Whole Genome Sequencing (WGS) as a unified platform for outbreak identification and resistance prediction in *Staphylococcus aureus*

Nicola Claire Gordon

Oriel College, Oxford

Nuffield Department of Medicine

A thesis submitted for the degree of Doctor of Philosophy

Trinity Term 2016

Whole Genome Sequencing (WGS) as a unified platform for outbreak identification and resistance prediction in *Staphylococcus aureus*

A thesis submitted for the degree of Doctor of Philosophy, Trinity Term 2016.

Nicola Claire Gordon, Oriel College, Oxford.

Staphylococcus aureus continues to present challenges to modern healthcare, due to its acquisition of antimicrobial resistance factors, its ability to cause invasive infections associated with significant morbidity and mortality, and its propensity for person-to-person transmission resulting in outbreaks. For outbreak investigation, current typing methods lack resolution, and the relatively slow turnaround times may hinder effective infection control intervention.

Whole-genome sequencing (WGS) is rapidly becoming faster and more affordable, offering increased resolution in a comparable timeframe. Having the entire genome means the data can also be explored for resistance and virulence testing.

In this thesis, I explore the use of WGS for investigating outbreaks and for predicting antimicrobial resistance phenotype. First, I establish the genetic diversity in the nasal *S. aureus* population at acquisition and after long-term carriage, to determine the effect of within-host diversity on outbreak WGS interpretation. I then use 15 well characterised *S. aureus* outbreaks to develop an WGS approach for outbreak investigation. I test this approach by applying WGS to a further 5 outbreaks where the infection control investigation was inconclusive. Combining the epidemiological and WGS data from all 20 outbreaks, I then evaluate whether the WGS data can be used to predict the possibility of a long-term carrier involved in maintaining an outbreak, and apply this in real-time using a rapid turnaround benchtop sequencer. Finally, I explore the use of WGS for predicting antimicrobial resistance phenotype.

I conclude that, for outbreaks, WGS has enhanced resolution compared to standard techniques and can give additional information to aid the outbreak investigation. For antimicrobial resistance, WGS is as sensitive and specific as routine testing methods. WGS provides a promising alternative to traditional culture and typing methods for enhancing our understanding of *S. aureus* and ultimately other pathogens.

Acknowledgements

I am grateful to my supervisors, Professors Sarah Walker, Tim Peto and Angela Kearns for their guidance, patience and supervision, and to Professor Derrick Crook for his mentorship.

I would also like to acknowledge and thank the many other people who have contributed data, help or advice (or all three). In no particular order: the staff of the PHE Staphylococcal reference laboratory, especially Margery Gannon for tolerating and assisting my rifling of her meticulously kept freezer contents, and Bruno Pichon for providing data; James Price and Kevin Cole in Brighton for digging out samples, results and data; John Paul for providing the initial funding nod and for his editorial input; the various infection control teams who contributed outbreaks, particularly Dr James Q Nash and his team in Ashford.

Thanks are due to the entire Modernising Medical Microbiology empire, but especially to Tanya Golubchik and Daniel Wilson for all their help and advice and for enduring my endless questions; Anna Shepherd, Jessica Hedges and Nicola de Maio for help with analyses; Luke Anson, Dai Griffith and Tonya Votintseva in the laboratory; John Finney for retrieving data on demand; and Ali Vaughan for everything.

Thanks are also due to Dr David Wareham, QMUL, for his early guidance and support.

Gratitude goes to my office confreres Nicole, David, Bernadette, Tim, Ana, Nick, and John for making the three years enormous fun.

Finally, I would like to thank Robert Gordon and Daniel Tallon for proofreading.

Funding

This work was supported by a Medical Research Council Doctoral Training Fellowship (MR/K023985/1).

Additional funding was via the UKCRC Modernising Medical Microbiology Consortium, including the NIHR Oxford Biomedical Research Centre, the UKCRC Translational Infection Research Initiative supported by Medical Research Council, Biotechnology and Biological Sciences Research Council and the National Institute for Health Research on behalf of the Department of Health (Grant G0800778) and the Wellcome Trust (Grant 087646/Z/08/Z), and Wellcome Trust core funding (Grant 090532/Z/09/Z). Sample collections in Brighton were supported by a grant from the Hospital Infection Society, and the Lausanne sequencing was funded by a grant from the Swiss Foundation for Research.

Ethical approval

This work was completed in conjunction with Public Health England as part of service development. No additional samples or patient information other than that obtained for routine analysis were used. Patient data has been anonymised as per information governance guidelines. Ethical approval for sequencing isolates from routine clinical samples, and linkage to patient data without individual patient consent, has been obtained from Berkshire Ethics Committee (10/H0505/83) and the UK National Information Governance Board (8-05(e)/2010). Public Health England has Patient Information Advisory Group approval to hold and analyse surveillance data for public health purposes under Section 60 of the Health and Social Care Act 2001.

Declarations and Attributions

The work contained within this thesis is my own, but could not have been undertaken without the direct and indirect involvement of many people. In general, sample collection was done by individual clinical teams, and initial processing by staff at the respective routine or reference laboratories.

The bioinformatics pipeline for assembling raw reads was developed by an in-house bioinformatics team, whose analysis scripts I have also utilised throughout for SNP finding and tree-building.

Chapter 2

Isolates for this study came from a collection previously established by Dr Ruth Miller. Dr Daniel Wilson provided advice on the study design. Dr Nicola de Maio kindly ran my BEAST output through his newly developed BASTA programme and provided the results for the construction of figure 2.10.

Chapters 3 and 4

Individual outbreak teams collected the samples and the initial epidemiological data. Further data was sought via a feedback questionnaire (see appendix) and by travelling to, or holding telephone case conferences with, the relevant departments to present and discuss the findings. The infection control leads for the outbreak teams were Dominique S Blanc, Jennifer Collins, Nick Cortes, Mark Cubbon, Frances K Gould, Peter J Jenks, Lily O' Connor, Jorge M Orendi, Karthik Paranthaman, Laurent Senn, Helen L Thomas, and Sarah Wyllie.

Additional background isolates for within-spa / MLST comparison were obtained from locally held collections which had been previously sequenced. Dr Anna Shepherd ran the remapping for the Lausanne outbreak.

Chapter 5

MRSA slopes were obtained from the routine laboratory. Luke Ansen performed the library preparation and ran the MiSeq. Dr Zam Iqbal and Phelim Bradley developed the Mykrobe programme, which was incorporated into the assembly pipeline by Dr Carlos Del Ojo Elias.

Chapter 6

Dr Tanya Golubchik contributed significantly to this work, writing the Typewriter script for using BLAST on the assembled data and performing multiple tests and iterations to get the final version, and initially noted the *blaZ* discrepancies which were further investigated by myself.

Initial phenotype data for the derivation set was obtained from the routine laboratories in Oxford and Brighton, and these isolates had been extracted and sequenced previously. Kevin Cole did some additional phenotypic testing of Brighton isolates. Marcus Morgan and his staff in the JR routine laboratory ran the Phoenix testing of the validation set. Professor Tim Peto provided figure 6.2.

Publications and conference proceedings

Publications

1. Whole Genome Sequencing reveals the contribution of long-term carriers in *Staphylococcus aureus* outbreak investigation. Gordon NC, Pichon B, Golubchik T et al. J Clin Microbiol 2017; 55:2188-2197

2. The Stealthy Superbug: the Role of Asymptomatic Enteric Carriage in Maintaining a Long-Term Hospital Outbreak of ST228 Methicillin-Resistant *Staphylococcus aureus*. Senn L, Clerc O, Zanetti G, Basset P, Prodhom G, Gordon NC et al, MBio 2016, 7:e02039-15

3. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and Mycobacterium tuberculosis. Bradley P, Gordon NC, Walker TM et al, Nat Commun. 2015, 6:10063

4. Whole-genome sequencing in hierarchy with pulsed-field gel electrophoresis: the utility of this approach to establish possible sources of MRSA cross-transmission. Moore G, Cookson B, Gordon NC et al. J Hosp Infect 2015, 90:38-45

5. Prediction of *Staphylococcus aureus* antimicrobial resistance by whole genome sequencing. Gordon NC, Price JR, Cole K et al. J Clin Microbiol 2014, 52:4

6. The usefulness of whole genome sequencing in the management of *Staphylococcus aureus* infections. Price J, Gordon NC, Crook D et al, Clin Microbiol Infect 2013,19:784-9

7. A pilot study of rapid benchtop sequencing of *Staphylococcus aureus* and Clostridium difficile for outbreak detection and surveillance. Eyre DW, Gordon NC, Golubchik T et al, BMJ Open 2012, 6:2(3)

Oral presentations

1. Whole genome sequencing for *Staphylococcus aureus* and beyond. British Society for Microbial Technology (BSMT) annual conference, London 2015.

2. Investigating *S. aureus* outbreaks using whole genome sequencing. ECCMID, Copenhagen 2015

3. Whole genome sequencing to study resistance in *Staphylococcus aureus*. BSAC Antimicrobial Resistance Mechanisms annual workshop, Birmingham 2014

4. Blinded validation study of genotype prediction of antimicrobial resistance from *Staphylococcus aureus* whole genome sequences. ECCMID, Barcelona 2014

5. Genotypic prediction of anti-microbial susceptibilities in *Staphylococcus aureus*. BIA Annual Scientific Meeting, London 2013

Poster presentations

1. Whole genome sequencing of an ST8 *S. aureus* outbreak in South East England. ISSSI, Chicago 2014

2. Prevalence of *blaZ* in UK clinical isolates of *Staphylococcus aureus* and identification of 3 novel function impairing mutations. ECCMID, Barcelona 2014

3. Genotypic prediction of antimicrobial susceptibilities in *Staphylococcus aureus*. IDSA Infection Week, San Diego 2012

4. Rapid bench top whole genome sequencing for investigation of a putative MRSA outbreak. ECCMID, London 2012

Table of Contents

List of Figures

List of Tables

List of abbreviations and acronyms

AST	antimicrobial susceptibility testing
bp	base pair
DNA	deoxyribonucleic acid
MCMC	Markov chain Monte Carlo method
MIC	minimum inhibitory concentration
MMM	Modernising Medical Microbiology
NHS	National Health Service
OUH	Oxford University Hospitals NHS Foundation Trust
PCR	polymerase chain reaction
PFGE	pulsed field gel electrophoresis
PHE	Public Health England (formally the Health Protection Agency)
SNV	single nucleotide variant
WGS	whole genome sequencing

List of genomic analysis tools used

Tool	Description	Website
BASTA	BAyesian Structured coalescent Approximation. Addition to the BEAST package (see below) allowing incorporation of prior knowledge of migration rates / ancestral locations.	https://bitbucket.org/nicofmay/basta-bayesian-structured-coalescent-approximation
BEAST	Bayesian Evolutionary Analysis Sampling Trees. Programme for reconstructing phylogenies, testing evolutionary hypotheses, and estimating time to most recent common ancestor (TMRCA). Uses Markov chain Monte Carlo Method (MCMC) to reconstruct ancestries, and gives posterior probability of the inferred phylogeny. Can use strict (lineages all evolve at same rate) or relaxed (lineages evolve at different rates) molecular clock models.	http://beast.bio.ed.ac.uk
BLAST	Basic Local Alignment Search Tool. A sequence is divided into shorter "query word" sequences, which are matched to a database. Matches are extended in either direction, and a score calculated based on the closeness of the match. BLASTn (standard nucleotide BLAST): query is a nucleotide sequence, returns the closest matching nucleotide sequences; tBLASTn (translated BLAST): query is a protein sequence to be matched against a database of translated nucleotide sequences (as distinct from BLASTp and BLASTx, which search a protein database).	https://blast.ncbi.nlm.nih.gov/Blast.cgi
Outbreaker	R package for reconstruction / estimation of transmission trees from pathogen sequences. Estimates dates of infection, mutation rates and imported cases and effective reproduction number using MCMC methods.	https://sites.google.com/site/therepiproject/r-pac/outbreaker
Path-O-Gen	Assesses temporal signal and "clocklikeness" of an estimated phylogeny. Roots the tree at the position most compatible with the assumption of a molecular clock and determines whether there is sufficient temporal signal for clock assumptions to be valid.	http://tree.bio.ed.ac.uk/software/pathogen/

PHAST	PHAge Search Tool: webserver for identifying, annotating and displaying prophage sequences. Query sequences (e.g. contigs from a de novo alignment) are searched against a database of bacterial phages, and examined for features indicating an incomplete or unannotated prophage.	http://phast.wishartlab.com
PhyML	Programme for creating maximum likelihood phylogenetic trees, using a heuristic approach and nearest neighbour interchange.	http://www.atgc-montpellier.fr/phyml
Plasmid-finder	Uses BLAST to search online plasmid database for query sequences from whole genomes or contigs.	https://cge.cbs.dtu.dk/services/PlasmidFinder/
Readseq	Reads and converts biosequences between a selection of common biological sequence formats, e.g. fasta, EMBL, GenBank.	http://www.ebi.ac.uk/Tools/sfc/readseq/
SAMtools	Set of utilities for reading, interacting and post-processing aligned reads as generated by e.g. Stampy (see below).	https://sourceforge.net/projects/samtools/
Stampy	Package for mapping short Illumina reads to a reference genome. Uses a fast hashing algorithm: the short reads from the query genome are matched against a hash table constructed from the reference genome. Once the closest match is identified, the query sequence is mapped to the relevant position in the genome.	http://www.well.ox.ac.uk/project-stampy
Velvet	For de novo assembly of reads. Breaks reads into kmers of length n, then constructs a De Bruijn graph from the reads to derive the most likely links between kmers to produce a contig.	https://www.ebi.ac.uk/~zerbino/velvet/

1 Introduction

1.1 Ogston's coccus

Staphylococcal species were probably observed as early as 1683 by Anton van Leewenhoek, a Dutch draper by trade and lens-maker by hobby, who observed oval and circular morphologies among the small living 'Animacules' found in the scrapings from his own teeth (1). Two centuries later, J. Cosser Ewart described the spirillum, bacillus, bacterium and micrococcus morphologies of bacteria, with micrococcus referring to those circular organisms which could be found in clusters or chains (2-4), although initially Ewart believed all forms to be different life cycle stages of the same species.

In 1879, Kocher demonstrated the presence of micrococci in pus (5); however, its pathological nature was in some dispute, with several parties, Kocher included, certain of its causality in infection, while others (notably William Watson Cheyne) protested its innocence (6).

The debate was settled by Alexander Ogston, who in 1881, while investigating the causes of surgical wound infection (7), reported the universal presence of micrococci in pus from acute suppurations (n=65), as well as its absence from chronic "cold" abscesses (n=13). He also noted that, while in some specimens the micrococci were arranged in chains as previously identified by Bilroth (8), in others they were arranged in clusters "like the roe of fish" "where there was neither beginning nor end", and so he differentiated between the streptococcal and staphylococcal genera. He named the clusters *Staphylococci*, from the Greek σταφυλή, staphulē, cluster of grapes (http://biblesuite.com/greek/2848.htm), and κόκκος, kokkos, grain or seed (http://biblesuite.com/greek/4718.htm).

Ogston went on to demonstrate convincingly the infectious nature of *Staphylococcus* by injecting pus containing the organisms into chloroformed guinea pigs, white mice and wild mice. The infected animals exhibited symptoms of blood poisoning (noted as "refused food, sat cowering, were listless and apathetic, with disordered coats"). Subsequent

sacrifice and dissection of the animals found Staphylococci at the injection sites and in their blood, and onward infection was demonstrated by injecting organisms thus recovered into further animals, resulting in similar symptoms of sepsis. Ogston's work also provided support for the antiseptic practices of Lister (9), as Ogston noted that pus treated with carbolic acid did not produce the same result on inoculation.

Offering a sop to his dissenters, Ogston's paper concludes that "lastly, there are possibly micrococci that do not produce suppuration", presumably identifying what are now grouped together as the coagulase negative staphylococci. The formal distinction between the two species was made in 1884, by Anton J Rosenbach, a German surgeon and perhaps more of a Latinist than a philhellene, who separated *Staphylococcus aureus* (Latin aureus, golden) from *Staphylococcus albus* (Latin, white) which, we now know, encompasses a large number of (mostly) non-pathogenic species (10).

1.2 *Staphylococcus aureus* in modern healthcare

More than a century since Ogston's discoveries, and despite enormous advances in sanitation and antiseptic techniques and the advent of antimicrobial drugs, *S. aureus* remains a major public health concern, particularly as a healthcare-associated pathogen. It has been reported as a cause of healthcare-associated outbreaks since the 1950s (11), and emerged as a significant cause of community clusters in the 1980s (12-14), although the apparent rise in community infections may coincide with increasing resistance, as infections caused by organisms not cured by first-line drugs are likely to persist and spread, and be subjected to further investigation and reporting.

As well as asymptomatic colonisation, *S. aureus* can cause a wide spectrum of diseases, including skin / soft tissue and wound infections (15, 16), bone / joint infections (17), respiratory tract infections (pneumonia, lung abscesses and ventilator-associated pneumonia (18, 19)) and endovascular infections such as endocarditis (20). It can be

particularly problematic in the presence of prosthetic material, where the formation of biofilms makes eradication difficult (21). Mortality in bloodstream infection can be up to 30% (22).

The problem has increased with the emergence of methicillin-resistant (MRSA) epidemic strains. In the UK in recent years EMRSA-15 and EMRSA-16 have dominated hospital-acquired infections (23, 24), while in the US, the highly transmissible and pathogenic USA300 MRSA clone is now the leading cause of community acquired skin and soft tissue infections (25). USA300 is a particular public health concern, with implications for antimicrobial use since the first-line antimicrobials for suspected staphylococcal disease in North America are now vancomycin and linezolid, usually regarded as second- or third-line drugs in other parts of the world. It is therefore important to recognise the factors promoting the survival and pathogenicity of epidemic strains, to ensure adequate surveillance and timely intervention to prevent similar widespread dissemination of emerging, resistant epidemic clones.

1.3 Outbreaks

1.3.1 Investigating outbreaks

There is almost universal human exposure to *S. aureus*, since it is chronically carried by 20% of the population, and intermittently by at least another 30% (26). Person-to-person spread may occur at low levels, resulting in a background diversity of strains within a population. Whether a strain causes disease or not is dependent on both host and pathogen factors. Outbreaks can occur in the community or in association with healthcare exposure, with rapid person-to-person spread causing both colonisation and disease. Approximately 500 putative *S. aureus* outbreaks per year are referred to Public Health England (PHE) for investigation (27), with significant impacts on the individuals and institutions involved.

Distinguishing between the relative contributions of bacterial and clinical factors to an outbreak is vitally important. Epidemiologically, an outbreak is defined as an increase in cases of disease over and above what would normally be expected in a population. Consequently, an outbreak may occur without person-to-person transmission, for example, an increase in post-operative wound infections due to poor surgical technique, or failure to give routine surgical prophylaxis. In practice, however, the instigation of a formal outbreak investigation relies primarily on determining the relatedness of isolates. Relatedness consistent with an outbreak due to person-to-person transmission is inferred when isolates with a common epidemiological link are identified as belonging to the same typing group. However, this approach may falsely impute relatedness unless the resolution of the typing method is sufficiently high. Similarly, the background diversity of *S. aureus* in the population must be understood to ensure that an apparent transmission event is not simply due to a frequently occurring clone prevalent within that population. These two factors are interdependent: where there is high background diversity, the occurrence of 2 or more highly similar isolates may be sufficient to identify an outbreak, however, where a successful clone has become established in a community, chance alone dictates that there is a high likelihood that isolates from any two individuals will belong to the same typing group. This is of particular importance if epidemiological data are incomplete.

1.3.2 Typing methods for outbreaks

Typing of *S. aureus* became important in the 1950s when the emergence of drug resistance prompted interest in the transmission of resistant strains and paved the way for current infection control practices. Initially, strains were assumed to be related on the basis of their antibiograms, since resistance was so rare that, if two epidemiologically related isolates were resistant to the same antimicrobial, it could be assumed that they originated from the same parent organism (28). As resistance became more common,

other methods were developed, based on phenotypic behaviour or genotypic analysis, including serotyping, gas chromatography, ribotyping and analysis of endonuclease restriction patterns. As the need for inter- as well as intra-hospital comparison increased, phage typing, which uses the pattern of lysis caused by a panel of bacteriophages, was adopted by most reference laboratories, and the development of an internationally standardised set of bacteriophages allowed comparison of strains across different geographical regions as well as within a hospital outbreak.

Problems with phage typing include technical difficulty, cost, lack of phage activity against some strains with the standard phage set, and, as epidemic strains of *S. aureus* emerged, lack of resolution, meaning that a combination of techniques in addition to phage typing had to be employed to distinguish accurately between related isolates (for example, amino-glycoside inactivating enzyme analysis plus plasmid fingerprinting (29, 30), restriction endonuclease analysis (31), radio-labelled protein gel electrophoresis (32)). To demonstrate that, for many putative outbreaks, phage typing alone is insufficient, Pekkala and colleagues (33) used restriction endonuclease analysis to identify 20 different strains of *S. aureus* in 23 patients initially suspected, due to epidemiology and phage typing, of forming a clonal cluster.

In a 1993 review of typing methods for bacteria in general, Maslow (34) considered pulsed field gel electrophoresis (PFGE) to be the preferred method with maximal typeability, reproducibility and discriminatory power. However, a study by Tenover et al one year later (35) evaluated 11 different typing methods (antibiogram, phage typing, biotyping, immunoblot, insertion sequence typing, multilocus enzyme electrophoresis, restriction enzyme analysis of plasmid (REAP) DNA, PFGE, field inversion gel electrophoresis (FIGE), restriction analysis of the coagulase gene, and restriction fragment length polymorphism analysis (RFLP)), applying each to a test set of organisms comprising four well characterised outbreaks and a number of epidemiologically unrelated strains. Despite the multitude of techniques, none emerged with clear superiority. PFGE and FIGE were

highly discriminatory, but each falsely allocated non-outbreak strains to the outbreak groups. Lack of reproducibility and subjective interpretation were also drawbacks with these methods.

PFGE was developed in 1984 by Schwarz and Cantor (36), and is based on the digestion of whole cell DNA by restriction enzymes with limited cleavage sites. The restriction enzyme digestion process yields 5-20 fragments of 30-2000kb, which can be separated in a gel by the application of an electric current. Intermittently pulsing (as in PFGE) or inverting (as in FIGE) the electric current increases the separation of fragments, improving the interpretation. Both methods can be applied across species. However, interpretation is somewhat subjective and results are improved if strains from the same cluster are analysed at the same time, on the same gel, limiting its usefulness in ongoing outbreaks.

The problems highlighted in Tenover's study were overcome to some degree by the development of guidelines for interpretation (37). The superiority of PFGE versus phage typing, if guidelines were applied, was subsequently demonstrated by Bannerman et al (38), who, in a test set of 300 isolates, found that 32% of the isolates were unreactive with the standard phage set, while PFGE had far superior discriminatory power. For this reason, by the late 1990s PFGE had replaced phage typing as the standard reference technique, and the standardisation of parameters, equipment, imaging software and pulsing conditions improved the reproducibility and allowed better cataloguing and comparison of strains.

Subsequently, multiple studies (as summarised by Leonard et al (39)) have demonstrated the superiority of PFGE versus a variety of techniques and PFGE remains the gold standard for reference laboratories worldwide. However, with the exception of well recognised patterns for globally important epidemic clones such as EMRSA-15 and -16, and USA300, it is less useful for global molecular epidemiology and tracing the origins of emerging strains (40), since the band patterns contain no information about the genetic

content. Table 1.1 summarises the typing methods which have been applied to *S. aureus* in previous studies.

In current reference laboratory practice, typing to assist with outbreak investigations is usually stratified. Initial strain characterisation is performed using multi-locus sequence typing (MLST), where the combination of distinct alleles in defined housekeeping genes determines the sequence type (41), and *spa* typing, which uses single locus sequencing of the repeat region of Staphylococcal protein A (*spa*) gene and assigns *spa*-type based on the number and order of repeat motifs (42, 43). Further typing by PFGE is only necessary if the strains are of the same MLST sequence and *spa*-type. In some cases, further testing may be required (for example typing of the SCC*mec* gene cassette associated with methicillin resistance) increasing both costs and turnaround times (44, 45). This is particularly true of the major epidemic clones which have limited genetic diversity and which may appear highly similar despite sharing no recent epidemiological link (46), highlighting the need for an alternative typing method.

TABLE 1.1 COMPARISON OF TYPING METHODS USED TO INVESTIGATE *S. AUREUS*

Typing Method	Description	Advantages	Disadvantages
Phenotype based methods			
Antibiogram	Compares susceptibility to a set of antimicrobials. Discriminatory power is increased if zone size, rather than categorical interpretation, is used.	Easy to perform. No specialist equipment required. Reproducible.	Lacks resolution, since resistance can arise from minor genetic events such as point mutation or loss/gain of a mobile element.
Phage typing	Utilises pattern of lysis caused by a panel of bacteriophages.	Good reproducibility if standard phage set used. Easy to interpret.	Technically complex, expensive. Standard phage set has limited activity against some strains. May lack resolution for epidemic strains.
Serotyping	Based on variation in cell surface antigen expression (e.g. lipopolysaccharides, membrane proteins, flagella, fimbriae). Tested using specific agglutination sera or via fluorescent / enzyme labelling assays.	Reproducible, easy to interpret.	Some strains may be un-typable. Poor discrimination. May be cross-reactivity with some antigens. No standardised *S. aureus* scheme.
Biotyping	Based on colonial morphology, metabolic activity and growth conditions (e.g. sugar fermentation, amino acid decarboxylation / deamination, pH tolerance, haemolysis).	Reproducible, easy to interpret. No special equipment required.	Poor discrimination. Variation in gene expression may affect phenotype.
Protein electrophoresis methods			
Immunoblot	Gel electrophoresis of cellular proteins. Product is reacted with labelled anti-staphylococcal antibodies and examined by Western blotting.	Less susceptible to running conditions than other protein electrophoresis methods.	Complex and laborious; subjective interpretation of banding pattern.
Multi-locus enzyme electrophoresis (MLEE)	Electrophoresis of cell extracts containing soluble enzymes. Variability in bands reflects differences in enzymic composition.	Highly reproducible, relatively easy to interpret.	Only moderately discriminatory. Technically demanding.
Plasmid analysis			
Plasmid finger printing	Types strains according to size and number of plasmids.	May provide additional resolution beyond the core genome.	Lacks resolution if used alone due to mobile nature of plasmids. Electrophoretic properties of plasmids may

			be affected by state of DNA (e.g. coiled, nicked, integrated).
Restriction endonuclease analysis of plasmid DNA (REAP)	Restriction endonuclease analysis of plasmid DNA; can distinguish between unrelated plasmids of same size based on restriction site location.	Improved resolution versus plasmid fingerprinting.	As for plasmid fingerprinting.
Genotyping methods			
Restriction endonuclease analysis (REA)	Digestion of whole DNA by standard restriction endonucleases, generating hundreds of fragments of ~0.5 to 50 kb, followed by electrophoresis and analysis of band pattern.	Reproducible.	Complex profile, may contain overlapping bands or bands generated from digestion of plasmids. Band separation is improved by the application of a pulsed or inverted electrical current (see below).
Pulsed field gel electrophoresis (PFGE)	As for restriction endonuclease analysis (REA), above, but fragments are separated by the application of a pulsed electrical which improves separation.	Increased resolution compared with REA; reproducible with standardised techniques.	Lacks discrimination for epidemic clones. Requires considerable expertise, labour intensive. Limited portability.
Field inversion gel electrophoresis (FIGE)	As for REA, but fragments separated by alternating inversion of the electrical field.	As for PFGE.	As for PFGE.
Radiolabelled protein gel electrophoresis	As for REA, but with the addition of carbon-14 to produce an autoradiograph.	Increased resolution compared with REA.	Profiles remain complex, requires considerable expertise.
Insertion sequence typing	Uses number and position of insertion sequences (IS) located in chromosome.	Reproducible.	May be complicated by plasmid carriage of insertion sequences, lacks discrimination.
Restriction fragment length polymorphism (RFLP) analysis	DNA digestion by endonucleases followed by electrophoresis. Southern blotting using labelled DNA probes is then used to identify fragments containing specific DNA sequences.	Reproducible (if standard set of probes used), easy to interpret.	Lacks discrimination, labour intensive. Requires costly reagents / equipment.
Ribotyping	As for RFLP, but using restriction digestion of 16s rRNA, 25s rRNA and tRNA.	Reproducible, easy to interpret.	Lacks discrimination. More useful for comparing relatedness of organisms over time.

PCR based methods			
Coagulase gene restriction analysis	PCR amplification and restriction analysis of coagulase gene, coded according to presence / absence of specific digest fragments. size and number of repeats	Reproducible, easy to interpret.	Lacks discrimination
Multi-locus sequence typing (MLST)	Uses allelic variation at seven house-keeping loci to give unambiguous sequence type.	Reproducible, high inter-laboratory agreement. Globally standardised typing scheme.	Lacks discrimination: more useful for tracking global epidemiology. Expensive.
spa-typing	PCR based typing of the hypervariable X region of the *spa* gene, based on number and type of repetitive sequences.	Reproducible, rapid. Globally standardised typing scheme.	Lacks discrimination. May not correlate with clonal relationships as defined by other typing methods (e.g. MLST)
Multi-locus VNTR analysis (MLVA)	Polymorphisms in chromosomal number tandem repeat (VNTR) elements	Rapid, reproducible. High throughput.	Lacks discrimination. No international standard protocol or nomenclature.
Other methods			
Aminoglycoside inactivating enzyme analysis	Analysis of amino-glycoside inactivating enzymes via nitrocellulose binding assay.	Reproducible.	Limited applicability (i.e. aminoglycoside resistant strains only). Lacks discrimination.
Gas chromatography	Analysis of fatty acid composition determined by gas-liquid chromatograpy	Reproducible	Lacks discrimination
Phage-head typing	Characterisation based on prophage morphology (e.g. PVL-carrying phage)	Reproducible. May be useful for sub-typing e.g. for epidemic clones which cannot be further discriminated by PFGE.	Limited applicability. Lacks discrimination when used as stand-alone method.
SCC*mec* *typing*	Based on variation in the *Staphylococcus* chromosome cassette SCC*mec*, which contains the *mec*A gene conferring methicillin resistance.	Reproducible. May give additional information regarding epidemiology of *mec*A gene. Standard nomenclature.	Limited applicability (MRSA only). High cost.

1.3.3 Cost of outbreaks

The costs of hospital outbreaks and the containment of MRSA have been the focus of several studies, particularly as part of the debate about whether preventing transmission is cost-effective. Rao et al (47) estimated that the total cost of eradicating MRSA carriage from a hospital during an outbreak which caused disease in 12 patients was approximately equal to the cost of treating a single infected patient ($9,984 for the eradication programme, including microbiological processing, cost of chlorhexidine and additional staffing versus $144,320 to treat 12 patients). The cost per patient varied depending on site of infection and whether or not ITU admission was required ($2,270 for a skin / soft tissue infection versus $16,060 for treatment of a bloodstream infection in ITU), although Rao et al did not take into account the patients' comorbidities and whether the infection was the cause of admission to ITU. Similarly, Mehtar et al (48) estimated that the cost of treating 12 infected patients during a five week outbreak was £6,440 (approximately £15,438 adjusted for inflation) compared with £6,495 for the containment programme including decolonisation of a further 18 patients and 3 staff members. Colonisation with *S. aureus* is widely accepted as a risk factor for infection (49-51), and therefore both authors make the case that prevention of additional cases by reducing MRSA transmission is likely to be cost-effective. Rao in particular argues that prevention of a single infection is worth the additional outlay of a control programme.

Coello et al (52) performed a case-control study to estimate the cost per patient of MRSA infection, comparing 67 infected surgical patients with age and co-morbidity matched uninfected controls. The costs considered included antimicrobial therapy, pathology costs, additional length of stay and additional testing. They found that the highest cost of an MRSA infection was due to additional length of stay, with infected patients spending a mean of 8.2 extra days in hospital compared with controls, at an additional cost of around £1,041 per patient.

The most comprehensive assessment of the cost of MRSA infection is by Kanerva and colleagues (53), who estimated that treating a patient with MRSA bacteraemia or pneumonia cost an additional €560, while managing a patient simply colonised with MRSA cost an additional €256, including decolonisation therapy, isolation, follow-up screening and the need for more expensive vancomycin prophylaxis for surgical procedures. The calculations did not include additional length of stay, which would increase the costs still further. The outbreak described by Kanerva, involving 144 infected patients and 152 colonised patients, also resulted in €1,183,808 of income loss due to bed closures.

Comparing the estimates of each of these studies is difficult due to varying currency, inflation, funding models and the different factors each has used in their analysis (for example, not all studies consider the added laboratory costs of an enhanced screening programme for preventing further cases during an outbreak). However, the total cost is likely to be in the region of £2,000-£4,000 per patient treated for infection (rather than colonisation) when each aspect of care is considered (table 1.2).

TABLE 1.2 ESTIMATED ADDITIONAL COST OF TREATING A PATIENT WITH HOSPITAL ACQUIRED INFECTION CAUSED BY MRSA

Study	Study summary	Estimated mean additional cost per patient treated	Cost adjusted for inflation	Adjusted cost in GBP	Comments
Rao et al, 1998 (47)	25 patients (12 infected, 13 colonised) in teaching hospital over 8 months	$9,526	$19,062	£12,528	Range: £2,984 for soft tissue infection to £21,117 for BSI requiring admission to ITU. Largest cost is increased LOS
Mehtar et al, 1989 (48)	30 patients (12 infected, 18 colonised) plus 3 colonised staff in general hospital over 5 week period	£392	£917	£917	Does not include additional costs accrued by microbiology or infection control departments
Coello et al, 1993 (52)	Case control study, 67 infected versus matched controls from surgical unit	£1,041	£1,880	£1,880	Largest cost is due to increased LOS
Plowman et al, 1999 (54)	UK government economic evaluation of cost of HAI (due to any organism)	£3,154	£4,846	£4,846	All HAI infection, not just MRSA
Rampling et al, 2001 (55)	8 infected patients on a surgical ward over 27 months	£3,066	£4,506	£4,506	Estimated from total cost of controlling the outbreak
Lee et al, 2004 (56)	10 central line infections on haemodialysis unit	$2,324	$2,921	£1,920	Most cases managed as outpatients
Kanerva et al, 2007 (53)	Large scale, hospital-wide outbreak, 144 infected patients, 152 colonised	€564	€639	£475	Does not include additional LOS. If adjusted for this at conservative estimate of 4.5 additional days, per patient cost is £2,464

BSI: bloodstream infection; LOS: length of stay.

Costs were adjusted for inflation using online calculators (http://www.usinflationcalculator.com/; http://www.thisismoney.co.uk/money/bills/article-1633409/Historic-inflation-calculator-value-money-changed-1900.html; http://www.lawyerdb.de/Inflationrate.aspx), and converted to GBP using the exchange rate as of 10/2/2015 (http://www.xe.com/currencyconverter/).

There is therefore compelling evidence that, for MRSA at least, there is a significant monetary cost to a hospital for a single infected patient, giving weight to the argument that early identification of an outbreak, followed by the instigation of relatively inexpensive methods (screening and decolonisation) to prevent further cases, is likely to be cost-effective.

None of these studies considers the effect on individual patients or the wider societal impact. As well as the increased morbidity, affected patients may suffer from side effects of treatment (e.g. reactions to antibiotics or decolonisation preparations), and there may be a psychological impact including the stigma of isolation and barrier nursing if they are affected by MRSA, or strains with particular virulence factors. Societal costs may include lost working days, and damage to hospital reputations. Although MRSA infections are likely to have the largest impact, MSSA can also cause similar effects (57).

1.3.4 Preventing transmission and controlling outbreaks

One of the major problems facing control programmes is that there is little evidence for which specific interventions are the most effective. Control of an outbreak requires the institution of a number of measures, and delineating the impact of each of these is difficult. In community outbreaks, control is frequently achieved by improved hygiene practices, particularly in outbreaks associated with contact sports or schools. This is cheap and easy, and, as environmental contamination is rarely identified as a source of ongoing transmission (58-61), additional measures may not be necessary. Decolonisation of non-symptomatic carriers is not generally required in outbreaks of skin / soft tissue infection (46, 59, 62, 63), although it is frequently instituted for carrier contacts in residential / day care facilities (64-66).

