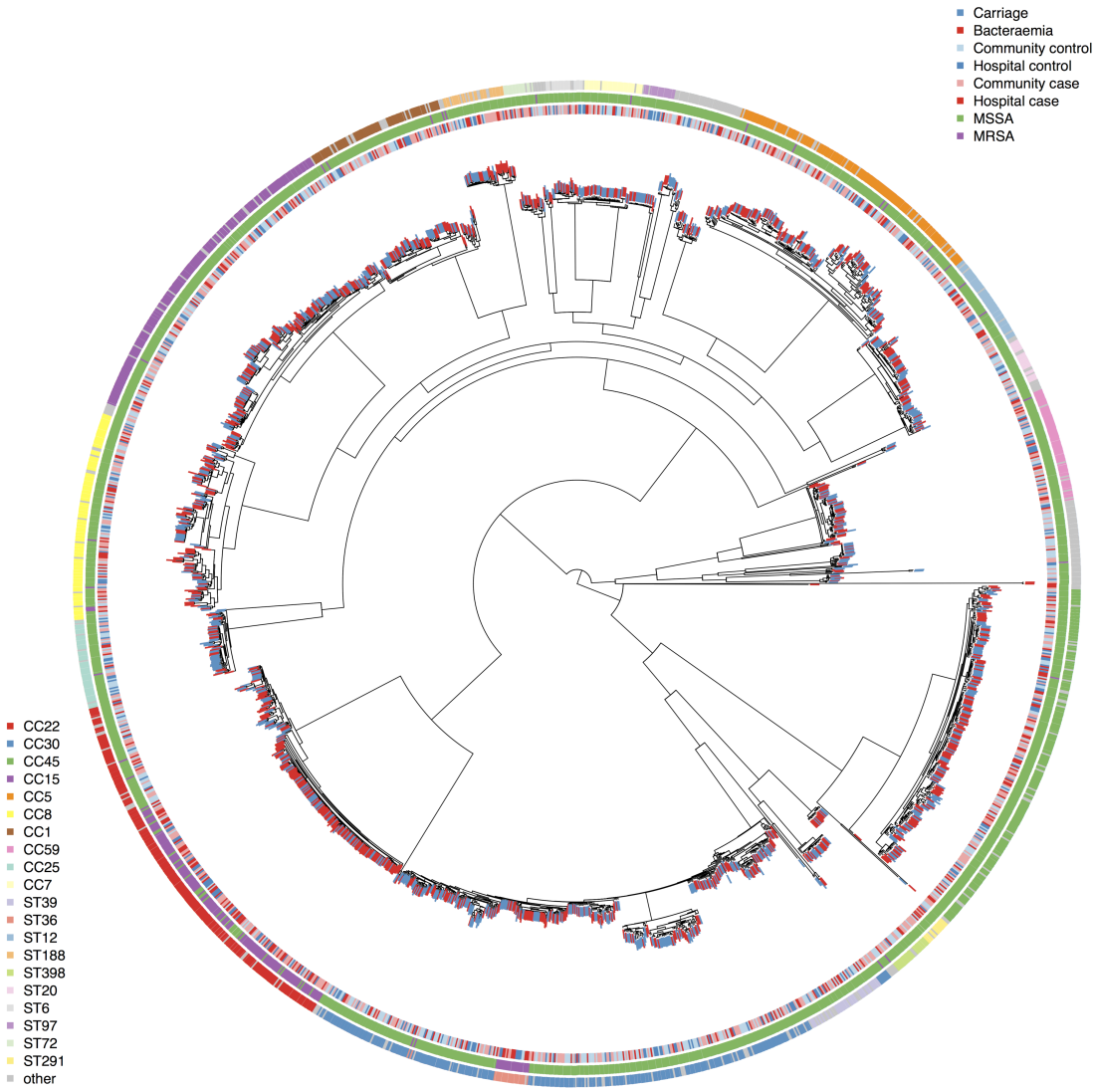


Figure 5.3: Maximum likelihood phylogeny of 2001 isolates from SAB and carriage. Colours at the tips of the tree represent isolate source (blue carriage, red bacteraemia). The inner ring indicates whether each isolate was community or hospital associated. The middle ring indicates whether each isolate was MRSA (purple) or MSSA (green). The outer ring indicates *in silico* MLST.

Formal testing for lineage effects with *bugwas*[30] confirmed the absence of strong lineage effects. PC3, which identifies the MRSA clade within the CC-22 lineage, was most strongly associated with case-control status (p=0.02), but this was not statistically significant after adjusting for multiple testing (Figure 5.4). When MRSA was included as a fixed effect in GEMMA, the association between PC3 and case control status remained the lineage most strongly associated with bacteraemia, but the significance of this association was reduced further (p=0.08). These findings

help resolve the conflicting findings from case-control studies which have found evidence supporting[32,33] and refuting[31] differing invasiveness between lineages. This comprehensive survey of SAB and carriage strongly supports the hypothesis that lineages of *S. aureus* do not differ substantially in their propensity to cause bacteraemia.

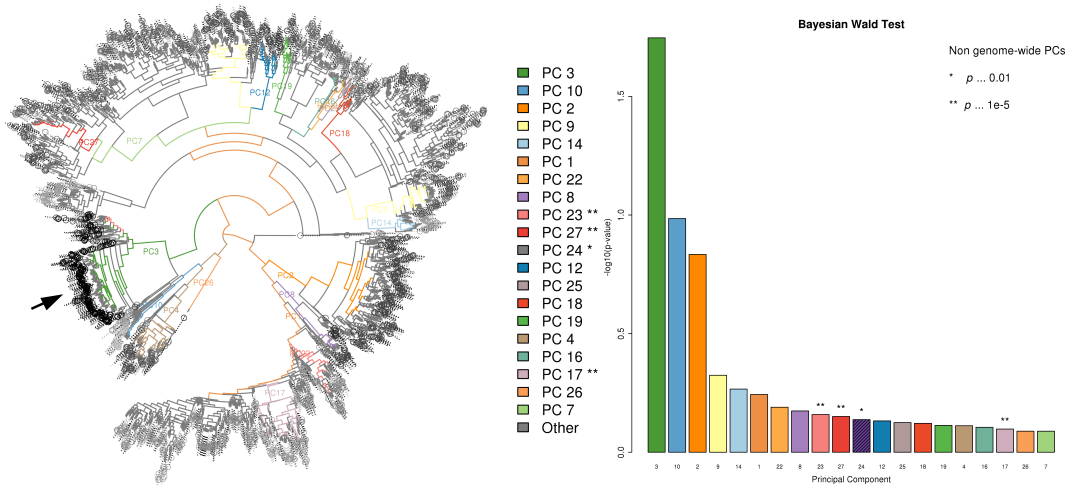


Figure 5.4: Association of Principle Components with case-control status. (A) The branches corresponding to 20 most significantly associated lineages are coloured on a maximum likelihood phylogeny of the study isolates. Lines at the tips of the tree represent the phenotype of each isolate (grey, carriage; black bacteraemia), while circles represent the phenotype predicted by the LMM (grey, carriage; black bacteraemia). **(B)** Significance ($-\log_{10}$ *p*-values) of the 20 most significantly associated PCs. A Bonferroni corrected threshold for significance is $10^{-4.6}$.

5.3.2 No individual SNPs are found in significant association with bacteraemia

The overall sample heritability was predicted to be low and not significantly different from 0 (2.1%, 95% CI 0-5.3%). Testing all identified SNPs for association with case/control status in 2001 isolates did not identify any single SNPs associated with bacteraemia at a genome wide level (Figure 5.5). As previously, we used GEMMA[34] to control for population structure in testing SNP associations.

The SNP with the strongest association was an A to T mutation at position 1497290, in the MRSA252 reference genome, encoding a phenylalanine to tyrosine substitution at codon position 99 in dihydrofolate reductase (*dfrB*). This mutation confers trimethoprim resistance.[20] It was relatively rare, being found in 41 cases

and 5 controls (OR 8.2, $p=10^{-5.6}$). This variant was correlated with other antibiotic resistance: 36/46 (78%) of isolates with this variant were also MRSA. This variant was found most commonly in isolates from CC-22 (63%) and ST-36 (22%).

The next top ranked SNP was an A to C mutation at MRSA252 position 1633189, encoding a valine to leucine substitution in SAR1555, a regulatory gene in the integrated prophage φSa2, showing an OR for disease of 2.3 ($p=10^{-5.5}$). The presence of this gene was one of 14 associated with worse outcome in *S. aureus* bacteraemia in a microarray study.[35] SNP 1633189 was commoner than 1497290, being found in 227 isolates. 163 of these (72%) were MRSA, and 164 were in CC-22, with ST-36 and other CC-30 isolates harbouring the majority of the remainder (9.3% and 3.5% each).

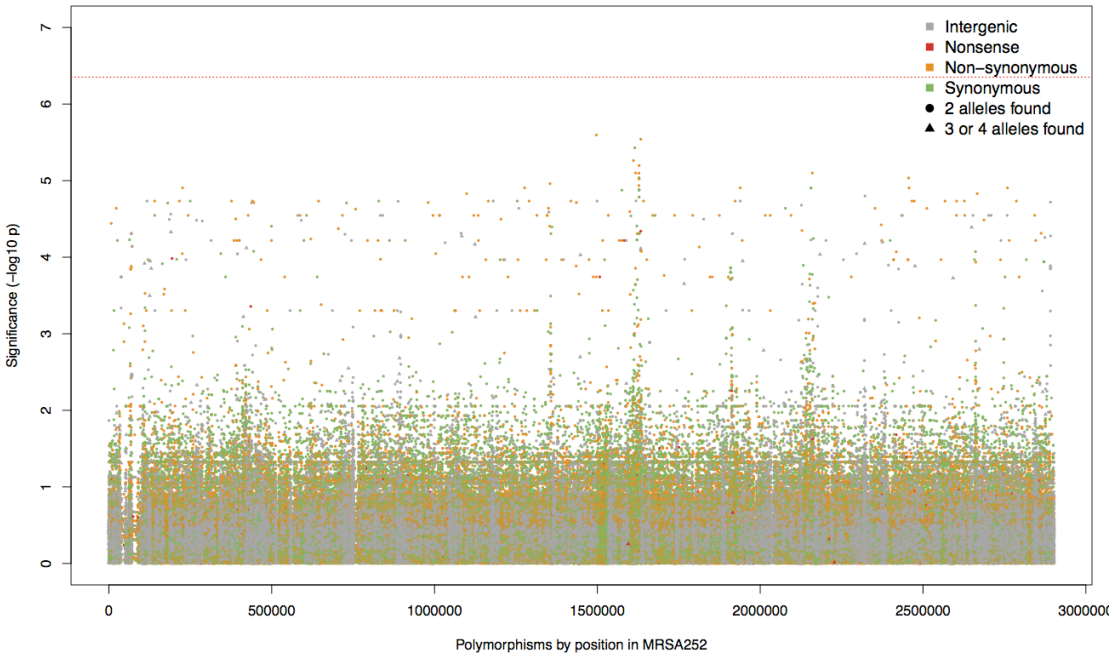
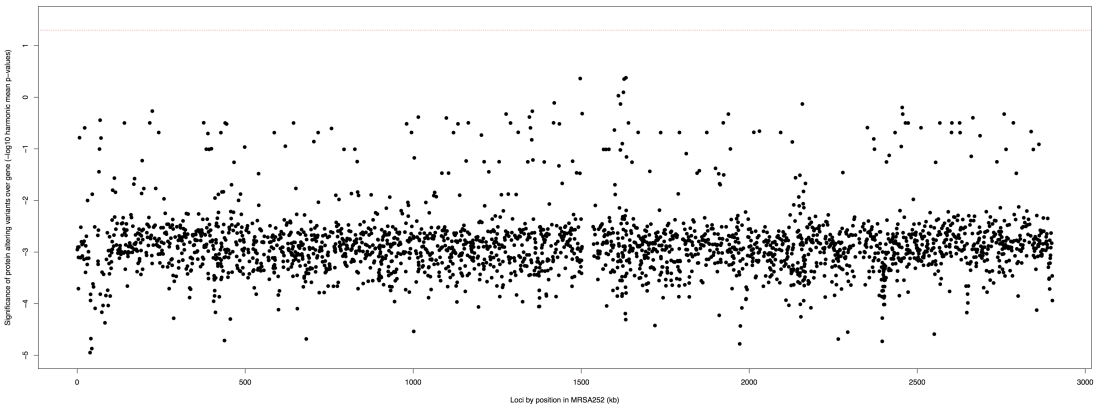


Figure 5.5: Manhattan plot of SNP associations with phenotype. The significance of each SNP ($-\log_{10} P$ value) is plotted according to location on the 2.9MB MRSA252 reference genome, with control for population structure. SNPs are coloured according to their predicted effect on protein, and whether the polymorphism was biallelic or tri/tetrallelic.

While no individual SNPs reached genome wide significance, we can apply a model averaging method to test whether a single gene or group of genes are enriched for variation. The GSEA methods used to test within-host variants in Chapter 4 depend upon the assumption that there is no population structure in the variant data, which is valid because the paired samples provide an effective control for strain background. An alternative approach in the presence of population structure is to test the harmonic mean *p*-value (HMP) across a gene or group of genes (Methods 3.6).[36]

The HMP was calculated for protein-altering variants across the genome, within each locus in the reference genome (MRSA252), as well as the groups of genes defined by ontological classifications in the Biocyc[37] database, and groups of genes under shared regulatory control, found in the Sammd[38] database. The HMP for each locus or group of genes represents the pooled evidence for the hypothesis that variation within that gene or region is associated with bacteraemia (Figure 5.6).



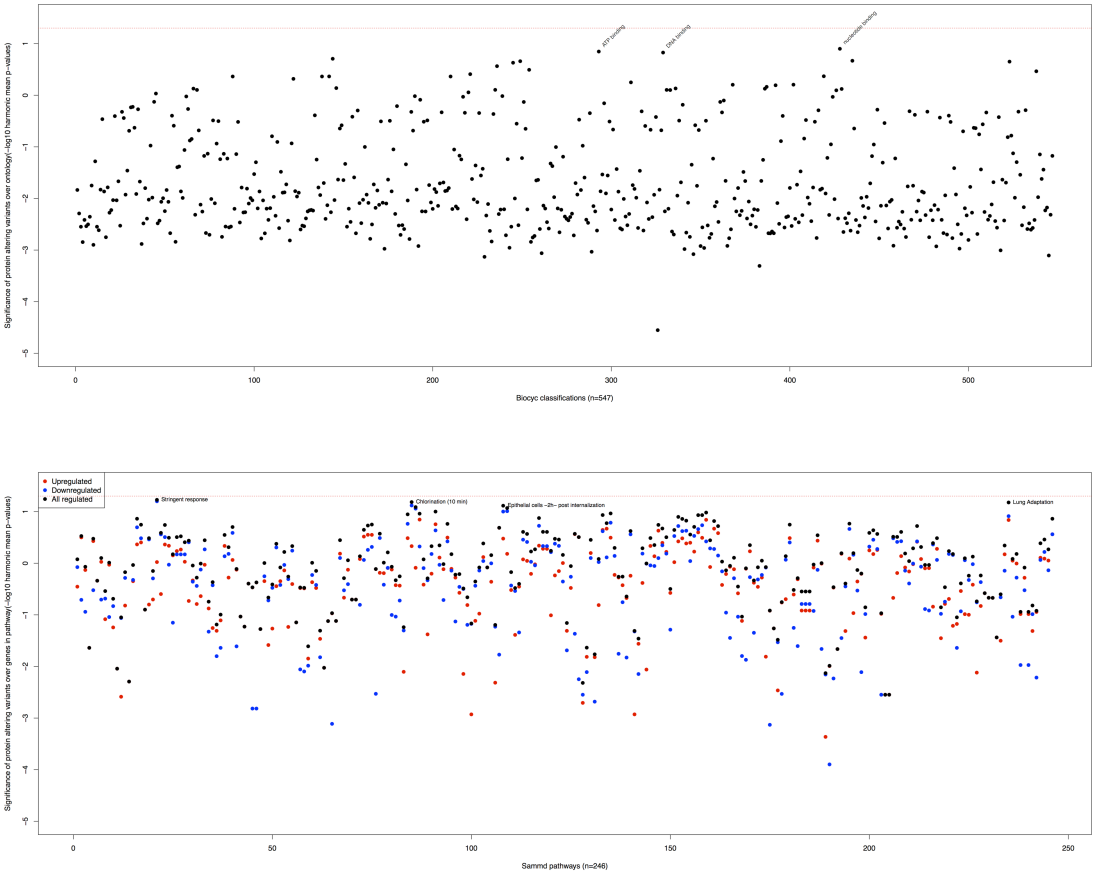


Figure 5.6: Manhattan plot of harmonic mean *p*-values for protein altering variants at each locus in MRSA252, 552 ontological classifications and 246 transcriptional pathways. Each point represents the $-\log_{10}$ adjusted *p* value, which is corrected for multiple testing through the proportion of all variants that are found in the gene or genes under investigation. The significance threshold, controlling for a family-wise error rate of 0.05, is plotted in red. **(A)** and **(B)** black points represent a gene **(A)** or set of genes with shared classification **(B)**. **(C)** Black points represent all genes with altered regulation in the study conditions, red points represent groups of u*p*-regulated genes and blue points represent groups of down-regulated genes. Groups and pathways approaching statistical significance are labelled.

The pooled evidence for an association between protein altering variants and bacteraemia was marginally statistically significant at HMP=0.043. No individual loci showed evidence for significant variants within the gene (Fig 5.6A). The Biocyc classifications with the strongest evidence for association with phenotype were: nucleotide binding (HMP=0.18), DNA-binding (HMP=0.18) and ATP-binding (HMP=0.20).

A number of transcriptional pathways showed somewhat stronger evidence when the evidence from variants was pooled across genes. The evidence for association

between bacteraemia and variants in genes with altered transcriptional activity after induction of the stringent response[39] gave HMP=0.13. There was support for association with variants in genes with altered transcription in different biological niches: genes with altered transcription after growth in mouse lung[40] and inside human epithelial cells[41] both gave HMP=0.14. These are also the transcriptional pathways with the largest regulons: the stringent response affects transcription of 800 genes across the genome; in lung adaptation 818 genes are differentially regulated; in epithelial cell internalisation 918 genes are differentially regulated. So while the localisation of the signal of genetic association with bacteraemia to genes with altered activity during adaption to tissue niches or in response to nutritional limitation is intriguing, it does not greatly help us narrow the location of that signal.

5.3.3 Kmer GWAS of SAB shows strong batch effects arising from refinements in sequencing technology

Testing for associations between the presence or absence 31bp kmers and case/control status in these 2001 isolates revealed unexpectedly significant associations between 5604 kmers that did not map to the *S. aureus* reference genome (Figure 5.7). BLAST alignment of these kmers to the National Center for Biotechnology Information (NCBI) microbial database revealed they aligned to a 5.4 kb *Escherichia* phage (phiX174, Genbank accession NC_001422.1). These kmers were found exclusively in cases sequenced prior to 2014.

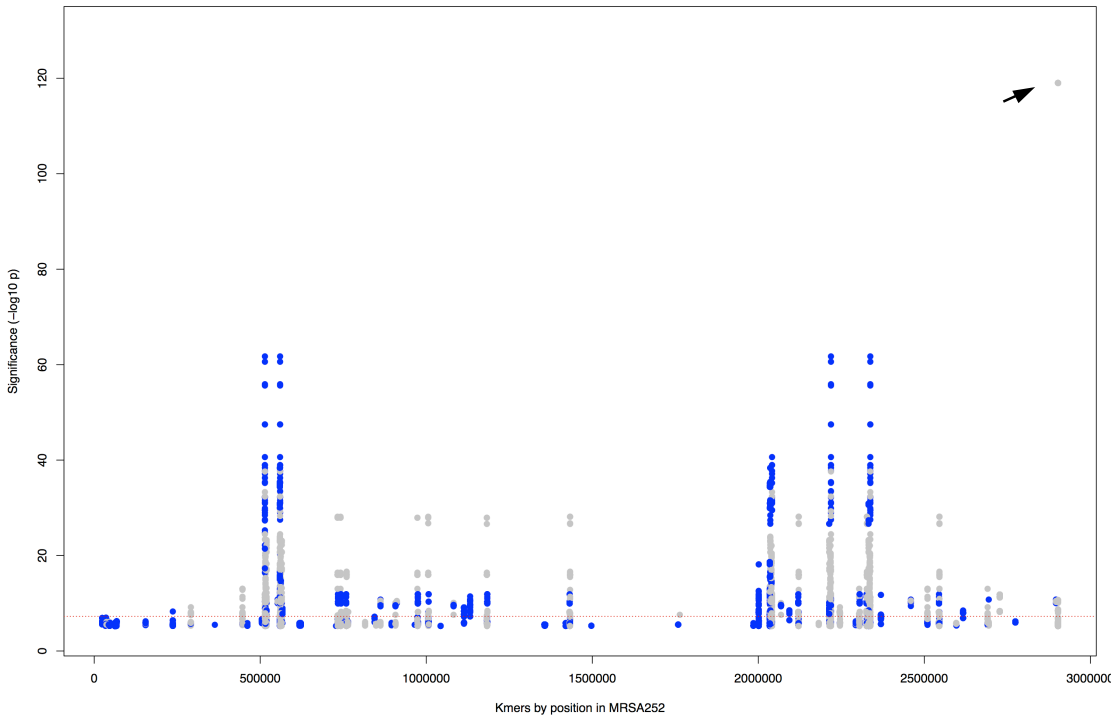


Figure 5.7: Significance (-log$_{10}$ *p*-value) of 10,000 kmers most significantly associated with *S. aureus* bacteraemia. The x-axis represents the location to which the kmer mapped in a reference genome (MRSA252). Kmers which did not map to the reference are placed at the right hand side. Kmers with a mapping quality score <10 are coloured grey, ≥10 coloured blue. A threshold of significance (adjusted for multiple testing) is plotted in red). An arrow highlights the highly significant kmers which did not map to MRSA252.

This observation was highly suggestive of confounding by batch-specific effects. We included 417 cases from relatively recently sequenced collections in this study. Although the costs of bacterial genome sequencing have fallen dramatically over the past two decades, they remain significant (approximately £40GBP per genome), and this represented an attempt to increase the value of previous work. However the inclusion of these isolates (all of which were cases) sequenced with slightly different technology risked batch effects.

One potential source of batch effect related to a differential processing by the WTCHG in 2011-2, during which time all bacterial libraries were 'spiked' by the inclusion of DNA from phiX174. This was done for internal quality control assessment of library preparation and sequencing. However reads related to this phage were not removed prior to returning sequencing data for pipeline

processing. This phage has no homology to MRSA252, showing no significant alignment of more than 30bp with the 323 complete *S. aureus* genomes in NCBI microbial BLAST database. Phage reads were therefore unmapped and did not affect SNP calling, but the reads were present in *de novo* assembly from which we generated kmers, creating an artefactual association. To remove this artefact, all kmers in the study were mapped to the reference sequence for phiX174 using Bowtie.[42] Kmers which mapped to phiX174 with a mapping quality over 20 were removed from the set of kmers before association testing in GEMMA. In total 5604 kmers in 12 patterns were removed.

Phage spiking was one systematic difference in sequencing protocols, however other differences existed, which may have given rise to additional confounding effects. These differences are summarised in Table 5.5.

	Batch 1 (2011)	Batch 2 (2011-2)	Batch 3 (2014)	Batch 4 (2015)
Total (*n*)	248	169	1263	321
Cases (*n*)	248	169	279	321
Controls (*n*)	0	0	984	0
Centre	Oxford	Oxford (*n*=104), Brighton (*n*=69)	Oxford (*n*= 984 controls, *n*= 119 cases), Plymouth (*n*=160 cases)	Oxford (*n*=199), Brighton (*n*=122)
Platform	Illumina HiSeq 2000	Illumina HiSeq 2000, changing to HiSeq 2500	Illumina HiSeq 2500	Illumina HiSeq 2500
Read length	100bp (all)	100bp (*n*=143) 151bp (*n*=26)	151bp (all)	151bp (all)
Other	PhiX174 added to sequencing library	PhiX174 added to sequencing library		

Table 5.5: Summary of sequencing batches, and illustrative list of possible sources of confounding effects. This list is not exhaustive: Illumina sequencing kits and WTCHG protocols are regularly refined**.**

After filtering kmers from phiX174, testing associations between kmer presence and *S. aureus* bacteraemia still showed kmers with strongly significant associations in multiple peaks across the genome (Figure 5.8). However the locations of some of these kmers were suggestive of further batch effects. The vast majority of significant kmers had multiple equally good mappings across the reference, and these mapped to regions with multiple copies in the reference: 16S rRNA (of which MRSA252 has 5 copies) and transposon insertion sequences ISSaur3 and IS5 (MRSA252 has 5 copies of both). All of these kmers were more frequently present in controls than cases.

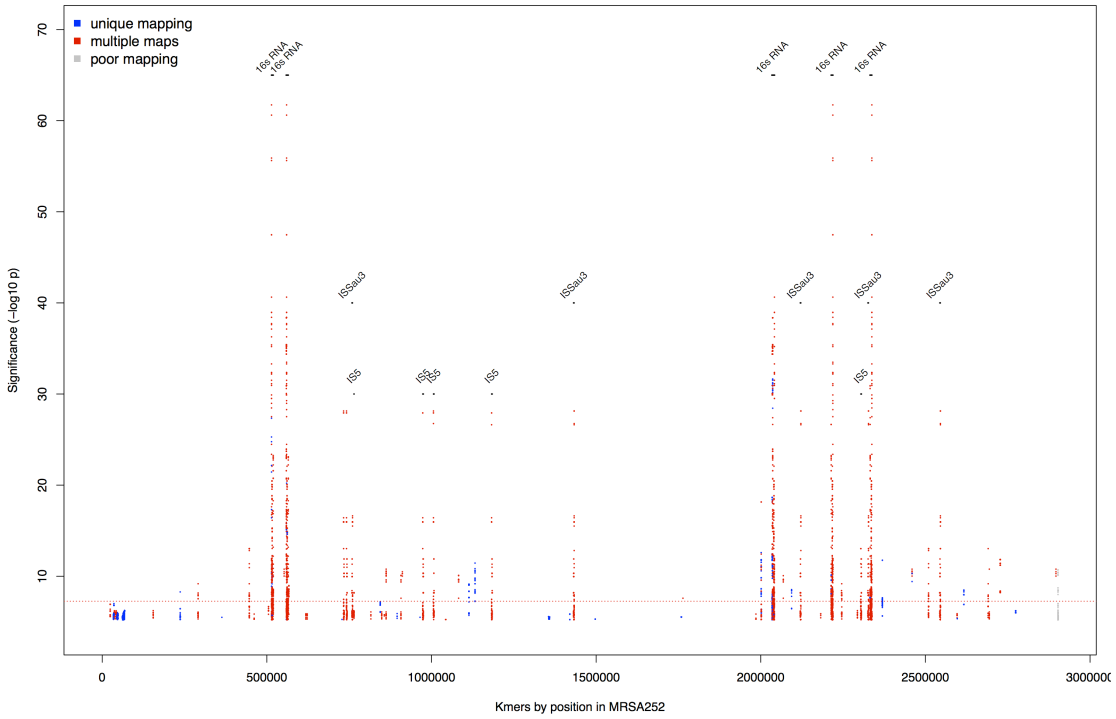


Figure 5.8: Significance (-log$_{10}$ *p*-value) of 10,000 kmers most significantly associated with *S. aureus* bacteraemia. The x-axis represents location on mapping to a 2.9MB reference genome (MRSA252). Kmers which did not map to the reference are placed at the right hand side. Kmers with a mapping quality score <10 are coloured grey, ≥10 coloured blue or red. Kmers coloured red had multiple equally good mappings to the reference, and kmers coloured blue mapped only once. A threshold of significance (adjusted for multiple testing) is plotted in red. Repeat regions with significant kmers are marked in black and annotated.

We hypothesised these kmers may be an artefact of varying read length in the sample set. Increasing read length will improve the assembly of repeat regions, and if an increase from 100bp to 151bp reads were sufficient to improve assembly of

these RNA and transposon sequences, this would manifest as increased odds of finding them in controls (all of which were sequenced with 151bp reads) compared to cases (626/1017, 61.5% sequenced with 151bp reads).

We tested this hypothesis by studying the 10,000 kmers most significantly significant associated with disease, and comparing the likelihood of finding each kmer in disease (association with phenotype from GEMMA), with the likelihood of finding the same kmer in a batch (tested using a χ^2 test with 3 degrees of freedom), or in a sequence with 151bp reads (tested using a χ^2 test with 1 degree of freedom) (Figure 5.9). This revealed that almost all the highest ranked kmers were more strongly associated with batch than phenotype, and there was a strong association with read length that was much more strongly significant than the association with phenotype. For a small number of kmers the association with phenotype was stronger than that seen with batch effect or read length, though this association was modest.

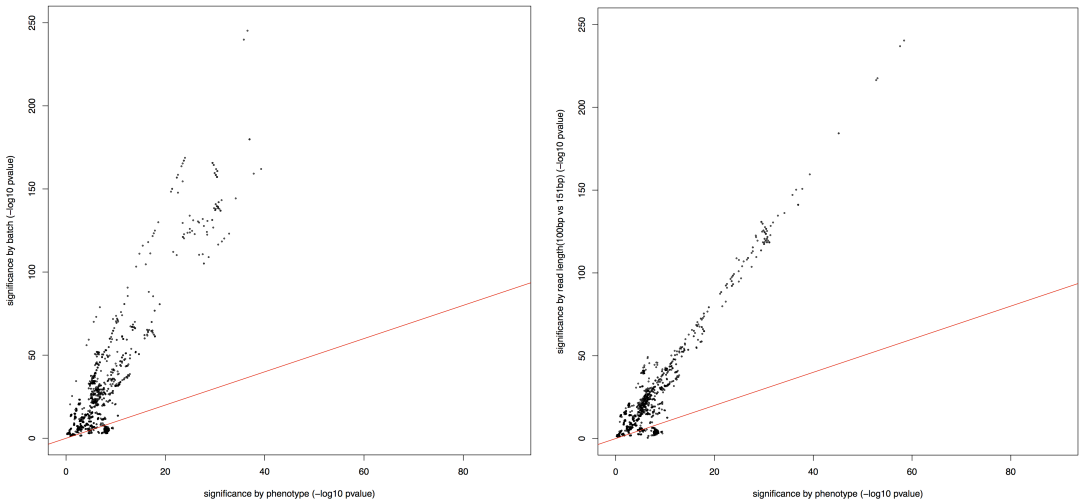


Figure 5.9: Association between kmers mapping to repeat regions with phenotype, compared to association with (left) sequencing batch and (right) read length. Negative $\log_{10}$ *p*-values of each test are plotted. A line of 1:1 correspondence is plotted in red.

In some respects, the assembly characteristics of the genomes were remarkably similar irrespective of read length. For example there was no difference in the length of the total assembled genome or the total number of kmers identified based on sequencing read length (Table 5.6). However the assembled genomes sequenced with longer (151bp) reads were assembled with a longer kmer length on average (Figure 5.10). Kmer length was selected on a per sample basis by Velvet to optimise assembly quality. This systematic difference likely conferred greater ability to span repetitive sequences with longer reads.

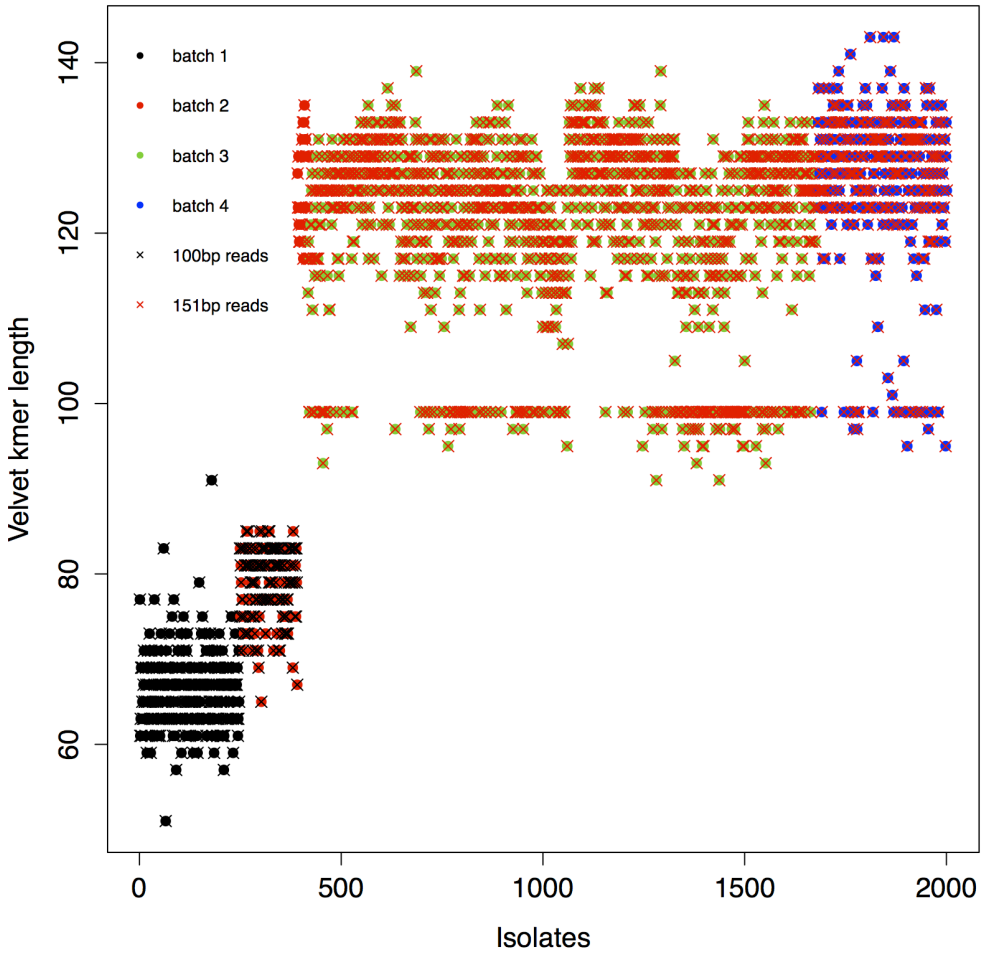


Figure 5.10: The kmer-length used for Velvet assembly plotted according to sequencing batch (solid circles) and read length (crosses).

	Bases in contigs of *de novo* assembly Med (IQR)	Bases in contigs >1000bp Med (IQR)	Number of unique 31bp kmers on dsk Med (IQR)
100bp	2,762,000 (2,727,000-2,811,000)	2,750,000 (2,715,000-2,795,000)	2,757,000 (2,722,000-2,805,000
151bp	2,763,000 (2,729,000-2,810,000	2,751,000 (2,718,000-2,796,000)	2,761,000 (2,726,000-2,807,000

Table 5.6: *De novo* assembly lengths and number of kmers in isolates with differing read lengths

5.3.4 Antibiotic resistance-conferring mutations and genes are associated with SAB in a kmer GWAS

In view of the strong confounding effects of read length on kmer presence, we excluded sequences with a read length <150 bp from the kmer GWAS, studying only isolates with 151bp read lengths, which reduced the sample size to 626 cases and 984 controls. In this reduced set, we found evidence of 23 kmers, occurring in 2 phylopatterns, significantly associated with *S. aureus* bacteraemia (Figure 5.11). These kmers, when present, mapped to a 52bp region in *dfrB*, and had 11.2-11.6-fold increased odds of being found in a disease causing, rather than carried, *S. aureus* (p=$10^{-9.0}$-$10^{-9.3}$). These kmers centred on MRSA252 position 1497290, where three known variants are capable of conferring trimethoprim resistance.[20] Thus these kmers captured the most significant variant found in the SNP GWAS. (5.3.2).

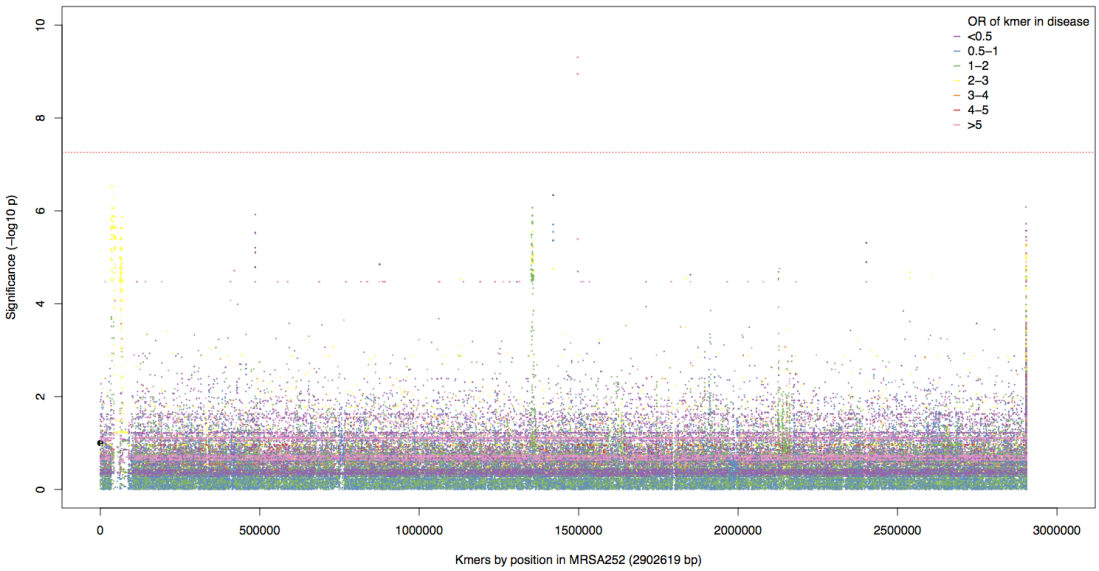


Figure 5.11: Association between kmers and bacteraemia. Each point represents a single kmer. The 10,000 kmers with the most significant association, and 200,000 randomly selected kmers are plotted by their position on mapping to MRSA252 and the -$\log_{10}$ *p*-value of their association with the phenotype, controlling for population structure. Points are coloured by their OR of being found in bacteraemia. A Bonferroni corrected threshold for significance is plotted in red.

Only the associations between these 23 kmers and SAB were genome-wide significant, but peaks of relatively strongly associated kmers were visible at 50kb, 500kb and 1.3Mb (Figure 5.11). As with SNP data, it is possible to pool the evidence from multiple kmers using the harmonic mean of the *p*-values to assess the evidence for association across a gene or locus. This was done by mapping kmers to a reference genome (MRSA252), and identifying the likely locus to which they mapped. 59% of kmers mapped to MRSA252. For kmers mapped to MRSA252, the pooled evidence for association between kmers and SAB was strong, HMP=$10^{-3.7}$.

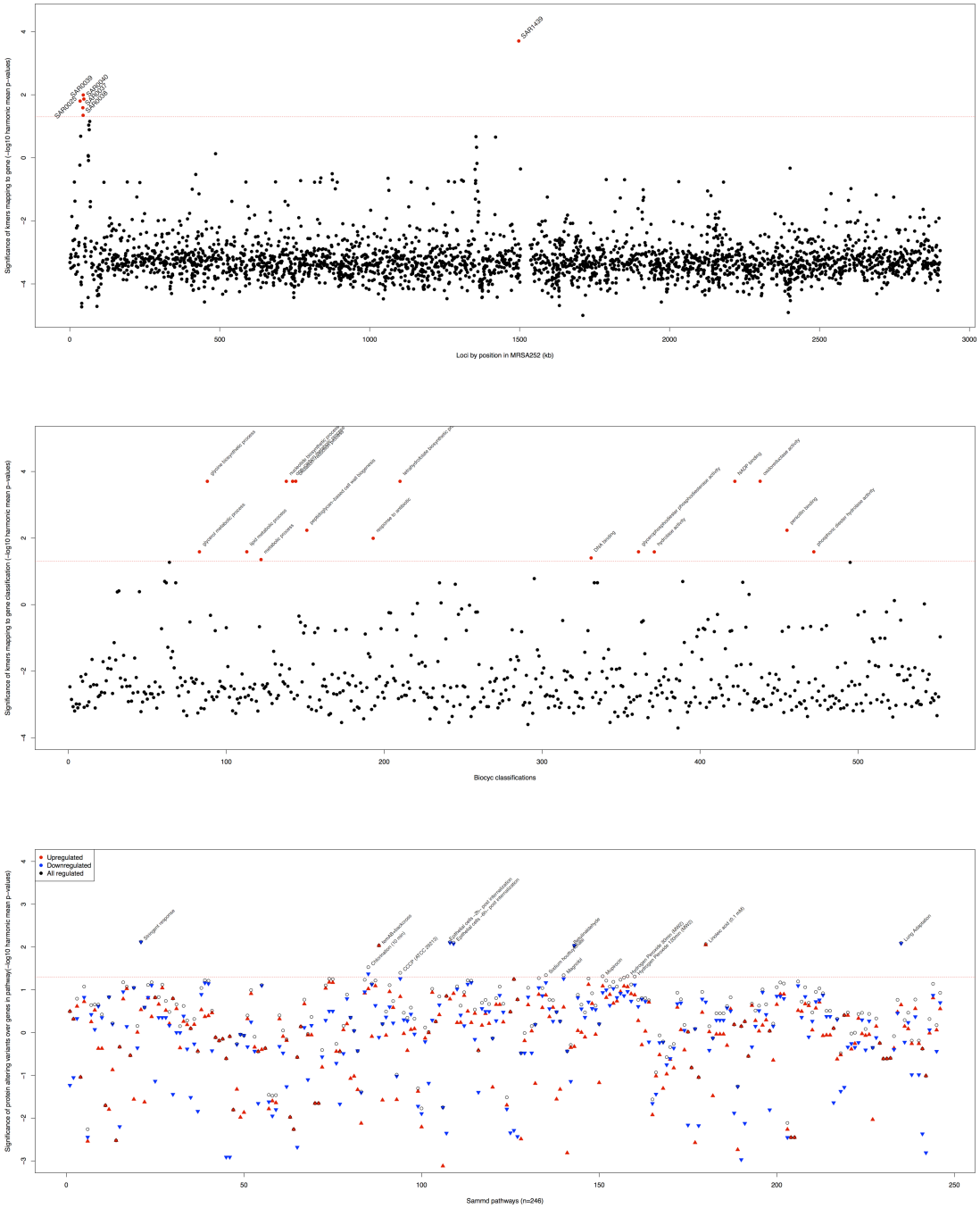


Figure 5.12: Manhattan plot of harmonic mean *p*-values for protein-altering variants in each locus in MRSA252, 552 ontological classifications and 246 transcriptional pathways. Each point represents the $-\log_{10}$ adjusted *p*-value, adjusted for multiple-testing by proportion of all variants that are found in the gene or genes in question. A threshold controlling the family-wide error rate at 0.05 is plotted in red. **(A)** and **(B)** black points represent a gene **(A)** or set of genes with shared classification **(B)**. **(C)** Black points represent all genes with altered regulation in the study conditions, red points represent groups of u*p*-regulated genes and blue points represent groups of down-regulated genes.

The evidence for kmers associated with disease, when pooled, was significant in a number of loci (Figure 5.12A). The strongest evidence was for SAR1439, or *dfrB* (HMP=$10^{-3.7}$). 5 additional loci in sccMec showed significant association (HMP=$10^{-1.3}$-$10^{-2.0}$): of these the strongest evidence was for *mecA* (SAR0039, HMP=$10^{-2.0}$), which encodes the low affinity PBP2 that confers methicillin resistance, followed by McrR1 (SAR0040, HMP=$10^{-1.9}$), a regulatory protein in sccMec.

17 gene classifications and 14 transcriptional pathways showed significant pooled evidence for association with disease. The gene classifications with strongest evidence included the tetrahydrofolate biosynthetic process, as well as glycine biosynthesis, nucleotide biosynthesis, one-carbon metabolism, redox processes, NADP binding and oxioreductase activity (all HMP=$10^{-3.7}$). The striking similarity in these HMP-values is due to the contribution of *dfrB* to all of them: harmonic mean *p*-values are strongly affected by the smallest values. The classification for genes involved in penicillin-binding, response to antibiotic and peptidoglycan-based cell wall biogenesis also demonstrated strong evidence of association (HMP $10^{-2.0}$ to $10^{-2.3}$): *mecA* kmers contributed strongly to these findings.

The transcriptional pathways with the strongest evidence of kmers associated with SAB are those with the strongest (though not statistically significant) evidence in the SNP association study, and also those with the largest regulons: exhibition of the stringent response, internalisation in epithelial cells and adaptation to the lung (all HMP=$10^{-2.1}$). The remaining pathways that have statistically significant pooled evidence for a genetic association with phenotype likewise have large regulons, with a median of 519 genes showing altered regulation in the study conditions.

In this analysis, genes and variants that cause antibiotic resistance are associated with bacteraemia, particularly elements that cause or are closely associated with methicillin resistance. Overall we found a higher rate of MRSA in SAB than in

carriage in our sampling set. It is possible that some of this disparity arises from a sampling bias, as CA SAB cases could have greater hospital exposure, and MRSA is most commonly – though not exclusively – a healthcare-associated organism in the UK.

We can attempt to control for sampling bias by including whether the isolate is HA and methicillin resistant as additional fixed effects in the LMM performed by GEMMA. If HA status alone is included, the overall estimated sample heritability is only slightly reduced from 2.1% to 2.0% (95% CI 0-7.1% for both). If methicillin resistance is included, it reduces much more to 0.45% (95% CI 0-2.5%). When both are included, the heritability is 0.5%, and the association between kmers in *dfrB* is no longer statistically significant, and there is no evidence of significant association between *dfrB* or *mecA* and disease when we pool evidence across kmers (Figure 5.13). The total evidence for association between kmers and phenotype is now low, with HMP=0.96 when the evidence from all kmers is pooled. The modest evidence for heritability in our sample thus partitions with MRSA status, and when MRSA is included as a fixed effect, no region of the genome shows evidence for association with case/control status when all carriage and bacteraemia samples are studied together.

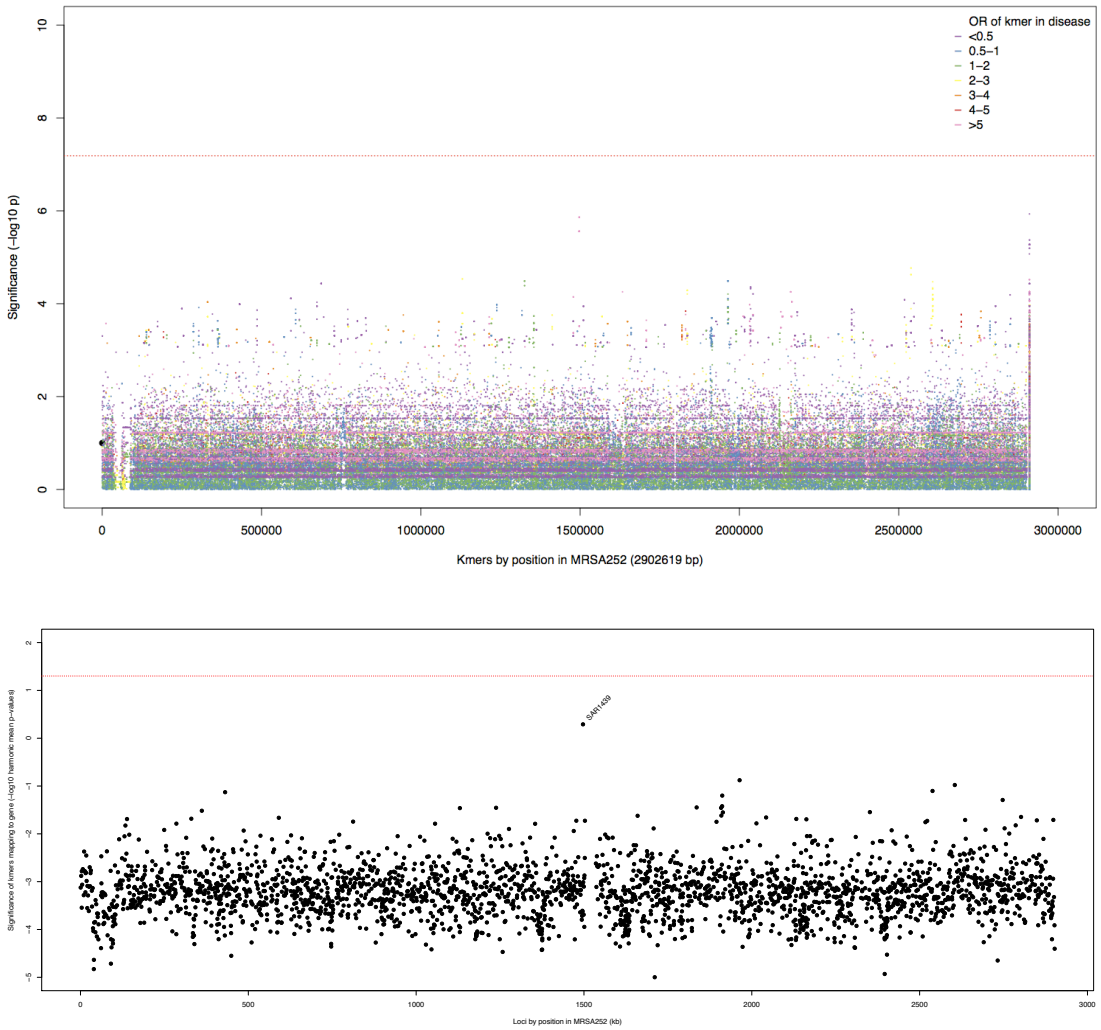


Figure 5.13: Kmer association testing results when methicillin resistance and HA/CA origin are included as covariates. (A) 10,000 most significant kmers and 200,000 randomly chosen kmers are plotted by their position on mapping to reference genome (MRSA252) and significance of association with phenotype. Each kmer is coloured by OR of being found in disease. A Bonferroni corrected threshold of significance is plotted in red. (B) Pooled evidence for association between kmers in each locus of the reference genome. A threshold controlling the family-wide error rate at 0.05 is plotted in red.