Outbreaks in hospitals are much more complex. Government and media reporting has frequently focussed on environmental contamination, with high-profile, expensive deep

cleaning programmes aimed at reassuring the public that action is being taken, although there is little correlation between documented hygiene scores and MRSA rates (67). Environmental contamination may be a factor, but a more important contributor is transient hand carriage by staff (68). Hand hygiene is of utmost importance, but the value of additional measures such as universal screening, isolation, aggressive decolonisation, and enhanced cleaning is unclear. Outbreaks may be due to a general breakdown in infection control, with an overall increase in transmission / infection caused by a variety of strains; an epidemic strain with the propensity for rapid dispersal such as EMRSA-15 or USA300; an environmental reservoir associated with poor cleaning; or a staff carrier. Staff carriage may be particularly problematic since staff might be transient carriers (especially on the skin / hands), "victims" of an epidemic clone, long-term carriers of an innocent strain, or the genuine source of an outbreak, particularly individuals with chronic skin conditions such as psoriasis or eczema (69, 70). Labelling a staff member as the source of an outbreak may have implications for their career and, although in some countries staff screening and exclusion are performed routinely (71), in the UK there is no universal policy for this and local negotiations must be made. Because of the potential impact, it is important that the evidence for implicating an individual staff member is robust.

In the UK, a variety of government directives have been introduced in an effort to reduce the rates of MRSA bacteraemia. The earliest of these initiatives was the introduction of surveillance, with obligatory reporting of MRSA bacteraemias introduced in 2003 (72). In 2004, the focus was on hospital cleanliness (the Cleaner Hospitals Improvement Programme (CHIP) (73) (74)) and hand hygiene (the Clean Your Hands Campaign (75)). These were aimed at reducing all nosocomial infections, however, specific mandatory targets for reducing MRSA bacteraemias were also introduced (76). The Saving Lives initiative in 2005 saw the implementation of care bundles for a variety of clinical practices (77), including care bundles for insertion and care of central and peripheral venous access devices, which were identified as risk factors for MRSA bloodstream infection.

Further generic infection control interventions include antimicrobial stewardship programmes (77, 78) and the controversial “bare below the elbows” campaign (79) embedded in the rather more prosaically named “Uniforms and Workwear” document from the Department of Health (80).

Finally, universal screening for MRSA carriage was instigated in 2010, although poor compliance, slow laboratory turnaround times and frequent patient transfers resulted in modification of the guidelines to include only admissions to high risk specialties (81, 82).

Overall, there has been a steady decline in MRSA bacteraemia rates since the peak in 2003 (83). However, while mandatory reporting has only recently been introduced for MSSA bacteraemias, the effect of control measures appears to be less marked (84).

Evidence for the effectiveness of each individual intervention is limited. In 1991, Boyce et al attempted to identify the most effective interventions for controlling MRSA, using a questionnaire to collect information regarding hospital infection control interventions which was then compared with the MRSA rates of those hospitals (68). No difference in MRSA prevalence in relation to any single one of 10 defined infection control interventions was found. However, in hospitals reporting fewer than 20 MRSA cases per year, infection control intervention of any sort resulted in total eradication of MRSA from the hospital, compared with a 71% reduction in MRSA cases in hospitals reporting 20-39 cases, and a 10% reduction in hospitals with more than 40 cases. Other factors associated with eradication or reduction in the Boyce study were: improved hand hygiene, isolation or cohort boarding of colonised patients, and cohort nursing. Aggressive decolonisation with oral antibiotics for carriers, used when other measures failed, was associated with improved eradication / reduction (95% versus 77% for non-treated carriers), but this was not statistically significant, and should be balanced with the risk of side effects (for patients with no clinical infection) and the risk of selecting for resistant organisms.

Of the various control measures introduced in the UK, only hospital purchasing of alcohol hand gel has been demonstrated to have an independent association with MRSA reduction (85, 86). The roles of screening and stewardship programmes in preventing transmission (rather than infections) are less clear (83).

The conclusion must therefore be that, apart from poor hand hygiene, the underlying contributors to *S. aureus* transmission and infection are multifactorial and there is limited evidence for the effectiveness of many of the interventions. This may be improved with a better understanding of the dynamics of an individual outbreak, including the nature of the organism involved. Crucial to this are rapid identification of an outbreak and access to rapid typing to identify individual transmission of a strain from an overall rise in background rate.

1.4 Multi-drug resistance

Much of the interest in *S. aureus* is due to the public health concern surrounding multi-drug resistance. Resistance to penicillin was observed in vitro in 1941 (87), and in vivo in 1944 (88). Efforts to circumvent resistance by using a beta-lactamase stable beta-lactam (usually flucloxacillin) were swiftly thwarted by the emergence of MRSA (89, 90), which is resistant to all beta-lactams except the newly developed cephalosporins, ceftaroline and ceftobiprole. Tracking MRSA rates, and particularly rates of MRSA bacteraemia as a marker for all MRSA-related disease, has become an integral part of surveillance and infection control systems in the UK, with each MRSA bacteraemia requiring a formal root-cause analysis. MRSA bacteraemia is particularly problematic as treatment options are limited to the glycopeptides (vancomycin or teicoplanin), which have been shown to be inferior to beta-lactams for MSSA (91, 92), daptomycin, which has been subject to an FDA warning due to eosinophilic pneumonia (93) and for which an optimal dosing regimen is yet to be established (94), or linezolid, an expensive alternative which in prolonged use

carries the risk of severe side effects such as peripheral neuropathy, optic nerve damage and neutropenia, and which has been associated with a higher mortality compared to vancomycin or daptomycin in the treatment of MRSA bacteraemia (95).

Resistance is also problematic in MSSA, restricting the choice of antimicrobial therapy in all but the most straightforward cases. Alternatives to beta-lactams are needed for patients with hypersensitivity, or in complex circumstances, for example bone / joint infections, where combinations of antibiotics may be preferred. The current armoury of antimicrobials is becoming severely limited by increasing resistance, and judicious use is paramount in preserving the currently available antibiotics while awaiting the development of new compounds.

There is a constant trade-off to be made between the individual need of a patient and restricting the use of antimicrobials to preserve their utility for the greater good. It is accepted that, in severe sepsis, survival is significantly improved by the timely administration of an effective antimicrobial, with the consequence that, in many cases, broad spectrum agents are administered for several days until culture and drug susceptibility results are available (96, 97). Reducing the time taken to provide susceptibility results should therefore lessen the amount of broad spectrum antimicrobials administered. This reduces the individual patient's risk of acquiring a secondary, multidrug-resistant infection, and their exposure to multiple drugs or combinations with the risks of side effects. It also has the wider benefit that more prudent use of broad spectrum antimicrobials should result in a reduction (or at least a stabilisation) in resistance rates.

Current microbiology methods require samples to be cultured for at least 24 hours before an organism is identified, with further susceptibility testing taking a further 24-48 hours. A number of methods are used to test phenotypic susceptibilities, and, should there be inconsistencies, isolates may require further testing at a reference laboratory. All of this results in delays to the timely rationalisation of antimicrobial therapy.

1.5 Whole genome sequencing for investigating *S. aureus*

1.5.1 Whole genome sequencing as a single test platform

To investigate fully a case of disease due to *S. aureus*, a variety of platforms and techniques are required, which may result in delays to patient treatment and / or delays in implementing appropriate infection control measures. If there is concern regarding a transmission event or outbreak, particularly of MRSA, further testing to confirm the relatedness of the samples is undertaken at a reference laboratory. Suspicion of a specific virulence factor, such as toxic shock syndrome or scalded skin syndrome, may also lead to further testing. Detecting an outbreak or virulence factor requires clinical suspicion and, unless infection control surveillance is suitably robust, many instances of transmission may go undetected, delaying the opportunity to intervene and reduce the impact of a transmissible or virulent strain.

In theory, it should be possible to predict the phenotypic behaviour of an organism from its genetic sequence, as well as being able to examine genetic relatedness and the presence of specific genes of interest. In recent years, the advent of next-generation whole genome sequencing has made it possible to obtain the entire genome of an organism, reliably and in a relatively short period of time, bringing the possibility of interrogating whole genomes on a much larger scale than was previously possible.

1.5.2 Sequencing technology

DNA sequencing was pioneered by Sanger (98). The Sanger method uses DNA polymerase to sequence a complementary strand, with labelled nucleotides incorporated at random positions causing chain termination and separation of the resultant strands by molecular weight to deduce the sequence. Initially, radio-labelled nucleotides were used and the product run on a gel, whereas in modern use fluorescent labelling and capillary

separation have increased the speed and capacity significantly. However, Sanger sequencing remains relatively slow and expensive, so for routine use, only short fragments can be sequenced, for example, regions of a genome that have been identified and amplified by PCR. To sequence an entire genome, a shotgun method is employed whereby multiple, random areas of the genome are sequenced, and overlapping regions identified to assemble the entire genome. The expense and time taken has meant that, until recently, sequencing of entire bacterial genomes has been limited to a handful of reference strains.

The advent of next-generation sequencing has opened up whole genome sequencing for routine research use, as it becomes equally cheap and rapid to sequence the entire genome as to sequence a specific region of interest. A number of different sequencing platforms exist, with differences in chemistry, sequencing technique, capacity and turnaround time.

Illumina technology (http://www.illumina.com/applications/sequencing.html) uses a sequencing-by-synthesis method. DNA is extracted and purified, and a sequencing library is prepared by fragmentation of the DNA into 150bp segments. This is followed by repair of the sheared ends, adhesion to a flow cell, fragment tagging and then the generation of millions of identical clusters by PCR of the adhered fragments. The incorporated nucleotides in the complementary strand are fluorescently labelled, and, at each incorporation step, laser excitation causes phosphorescence of a different wavelength for each nucleotide added. This can then be decoded to deduce the sequence.

The size of the clusters minimises the error rate, as a majority call is generated for each cluster. The data generated include information about the cluster density so that the quality of each base read can be assessed. The output is multiple, overlapping reads of 150bp lengths, which are assessed for quality and assembled to allow interrogation of the genome.

Two methods are used to assemble the reads. The first is mapping to a reference, where the reads are aligned using a mapping programme such as Stampy (99) against a closed genome of a similar strain generated by Sanger sequencing to provide a scaffold for the reads, with assessment of quality for each site (i.e. depth of coverage and consistency of call). This generates multiple overlapping sequence fragments, usually with a depth of coverage of 50-100 reads at each site, allowing accurate identification of variants. The main drawback with reference-based mapping is that only sites present in the reference genome can be assessed - the "core" genome (common to all subtypes of a species) - and any mobile genetic elements (MGEs) that happen to be in the reference strain. Any other MGEs (phages, plasmids, insertional sequences) cannot be identified and are excluded from the mapping process, although they may be assembled separately, for example by mapping to a plasmid database.

Similarly, there must be a relatively high degree of homology between the reference and the mapped genome, as large degrees of divergence will result in less coverage and the potential for missing data. This may be difficult if no suitable reference strain exists, or if little is known about the origin / strain type of the query genome prior to assembly. It also means that species must be known prior to assembly. A further problem is dealing with repetitive regions, either within the same gene (such as the *spa* gene in *S. aureus*) or repetitive motifs which occur at multiple sites on the genome.

The alternative method is to assemble the genome de novo, i.e. without a reference genome, using software such as Velvet (100). This generates multiple sequence fragments (contigs) of varying lengths which can then be interrogated; however, it is computationally more complex and does not allow for full quality evaluation of variants. De novo assembly is able to assemble the accessory genome, but is also unable to assemble repetitive regions and, rather than producing a closed genome, results in up to 100 contigs which can then be examined for genes / regions of interest.

At present, both methods are generally required to assess both the core and accessory genomes. The development of sequencing platforms able to generate much longer reads (for example, the PacBio single molecule real-time sequencing platform, http://www.pacificbiosciences.com/products/smrt-technology/smrt-sequencing-advantage/) is likely to resolve some of the assembly issues in the near future, but at present these methods are too prohibitively expensive for routine use.

Current platforms require relatively large amounts of DNA for cluster generation (2-30ng), with the result that sequencing usually requires traditional culture of an organism in order to provide enough DNA for library preparation. Newer technologies able to sequence a single molecule more cheaply than the PacBio platform are being developed (101), bringing closer the prospect of direct sequencing from clinical samples. Coupled with the development of rapid benchtop sequencers which can sequence a bacterium in 27 hours, it becomes possible to obtain the entire genome sequence of an organism in a time frame comparable to the different modalities used for susceptibility testing, virulence profiling and typing, and to use the sequence to provide the same information given by those tests. This has the potential to transform current delivery of microbiology, infection control and public health services.

Timely information on resistance profile should ensure appropriate use of antimicrobials, and identifying previously unsuspected transmission events should allow prompt infection control interventions. This could change the current infection control approach from one of surveillance / clinical suspicion to active management, since a potential outbreak could be identified as soon as the first secondary case occurs, allowing rapid intervention and the prevention of onward spread.

1.6 Aim of thesis

The aim of this thesis is to apply recent advances in whole genome sequencing (WGS) to the investigation of *S. aureus*, to assess how the increased resolution of whole genome sequencing might be able to alter and / or improve the current standard outbreak investigation pathway, to establish whether WGS can lead to a better understanding of how *S. aureus* is transmitted, and to examine the potential for the use of WGS in determining antimicrobial susceptibilities.

2 Establishing within-host diversity at acquisition of *S. aureus* nasal carriage

2.1 Introduction

WGS permits resolution of sample variation down to the level of individual SNVs. In the outbreak setting, this means that rather than allocating strains to a particular category or type defined artificially on the basis of a small number of genes or loci (102), the decision on whether isolates are related or not is based on the genetic distances between them and whether the estimated time to their most recent common ancestor is consistent with a recent transmission event.

In the past, carriage of *S. aureus* has been assumed to be highly clonal, and, based on this assumption, study and screening protocols frequently examine only 1-2 colonies per sample (103). Similarly, in the outbreak setting, clinical laboratories routinely subculture only 1-5 colonies from a sample to process, store and / or submit to the reference laboratory.

However, several studies have shown that at a single time-point, there may be considerable diversity in the carried *S. aureus* population in an individual. In hospital studies, patients may be colonised by different strains at different sites (104) (105) due to multiple transmission events, or may acquire and lose different strains during their inpatient stay (106). More recent studies have found co-colonisation with multiple, distinct *spa*-types in community nasal carriage at a single time-point (103), and even where colonisation is with a single *spa*-type, there may be considerable genetic diversity within that population, with up to 40 SNVs between colonies identified (107).

This “cloud” of diversity, where an individual carries multiple clones of *S. aureus*, leads to concern about the use of WGS in the outbreak / transmission setting, as genetically diverse isolates could be falsely excluded from an outbreak when in fact they may have originated from the same source cloud. Although this is also true of standard typing

methods, where single colony sampling may miss transmission events if the donor population consists of more than 1 *spa*- or MLST sequence- type, it poses particular problems for the interpretation of WGS results as there may be a false assumption that WGS results are more reliable because of the higher resolution, even though within-host diversity may be of the same order of magnitude as within-*spa*-type diversity (108).

In addition, the diversity of the transmitted population will be affected by the proportions of each clone in the donor, the number of bacteria transmitted and the degree to which transmission is dependent on the fitness or otherwise of a particular clone. Mutations affecting adaptation to the nasal / skin environments, or causing resistance to antimicrobials, may promote preferential transmission or persistence of a specific variant. Host factors may also be relevant, for example, the increased *S. aureus* dispersal associated with upper respiratory tract infections (109, 110) or chronic skin conditions such as eczema or psoriasis, as well as poor host hygiene behaviours, particularly non-adherence to hand-washing protocols.

If onward transmission relies on the fitness of a single successful clone, outbreak isolates would be expected to be near identical. However, where transmission occurs from a host with a significant cloud of diversity, it is possible that different strains will be passed on to different recipients, giving rise to a degree of diversity across the outbreak which reflects the heterogeneity of the donor population. There is therefore concern that, in this context, there will be considerable diversity both within a recipient's population and between the populations from different recipients, which may result in falsely excluding cases and transmission links unless multiple samples from each case are sequenced (111, 112). This would increase costs significantly.

However, the *S. aureus* outbreaks investigated using WGS to date have been relatively clonal, with increased diversity observed only where there has been a hyper-mutator (113) or staff carrier (114). This leads to the hypothesis that colonisation begins with the acquisition of a minimally diverse population, followed by accumulation of diversity over

time. If this is the case, WGS in the outbreak setting becomes more feasible and multiple sampling may not be required.

To test this hypothesis, this chapter investigates the population diversity of *S. aureus* at acquisition and following long-term carriage, and discusses the implications for using WGS in the investigation of outbreaks.

2.2 Methods

2.2.1 Sampling frame

Individuals were identified from a long-term population carriage study of *S. aureus* nasal carriage in 571 adults attending general practices in Oxfordshire (115). Subjects had nasal swabs taken on entry to the study, at one and two months, and then every two months for a median of 2 years. Following collection, nasal swabs were initially incubated in 5% NaCl enrichment broth to maximise recovery of all strains present, before being inoculated on to *S. aureus* selective chromogenic agar (SASelect®, Bio-Rad, Limerick, Ireland). A meta-population from the SASelect® plate, taken by sweeping a swab across the plate, was stored in glycerol at -80ºC.

To investigate the difference in population diversity between acquisition and long-term carriage populations, study participants were identified who were negative for nasal carriage at recruitment and then for at least 3 consecutive swabs, before acquiring a single strain (identified by *spa* typing) and carrying it consistently for a minimum of one year (positive on at least 6 consecutive bi-monthly swabs).

Demographic data, *spa*-type and antibiotic use were obtained from the study database with the permission of the study authors.

2.2.2 Culture and DNA extraction

The first- and last-available positive samples for each individual were retrieved from the original stored meta-population. Samples were plated on Columbia blood agar (CBA) and incubated overnight at 37ºC. For each time-point, 8 individual colonies were selected and sub-cultured on to a further CBA plate and again incubated overnight at 37ºC. Twelve colonies were sampled from the first and last time-points from one individual (1218), plus a single colony from each of the intervening time-points.

DNA was extracted using the QuickGene commercial kit (Fujifilm, Tokyo, Japan) and sequenced on the Illumina HiSeq 2000 platform (San Diego, CA, USA), generating 150bp reads.

2.2.3 Assembly and genetic analysis

Reads were mapped to the MRSA252 reference strain (GenBank accession no. BX571856.1) using Stampy v1.0.18 (99). Assembled genomes were then aligned and SNVs called using SAMtools mpileup (116). Quality filtering parameters were set to include only SNVs supported by at least 5 reads, and SNVs occurring in known MRSA252 mobile genetic elements were excluded from the analysis. Maximum likelihood trees were constructed from the mapped genomes using PhyML (117) under a Jukes Cantor model. SNVs were annotated using Geneious 7.1.7 (http://www.geneious.com/).

To investigate variations in the accessory genome, reads were also de novo assembled using Velvet v1.0.18 (100), with quality parameters set to exclude samples if less than 50% of the genome was assembled in contigs of >1kb. The *de novo* assembled genomes were then interrogated using BLAST+ v2.2.28 blastn (118) to identify nucleotide sequences matching genes from a previously constructed query panel (119). Velvet assemblies were also analysed using PHAST (120) and PlasmidFinder 1.2 (121) to look for variation in phage and plasmid carriage.

2.2.4 Statistical analysis

Statistical analysis was performed using Stata v13.1 (http://www.stata.com). A pairwise distance matrix was constructed for each subject. Pairwise differences were compared between early and late time-points within each subject using the 2-group t-test. Time of acquisition was estimated by performing a linear regression analysis of root to tip distance against time from the most recent negative sample.

2.2.5 Estimation of molecular clock rate

To ascertain whether the diversity observed within a single host at acquisition and over time is consistent with current clock estimates, samples for each patient were analysed using BEAST v 1.8.1. (122). A Nexus alignment was generated from the alignment fasta file using Readseq (http://www.ebi.ac.uk/Tools/sfc/readseq/) and loaded in to BEAUti (http://beast.bio.ed.ac.uk/BEAUti) to set parameters. A simple HKY substitution model was used, with a strict molecular clock and coalescent model with a constant population size. Priors were set as follows: nucleotide frequencies: uniform prior; κ: log-normal prior with mean 1 and SD 1.25 on a logarithmic scale. Operators were set to auto-optimise. MCMC chain length was set to 10,000,000 with sampling every 1000 iterations and a burn in of 100,000 iterations, to obtain an effective sample size (ESS) of >200 for each iteration. Each sample was run in duplicate and the chains merged to obtain the final result. The output was inspected for quality and ESS using Tracer (http://beast.bio.ed.ac.uk/Tracer), annotated using TreeAnnotator (http://beast.bio.ed.ac.uk/treeannotator) and visualised using FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

For subjects 1218 and 1219, temporal signal was assessed using regression of root-to-tip distance against sample date using Path-O-Gen v1.4

(http://tree.bio.ed.ac.uk/software/pathogen/).

2.3 Results

2.3.1 Study subjects

8 subjects were identified with ≥ 3 consecutive negative nasal swabs followed by ≥ 1 year of consistently positive swabs. All isolates were MSSA, belonging to 7 *spa*-types and representative of 5 sequence types and 4 clonal complexes (table 2.1). Median time between first positive and most recent positive was 490 days (range 358-727). Two subjects received antibiotics in the six months preceding their first positive swab, and 1 subject had three courses of antibiotics during their carriage period (flucloxacillin, amoxicillin and ampicillin).

2.3.2 Analysis of diversity over time

In total, 135 isolates were successfully sequenced. One isolate (from 1219, early sample) failed quality checks due to contamination and was excluded from further analysis. In 6/8 subjects, there was a significant increase in mean pairwise diversity between the first and last samples (table 2.2, figure 2.1). In 1/8, a small increase was not significant: interestingly this subject had the highest diversity at their first sample. In the final subject, there was a small decrease in mean pairwise difference, which was not statistically significant. In this individual, there was also a change in *spa*-type from t091 to t9726. The BURP (Based Upon Repeat Pattern) distance between these *spa*-types is 5, considered closely related, and therefore this sample was included in the analysis. This was confirmed by the phylogenetic tree, which demonstrated that the core genomes were highly similar (figure 2.2).

TABLE 2.1 SAMPLE CHARACTERISTICS

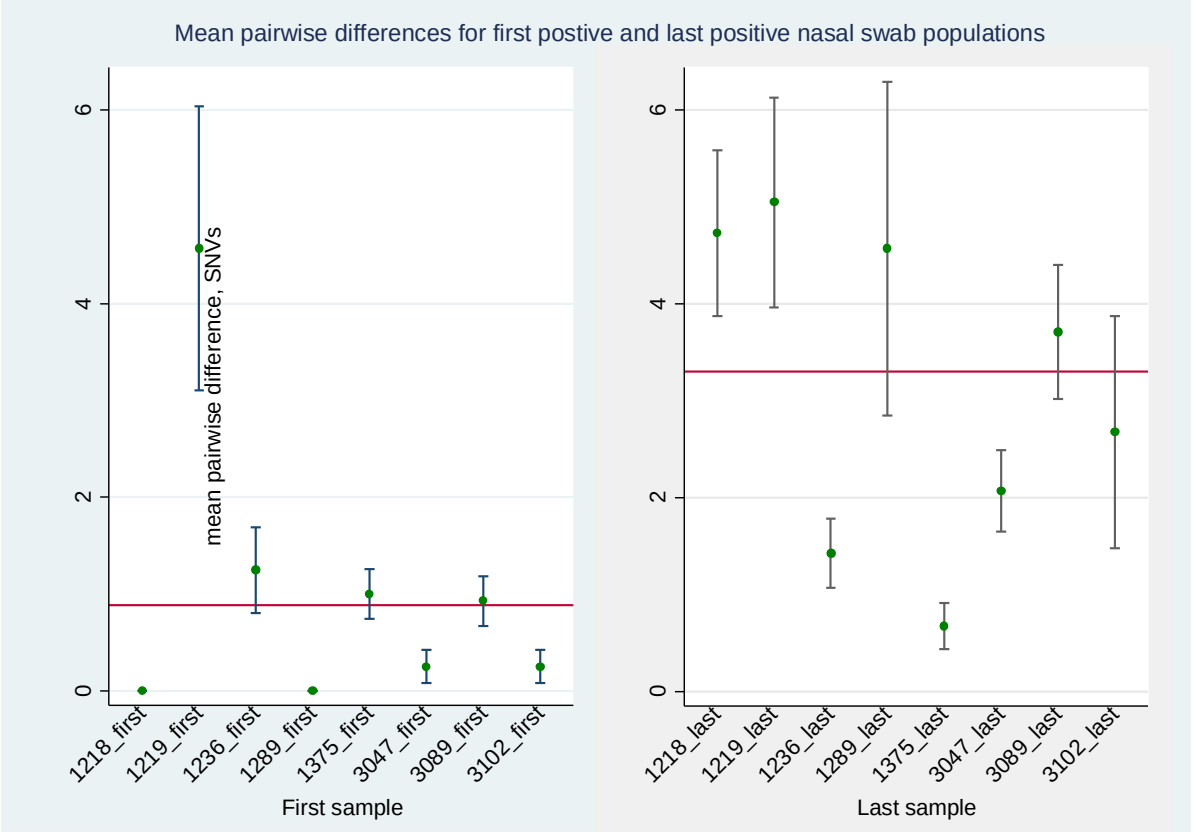
Subject	No of negative swabs prior to first positive	Months from enrolment to first positive swab	Days between first and most recent positive swabs	*spa*-type	MLST	Antibiotic use
1218	13	24	358	t012	30	None
1219	12	22	727	t012	30	Co-amoxiclav month 19
1236	13	26	543	t021	30	Penicillin month 20
1289	3	8	490	t7060	39	None
1375	6	8	483	t091 → t9726*	7	None
3047	4	8	360	t189	188	Flucloxacillin month 10, amoxicillin month 16, ampicillin month 18
3089	5	10	421	t015	45	None
3102	3	4	610	t382	39	None

*BURP distance 5, considered to be closely related (123, 124)

TABLE 2.2 MEAN PAIRWISE DIFFERENCES BETWEEN FIRST AND LAST SAMPLES

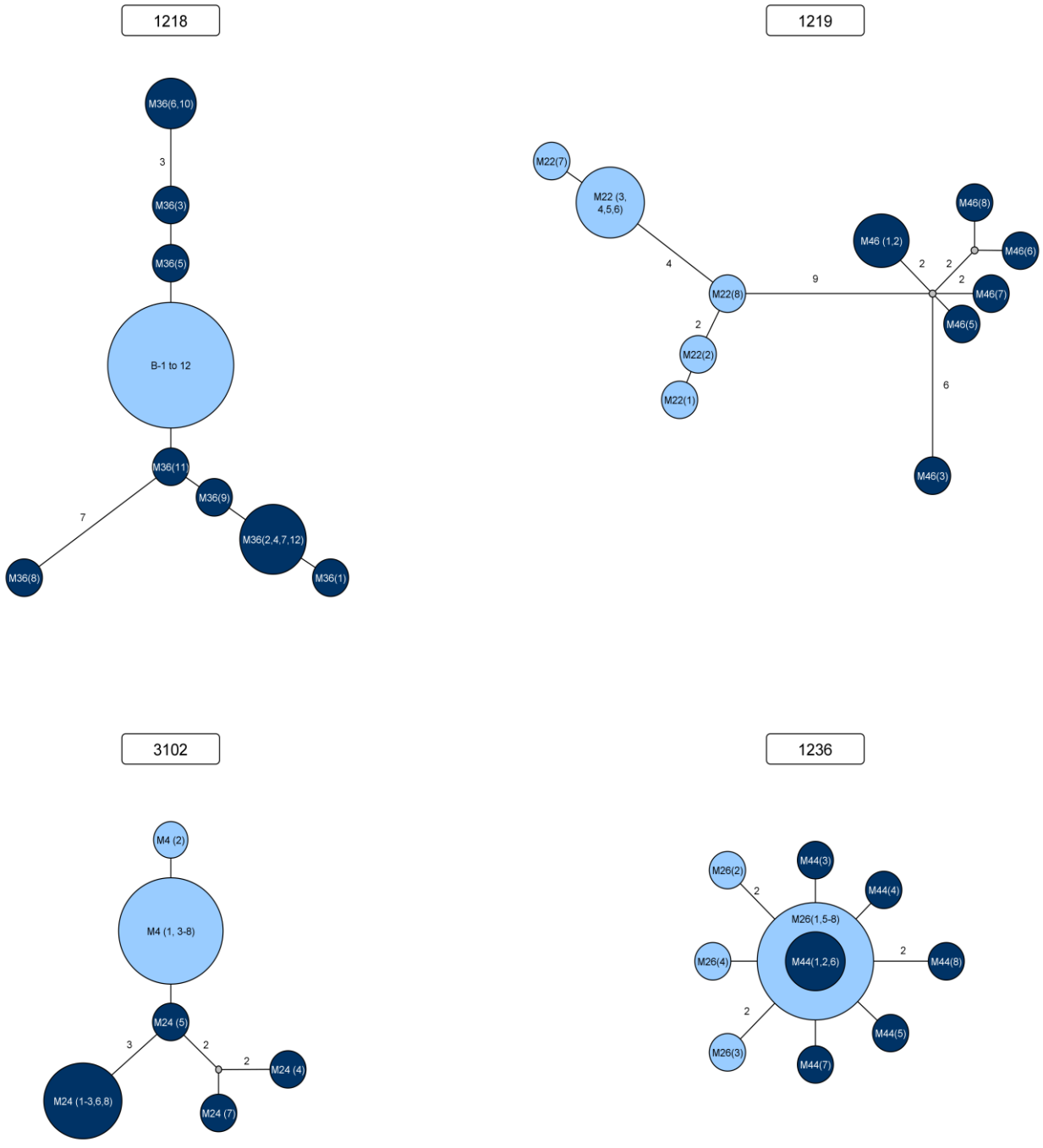
Subject	Mean pairwise difference, SNVs (95% CI)		P value
	First sample	Last sample	
1218 (B)	0	4.73 (3.87-5.58)	<0.0001
1219 (A)	4.57 (3.10-6.04)	8.89 (6.14-11.65)	0.0063
1236	1.25 (0.81-1.69)	1.43 (1.07-1.79)	0.5222
1289	0	4.57 (2.85-6.29)	<0.0001
1375	1 (0.74-1.26)	0.68 (0.44-0.92)	0.0656
3047	0.25 (0.08-0.42)	2.07 (1.65-2.49)	<0.0001
3089	0.93 (0.67-1.19)	3.71 (3.02-4.41)	<0.0001
3102	0.25 (0.08-0.42)	2.68 (1.48-3.88)	<0.0001

FIGURE 2.1 MEAN PAIRWISE DIFFERENCES FOR FIRST POSITIVE AND LAST POSITIVE NASAL SWAB POPULATIONS



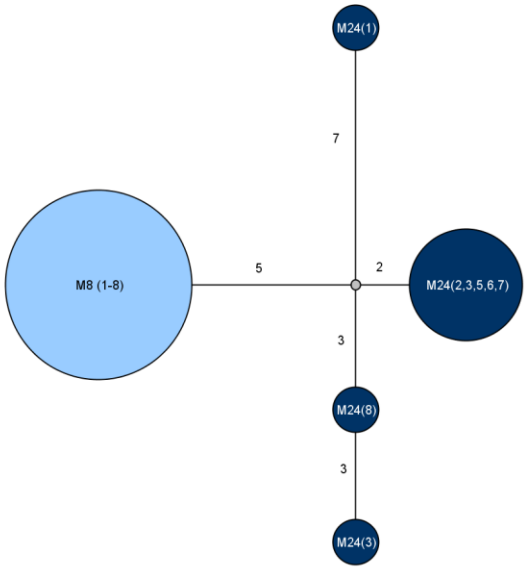
Horizontal red lines indicate overall mean.

FIGURE 2.2 PHYLOGENETIC TREES SHOWING SINGLE COLONY PICKS FROM EARLY (LIGHT BLUE) AND LATE (DARK BLUE) SAMPLES

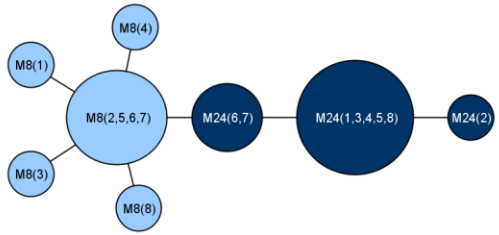


M12(1): month 12, colony 1.

1289

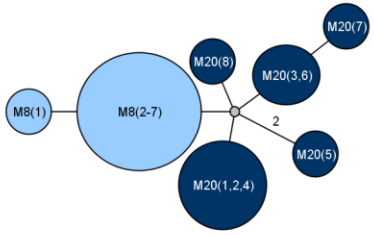


1375

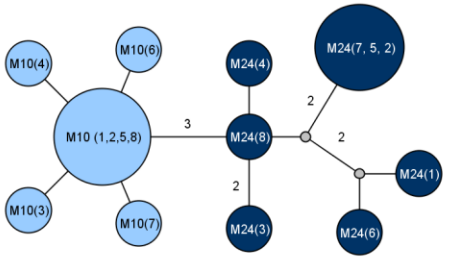


t091 t9726

3047



3089



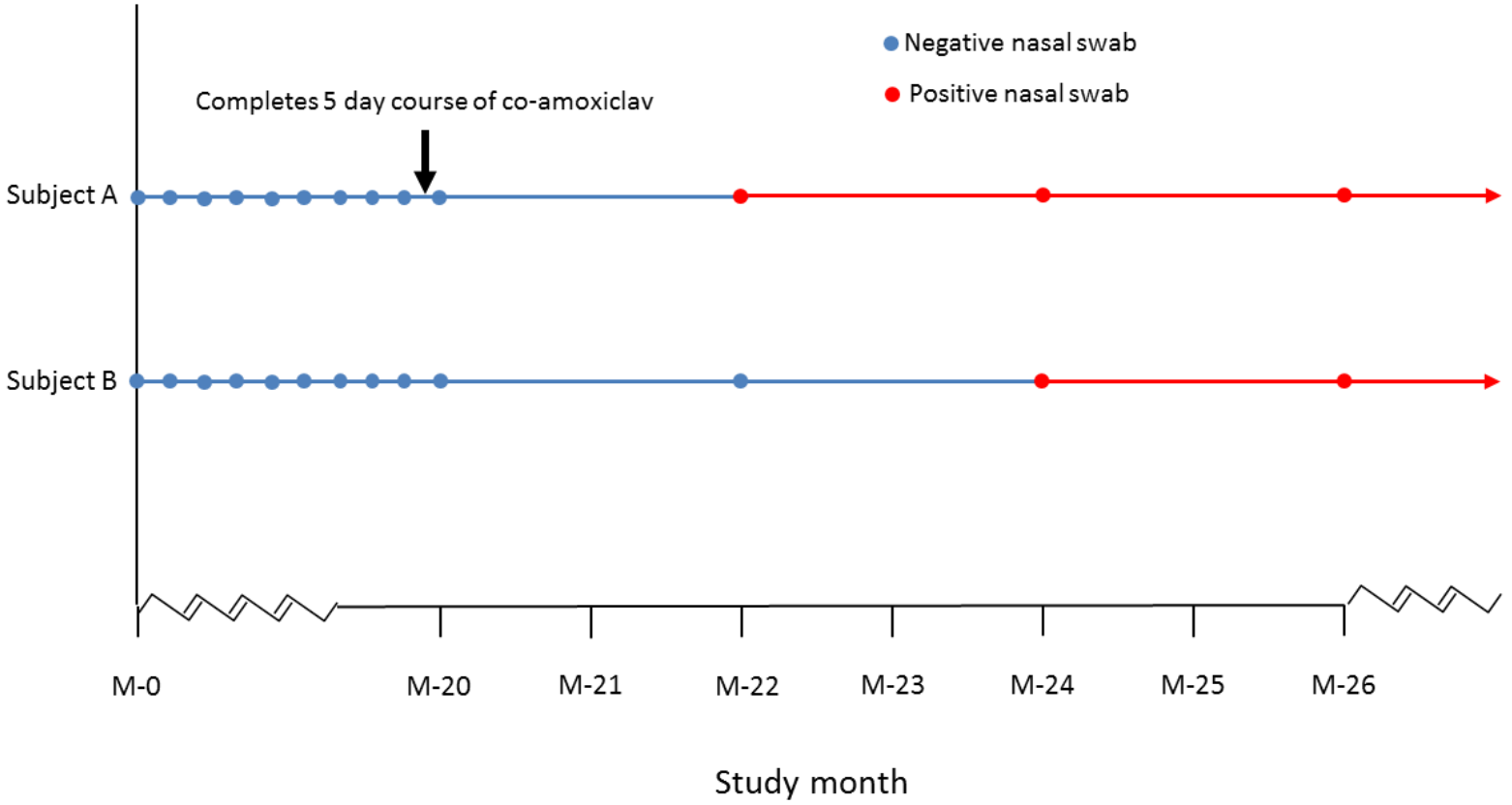
Unrooted phylogenetic trees for all subjects are shown in figure 2.2, and demonstrate a high degree of clonality in the early population in the majority of subjects, with a single strain observed in 2 subjects. In the 7 subjects where diversity increased, tree branching occurred in several directions, indicating the emergence of several clones. In subject 1375, diversity decreased (although not significantly) and all later isolates occurred along the same branch, possibly indicating the emergence of a new, dominant strain with a different *spa*-type. Of the 5 SNVs occurring in the late vs the early samples in this subject, 2 were in intergenic regions, 1 was synonymous, 1 occurred in a hypothetical protein and 1 resulted in a proline to serine substitution in a putative transferase gene. None of these mutations would be anticipated to have a significant evolutionary advantage. In subject 1236, 3 later isolates were identical to the majority population from the early sample, indicating long-term persistence of this clone.

Overall, there was a mean increase in pairwise distance from 0.88 SNVs (95%CI 0.65-1.11) to 3.30 (95%CI 2.92-3.68) between the first and last samples ($p<0.0001$)

2.3.3 Identification of donor-recipient pair

Two of the subjects described above, 1218 (hereafter subject B) and 1219 (hereafter subject A), shared the same surname and address and were thus presumed to be household contacts. Subject A had 10 negative swabs before receiving a course of co-amoxiclav in month 19. The last antibiotic dose was taken the day before the 11th swab, which was negative. However, the 12th swab, 2 months later, was positive for MSSA (*spa*-type t012) and the participant continued to return positive swabs, all of which were *spa*-type t012, for a further 2 years. Subject B had 13 negative swabs and became positive 2 months after 1219, also with *spa*-type t012, which was consistently carried for 12 months (figure 2.3).

FIGURE 2.3 TIMELINE SHOWING ACQUISITION OF MSSA, SPA-TYPE T012 BY 2 MEMBERS OF THE SAME HOUSEHOLD

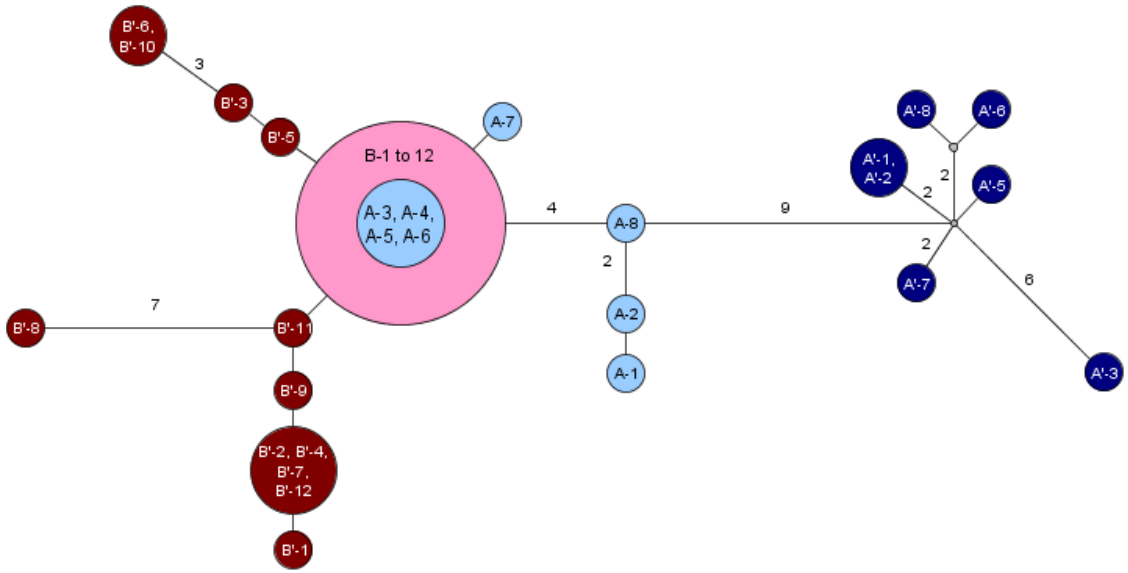


To assess whether B had acquired the strain from A, and to determine if there was any evidence of ongoing transmission, isolates from these two subjects were aligned to reference genome MRSA252 as above, and a PhyML tree constructed (figure 2.4). Further single colonies were taken from B for each intervening time-point up until loss of the strain at month 46: these are plotted in figure 2.5.

Analysis of the phylogeny revealed that subject B had an entirely clonal early population which overlapped with the dominant strain in the first sample from A. The late samples from both individuals diverged into separate branches of the tree, with no overlapping clades. Although only a single colony was sequenced for each of the interim time-points between B's early and late samples, these all occurred within B's clade and there was no overlap with any of the later sequences from subject A.

To determine whether the data supported acquisition by patient B within the time-frame estimated, a phylogenetic tree for B's isolates was constructed, including unrelated t012 isolates obtained from local isolate collections to provide an outgroup. Root to tip distances were calculated and a linear regression performed against time of sampling measured from date of last negative swab (figure 2.6). The confidence interval is in keeping with acquisition after the time of the negative swab.

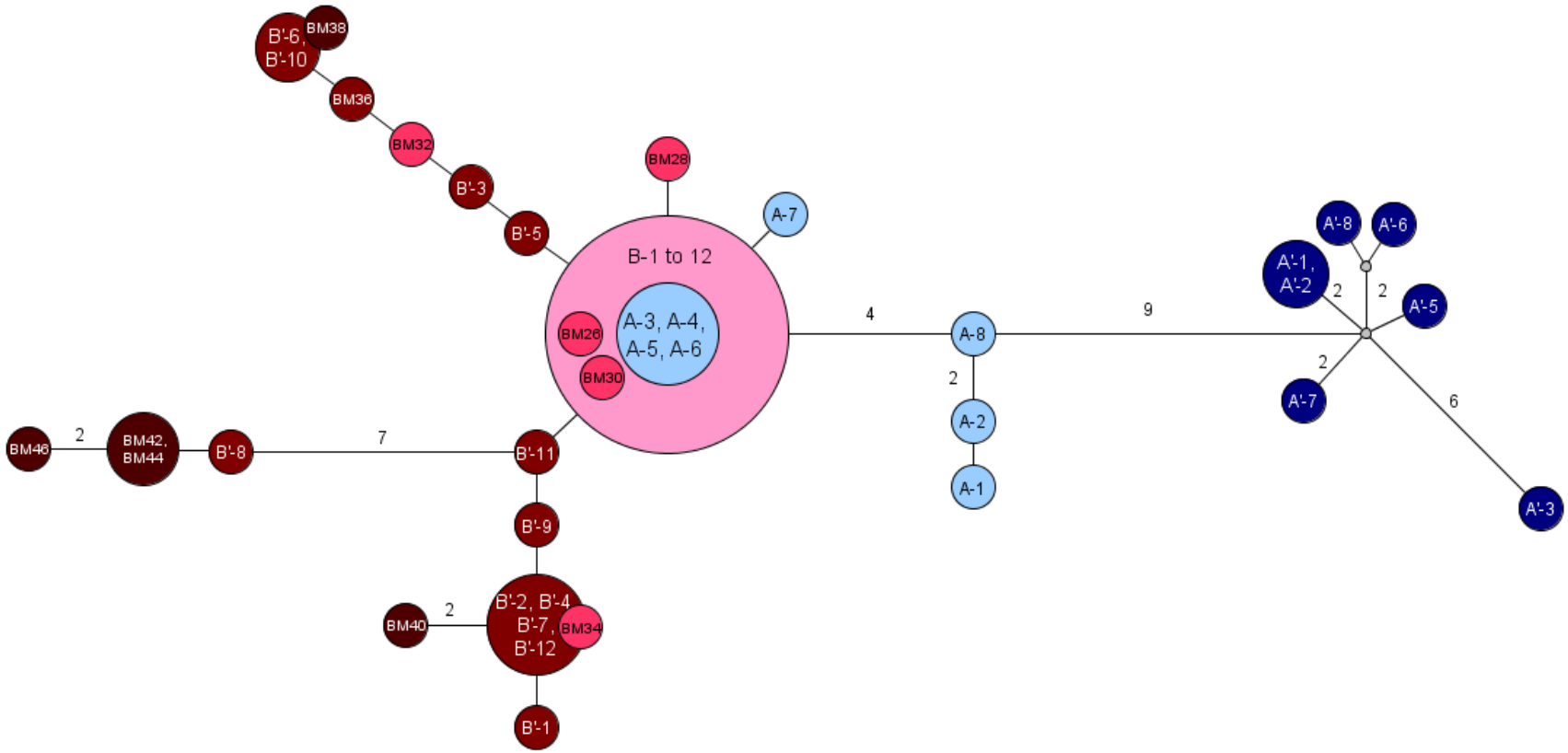
FIGURE 2.4 UNROOTED PHYML TREE SHOWING EARLY AND LATE SAMPLES FROM HOUSEHOLD SUBJECTS A AND B



Pale blue: early samples, A; dark blue: late samples, A; pale pink: early samples, B; dark red: late samples, B.

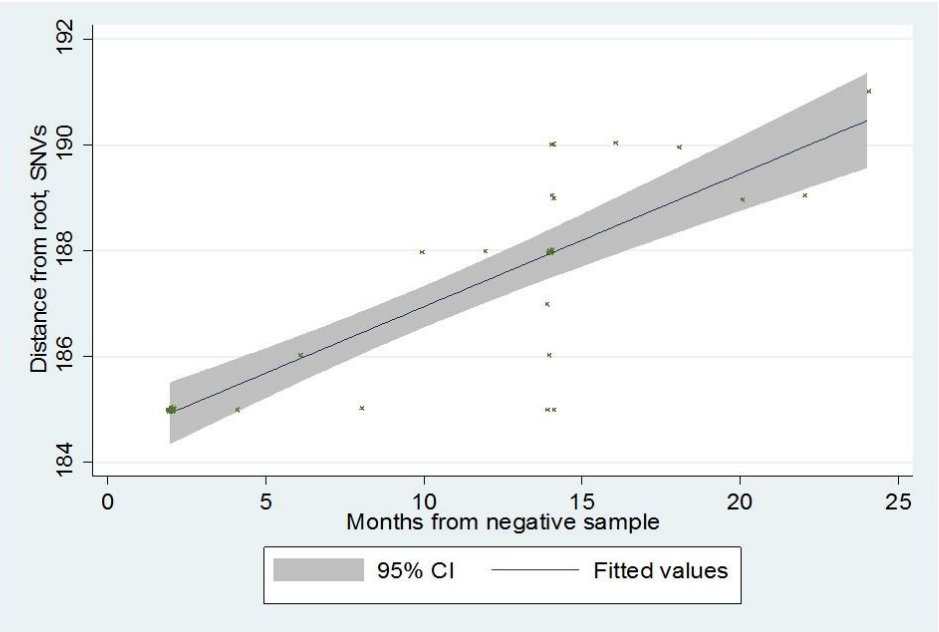
Numbers in circles indicate single colony pick number. Branch lengths are given in SNVs.

FIGURE 2.5 PHYML TREE AS IN FIGURE 4, SHOWING SINGLE PICKS FOR ALL AVAILABLE SAMPLES FOR SUBJECT B



BM26: subject B, month 26. Rose pink circles indicate single colony samples from subject B taken between the early (pale pink) and late (dark red) samples chosen for initial study; burgundy circles are samples taken after the late sample.

FIGURE 2.6 LINEAR REGRESSION PLOT OF ROOT TO TIP DISTANCES OF ISOLATES FROM SUBJECT B AGAINST TIME FROM LAST NEGATIVE SWAB (R2=0.71)



12 colonies were sequenced for time-points 62 days and 427 days. A single colony was sequenced for all other time-points.

2.3.4 SNV analysis

Across all study subjects, there was a total of 96 SNVs between early and late isolates. 30 (31%) were intergenic and 19 (14%) occurred in hypothetical proteins. 47 SNVs occurred in known functional regions: 11 of these were synonymous and 36 were non-synonymous (table 2.3).

SNVs were distributed throughout the genome, and, excluding known mobile elements in the reference genome, there was no significant clustering of SNVs in variable regions (figure 2.7), no SNV occurred more than once, and no gene had SNVs in different individuals, supporting natural genetic drift rather than recombination events.

2.3.5 Resistance / virulence profiling

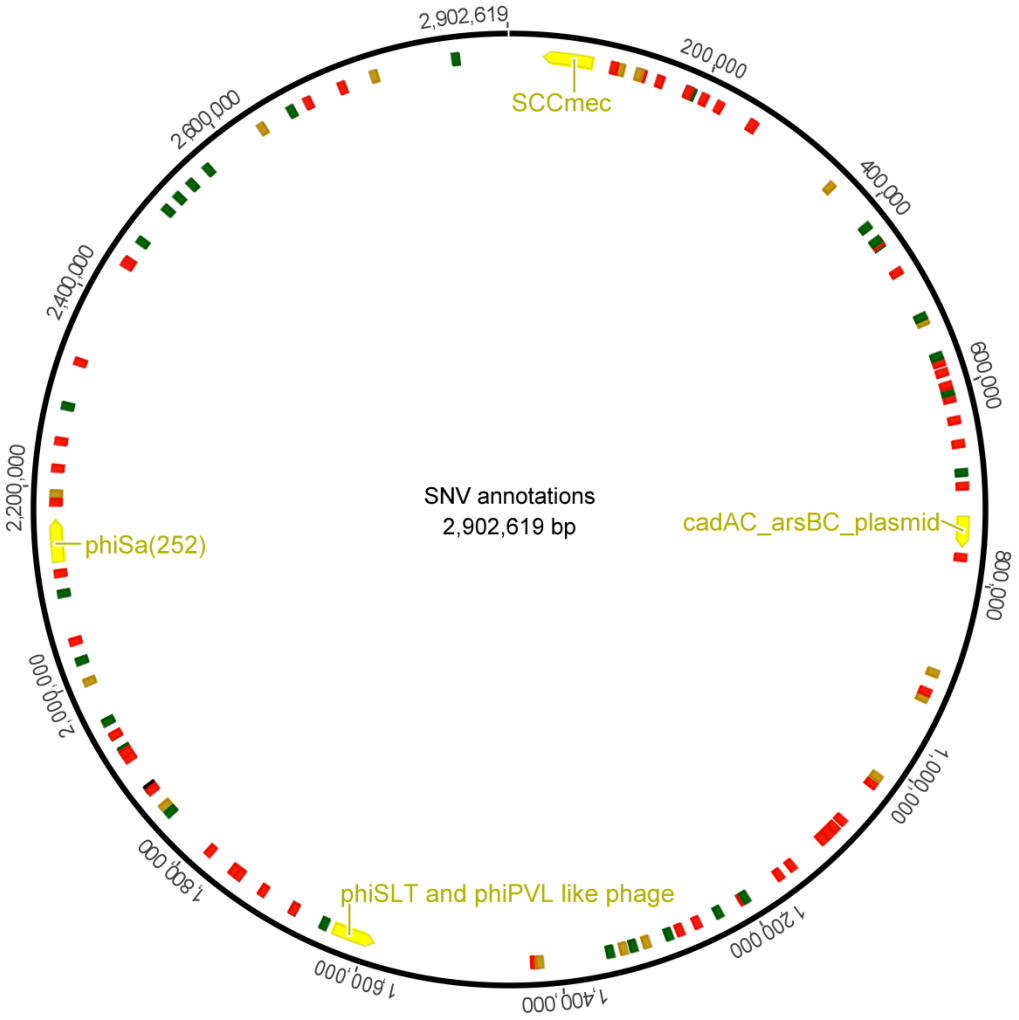
All isolates were MSSA. Out of 135 sequenced colonies, all but one carried *blaZ* (coding for penicillinase), reflecting the near universal prevalence of penicillin resistance (125). No other resistance genes were present. One subject received three antibiotics courses during the period of carriage, including a course of flucloxacillin; however, their carriage status was unaffected even though the strain was methicillin susceptible.

On virulence gene profiling, the isolates from 6/8 subjects had consistent profiles over time across all early and late isolates. Isolates from 3 subjects (1218, 1219 and 1289) harboured the toxic shock syndrome toxin 1 (TSST1) gene, a superantigen responsible for the majority of cases of staphylococcal toxic shock syndrome which results in multi-organ failure (126). The gene was consistently present in all isolates from these individuals. Isolates from 1 subject (1289) were positive for the Panton-Valentine leucocidin toxin, which is associated with skin and soft tissue infection (127). Again this was carried consistently in all isolates.

TABLE 2.3 POSITION AND FUNCTION OF NON-SYNONYMOUS SNVS BETWEEN EARLY AND LATE POPULATIONS

Subject ID	SNV position (relative to MRSA 252)	No. of strains with SNV	Base substitution	Gene	Function	Residue substitution
1218	2185658	2	A->G	*agr*C	Signal dectecting component of the agr autoinducer peptide-quorum sensing system	287:N=>S
	2248213	4	G->A	SAR2177	Similar to cardiolipin synthetase in bacillus sp.	333:G=>S
	2445919	1	A->G	*ure*D	Urease accessory protein	44:T=>A
1219	223765	1	C->T	SAR0196	Type 1 restriction enzyme	44: R=>C
	260948	7	C->T	*fad*A	Thiolase	346:G=>S
	575458	1	G->A	*clp*C	Stress response-related Clp ATPase	521:S=>N
	613842	2	C->T	SAR0562	Deoxyadenosine kinase	7:A=>T
	1102664	7	A->G	*pts*H	Phosphocarrier protein HPr	13:I=>V
	1205176	1	A->G	SAR1160	Cell division protein	246:S=>G
	1785741	7	C->T	*que*A	S-adenosylmethionine--tRNA ribosyltransferase-isomerase	337:S=>T
	2442881	1	A->C	*ure*C	Urease subunit alpha	194:N=>T
	2692930	1	C->T	*srt*A	Sortase	61:A=>T
	2731558	6	T->C	*crt*N	Squalene synthase	101:D=>G
1236	262024	1	G->A	*fad*B	Fatty oxidation complex protein	752:R=>C
	1427101	1	A->G	*mpr*F	Probable transmembrane protein	19:T=>A
	2331187	2	T->A	*roc*F	Arginase	115:K:I
1289	157974	8	A->G	SAR0144	ABC transporter ATP-binding protein	237:V=>A
	660600	8	C->A	SAR0611	Phophohydrolase	287:P=>H
	1085826	1	A->G	*pur*C	Phosphoribosylaminoimidazole-succinocarboxamide synthase	137:M=>V
	1256204	1	A->G	*fab*D	Malonyl CoA-ACP transacylase	136:D=>G
	1275020	1	G->A	SAR1222	Succinyl-CoA synthetase subunit alpha	246:G=>R
	1755053	1	G->A	SAR1693	O-methyltransferase	192:P=>S
	1918625	8	T->C	SAR1836	Dipeptidase PepV	262:T=>A
1375	205791	6	C->T	SAR0181	4'-phosphopantetheinyl transferase	71:P=>S
3047	138578	3	T->A	SAR0124	Siderophore biosynthesis protein	571:I=>N
	775766	1	C->G	SAR0740	Cobalamin synthesis protein	262:A=>G
3089	187673	2	C->T	SAR0169	Aldehyde dehydrogenase	353:A=>V
	599222	1	C->A	*rps*G	30S ribosomal protein S7	38:T=>K
	635717	8	A->C	SAR0581	Ketoacyl-CoA thiolase	153:I=>L
	917246	5	A->G	SAR0879	NifU-like protein	55:I=>V
	1144032	1	C->A	SAR1097	Methylase	159:P=>S
	1159315	3	C->T	*phe*T	Phenylalanyl-tRNA synthetase subunit beta	565:A=>V
	1875793	1	G->A	SAR1801	D-3-phosphoglycerate dehydrogenase	stop codon
	1910724	3	G->T	SAR1827	Transposase	14:A=>D
3102	1093746	1	A->G	*pur*H	Bifunctional phosphoribosylaminoimidazolecarboxamide formyltransferase / IMP cyclohydrolase	480:V=>I
	2221373	1	C->A	SAR2151	RNA binding protein	325:I=>R

FIGURE 2.7 DISTRIBUTION OF SNVS PLOTTED AGAINST MRSA 252 REFERENCE GENOME



Prophages, excluded from comparative SNV analysis, are shown in yellow. Red indicates non-synonymous SNV in coding region, green indicates synonymous SNV, brown indicates SNV in intergenic region.

There was variation in the presence / absence of virulence genes *sak*, *chp* and *scn* in subjects 1218(B) and 1219(A). These genes form part of the immune evasion cluster (IEC) present in most human isolates and are usually located on β-haemolysin-converting or φSa3 family bacteriophages (128). In these isolates, the presence / absence of IEC elements corresponded with the presence / absence of a φSa3-like phage identified using PHAST.

2.3.6 Analysis of molecular clock using BEAST

Under a strict molecular clock model, the median rate of substitutions per site per year for all subjects was estimated to be 1.16×10^{-6} (IQR 9.05×10^{-7}-2.08×10^{-6}), or 3.37 substitutions per genome per year (IQR 2.63-6.03). Estimates for each individual subject are given in table 2.4.

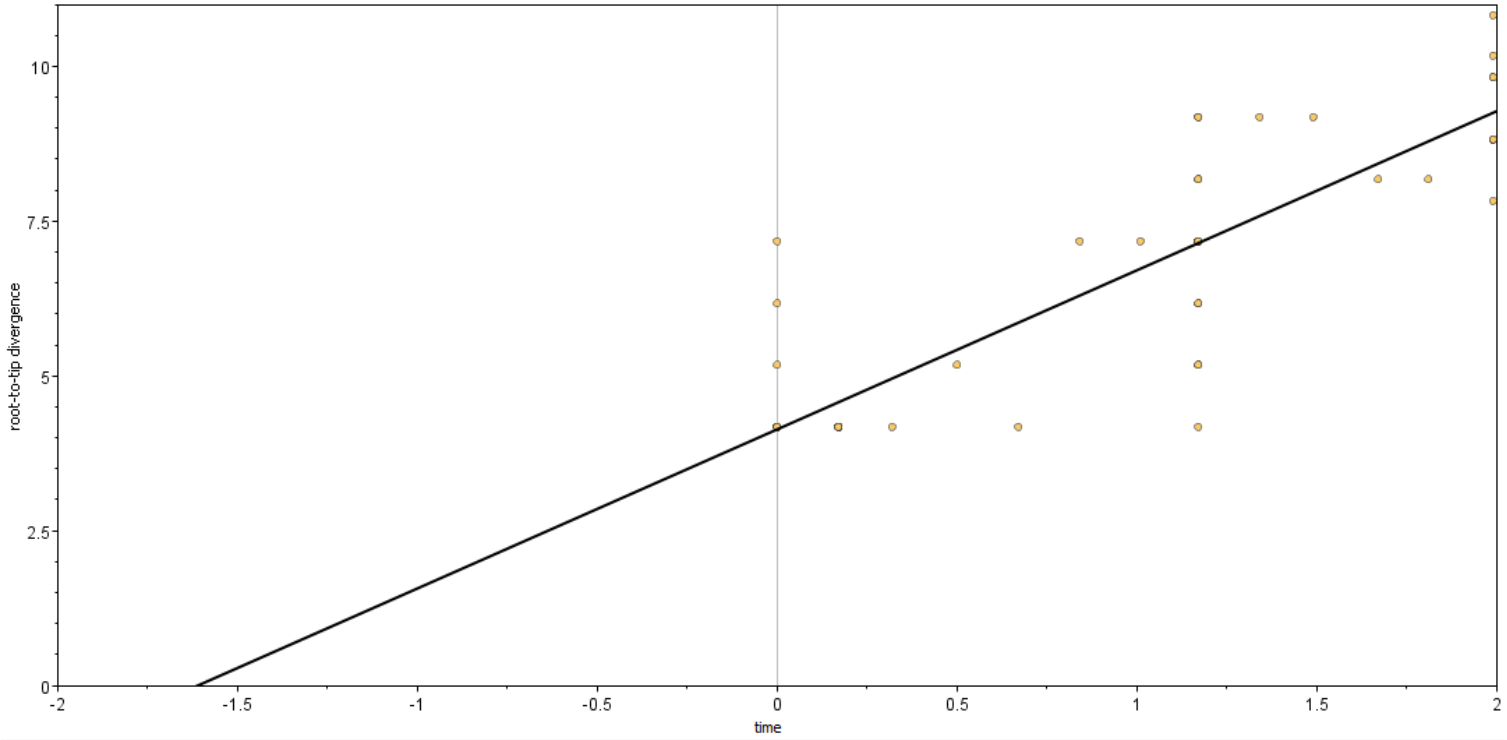
2.3.7 Donor-recipient household pair

To investigate the transmission dynamics between the household pair A and B, root-to-tip divergence was estimated from the PhyML tree and assessed for a temporal signal using Path-O-Gen. A strong temporal signal was observed (figure 2.8), with a positive correlation coefficient (r=0.86).

TABLE 2.4 SUBSTITUTION RATES MODELLED BY BEAST, STRICT MOLECULAR CLOCK

Subject	Estimated clock rate (substitutions per site per year)	95% Bayesian credible interval	Substitutions per genome per year
1218	9.34×10^{-7}	4.53×10^{-7} - 1.46×10^{-6}	2.71
1219	2.91×10^{-6}	1.42×10^{-6} - 4.54×10^{-6}	8.45
1236	3.15×10^{-7}	1.41×10^{-9} - 7.65×10^{-7}	0.91
1289	2.42×10^{-6}	1.15×10^{-6} - 3.93×10^{-6}	7.02
1375	1.24×10^{-6}	3.13×10^{-7} - 2.29×10^{-6}	3.60
3047	1.07×10^{-6}	2.36×10^{-7} - 2.10×10^{-6}	3.11
3089	1.74×10^{-6}	5.84×10^{-7} - 3.00×10^{-6}	5.05
3102	8.76×10^{-7}	2.86×10^{-7} - 1.55×10^{-6}	2.54

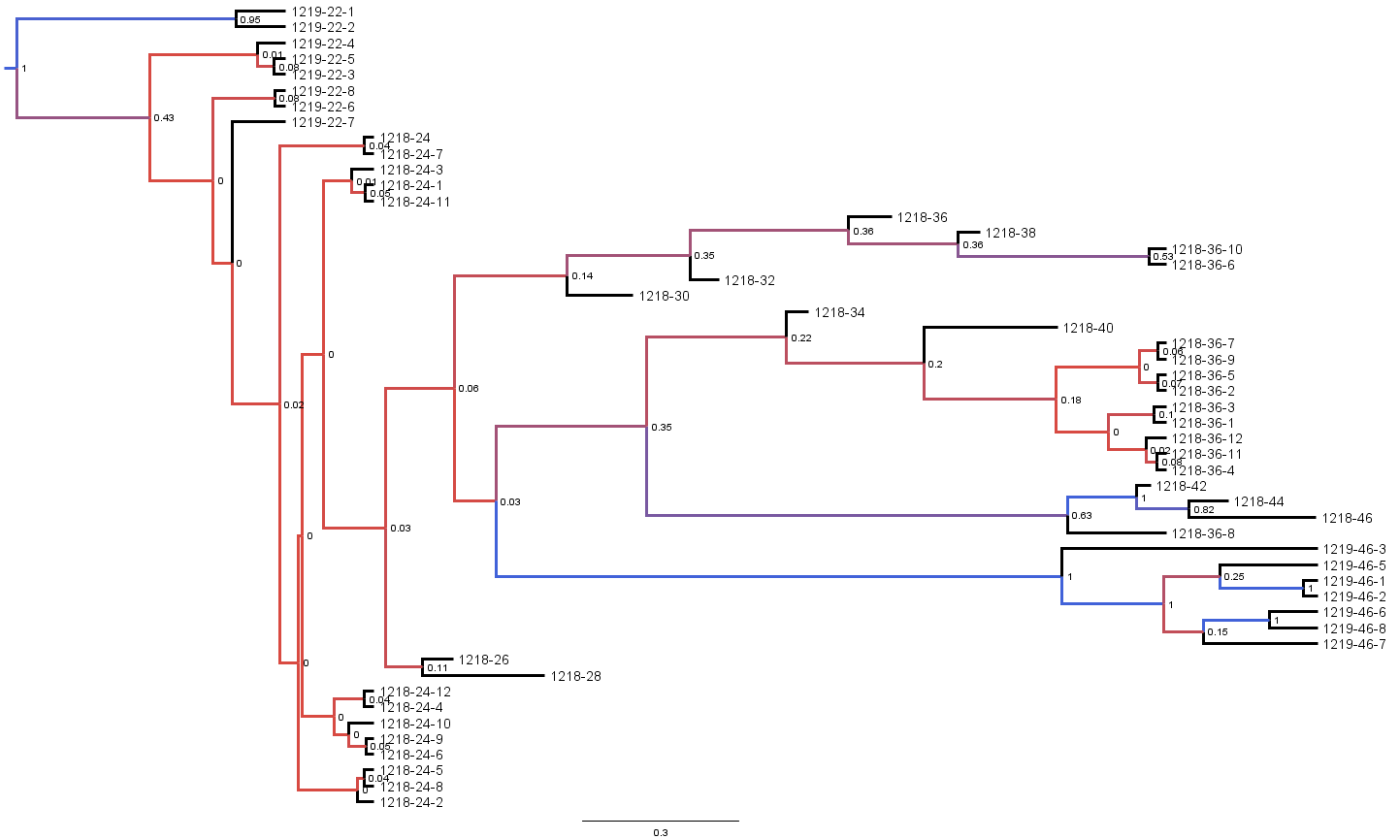
FIGURE 2.8 REGRESSION OF ROOT-TO-TIP DISTANCES AGAINST TIME OF SAMPLE (IN YEARS FROM EARLIEST SAMPLE) FOR ALL A AND B ISOLATES



Using BEAST to estimate a maximum clade credibility tree for the household pair gave a substitution rate of 1.19×10^{-6} per site per year (95% Bayesian credible interval 7.02×10^{-7} – 1.70×10^{-6}), or 3.46 SNVs per genome per year. The BEAST reconstruction is shown in figure 9. This model raises the possibility that after transmission to B, subject A is subsequently re-infected by a clade from B, which is not in keeping with the PhyML estimated phylogeny. However, the posterior support for the late A branch arising from a B clade is very low (0.03). On manual inspection of the SNVs, none of A's SNVs are shared by later clades from B and the overall likelihood of the BEAST reconstruction reflecting actual events is low.

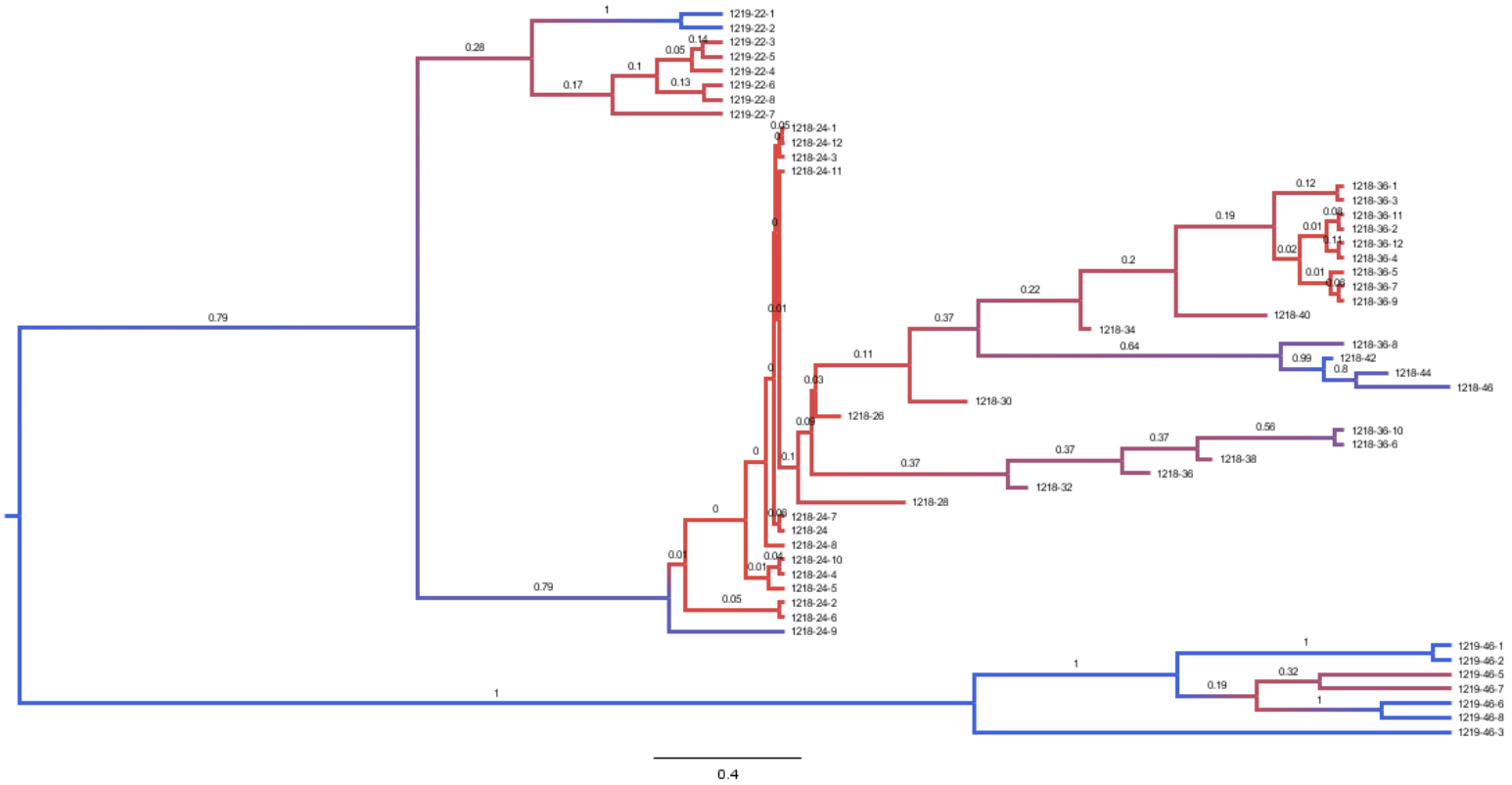
To address this problem, the analysis was repeated using the recently developed BASTA (BAyesian STructured coalescent Approximation) package (129). Subjects A and B were designated as separate demes in order to increase the prior probability that the later populations from the 2 subjects are independent. The results are shown in figure 2.10. The model gives a high posterior probability that after acquisition subject B's population evolved independently of subject A, and does not support the hypothesis that there was re-infection of subject B.

FIGURE 2.9 MAXIMUM CLADE CREDIBILITY TREE FOR SUBJECTS A AND B



Branches are coloured according to posterior probability. Branches with a high posterior probability (>0.9) are blue / black, those with a low posterior probability (<0.3) are red.

FIGURE 2.10 MAXIMUM CLADE CREDIBILITY TREE FOR SUBJECTS A AND B, MODELLED USING BASTA



Branches are coloured according to posterior probability (high: blue/black, low: red).

2.4 Discussion

2.4.1 Accumulation of diversity over time

These results support the hypothesis that in newly acquired nasal carriage, there is lower diversity than in long-term established carriage. In 6/8 individuals, there was a significant increase in pairwise diversity over time. In 2 subjects, the initial populations were entirely clonal and in one there was a single isolate 1 SNV distant from the other clonal isolates. One of the patients (1219) had relatively high diversity in their early sample compared with the other subjects. Interestingly, this patient had received a course of co-amoxiclav just prior to their preceding negative swab. It is therefore possible that this was a false negative, with suppression due to the anti-microbial and that they had in fact acquired carriage in the 4 months prior to the first positive. This may explain the relatively increased diversity observed.