5.3.5 Cryptic population structure exists in *S. aureus* carriage, and specific variants are associated with healthcare-associated carriage.

In the previous section we found limited evidence for bacterial genetic association with SAB in the population as a whole. There are good reasons to think that HA SAB will have a greater non-bacterial contribution to disease, as ill hosts, medical intervention and antibiotic selection pressure are all at play. A GWAS limited to CA carriage and SAB is therefore a logical sub-group analysis and we would expect this

subset to be enriched for any bacterial contribution to disease. However, given the sampling differences between CA cases and controls, there is a risk of confounding by sampling bias in this analysis, if CA and HA carriage strains differ beyond a lineage effect. Therefore we conducted a GWAS of 984 carried isolates (330 HA, 654 CA), to examine for bacterial genetic associations with HA strains found in carriage.

Lineage testing with *bugwas* demonstrated no PCs associated with HA carriage at a genome wide level. The strongest lineage association was for PC3, which identifies MRSA CC-22 (p=0.07) (Fig 5.14).

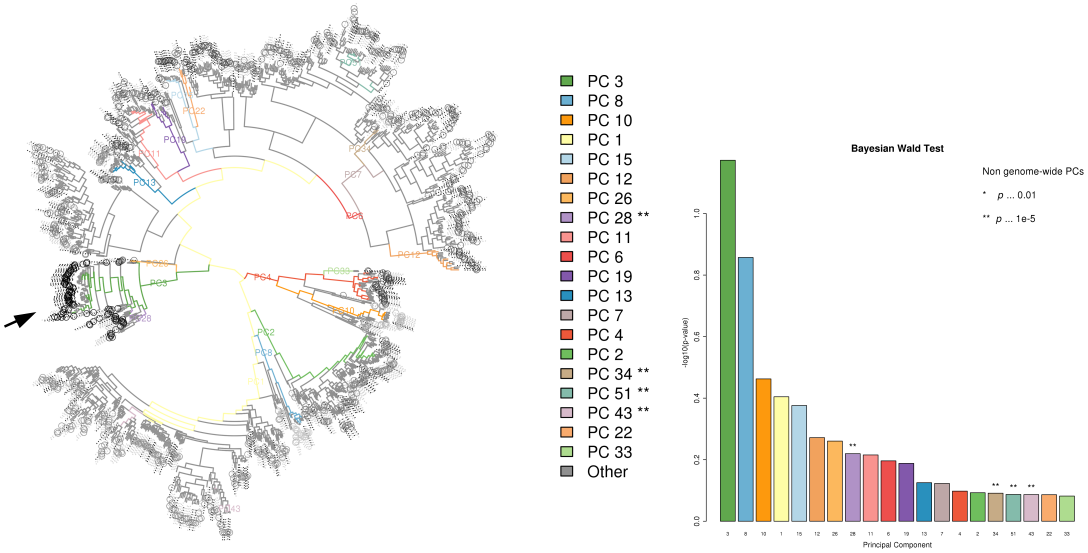
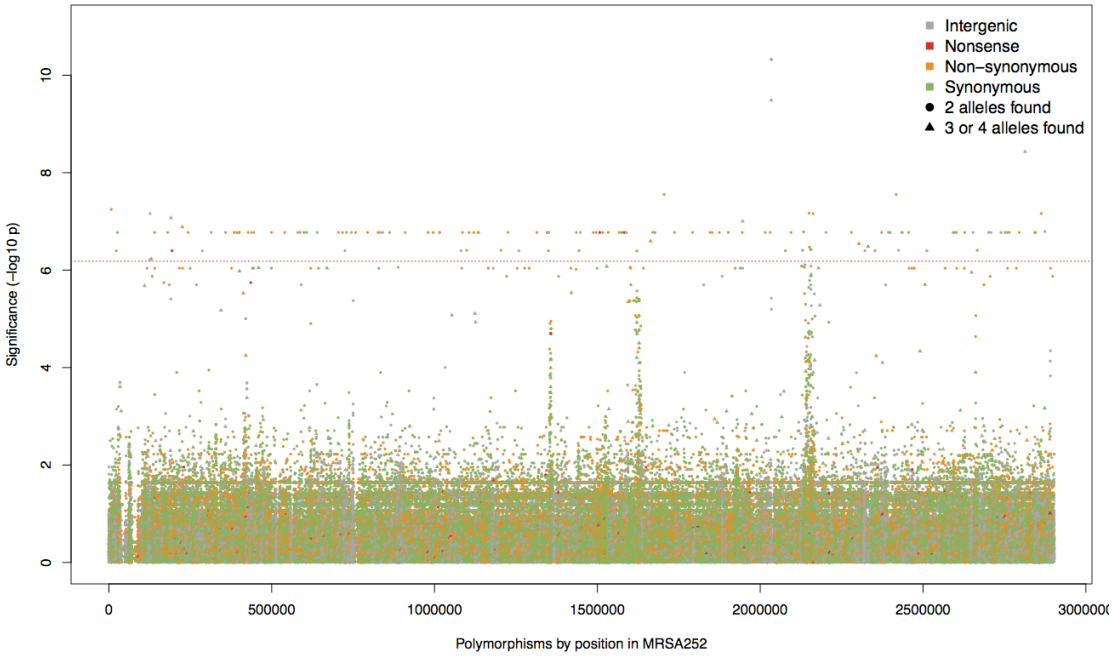


Figure 5.14 Lineage associations in 984 isolates between CA carriage (*n*=654) and HA carriage (*n*=330). (A) The branches corresponding to 20 most significantly associated lineages are coloured on a maximum likelihood phylogeny of the study isolates. Lines at the tips of the tree represent the phenotype of each isolate (grey, CA carriage; black HA carriage), while circles represent the phenotype predicted by the LMM (grey, CA carriage; black HA carriage). **(B)** Significance (-$\log_{10}$ p-values) of the 20 most significantly associated PCs. A Bonferroni corrected threshold for significance is $10^{-4.3}$.

On testing for SNP associations we found 134 SNPs significantly associated with HA carriage. However, there was evidence of population structure, with a strong horizontal band of alleles showing identical evidence for association (Fig 5.15A).

This is strongly suggestive of linkage disequilibrium and population structure not completely accounted for by the LMM. LMM are able to control for lineages, and cryptic population structure through the use of relatedness matrix, but they make the assumption that closely related isolates are unlikely to have large differences in phenotype. This assumption means that LMM may fail to account for lineage effects arising from closely related strains which vary significantly in frequency between phenotypes. In this case the SNPS in LD are found in the predominantly HA-MRSA isolates within CC-22. This may represent either a true association with that lineage, or confounding by sampling bias.



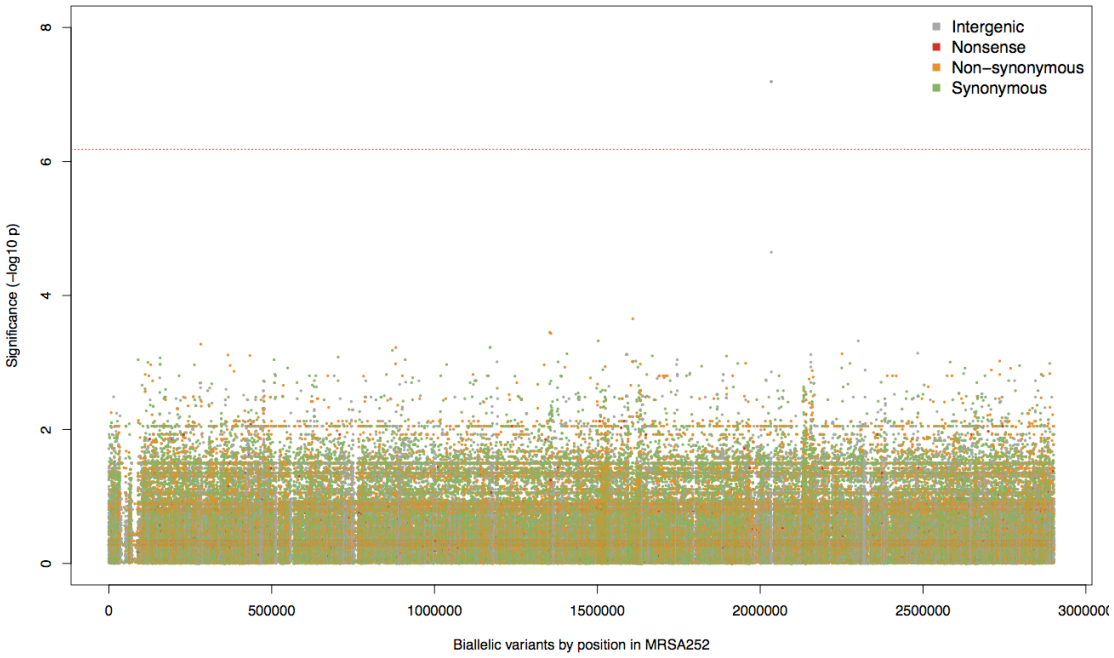
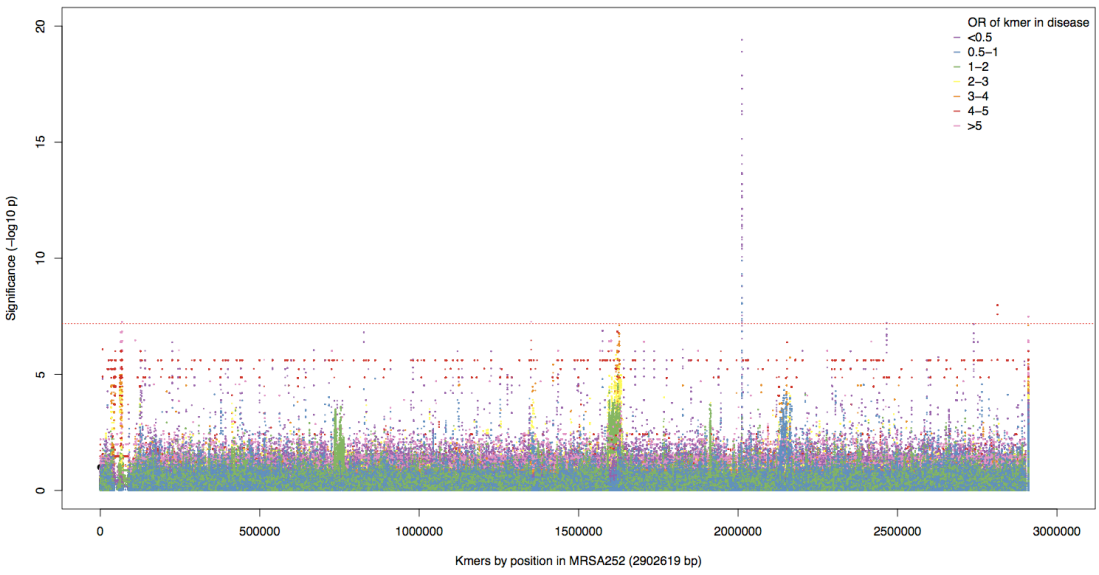


Figure 5.15: Manhattan plot of SNP associations with HA vs CA carriage. The significance ($-\log_{10}$ *p*-value) of each SNP is plotted according to its location on the 2.9 MB MRSA252 reference genome, with control for population structure. SNPs are coloured according to their predicted effect on protein, and whether the polymorphism was biallelic or tri/tetrallelic. A Bonferroni-adjusted threshold for significance is plotted in red. **(A)** Testing all polymorphisms for association with HA carriage, controlling for population structure. **(B)** Testing biallelic variants for association with HA carriage, controlling for methicillin resistance and population structure.

When MRSA status was included in the model, almost none of these variants remained significant. Only 3 biallelic SNPS remained significantly associated with HA vs CA carriage, (Fig 5.15B). These were 3 non-coding variants, localised to the intergenic region between two tRNA genes (tRNA-Leu and t-RNA-Gly), in complete LD with one another. Using SNP imputation for missing calls, these SNPs had OR 2.1 for association with HA vs CA carriage ($p=10^{-7.2}$). However, these SNPs were only called in 379/984 sequences (38.5%). When imputed calls at these sites were excluded, the OR for finding these SNP in HA vs CA carriage was 0.87, suggesting that the observed association was a product, and possibly an artefact of SNP imputation.

All control strains were sequenced with 151bp reads, allowing their inclusion in a kmer-based analysis without introducing batch effects arising from read length. Using 31bp kmers for association testing we found 184 kmers mapping across several regions strongly associated with HA vs CA carriage (Figure 5.16A). The most significant of these were 123 kmers mapping to MRSA252 between bases 2011974 and 2012190, which are found less often in HA than CA carriage (OR 0.02-0.58, p=$10^{-7.3}$ to $10^{-19.4}$) (Fig 5.16). This 217bp region is part of SAR1932, a 330 amino acid foldase protein (*prsA*), which performs post-translational folding of extra-cellular protein, a key role in the secretion of proteins. These kmers map to the peptidylprolyl isomerase region, which is predicted by the UniProt annotation database[43] to catalyse structural change of secreted proteins. A further 26 kmers mapped to *sasG*, a surface protein and putative virulence factor,[44] with OR 7.2 for association with HA vs CA carriage (p=$10^{-7.5}$). There were also several peaks of near-genome wide significance at 50kb and 1.6Mb.



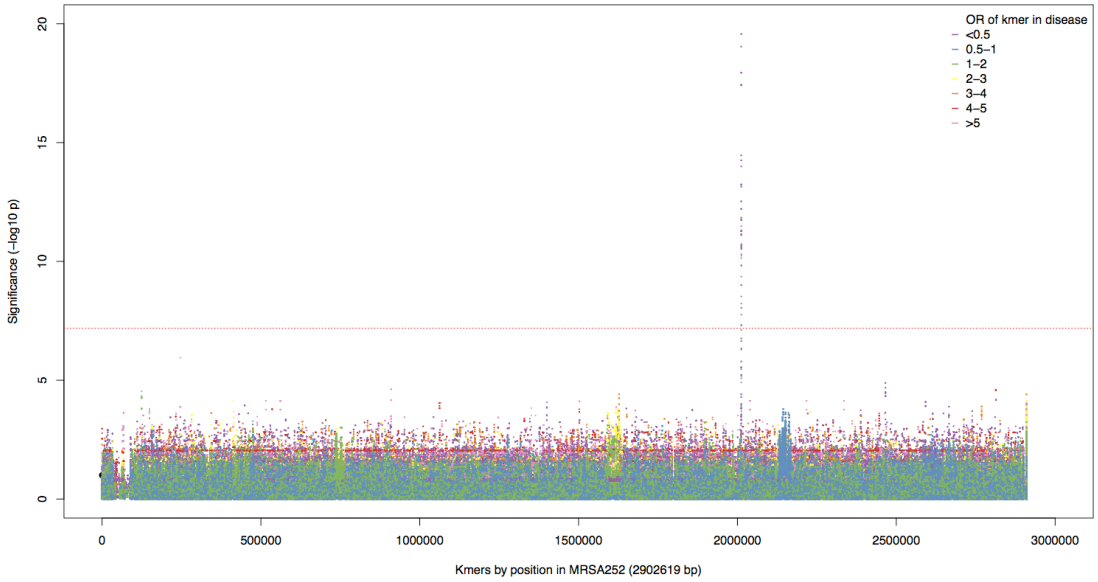


Figure 5.16: Association between kmers and HA vs CA carriage. Each point represents a single kmer. Kmers are plotted by their position on mapping to MRSA252 and the -$\log_{10}$ *p*-value of their association with HA vs CA carriage. Points are coloured by their OR of being found in HA carriage. A Bonferroni corrected threshold for significance is plotted in red. Association testing is performed with control for relatedness **(A)**, and with control for relatedness and methicillin resistance **(B)**.

When methicillin resistance was included as a covariate, the significance of most kmer associations with HA carriage decreased, but we continued to observe strong evidence for association for kmers mapping to *prsA*. 103 kmers mapping to the same 217bp region were found less often in HA carriage (OR 0.02-0.58, $p=10^{-7.3}$ to $10^{-19.6}$) (Fig 5.16). When evidence of kmer associations was pooled across loci, only this locus had significant evidence of association with HA vs CA carriage ($p=10^{-13.1}$). The *prsA* kmers were found to have variable presence in many lineages, though they were most commonly absent in CC-22, both MSSA and MRSA strains (Fig 5.17).

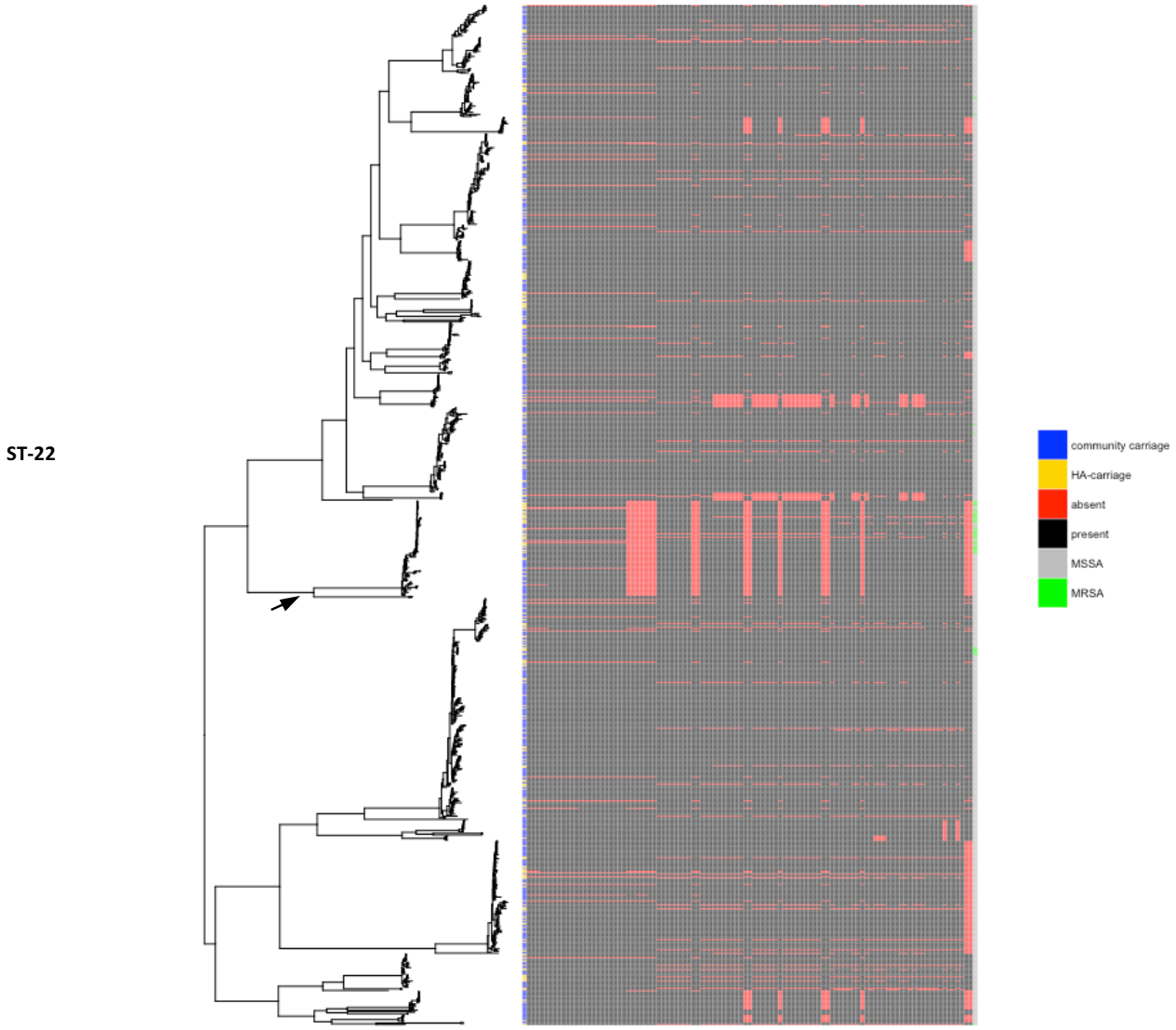


Figure 5.17 Relationship between phylogenetic tree and variation in *prsA* for 984 carried *S. aureus* isolates. Community carriage isolates are marked in blue, and healthcare-associated carriage isolates in gold along the left edge. Kmers significantly associated with HA carriage are plotted by their presence (black) or absence (red) in each isolate, in decreasing significance of association left to right. MRSA isolates are marked in green on the right edge.

The surface bound foldase protein *prsA* has been implicated as a secondary resistance factor for both beta-lactam and glycopeptide antibiotics. *In vitro* assays have shown that *S. aureus* can survive despite disruption to *prsA*, but strains with *prsA* disruption are more susceptible to oxacillin.[45] PrsA expression is controlled by VraRS, a cell wall stress regulatory system, and upregulation of *prsA* accompanies

the acquisition of a glycopeptide-intermediate susceptibility phenotype.[45] This protein has been demonstrated to regulate expression of a methicillin-resistant phenotype independently of *mecA*, reducing the membrane quantity of PBP2A without altered transcription of *mecA*.[46] Additionally, *prsA* facilitates the secretion of proteases and phospholipases, and the secretion of these virulence factors is decreased when *prsA* is deleted.[47]

This study demonstrates that while no lineage is strongly associated with HA carriage, there is population structure that distinguishes HA and CA carriage which is not completely controlled by LMM. We also find that, even after controlling for methicillin resistance in addition to bacterial lineage, variation in *prsA*, a modulator of methicillin susceptibility, strongly predicts a carried strain is of hospital origin.

5.3.6 SNPs in MRSA-associated mobile genetic elements and healthcare-associated foldase variants are associated with "community" associated SAB

We next investigated for bacterial genetic factors associated with CA bacteraemia vs CA carriage by studying 1285 isolates, 654 controls from CA carriage and 631 cases from CA-SAB. Having demonstrated population structure associated with HA strains, careful control for possible sampling bias formed a critical step in the analysis. As with the total set of 2001 isolates, the healthcare-associated MRSA CC-22 lineage showed the strongest association with bacteraemia (p=0.03)(Fig 5.18).

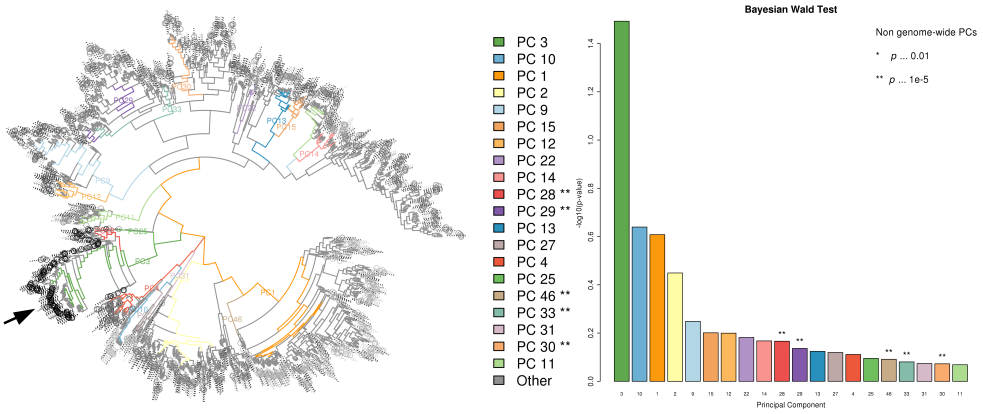


Figure 5.18 Lineage associations in 1285 isolates between CA carriage (*n*=654) and CA bacteraemia (*n*=631).). (A) The branches corresponding to 20 most significantly associated lineages are coloured on a maximum likelihood phylogeny of the study isolates. Lines at the tips of the tree represent the phenotype of each isolate (grey, CA carriage; black CA bacteraemia), while circles represent the phenotype predicted by the LMM (grey, CA carriage; black CA bacteraemia). **(B)** Significance ($-\log_{10}$ *p*-values) of the 20 most significantly associated PCs. A Bonferroni corrected threshold for significance is $10^{-4.4}$.

Testing all SNPs for association demonstrated 120 SNPs significantly associated with CA-SAB vs CA carriage (Figure 15.19A), with particular peaks at the sites of two integrated prophages, φSa2 and φSa3. As in section 5.2.5 above, the distribution of significant variants was strongly suggestive of incomplete control for population structure. When MRSA was included as a fixed effect in the GEMMA analysis, no variants were significantly associated with CA-SAB. The variants seen in association with CA-SAB can thus be explained by the difference in rates of MRSA – and MGEs associated with MRSA lineages – in this population.