Although sampling more than 8 colonies may have identified greater diversity as modelled by Paterson and colleagues (112), this would be true of the later samples and thus the overall conclusion would be expected to be the same.

2.4.2 Comparison with previous studies

Several studies have demonstrated the diversity in the *S. aureus* population carried by an individual, with up to 40 SNVs reported between single colony picks from a primary culture plate (107). Young et al (130) demonstrated the evolution of a *S. aureus* strain in a single patient over a 2 year period, with an accumulation of SNVs including a nonsense mutation causing a premature stop codon which may have contributed to the apparent increase in pathogenicity of the strain. Their study indicates that there is considerable flux in the population structure over time, with accumulation of mutations and several evolutionary “dead ends”, where branches of the tree have failed to persist in later sampling. These are presumably due to purifying selection in which deleterious mutations are removed from

the population, while beneficial mutations move to fixation, resulting in a population homogeneous for the advantageous mutation(s) (131). Synonymous mutations occurring in the genotype with the favourable mutations will also persevere in the emerging population. It might therefore be expected that an initially diverse population will move towards homogeneity over time as a result of purifying selection.

Previous studies have also assumed that the founder population in a transmission recipient reflects the diversity of the donor population. In a study by Tong and colleagues (132), high levels of MRSA transmission in a limited resource setting were found to be linked to the admission of a single index patient. This patient was positive on admission for MRSA of a unique, new clade within the locally prevalent ST239 complex, and prospective sampling of patients on the same ward demonstrated the emergence of this clade as the dominant MRSA strain. Multiple samples taken throughout the index case's stay were sequenced, including multiple colonies from 4 time-points. These were found to overlap with several of the recipients' strains, indicating likely transmission links. The authors argue that the diversity in the founder population, combined with an absence of purifying selection as modelled by Rocha (131), resulted in a high level of diversity in the recipients.

Similarly, Harris and colleagues described an outbreak on a neonatal unit, where the diversity observed in the outbreak cases was attributed to long-term nasal carriage in a staff member (114). Despite strong epidemiological evidence for an outbreak, isolates were relatively distant compared to previous outbreak study in the same unit (113). This was resolved by sequencing multiple colonies from the putative staff source, who was found to have a diverse cloud with up to 10 SNVs between strains. It was concluded that the staff member was the driver of the outbreak, causing repeated introductions of the strain despite intervening periods where no colonised patient was on the ward.

This problem of the "cloud" has also been highlighted by Worby and colleagues (111), who are concerned that studies using WGS data have assumed, on the basis of minimal

evidence, that the transmission network is topologically similar to estimated phylogenetic trees. Since these are predominantly based on sequencing single isolates from each case and do not take into account within host diversity and heterogeneous transmission, the conclusion may be flawed. In modelling studies, they demonstrate that, in the scenario where each host has a “cloud” of diversity, using a single isolate for each infected person (defined as infrequent sampling) results in reconstructions equivalent to models which select sources at random. Even with near perfect sampling, the models still result in too much uncertainty to identify individual transmission routes.

Most recently, Paterson and colleagues (112) expressed concern over the use of single isolates in outbreak / transmission studies. In their study, deep sequencing of animal cases and staff within a veterinary hospital outbreak identified a significant within-host cloud, with variation in diversity over time.

Paterson’s concern echoes the collective conclusion from the studies outlined: WGS may lack utility in the outbreak setting, as results are likely to be confounded by within host diversity, and to overcome this, it may be necessary to perform multiple sampling and sequencing from individuals, with 20 colonies being the upper limit used at a single time-point in the studies above.

Since routine practice in most clinical laboratories is to subculture and send only 1-5 colonies from each sample to the reference laboratory, robust population analysis of previous outbreaks becomes impossible, and laboratories would need to change practice to allow sequencing to be used to maximal potential for future investigations. Even more importantly, current bioinformatic analysis tools are optimised for single colony sequencing, so it is clear that the requirement for multiple strains may rapidly escalate costs (up to x20 for each outbreak if all cases require multiple sequencing), taking WGS back to the realms of research-only use.

However, the results from my study dispel some of this pessimism. In the cases studied, there is minimal diversity (and sometimes a single clone) observed within the founder population. In the majority of cases, diversity increases rather than decreases with long-term carriage, and there is little evidence for the effect of purifying selection.

The data from my study would also be in keeping with a small transmission bottleneck size as modelled by Worby, where the expected pairwise difference after a transmission event is <2 for bottleneck size <10 and mutation rate $\mu = 0.0001$, or bottleneck size <2 and $\mu = 0.0005$. These estimates are similar to the pairwise differences seen in the early samples.

Also, in Worby's model, the bottleneck is imposed after 1000 generations. Using an estimated bacterial generation time of 30mins, this equates to a single transmission event after approximately 21 days, which is a relatively long time in the outbreak context. It also assumes a single chain of transmission and is repeated consecutively for 25 transmission events, which equates to transmission over prolonged period (i.e. over 500 days). This is very different from the acute outbreak scenario, where multiple transmission events take place over a short period of time. In this case, there would therefore be insufficient time for each new host to accumulate diversity, increasing the likelihood of a relatively homogeneous population at each new transmission. Overall, in this more realistic scenario, pairwise differences between individuals would be expected to be small, and added to even basic epidemiological information such as whether hospital inpatient stays overlapped, results using WGS on single colonies are still likely to be more informative than those obtained by standard investigation. Similarly, in an outbreak it is frequently sufficient to confirm, with high confidence, that transmission has occurred and that cases are linked, without needing to identify individual transmission links. In hospitals, some putative transmission events may be easily excluded on the basis of simple epidemiological data such as admission / discharge dates, since it can be readily deduced that patients were not in close contact if their ward stays did not overlap.

The serendipitous occurrence of a household transmission event followed by long-term carriage in two subjects in this study provides further evidence. In this pair, where an individual most likely acquired the strain from their household contact, the initial population in the presumed recipient was entirely clonal and reflected the dominant population of the donor.

This may indicate that both subjects acquired their strains from the same point source (for example, an unsampled household member), but given the time interval and exposure history it is more likely that A donated a single strain, representative of their dominant population, to B. Subsequently, despite ongoing exposure, the *S. aureus* strain in both individuals has evolved and maintained separate populations.

This pair is also an illustration of the pitfalls of relying on modelling data alone. Using BEAST, a software package frequently used to reconstruct phylogenies based on sequence variation and sampling time, the model predicted that subject A was re-infected by a clade from subject B, even though manual inspection revealed that none of A's SNVs occurred in later clades from B. This is presumably because the model does not take into account the fact that the sampling was from 2 different individuals.

An alternative method has recently been developed by a colleague and released as a software package called BASTA. This was primarily developed to reconstruct ancestries with reference to the geographical source of the genomes, allowing inference of migration rates and ancestral location. For example, if isolates X and Y were both sampled in England, there is a higher prior probability that they are more closely related to each other than to isolate Z which was sampled in Australia. For the transmission pair in this study, rather than dividing the sequences by geographic location, I treated each of the subjects as a separate location, or "deme", allowing the model to specify a higher prior probability that the later isolates from subjects A and B arose from their earlier populations. The results are much more in keeping with the epidemiological data and visual inspection of the mutations.

This provides a useful case study for the potential application of this software to comparing within-host populations, and also demonstrates the importance of interpreting and testing modelling data in the context of available epidemiological data.

Overall, these results suggest that where an outbreak occurs over a short period of time and does not involve a persistently colonised carrier, minimal diversity may be expected across the outbreak samples as there is insufficient time for diversity to accumulate. Rather than acquiring a cloud which then undergoes purifying selection, it appears from this study that there is a bottleneck prior to or at acquisition, resulting in acquisition of a minimally diverse population which accumulates mutations before potentially undergoing purifying selection at a much later stage. This is a somewhat different scenario, and hence conclusion, from that of Tong et al.

These results are reassuring from the perspective of single colony sequencing, as from the data above one would predict that, shortly after acquisition, the genetic distance between colonies from the same population will be minimal and the assumption that a single colony is representative of the rest of the population will be valid.

2.4.3 Explanations for founder effect

It may be that the founder strain has a selective advantage which favours transmission to and establishment in a new host, which may explain the lack of diversity in the initial population. In the Tong study, the emergent ITU clade was found to carry the *qac* gene, carried on a plasmid, which is associated with tolerance for and possible resistance to disinfectants such as chlorhexidine, commonly used in the hospital setting. This was not present in earlier clades and thus there may have been a selective advantage in the acquisition of this gene.

An alternative explanation is that the results are due to sampling artefact and that, by chance, the time-points chosen have coincided with a significant evolutionary event. The

early sampling times may have coincided with a selection event immediately preceding sampling and have captured a more homogeneous population due to removal of deleterious mutations or emergence of advantageous mutations. Similarly, later samples may have simply captured a more diverse population immediately before a purifying selection event.

However, against this is the finding by Rocha (131) that, for closely related strains, there is a low likelihood of a deleterious mutation occurring in such short time frames: assuming that the negative swab preceding acquisition is a true negative, there is therefore insufficient time for purifying selection to occur. Consequently, if the founding population is truly diverse, sampling at 2 months should still capture that diversity, given current estimated mutation rates.

A further possibility is that there is repeated loss / acquisition in the study subjects from ongoing exposure to an unsampled source (for example, their spouse, pet or other household contact), and although it is assumed that the sampling reflects true recent acquisition followed by persistent carriage, this can only be confirmed by more frequent sampling. It is reassuring that for the household pair 1218 and 1219, there is no overlap between the later populations, suggesting that there has been emergence of separate populations with no transmission events in the interim. Repeated loss and acquisition may be more of an issue in situations where eradication or suppression therapy is attempted, as initial populations may be removed and a related strain reintroduced, especially if there is ongoing exposure to the original source, for example in a household. In this study, there were no hospital admissions recorded for these subjects and since none of the strains was MRSA, there is no reason to think that they have been exposed to eradication regimens.

Another possibility is that the initial transmitted population was more diverse but rapidly underwent genetic shift due to host-pathogen interactions, with quick emergence (within 2 months) of the strain best adapted for its new host, followed by gradual genetic drift and accumulation of null mutations over time, ultimately resulting in increased diversity. This

hypothesis is not supported by the 1218-1219 pair, since their early populations were very similar, although it is not known whether this presumed household transmission is between spouses or siblings, as, if it is the latter, the host factors may be highly similar, resulting in initial establishment of the same clone.

The most likely explanation for the results above is that, in community transmission and carriage, there is a narrow bottleneck effect with a dominant strain being established in the recipient which does not fully reflect the diversity of the donor population, as only a small proportion of the cloud is transmitted, possibly related to a selective advantage of the dominant strain favouring transmission and establishment in the nares. This is followed by gradual genetic drift with a slow accumulation of diversity over time.

One limitation of the study is that the isolates were all nasal swabs from community subjects, and the findings may not be applicable to hospital outbreaks, where the *S. aureus* reservoir is different and subject to a range of selective pressures such as antibiotic use and cleaning agents. Similarly, the results may not be applicable where transmission results in infection, or where there is colonisation / infection of a different site such as a wound, throat or skin. The comparative behaviour of *S. aureus* populations at these sites has yet to be studied.

2.4.4 Estimation of the molecular clock

The molecular clock estimates are also relatively slow compared to previous estimates (1.16×10^{-6} per genome per year, compared with 3.3×10^{-6}) (133). This may reflect the fact that after initial establishment, there is stabilisation of the population maintained by host-pathogen interactions. In the absence of selection pressures such as antibiotic use or the introduction of competing strains, a stable population may emerge with a low mutation rate, so that even at one year there is minimal diversity.

2.4.5 Significance of SNV positions and resistance/virulence MGEs

There was no clustering of SNVs within the genome, and the proportions of synonymous / non-synonymous SNVS are similar to those seen in other studies (107, 130).

Prevalence of TSST-1 varies widely according to geography / epidemiological setting, usually reflecting its presence within the local circulating clones (134), with a higher prevalence in the hospital setting (135-139). There was a high prevalence of TSST-1 in the individuals in this study (3/8, 38%). This presumably reflects its frequent occurrence in clonal complex 30 (140, 141), since 2 of the TSST-1 positive samples belonged to this lineage.

Full clinical details for the subjects are not available, but there were no hospital admissions recorded during the study period for these subjects. Similarly, although community MSSA carriage of PVL remains relatively low in the UK (142, 143), the gene was found in one individual in this cohort. Again there were no hospital admissions or episodes of antibiotic use recorded for this subject to suggest they had recurrent infection.

In vivo variation in mobile genetic elements within a population has previously been reported (107, 144, 145), including for the φSa3 family as observed in this case. This may confound genomic comparisons unless these regions are identified and taken into account when comparing genomes at the individual SNV level.

2.5 Conclusion

The results confirm the hypothesis that there is little diversity in the founder population, and suggest that the accumulation of diversity in the course of a short-term outbreak is likely to be minimal, assuming that there is no repeated transmission from a long-term carrier.

One implication of this is that it may be possible to deduce the nature of an outbreak from the genetic data, even with limited sampling and sequencing. In a point source outbreak where there is rapid person-to-person spread, there will be minimal genetic diversity, and isolates will be clustered very closely together compared with non-outbreak isolates. However, if there is ongoing transmission from, for example, a staff member with long-term carriage and a cloud of diversity, outbreak isolates may be more genetically diverse.

This hypothesis will be examined in the following chapters.

3 Calibrating WGS as a tool for outbreak investigation

3.1 Introduction

To date, several studies have used WGS to investigate disease outbreaks due to *S. aureus* (113, 114, 146) and background transmission resulting in carriage (108, 132). However, these studies have focused entirely on intensive care units, and particularly neonatal intensive care units. It now seems inarguable that, in comparison with standard typing techniques (spa, PFGE, phage), WGS has enhanced resolution and therefore can demonstrate, with a very high degree of certainty, whether isolates are closely related to each other or otherwise. Consequently, there is much enthusiasm for the adoption of WGS into routine outbreak investigation. However, at present there is no consensus for interpreting results, or for understanding how WGS data relates to the results obtained using standard techniques, particularly in the wider hospital setting or in the community.

The results from chapter 2 indicate that, in the household setting at least, there is minimal diversity in the founder population. This provides reassuring evidence that, even if single colonies are sequenced, where the organism is rapidly colonising / infecting new hosts in an outbreak setting, there may be insufficient time for the accumulation of diversity, and WGS can therefore provide additional, useful information for the investigating clinicians.

To test this, WGS was applied to 15 well characterised outbreaks previously investigated by the UK PHE laboratory using standard epidemiology and typing techniques.

3.2 Methods

3.2.1 Sampling frame

Outbreaks were chosen in collaboration with the PHE reference laboratory, to provide a range of sequence types representative of community and hospital settings, and to include both MRSA and MSSA. 15 “resolved” outbreaks were selected, for which the

epidemiological links between patients were well described and which provided evidence for short transmission chains (i.e. fewer than 20 cases), and where the genetic relatedness had been confirmed by *spa* typing, MLST and PFGE (indistinguishable or closely related) (37). Participating hospitals were sent a proposal summary outlining the project, (see appendix 1), and the infection control lead was asked to give permission for the release of isolates and initial epidemiological data from the reference laboratory.

Isolates were retrieved from single colony frozen stocks held at the PHE reference laboratory, Colindale, London. For each outbreak, samples included both clinical isolates and nasal screening swabs.

Epidemiological information was obtained from each infection control team. This included data obtained as part of the routine investigation (date of specimen, sample type, ward location, date of admission, date of discharge, previous screening results where applicable, ward spell data including outpatient and day case attendances, clinical specialty). For community outbreaks, the data included social and household contacts and the results of extended screening.

For each outbreak, additional background samples were also included in the analysis. Two types of background samples were analysed:

1) all isolates obtained from the unit during the outbreak investigation, even if not directly epidemiologically linked. This was to estimate expected genetic diversity outwith the outbreak strain, to screen for overall increased rates of *S. aureus* transmission, and to ensure that the apparently outbreak strains were not simply part of an ongoing epidemic clonal expansion.

2) non-epidemiologically linked isolates matched for *spa*-type and / or MLST, obtained from collections held by the reference laboratory or from local collections held in Oxford. This was to provide a comparison for expected within *spa*-type distances, and to provide an outgroup for phylogenetic analysis.

3.2.2 Extraction and sequencing

DNA was extracted and sequenced as previously described from a single colony sub-cultured on to Columbia blood agar and incubated for 18-24hrs. Sequencing was performed using the Illumina HiSeq platform, except for 4 outbreaks which were sequenced in real time using the benchtop Illumina MiSeq.

3.2.3 Phylogenetic analysis

For all outbreaks, reads were aligned using Stampy v1.0.17 to reference genome MRSA252 (GenBank: NC_002952), chosen as a closed, fully annotated, Sanger sequenced exemplar organism, with masking of known mobile regions.

To assess the effect of the choice of reference mapping organism on SNV distances and phylogeny, confirmed outbreaks isolates were also mapped to a closed reference genome obtained from GenBank for the nearest possible MLST / clonal complex match. MSSA outbreak isolates were also mapped to MSSA476 (NC_002953.3).

SNVs were identified across all mapped non-repetitive sites with masking of known mobile elements as described in chapter 2, and maximum likelihood trees estimated using PhyML.

3.2.4 Analysis of MGEs

Assembled genomes were blasted for a panel of resistance and virulence determinants as described in chapter 2.

To excluded contributions from unannotated mobile elements, SNV clustering was assessed using the ape package in R to plot SNV density against genome position. Where there was a high density of SNVs in a single region, isolates were analysed using PHAST

(120) and PlasmidFinder (121) to search for mobile genetic elements which had not been identified from the MRSA252 reference genome.

3.2.5 Outbreak analysis

The index case was defined as the earliest case in each cluster. This could predate the start of the outbreak investigation as some cases were identified retrospectively. Outbreak cases were defined by the reference laboratory as those sharing an identical PFGE pulsotype and a definite epidemiological link to the index or secondary cases (>24hr stay in same ward, or sharing same household, classroom or similar community situation with prolonged contact (e.g. childcare)). For each isolate, the genetic distance in SNVs was calculated both from the index case and from the next nearest neighbour.

For analysis of SNV thresholds, isolates were considered to be definite outbreak isolates if they met the criteria above, and were closer to the index case than the *spa* / MLST matched comparator strains. If isolates were more distant than the *spa* isolates, antibiograms were examined and further epidemiological information sought.

The degree of epidemiological linkage was defined for each pair of cases within an outbreak as follows:

- Direct ward: >24 hr overlapping stay in same ward
- Indirect ward: same ward but inpatient stay did not overlap
- Healthcare worker-patient: HCW with direct patient contact with case
- Household: living / sleeping in same house
- Social: friend / other family member with regular contact but not living in same house
- School: attending same school during outbreak period

- Indirect: occurring within same outbreak but no direct contact identified between individuals (for example, family members of ward cases are considered indirect contacts of the other direct ward cases)

Statistical analyses were performed using Stata v13.1.

3.2.6 Comparison with standard results from infection control investigation

For each outbreak, results were returned to the infection control team to assess congruence between the genetic data and the conclusion from the original investigation (see appendix 2 for example feedback proforma). Teams were asked to assess whether the WGS analysis agreed with their analysis and whether it provided any additional information or benefit to the outbreak investigation. Where the conclusion differed, additional epidemiological information was sought to try and account for the results, and, if necessary, isolates were re-sequenced from the frozen stock for confirmation.

3.2.7 Outbreak simulation using Outbreaker

Outbreak F occurred in a neonatal unit in a local Oxford hospital. Enhanced screening was undertaken and detailed epidemiological data was available as the notes and ward records could be inspected directly. Because it was a neonatal unit, there was unequivocal evidence for exposure time and there were obviously no prior hospital or community exposures as all patients were admitted directly to the unit following delivery. This outbreak was therefore chosen for further analysis using the R Outbreaker package (147), to test whether the combination of genetic and epidemiological data could reconstruct credible transmission pathways.

The MRSA252 reference genome was used to construct an alignment with SNVs simulated at equivalent sites for each outbreak isolate. The alignment was uploaded into

R, and the Outbreaker package run using an exponential decline model to obtain the generation time distribution (time between colonisation and transmission to a secondary case). 4 parallel runs of 10,000,000 iterations were performed to obtain estimated dates of infection and R values (the estimated number of secondary cases per infected individual). The putative transmission tree constructed using these values was visualised using the transgraph function in the package.

3.3 Results

3.3.1 Outbreak / sample characteristics

101 outbreak isolates from 15 outbreaks were sequenced, along with 59 *spa*-matched comparator isolates and 144 background isolates. Outbreak descriptions are given in table 3.1.

TABLE 3.1 EPIDEMIOLOGICAL SUMMARY OF 15 OUTBREAKS INVESTIGATED BY WGS

Outbreak	Outbreak category	Reason for investigation	Phenotype	MLST	*Spa*-type	Duration (days)	No of cases
A	Hospital, maternity unit	PVL positive disease in index case	MRSA	ST772	t657	70	6
B	Hospital, maternity unit	Scalded skin syndrome in neonates	MSSA	ST2434	t346	70	9
C	Hospital, neonatal unit	Increase in MRSA colonisation	MRSA	ST22	t5892	43	4
D	Hospital, neonatal unit	Increase in MRSA colonisation	MRSA	ST30	t019	57	6
E	Hospital, neonatal unit	Increase in MRSA colonisation	MRSA	ST59	t216	8	3
F	Hospital, neonatal unit	MRSA bacteraemia	MRSA	ST88	t5973	65	8
G	Hospital, multiple wards	Increase in *S. aureus* wound infections	MRSA	ST241	t037	98	11
H	Hospital, single ward	Increase in MRSA colonisation	MRSA	ST22	t032	367	5
I	Hospital, single ward	Increase in MRSA colonisation	MRSA	ST8	t008	88	16
J	Hospital, single ward	Increase in MRSA wound infections	MRSA	ST22	t022	19	8
K	Household	PVL positive disease in index case	MRSA	ST30	t019	195	5
L	Household	PVL positive disease in index case	MRSA	ST30	t019	20	4
M	Household	PVL positive disease in index case	MRSA	ST30	t019	8	3
N	Household	PVL positive disease in index case (148)	MRSA	ST30	t019	44	8
O	School	PVL positive disease in index case	MSSA	ST121	t645	74	5

10 (67%) outbreaks were hospital associated. Of these, 4 were on neonatal units, 3 were on single adult wards, 1 involved multiple adult wards, and 2 were associated with maternity units. 5 (33%) outbreaks were community associated: 4 in households and 1 in a school. Reasons for instigating an outbreak investigation were: index case with severe / invasive disease, including PVL producing strains (7 outbreaks), increase in MRSA carriage identified by routine hospital screening (5 outbreaks), increase in *S. aureus* wound infections (2 outbreaks), and staphylococcal scalded skin syndrome in infants (1 outbreak). 2 outbreaks (13%) were due to MSSA, and the remaining 13 (87%) were due to MRSA.

The median number of cases for each outbreak was 6 (IQR 4-8). Median duration was 65 days (IQR 20-88).

Outbreak samples were representative of 9 different clonal complexes, 9 MLST types and 11 *spa*-types, reflecting the global diversity of *S. aureus* in the human population (102). As would be expected for the UK, given that the majority of outbreaks were MRSA, clonal complexes 22 and 30 occurred more frequently than other complexes (149).

For 14/15 outbreaks, *spa*-matched controls were sourced from existing sequenced collections (range 1-6 controls per outbreak). Outbreak L was caused by spa-type t5973, which is rare in the UK (146), and no *spa*-matched controls were available. For this outbreak, 6 MLST-matched isolates (ST88) were sourced from in-house collections to use as comparator strains in lieu of *spa*-matching.

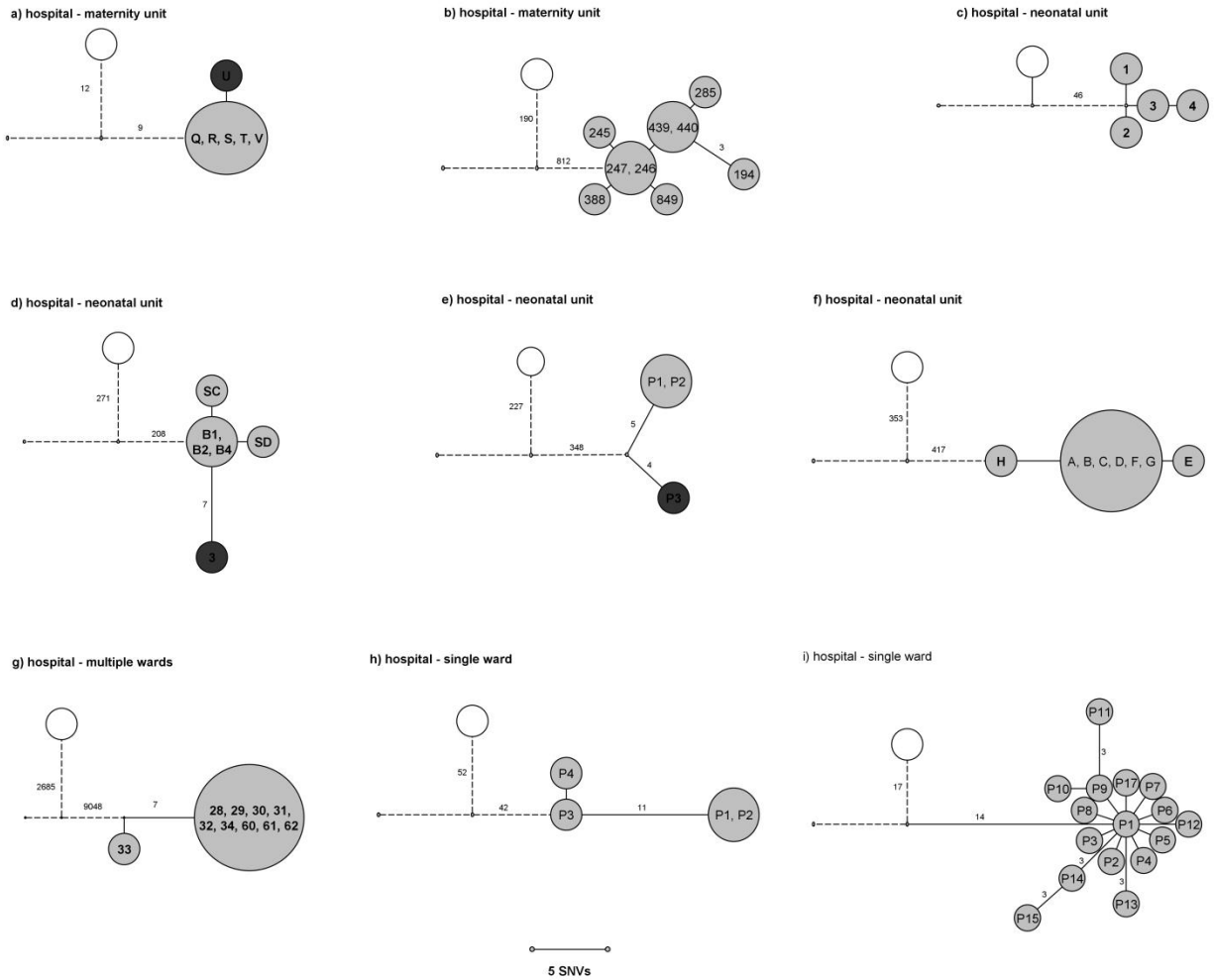
For all outbreaks, >80% coverage of the MRSA 252 reference genome was achieved.

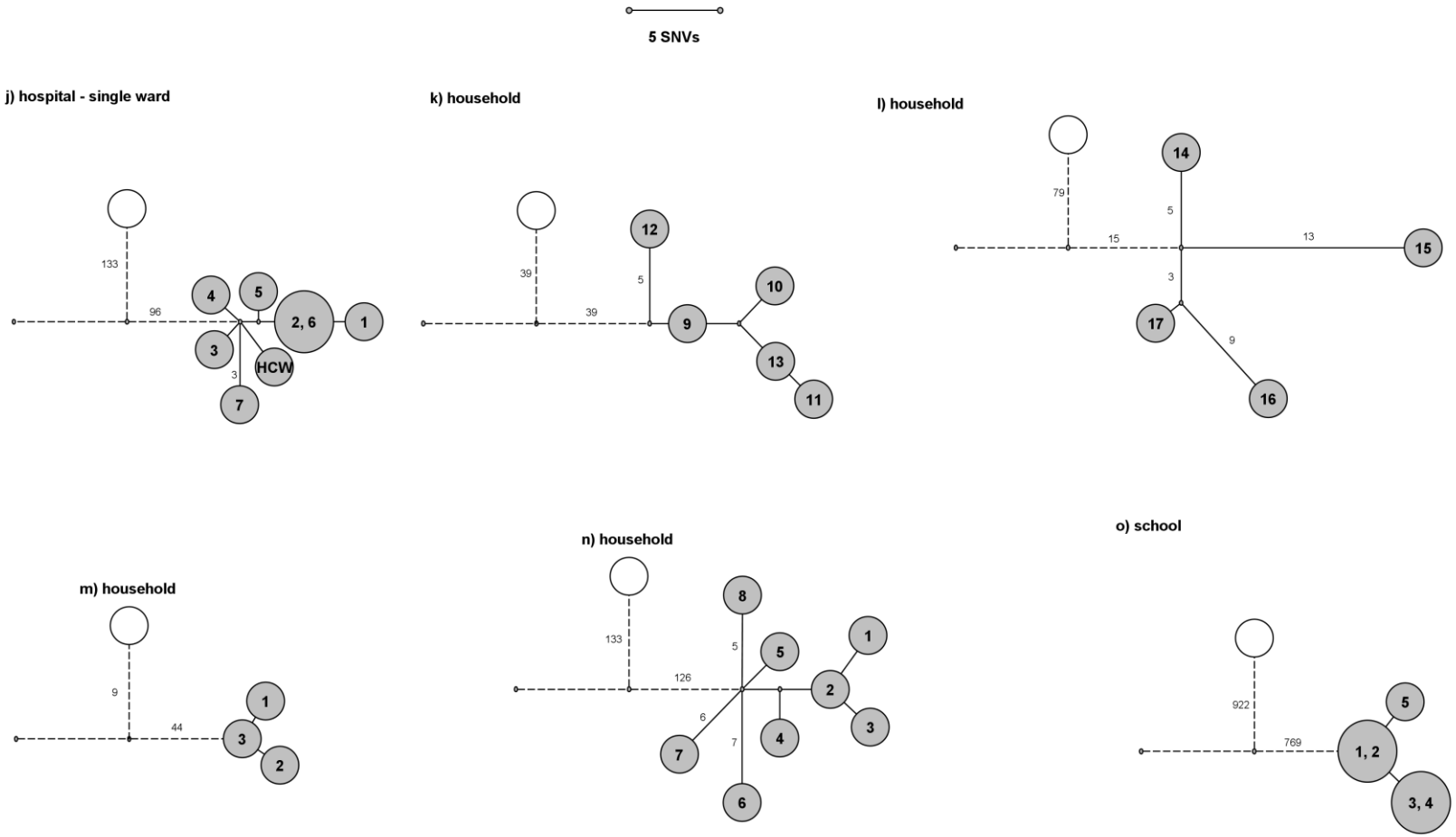
3.3.2 Phylogenetic analysis

Phylogenetic trees for the outbreaks are shown in figure 3.1, with the nearest unrelated background isolate shown for comparison. In all except one outbreak (outbreak I), isolates

were relatively clonal and clustered more closely to the index case than to the *spa* / MLST matched outgroup. In outbreak I, one isolate was more distant from the index than the nearest *spa* matched isolate, while the remaining 15 isolates were clustered very closely (within 5 SNVs of the index, figure 3.2). This isolate also had a different antibiogram, and lacked the *blaZ, qacA* and *dhfrA* genes found in the other isolates. No mutations were found in the *mut*S/*mut*L region to suggest it was a hypermutator, and examination of SNV positions did not find any clustering of SNV to suggest a significant recombination event (figures 3.3 and 3.4). Isolates from this outbreak belong to ST8, which is relatively common in the UK. Consequently, it is highly likely that this individual acquired the strain from a separate source to the rest of the outbreak cases, and this isolate was excluded from the SNV threshold analysis.

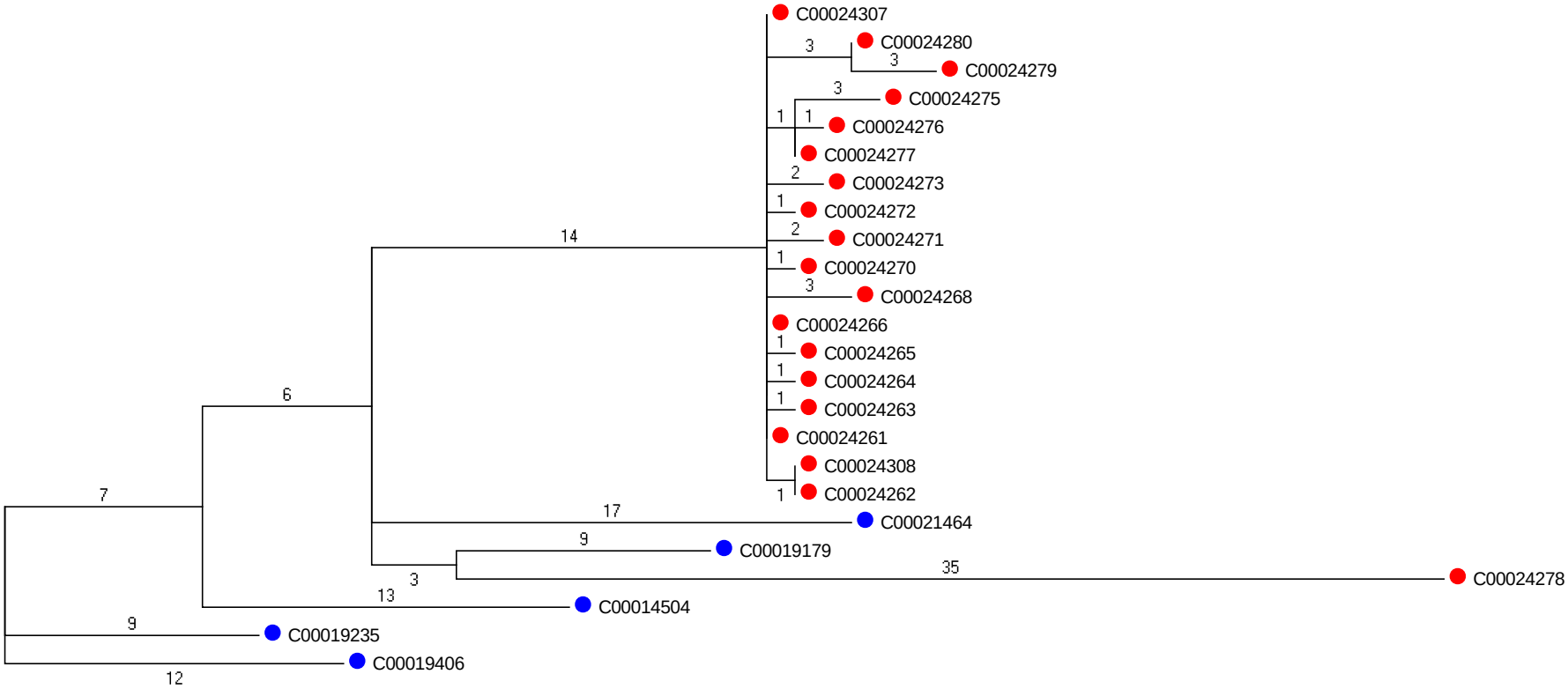
FIGURE 3.1 PHYLOGENETIC TREES FOR 15 OUTBREAKS ANALYSED BY WGS





Outbreak isolates are shown in grey, with dark grey circles indicating family members of cases affected by a hospital outbreak. Empty circles represent the nearest non-outbreak spa-matched isolates. Genetic distances are indicated in SNVs.

FIGURE 3.2 PHYLOGENETIC TREE FOR OUTBREAK I SHOWING NEAREST UNRELATED SPA-MATCHED ISOLATES



Red circles indicate putative outbreak strains, and blue circles indicate unrelated, spa-matched isolates. Isolate C00024278 (Patient 16) clusters more closely to the unrelated isolates than to the rest of the outbreak isolates.