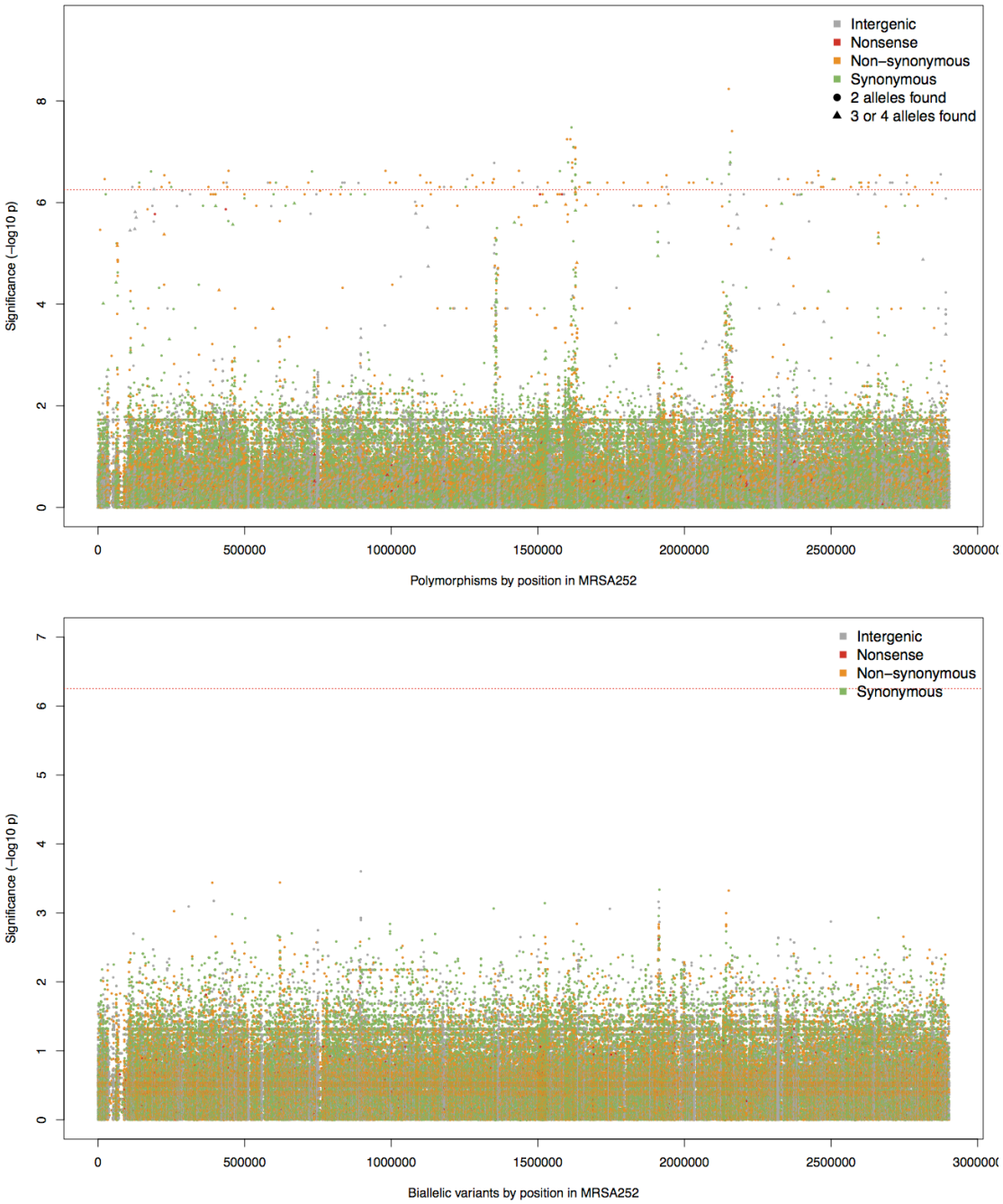


Figure 5.19: Manhattan plot of SNP associations with CA-SAB vs CA carriage. The significance (-$\log_{10}$ *p*-value) of each SNP is plotted according to its location on the 2.9 MB MRSA252 reference genome, with control for population structure. SNPs are coloured according to their predicted effect on protein, and whether the polymorphism was biallelic or tri/tetrallelic. A Bonferroni-adjusted threshold for significance is plotted in red. **(A)** Testing all polymorphisms for association with CA SAB, controlling for population structure. **(B)** Testing biallelic variants for association with CA SAB, controlling for methicillin resistance and population structure.

A kmer GWAS was performed on CA controls versus all CA cases sequenced with reads of 151bp (*n*=410) (Figure 5.20). Again, when MRSA was not included as a

covariate, peaks of *p*-values suggestive of significant association occurred, here in the region of sccMec, at 1.3Mb and in unmapped kmers. When we controlled for MRSA, the majority of these associations disappeared, leaving only a single peak of significant kmers at mapping to *prsA* (Figure 5.20B). Pooling of evidence using harmonic mean *p*-values found no evidence for significant association with CA SAB vs carriage outside this locus.

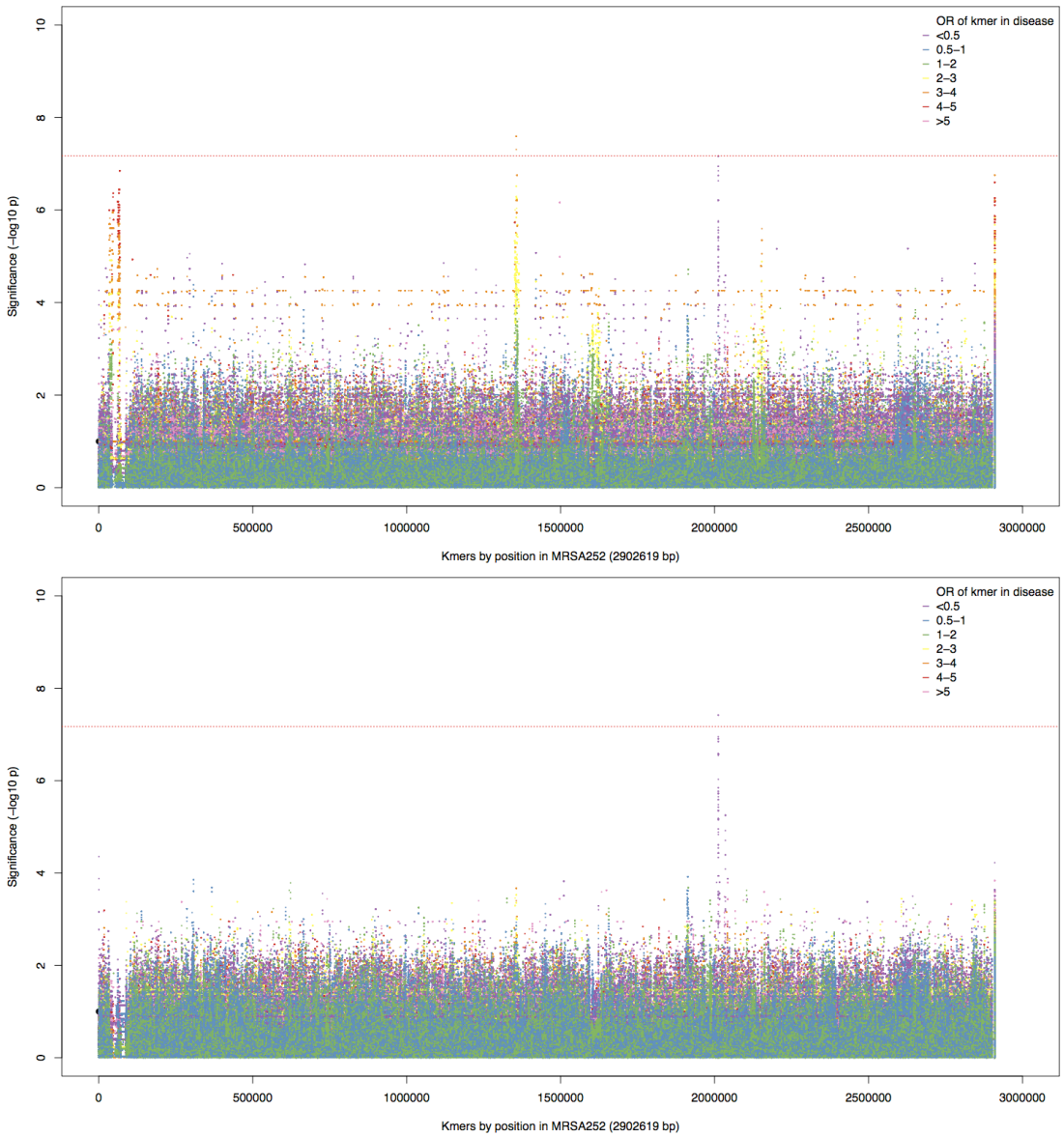


Figure 5.20: Association between kmers and CA SAB vs CA carriage. Each point represents a single kmer. Kmers are plotted by their position on mapping to MRSA252 and the -$\log_{10}$ *p*-value of their association with the phenotype. Points are coloured by their OR of being found in CA SAB. A Bonferroni corrected threshold for significance is plotted in red. Association testing is performed with control for relatedness **(A)**, and with control for relatedness and methicillin resistance **(B)**.

Kmers mapping to *prsA* had a lower odds of being found in CA SAB than CA carriage. The most significantly associated kmers had OR 0.06 being found in CA SAB. They showed the same direction of effect as the association between *prsaA* kmers and HA vs CA carriage, but the effect was not quite as strong: we found OR of 0.02 for these kmers being found in HA carriage. This is consistent with the association between CA SAB and *prsA* kmers being due to confounding by cryptic HA cases.

Thus, in addition to SNP associations arising from the population structure of MRSA, our CA bacteraemia isolates were relatively enriched for variants that differentiated HA from CA carriage. Our case definition, which is among the more stringent operational definitions of CA bloodstream infection, nevertheless misidentifies some cryptic HA cases as CA. WGS is able to demonstrate variants associated with HA carriage – and thus suggestive of a HA origin of the isolate – even when the exposure is relatively remote from admission.

5.3.7 Total heritability of *S. aureus* bacteraemia is low, and mostly accounted for by differences in MRSA rates.

In our main analysis of carriage and SAB, as well as two sub-analyses of CA vs HA carriage and CA carriage vs CA SAB, we found few variants associated with our phenotypes of interest. Almost all our findings were better accounted for by differing rates of MRSA lineages identified with these phenotypes. The overall sample heritability – that is, the proportion of variation in sample phenotype that GEMMA estimates is explained genetic variation – was low in all 3 studies, and was lower still when methicillin resistance was included as a covariate in the model (Table 5.7).

Similarly, the pooled evidence for any variants associated with the studied phenotype was drastically reduced when we controlled for MRSA. One exception to this was the combined evidence for association between kmers and HA vs CA

carriage: while the total heritability of this phenotype was estimated to be low after adjustment for MRSA frequency (95% confidence interval 0-1%), there was strong evidence for association between kmers mapping to *prsA* and that (albeit very small) component of heritability.

This demonstrates that when all SAB cases are compared with carriage, the total heritability of the phenotype is low. This does not prevent us from identifying associations at the kmer level, but it does suggest the overall impact of these variants is low.

	Sample heritability (estimate, 95% CI)		Sample heritability adjusted for covariates (estimate, 95% CI)		Pooled evidence for variants associated with phenotype (*p*-value)		Pooled evidence for variants associated with phenotype, adjusted for covariates (*p*-value)	
	SNPs	Kmers	SNPs	Kmers	SNPs	Kmers	SNPs	Kmers
Bacteraemia vs nasal carriage (*n*=2001)	2.1% (0 - 5.3)	2.1% (0 - 7.1)	0.4% (0 - 1.6)	0.5% (0 - 2.8)	0.04	$10^{-3.9}$	1	0.96
HA carriage vs CA carriage, (*n*=984)	2.8% (0 - 9.4)	9.2% (0 - 21.8)	$10^{-6.2}$ (0 - 10^{-2})	$10^{-4.3}$ (0 - 10^{-2})	$10^{-4.0}$	$10^{-13.0}$	1	$10^{-13.2}$
CA bacteraemia vs CA controls (*n*=1285)	2.1% (0 - 5.9)	1.2 % (0 - 4.7)	0.2% (0 - 1.1)	0.2% (0 - 1.3)	$10^{-3.6}$	$10^{-2.7}$	1	0.08

Table 5.7: Predicted heritability (estimate and 95% confidence interval) in each phenotype, with and without adjustment for covariates. After controlling for covariates, the pooled evidence for association is significant only between kmers in *prsA* and HA carriage vs CA carriage.

5.3.8 Analysis of recent variants shows enrichment in the GraS regulon in *S. aureus* isolated from the bloodstream

This study of common genetic variants across population of *S. aureus* found in carriage and bacteraemia did not reveal evidence of genetic variation increasing the odds of bacteraemia. A population-based GWAS will lack power to find such evidence if there is variation associated with the phenotype of interest that is not

common between strains. The majority of variants found in *S. aureus* are rare: 49.2% of SNPs have a minor allele frequency of <1%[30]. If multiple variants impacting on the phenotype occur in close proximity in the genome, our power is improved by the use of kmers, as kmers of the unaffected (or 'wild type') sequence will have greater significance when they span multiple risk variants, but this is restricted to regions less than or equal to the kmer length (31bp in this instance).

In Chapter 4, we identified a signal of adaptation by identifying variation arising across patients in particular loci and transcriptional pathways. We hypothesised that recently arising, polygenic variation in key loci or pathways may be associated with bacteraemia at a population level.

Direct comparison of variation in carriage and control isolates at a gene level would be subject to confounding by population structure. Variants related to population structure occur on the deep branches of the phylogeny (Figure 5.4). To replicate the analysis performed in chapter 4, we identified and compared variation between the most closely related cases and controls. Matching cases and controls by genetic background guards against the confounding effects of population structure. We hypothesised that these variants are the most recent, and will include variation arising within hosts during carriage and disease.

From the maximum likelihood phylogeny of 2001 carriage and bacteraemia isolates constructed in RAxML (Figure 5.3), we identified 270 isolates whose terminal branch lengths showed a per base divergence less than 6×10^{-5}, and identified the two closest relatives to each isolate (Figure 5.21). Variation between these isolates was determined using mapping to a single common reference and by comparing *de novo* assemblies with the Cortex variant caller. Variants where A possessed a different allele to these close relatives (ie variants unique to A within the trio) were inferred to occur on the branch leading to A, and these were annotated using the

methods previously described. The branch length selected corresponded to fewer than 60 unique variants in each isolate, which is the approximate upper bound of within host variation we observed in chapter 4 (though much higher than the mean of 5.7 variants observed per individual).

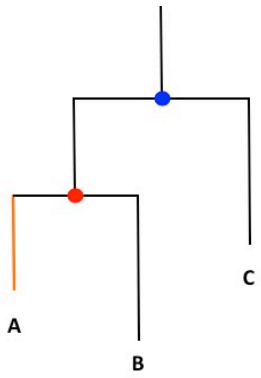


Figure 5.21: Recent variation identified by comparison of individual isolates and nearest relatives according to ML phylogeny. Isolate for study (A) is identified by relatively short terminal branch (orange). Isolates for comparison are identified as the isolate with the shortest descending branch from A's ancestral node (B), and the isolate with the shortest descending branch from the ancestral node of A and B (C). Variants where the allele found in A differs from that in B and C are inferred to occur on the terminal (orange) branch to A.

In total 1570 variants on the terminal branches to disease isolates and 1442 variants on the terminal branches to carriage isolates were identified. There were no significant differences in the number of variants found, or the rate of discovery of non-synonymous, synonymous or non-coding variants between the two groups.

		Carriage	Bacteraemia		All
Isolates (*n*)		143	127		270
Isolates with any unique variants (*n*)		94	96	*p*=0.15*	190
Total number of unique variants found, (*n* (range))		1442 (1-47)	1570 (1-52)		3012 (1-52)
Median number of unique variants (median, (range))		10.5 (3-25)	11 (4-29)	*p*=0.42**	3012 11 (3-27)
Variants by predicted effect on protein	Protein altering	812	917	*p*=0.35***	1729
	Synonymous	246	270		516
	Non coding	384	383		767

Table 5.8: Unique variants arising in carriage and bacteraemia. Observations between carriage and disease isolates compared with * χ^2 test with 1 degree of freedom, ** Mann-Witney test, *** χ^2 test with 2 degrees of freedom

These variants were tested for enrichment in loci, ontologies and transcriptional pathways using the GSEA methods applied in Chapter 4 (Figure 5.22). In both carriage and disease isolates we found significant enrichment for variation in *agrC*, the sensor component of the response arm of the Accessory Gene Regulator quorum-sensing system. This locus was 17- and 14-fold enriched for protein altering variants unique in disease and carriage isolates respectively (p=$10^{-7.4}$ and $10^{-5.3}$)(Figure 5.22A). No other loci showed significant enrichment.

No ontological classifications as described in the Biocyc database showed significant enrichment (Figure 5.22B). Two classifications showed near significant results – DNA integration in carriage and transpose activity. No variants were found in either of these classifications in carriage or disease, and these results represent depletion of variation compared to that expected under a neutral model.

C

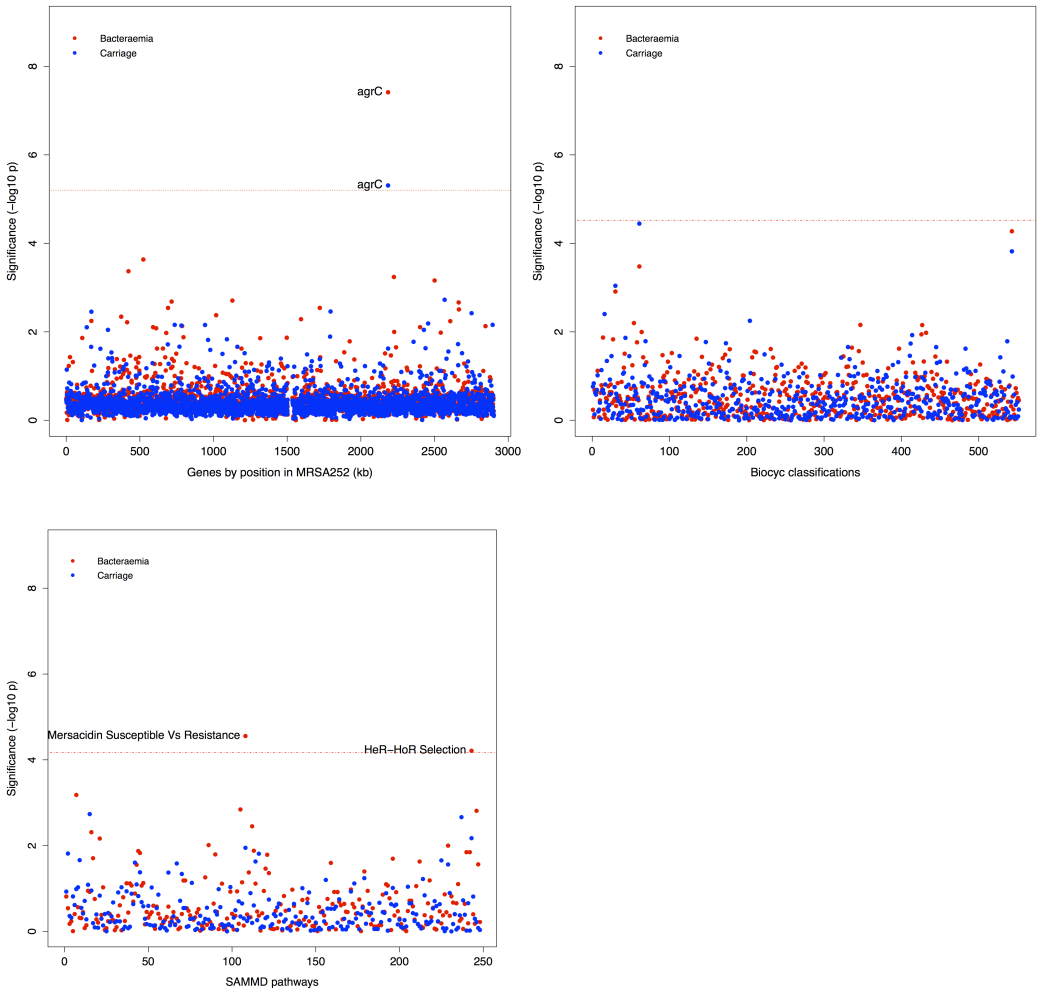


Figure 5.22: GSEA of unique variants found in 124 isolates from bacteraemia and 146 isolates from carriage. Significance of enrichment or depletion of protein altering variants in **(A)** loci, **(B)** ontological classifications from Biocyc database and **(C)** transcriptional pathways from SAMMD database. Enrichment among carriage isolates plotted in blue, enrichment among disease isolates plotted in red. A Bonferroni-corrected threshold for significance is plotted in red.

Two SAMMD transcriptional pathways showed significant enrichment for variants in disease only (Fig 5.22C). Variation in genes with reduced expression in a Mersacidin susceptible strain was enriched 2.7-fold, while no variants were seen in genes up-regulated in this strain (p=$10^{-4.6}$). In carriage isolates no significant enrichment or depletion was found in this pathway (p=0.01).

This transcriptional pathway was identified by comparing a common laboratory strain (SG511-Berlin) with a German MRSA clone (SA137/9A). A comparison of transcription in these strains revealed divergent regulation of 26 genes under the

control of the two-component regulatory system GraRS.[48] Further characterisation of SG511-Berlin and SA137/9A showed that the former had acquired an insertion in the *graS* sensor gene, and this complementation of this gene reversed the phenotypic differences in antibiotic sensitivity (including sensitivity to Mersacidin) observed between these strains.[48] Further characterisation of GraRS has demonstrated that it plays a key role in resistance to cationic antimicrobial peptides.[49] GraRS is specifically induced by CAMP release by platelets, but not by exposure to CAMP release from neutrophils or in response to vancomycin. The variants contributing to this enrichment in disease strains include variants in two other regulatory loci (*agrC* and *vicK*), capsule synthesis genes, urease, surface transport proteins, *mecA* and the D-alanine transport protein *dltB*.

It is notable that this enrichment included variation in the methicillin resistance conferring gene *mecA*, and that the pathway has also been implicated in both resistance to mersascidin – an antibiotic peptide that inhibits peptidoglycan synthesis[50] – and development of the vancomycin-intermediate *S. aureus* phenotypes.[49] Variants in this pathway are found in 21 of 96 disease isolates, but 15/21 (71%) of these were MRSA, suggesting this enrichment might be relatively specific to MRSA.

The second pathway showing significant enrichment in disease was the set of genes with increased transcription in transition to homogenous methicillin resistance, showing 3-fold enrichment (p=$10^{-4.2}$). *S. aureus* possessing *mecA* but demonstrating heterotypic methicillin resistance (HeR) underwent selection of methicillin resistance through exposure to an oxacillin-induced stress response, and gene expression in these isolates was compared with expression in bacteria exhibiting homotypic resistance (HoR).[51] The up-regulated genes included *agrC*, and through knock-out and complementation experiments the authors conclude that Agr expression is a necessary and mechanistic component of the stress-response

mediated, beta-lactam induced mutagenesis that determines homogenenous expression of methicillin resistance.[51] The variants contributing to this enrichment occur in *agrA* as well as *agrC*, *mecA*, capsular synthesis enzymes, and the staphylokinase and autolysin toxin genes. 13 of the 16 isolates contributing to this signal were MRSA (81%).

Overall we found enrichment for recent variation among disease isolates in genes whose expression is regulated in response to the innate immune antimicrobials, as well as peptide antibiotics and methicillin. It is possible these facilitate survival in the bloodstream or under selective pressures of antibiotics, however we must be cautious about drawing a strong conclusion about differential signal of enrichment between disease and carriage isolates given the relatively higher number of MRSA isolates in our sample. In fact, MRSA isolates form a relatively high proportion of the bacteria with a short final branch length in this population – being 45% of disease and 25% of carriage isolates included in this unique variant analysis, compared with 14% and 6% of the GWAS study population. 66.1% and 65.3% of variants identified in disease and carriage isolates respectively were identified in MRSA isolates. This is likely a result of the highly clonal MRSA lineages, resulting in more detailed sampling of this population and a greater opportunity to observe fine differences and recent variation between isolates.

5.4 Discussion

S. aureus bacteraemia is a serious and common infection, with an overall mortality of around 20% at 30 days, even with appropriate treatment. Whether or not the *S. aureus* bacterial genome contributes to the development of invasive bacteraemia is an important question for our understanding of *S. aureus* pathogenesis, and the rational design of treatment and prevention. A large body of evidence provides conflicting evidence: we present a large bacterial GWAS to add clarity to this issue.

We found a very limited evidence for contribution of bacterial genome to odds of invasive disease when all cases of SAB are considered. We found marginal evidence for an association between MRSA CC-22 isolates and bacteraemia, but this effect may be better explained by hospital-based sampling than by the lineage itself. At a population level we found evidence of antimicrobial resistance – conferred both by SNPs and gene presence – associated with increased odds of invasive disease. However these effects were not significant if methicillin resistance and hospital-based sampling were included as fixed effects in the LMM. The biggest impact on predicted sample heritability was that of controlling for MRSA

While we found a significantly higher proportion of MRSA in *S. aureus* bacteraemia than in nasal carriage, we cannot exclude a sampling artefact as the reason for this excess among CA *S. aureus* bacteraemia. The decision to exclude CA controls with overnight hospital admission in the previous 12 months created possible sampling bias, as hospital exposure may be systematically less in CA controls. This decision was made to ensure CA controls truly represented bacterial isolates circulating in the community. A more lenient definition of CA control would have prevented this bias but at the cost of increasing hospital related samples in our carriage population. Finally, it is likely that an imbalance of MRSA rates (and other hospital exposure effects) would occur between case and control groups even with a more lenient definition of community carriage, as 60% of our community controls come from a study of community carriage (Fig 5.1, 5.2) in which the overall detection of MRSA was less than 1%.[18]

Even controlling for population structure, we saw associations that were likely due to differences of MRSA rates in case and control groups, and these associations were lost when methicillin resistance was included as a fixed effect in the LMM. Controlling for MRSA in this manner may be overly conservative. If the genetic mechanism responsible for methicillin resistance is also able to affect the odds of

invasive disease, then by controlling for the presence of that element we are over-correcting. It has been postulated that antibiotic resistance incurs a fitness cost, and that this cost is only is only worth paying in an environment where antibiotic selection pressure ensures that resistance confers a selective advantage. However other evidence has demonstrated that the expression of altered PBP2a directly supresses Agr activity, and Agr function is restored by deletion of *mecA*.[14] Agr suppression has been demonstrated to enhance intracellular survival,[52] and survival in serum,[16] while reduction in the Agr controlled expression of PSMs impairs neutrophil mediated clearance of *S. aureus* from the bloodstream.[53,54]

The inclusion of MRSA as a covariate may thus have biased our results to the null, an example of over-adjustment bias.[55] In particular, the complete confounding of *mecA* with both methicillin resistance and a hypothesised mechanism for increased bacterial survival in the bloodstream made it impossible to disentangle any possible causality by adjustment or stratification (Figure 5.23). For this reason we could not exclude an association between *mecA* and an increased risk of SAB.

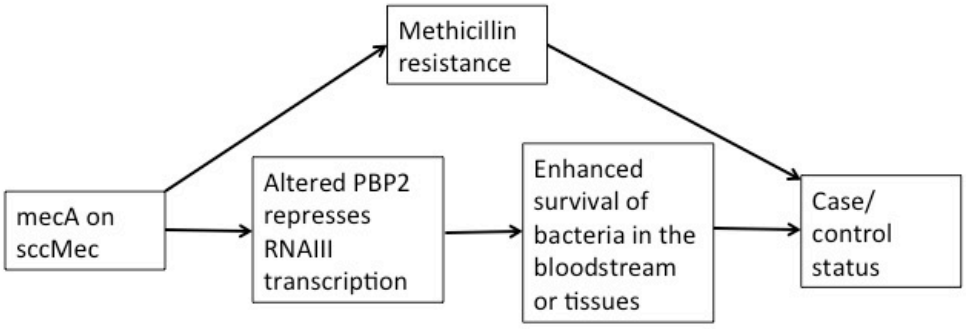


Figure 5.23: Directed acyclic graph demonstrates complete confounding of sccMec with resistance and possible gene directed effect on bacteraemia. A hypothesised causal pathway from *mecA* to the isolation of bacteria from the bloodstream is depicted.

With this caveat of not being able to exclude a *mecA* association, we found a paucity of evidence for a contribution of bacterial genotype to bacteraemia, with total contribution of the bacterial genome to phenotype estimated to be 0.5% in an LMM

implemented in GEMMA. This indicates there are not bacteraemia-specialist *S. aureus* isolates. At a population level, this evidence suggests that selection does not favour traits enhancing bacteraemia. This may in part be because the high mortality of SAB means this disease provides limited opportunity for transmission.

Hypothesising that CA SAB would have a higher bacterial contribution, we performed a sub-study excluding isolates from HA carriage and bacteraemia. Here we again found that overall heritability is low, estimated to be 0.2% when controlling for MRSA. Kmers with in the foldase protein *prsA* remained significant after controlling for MRSA, but these same kmers were much more significantly associated with HA carriage than CA SAB, and this result likely stemmed from differences in antimicrobial resistance pressure mediated through hospital exposure in CA cases and controls.

The findings of higher rates of MRSA, and the association of *prsA* kmers with CA bacteraemia suggests that epidemiological definitions of "CA" *S. aureus* bacteraemia still include some cases where strain acquisition arises from healthcare exposure. This suggests the interval between carriage acquisition and invasive disease can be several months long. These findings highlight the importance of controlling for population structure in bacterial GWAS to avoid multiple sources of confounding. Two recent *S. aureus* GWAS have focussed on single lineages defined by CC, including CC-22, to avoid confounding by population structure.[17,56] Our results suggest that the single lineages approach is still subject to confounding by population structure in *S. aureus*.

While this study did not find evidence of *S. aureus* genomic variation associated with SAB, we cannot conclude variation in the bacterial genome does not affect the odds of bacteraemia. Our study is a comprehensive GWAS based on short-read sequencing, and this technology cannot access the entire *S. aureus* genome. In

particular, repetitive elements of the genome are neither reliably called by mapping nor completely recovered from *de novo* assembly. In *S. aureus* these repeat regions include the ligand-binding regions of MSCRAMMs, genes that are thought to play an important role in tissue attachment, invasion and immune evasion. These genes are important candidates for playing a causal role in bacteraemia. For example, small studies sequencing a portion of *fnbA* have suggested variation within a repetitive region of gene affects the ability of *S. aureus* to adhere to medical devices in the bloodstream through enhanced binding to fibrinogen.[57,58] Enhanced fibrinogen binding has recently been demonstrated in a group of CC-22 MRSA isolates associated with increased risks of endocarditis.[59] So the technology we have been able to use here is unable to resolve some of the most interesting regions of the genome.

Our null result may also arise from choosing too broad a phenotype. We chose bacteraemia as the phenotype to study for several reasons: it is an objective phenotype, not subject to investigator bias; it is an important phenotype, in terms of both morbidity and mortality; finally, it is plausible that bacterial factors could contribute to this phenotype, as survival in the bloodstream represents a significant challenge in terms of metabolic stress and immune challenge when compared with the usual niche of *S. aureus*.

However, SAB may not represent a single entity. There is evidence from large cohort studies that outcomes of SAB vary according to the focus of infection.[6,7] Some of these, such as vascular catheter infection, probably have a very limited bacterial contribution, and will dilute any signal in our study. Likewise if the genetic basis for one focus (e.g. soft tissue infection) differs from that seen in another (e.g. respiratory infection), our study may fail to detect this.

In addition to a study for common variants across the population associated with SAB, we performed study of relatively recent variation in SAB compared with carriage. Using this approach, we aimed to find any signal of recent adaptation associated with SAB, including variation within single or linked genes across individuals.

We found enrichment in the sensor component of Agr, but this was enriched in the recent variants from both carriage and disease isolates. We found that among isolates from bloodstream infections there was enrichment for recent variation in the genes under control of GraRS, which mediates protection from CAMPs. This is related to the regulons with the greatest enrichment for variation separating carriage and disease isolates in chapter 4, where genes with altered regulation in response to specific platelet derived CAMPs were significantly enriched. However we did not find an identical signal in this analysis to that seen in our study of within-host evolution using paired isolates in Chapter 4. In particular, we did not find enrichment for cell wall anchored proteins or the response effector of the Agr system.

There are several possible reasons for this discrepancy. Isolates for the study of recent variation were selected based on finding relatively closely related isolates in the study population. This included isolates with up to 52 unique variants compared with their closest neighbour. While we found up to 60 variants separating carriage and disease populations within hosts, the mean number of variants found between populations was 5.7. Studies of stable carriage have estimated that *S. aureus* accrues an average of 8 variants per year,[60] but this may be higher in disease.[61] Without sequential samples, we cannot know how many of the variants unique to a strain represent within-host variation, and with the variable rates of within-host evolution it is difficult to predict how recent the variation we find is. There is likely to be significant noise to signal ratio, which may obscure our

ability to find true enrichment.

In this analysis, the isolate under study and its closest neighbours were relatively distinct compared to the isolates studied in chapter 4. We have found the sensitivity of the Cortex variant caller is reduced by increased distance between the reference used and the isolates under investigation (See Appendix for further details). The distance between isolates and their closest relatives had a median base divergence of 1.3×10^{-4} (estimated 397 SNPs). This may have affected the performance of Cortex variant caller in this analysis compared with our previous study of within-host evolution.

Overall this study finds no firm evidence for bacterial genomic variation associated with the phenotype of *S. aureus* bacteraemia.

References chapter 5

1. Lowy FD. *Staphylococcus aureus* infections N Engl J Med 1998 Aug 20;339(8):520-32.

2. Tong SY, Davis JS, Eichenberger E, Holland TL, Fowler VG Jr. *Staphylococcus aureus* infections: epidemiology, pathophysiology, clinical manifestations, and management. Clin Microbiol Rev. 2015 Jul;28(3):603-61.

3. Anderson DJ, Moehring RW, Sloane R, Schmader KE, Weber DJ, Fowler VG,Jr, Smathers E, Sexton DJ. Bloodstream infections in community hospitals in the 21st century: A multicenter cohort study PLoS One 2014 Mar 18;9(3):e91713.

4. Shorr AF, Tabak YP, Killian AD, Gupta V, Liu LZ, Kollef MH. Healthcare-associated bloodstream infection: A distinct entity? insights from a large U.S. database Crit Care Med 2006 Oct;34(10):2588-95.

5. Annual Epidemiological Commentary: Mandatory MRSA, MSSA and *E. coli* bacteraemia and *C. difficile* infection data 2016/7, Public Health England 2017 [online].

6. Thwaites GE, United Kingdom Clinical Infection Research Group (UKCIRG). The management of *Staphylococcus aureus* bacteremia in the United Kingdom and Vietnam: A multi-centre evaluation PLoS One 2010 Dec 13;5(12):e14170.

7. Kaasch AJ, Barlow G, Edgeworth JD, Fowler VGJ, Hellmich M, Hopkins S, Kern WV, Llewelyn MJ, Rieg S, Rodriguez-Bano J, et al. *Staphylococcus aureus* bloodstream infection: A pooled analysis of five prospective, observational studies. The Journal of Infection 2014 Mar;68(3):242-51.

8. Lenz R, Leal JR, Church DL, Gregson DB, Ross T, Laupland KB. The distinct category of healthcare associated bloodstream infections. BMC Infect Dis. 2012 Apr 9;12:85.

9. Messina JA, Thaden JT, Sharma-Kuinkel BK, Fowler VG,Jr. Impact of bacterial and human genetic variation on *Staphylococcus aureus* infections PLoS Pathog 2016 Jan 14;12(1):e1005330.

10. Peacock SJ, Moore CE, Justice A, Kantzanou M, Story L, Mackie K, O'Neill G, Day NP. Virulent combinations of adhesin and toxin genes in natural populations of *Staphylococcus aureus*. Infect Immun. 2002 Sep;70(9):4987-96.

11. Lindsay JA, Moore CE, Day NP, Peacock SJ, Witney AA, Stabler RA, Husain SE, Butcher PD, Hinds J. Microarrays reveal that each of the ten dominant lineages of *Staphylococcus aureus* has a unique combination of surface-associated and regulatory genes J Bacteriol 2006 Jan;188(2):669-76.

12. Feil EJ, Cooper JE, Grundmann H, Robinson DA, Enright MC, Berendt T, Peacock SJ, Smith JM, Murphy M, Spratt BG, et al. How clonal is *Staphylococcus aureus*? J Bacteriol 2003;185(11):3307-3316.

13. Fowler VG,Jr, Nelson CL, McIntyre LM, Kreiswirth BN, Monk A, Archer GL, Federspiel J, Naidich S, Remortel B, Rude T, et al. Potential associations between hematogenous complications and bacterial genotype in *Staphylococcus aureus* infection J Infect Dis 2007 Sep 1;196(5):738-47.

14. Painter KL, Krishna A, Wigneshweraraj S, Edwards AM. What role does the quorum-

sensing accessory gene regulator system play during *Staphylococcus aureus* bacteremia? Trends Microbiol 2014 Dec;22(12):676-85.

15. Das S, Lindemann C, Young BC, Muller J, Österreich B, Ternette N, Winkler AC, Paprotka K, Reinhardt R, Förstner KU, Allen E, et al. Natural mutations in a *Staphylococcus aureus* virulence regulator attenuate cytotoxicity but permit bacteremia and abscess formation Proc Natl Acad Sci U S A. 2016. 113(22):E3101-10.

16. Laabei M, Uhlemann AC, Lowy FD, Austin ED, Yokoyama M, Ouadi K, Feil E, Thorpe HA, Williams B, Perkins M, et al. Evolutionary trade-offs underlie the multi-faceted virulence of *Staphylococcus aureus* PLoS Biol 2015 Sep 2;13(9):e1002229.

17. Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, Sloan TJ, Williams P, Endres JL, Bayles KW, Fey PD, et al. Predicting the virulence of MRSA from its genome sequence Genome Res 2014 May;24(5):839-49.

18. Miller RR, Walker AS, Godwin H, Fung R, Votintseva A, Bowden R, Mant D, Peto TE, Crook DW, Knox K. Dynamics of acquisition and loss of carriage of *Staphylococcus aureus* strains in the community: the effect of clonal complex. J Infect. 2014 May;68(5):426-39.

19. Everitt RG, Didelot X, Batty EM, H, et al. Mobile elements drive recombination hotspots in the core genome of *Staphylococcus aureus* Nat Commun 2014 May 23;5:3956.

20. Gordon NC, et al. 2014. Prediction of *Staphylococcus aureus* antimicrobial resistance by whole-genome sequencing J Clin Microbiol 52(4):1182-91.

21. Finney JM, Walker AS, Peto TE, Wyllie DH. An efficient record linkage scheme using graphical analysis for identifier error detection. BMC Med Inform Decis Mak. 2011 Feb 1;11:7.

22. Cardoso T, Almeida M, Friedman ND, Aragão I, Costa-Pereira A, Sarmento AE, Azevedo L. Classification of healthcare-associated infection: a systematic review 10 years after the first proposal. BMC Med. 2014 Mar 6;12:40.

23. Miller R, Walker AS, Knox K, Wyllie D, Paul J, Haworth E, Mant D, Peto T, Crook DW. 'Feral' and 'wild'-type methicillin-resistant *Staphylococcus aureus* in the United Kingdom. Epidemiol Infect. 2010 May;138(5):655-65.

24. Gamblin J, Jefferies JM, Harris S, Ahmad N, Marsh P, Faust SN, Fraser S, Moore M, Roderick P, Blair I, Clarke SC. Nasal self-swabbing for estimating the prevalence of *Staphylococcus aureus* in the community. J Med Microbiol. 2013 Mar;62(Pt 3):437-40.

25. Otter JA, Herdman MT, Williams B, Tosas O, Edgeworth JD, French GL. Low prevalence of meticillin-resistant *Staphylococcus aureus* carriage at hospital admission: implications for risk-factor-based vs universal screening. J Hosp Infect. 2013 Feb;83(2):114-21.

26. Fuller C, Robotham J, Savage J, Hopkins S, Deeny SR, Stone S, Cookson B. The national one week prevalence audit of universal meticillin-resistant *Staphylococcus aureus* (MRSA) admission screening 2012. PLoS One. 2013 Sep 12;8(9):e74219.

27. Miller RM, Price JR, Batty EM, Didelot X, Wyllie D, Golubchik T, Crook DW, Paul J, Peto TE, Wilson DJ, et al. Healthcare-associated outbreak of meticillin-resistant *Staphylococcus aureus* bacteraemia: role of a cryptic variant of an epidemic clone. J Hosp Infect. 2014 Feb;86(2):83-9.

28. Wyllie DH, Walker AS, Miller R, Moore C, Williamson SR, Schlackow I, Finney JM, O'Connor L, Peto TE, Crook DW. Decline of meticillin-resistant *Staphylococcus aureus* in Oxfordshire hospitals is strain-specific and preceded infection-control intensification. BMJ Open. 2011 Aug 27;1(1):e000160.

29. Knight GM, Budd EL, Whitney L, Thornley A, Al-Ghusein H, Planche T, Lindsay JA. Shift in dominant hospital-associated methicillin-resistant *Staphylococcus aureus* (HA-MRSA) clones over time. J Antimicrob Chemother. 2012 Oct;67(10):2514-22.

30. Earle SG, Wu CH, Charlesworth J, Stoesser N, Gordon NC, Walker TM, Spencer CC, Iqbal Z, Clifton DA, Hopkins KL, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies Nat Microbiol 2016 Apr 4;1:16041.

31. Feil EJ, Cooper JE, Grundmann H, Robinson DA, Enright MC, Berendt T, Peacock SJ, Smith JM, Murphy M, Spratt BG, et al. How clonal is *Staphylococcus aureus*? J Bacteriol 2003;185(11):3307-3316.x

32. Wertheim HF, van Leeuwen WB, Snijders S, Vos MC, Voss A, Vandenbroucke-Grauls CM, Kluytmans JA, Verbrugh HA, van Belkum A. Associations between *Staphylococcus aureus* Genotype, Infection, and In-Hospital Mortality: A Nested Case-Control Study. J Infect Dis. 2005 Oct 1;192(7):1196-2000.

33. Messina JA, Thaden JT, Sharma-Kuinkel BK, Fowler VG,Jr. Impact of bacterial and human genetic variation on *Staphylococcus aureus* infections PLoS Pathog 2016 Jan 14;12(1):e1005330.

34. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies Nat Genet. 2012 Jun 17;44(7):821-4.

35. Gill SR, McIntyre LM, Nelson CL, Remortel B, Rude T, Reller LB, Fowler VG Jr. Potential associations between severity of infection and the presence of virulence-associated genes in clinical strains of *Staphylococcus aureus*. PLoS One. 2011 Apr 26;6(4):e18673.

36. Wilson DJ, The harmonic mean *p*-value and model averaging by mean maximum likelihood. (Preprint BioRxiv, August 2 2017, doi: https://doi.org/10.1101/171751).

37. Caspi R, Billington R, Ferrer L,et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases Nucleic Acids Res. 2016 Jan 4;44(D1):D471-80.

38. Nagarajan V and Elasri M. SAMMD: Staphylococcus aureus microarray meta-database. BMC Genomics. 2007; 8(1):351.

39. Anderson KL, Roberts C, Disz T, Vonstein V, Hwang K, Overbeek R, Olson PD, Projan SJ, Dunman PM. Characterization of the *Staphylococcus aureus* heat shock, cold shock, stringent, and SOS responses and their effects on log-phase mRNA turnover. J Bacteriol. 2006 Oct;188(19):6739-56.

40. Chaffin DO, Taylor D, Skerrett SJ, Rubens CE. Changes in the *Staphylococcus aureus* transcriptome during early adaptation to the lung. PLoS One.2012;7(8):e41329.

41. Garzoni C, Francois P, Huyghe A, Couzinet S, Tapparel C, Charbonnier Y, Renzoni A, Lucchini S, Lew DP, Vaudaux P, Kelley WL, Schrenzel J. A global view of *Staphylococcus aureus* whole genome expression upon internalization in human epithelial cells. BMC Genomics. 2007 Jun 14;8:171.

42. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2 Nat Methods. 2012 Mar 4;9(4):357-9.

43. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017 Jan 4;45(D1):D158-D169.

44. Foster TJ, Geoghegan JA, Ganesh VK, Höök M. Adhesion, invasion and evasion: The many functions of the surface proteins of staphylococcus aureus Nature Reviews Microbiology 2013;12(1):49 – 62.

45. Jousselin A, Renzoni A, Andrey DO, Monod A, Lew DP, Kelley WL. The posttranslocational chaperone lipoprotein PrsA is involved in both glycopeptide and oxacillin resistance in Staphylococcus aureus. Antimicrob Agents Chemother. 2012 Jul;56(7):3629-40.

46. Jousselin A, Manzano C, Biette A, Reed P, Pinho MG, Rosato AE, Kelley WL, Renzoni A. The Staphylococcus aureus Chaperone PrsA Is a New Auxiliary Factor of Oxacillin Resistance Affecting Penicillin-Binding Protein 2A. Antimicrob Agents Chemother. 2015 Dec 28;60(3):1656-66.

47. Wiemels RE, Cech SM, Meyer NM, Burke CA, Weiss A, Parks AR, Shaw LN, Carroll RK. An Intracellular Peptidyl-Prolyl cis/trans Isomerase Is Required for Folding and Activity of the *Staphylococcus aureus* Secreted Virulence Factor Nuclease. J Bacteriol. 2016 Dec 13;199(1).

48. Sass P, Bierbaum G. Native graS mutation supports the susceptibility of *Staphylococcus aureus* strain SG511 to antimicrobial peptides. Int J Med Microbiol. 2009 Jun;299(5):313-22.

49. Yang SJ, Bayer AS, Mishra NN, Meehl M, Ledala N, Yeaman MR, Xiong YQ, Cheung AL. The *Staphylococcus aureus* two-component regulatory system, GraRS, senses and confers resistance to selected cationic antimicrobial peptides. Infect Immun. 2012 Jan;80(1):74-81.

50. Brötz H, Bierbaum G, Markus A, Molitor E, Sahl HG. Mode of action of the lantibiotic mersacidin: inhibition of peptidoglycan biosynthesis via a novel mechanism? Antimicrob Agents Chemother. 1995 Mar;39(3):714-9.

51. Plata KB, Rosato RR, Rosato AE. Fate of mutation rate depends on agr locus expression during oxacillin-mediated heterogeneous-homogeneous selection in methicillin-resistant *Staphylococcus aureus* clinical strains. Antimicrob Agents Chemother. 2011 Jul;55(7):3176-86.

52. Tuchscherr L, Loffler B. Staphylococcus aureus dynamically adapts global regulators and virulence factor expression in the course from acute to chronic infection Curr Genet 2016 Feb;62(1):15-7.

53. Wang R, Braughton KR, Kretschmer D, et al. Identification of novel cytolytic peptides as key virulence determinants for community-associated MRSA. Nat Med. 2007 Dec;13(12):1510-4.

54. Cheung GY, Kretschmer D, Duong AC, et al. Production of an attenuated phenol-soluble modulin variant unique to the MRSA clonal complex 30 increases severity of bloodstream infection. PLoS Pathog. 2014 Aug 21;10(8):e1004298.

55. Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiology. 2009 Jul;20(4):488-95.

56. Recker M, Laabei M, Toleman MS, et al. Clonal differences in *Staphylococcus aureus* bacteraemia-associated mortality. Nat Microbiol. 2017 Aug 7. doi: 10.1038/s41564-017-0001-x.

57. Lower SK, Lamlertthon S, Casillas-Ituarte NN, Lins RD, Yongsunthon R, Taylor ES, DiBartola AC, Edmonson C, McIntyre LM, Reller LB, et al. Polymorphisms in fibronectin binding protein A of *Staphylococcus aureu*s are associated with infection of cardiovascular devices Proc Natl Acad Sci U S A 2011 Nov 8;108(45):18372-7.

58. Hos NJ, Rieg S, Kern WV, Jonas D, Fowler VG, Higgins PG, Seifert H, Kaasch AJ. Amino acid alterations in fibronectin binding protein A (FnBPA) and bacterial genotype are associated with cardiac device related infection in *Staphylococcus aureus* bacteraemia. J Infect. 2015 Feb;70(2):153-9.

59. Marbach H, Boakes E, Lynham S, Ward M, Otter JA, Edgeworth JD. Identification of a distinctive phenotype for endocarditis-associated clonal complex 22 MRSA isolates with reduced vancomycin susceptibility. J Med Microbiol. 2017 May;66(5):584-591.

60. Golubchik T, Batty EM, Miller RR, Farr H, Young BC, Larner-Svensson H, Fung R, Godwin H, Knox K, Votintseva A, et al. Within- host evolution of staphylococcus aureus during asymptomatic carriage. PloS One 2013;8(5):e61319.

61. Young BC, Golubchik T, Batty EM, Fung R, Larner-Svensson H, Votintseva AA, Miller RR, Godwin H, Knox K, Everitt RG, et al. Evolutionary dynamics of staphylococcus aureus during progression from carriage to disease. Proc Natl Acad Sci U S A 2012;109(12):4550.

Chapter 6

6 Genome-wide association study reveals the bacterial determinants of pyomyositis caused by *Staphylococcus aureus*

6.1 Pyomyositis

Microbial genome sequencing and the development of methods for bacterial genome-wide association studies present new opportunities to discover the bacterial genetic basis of serious infection in major pathogens.[1,2,3,4,5,6] Pyomyositis is a severe infection of skeletal muscle, most commonly seen in children in the tropics, associated with malnutrition, trauma and immunodeficiency.[7,8,9] Pyomyositis is caused by a single bacterial pathogen, *Staphylococcus aureus*, in up to 90% of cases.[7,8,9,10] Such predominance, along with the doubling of incidence in the USA following the emergence of community associated methicillin resistant *S. aureus* (CA-MRSA),[11] suggests that lineages may vary in their propensity to cause pyomyositis. Molecular and genetic investigation of *S. aureus* strains found in pyomyositis has been limited, in part due to resource constraint in regions where pyomyositis is most prevalent.[12]

One hypothesised but disputed bacterial determinant is the *S. aureus* toxin Panton-Valentine leukocidin (PVL). PVL is a pore-forming toxin of myeloid cells, associated with increased inflammation and severity in bone infections.[13] PVL has been reported as a cause of severe *S. aureus* pneumonia,[14] but this association is

contested, as other mutations with demonstrable effects on bacterial toxin expression are present in the same strains.[15] Small case series testing for candidate virulence genes report a high prevalence of PVL in *S. aureus* isolated from pyomyositis.[11,16,17] However these series only assayed for 4-20 specific toxin genes (usually by sequencing gene fragments), and lack control strains to account for local bacterial population structure. A recent meta-analysis found no evidence for an increased rate of musculoskeletal infection or other invasive disease from PVL positive strains compared to controls.[18] Single gene associations have previously been demonstrated between *S. aureus* and invasive disease,[19] but these associations failed to replicate when strain background was included in the analysis.[20] The role of PVL is therefore controversial, with opinion divided as to whether this toxin is an important virulence factor, or merely a strain marker, linked to unidentified genetic determinants.[21,22]

Whole genome sequencing, and advances in bacterial genome-wide association studies (GWAS) offer a means to resolve these uncertainties and investigate the relationship between bacterial genomes and disease. However bacterial GWAS face specific challenges: complex bacterial genomic variation – including the movement of mobile genetic elements and large accessory genomes – cannot be captured by studying single nucleotide polymorphisms, reducing the applicability of well-established human GWAS methods, while strong population structure, arising from clonal reproduction, creates confounding effects.[3,5,6] Linear mixed models (LMM), the current state-of-the-art in controlling for population structure, have been adapted to additionally identify lineage-phenotype associations for the purposes of bacterial studies, a modification that counterbalances the loss in power attributable to strong population structure in bacteria.[3] Applying these tools to large data sets has demonstrated the feasibility of bacterial GWAS by successfully identifying genetic determinants of antimicrobial resistance and host adaptation.[1,2,3,4] No

bacterial studies have yet shown a direct link between bacterial genetic variants and infections in humans, but in the long run these tools open the possibility of dissecting less well understood phenotypes like bacterial virulence.[6]

To investigate the bacterial genetic basis of this infection, we conducted a genome-wide association study of 101 bacterial isolates cultured from pyomyositis in children attending the Angkor Hospital for Children, in Siem Reap, Cambodia, and 417 isolates cultured from nasal carriage in children without pyomyositis attending the same hospital.