FIGURE 3.3 DOTPLOT SHOWING DISTANCES FROM INDEX CASE FOR PUTATIVE OUTBREAK ISOLATES COMPARED WITH SPA-MATCHED ISOLATES FOR OUTBREAK I

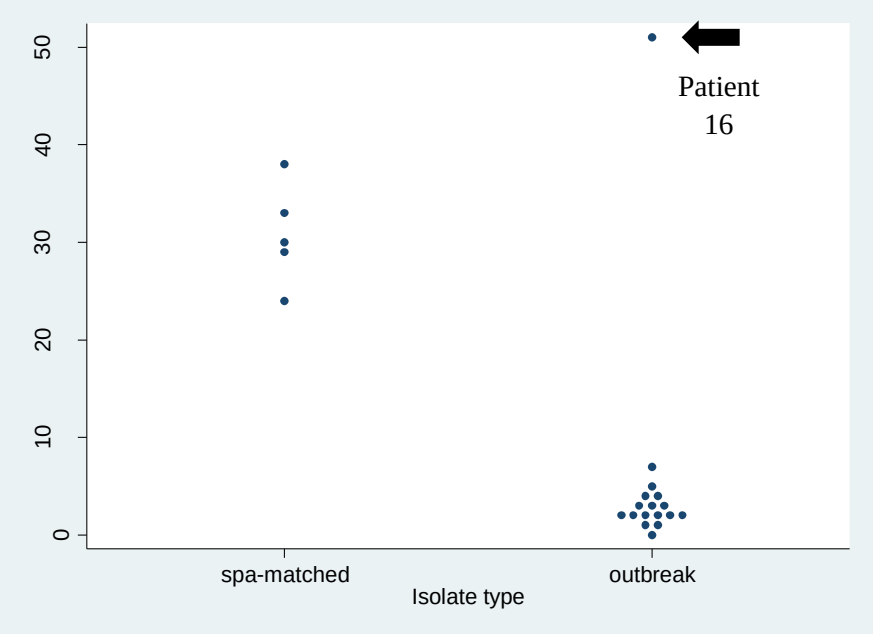
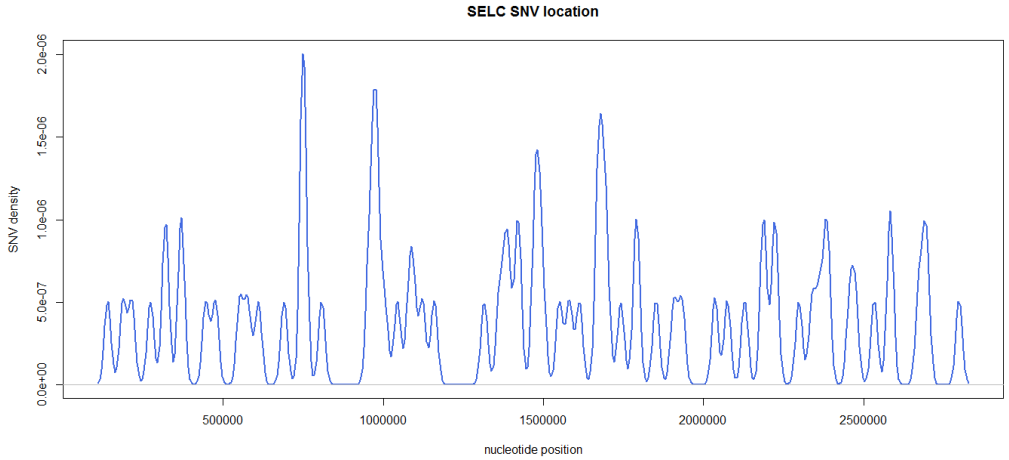


FIGURE 3.4 SNV DENSITY VS NUCLEOTIDE POSITION FOR SNVS IDENTIFIED ACROSS ALL OUTBREAK ISOLATES IN OUTBREAK I



Bin width = 100 (i.e. SNV density per 100 bps)

SNVs are spread across the genome with no areas of high density to suggest a recombination event.

3.3.3 SNV thresholds

The overall mean pairwise difference across all outbreak sample pairs was 3.7 SNVs (95%CI 3.3-4.1), compared with 4444 SNVs for spa-matched pairs (95% CI 2492-6395) and 30192 SNVs for non-outbreak isolates from the same units (95% CI 29781-30603). No additional outbreak isolates were identified in the background sampling.

All outbreak isolates were 18 SNVs or fewer from the index case. All except one were within 10 SNVs of their nearest neighbour (figure 3.5). The outlying isolate was 17 SNVs from its nearest neighbour and came from a household cluster identified due to PVL positive disease in the index case. 90% of isolates were within 6 SNVs of their nearest neighbour.

3.3.4 Factors affecting within-outbreak diversity

Linear regression analysis revealed no significant association between within-outbreak mean pairwise difference and any of outbreak type (p=0.919), outbreak duration (p=0.306), reason for outbreak investigation (p=0.966), or MRSA phenotype (p=0.258).

However, there was a significant difference in the overall mean pairwise difference for healthcare associated (3.05 SNVs, 95%CI 2.70-3.41) versus community (household / school) outbreaks (7.14 SNVs, 95% CI 5.73-8.55), p<0.001, figure 3.6.

FIGURE 3.5 PERCENTAGE OF OUTBREAK ISOLATES INCLUDED BY SNV THRESHOLDS

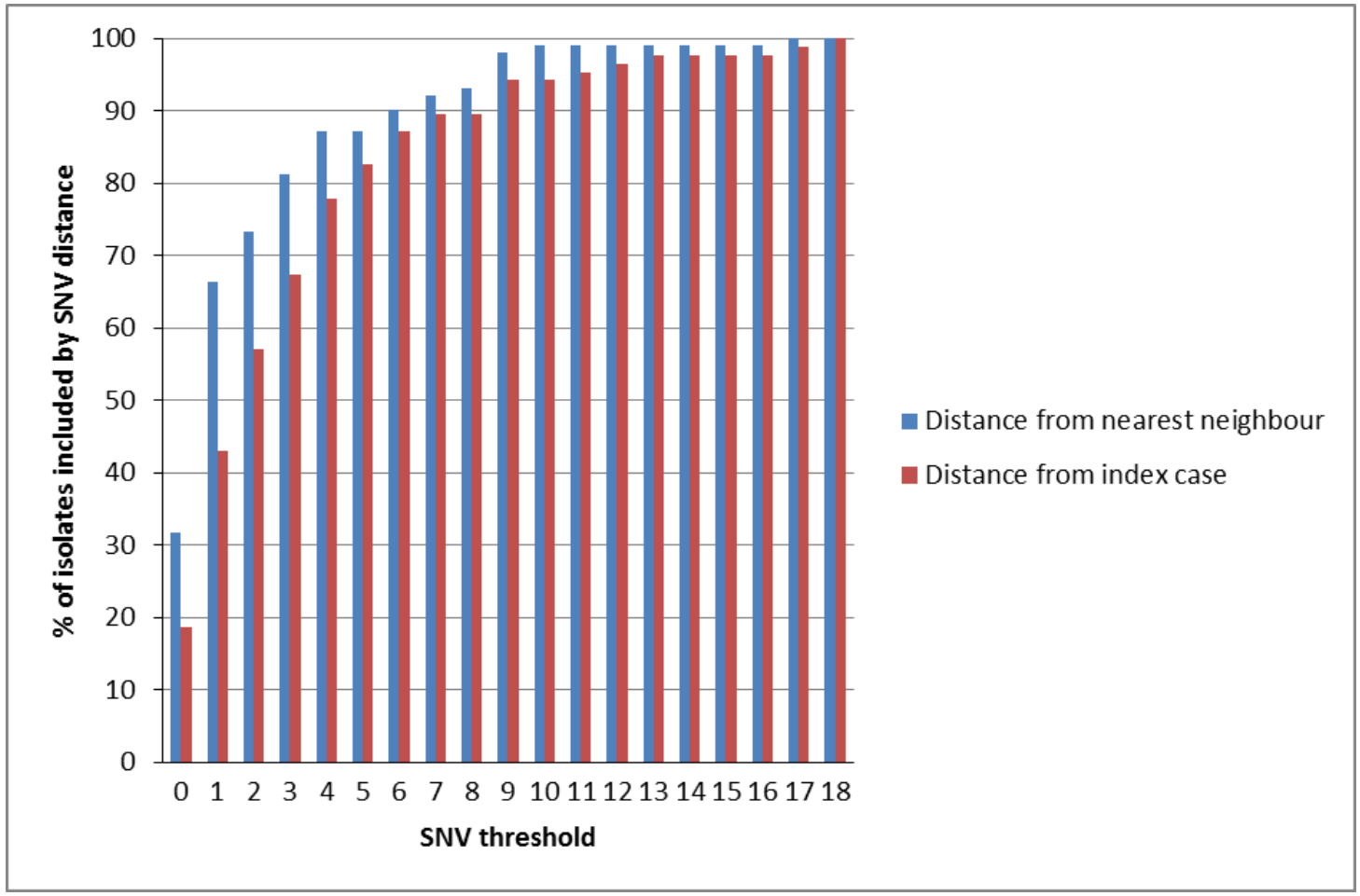
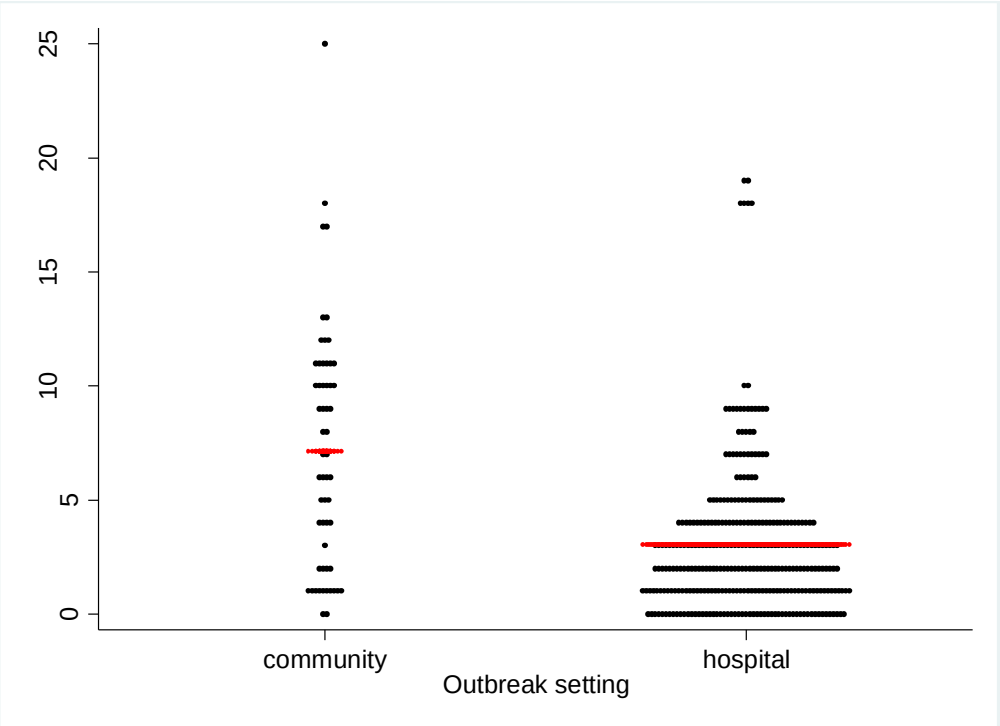
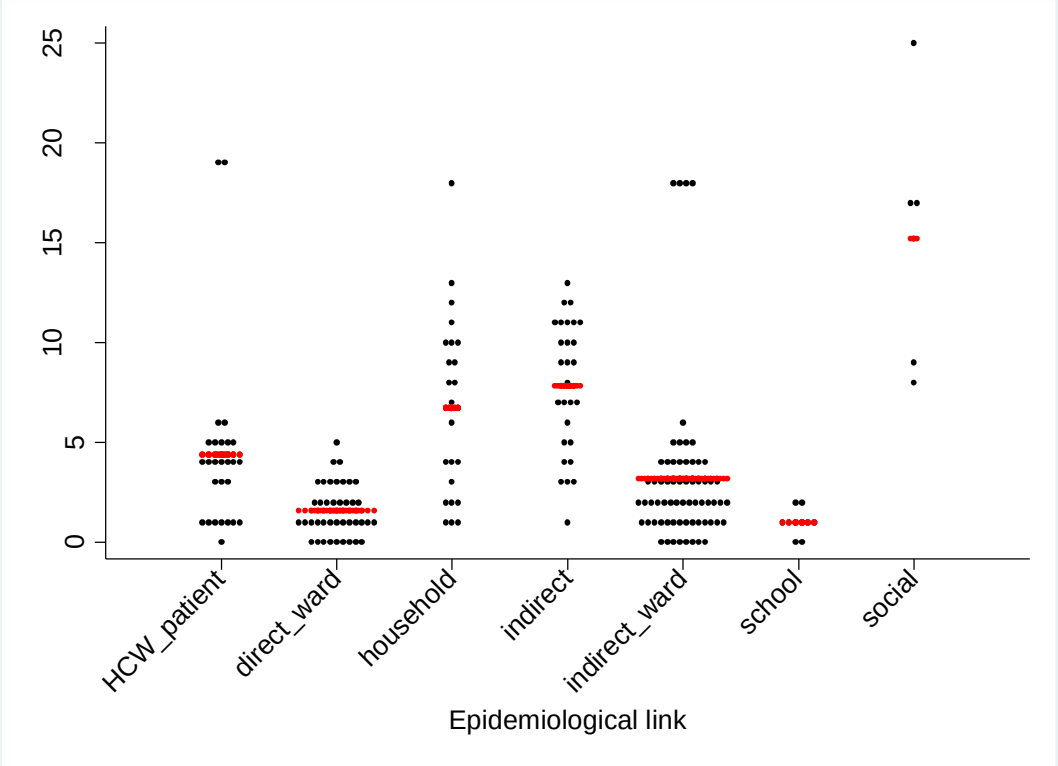


FIGURE 3.6 DOTPLOT OF ALL PAIRWISE DIFFERENCES FOR COMMUNITY / HOSPITAL SETTINGS



The overall mean for each setting is plotted in red.

FIGURE 3.7 PAIRWISE DIFFERENCES BY EPIDEMIOLOGICAL LINK



Mean pairwise difference for each category is indicated in red.

3.3.5 Effect of mapping

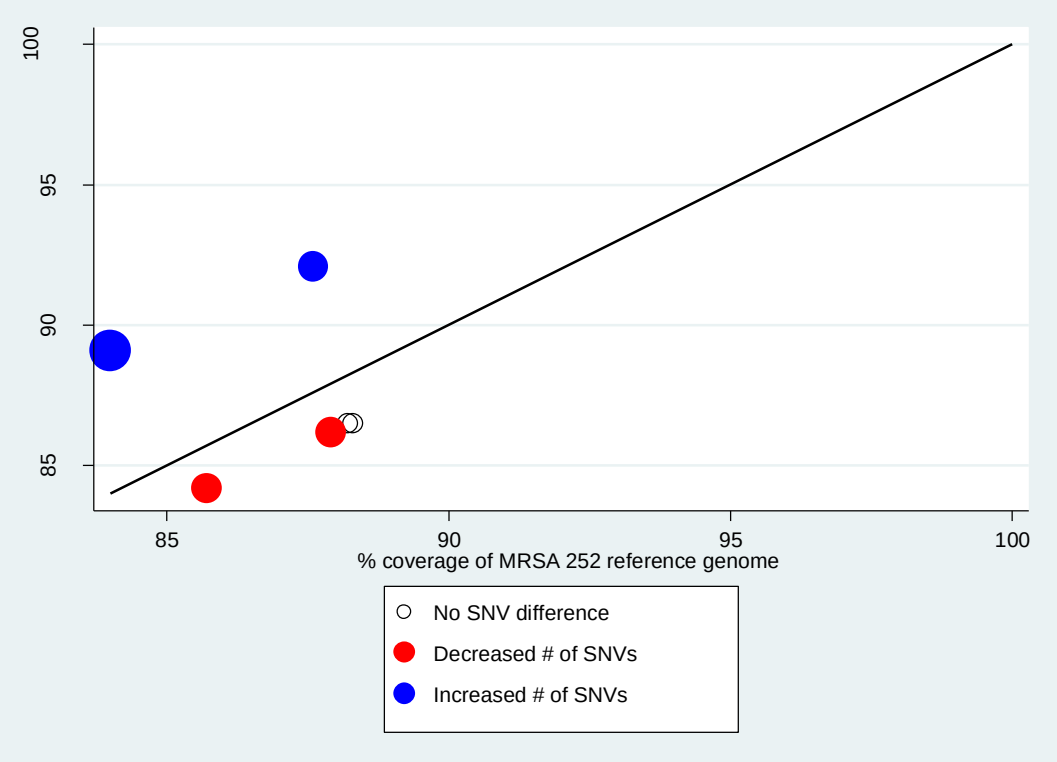
6 outbreaks were remapped to alternative reference genomes of the same or similar clonal complex as the outbreak strain. For the remaining outbreaks, either MRSA252 was the closest available reference genome, or there was no suitable, closer reference genome available.

Across the 6 outbreaks, mapping to MRSA252 yielded a total of 66 SNVs, and mapping to the alternative reference yielded 68 SNVs (table 3.2). The median pairwise difference for the outbreaks was 2.80 for MRSA252 mapping and 2.77 for alternative reference mapping. In 2 outbreaks, there was no difference in the SNVs identified between standard mapping (to MRSA252) and within-clonal complex mapping. Standard mapping (to MRSA252) identified one additional SNV position in 2 outbreaks, while within-clonal complex mapping identified 3 additional SNVs in 1 outbreak, and one additional SNV in the final outbreak. In each case, the increase in SNVs was in the same direction as the increase in percentage of the reference genome covered (figure 3.8). There was no effect on overall tree phylogeny for any of the outbreaks.

TABLE 3.2 DIFFERENCE IN SNVS IDENTIFIED BY MAPPING TO WITHIN-CLONAL COMPLEX REFERENCE GENOME, COMPARED WITH STANDARD MAPPING TO MRSA 252

Outbreak	CC-match reference genome	% coverage		No of SNVS identified			Mean pairwise difference	
		MRSA 252	CC-match	MRSA 252	CC-match	Difference	MRSA 252	CC-match
B	MSSA476	84.0	89.1	4	7	-3	1.97	1.75
C	EMRSA 15	88.3	86.5	4	4	0	2.17	2.17
H	EMRSA 15	88.2	86.5	20	20	0	11.7	11.9
I	USA300	87.6	92.1	24	25	-1	3.43	3.65
J	EMRSA 15	87.9	86.2	12	11	+1	3.61	3.36
O	*S. aureus* Newman	85.7	84.2	2	1	+1	1.00	0.60

FIGURE 3.8 PERCENTAGE OF REFERENCE GENOME COVERED FOR MRSA 252 AND CC-MATCHED REFERENCES

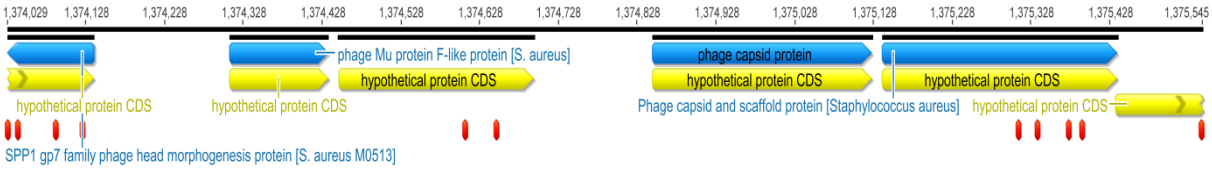


Size of circle indicates number of SNVs differing between the mapped genomes, with blue circles indicating increased number of SNVs identified when mapped to the CC-matched reference and red indicating a decrease in the number of SNVs.

3.3.6 Variation in MGEs

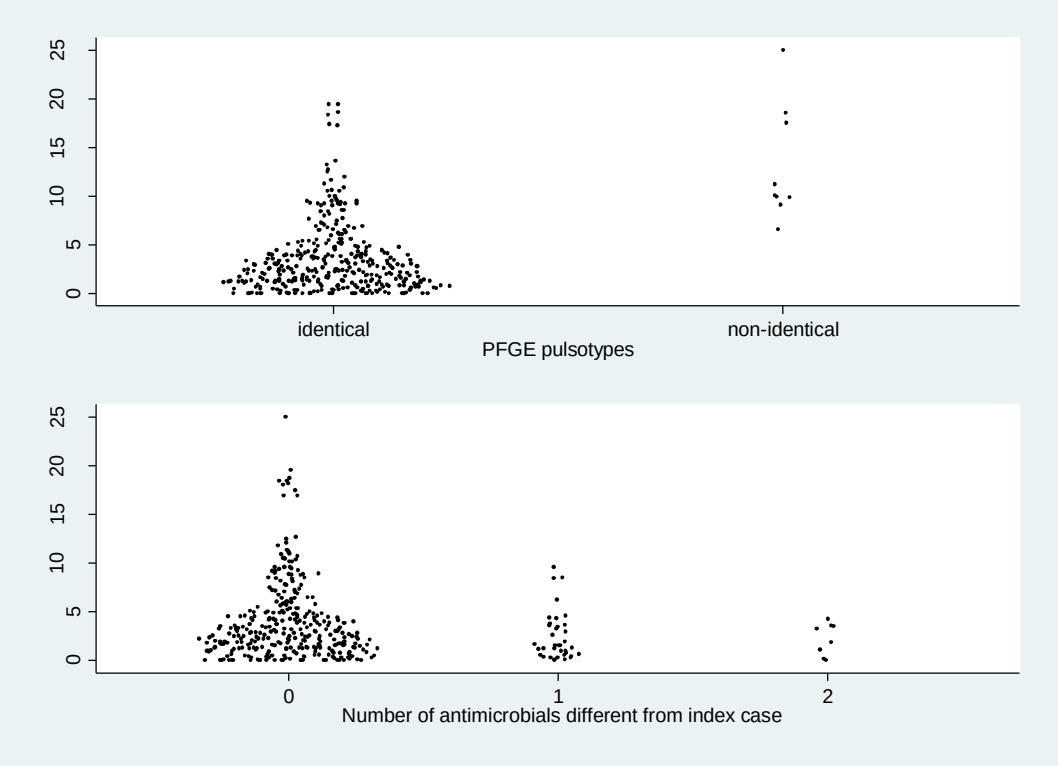
In one outbreak, 4 isolates were identified as having direct ward contact and shared the same PFGE. Mean pairwise difference was 11 SNVs, with a maximum distance of 19 SNVs between isolates. This was higher than for other outbreaks involving direct ward contact. Isolates shared the same SCC*mec* element and no *mut*S/*mut*L mutations (indicative of a hyper-mutator (113)) were identified. Analysis of SNV density identified that, out of 23 SNVs, 11 (48%) occurred in a single 1516 bp region. Examining this region with PHAST and PlasmidFinder did not identify any intact mobile elements, however BLAST searches of the open reading frames identified 4 putative phage proteins in the region. Eight of the 11 SNVS in this region occurred in 2 of these putative phage proteins (figure 3.9), potentially suggesting a recombination event within the outbreak rather than an accumulation of mutations. Masking this region as for known mobile elements resulted in a mean pairwise difference of 3.4 SNVs with a maximum distance between isolates of 8 SNVs across the outbreak.

FIGURE 3.9 VARIABLE REGION CONTAINING PHAGE PROTEINS



SNVs are indicated in red.

FIGURE 3.10 PAIRWISE SNV DIFFERENCES FOR OUTBREAK ISOLATE PAIRS WITH A) DIFFERING PFGE PULSOTYPES B) DIFFERING ANTIBIOGRAMS



In 3/15 resolved outbreaks, there was a difference in antimicrobial susceptibility across the outbreak isolates, which was confirmed by the presence / absence of mobile resistance determinants using BLAST on the whole genome sequences. In outbreak B, there was a difference in tetracycline susceptibility, with 4 isolates carrying a rep7 plasmid containing the *tet*K gene (150). In outbreak F, there was variation in resistance to both tetracycline and penicillin: one isolate was susceptible to both, 2 isolates were resistant to both, and the remaining isolates were resistant to penicillin and susceptible to tetracycline. The phenotypes corresponded to the presence / absence of *tet*K and / or *blaZ* in the respective genomes. In outbreak N, there was variation in susceptibility to penicillin corresponding to the presence / absence of *blaZ*.

In spite of the differences in phenotype, isolates could still be identified as belonging to the outbreak on comparison of the mapped genomes, as the resistance genes were carried on mobile elements and therefore masked. The mean pairwise difference between isolates sharing the same antibiogram was in fact higher than for isolates with differing antibiograms (3.9 SNVs (95% CI 3.4-4.3) compared with 2.3 SNVS (95% CI 1.5-2.9), p=0.0055, figure 3.10)

Similarly, 2 outbreaks contained isolates which differed from the index case by 1 or more bands on PFGE. The mean pairwise difference between isolates with identical PFGE pulsotypes was 3.4 SNVs (95% CI 3.1-3.8), compared with 12.9 SNVS (95% CI 8.3-17.4) between isolates with differing pulsotypes, p<0.0001 (figure 3.10)

3.3.7 Comparison with routine results

As described for outbreak I, WGS was able to exclude isolates from the putative cluster by demonstrating a greater distance than isolates known to be unrelated. For one further outbreak, all definite outbreak isolates were within 9 SNVs of the index case. Two further isolates were submitted for WGS analysis as they shared the outbreak PFGE pulsotype and came from a hospital in the same geographical region, although there were no direct epidemiological links: these were defined as "possibly related" by the reference laboratory. Using WGS, these isolates were found to be 15 and 23 SNVs distant from the rest of the cluster, in keeping with a relatively recent common ancestor but not direct transmission. Although the individuals from whom the comparator samples were taken from had not been in the outbreak hospital, there was occasional transfer of patients between the hospitals and therefore the WGS results were in keeping with the epidemiology.

Feedback from the infection control teams indicated that, in 7/15 outbreaks, WGS was felt to have enhanced the routine investigation. In outbreaks A and F, sequencing was performed in real time using the Illumina MiSeq platform, allowing rapid confirmation of the outbreak and directly influencing the investigating teams' responses. Putative outbreak isolates were excluded in outbreaks G and I as described above. In outbreak G, one of the putative outbreak isolates arose several weeks after the outbreak was thought to be controlled. The isolate was closely related on WGS, confirming with a high degree of certainty that the outbreak strain was still present in the unit.

In outbreaks B and N, confirmation by WGS of a definite outbreak prompted the investigating teams to instigate enhanced surveillance, including staff screening, to ensure that the outbreak was under control.

Finally, in outbreak J, two isolates with the outbreak antibiogram were submitted for investigation and were found to have identical PFGE pulsotypes, although different from the outbreak strain. The team were concerned that this represented a second, previously

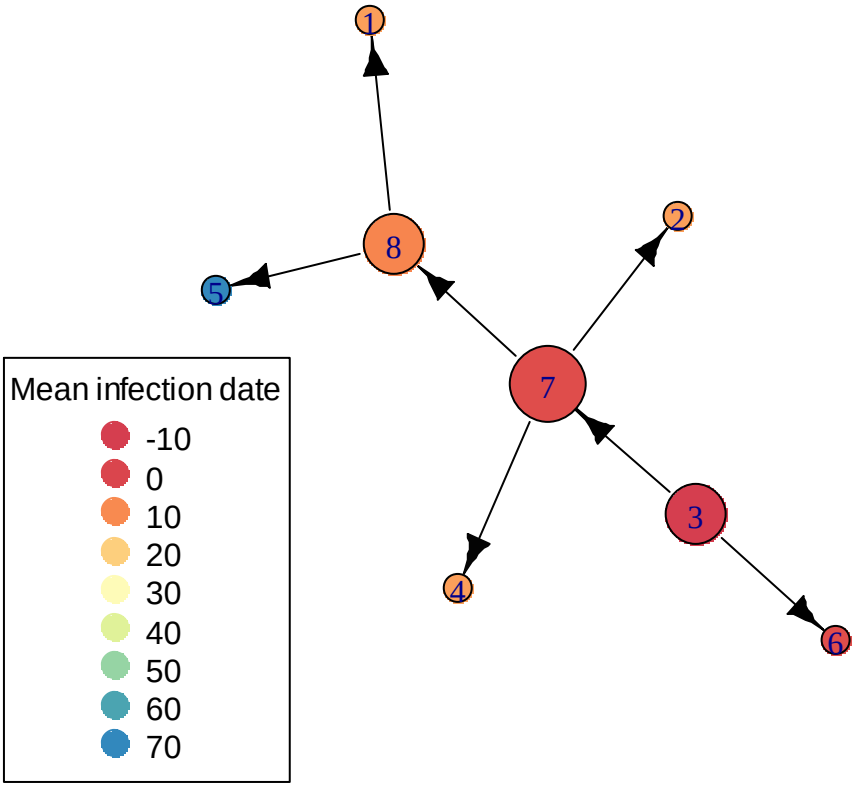
undetected outbreak; however, these two isolates were 83 SNVS apart, spread across the genome, and this was sufficient to exclude direct transmission.

3.3.8 Comparison with Outbreaker software

The transmission network inferred by Outbreaker is shown in figure 3.11. From the stimulation, there is high posterior support (>0.90) for a network starting with case 3, with onward transmission to cases 6 and 7. There is further transmission from case 7 to cases 2, 4 and 8, and from case 8 to cases 1 and 5.

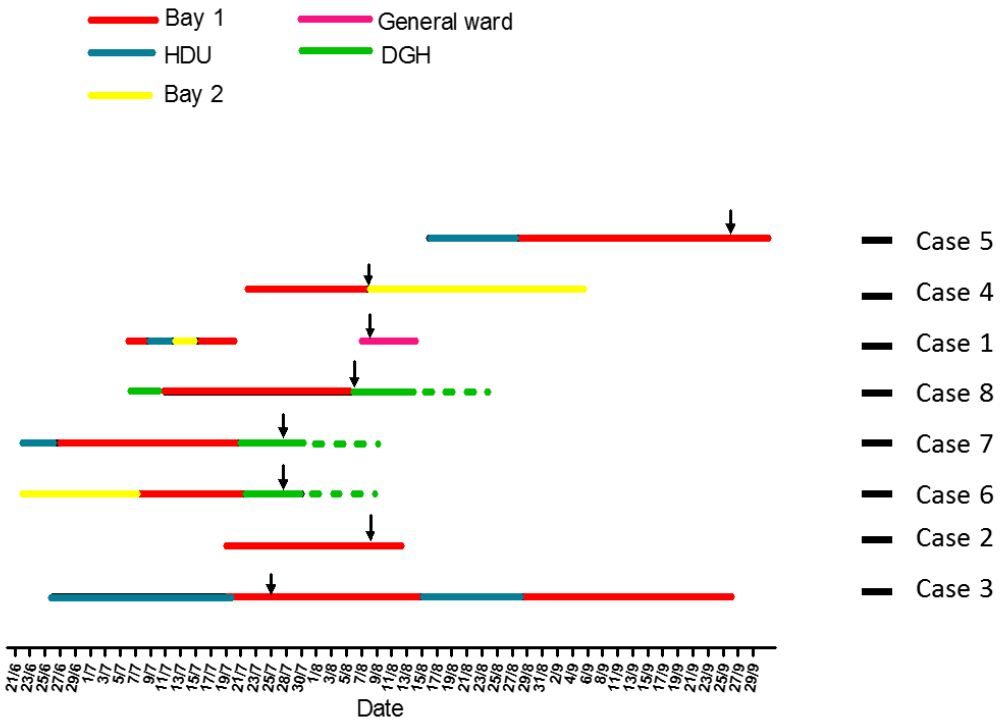
A timeline of the inpatient stays (starting from time of birth) is shown in figure 3.12. From the timeline, cases 2,3,4 and 8 have direct overlapping ward stays, suggesting transmission between these cases. However, the most likely transmission route for the bacteraemia case (case 5) is from case 3, as this is the only case with whom they had direct ward contact. Case 8, the putative source for case 5 identified by Outbreaker, had been transferred to another hospital shortly before case 5 was born.

FIGURE 3.11 TRANSMISSION NETWORK INFERRED BY OUTBREAKER



Nodes are coloured by estimated infection date in days from index case, and sized according to the probable number of secondary cases they have caused. All transmission pathways have an estimated posterior support of >0.90.

FIGURE 3.12 TIMELINE OF INPATIENT STAYS FOR OUTBREAK F



Black arrows indicate positive MRSA sample. Dashed horizontal lines indicate cases where the date of discharge is unknown (due to transfer to a different hospital).

3.4 Discussion

3.4.1 Current outbreak investigation

Identification of the cause of an outbreak is usually necessary in order to control the source and to prevent further exposures and cases. Even when an outbreak has self-terminated, information can be gained which may help to prevent future outbreaks by influencing policies or improving prevention strategies (151).

The general approach to outbreak investigation relies on identifying an increased number of cases, either through physician awareness, laboratory surveillance or central reporting and monitoring. In *S. aureus*, this is usually through routine surveillance for MRSA carriage (152), post-operative wound infections (153) or bacteraemias (154), or active case finding following severe or invasive disease.

Outbreak case definitions are based on confirming the genetic relatedness of the isolates, which is usually based on PFGE band patterns. However, this may lack both sensitivity and specificity, since band changes caused by recombination events (gene re-assortment, insertions, deletions and gain/loss of mobile elements) may not necessarily reflect changes or otherwise in the core genome. For *S. aureus*, isolates are considered clonally related if they vary by six DNA fragments (bands) or fewer (155), but, conversely, unrelated global epidemic clones such as the USA300 strain may be highly conserved with identical pulsotypes, even when there is minimal epidemiological linkage (46). In the UK, isolates belonging to the major epidemiological lineages (EMRSA-15, EMRSA-16) have highly conserved pulsotypes, resulting in heavy reliance on the epidemiological investigation and the attendant risk of incorrectly inferring transmission links.

The increased resolution of WGS makes it a highly attractive alternative to PFGE, however, as discussed in the preceding chapter, its use has been questioned due to concern around within-host variability (111, 112).

3.4.2 Use of WGS in outbreak / transmission studies

Several studies have used WGS to investigate the transmission dynamics of *S. aureus*, all of them in unique settings. Harrison and colleagues (156) identified bovine / human and sheep / human transmission of *mecC* MRSA isolates, found predominantly in livestock (157). SNV distances in this case were in keeping (0-5 SNVs) with the results above, despite the differing host species. Altman and colleagues found 0 SNVs between MRSA isolates from a liver transplant donor and recipient (158), and a similar study found a single SNV between liver and lung donor / recipient isolates (159), confirming solid organ transplantation as the source of MRSA bacteraemia in the recipients. Other studies in the hospital setting (108, 160) have found that the increased resolution of WGS data highlighted a lower than expected rate of MRSA transmission in the context of screening programmes, where transmission events had previously been assumed on the basis of epidemiological data and spa-type or MLST plus antibiogram. As discussed above, variation in MGEs means that antibiograms may not reflect core genome relatedness, so excluding isolates on the basis of antibiogram as was done in the latter study may have reduced the number of transmission events / outbreaks identified.

Transmission studies in the community setting have focused on the hugely problematic USA300 strain in an attempt to understand the success of this clone (161, 162). The first of these found more diversity than expected within a large outbreak of ST 8 MRSA in New York, and was able to exclude some putative transmission events previously identified using *spa*-typing, MLST and / or PFGE. This study found up to 23 SNVs separating isolates from the same household. In the second study, household isolates from San Diego and New York differed by a mean of 17.6 and 12 SNVs respectively. Together with the longitudinal data presented in chapter 2, this suggests that long-term persistence of the clone within households could potentially act as a reservoir for further transmission in communities.

The SNV distances reported in the San Diego / New York study are in keeping with the distances observed in the household clusters in my study, supporting the hypothesis that there is greater diversity within a household outbreak compared to a hospital outbreak, either as a result of a cloud in the donor case or due to the time difference between acquisition and sampling. Finding evidence of this in a clone other than USA300 suggests that it is due to epidemiological factors, rather than the rate of evolution of a particular clone. My results should therefore be readily applicable to such transmission studies, allowing ready identification of isolates which are highly related and allowing better interpretation of genetic distances.