6.2 Materials and Methods

6.2.1 Pyomyositis strains

Pyomyositis cases were identified from the Angkor Hospital for Children between January 2007 and November 2011. Attendances in children (aged ≤ 16 years) were screened using clinical coding (ICD-10 code M60 (myositis)) and isolation of *S. aureus* from skeletal muscle abscess pus. Clinical notes review was used to confirm cases, and bacterial strains cultured by the routine clinical microbiology laboratory were retrieved from the local microbiology biobank. 106 clinical episodes of pyomyositis were identified, in 101 individuals, and we included the earliest episode from each individual.

6.2.2 Carriage strains

S. aureus nasal colonisation isolates were collected from two cohort studies undertaken at Angkor Hospital for Children in Siem Reap, Cambodia. The first were selected from a collection characterising nasal colonization in the period between September and October, 2008, which has previously been described using MLST.[23] The swabs had been saved at -80°C since the study, these samples were re-examined for the presence of *S. aureus* using selective agar and CLSI guidelines for

susceptibility testing.

A second cohort study of *S. aureus* nasal carriage was undertaken between the 2nd-7th of July 2012. Eligible subjects were children (≤16 years) attending as an outpatient at Angkor Hospital for Children with informed consent, there were no exclusion criteria. Nasal swabs were performed on participants; using one sterile cotton-tipped swab (pre-moistened with PBS) for each participant. The swab was placed in each nostril in turn, and 3 full rotations of the swab were made in the anterior nares before the ends were broken into bottles containing sterile PBS and transported to the laboratory. The swabs were plated onto selective Mannitol Salt agar and cultured at 37°C. The CLSI standards were followed for susceptibility testing and bacteria stored in TSB Glycerol at -80°C.

Controls were randomly selected from carriers in these two cohorts using the excel randomization function: 222 of 519 from the 2008 cohort and 195 of 261 from the 2012 cohort. Cases and controls are summarised in Table 5.1

	Pyomyositis (2007-2012)	Nasal carriage (2008)	Nasal carriage (2012)
Number of isolates	101	222	195
Age (med, IQR)	7.8 years	5.9 years	6.3 years
	(4.2-11.8)	(2.5- 8.3)	(4.1- 9.9)
Male (n (%))	66/97 (68%)	122/221 (55.2%)	105/195 (53.8%)
MRSA (n (%))	0 (0%)	61 (27.5%)	52 (26.7%)

Table 6.1: Isolates included in this study.

6.2.3 Whole genome sequencing

For each bacterial culture, a single colony was sub-cultured and DNA was extracted as described in Methods 3.1.3, then sequenced at the Wellcome Trust Centre for Human Genetics, Oxford on the Illumina (San Diego, California, USA) HiSeq 2500 platform, with paired-end reads 150 bp long.

We used Velvet[24] to assemble reads into contigs *de novo*. Velvet optimiser selected a median kmer length of 123bp for assembly (IQR 119-127). Assemblies were a median of 2.81 MB long (IQR 2.78-2.84), and these had a median of 2,792,000 bases in contigs over 1000bp (IQR 2,770,000-2,826,000).

Multilocus sequence types were obtained using BLAST[25] of *de novo* assemblies, as described in Methods 3.2.4. Where multiple MLSTs were identified which differed by only a single-locus variant, or if ST could not be determined *in silico* but 6 of 7 loci were found to be identical to a central strain, these are grouped as a Clonal Complex (CC).[26] Antibiotic sensitivity was predicted as described in Methods 3.2.4

We used Stampy[27] to map reads, applying filters as described in Methods 3.2.2. The reference selected was USA300_FPR3757 (Genbank accession number CP000255.1[28]).

6.2.4 Genome-wide association study

Phylogeny was constructed with RAxML[29] and ClonalFrameML[30] as described in Methods 3.3.1. Sample heritability, kmer-based GWAS were performed using GEMMA[35], bacterialGWAS[3] and *bugwas*[3] methods as described in Methods 3.5. Kmer counting with dsk[31] identified a median of 2,801,000 kmers per isolate (IQR 2,778,000-2.837,000).

Associations between kmers and PCs were considered significant for *p*-values below 0.05, adjusted for multiple testing by Bonferroni method.[32] We identified 518 PCs and 236,627 sets of kmers with unique patterns of presence or absence in the phylogeny ("phylopatterns"). These led to thresholds of 9.7×10^{-5} and 2.1×10^{-7} respectively for genome wide significance.

The genetic content of kmers was determined by aligning kmers to well-annotated closed reference sequences using BLAST[25] as well as mapping kmers to a reference

strain using the read-mapping software Bowtie 2.[33] We used a draft assembly for kmer mapping. Areas of homology to well-annotated reference were identified by aligning sequences with Mauve.[34]

6.2.5 Toxin quantification

Production of PVL toxin was quantified using a sandwich ELISA, as described in Methods 3.1.4. The limit of detection was 0.014 ng/ml. All isolates were tested in triplicate and the mean value taken.

6.3 Results

6.3.1 Pyomyositis prevalence varies strongly by bacterial lineage

Phylogenetic analysis revealed striking differences in the observation of pyomyositis cases within strains (Figure 6.1). In this study, carriage genomes encompass multiple global lineages of *S. aureus*, including clonal complex (CC)-1, CC-5, CC-15, CC-45 and CC-121. Strain composition of carried *S. aureus* was stable over time, with no major lineages arising or disappearing between 2008 and 2012. In marked contrast to the range of lineages represented in carriage, 86/101 (85%) of pyomyositis genomes were found in CC-121, and none were found in the next two most commonly carried strains (ST-834 and CC-45).

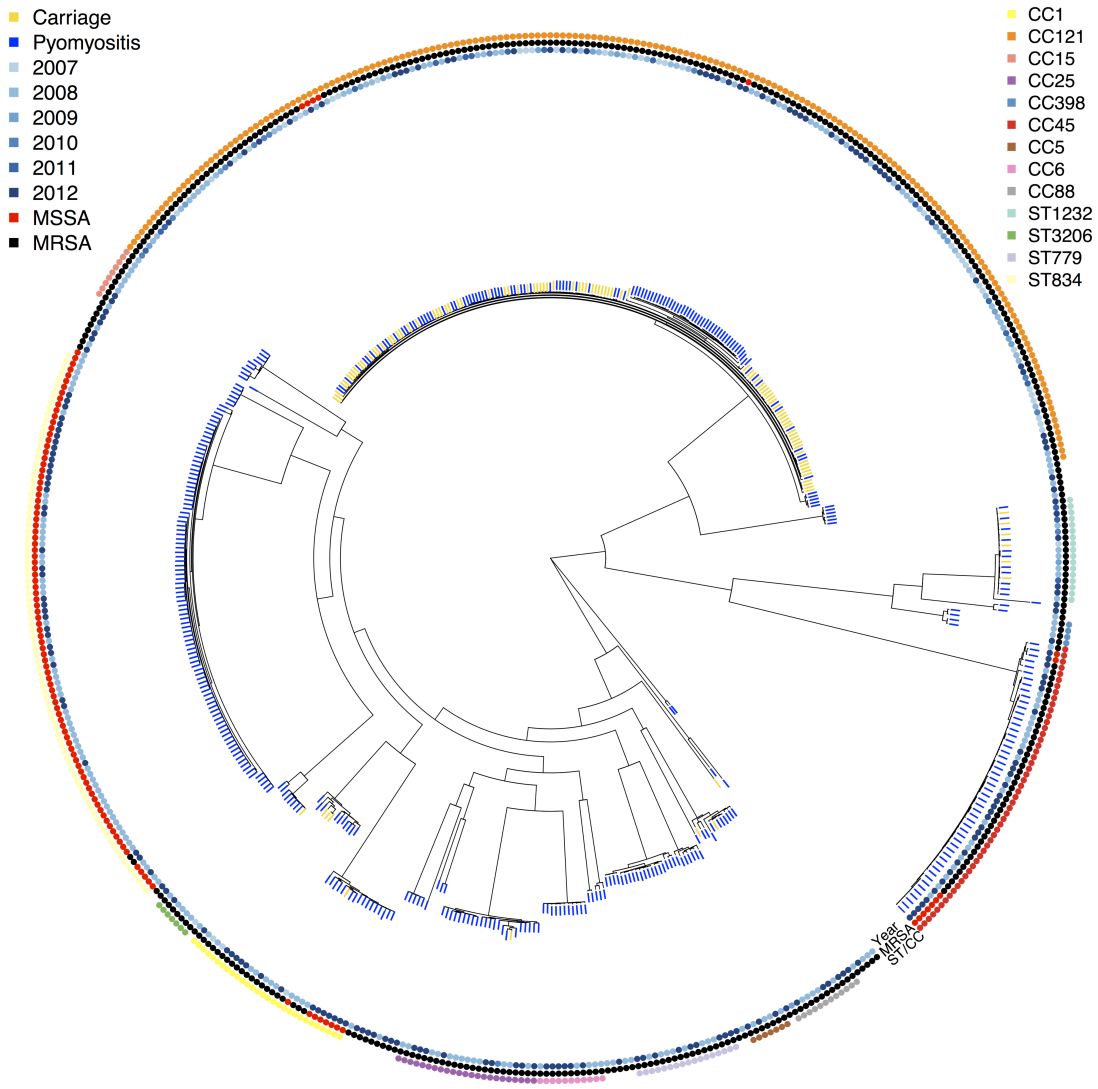


Figure 6.1: Maximum likelihood phylogenetic tree of *S. aureus* from children in Cambodia shows global diversity of strains, stable over the study period. Phylogeny was constructed from the entire genome sequences after mapping to a reference sequence (USA300_FPR3757). Tips of tree are marked according to whether isolate was found in carriage or pyomyositis. The innermost ring indicates year of isolation (lightest blue is earliest). The middle ring indicates whether isolate was phenotypically methicillin resistant. Outmost ring indicates ST/CC determined *in silico*.

A formal test using *bugwas*[3] found strong evidence for differences in incidence of pyomyositis between strains, defined in terms of the principal components (PCs) of bacterial genetic variation. PC 1, which distinguished CC-121 from ST-834, showed the strongest association ($p=10^{-29.6}$, Wald test), followed by PC20, which differentiated a subclade within CC-121 where no cases were seen ($p=10^{-13.9}$), and PC 2, which distinguished CC-45 ($p=10^{-4.9}$) (Fig 6.2). The heritability of case/control

status in the sample was estimated by a linear mixed model implemented in GEMMA[35] to be 63.8% (95% CI 49.2-78.4%), indicating a strong predictive relationship between bacterial genome and case/control status.

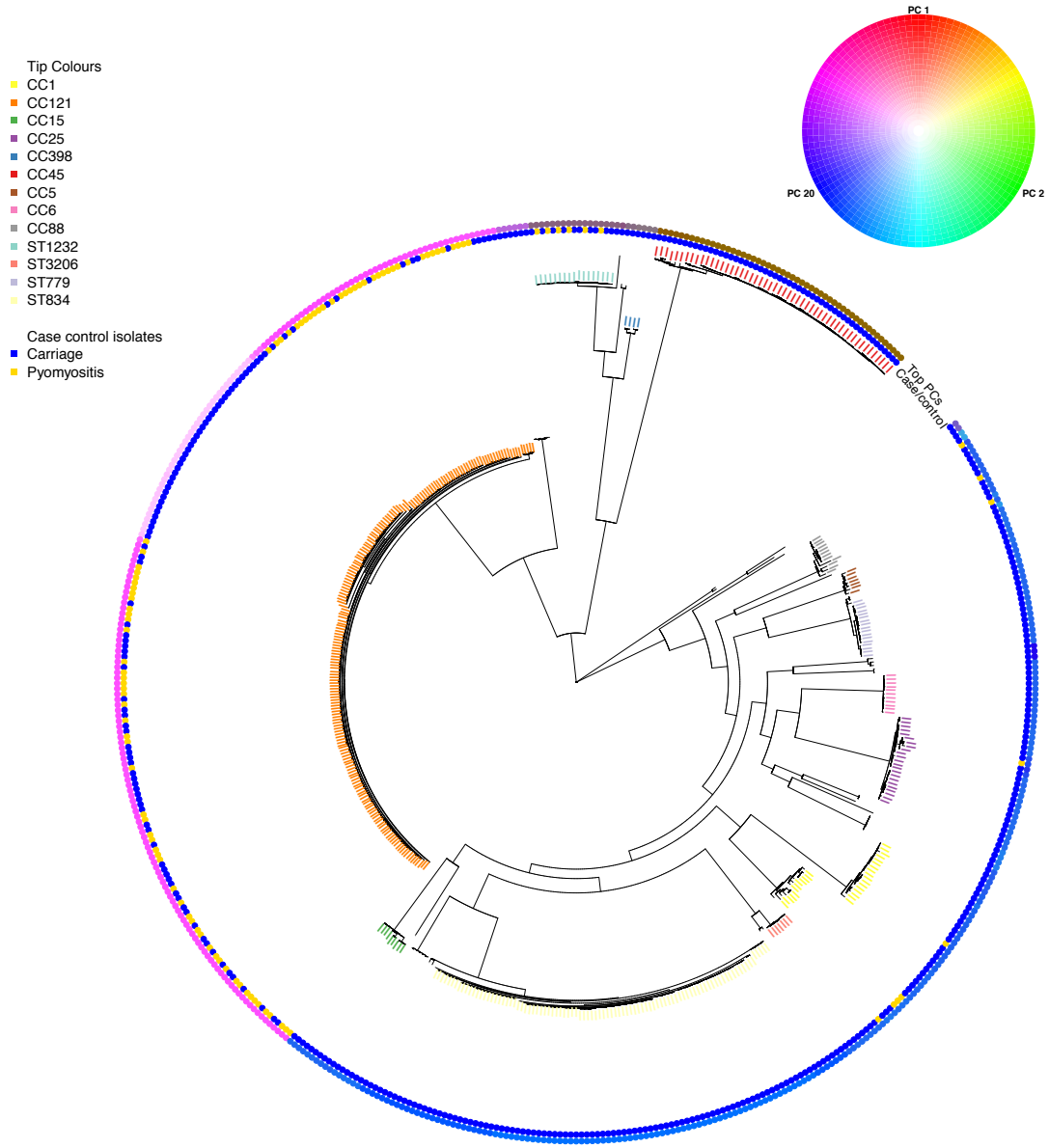


Figure 6.2: Maximum likelihood phylogenetic tree of *S. aureus* from children in Cambodia shows strong strain-to-strain variation in pyomyositis rates. Phylogeny was constructed from the entire genome sequences after mapping to a reference sequence (USA300_FPR3757). Tip colours correspond to the multi-locus sequence type (ST) or clonal complex (CC). The inner ring indicates case/control status: pyomyositis (gold) or nasal carriage (blue). The outer ring depicts the projection of each bacterial genome on to the three principal components (PCs) most significantly associated with case/control status. The projection of each genome on to PCs 1, 2 and 20 are mapped on to the red, green and blue colour channels respectively.

6.3.2 Kmers associated with disease reveal strong locus association with pyomyositis

To fully capture both variability in the accessory genome and non-SNP variation, we used a kmer-based approach[1] in which every unique 31 base pair DNA sequence (i.e. each 31mer) in the *de novo* assembly of each bacterial genome was identified, and tested for association with pyomyositis case/control status, again using LMM to control for population structure. We found 10,744,013 unique kmers, whose presence or absence across bacterial genomes occurred in 236,627 distinct phylopatterns. In total, 9175 kmers comprising 432 distinct phylopatterns were significantly associated with case/control status after correction for multiple testing (p=$10^{-6.8}$-10^{-21}) (Figure 6.3A). Of these, 9173/9175 (99.98%) were found more frequently in pyomyositis. Kmers found more frequently in disease showed odds ratios (OR) ranging from the modest (OR=2.7, p=$10^{-6.8}$) to the dramatic (OR=139.8, p=10^{-21}).

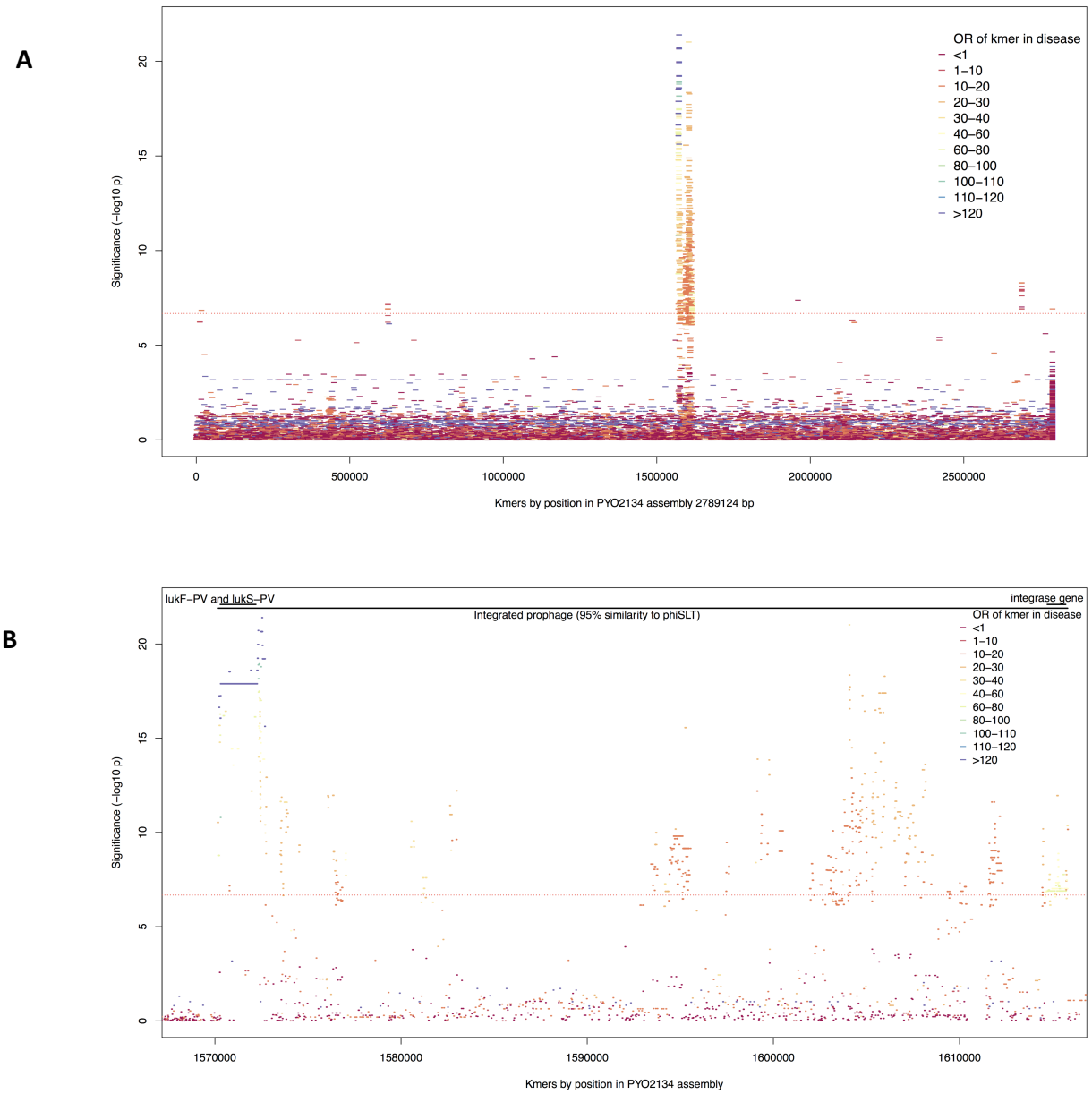


Figure 6.3: Kmers associated with pyomyositis and carriage. (A) All significant kmers and a random selection of 100,000 non-significant kmers were mapped to the assembly of one ST-121 pyomyositis isolate (PYO2134) using bowtie. Each horizontal segment represents a kmer, plotted by the mapped location and the significance of the association between that kmer and disease (negative $\log_{10}$ *P* value). Kmers are coloured by the odds ratio (OR) of that kmer and disease in the study. A Bonferroni-adjusted threshold for significance is plotted in red. **(B)** The region between 1.57 MB and 1.62 MB on the reference assembly is shown in greater detail. Annotations determined by similarity to reference sequences on BLAST and alignment of the reference to USA300 using Mauve. Bonferroni-adjusted threshold for significance is plotted in red.

The vast majority of significant kmers (8993/9176, 98.0%) localized to one 45.7kb region when mapped to the draft genome assembly of one ST-121 pyomyositis isolate, PYO2134 (Figure 6.3B). The kmers with highest OR of being found in pyomyositis mapped to the genes *lukF-PV* and *lukS-PV* and the region immediately upstream of the *lukS-PV* coding sequence. A total of 1,630 kmers representing variants in complete LD and covering the entire coding sequence of these genes were highly significantly associated with disease, being found in 98/101 (97%) of pyomyositis isolates and 84/417 (20%) of carriage isolates, OR=130 (p=10^{-18}). Kmers mapping to the 1902bp region immediately upstream of *lukSF-PV* were still more strongly associated with disease: kmers mapping to this region were observed in all *lukSF-PV*-carrying disease isolates (98/101, 97.0%), and variably present in a smaller number of carriage isolates (79-83/417, 18.9-19.9%), OR 132-140 (p=10^{-19} - 10^{-21}) (Figure 6.4)

The highest OR was for kmers 244bp upstream of the translation start site. A conditional analysis in GEMMA including the presence or absence of *lukSF-PV* as a covariate found the remaining estimated heritability for case/control status to be only 1.0% (95% CI 0-4.4%), and when the presence of both *lukSF-PV* and high risk kmers upstream of *lukSF-PV* were included together, this estimate of heritability dropped further to 8.3×10^{-7}. The high estimated heritability could therefore be fully explained by the presence of the *lukSF-PV* genes and upstream (non-coding) sequence.

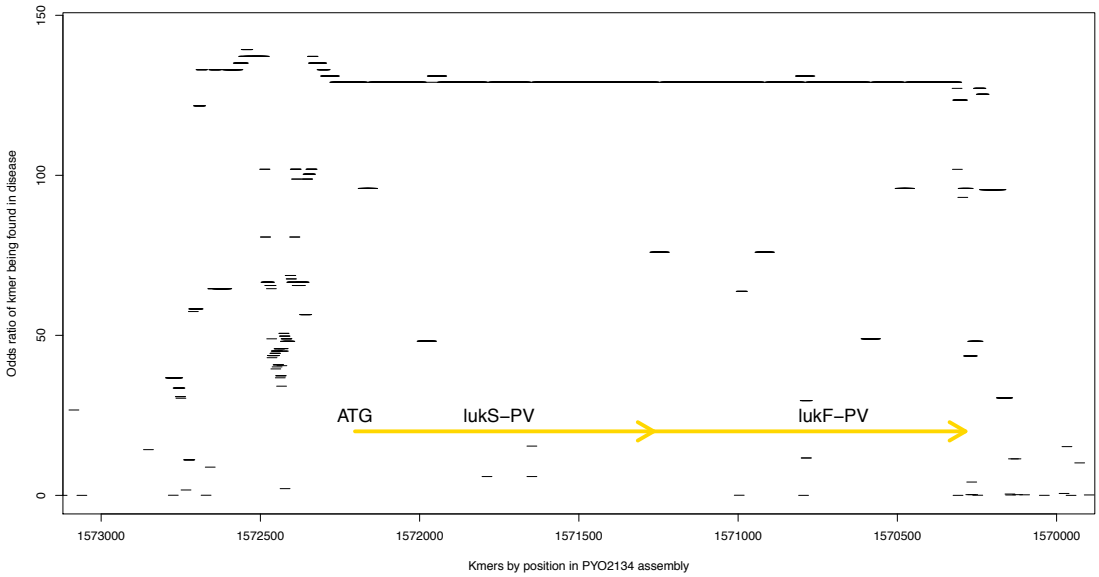


Figure 6.4: The odds ratio of association with disease for kmers that map to the PVL locus and adjacent region. The coding sequences *(lukS-PV* and *lukF-PV*) are marked by gold arrows. The *lukS-PV* translation start site is position 1572201 in the draft assembly.

The *lukSF-PV* genes were found in nearly all cases, indicating they are all but necessary for disease. However, they do not have complete penetrance, as they were found in carriage as well as disease, indicating they are not sufficient for disease. While host factors are likely to play a role, we find evidence suggesting bacterial genetic basis for some incomplete penetrance in strains where *lukSF-PV* is prevalent. The *lukSF-PV* genes were found in all ST-1232 and ST-3206 strains, and found variably present in the major pyomyositis-causing clade CC121 (Figure 6.5). The *lukSF-PV* upstream region showed variation in carriage strains from CC-121, ST-1232 and ST-3206: this variation, identified by the absence of kmers mapping upstream of PVL, is seen arising independently in multiple lineages (Figure 6.5). The presence of these kmers in all PVL positive strains cultured from pyomyositis suggests that conservation of this region is important to the role of PVL in pyomyositis. This region is immediately upstream of PVL transcription, and may play a role in PVL transcriptional regulation.

The strong association with *lukSF-PV* not only explains strong strain-to-strain variability in pyomyositis, it also explains the majority of cases that occur low-incidence strains. These genes occur only sporadically in CC-1, CC-88 and CC-25, but when they do occur in these low incidence strains, they are found in 5 of 7 cases, and 1 of 57 controls (Table 6.2).

Lineage	Isolates (*n*=518; 101 cases, 417 controls)		lukSF-PV kmers found (%)		lukSF-PV upstream region kmers found (%)	
	Cases	Controls	Cases	Controls	Cases	Controls
CC-121	86	109	98.8	59.6	98.8	57.8
ST-1232	6	11	100	100	100	72.7
ST-3206	1	6	100	100	100	50
CC-1	4	25	75	4	75	4
CC-25	1	22	100	0	100	0
CC-88	2	10	50	0	50	0
Unnamed ST	1	0	100	NA	100	NA

Table 6.2: Frequency of pyomyositis, kmers mapping to lukSF-PV and kmers mapping to the upstream non-coding region in all pyomyositis causing lineages in the study population

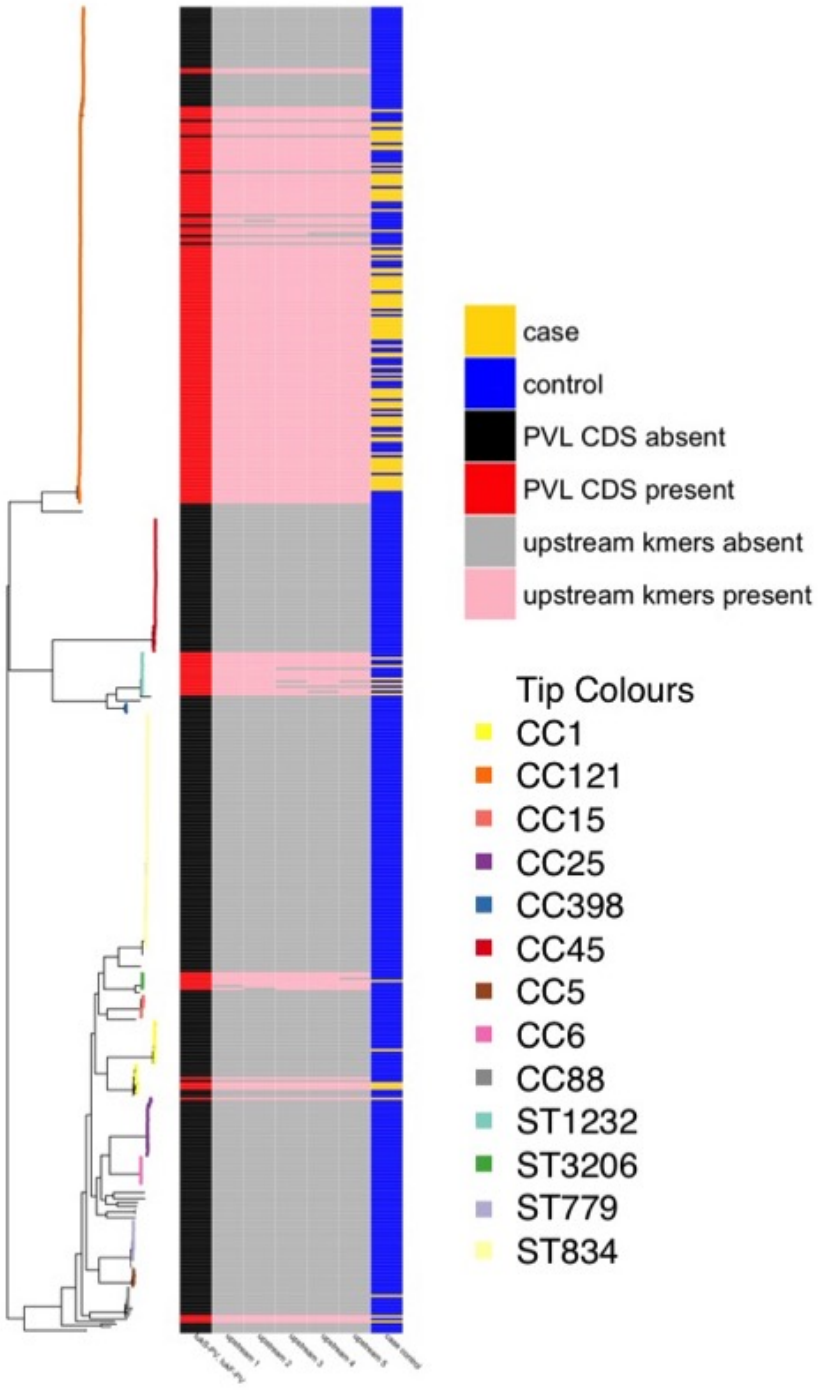


Figure 6.5: The correspondence of disease-associated kmers, phylogeny and disease. Maximum likelihood phylogeny is annotated with tip colours representing Multi-locus sequence type (ST) or Clonal Cluster (CC). The left most-column indicates whether kmers mapping to the PVL locus were present (red) or absent (black). Columns 2-6 each represent one of five distinct patterns of kmers mapping to the upstream region (in order of increasing OR), and indicate whether kmers in that pattern were present (pink) or absent (grey). The right-hand column indicates whether each genome was recovered from an individual with pyomyositis (gold) or nasal carriage only (blue).

6.3.3 Changes in PVL expression are associated with high risk kmers.

We tested the impact of the highest risk kmers on PVL production by quantifying LukS-PV expression using quantitative ELISA.

We tested the impact of the genotype on PVL expression by quantifying PVL secretion in isolates from pyomyositis (*n*=32) and carriage (*n*=18). We tested isolates without the PVL locus (*n*=7; 4 from carriage, 3 from pyomyositis) and isolates with kmers which showed the strongest association with disease (*n*=34; 29 from pyomyositis, 5 from carriage). PVL positive and negative isolates were randomly selected from lineages in which pyomyositis was observed. We also tested all isolates found in carriage which lacked at least one of these kmer patterns, but in which the PVL coding sequence was identified (*n*=9). Production of PVL was tested using ELISA and quantified by comparison with known concentrations of recombinant PVL component protein (LukS-PV).

As expected, strains without the PVL locus had lower levels of LukS-PV detected than strains with PVL genes present (Figure 6.6). Notably, some isolates from strains with low frequency of pyomyositis had low levels of PVL expression even when the locus was present (CC1 and CC25).

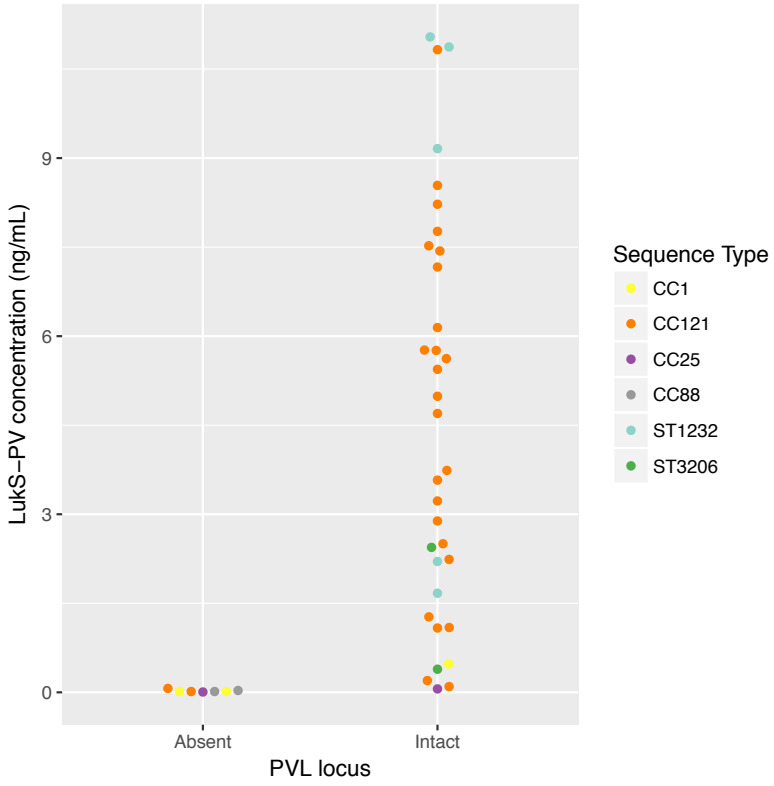


Figure 6.6: Expression of PVL from isolates with and without PVL coding sequence. The mean concentration of LukS-PV is plotted on the Y-axis. Isolates are grouped by presence of absence of PVL coding sequence on whole genome sequencing. Points are coloured by lineage.

Strains with variation in the high OR kmers upstream of PVL were found in 3 lineages: CC-121, ST-1232 and ST-3206. We compared isolates in these lineages with and without alterations to the upstream kmers for evidence of altered PVL secretion accompanying these changes (Figure 6.7). Strains with alteration to the upstream kmers had overall lower toxin expression (median 0.93ng/ml, IQR -0.24 to 3.1), than that seen in isolates with an intact upstream region (median 4.8ng/ml, IQR 2.3-7.65).

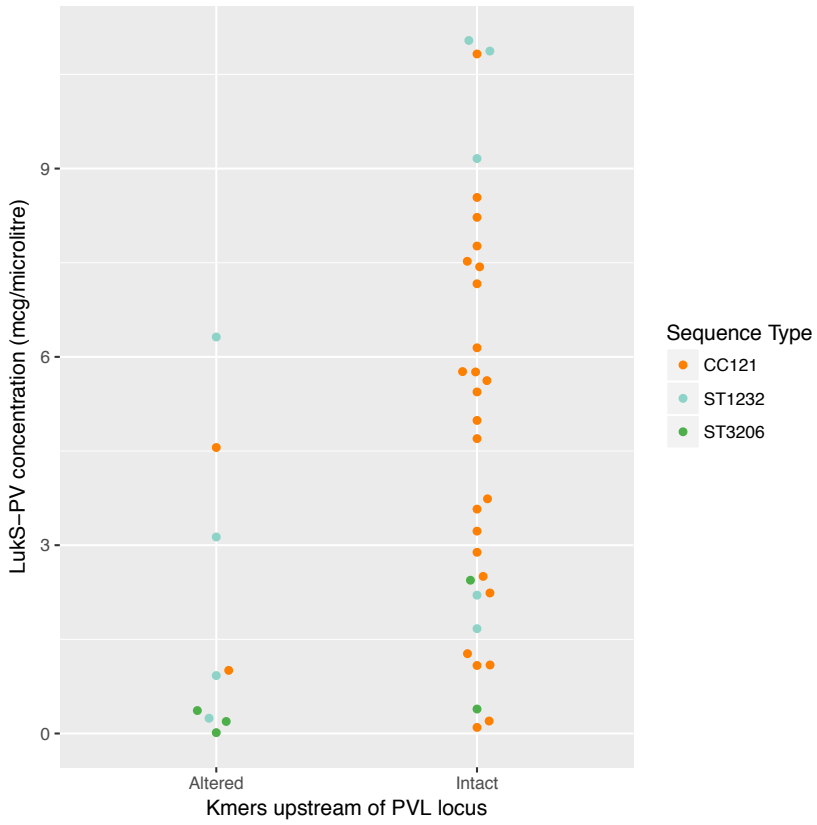


Figure 6.7: Expression of PVL from isolates with and without alteration to kmers upstream of PVL CDS. The mean concentration of LukS-PV is plotted on the Y-axis. Isolates are grouped by presence of absence of high OR kmers upstream of PVL CDS on whole genome sequencing. Points are coloured by lineage.

PVL expression appeared to vary with both upstream kmer change and with strain type, being higher in ST-1232 and CC-121 than ST-3206 (Figure 6.6, Figure 6.7). We modelled log-transformed LukS concentrations as a function of the presence of the upstream kmers and the CC. Both the variation of upstream kmers and strain type were significantly associated with PVL altered expression (p=0.02 and p=0.003 respectively, χ^2). Independent of strain, upstream kmer presence was associated with a 3.4-fold (95% CI 1.17-9.6) increase in LukS supernatant concentration relative to the absence of such kmers (p=0.025, Welch's t-test).

6.4 Discussion

In this chapter I present a bacterial GWAS that uncovers powerful evidence of association between a highly distinctive infection in children and a bacterial genetic determinant. A single coding region and the adjacent regulatory sequence are all but necessary for the development of pyomyositis: sporadic presence of this gene is associated with pyomyositis in otherwise low-frequency strains, and its absence is associated with carriage in a high propensity strain. This effect is not confined to altering the propensity for muscle infection within a single lineage, but is seen across multiple distinct *S. aureus* lineages.

The *S. aureus* toxin PVL is cytotoxic to myeloid cells, which form a first line of defence against bacterial infection.[20,36] PVL-producing *S. aureus* strains have increased duration of bacteraemia in a rabbit model of sepsis,[36] and the toxin has been found strongly bound to necrotic muscle in myositis associated with necrotizing fasciitis,[37] raising the hypothesis PVL may have both an enhanced capacity to seed to muscles via the bloodstream, and tropism for muscular infection. While PVL has long been thought an important *S. aureus* virulence factor,[13,38,39] its role in invasive disease has been controversial, and case control studies show heterogeneous results.[18,21,22] We demonstrate that the association between PVL and pyomyositis includes a variably present upstream non-coding region, the absence of which is associated with significant reduction in PVL expression. This may account for conflicting results in studies that only assay gene fragments (e.g. by PCR).

Our findings may explain trends in pyomyositis incidence: an increasing incidence in the USA following the emergence of CA-MRSA USA300 strains has been noted,[11] but the role of PVL could not be established because it presence was almost completely confounded by the presence of antibiotic (methicillin) resistance and

other strain (CC8)-linked virulence factors. Here we are able to separate the correlation with PVL from that of antibiotic resistance because of the independent occurrence of PVL in distinct strains, finding strong evidence for the role of PVL across strains, but in particular in CC-121, a largely methicillin susceptible lineage unrelated to CC-8. While CA-MRSA carriage is widespread in this population (most commonly with ST-834 lineage, but also with CC-121 strains, (Figure 6.1)),[23] we find no cases of pyomyositis due to MRSA strains. Further, as the incidence of PVL varies geographically, these findings may also help illuminate geographic differences in the incidence of pyomyositis.[7,8],[18]

These results have important implications for pyomyositis treatment and prevention. They establish staphylococcal pyomyositis as a disease whose causation depends critically on expression of a single toxin. This property is shared by toxin-driven, vaccine-preventable diseases such as tetanus and diphtheria. Therefore, the generation of neutralizing anti-toxin antibodies against PVL by vaccination, which is feasible,[40] may protect human populations against this tropical disease

Further, although beta-lactam antibiotics commonly used to treat *S. aureus* increase PVL production, antibiotics which affect bacterial protein synthesis (clindamycin and linezolid) decrease PVL activity and are recommended in severe PVL associated infection.[41,42,43] This advice is largely based on expert opinion[42,43] and the extrapolation of *in vitro* observations.[41] The rare and sporadic nature of necrotizing staphylococcal pneumonia and necrotizing fasciitis are a significant challenge to rigorous testing of this advice through a randomised clinical trial. Given the demonstrated importance of PVL expression in pyomyositis, this condition would be an ideal focus for a trial of the utility of protein synthesis inhibiting antibiotics in PVL-associated *S. aureus* disease.

This study demonstrates that microbial genomics are a powerful tool to explore the pathogenesis of human infections. We show that infections can be attributed to specific genes through genome wide association study of highly distinctive phenotypes.

References chapter 6

1. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, Bentley SD, Maiden MC, Parkhill J, Falush D. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in campylobacter Proc Natl Acad Sci U S A. 2013 Jul 16;110(29):11923-7.

2. Chewapreecha C, Marttinen P, Croucher NJ, Salter SJ, Harris SR, Mather AE, Hanage WP, Goldblatt D, Nosten FH, Turner C, et al. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes PLoS Genet. 2014 Aug 7;10(8):e1004547.

3. Earle SG, Wu CH, Charlesworth J, Stoesser N, Gordon NC, Walker TM, Spencer CC, Iqbal Z, Clifton DA, Hopkins KL, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies Nat Microbiol. 2016 Apr 4;1:16041.

4. Lees JA, Vehkala M, Valimaki N, Harris SR, Chewapreecha C, Croucher NJ, Marttinen P, Davies MR, Steer AC, Tong SY, et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes Nat Commun. 2016 Sep 16;7:12797.

5. Falush D. Bacterial genomics: Microbial GWAS coming of age Nature Microbiol. 2016;1(5):16059.

6. Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: Lessons from human GWAS Nat Rev Genet. 2017 Jan;18(1):41-50.

7. Chauhan S, Jain S, Varma S, Chauhan SS. Tropical pyomyositis (myositis tropicans): Current perspective Postgrad Med J. 2004 May;80(943):267-70.

8. Verma S. Pyomyositis in children Curr Infect Dis Rep. 2016 Mar;18(4):12,016-0520-2.

9. Bickels J, Ben-Sira L, Kessler A, Wientroub S. Primary pyomyositis J Bone Joint Surg Am. 2002 Dec;84-A(12):2277-86.

10. Moriarty P, Leung C, Walsh M, Nourse C. Increasing pyomyositis presentations among children in queensland, australia Pediatr Infect Dis. J 2015 Jan;34(1):1-4.

11. Pannaraj PS, Hulten KG, Gonzalez BE, Mason EO,Jr, Kaplan SL. Infective pyomyositis and myositis in children in the era of community-acquired, methicillin-resistant *Staphylococcus aureus* infection Clin Infect Dis. 2006 Oct 15;43(8):953-60.

12. Borges AH, Faragher B, Lalloo DG. Pyomyositis in the upper negro river basin, Brazilian Amazonia Trans R Soc Trop Med Hyg. 2012 Sep;106(9):532-7.

13. Bocchini CE, Hulten KG, Mason EO,Jr, Gonzalez BE, Hammerman WA, Kaplan SL. Panton-Valentine leukocidin genes are associated with enhanced inflammatory response and local disease in acute hematogenous *Staphylococcus aureus* osteomyelitis in children Pediatrics. 2006 Feb;117(2):433-40.

14. Labandeira-Rey M, Couzon F, Boisset S, Brown EL, Bes M, Benito Y, Barbu EM, Vazquez V, Hook M, Etienne J, et al. *Staphylococcus aureus* Panton-Valentine leukocidin causes necrotizing pneumonia Science. 2007 Feb 23;315(5815):1130-3.

15. Villaruz AE, Bubeck Wardenburg J, Khan BA, Whitney AR, Sturdevant DE, Gardner DJ, DeLeo FR, Otto M. A point mutation in the agr locus rather than expression of the Panton-Valentine leukocidin caused previously reported phenotypes in *Staphylococcus aureus* pneumonia and gene regulation J Infect Dis. 2009 Sep 1;200(5):724-34.

16. Sina H, Ahoyo TA, Moussaoui W, Keller D, Bankole HS, Barogui Y, Stienstra Y, Kotchoni SO, Prevost G, Baba-Moussa L. Variability of antibiotic susceptibility and toxin production of *Staphylococcus aureus* strains isolated from skin, soft tissue, and bone related infections BMC Microbiol. 2013 Aug 8;13:188,2180-13-188.

17. Garcia C, Hallin M, Deplano A, Denis O, Sihuincha M, de Groot R, Gotuzzo E, Jacobs J. *Staphylococcus aureus* causing tropical pyomyositis, amazon basin, peru Emerg Infect Dis. 2013 Jan;19(1):123-5.

18. Shallcross LJ, Fragaszy E, Johnson AM, Hayward AC. The role of the Panton-Valentine leucocidin toxin in staphylococcal disease: A systematic review and meta-analysis Lancet Infect Dis. 2013 Jan;13(1):43-54.

19. Peacock SJ, Moore CE, Justice A, Kantzanou M, Story L, Mackie K, O'Neill G, Day NP. Virulent combinations of adhesin and toxin genes in natural populations of staphylococcus aureus Infect Immun. 2002 Sep;70(9):4987-96.

20. Lindsay JA, Moore CE, Day NP, Peacock SJ, Witney AA, Stabler RA, Husain SE, Butcher PD, Hinds J. Microarrays reveal that each of the ten dominant lineages of *Staphylococcus aureus* has a unique combination of surface-associated and regulatory genes J Bacteriol. 2006 Jan;188(2):669-76

21. Otto M. A MRSA-terious enemy among us: End of the PVL controversy? Nat Med. 2011 Feb;17(2):169-70.

22. Day NP. Panton-Valentine leucocidin and staphylococcal disease Lancet Infect Dis. 2013 Jan;13(1):5-6.

23. Nickerson EK, Wuthiekanun V, Kumar V, Amornchai P, Wongdeethai N, Chheng K, Chantratita N, Putchhat H, Thaipadungpanit J, Day NP, Peacock SJ. Emergence of community-associated methicillin-resistant *Staphylococcus aureus* carriage in children in Cambodia. Am J Trop Med Hyg. 2011 Feb;84(2):313-7

24. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de bruijn graphs Genome Res. 2008 May;18(5):821-9.

25. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool J Mol Biol. 1990 Oct 5;215(3):403-10.

26. Feil EJ, Cooper JE, Grundmann H, Robinson DA, Enright MC, Berendt T, Peacock SJ, Smith JM, Murphy M, Spratt BG, et al. How clonal is Staphylococcus *aureus*? J Bacteriol. 2003;185(11):3307-3316.

27. Lunter G, Goodson M. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads Genome Res. 2011 Jun;21(6):936-9.

28. Diep BA, Gill SR, Chang RF, Phan TH, Chen JH, Davidson MG, Lin F, Lin J, Carleton HA, Mongodin EF, et al. Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant *Staphylococcus aureus* Lancet. 2006 Mar 4;367(9512):731-9.

29. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of

large phylogenies Bioinformatics. 2014 May 1;30(9):1312-3.

30. Didelot X, Wilson DJ. ClonalFrameML: Efficient inference of recombination in whole bacterial genomes PLoS Comput Biol. 2015 Feb 12;11(2):e1004041.

31. Rizk G, Lavenier D, Chikhi R. DSK: K-mer counting with very low memory usage Bioinformatics. 2013 Mar 1;29(5):652-3.

32. OJ Dunn. Estimation of the medians for dependent variables. Ann. Math. Stat. 1959 30, 192–197.

33. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2 Nat Methods. 2012 Mar 4;9(4):357-9.

34. Darling AC, Mau B, Blattner FR, Perna NT. Mauve: Multiple alignment of conserved genomic sequence with rearrangements Genome Res. 2004 Jul;14(7):1394-403.

35. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies Nat Genet. 2012 Jun 17;44(7):821-4.

36. Diep BA, Palazzolo-Ballance AM, Tattevin P, Basuino L, Braughton KR, Whitney AR, Chen L, Kreiswirth BN, Otto M, DeLeo FR, et al. Contribution of Panton-Valentine leukocidin in community-associated methicillin-resistant *Staphylococcus aureus* pathogenesis PLoS One. 2008 Sep 12;3(9):e3198.

37. Lehman D, Tseng CW, Eells S, Miller LG, Fan X, Beenhouwer DO, Liu GY. *Staphylococcus aureus* Panton-Valentine leukocidin targets muscle tissues in a child with myositis and necrotizing fasciitis Clin Infect Dis. 2010 Jan 1;50(1):69-72.

38. Boakes E, Kearns AM, Ganner M, Perry C, Hill RL, Ellington MJ. Distinct bacteriophages encoding Panton-Valentine leukocidin (PVL) among international methicillin-resistant *Staphylococcus aureus* clones harboring PVL J Clin Microbiol. 2011 Feb;49(2):684-92.