Six studies have specifically used WGS in *S. aureus* outbreak investigations. The first two were conducted as a pilot study for this thesis (146), and the outbreaks are included above (outbreaks A and F). Of the remaining studies, 3 involved neonatal units and used WGS to confirm relatedness in unusual situations where there was an epidemic clone (163), a hyper-mutator (113), and where a staff member was thought to be responsible for a prolonged outbreak (114). These last 2 were from the same unit, although occurring at different times. In total 4/6 of the published outbreak investigations have involved neonatal units, and the applicability of findings to other outbreak settings cannot be assumed.

In the sixth study, WGS excluded transmission of MRSA in a US professional football team (164). MRSA, and particularly the USA300 clone, is a major cause of morbidity in USA football players at both the school / college and professional level (165-167). Outbreaks (and football) excite considerable public interest, hence the use of WGS to resolve this outbreak. However, although refuting this outbreak may help to inform future investigations and there may be financial implications for the teams involved (168), the results from this study clearly cannot be used for calibration of other outbreaks.

The 15 outbreaks in this study are likely to reflect current public health concerns rather than the true epidemiology of *S. aureus* outbreaks. The majority of the outbreaks included were caused by MRSA despite its low prevalence in the UK, since active surveillance and

awareness of MRSA means that MRSA outbreaks are more likely to be detected and followed up (169, 170). Likewise the study is also enriched for PVL positive outbreaks (5/15 outbreaks), as finding PVL positive disease triggers active case finding in the UK and hence a higher likelihood of finding an outbreak than for PVL negative disease (171). Despite these limitations, the inclusion of neonatal and adult wards, hospital and community outbreaks, and a range of clonal complexes means that these results should be broadly applicable.

3.4.3 Effect of epidemiological linkage on SNV thresholds

The results from my study demonstrate that there is minimal within-outbreak diversity in a range of epidemiological settings and genotypes, and a high degree of clonality within an outbreak compared with non-epidemiologically linked, spa-matched isolates. The results support the use of WGS in the outbreak setting, as there is minimal evidence for confounding by a cloud of within-host diversity, and the results are in keeping with the results in chapter 2, namely, there is a narrow bottleneck resulting in minimal diversity at acquisition. The results also allay concern regarding multiple picks: in these outbreaks, isolates were obtained from frozen stocks held by the reference laboratory which had been passaged at least twice prior to sequencing (when sent to reference laboratory and when re-grown from frozen stocks). If there had been significant within-host diversity in the individuals affected by the outbreak it is unlikely that selection from the primary plate always captured the isolate closest to the outbreak clone, although serendipity cannot be excluded.

A SNV threshold of 6 identified 90% of definite outbreak isolates, and all except one isolate were within 11 SNVS of their nearest neighbour. The outlying isolate belonged to a household cluster, and is in keeping with a time to common ancestor of 1-2 years. In this case, there may have been prolonged carriage prior to time of sampling, explaining the relative distance in this outbreak.

This is supported by the analysis of factors affecting SNV distance. There is not only a significant difference in mean pairwise difference for community versus hospital outbreak settings, but also a significantly increased number of SNVs observed between isolates where the epidemiological link was social or household. As might be expected, there was also a significantly increased number of SNVs observed between indirect epidemiological contacts compared with direct epidemiological contacts, as these suggest other intervening cases in the transmission chain.

These differences in SNV distances presumably reflect the greater uncertainty over the time between acquisition and time of sampling compared with hospital based links where acquisition is more likely to be recent. For hospital outbreaks in particular, there is a limited exposure window and a high likelihood that patients are sampled soon after acquisition. Where there is active surveillance, there may be a maximum of one week between acquisition and the next screening swab, and therefore there is minimal time for the accumulation of SNVs. For social and household contacts, there may have been multiple exposures over a prolonged period, and time of acquisition may predate time of sampling by months or even years, allowing diversification of their populations.

The association between epidemiological link and SNV distance may mean that is it possible to elucidate the origin of an outbreak: if all strains are virtually identical then this suggests a single introduction followed by short transmission chains, as one might expect from a hospital ward where an MRSA positive patient is admitted for a short time and transmits to one or two other patients who are then sources for the next transmission episodes. Where more diversity is seen, this may be due to a prevalent circulating strain (in which case it may be necessary to expand the sampling frame to include weakly linked isolates to see if this strain is present with the same degree of diversity e.g. in the rest of the hospital) or due to donation from a long-term carrier such as a staff member or previously colonised patient with a prolonged hospital stay (132). This would be a significant advantage over current typing techniques, as the increased information

provided by the WGS data may be able to inform the outbreak investigation and aid decision about whether to screen staff and / or family members. This is important from the occupational health perspective as the consequences on the staff member may be considerable and any intervention should be founded on robust evidence.

These results endorse the hypothesis proposed in chapter 2. Despite the evidence for within-host diversity discussed in chapter 2, this did not appear to be relevant in the outbreaks investigated. It may simply be that, in contrast to the situation described by Koser et al (113), none of these outbreaks incorporated a long-term carrier and they therefore represent point source transmission where all strains might be expected to be genetically near-identical. However, if a "super-spreader" is suspected, multiple sampling of likely donors, or sequencing multiple picks from the suspected donor, may be of use in establishing them as the source. This would be more cost-effective than processing multiple samples for all cases, and would not need to be done for all outbreaks.

In the high transmission setting described by Tong and colleagues (132), there is no decolonisation or isolation of affected patients, allowing large-scale propagation of clones from a single long-term carrier. This, one would hope, is different from the situation in the UK, where frequent screening of inpatients, decolonisation and isolation precautions means that transmission on this scale is infrequent. Hence sequencing multiple isolates to the same extent should rarely be necessary.

3.4.4 Effect of mapping on SNV distances

Perhaps counter to expectations, I found that mapping to alternative reference genomes has minimal effect on the final results. The maximum number of additional SNVs obtained by remapping was 3, for an increase in reference genome coverage of 5% (from 84% to 89%).

Since all outbreaks achieved >80% coverage of the reference genome in all isolates, this provides evidence that use of a single reference genome in this context is acceptable. Where there is <80% coverage of the chosen reference genome, it may be prudent to remap to a closer reference genome. This provides evidence that generic SNV distances used in previous studies are broadly applicable and are not necessarily contingent on the mapping method used. This finding also means that genetic data can be mapped as soon as it is available without the need for initial typing or screening to decide which reference to map to. This will reduce turnaround times, with the caveat that isolates may need to be remapped if the conditions above are not met.

Importantly, including the unmapped regions may lead to falsely excluding isolates, as these regions are likely to contain mobile genetic elements which, as noted above, may vary within an outbreak.

3.4.5 Variation in MGEs

In 3 outbreaks, there was a difference in antibiogram despite close genetic relatedness. Interestingly these did not result in a different PFGE band pattern, although in all cases the difference in resistance was due to the presence or absence of resistance genes carried on mobile genetic elements rather than altered targets at the chromosomal level.

Antibiogram differences within an individual have been reported by Stanczak-Mrozek and colleagues (144). In this study, screening swabs from 9/38 MRSA carriers were found to contain more than one antibiogram on examination of multiple picks, correlating to differences in phage / plasmid profiles. A further 15 carriers had variation in mobile elements which did not result in an altered phenotype. WGS analysis of a single carrier with multiple phenotypes revealed no SNV differences at the core genome level. Additionally, data from animal studies have demonstrated rapid and frequent in vivo

transfer of phages and plasmids between isolates, even in the absence of antibiotic selective pressure (145).

These findings support the results seen in the outbreak study above, where the core genomes were closely related despite variation in MGEs. This highlights the importance of taking horizontal gene transfer into account when considering genetic relatedness. Investigators should also reconsider the use of antibiograms to screen for the possibility of an outbreak, since this may result in isolates being falsely excluded from the investigation. Similarly, surveillance programmes based on detection of isolates with similar antibiograms may result in some outbreaks remaining entirely undetected.

It is interesting to note that where a difference in MGEs was observed, there was no difference in PFGE pulsotype for these isolates. Also, for isolates with a different antibiogram to the index case, the mean pairwise difference was much lower than for isolates with the same pulsotype. This could be due to the small number of samples with differing antibiograms, however it demonstrates that changes affecting phenotypic behaviour are not necessarily reflected by changes in the core genome. This must be taken into account when using WGS in outbreak investigations, with exclusion of known MGEs and screening for SNV density to identify unannotated MGE regions.

3.4.6 Comparison with routine results

In 7 of the outbreaks, the WGS results offered increased information to the investigating teams either by including or excluding isolates, or by providing additional confirmation of an outbreak involving an epidemic clone where PFGE typing alone was insufficient to distinguish between outbreak and non-outbreak isolates. In two outbreaks, data was available in real-time, providing the infection control teams with robust results which were used to reinforce good practice within the units and help contain the outbreaks (at least in the view of the infection control team).

One further putative outbreak was investigated in real-time as part of this study following a death due to MRSA bacteraemia on a psychiatric unit. Screening identified 6 MRSA carriers with epidemiological links to the index case and hence thought to be part of a cluster. In this outbreak, the WGS results were available before the results of routine analysis and identified only 1 genetically related case. There was extremely close contact between the index case and the secondary case and this was therefore considered to be an isolated transmission event rather than a true outbreak. The outbreak was therefore excluded from the analysis and there was no infection control intervention other than offering decolonisation to the carriers. Strains from the remaining 5 MRSA carriers differed by >10,000 SNVs from the index case, and were subsequently confirmed by the reference laboratory as belonging to different MLST sequence types. The rapid availability of the WGS results (before the reference laboratory results) is evidence of the progression in technology which is likely to bring WGS into routine use for *S. aureus* outbreaks in the near future.

Of note, the isolates included in this study were either from clinical samples (blood cultures, wound swabs, pus samples) or from nasal carriage, which is the predominant method of screening for asymptomatic colonisation. Consequently, the effect of colonisation at other body sites (e.g. throat, groin) on outbreak genetics is unknown. Within-host diversity of populations obtained from different sites has been reported in community (172) and hospital (114) transmission studies, but in this study the close clustering of each outbreak suggests that the outbreak population closely mirrored that found in nasal carriage. It is however possible that screening several sites may have detected more carriers / possible transmission links, however the results do suggest that the nasal carriage is likely to contribute to outbreak propagation, as well as being a risk factor for disease in the individual patient (26, 49, 130).

3.4.7 Comparison with Outbreaker Software

The Outbreaker package is one of the most widely used approaches for outbreak simulation based on sequencing data. The algorithm uses the genetic data and sample dates to reconstruct a possible transmission network, and gives statistical support for the network edges imputed. To test the validity of the Outbreaker results I applied it to a very well characterised, local neonatal unit outbreak, as there was inarguable evidence for dates of exposure, and the date of acquisition, even if acquired vertically, could not predate the birth date.

In the Outbreaker simulation, the predicted transmission pathways corresponded poorly with the available epidemiological data, and the presumed transmission pathway for the invasive, bacteraemia case (from case 3 to case 5) was not identified.

The situation is similar to that observed for the BEAST simulation in chapter 2. Using genetic data in isolation may be mathematically appealing but the evidence is improved if epidemiological data is taken into account. For the BEAST analysis, the result was more plausible when epidemiological data was included, and it is possible that a similar expansion of the Outbreaker software, allowing known exposures to be taken into account, may improve its performance. Consequently, WGS is unlikely to replace entirely traditional "shoe leather" epidemiology, but may be able to help direct the investigation since the combination of WGS and epidemiology may allow more robust conclusions to be drawn.

3.5 Conclusion

In conclusion, this chapter uses a range of definite, resolved and well characterised outbreaks to calibrate WGS. Although use of WGS in outbreak investigation has been reported, this study applies WGS in a systematic manner to a series of known outbreaks, rather than as a single case study, and addresses the potential problems caused by choice of reference genome and mobile genetic elements. The study demonstrates a difference

in within-outbreak diversity between community and hospital outbreaks, providing an insight into the transmission dynamics in each context, and also shows an association between epidemiological contact and SNV distance which may enable genomic data to be used to elucidate possible transmission routes in outbreaks where epidemiological data are lacking. The application of SNV thresholds to unresolved outbreaks will be addressed in chapter 4.

4 Complex outbreaks

4.1 Introduction

Standard outbreak surveillance comprises monitoring of overall MRSA infection and colonisation rates, with newly reported cases notified to infection control teams who review epidemiological data, identify possible related cases and determine whether a full outbreak investigation is warranted. If epidemiological links are identified between patients, isolates are submitted for typing. However, in large scale outbreaks, tracing epidemiological links may rapidly become overwhelmingly complex as there may be multiple links of varying importance between patients. In addition, patients are frequently moved between wards / units, or discharged and readmitted, a problem previously highlighted in MRSA screening programmes (173). Where there is a large time difference between isolates of an epidemic clone, it may be impossible to confirm with sufficient certainty whether there has been failure to control the initial outbreak or a re-introduction of a similar strain (46).

In chapter 3, I demonstrated that analysis of WGS data was able to distinguish correctly outbreak from non-outbreak isolates compared with standard methods. To assess the added value of WGS in complex outbreaks, I have applied the WGS approach developed in chapter 3 to investigate 5 further outbreaks considered to be “unresolved” by standard typing. These outbreaks were either prolonged, with uncertainty about whether later cases were due to re-emergence of the original strain or a new introduction of a similar strain, or contained a large number of cases such that tracing individual patient-patient links had become prohibitively complex. Four of these outbreaks were still ongoing at the time of analysis, allowing real-time feedback and discussion with the investigating teams.

In chapter 2, I addressed the issue of within-host diversity and whether this might affect interpretation of an outbreak. The results demonstrated that there is limited diversity in early acquisition of nasal carriage. Rather than confound the outbreak investigation, I have hypothesised that an outbreak with higher diversity might indicate a long-term carrier as a

source, reflecting the diversity in that individual accumulated over time. To test this, I have assessed the WGS and epidemiological data for all 20 outbreaks (including those used in chapter 3) for the contribution of a possible long-term carrier (LTC), and analysed the sequences using BEAST, to see if the WGS information can be used to predict the involvement of an LTC in maintaining an outbreak.

4.2 Methods

4.2.1 Sampling frame

Five unresolved outbreaks were identified for investigation in conjunction with the PHE Staphylococcus reference laboratory, to reflect a number of epidemiological situations. These were 1) rapid clonal expansion in a low transmission setting, 2) a nursing home setting with prolonged contacts between patients, 3) a prolonged outbreak with possible multiple introductions, 4) a prolonged outbreak in a suspected high transmission setting, and finally 5) a large scale clonal expansion over >1 year involving multiple hospitals and multiple introductions. Details are given in table 4.1.

4.2.2 Sequencing and phylogenetic analysis

Single colonies from frozen samples stored by the reference laboratory were extracted and sequenced as described previously. Reads were assembled to MRSA252, masked for repetitive regions and known mobile elements, and phylogenetic trees constructed as described in previous chapters. For outbreak P, reads were also assembled to the index case, which had previously been whole genome sequenced and assembled using a combination of mapping and de novo assembly to produce a closed reference genome (GenBank: HE579069.1 (174)).

TABLE 4.1 EPIDEMIOLOGICAL FEATURES OF COMPLEX OUTBREAKS ANALYSED BY WGS

Outbreak	Category	Reason for outbreak investigation	Phenotype	Clonal complex / MLST	Duration	No of cases	Question to be addressed by WGS
P	Hospital: multiple wards	Increase in MRSA colonisation	MRSA	ST 228	3 months	54	Does WGS give any further information regarding transmission pathways?
Q	Community: nursing home	2 clusters of MRSA infections occurring 9 months apart	MRSA	CC30	9 months	9	Is this a single outbreak or is it two introductions of a prevalent strain (CC30, widespread in the UK)?
R	Hospital: surgical unit	Increase in MSSA wound infections	MSSA	ST 2021	6 months	6	Is there an outbreak strain or is this outbreak simply due to high MSSA transmission secondary to breakdown of infection control?
S	Hospital: neonatal unit	Recurrent MRSA infection / colonisation	MRSA	CC22	4 years	43	Is there evidence for a long term HCW carrier given the repeated MRSA outbreaks on this unit over a period of 4 years?
T	Trust: multiple wards in three hospitals	Increase in mupirocin resistant MRSA	MRSA	CC8 / ST 8	18 months	271	Does WGS give any further information in the context of a large scale clonal expansion?

4.2.3 Outbreak analysis

To analyse the outbreaks and determine whether isolates were likely to be related in the absence of a known direct epidemiological link, the thresholds determined in chapter 3 were applied to each outbreak. Isolates within 12 SNVs of their nearest neighbour were considered to belong to the outbreak, while isolates within 6 SNVs were considered to be highly likely to have a close epidemiological link. Where isolates were within 6 SNVs, epidemiological information was examined to determine if there was a direct or indirect link. The results were discussed with the investigating teams to determine the influence, if any, on the outbreak investigation.

For all outbreaks, background samples were obtained from the wider hospital to prevent confounding by local clonal expansion. Details of the background isolates are given in table 4.2 below.

For outbreak P, over 500 cases were identified over an 18-month period as the strain became established in the hospital. To evaluate whether WGS could identify transmission in the initial outbreak period, before the strain became endemic, only isolates from the first three months were obtained for sequencing. Outbreak isolates in this case were identified by the investigating team as highly related if they shared the same *spa*- and double locus sequencing type (DLST) (175), rather than PFGE.

To determine whether WGS data could be used to identify possible transmission events in the context of a prolonged, complex outbreak involving multiple hospitals (outbreak T), the 6 SNV threshold was used to provide a cut-off for epidemiological investigation. A phylogenetic tree was constructed and used to identify unique clades, and possible epidemiological links sought only for patients within the same clade. Data regarding admission / discharge dates and outpatient attendances were collected from the hospital databases. Finally, the results were shared with the infection control team to compare with the outcome from the standard investigation.

4.2.4 Estimation of time to most recent common ancestor (TMRCA) for all outbreaks

All twenty outbreaks (calibration outbreaks from chapter 3 plus complex outbreaks from this chapter) were analysed to assess the possibility of the involvement of a long-term carrier (LTC, carrying ≥6 weeks).

The possibility of an LTC was defined as follows:

1) LTC not suspected: direct contact between cases (e.g. ward, household or school), each with no history of pre-existing staphylococcal disease
2) evidence for a pre- or peri-outbreak LTC: either ≥1 case with prior history of recurrent staphylococcal disease, or non-overlapping hospital stays (cases occurring in patients on the same ward after a case-free interval, or cases occurring in patients in the same hospital but not on the same ward, indicating a possible healthcare-worker carrier)
3) evidence of a post-outbreak LTC: ≥1 case with positive nasal swab >6 weeks after initial swab.

To evaluate the relationship between outbreak diversity and the likelihood of a long-term carrier, taking into account outbreak duration, time to most recent common ancestor (TMRCA) was assessed using BEAST v1.8.1 (122). A standardised simple HKY substitution model was applied with a coalescent model and constant population size. The substitution rate was standardized for all outbreaks at a rate of $3.3x10^{-6}$ per genome per year (133). Priors were set as follows: nucleotide frequencies: uniform prior; κ: log-normal prior with mean 1 and SD 1.25 on a logarithmic scale. Operators were set to auto-optomise. MCMC chain length was set to 10,000,000 with sampling every 1000 iterations and a burn in of 100,000 iterations, to obtain an effective sample size (ESS) of >200 for each iteration. Each sample was run in duplicate and the chains merged to obtain the final results. To control for differences in outbreak duration, outbreaks were

censored at six months, and the (censored) outbreak duration subtracted from the calculated TMRCA to obtain a duration-adjusted TMRCA.

The ability of TMRCA to differentiate between outbreaks where an LTC was involved compared with outbreaks with no evidence for an LTC was evaluated using a receiver-operating-characteristic curve analysis.

TABLE 4.2 DETAILS OF BACKGROUND / COMPARISON ISOLATES

Outbreak	Source of background isolates
P	All ST 228 isolates from the hospital 9 months prior to and during the outbreak
Q	Unrelated t019 clone isolates from same geographical region
R	All MSSA from wound infections over the outbreak period
S	MRSA bacteraemia isolates from the outbreak period, plus hospital-wide MRSA isolates from a 2 week period during the investigation
T	All mupirocin resistant isolates from trust during the outbreak period

4.3 Results

4.3.1 Outbreak P

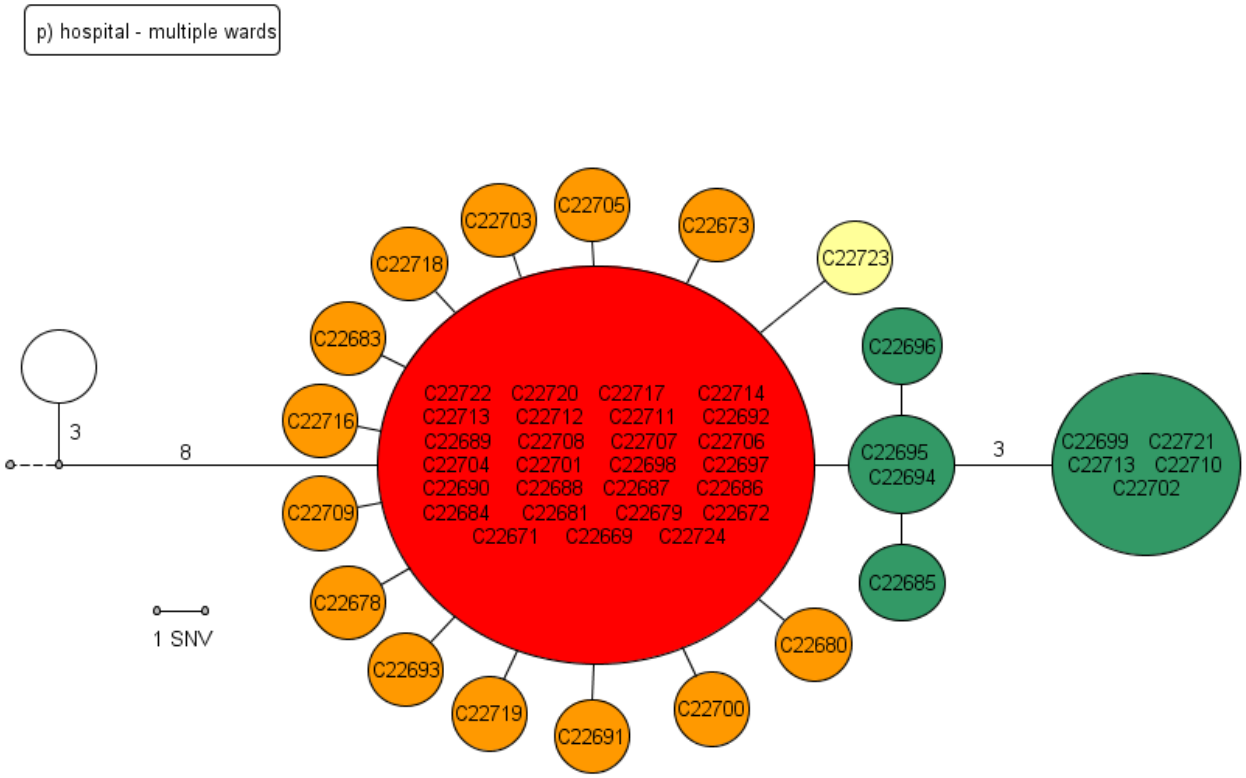
54 putative outbreak isolates were sequenced. Four were >50 SNVs distant from the remaining isolates and were considered to be sporadic. Across the remaining 50 isolates, the mean pairwise distance was 1.17 SNVs.

27/50 outbreak isolates were identical at the core genome level (figure 4.1). Thirteen isolates had single, unique SNVs differentiating them from this predominant clade, and 1 isolate had 2 unique SNVs. The remaining 9 isolates formed a separate, smaller sub-cluster (clade 2).

8 wards were involved in the outbreak. However, fewer than half the patients (22/50) remained on a single ward during their inpatient stay. The remaining 28 had inpatient stays involving up to 4 different wards and in many cases had multiple transfers between wards, resulting in almost all patients having multiple direct and indirect ward links (figure 4.2).

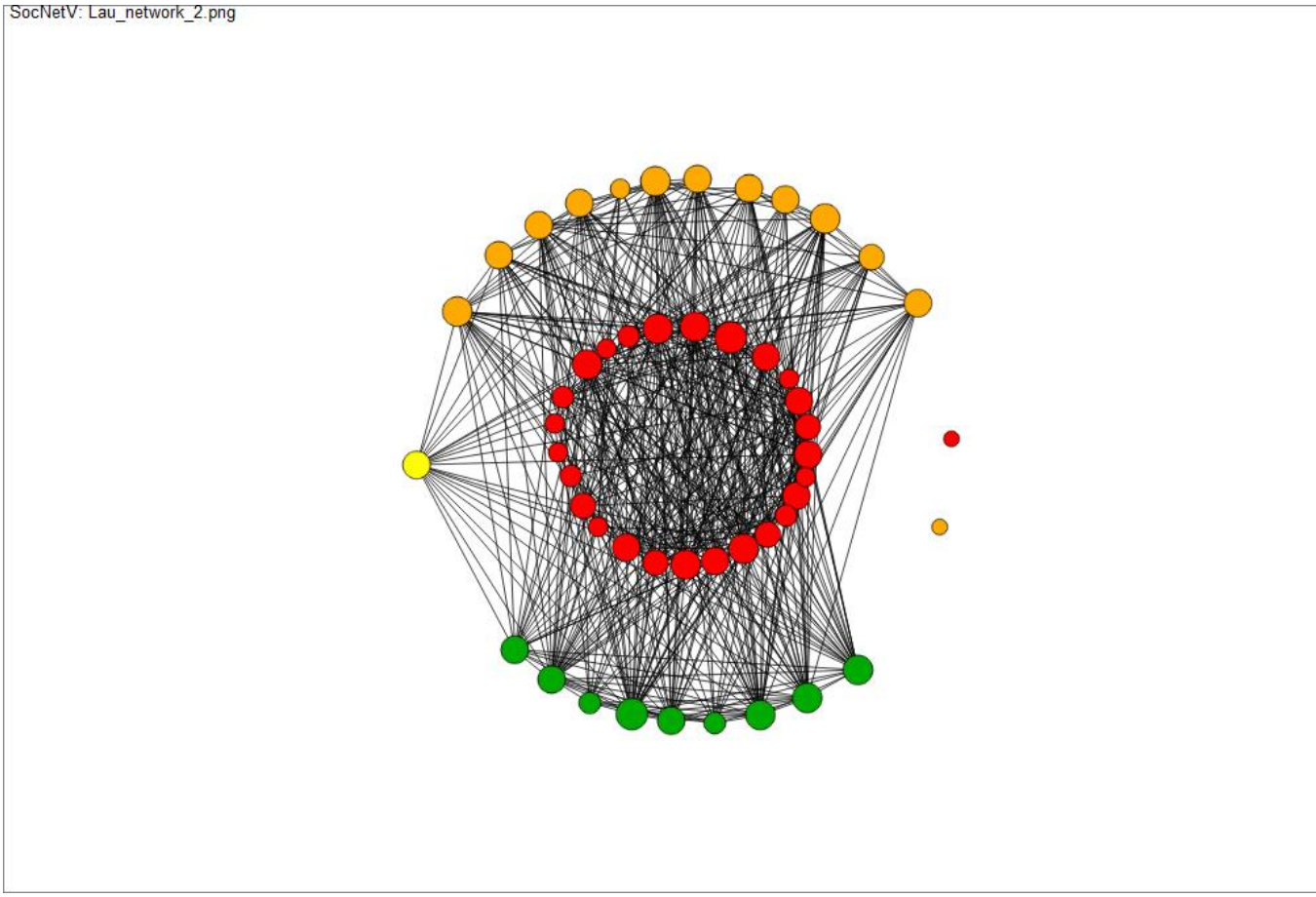
Clade 2 involved 2 wards, with all patients having a direct ward link with at least one other case in the clade, demonstrating a correlation between genetic and epidemiological clusters (figure 4.3).

FIGURE 4.1 PHYLOGENETIC TREE FOR OUTBREAK P



Red: clade 1; yellow: clade 1 derived singletons; green: clade 2.

FIGURE 4.2 NETWORK OF WARD LEVEL LINKS BETWEEN CASES

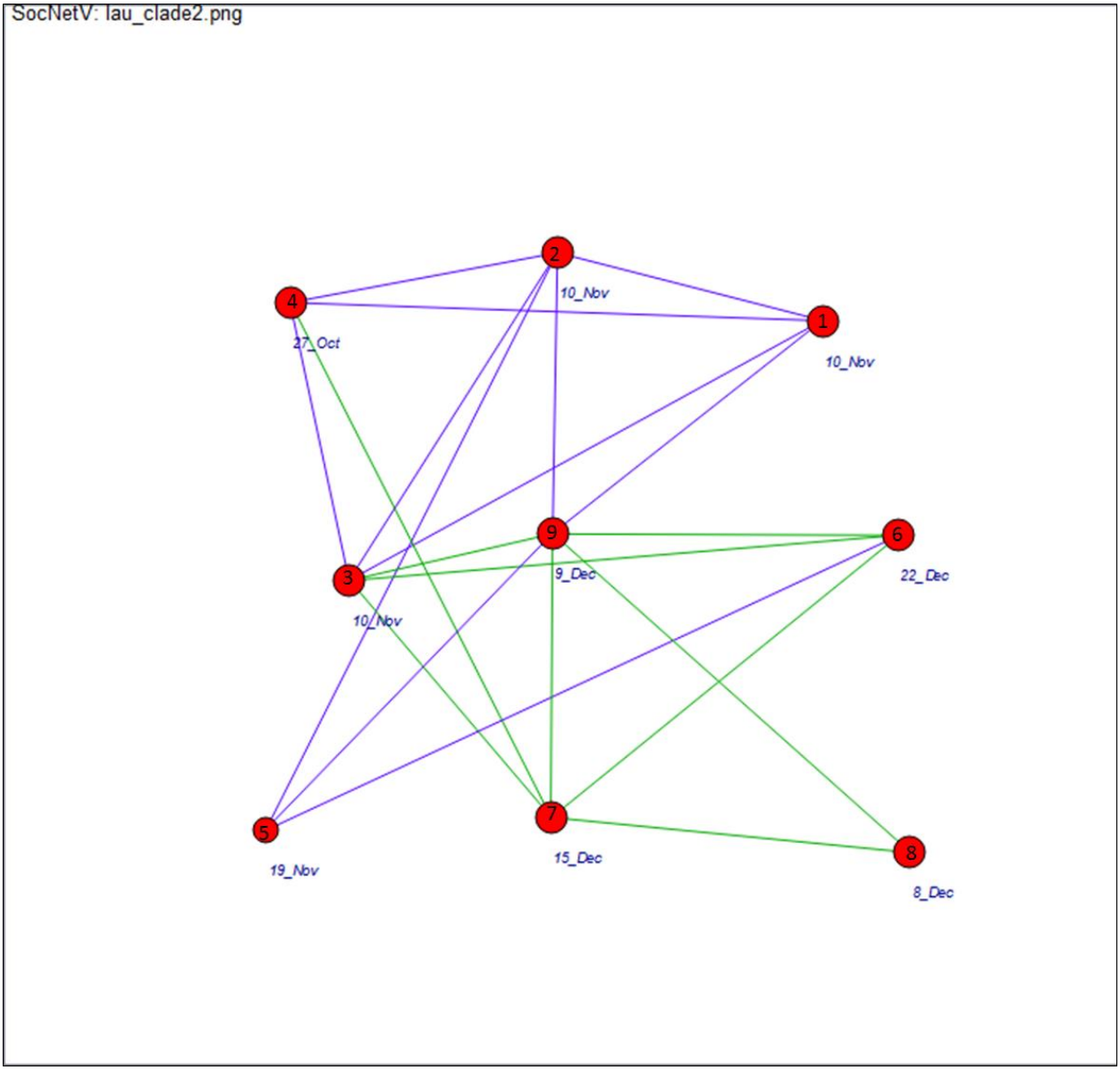


Red: clade 1; yellow: clade 1 derived singletons; green: clade 2.

Two cases have no ward links: one case was an ambulatory care patient, and the other was identified on a paediatric ward.

Size of circle is proportional to number of links.

FIGURE 4.3 NETWORK OF DIRECT WARD LINKS BETWEEN CASES IN CLADE 2, WITH SAMPLE DATE



Edge colour denotes ward: blue: ward 17, green: ward 5.

All cases except 1 and 5 had previous negative screens.

Mapping to the index case did not affect the phylogeny. Branch length was increased by 1 SNV for one isolate, but reduced by 1 SNV for one singleton and for the 5 identical isolates in clade 2. Overall mean pairwise difference was 1.20 SNVS, compared with 1.17 SNVs for mapping to MRSA 252.

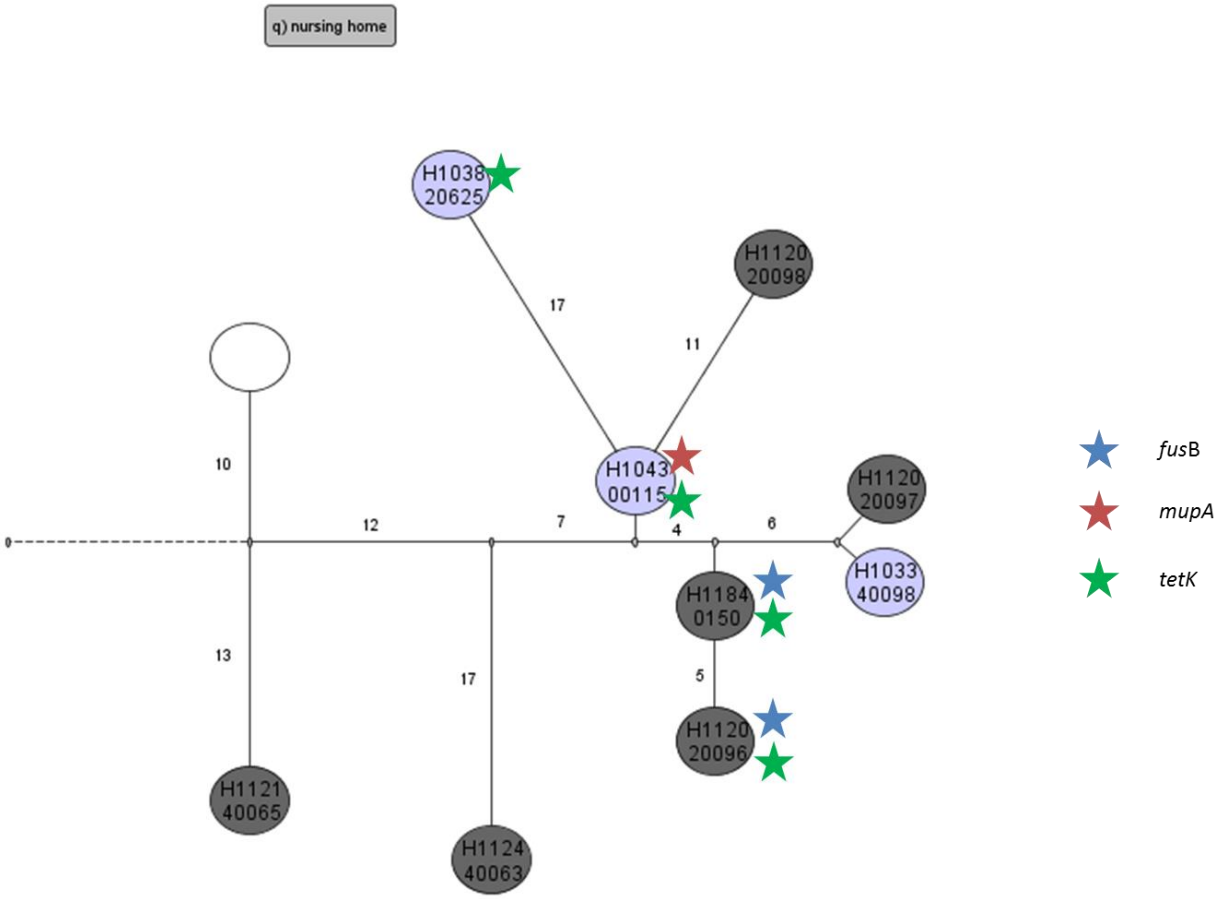
All of the outbreak strains contained *aacA-aph*D, *blaZ*, *mec*A and *erm*A, conferring resistance to gentamicin, penicillin, methicillin and erythromycin respectively. In addition, all of the outbreak isolates contained a combination of *qac*A, and a valine-to-phenylalanine mutation at position 588 of the *ile*S gene, associated with reduced susceptibility to antiseptics and low-level mupirocin resistance (176). Only 11/21 (52%) of the non-outbreak comparator strains had this combination. There was no difference in virulence profiles between the outbreak and non-outbreak strains.