39. Kurt K, Rasigade JP, Laurent F, Goering RV, Zemlickova H, Machova I, Struelens MJ, Zautner AE, Holtfreter S, Broker B, et al. Subpopulations of *Staphylococcus aureus* clonal complex 121 are associated with distinct clinical entities PLoS One. 2013;8(3):e58155.

40. Adhikari RP, Kort T, Shulenin S, Kanipakala T, Ganjbaksh N, Roghmann MC, Holtsberg FW, Aman MJ. Antibodies to *S. aureus* LukS-PV Attenuated Subunit Vaccine Neutralize a Broad Spectrum of Canonical and Non-Canonical Bicomponent Leukotoxin Pairs. PLoS One. 2015 Sep 14;10(9): e0137874.

41. Turner CE, Sriskandan S. Panton-Valentine leucocidin expression by *Staphylococcus aureus* exposed to common antibiotics J Infect 2015 Sep;71(3):338-46

42. Liu C, Bayer A, Cosgrove SE, Daum RS, Fridkin SK, Gorwitz RJ, Kaplan SL, Karchmer AW, Levine DP, Murray BE, et al. Clinical practice guidelines by the Infectious Diseases Society of America for the treatment of methicillin-resistant *Staphylococcus aureus* infections in adults and children: Executive summary Clin Infect Dis. 2011 Feb 1;52(3):285-92.

43. Guidance on the diagnosis and management of PVL-associated *Staphylococcal aureus* infections (PVL-SA) in England, 2nd Edition. 2008. Public Health England [online].

Chapter 7

7 Conclusions and future work

In this thesis I have presented three studies investigating the relationship between *S. aureus* genomic variation and invasive disease.

7.1 Within-host evolution of *S. aureus* is a promising avenue for understanding invasive disease

By studying the bacterial populations evolving in hosts with both *S. aureus* nasal carriage and invasive *S. aureus* infection, we found that extensive within-host variation occurs, and confirmed that invasive isolates represent distinct clades, which are usually derived from the carriage population. The extent of variation within infected hosts exceeded that previously found in stable carriage, and showed a relaxation of the purifying selection observed in carriage populations.

7.1.1 Summary of key findings

While the variants described in these populations were almost entirely unique, by comparing genes and groups of genes with multiple variants across individuals, we were able to systematically describe the variants that separated carriage and disease populations. Among these variants we found the transcriptional regulatory gene AgrA and genes controlled by virulence regulators Agr and Rsp were significantly enriched for protein altering variants. Additionally, genes with down-regulated expression in response to CAMPs were strongly enriched for variation. This result was particularly driven by changes in cell adhesion and pathogenesis

genes, particularly ClfA, ClfB, FnbA, Spa and Bbp.

The findings were specific to the variation seen within infected hosts, and differed from the signals of adaptation in stable carriage, across the species and among the recently derived variants from a collection of carried and invasive bacteria. However this enrichment was not specific to the site of infection. While enrichment in pathogenesis genes was specific to variants derived within infected populations, variation in transcriptional regulatory genes and their effector arms was observed both among variants derived in the infected site and within the carriage population. The within-host adaptation seen during infection was thus seen to alter the genes that regulate and effect shifts in bacterial behaviour from adhesion to toxin expression, as well as the adhesins under that regulatory control. This adaptation was not universal – the signal of adaptation derives from a minority of patients – but these results illustrate the adaptation occurring in some individuals.

7.1.2 Implications and Future work

The carriage and disease populations studied were contemporaneously rather than sequentially isolated and so causal conclusions cannot be drawn, but they provide several insights that can be carried forward to further our understanding of the mechanisms of pathogenesis of invasive *S. aureus* disease.

Firstly we see that natural populations are a rich resource for investigating the genetic basis of pathogenesis. The use of natural populations has the potential to introduce confounders, both through strain level differences and host factors. This design of taking paired isolates is a powerful control for these confounders, but generating an appropriate sampling frame is a challenge. Having demonstrated the power of the paired sample approach, we are beginning preparatory work to repeat this analysis in paired carriage and disease isolates from a cohort of nearly 300 patients sampled during a randomised, controlled trial in individuals with *S. aureus*

bacteraemia. Adding further power to this study, this RCT cohort has patient level data available, including focus of infection and host baseline morbidity (e.g. immunosuppression). This will allow us not only to perform a validation study of the findings reported here, but using the enhanced patient data will allow us to investigate whether within-host adaptation varies with the focus of bloodstream infection or host health.

Secondly, we can direct and interpret more traditional molecular microbiology using the findings of observational genomic studies. Initial work was performed on two Rsp mutants identified arising within hosts, including one identified in this study and one described during longitudinal sampling.[1] These strains demonstrated significant reductions in haemolysis and host cytotoxicity measured *in vitro* but no reduction in their survival in whole human blood.[2] A second paper on the transcriptional and phenotypic impact of laboratory constructed Rsp mutants was published concurrently with this work. These authors observed the same reduction in haemolysis and toxin production and abolished lethality in a mouse model and, from the alteration in phenotype, concluded "Rsp was essential for the development of bacteremia and skin infection".[3] This demonstrates the value of natural populations to contextualise *in vitro* observations.

In collaboration with Dr David Wyllie, we have begun characterisation of genotyped isolates from carriage and disease using two *in vitro* phenotypes: bacterially mediated toxicity of lymphoid cells, and complement inhibition. One limitation of the paired sample approach to study within-host adaptation is that important variation may well have occurred prior to sampling and be common to all isolates. Indeed, in our case study of within-host evolution of *S. aureus* that incorporated longitudinal sampling, an important variant preceded invasive disease by over a month.[1] To address this limitation, these phenotypic assays will incorporate a "population control" isolate, that is a *S. aureus* strain isolated from stable carriage

which is shown on whole genome sequencing to be closely related to the paired patient isolates. A future study of the impact of within-host mutation on transcriptional activity is also proposed. Through these methods we will investigate systematic changes in bacterial phenotype, defined by transcriptional activity, that distinguish carried from invasive bacteria.

Thirdly, and more speculatively, if these findings are validated, as well as providing insights for further understanding of the pathogenesis of *S. aureus* infection, they open a possibility of new, individualised diagnosis or even therapeutics. The pathogenesis of some *S. aureus* infections appears to result from toxin formation and direct tissue damage,[4] as well as over stimulation of the host immune response, leading to sepsis and multi-organ dysfunction.[5] Our data add to a body of work demonstrating that in some cases, *S. aureus* infection is associated with reduced toxin expression.[6,7,8] There may be multiple identifiable sub-types of *S. aureus* infection, in some of which an enhanced immune response – possibly by promoting immune mediated bacterial clearance – is beneficial rather than deleterious to the host.

This personalised approach to diagnosis and therapeutics has been recently further advanced by work from Recker and colleagues.[9] In a study of the bacterial determinants of mortality from SAB, they investigated two lineages of *S. aureus* CC-30 and CC-22, demonstrating heterogeneity within both lineages in biofilm formation, as well as strong differences both between and within lineages in toxicity *in vitro*. They further find that a combination of patient co-morbidity, bacterial genotype and *in vitro* phenotype can be used by machine learning to predict mortality associated with bacteraemia, finding that bacterial toxicity was predictive of mortality in CC-22 but not CC-30.[9]

Identifying sub-types of infection, termed endotypes,[10] based on bacterial phenotype is the microbiological complement to current strategies studying the human sepsis response. Recent studies of sepsis have focussed on the transcriptome of sepsis, finding evidence of regulatory loci implicated in sepsis,[11] and described gene expression signatures that distinguish 4 sepsis endotypes predictive of mortality.[12] We propose that bacterial endotypes may similarly have the potential to inform treatment strategies.

7.2 Two *S. aureus* virulence GWAS produce dramatically different results

In addition to studying in fine detail the most recent variants that characterise *S. aureus* disease, in chapters 5 and 6 I have presented two GWAS of population wide variants in *S. aureus* disease. These demonstrated contrasting findings from two distinct phenotypes of *S. aureus* infection.

7.2.1 Summary of key findings

GWAS of pyomyositis in a paediatric population from a tropical region revealed that pyomyositis is a disease which is almost entirely dependent upon the expression of a single toxin gene. The phenotype of pyomyositis had considerable sample heritability, and virtually the entire signal of heritability was localised to the PVL coding sequence and the non-coding region immediately upstream.

By applying GWAS methods we were able to identify non-coding variation that was associated with increased OR of disease, and to demonstrate this variation was associated with alteration in protein production *in vitro*. No PVL transcriptional promoter sites have been definitively identified, although two Pribnow boxes have been described in the 5' flanking region 143-148bp and 166-171bp from the initiation codon.[13] Our study provides evidence that sites with important functional effect on expression are likely to be located in the 244bp region upstream of *lukS-PV*. In this study, the promise of bacterial GWAS for virulence has delivered clarity

on a subject of previous uncertainty.

In contrast to pyomyositis, a large GWAS of *S. aureus* bacteraemia and carriage finds low heritability, limited lineage effects and no robust evidence of bacterial genetic content or variation associated with increased risk of disease. The strongest evidence of association was between the antibiotic resistance gene *mecA*, and a frequently co-carried trimethoprim resistance SNP. Sampling bias between hospital and community settings may explain these findings, though we could not exclude a *mecA*-mediated effect.

7.2.2 Implications and future work

PVL is phage-encoded gene, though this phage is frequently found integrated into the bacterial chromosome. Thus the causative gene for pyomyositis comes not from the bacterial genome, but from a bacterial parasite. Diverse important human diseases are likewise due to to the phage encoded toxins, including shiga-toxin associated dysentary[14], cholera[15], diphtheria[16] and botulism[17]. In a sense then, these bacterial diseases are viral in origin, though the expression of these viruses is bacterially mediated infections. Many exotoxins arise from bacterial phages, but these phages are widespread in nature, including in environments where the bacterial hosts associated with human disease are absent, suggesting these phages confer additional survival benefits to bacteria.[18] This observation has led to the hypothesis that bacterial virulence arises as an accidental effect of viral predation.[19]

In addition to this insight into the evolutionary origins of disease, the findings of this study have great practical importance. Critically, other toxin-mediated diseases, including diphtheria and tetanus, have been successfully controlled by vaccines that raise protective anti-toxin antibodies.[20] *S. aureus* vaccines have thus far failed to deliver protection in phase III studies where the end point has been bacteraemia.[21,22,23] Our data suggest that individuals at risk of pyomyositis

specifically may benefit from vaccination with PVL-specific antibody response. A vaccine which produces such antibodies has been developed.[24] A study in a setting where the incidence of pyomyositis is significant could be undertaken to investigate for a benefit to vaccination. This would likely need to be a tropical setting. Cambodia and other centres in South East Asia where the Mahidol Oxford Tropical Medicine Research Unit is active would be an appropriate setting. Data from other another MORU centre in Laos confirms PVL positive strains dominate skin and soft tissue infections in the region.[25]

While our GWAS of SAB yielded a largely negative result, these findings nevertheless have important implications. Firstly, they shed light on the many catalogued virulence factors in *S. aureus*. There are extensive, conflicting findings for and against a role for dozens of loci in *S. aureus* bacteraemia, and our large, well-controlled study is strong evidence that variation in the presence of these genes is unlikely to contribute to the development of bloodstream infection.

Poor control for population structure has likely played a role in false positive results for gene associations with SAB from case control studies. The existence of population structure must be recognised in the design of future *S. aureus* GWAS. We have shown genetic variants associated with HA carriage are found in cases categorised as community associated when admissions over the previous 12 weeks are taken into account. These findings suggest that HA *S. aureus* carriage can persist for many months following exposure, and that epidemiological classifications based on recent hospital exposure lead to significant misclassifications.

Although we found that *S. aureus* bacteraemia as a single entity has limited heritability, this finding may be due to dilution of true effects if multiple endotypes exist. SAB likely arises from a number of mechanisms: *S. aureus* entry to the bloodstream can be facilitated by vascular catheters or soft tissue infection as well

as occurring by cryptic means. SAB episodes occurring via these mechanisms likely vary in their pathogenesis and possibly vary in the bacterial determinants of disease. It remains to be determined if distinct endotypes of SAB can be identified. The existence of such distinct phenotypes is suggested by the result of studies of *S. aureus* cardiac device infection. Studies of candidate gene sequences found associations between variation in *fnbA* and the ability of *S. aureus* – after invading the bloodstream – to adhere to and infect implanted cardiac devices,[26,27] though the same variants were not associated with prosthetic joint infection.[28]

FnbA, like many proposed virulence determinants, contains repetitive sequence, which cannot be accurately and reliably accessed by short read sequencing. Thus the regions with some of the highest prior probability of affecting virulence can not be tested by the sequencing technology that is currently the best value for large scale GWAS. The use of longer read sequencing platforms (e.g. PacBio) would increase costs approximately 10x higher, making the sample sizes studied here extremely expensive. The accuracy of strand sequencing (e.g. Nanopore) remains a limiting factor for resolving SNP level differences. The combination of high accuracy short read sequencing reads with long reads (e.g. from Nanopore) to disentangle repetitive regions may offer a technical solution, though the costs would remain significant.[29]

Thus, there are two avenues for further study of the determinants of SAB, which may be undertaken in parallel. Firstly we can refine our bacteraemia phenotype. Characterising bacteraemia by its route of onset or focus of infection may allow us to detect bacterial genetic factors related to more precisely specified disease manifestations.

Secondly, we can attempt to refine our technical resolution of the cell wall proteins and other other repetitive regions of greatest interest by employing technologies

that resolve the sequence of these regions. Mapping based variant detection and variant detection from *de novo* assemblies both provided substantial increases in sensitivity when a closely related reference was employed, even when the "self" reference was an assembly based on short read sequencing (See Appendix 1). A judicious use of long-read sequencing to expand our repertoire of closed or nearly-closed reference sequences would improve our rates of calling variation in repeat regions. A drawback to this method would be the loss of inter-group comparability of specific variants.

7.3 Going beyond the bacterium

Our failure to identify bacterial genetic associations with SAB may be a true null result. While PVL is all but necessary for pyomyositis, these genes were originally viral rather than bacterial in origin. *S. aureus*, existing as it does as a common commensal may not be evolved to cause disease in humans at all, and infections – arising due to random events or host vulnerability – may be a threat to both host and pathogen. Alternately, if specific variants do increase the propensity of *S. aureus* to cause bacteraemia these may fail to be transmitted (e.g. because of host death), or even be selected against, and therefore fail to disseminate through the population to be found in the numbers needed to be identified by GWAS. Thus long-term evolutionary pressures may not shape the bacterial population to favour infection, and our results favour this hypothesis

While chance and non-genetic host factors almost certainly play a role in SAB, it is notable that human GWAS has found only limited evidence for host genetic heritability in bacterial infection. Human GWAS of susceptibility to bacterial infection has found limited evidence for a human genetic basis for infection, often localising to HLA loci, and in meningococcal disease to complement factors that confer protection from disease.[30,31] The complex possibility of interactions between

human and bacterial genomes warrants exploration, and the future of investigations into bacterial pathogenesis in humans may be joint human-bacterial GWAS. Large sample sizes and novel statistical methods will be needed to untangle these complex relationships.

7.4 Conclusion

The two approaches followed in this study – population GWAS and study of within-host evolution – have complementary strengths for understanding bacterial pathogenesis. Both control for bacterial population structure, incorporating strain relatedness in the model or matching closely related carried and invasive strains. Within-host evolution has the additional advantage of incorporating control for host factors, as the paired isolates are obtained from a single host.

GWAS are able to identify genes or variants in common across the population. We have seen differing success in two phenotypes, and the success of further GWAS will be critically dependent upon well-defined phenotypes. Within-host studies too are able to identify adaptation by pooling unique variants across genes and pathways. The paired sampling frame used in within-host study of *S. aureus* invasive disease is harder to replicate. An approximation of the within-host GSEA method, studying variants unique to carriage and invasive isolates from a population, might deliver insights if we can achieve fine enough sampling such that the variants that separate isolates are likely to have arisen within-host.

Like *S. aureus*, other major bacterial human pathogens are more commonly found in asymptomatic carriage than disease, including *Escherichia coli*, *Streptococcus pneumoniae*, and *S. pyogenes*. The methods applied herein, especially studying within-host evolution by comparing carried and invasive isolates are a promising avenue for better understanding of pathogenesis in these and similar organisms.

References chapter 7

1. Young BC, Golubchik T, Batty EM, Fung R, Larner-Svensson H, Votintseva AA, Miller RR, Godwin H, Knox K, Everitt RG, et al. Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. Proc Natl Acad Sci U S A 2012;109(12):4550.

2. Das S, Lindemann C, Young BC, et al. Natural mutations in a *Staphylococcus aureus* virulence regulator attenuate cytotoxicity but permit bacteremia and abscess formation Proc Natl Acad Sci U S A. 2016. 113(22):E3101-10.

3. Li T, He L, Song Y, Villaruz AE, Joo HS, Liu Q, Zhu Y, Wang Y, Qin J, Otto M, et al. AraC-type regulator rsp adapts *Staphylococcus aureus* gene expression to acute infection Infect Immun 2015 Dec 28.

4. Tong SY, Davis JS, Eichenberger E, Holland TL, Fowler VG Jr. *Staphylococcus aureus* infections: epidemiology, pathophysiology, clinical manifestations, and management. Clin Microbiol Rev. 2015 Jul;28(3):603-61.

5. Gotts JE, Matthay MA. Sepsis: pathophysiology and clinical management. BMJ. 2016 May 23;353:i1585.

6. Laabei M, Uhlemann AC, Lowy FD, Austin ED, Yokoyama M, Ouadi K, Feil E, Thorpe HA, Williams B, Perkins M, et al. Evolutionary trade-offs underlie the multi-faceted virulence of *Staphylococcus aureus* PLoS Biol 2015 Sep 2;13(9):e1002229.

7. Rose HR, Holzman RS, Altman DR, Smyth DS, Wasserman GA, Kafer JM, Wible M, Mendes RE, Torres VJ, Shopsin B. Cytotoxic virulence predicts mortality in nosocomial pneumonia due to methicillin-resistant *Staphylococcus aureus* J Infect Dis 2015 Jun 15;211(12):1862-74.

8. Jenkins A, Diep BA, Mai TT, Vo NH, Warrener P, Suzich J, Stover CK, Sellman BR. Differential expression and roles of *Staphylococcus aureus* virulence determinants during colonization and disease. MBio. 2015 Feb 17;6(1):e02272-14.

9. Recker M, Laabei M, Toleman MS, et al. Clonal differences in *Staphylococcus aureus* bacteraemia-associated mortality. Nat Microbiol. 2017 Aug 7. doi: 10.1038/s41564-017-0001-x.

10. Russell CD, Baillie JK. Treatable trains and therapeutic targets; Goals for systems biology in infectious disease. Current opinion in Systems Biology. 2017. 2:140-146.

11. Wang L, Ko ER, Gilchrist JJ, Pittman KJ, et al. Human genetic and metabolite variation reveals that methylthioadenosine is a prognostic biomarker and an inflammatory regulator in sepsis. Sci Adv. 2017 Mar 8;3(3):e1602096.

12. Scicluna BP, van Vught LA, Zwinderman AH, et al. Classification of patients with sepsis according to blood genomic endotype: a prospective cohort study. Lancet Respir Med. 2017 Aug 29. pii: S2213-2600(17)30294-1.

13. Rahman A, Izaki K, Kato I, Kamio Y. Nucleotide sequence of leukocidin S-component gene (lukS) from methicillin resistant *Staphylococcus aureus*. Biochem Biophys Res Commun. 1991 Nov 27;181(1):138-44.

14. Krüger A, Lucchesi PM. Shiga toxins and stx phages: highly diverse entities. Microbiology.

2015 Mar;161(Pt 3):451-62.

15. Davis BM, Waldor MK. Filamentous phages linked to virulence of *Vibrio cholerae*. Curr Opin Microbiol. 2003 Feb;6(1):35-42.

16. Freeman VJ, Morse, IU. Further observations on the change to virulence of bacteriophage-infected avirulent strains of *Corynebacterium diphtheria* J Bacteriol. 1952 Mar;63(3):407-14.

17. Carter AT, Peck MW. Genomes, neurotoxins and biology of *Clostridium botulinum* Group I and Group II. Res Microbiol. 2015 May;166(4):303-17.

18. Casas V, Maloy S. Role of bacteriophage-encoded exotoxins in the evolution of bacterial pathogens. Future Microbiol. 2011 Dec;6(12):1461-73.

19. Erken M, Lutz C, McDougald D. The rise of pathogens: predation as a factor driving the evolution of human pathogens in the environment. Microb Ecol. 2013 May;65(4):860-8.

20. World Health Organisation *General immunology.* Module 1: The immunological basis for immunization series. 1993. Accessed from: www.who.int/vaccines-documents/DocsPDF-IBI-e/mod1-e.pdf

21. Fowler VG Jr, Proctor RA. Where does a *Staphylococcus aureus* vaccine stand? Clin Microbiol Infect. 2014 May;20 Suppl 5:66-75.

22. Proctor RA. Recent developments for *Staphylococcus aureus* vaccines: clinical and basic science challenges. Eur Cell Mater. 2015 Dec 2;30:315-26.

23. Giersing BK, Dastgheyb SS, Modjarrad K, Moorthy V. Status of vaccine research and development of vaccines for *Staphylococcus aureus*. Vaccine. 2016 Jun 3;34(26):2962-6.

24. Adhikari RP, Kort T, Shulenin S, Kanipakala T, Ganjbaksh N, Roghmann MC, Holtsberg FW, Aman MJ. Antibodies to *S. aureus* LukS-PV Attenuated Subunit Vaccine Neutralize a Broad Spectrum of Canonical and Non-Canonical Bicomponent Leukotoxin Pairs. PLoS One. 2015 Sep 14;10(9): e0137874.

25. Yeap AD, Woods K, Dance DAB, Pichon B, Rattanavong S, Davong V, Phetsouvanh R, Newton PN, Shetty N, Kearns AM. Molecular Epidemiology *of Staphylococcus aureus* Skin and Soft Tissue Infections in the Lao People's Democratic Republic. Am J Trop Med Hyg. 2017 Aug;97(2):423-428.

26. Lower SK, Lamlertthon S, Casillas-Ituarte NN, Lins RD, Yongsunthon R, Taylor ES, DiBartola AC, Edmonson C, McIntyre LM, Reller LB, et al. Polymorphisms in fibronectin binding protein A of *Staphylococcus aureu*s are associated with infection of cardiovascular devices Proc Natl Acad Sci U S A 2011 Nov 8;108(45):18372-7.

27. Hos NJ, Rieg S, Kern WV, Jonas D, Fowler VG, Higgins PG, Seifert H, Kaasch AJ. Amino acid alterations in fibronectin binding protein A (FnBPA) and bacterial genotype are associated with cardiac device related infection in *Staphylococcus aureus* bacteraemia. J Infect. 2015 Feb;70(2):153-9.

28. Eichenberger EM, Thaden JT, Sharma-Kuinkel B, et al. Polymorphisms in Fibronectin Binding Proteins A and B among *Staphylococcus aureus* Bloodstream Isolates Are Not Associated with Arthroplasty Infection. PLoS One. 2015 Nov 25;10(11):e0141436.

29. Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, McCombie WR. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. Genome Res. 2015 Nov;25(11):1750-6.

30. Chapman SJ, Hill AV. Human genetic susceptibility to infectious disease. Nat Rev Genet. 2012 Feb 7;13(3):175-88

31. DeLorenze GN, Nelson CL, Scott WK, Allen AS, Ray GT, Tsai A, Quesenberry CP, Fowler VG. Polymorphisms in HLA class II genes are associated with susceptibility to *Staphylococcus aureu*s Infection in a white population J Infect Dis 2016 Mar 1;213(5):816-23.

Appendix 1

The cortex variant caller is a powerful tool for identifying both SNP and indel variants in a group of isolates. It uses a reference graph to identify and locate variation.

We applied this tool as one of 3 methods of identifying variation among a group of closely related isolates in chapter 4, and among related (but more distantly) isolates in chapter 5.

Several experiments conducted during the analysis for Chapter 4 demonstrated the variable sensitivity of the cortex variant caller according to both the choice of reference, and the degree of relatedness of the group under comparison.

I Choice of reference

For each group of isolates from an individual, we selected the assembled genome of one carriage isolate to use as a reference. We compared the performance of the cortex variant caller for each set of isolates using a) the patient specific reference and b) a randomly selected reference.

	Patient specific reference	Random reference
C vars identified	602	416
B vars identified	486	286
D vars identified	153	114
Total vars identified	1241	816

Using the Cortex var caller with a random reference had only 66% sensitivity

compared to performance using of a patient specific reference. All categories of variation were affected, but the B variants – those that separate carriage and disease, and which formed the basis of our findings – suffered the greatest loss of sensitivity, finding just 59% of those detected using a well matched reference.

II Varying relatedness of comparison group

When the "nearest neighbour" was included in the analysis of variation within each set of isolates we likewise found reduced sensitivity to within-host variation.

	Patient isolates only	Patient isolates plus nearest neighbour
C vars identified	602	484
B vars identified	486	411
D vars identified	153	145
Total vars identified	1241	1040

Here we found that, even using a patient specific reference, the inclusion of a more distantly related isolate in the set of isolates for comparison reduces the sensitivity of the Cortex var caller. The overall sensitivity was 84%.