4.3.2 Outbreak Q

9 putative outbreak isolates were sequenced. The first 3 cases occurred over a three-month period, with a further 6 cases occurring six months after the third case. It was unclear from standard typing results whether this was the original strain or a new introduction of a similar strain.

6 cases were within the 12 SNV threshold for belonging to an outbreak. 2 cases were definitely excluded on the basis of greater distance from the outbreak strains than the comparator strain (figure 4.4). One case was within 17 SNVs of the outbreak strain and was considered possibly related. One of the isolates from the first wave of the outbreak was only 3 SNVs distant to one of the second wave isolates, confirming a definite transmission between these 2 cases and that the original outbreak had not been controlled.

FIGURE 4.4 PHYML TREE FOR OUTBREAK Q



First wave samples are indicated by light grey, second wave by dark grey. Isolates H112440063 and H112140065 are more distant than the nearest unlinked isolate (indicated by empty circle) and are therefore considered to be sporadic. H103820625 is possibly linked.

The results of the WGS analysis confirmed that 6/9 isolates belonged to the outbreak, with distances between early and later isolates compatible with a close epidemiological link and indicating that the initial outbreak had not been controlled. 2 isolates were excluded on the basis of their distance compared with unlinked isolates of the same *spa* -type.

Variation was seen in the carriage of *fus*A (present in 2 isolates) *mup*A (1 isolate) and *tet*k (four isolates), annotated on the PhyML tree above.

Mean pairwise distance for the definite outbreak strains (within 12 SNVs) was 12.4 SNVs.

4.3.3 Outbreak R

92 isolates were submitted to the reference laboratory. Samples included patient clinical isolates, staff nasal screening swabs and environmental samples. 6 isolates belonged to the suspected outbreak strain (ST 2021), and a further 60 isolates belonged to 15 other MLST sequence types (figure 4.5). All isolates were sequenced as there was concern that there were multiple clusters in the background sampling suggesting a high level of background transmission.

Mean pairwise difference for the outbreak strain was 15.6 SNVs. There were two genetic clusters of isolates within 12 SNVs and therefore considered to be definitely related (patients 2, 5 and 6, and patients 1 and 4, figure 4.6). The two clusters were separated by 14 SNVs, and one isolate (from patient 3) was 13 SNVs from its nearest neighbour and therefore possibly related.

FIGURE 4.5 OVERALL PHYLOGENY OF MSSA ISOLATES COLLECTED DURING THE OUTBREAK PERIOD

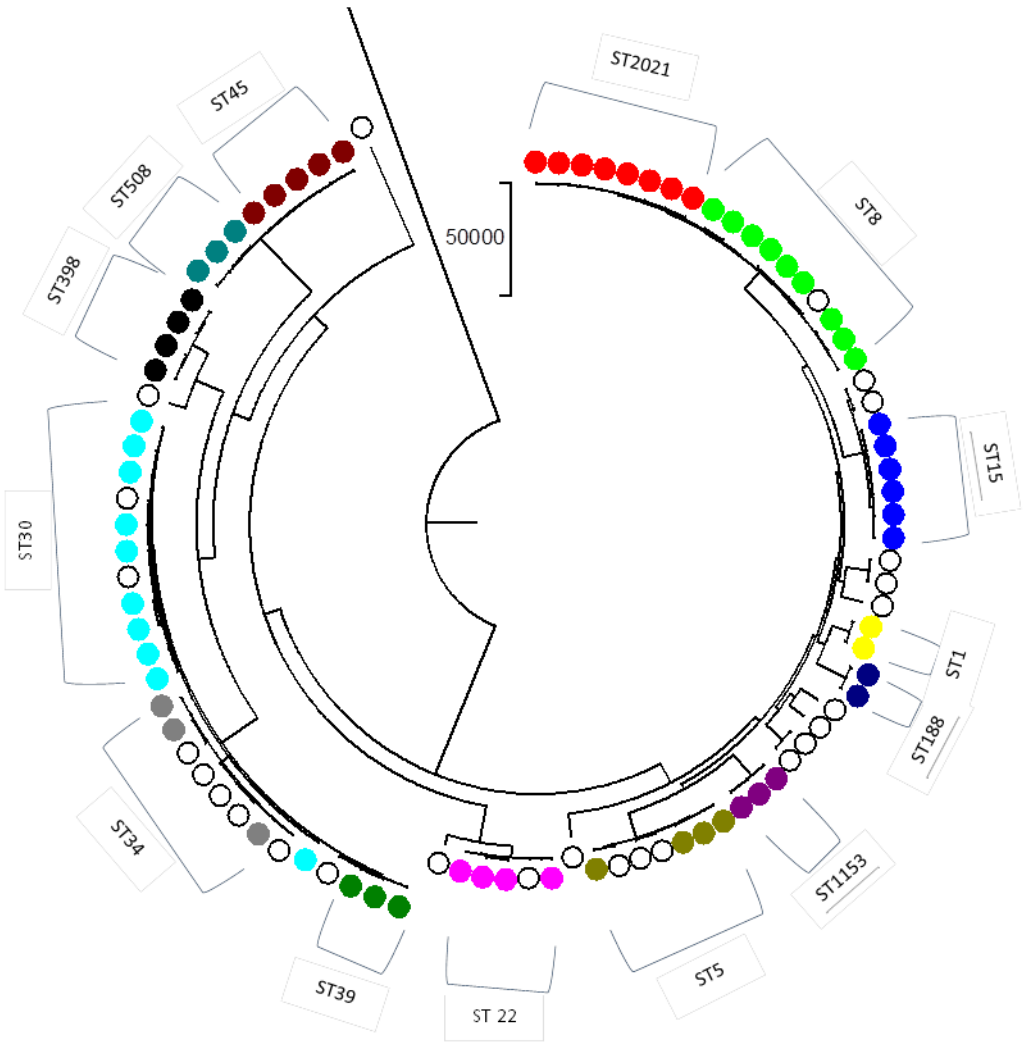
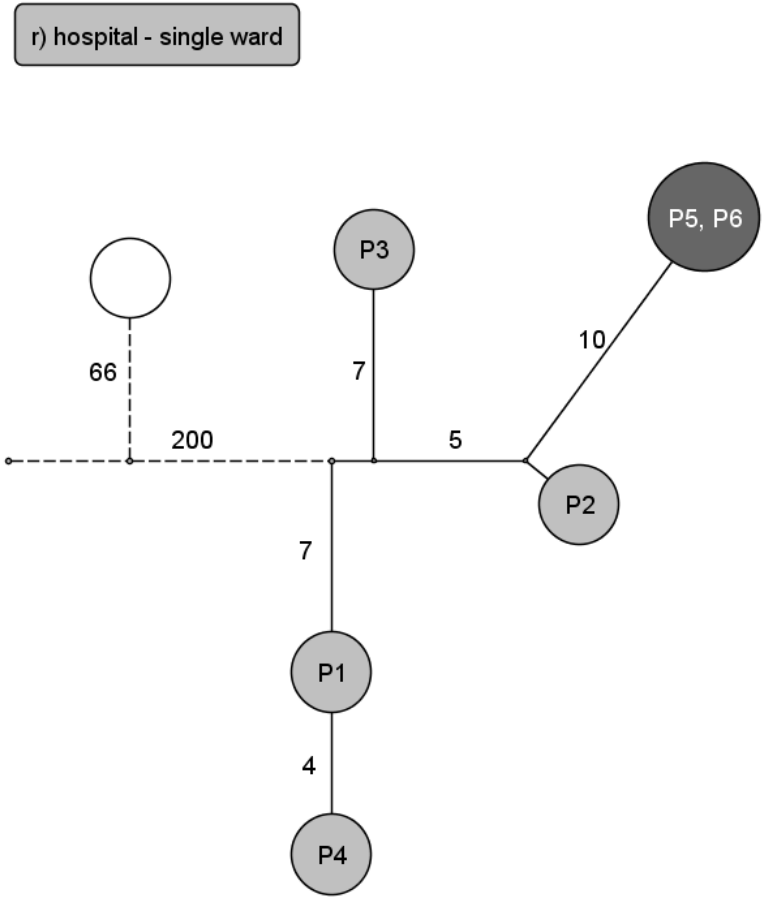


FIGURE 4.6 PHYML TREE FOR OUTBREAK STRAIN ST2021



Light grey: epidemiological cluster 1; dark grey: epidemiological cluster 2.

There were 411 days between the earliest and last isolates. Median duration of inpatient stay was 156 days (range 51-277). There were two epidemiological clusters: cases 1-4 had overlapping ward stays over a period of 214 days (epidemiological cluster 1), while cases 5 and 6 also had overlapping ward stays over a period of 54 days (epidemiological cluster 2). There was a single day between case 2 being discharged and case 6 being admitted.

Among the non-ST2021 strains, genetic pairs were identified in 6 of the MLST groups (table 4.3). However, none of these were from patient-patient links. Two were staff-patient, 2 were staff-staff, 1 was environment-staff and one was environment-patient. No staff member was found to carry the outbreak strain, nor was it found in any of the environmental samples.

TABLE 4.3 EPIDEMIOLOGICAL DETAILS FOR GENETICALLY LINKED, NON-OUTBREAK ISOLATES FROM OUTBREAK R

MLST	Isolate ID	Date of sample	Source	SNV difference
ST8	NECT-467	24/1/12	Theatre settle plate	12
	NECT-485	24/1/12	Staff-2	
ST30	NECT-042	25/1/12	Staff-4	2
	NECT-479	8/3/12	Patient-A	
ST188	NECT-156	22/9/12	Patient-B	2
	NECT-152	17/10/12	Bath	
ST15	NECT-296	22/10/12	Staff-5	2
	NECT-261	11/12/12	Patient-C	
ST398	NECT-305	22/10/12	Staff-6	4
	NECT-306	22/10/12	Staff-7	
ST not found	NECT-388	16/9/11	Staff-10	9
	NECT-481	24/1/12	Staff-11	

4.3.4 Outbreak S

43 cases were identified over a four-year period: 13 from 2009, 10 from 2011, 14 from 2012 and 6 from 2013. No isolates were available for 2010: although there were MRSA cases during this time, a connection to the 2009 outbreak was not suspected and the isolates were not stored.

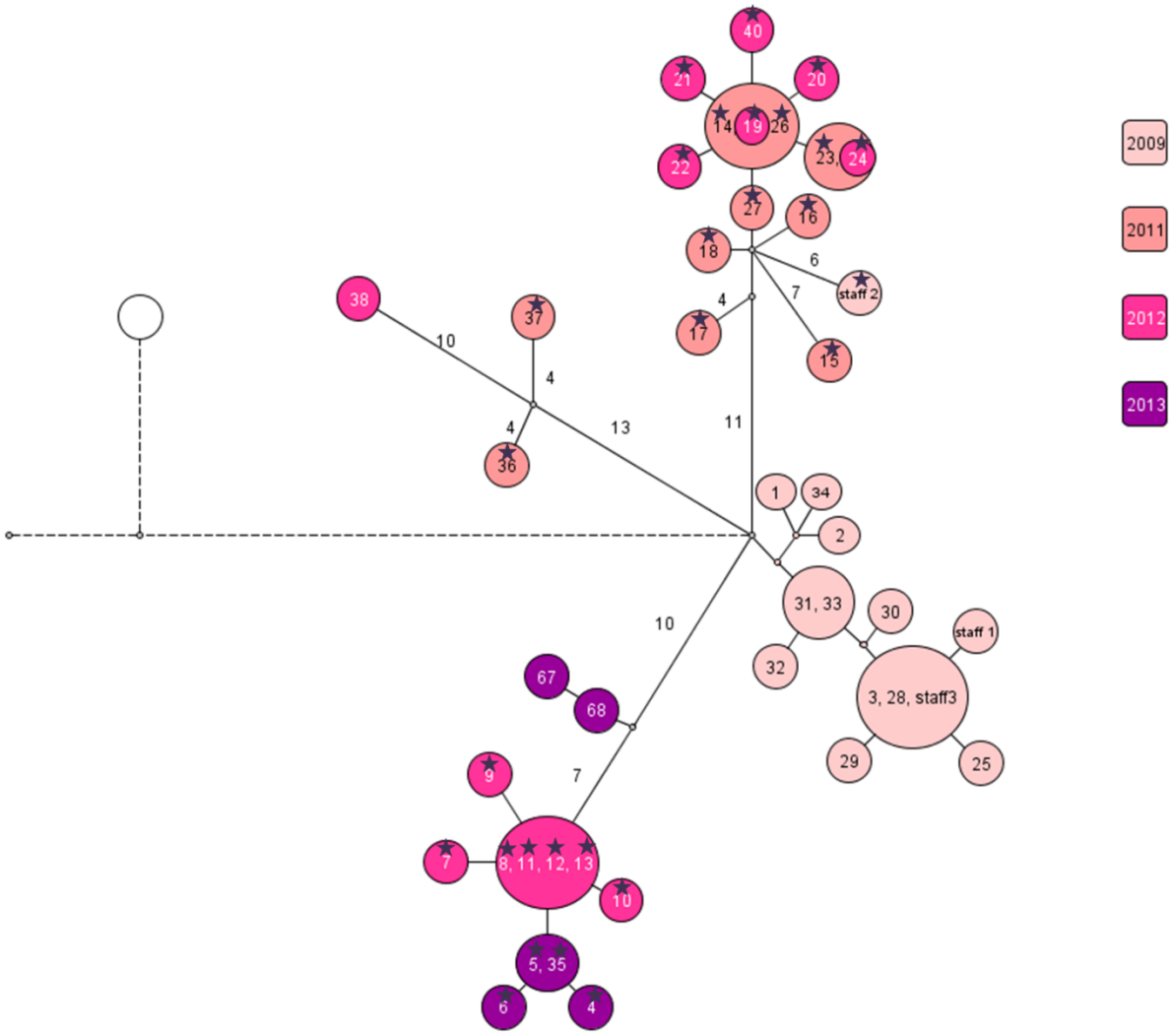
Overall mean pairwise distance was 19.8 SNVs (95% CI 19.0-20.6). The phylogenetic tree shows four clades, separated by a maximum of 32 SNVs (figure 4.7). Clade 1 contained all except one of the isolates from 2009. Clade 2 contained isolates from 2011 and 2012, clade 3 contained isolates from 2012 and 2013, and clade 4 contained 3 isolates, two from 2011 and a single, more distant isolate from 2012.

Two other isolates from this *spa*-type were found in the background hospital isolates, however they were >150 SNVs distant and therefore considered to be unrelated. The strain was not found in bacteraemia isolates stored by the hospital over the outbreak period.

5 staff members were found to be MRSA carriers, and three of these were positive for the outbreak strain.

There was variation in the presence / absence of *erm*C, which confers resistance to erythromycin. Only a single sample (staff member 2) from the 2009 isolates was positive for *ermC,* compared with 27 / 30 (90%) isolates from 2011-2013 (indicated on figure 4.7).

FIGURE 4.7 PHYML TREE OF OUTBREAK S



Colour indicates year of sample. Star indicates presence of *erm*C.

4.3.5 Outbreak T

During the outbreak investigation period, 277 mupirocin resistant MRSA isolates were sent to the reference lab, of which 227 were *spa*-type t008. Two were identified as USA300 by PFGE and were PVL positive: these were >3500 SNVs from the outbreak strain. Of the 225 non-USA300 isolates, 4 (1.8%) were >100 SNVs distant and were considered to be sporadic.

The remaining isolates were highly clonal, with a maximum of 14 SNVs between nearest neighbours and an overall mean pairwise difference of 15.6 SNVs (figure 4.8).

Using a 6 SNV threshold to define closely related isolates, the phylogenetic tree was collapsed into 9 clades containing 3 or more isolates, separated by 6 or more SNVs from the next nearest neighbour, plus 6 pairs of isolates within 6 SNVs of each other (figure 4.9)

The largest clade contained 94 isolates, and included strains from all three hospitals over the entire outbreak period (figure 4.10). For this clade, WGS data provided minimal additional data compared to the routine investigation.

FIGURE 4.8 PHYML TREE FOR ALL OUTBREAK T ISOLATES

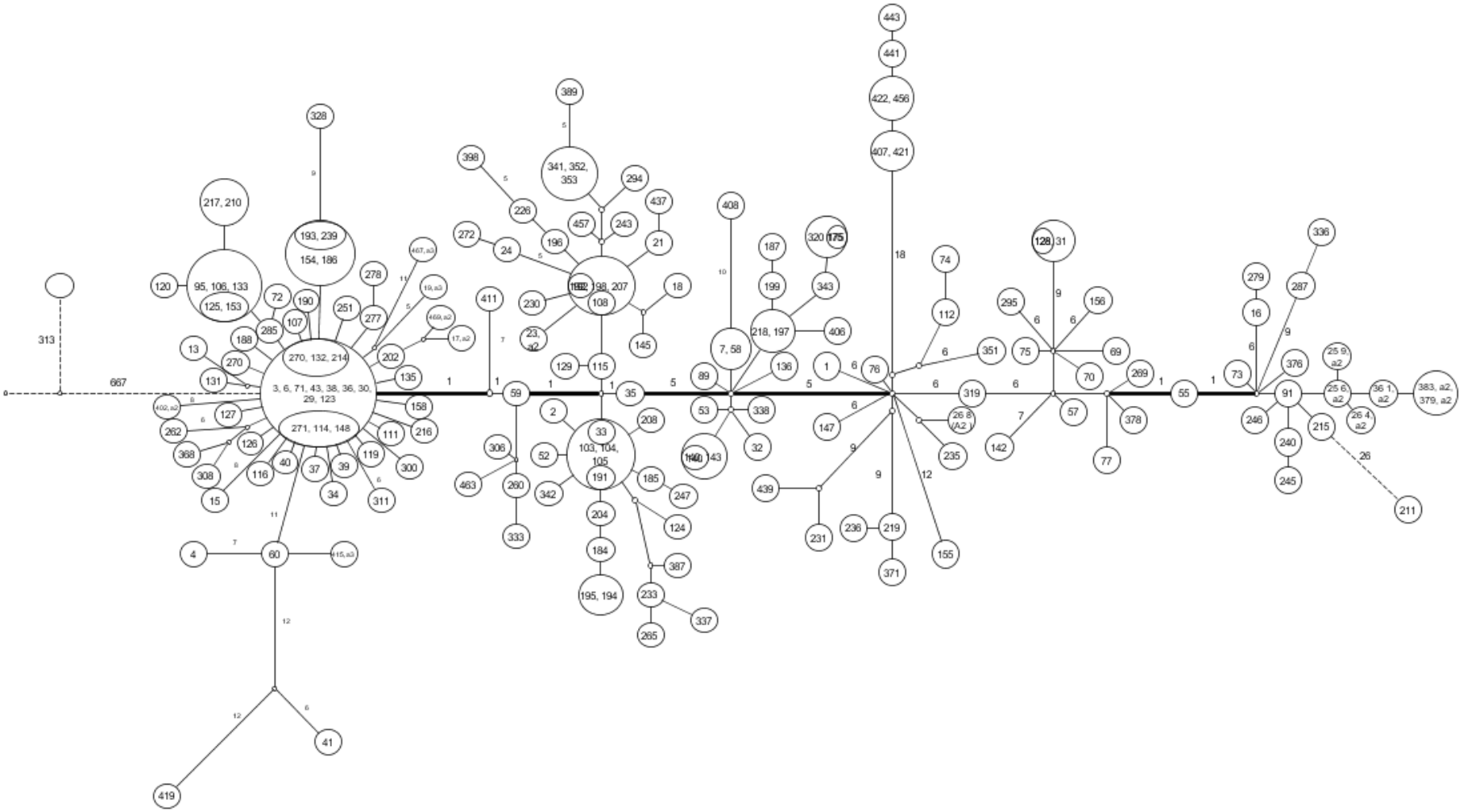


FIGURE 4.9 COLLAPSED PHYML TREE INDICATING CLADES CONTAINING 2 OR MORE HIGHLY SIMILAR ISOLATES (WITHIN 6 SNVS)

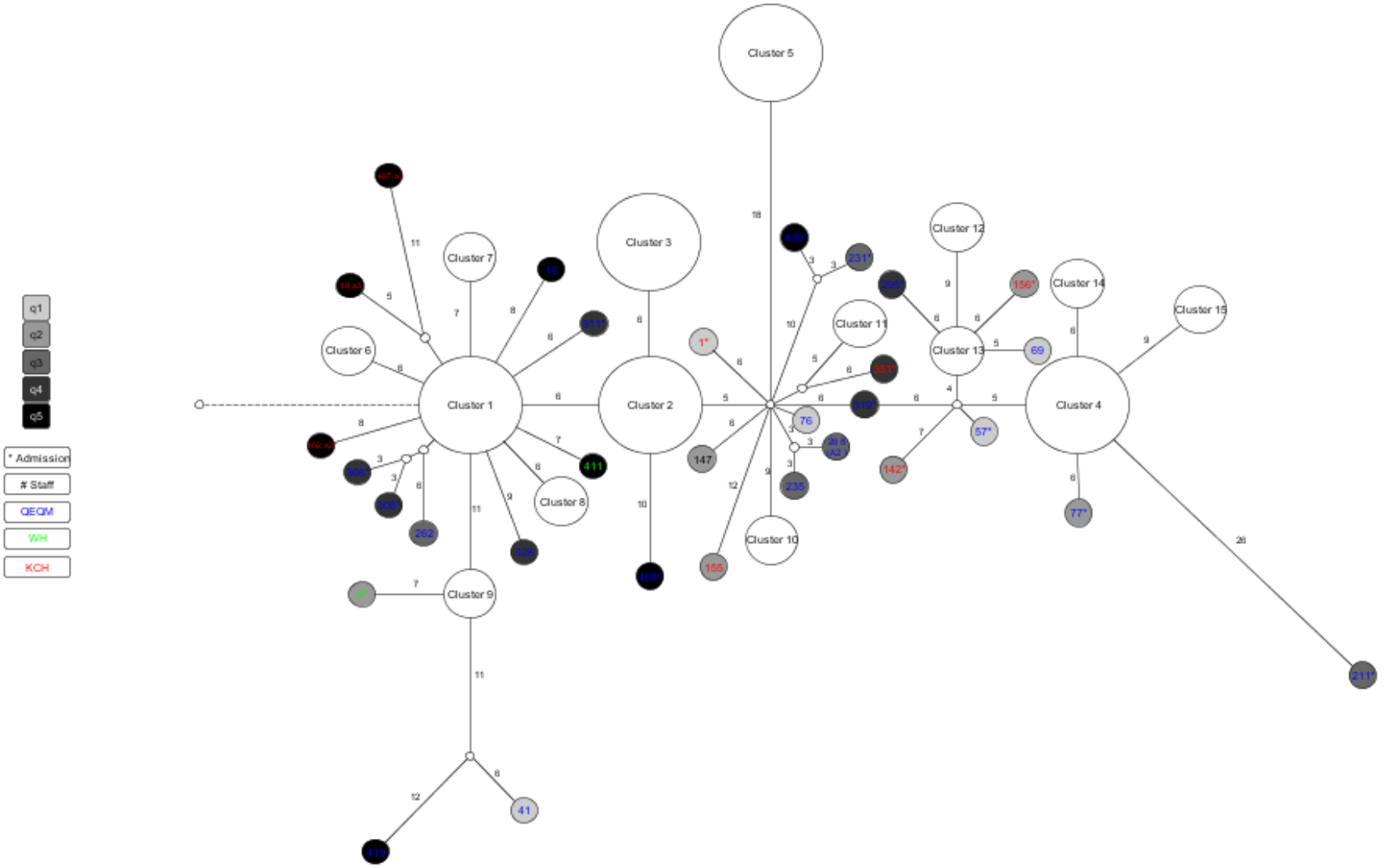
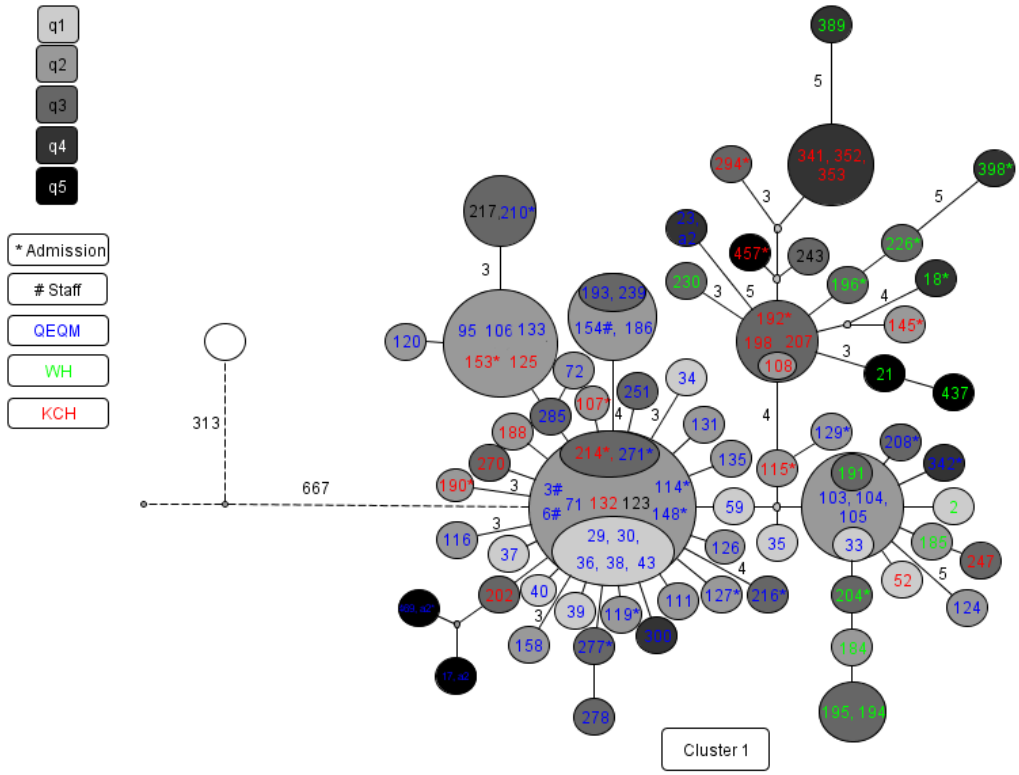


FIGURE 4.10 PHYML TREE FOR CLADE 1



Shading indicates time of sampling (pale grey: quarter 1->black: quarter 5).

Label colour indicates hospital in which sample was taken, and asterisk indicates that the sample was taken on admission.

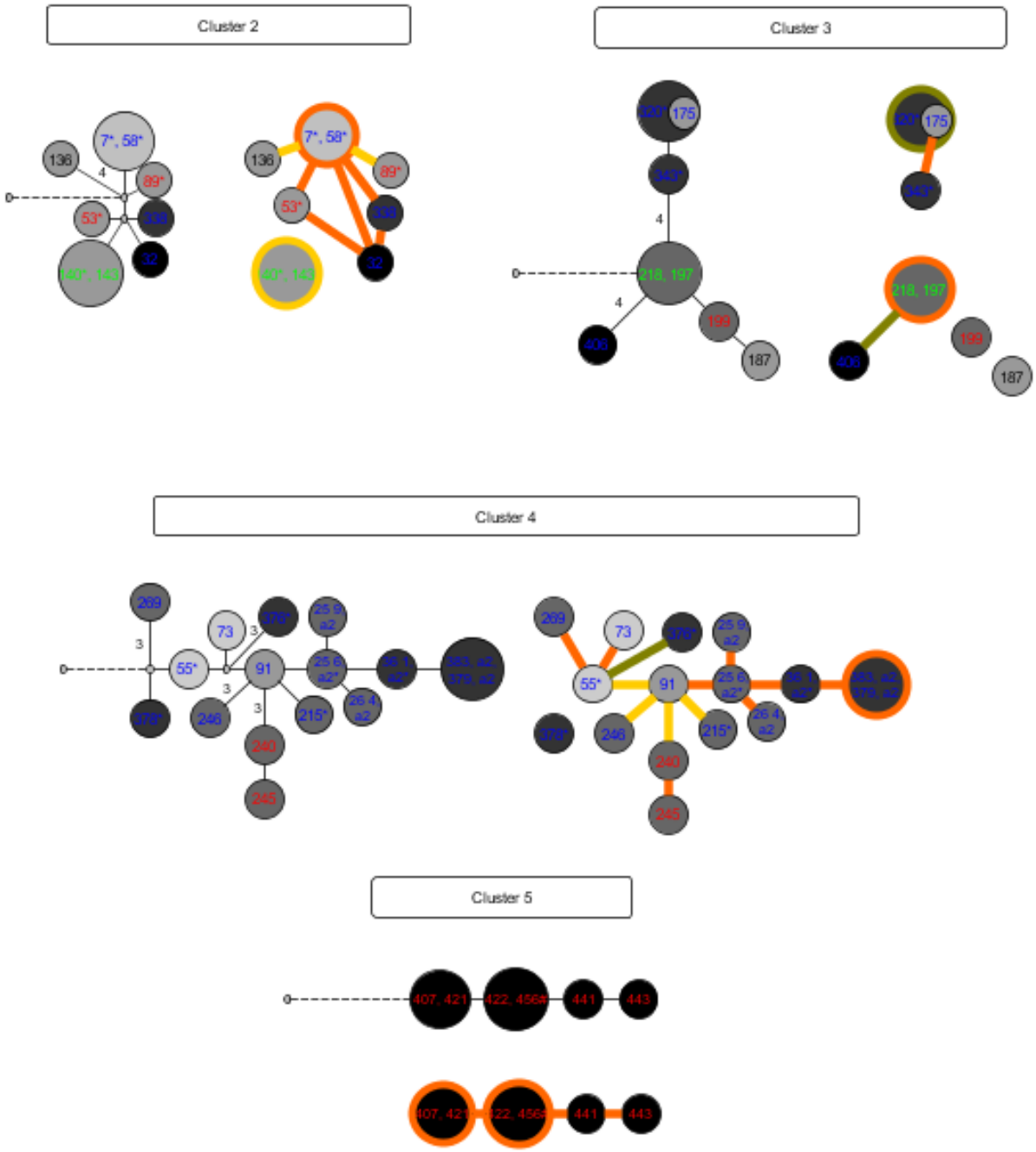
The remaining 8 larger clades (3 or more patients) comprised 54 patients. Examining only the links within a clade, direct ward links (overlapping ward stay) were found for 32 patients, indirect ward links (non-overlapping ward stay or overlapping hospital stay on different wards) were found for 17. There was a single healthcare worker-patient link. For four patients, no link could be found for any of the patients in their clade. PhyML trees with epidemiological links are shown for the clusters containing more than 3 cases in figure 4.11.

There were also 31 singletons which did not form part of a distinct clade. In 16 of these, patients were positive on admission and had therefore presumably acquired the strain on a previous admission (or possibly in the community). There was an association between clade and hospital in which sample was obtained (figure 4.12), although this was confounded by the fact that, in many cases, patients were transferred between units or had had previous admissions to another hospital within the region.

Unlike USA300, all isolates were negative for PVL and the arginine catabolic mobile element (ACME) cluster. All were positive for *mec*A, carried on SCC*mec*IVc, distinct from USA300 which harbours SCC*mec*IVa.

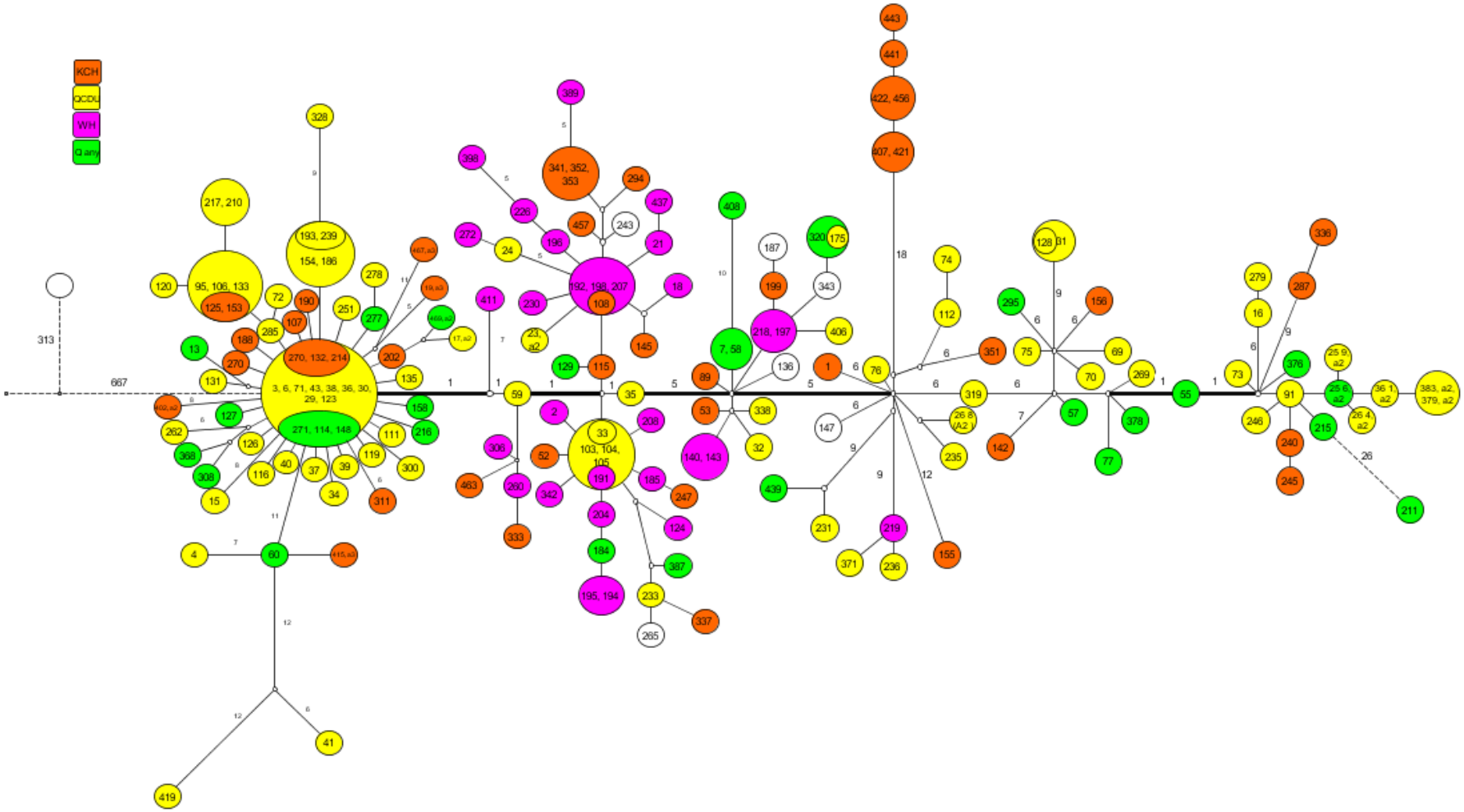
223/225 (99%) isolates were positive for immune evasion cluster genes *sak* and *scn*, and encoded enterotoxin gene A (*sea).* 213/225 (95%) harboured a plasmid carrying *blaZ* and *qac*A, associated with reduced chlorhexidine / antiseptic susceptibility. 189/225 (84%) contained a plasmid carrying *mup*A, *dfr*A and *erm*C genes, conferring resistance to mupirocin, trimethoprim and erythromycin respectively.

FIGURE 4.11 PHYML TREES WITH EPIDEMIOLOGICAL LINKAGE FOR CLADES 2-5



Orange: direct ward link, yellow: indirect ward link, green: direct hospital link.

FIGURE 4.12 PHYML TREE INDICATING WARD ON WHICH SAMPLE WAS OBTAINED

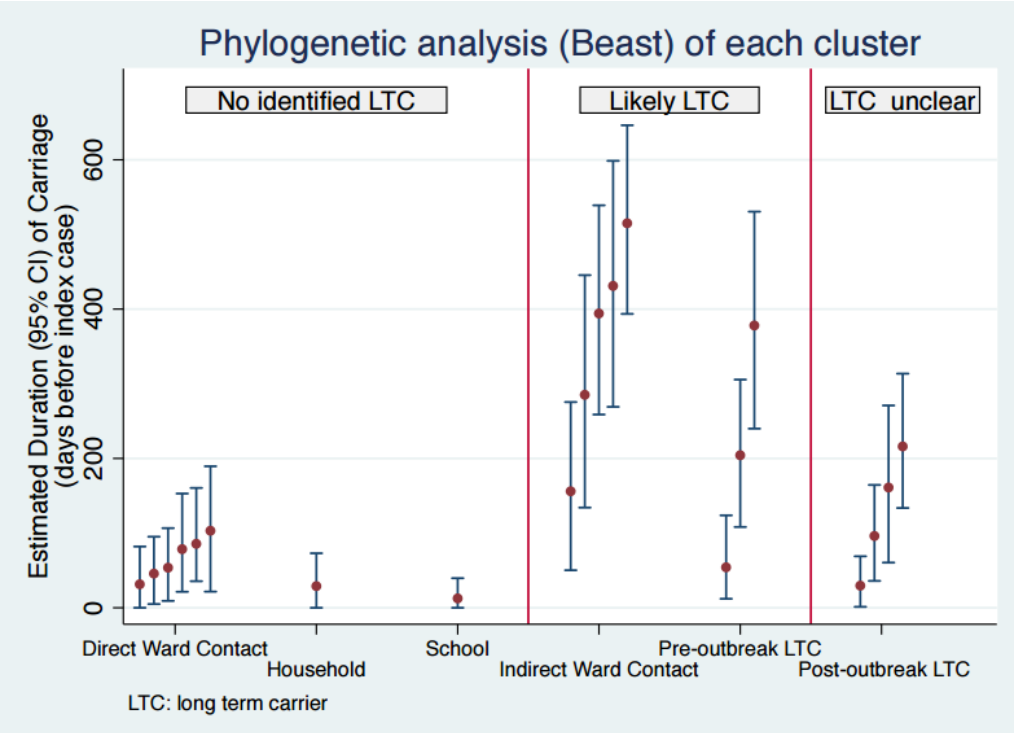


Yellow: hospital 1, admission unit; orange: hospital 1, other wards; pink: hospital 2; green: hospital 3; no colour: day case / community source

4.3.6 Use of duration adjusted TMRCA to predict the involvement of a long-term carrier

Twelve outbreaks (60%) had epidemiological evidence of an LTC: 3 included cases with a history of recurrent staphylococcal disease (pre-outbreak LTC), 4 included cases with positive nasal swabs >6 weeks after their initial swab (post-outbreak LTC), and in 5 outbreaks, an LTC was suspected due to non-overlapping ward stays of affected patients (peri-outbreak LTC). Mean duration-adjusted TMRCA for outbreaks with a suspected or proven LTC was 243 days (95% CI 143-343) compared with 55 days (95% CI 28-81) for outbreaks with no evidence for an LTC (p=0.004, figure 4.13). Analysis of the receiver-operating-characteristic curve for outbreaks with convincing evidence for an LTC (8 outbreaks, see figure 4.13) versus outbreaks with no evidence of an LTC (8 outbreaks) gave an AUC of 0.953 (95% CI 0.851-1). Using the Youden index to select the optimal threshold gave a cut-off value of 156 days, with a sensitivity of 0.875 and a specificity of 1.

FIGURE 4.13 DURATION-ADJUSTED TMRCA FOR OUTBREAKS WITH: NO EVIDENCE OF AN LTC, LIKELY LTC, OR LTC INVOLVEMENT UNCLEAR



4.4 Discussion

4.4.1 WGS confirms clonal expansion (Outbreak P)

This outbreak involved a low transmission setting (a Swiss hospital with low rates of MRSA) with a large increase in the incidence of MRSA following the transfer of an MRSA positive patient from another hospital. The strain was identified as ST228, also known as the South German / Italian clone, prevalent in Geneva and Lausanne but not in other parts of Switzerland. Although it had previously been seen in this hospital and had been responsible for 2 prior outbreaks in the early noughties, it became rare until a re-introduction in 2008 (174). The initial increase in cases occurred over a three-month period following the transfer of a known carrier from a hospital outside the region. Although the index case was rapidly identified and intensive infection control procedures instituted, the hospital experienced a significant increase in MRSA colonisation resulting in the infection / colonisation of over 500 patients over the next 18 months.

Even when mapped to the index case to maximise resolution, the outbreak was markedly clonal with a highly conserved core genome. Across the outbreak only 21 individual SNVs were identified: these were spread throughout the genome, and the majority occurred sporadically in single isolates only and were therefore not informative. There was a single genetic cluster, one SNV distant from the main clone, for which direct ward links could be traced for all patients and for which WGS proved more informative than standard typing.

Overall, the results convincingly argue for a rapid clonal expansion of this strain. This provides evidence for the conclusions from the preceding chapters, that in many outbreak settings there is insufficient time for diversity to accumulate within each individual and the effect of the cloud is minimal. The results suggest a point source introduction, with rapid person to person spread.

This outbreak also highlights the importance of ensuring an appropriate sampling frame, as a close epidemiological link might be assumed between any of the individuals in this

outbreak on the basis of genetic similarity, when in fact the hospital-wide prevalence of the strain means that any two MRSA isolates from the hospital over this period were highly likely to be similar even in the absence of direct epidemiological links.

The mapping results provide further evidence for the use of a single standard reference strain to identify SNVs in the core genome, as there was virtually no difference between the SNV distances for each reference, and phylogeny was entirely unchanged.

One further area of interest is the successful establishment of this clone despite intensive infection control efforts. There may have been an overall failure of infection control, but this might be expected to result in an increase in a number of MRSA strains rather than this one in particular. A contributing factor may be low-level mupirocin resistance, which was identified phenotypically and confirmed by the identification of the relevant point mutation in the *ile*S gene associated with low level resistance (176). This may have allowed evasion of eradication attempts with minimal fitness cost to the organism. Similarly, the universal presence of *qac*A may also have been a factor in driving the success of the clone.

4.4.2 WGS confirms relatedness in isolates 6 months apart (Outbreak Q)

In this outbreak, a cluster of MRSA infection / colonisation was observed in a nursing home, with re-emergence of the same lineage after an interval of almost 6 months. The strain was *spa*-type t019, known as the South West Pacific clone due to its endemicity in this geographic region (177) . In the UK, it is carried not infrequently by healthcare workers from the Philippines (178), and in this outbreak, given the re-emergence of the strain after a prolonged interval, the source was suspected to be one or more of the healthcare workers in the home, or possibly a long-term resident. Although isolates were deemed related by PFGE (7 pulsotype A, 1 pulsotype AA), differences in susceptibility were noted to fusidic acid, tetracycline and mupirocin. It was therefore uncertain whether the outbreak

could be related to a single staff member / resident or whether there had been repeated introductions (for example, due to residents being admitted to local hospitals and then coming back in to the home). The results of the WGS analysis confirmed that 6/9 isolates belonged to the outbreak, with distances between early and later isolates compatible with a direct epidemiological link and indicating that the initial outbreak had not been controlled. 2 isolates were excluded on the basis of their distance compared with unlinked isolates of the same *spa*-type, and one isolate was deemed to be possibly related on the basis of its genetic distance from the rest of the outbreaks isolates. Given the comparative rarity of the strain in the UK, it is possible that these 3 isolates were introduced by different healthcare workers from the Philippines, carrying non-identical but related strains.

All isolates were found to be positive for PVL, which is commonly seen in this lineage, however no severe or recurrent disease was noted.

In this cluster, there was increased mean pairwise diversity compared with the resolved hospital outbreaks studied in chapter 3. The phylogeny is more in keeping with those seen in the resolved household clusters, where there may have been long-term carriage and exchange of strains on multiple occasions, or with repeated introductions / transmissions from a long-term carrier e.g. a member of staff or long-term resident.

It is interesting to note lack of correlation between date difference between isolates and number of SNVs, indicating that despite the expected increased diversity over time there is conservation of the initial founder strain.

The nearest comparator *spa*-matched strain, although not directly related to this outbreak, is comparatively close to the outbreak strain. On further investigation, this isolate was found to originate from an unrelated individual from the Philippines, which may explain its relative proximity compared with the *spa*-matched strains for other outbreaks.

4.4.3 WGS confirms outbreak in a high transmission setting (Outbreak R)

This outbreak occurred in the setting of a cardiothoracic ward with an increase in the number of MSSA sternal wound infections. Ongoing active surveillance was instituted, with all isolates from wounds, staff screening and patient screening collected over a twelve-month period sent to the reference laboratory for typing. Standard investigation identified a small cluster (n=6) of isolates of identical pulsotype (ST2021, t008), however several other clusters were also noted and it was unclear whether there was ongoing transmission along with an outbreak strain (i.e. an overall breakdown in infection control rather than a point source).

Despite meeting the epidemiological criteria for an outbreak (an increase in cases above the expected rate), only a small number of the isolates were attributable to the same source. The presence of several closely related pairs is evidence for other episodes of person-to-person transmission in the unit, but none of these strains became established or caused large outbreaks. The majority of strains were sporadic, with >100 SNVs between isolates of the same MLST sequence type.

The overall tree phylogeny and SNV distances are more similar to those in a community / household cluster, which might be analogous to a long stay unit where there are multiple opportunities for transmission over time and longer intervals between sampling times. It may also fit with multiple introductions from a staff member who has long-term carriage.

Confirming the relatedness of the staff-patient sample pairs in the other MLST isolates was of value to the investigating team. Direction of transmission can only be inferred from dates, but, regardless of whether these were patient-staff or staff-patient transmissions, the results provided an added indication that overall infection control measures were inadequate, allowing onward spread of the outbreak and other clones.

4.4.4 WGS confirms recurrence of strain over 4 years (Outbreak S)

This was a prolonged MRSA outbreak on a neonatal unit, with reappearance of the same strain over a period of at least 4 years. Previous studies have found that MRSA colonisation in neonatal units is related to maternal carriage, introduction of strains from out-born neonates, and horizontal transmission by HCWs (179, 180). However, as this was not a prevalent community strain its persistence over several years, despite intense infection control measures and moving the unit to a new site, meant that there was a high suspicion of a colonised HCW as the source. Staff screening was initiated in the latter stages of the outbreak, but this was done completely anonymously and although there were positive samples, these could not be traced to individual staff members.

Although this was not a virulent clone and there was no associated increase in MRSA invasive infections, there were considerable concerns that high MRSA carriage rates would ultimately lead to an increase in MRSA disease rates. There was a negative effect on staff morale, both on the unit and among the infection control team, as well as significant parental concern.

The phylogeny suggests at least three separate introductions followed by short transmission chains. This would fit with a staff carrier introducing the strain to the unit followed by short chains of onward transmission by e.g. transient staff hand carriage. The alternative explanation is that this is a prevalent clone in the region and has been reintroduced on separate occasions by the admission of a new, colonised neonate. However, the clone was not found in the background samples and was not noted by the reference laboratory from other samples sent from that region.

Following analysis, the results were discussed with the investigating infection control team. Since the WGS data supported the team's suspicion that a staff member was involved, it was decided to progress with formal staff screening to confirm / refute this. This provides an example of WGS data being used to influence the hospital response, supporting the

suspected conclusion from the routine investigation with sufficient evidence to present at board level to permit staff screening / enhanced surveillance to be mandated.

However, shortly after the decision to proceed with screening, further information came to light regarding a staff member with a long-term health condition. This individual was known to be an MRSA carrier from a previous job at a different hospital but this information had not been shared with the infection control team or occupational health. Deterioration in this individual's health resulted in a prolonged absence from work which coincided with control of the outbreak.

On the basis of this, no further staff screening was implemented. The individual was unable to return to work due to their own health problems and so eradication was not attempted. Since their absence, no further cases have been detected.

It is tempting to speculate that had these data been available at an earlier stage, more intensive staff screening may have been instituted and earlier control of the outbreak achieved.

4.4.5 WGS provides enhanced resolution in the context of a large-scale clonal expansion (Outbreak T)

The final complex outbreak was a large increase in mupirocin resistance among MRSA isolates first noticed in a hospital in 2012. Retrospective analysis of resistance patterns indicated that the rise had started some months previously but because it was hospital-wide rather than focused in a single ward, an outbreak was not suspected. Mupirocin-resistant MRSA were sent to the reference laboratory and were identified as belonging to the Lyon clone (*spa*-type t008, SCC*mec*IVc, PVL negative). This strain is prevalent in certain parts of Europe, particularly the South of France, but is usually seen only sporadically in the UK (181).

The strain belongs to MLST sequence type 8 / clonal complex 8 which includes the highly successful USA300 clone. The resistance profile and rapid expansion observed in this outbreak led to concern that it could become similarly successful, however the WGS analysis demonstrates substantial differences and this is corroborated by the fact that there was no accompanying rise in invasive or severe cases with this clone.

Since mupirocin and chlorhexidine are the mainstay of MRSA eradication therapy in the UK (182), it is likely that resistance to these compounds may have helped to drive the successful establishment of the clone in this region. This is supported by the reduction in its prevalence following the discontinuation of mupirocin use for decolonisation by the hospitals involved.

Similar to outbreak P, in this outbreak there was a high degree of clonality which limited the use of WGS, particularly for early samples. However, WGS provided considerable added resolution compared to PFGE, and allowed the identification of multiple sub-clusters which were found to be compatible with epidemiological information. This allowed a more structured approach to the epidemiological data, with specific examination of closely related clusters, thus helping to reduce the enormous burden of tracing multiple, complex epidemiological links.

4.4.6 Duration-adjusted TMRCA predicts outbreaks where a long-term carrier is suspected

There was a significant difference in duration-adjusted TMRCA between outbreaks with and without evidence of an LTC, confirmed by ROC analysis. This is in keeping with the results from chapter 2, where greater diversity in an individual was seen with prolonged carriage. Similarly, greater diversity in an outbreak may indicate the involvement of an LTC, with outbreak diversity reflecting the donor cloud.

The longest TMRCAs were in hospital outbreaks with indirect links between cases (i.e. non-overlapping ward stays). In these outbreaks, the most likely reason for the occurrence of the outbreak strain in a new case was thought to be either re-introduction from the community (outbreak G) or a staff member with long-term carriage (outbreaks A, I, N and S). Staff carriage was proven in one outbreak (by sampling and subsequent termination of the outbreak on exclusion of the HCW), whilst in the remaining outbreaks HCWs were either not sampled, or HCW sampling was anonymised and positive results could not be linked definitively with the suspected carrier.

4.5 Conclusion

Using the information from the 15 resolved outbreaks, I have been able to establish SNV thresholds which indicate whether isolates are likely to have a close epidemiological link and therefore form part of an outbreak. Applying this information to complex, unresolved outbreaks, WGS data was able to clarify or confirm the results.

Using data from all the outbreaks, I have shown how genetic data and epidemiological data can be applied to interpret the TMRCA, to predict the possible involvement of a long-term carrier.

5 Real-time detection of MRSA outbreaks

5.1 Introduction

MRSA screening is the backbone of MRSA control policies in the UK. Screening aims to identify carriers to ensure appropriate precautions are taken for that individual (e.g. pre-operative decolonisation, use of effective antimicrobial prophylaxis, and empirical anti-MRSA treatment for infections) to prevent transition from asymptomatic carriage to clinical infection, and on the wider level to prevent transmission to other patients (e.g. isolation and barrier nursing).

Oxford University Hospitals (OUH) NHS trust implements a universal screening policy. Nasal swabs are taken from all adult and paediatric elective and emergency admissions, with further swabs taken if skin lesions are present – for example wounds, ulcers, enteral feeding tube sites or other device sites. Due to poor compliance with screening on the admissions units, patients are re-screened on transfer to other wards, but despite this, overall trust-wide compliance with screening is approximately 54% (79% of elective admissions and 49% of emergency admissions screened). This is broadly in keeping with the figures reported nationally (81), and is the rationale behind continuing with universal screening even though there has been a UK-wide move towards targeted screening as a more cost-effective measure (82).

The first positive sample from a newly identified MRSA colonised or infected patient has a full antimicrobial susceptibility profile recorded by the clinical microbiology laboratory and is stored on an agar slope. The results are sent to the infection control team, who provide advice regarding infection control precautions and maintain surveillance databases to detect and track clusters of cases which may require further investigation.

An increase in MRSA notifications from the same ward or unit triggers an infection control investigation. Antibiograms are examined and epidemiological data collected, and, if transmission is suspected, isolates are sent to the reference laboratory for typing.

Although some interim actions may be taken while awaiting typing results, full control measures are not implemented until the reference laboratory confirms the relatedness of samples, usually by PFGE.

Incorporating WGS into routine practice as part of laboratory identification and resistance profiling would also mean the data is available for investigation of outbreaks. This would alter the current trajectory of outbreak investigation from a reactive to proactive process, with epidemiological investigation initiated by the identification of related isolates.

In this chapter, I used rapid benchtop sequencing in close-to-real-time for the newly identified positive MRSA isolates in OUH trust, to compare with routine surveillance and assess whether WGS would allow earlier detection of outbreaks.

5.2 Methods

5.2.1 Sampling frame and sequencing

MRSA screening and clinical sample isolates from newly positive adult patients were collected weekly from the routine laboratory over a period of four months. Isolates were excluded if the sample originated from a GP surgery, or if the patient was a day case only with no previous hospital inpatient stays. Isolates were grown overnight on Columbia blood agar and DNA extracted as previously described. Isolates were sequenced on the MiSeq platform in batches of 12, which allowed for a minimum of 1 MiSeq run per week.

5.2.2 Assembly and analysis

MLST and predicted antimicrobial susceptibility results were obtained from raw reads using Mykrobe (183). This software utilises De Bruijn graphs to find similarities between raw reads and reference sequences without requiring assembly of the whole genome, resulting in a computational time of 1-2 minutes compared with several hours for *de novo*

assembly using e.g. Velvet. Using Mykrobe allowed bypassing of the standard analysis pipeline (and data queue) used in the previous chapters, to allow same day analysis of results.

Isolates identified as belonging to the same MLST sequence type were aligned to the MRSA252 reference sequence and SNVs identified (with masking of mobile elements) using Stampy as described previously. A 20 SNV threshold was chosen to identify possibly related isolates: this threshold was set higher than the SNV distances observed in the previous chapters to maximise sensitivity and capture possible historic transmissions. Epidemiological data for these isolates was examined using the hospital electronic patient record to determine if there were direct ward / indirect ward contacts. Isolates identified as closely related were reported to the relevant Infection Control team.

5.2.3 Time to most recent common ancestor

Time to most recent common ancestor (TMRCA) was estimates using BEAST v1.8.1 as described in chapter 4, with a simple HKY substitution model, constant population size and a substitution rate of 3.3×10^{-6} substitutions per genome per year. The time between first and last samples was subtracted from the calculated TMRCA to obtain a duration-adjusted TMRCA.

5.3 Results

5.3.1 Samples

89 MRSA isolates were collected over a 4-month period. Eight isolates had unidentified, sporadic MLST sequence types as identified by the Mykrobe database. The remaining 81 isolates belonged to 14 sequence types: the most common sequence type was ST22 (46 isolates, 52%), followed by ST5 (6 isolates, 7%), ST30 and ST927 (5 isolates each, 6%).

5.3.2 Turnaround times

Isolates were batched on a weekly basis to permit cost-effective use of the MiSeq. The maximum time between a positive sample and final result was 12 days for a sample taken on a Friday: this sample was not confirmed as MRSA on the routine bench till the following Monday and hence missed the sequencing run for that week, adding a further 7-day delay. All other isolates were sequenced within 7 days of sampling. The workflow is shown in figure 5.1.

5.3.3 Analysis of clusters

Seven clusters were identified. Four of these contained only 2 patients (in ST5, ST8, ST22 and ST30). All 8 patients were found to be positive on admission screening, and all had hospital inpatient stays in the previous 6 months, suggesting that acquisition had occurred during a previous admission. However, none of the pairs had direct epidemiological links as defined in chapter 3, and for 3 of the pairs the inpatient stays were in different hospitals within the trust. Details of the epidemiological links are given in table 5.1.

FIGURE 5.1 WORK FLOW FOR WGS OF MRSA SCREENING SWABS

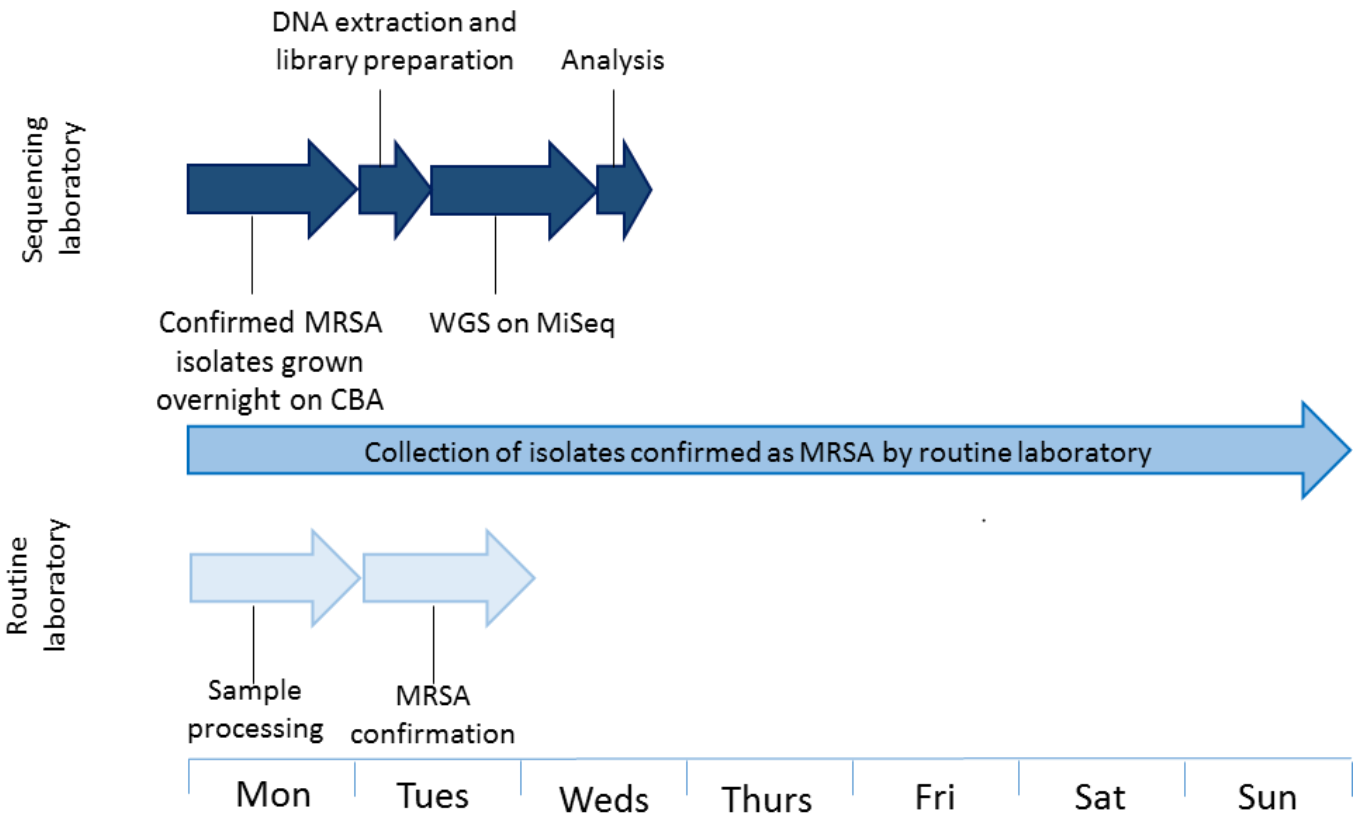


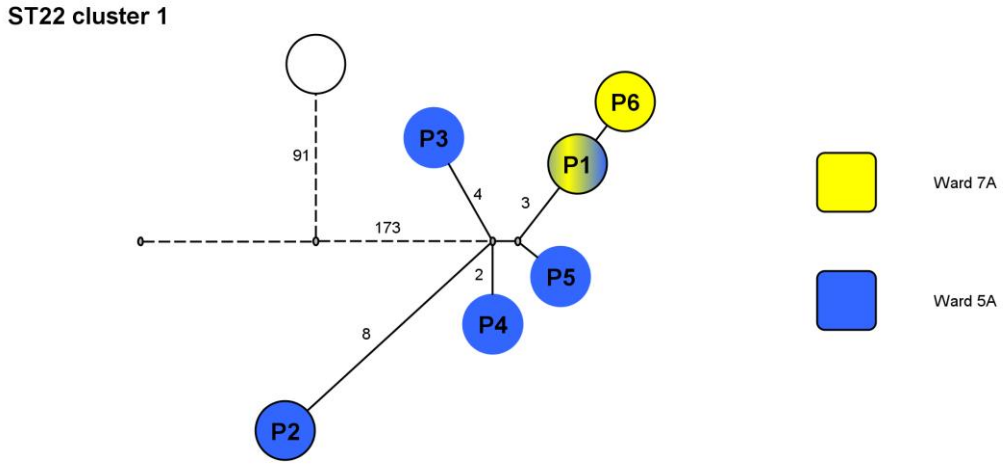
TABLE 5.1 SNV DISTANCES AND EPIDEMIOLOGICAL LINKS FOR CLOSELY RELATED SAMPLES DETECTED BY REAL-TIME SEQUENCING

Cluster	MLST	SNV distance	Time between samples	Epidemiological link
Real-time 1	ST8	6	3 days	Non-overlapping inpatient stay in Care of the Elderly wards in different hospitals.
Real-time 2	ST30	1	12 days	Non-overlapping inpatient stays in different specialties in different hospitals.
Real-time 3	ST22	3	5 days	Non-overlapping inpatient stays in different specialties in different hospitals.
Real-time 4	ST5	1	9 days	Non-overlapping inpatient stays in different surgical specialty wards, same hospital.

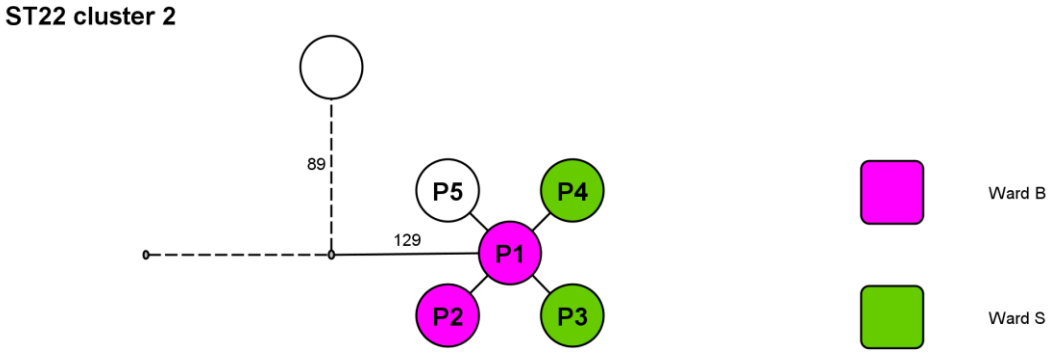
Three larger clusters of n>2 were identified. Two of these occurred in ST22 (labelled ST22 cluster 1 (6 patients), and ST22 cluster 2 (5 patients) respectively), while ST297 contained a single cluster of 4 patients. Phylogenetic trees for each of these are shown in figure 5.2.

ST22 cluster 1 comprised 6 patients. Five of the six samples were screening swabs, and 1 was a urine sample. Four patients were positive on routine admission screening to general medical wards, and one patient was found to be positive on pre-admission screening prior to a surgical procedure. The sixth patient was negative on admission but was positive on screening during their inpatient stay, when screening was performed prior to a surgical procedure.

FIGURE 5.2 PHYLOGENETIC TREES FOR CLUSTERS OF N>2 DETECTED BY WGS OF MRSA SCREENING SAMPLES

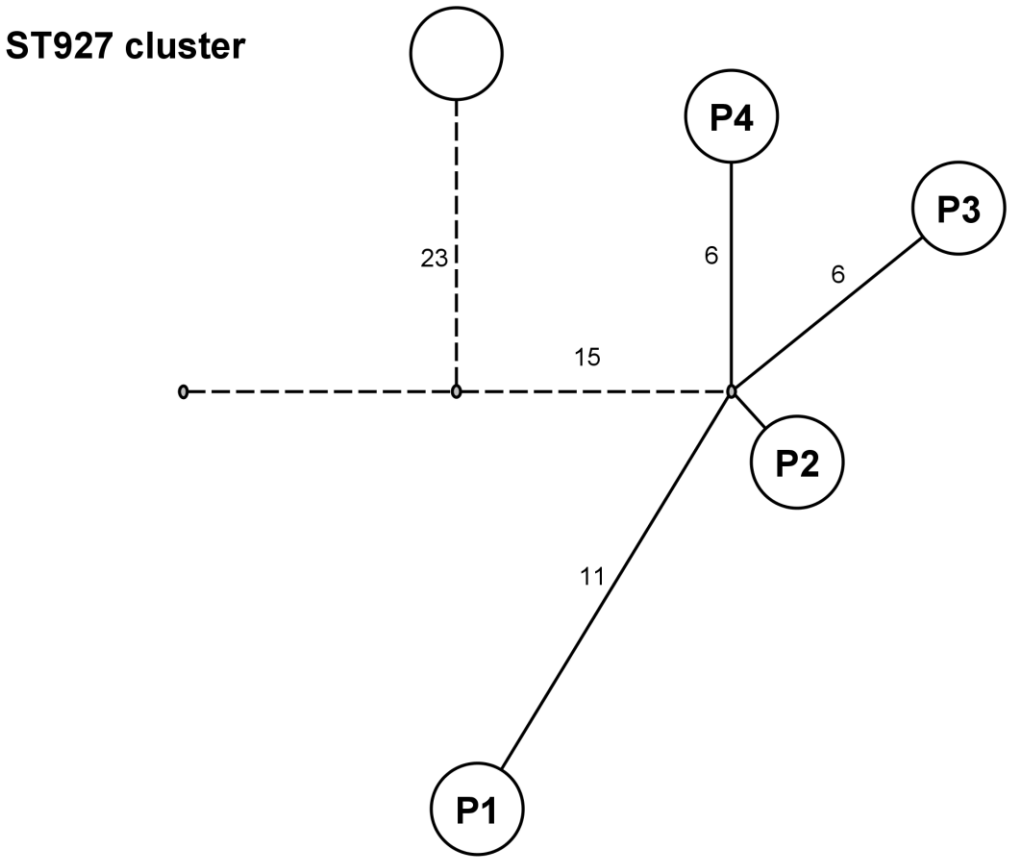


Distances are given in SNVs. The empty circle indicates the closest non-outbreak isolate of the same MLST. Black outline indicates patients with direct ward contacts: patient 2 with patient 1, and patient 1 with patient 6.



None of the patients in ST22 cluster 2 had direct ward contacts.

Figure 5.2 continued



None of the patients in the ST927 cluster had overlapping ward stays.

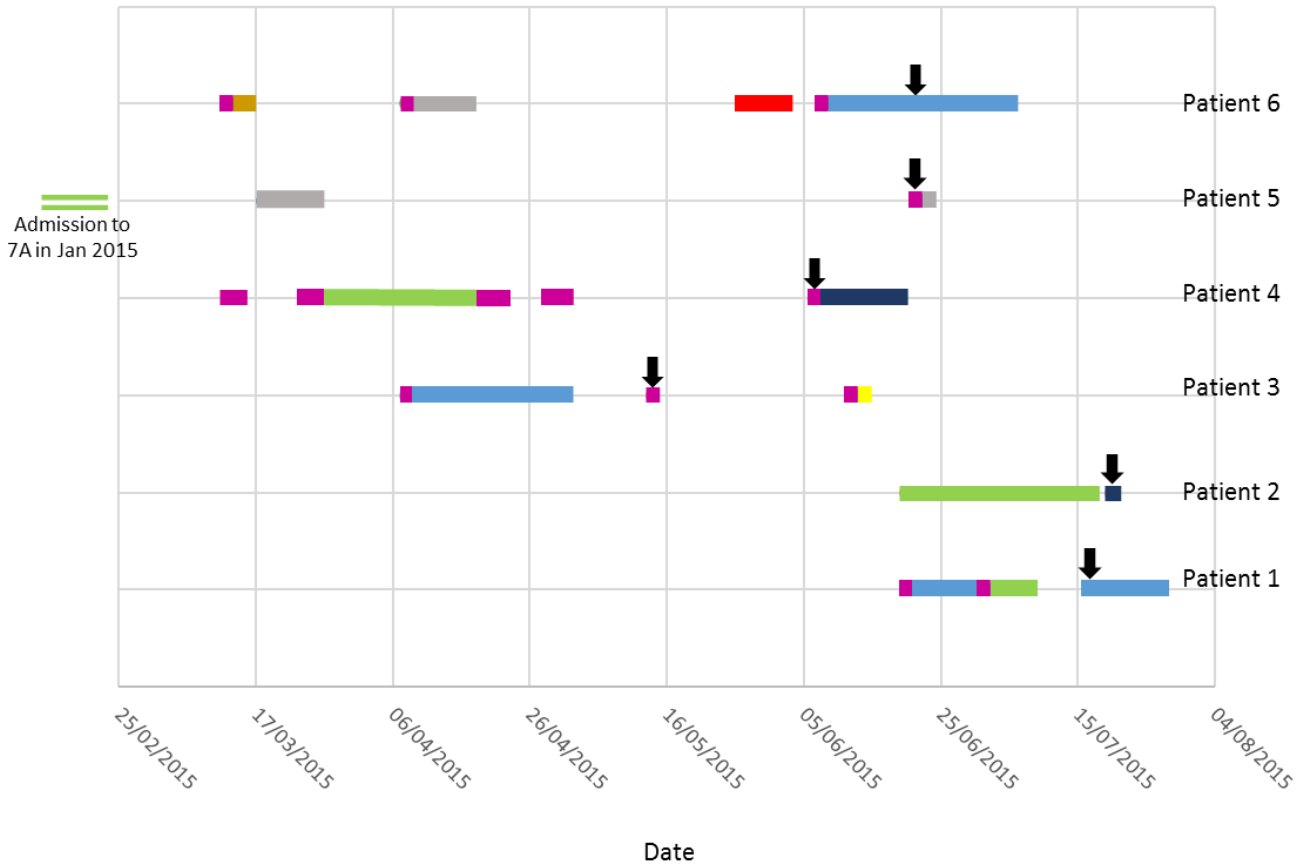
All six patients had been inpatients in acute general medicine at the John Radcliffe hospital in the 6 months prior to their positive MRSA swab. Three were inpatients on ward 5A (short stay ward), 2 were inpatients on ward 7A (acute medicine), and 1 had been an inpatient on both wards. A timeline is given in figure 5.3. Two patients were also inpatients on the care of the elderly care unit (Bedford ward) in the same time period, prior to their admission to general medicine.

Patients 1, 3, 5 and 6 all had stays on ward 5A, suggesting the ward as a potential source of acquisition. Patient 1 had an overlapping ward stay on 5A with patient 6 (direct ward contact) when 6 was known to be positive. Patient 1 was then transferred to ward 7A, where they had direct ward contact with patient 2, suggesting a potential transmission chain of patient 6 -> patient 1 -> patient 2. Patient 5 had a previous inpatient stay on ward 5A, suggesting acquisition during that inpatient episode. Patient 4 was positive on admission to ward 7C, suggesting prior acquisition during their previous stay on ward 7A or the emergency admission unit (EAU).

The WGS data and the epidemiological data both suggest a common exposure within the AGM service, with patient movement allowing transfer of the strain between wards.

Of interest, 4/6 of the cluster also had had outpatient appointments with ophthalmology in the preceding 6 months, raising the possibility of an alternative point source allowing introduction of the strain into the AGM cohort. However, 3 of the patients had negative admission screens in the month following their clinic appointment, making this a less likely source of acquisition.

FIGURE 5.3 TIMELINE OF ST22 CLUSTER 1 INPATIENT STAYS



Lines are coloured according to ward. Green: ward 7A, blue: ward 5A, purple: EAU, grey: elderly care ward, red: surgical ward, navy: ward 7C. Black arrows show date of positive MRSA screen.

The second ST22 cluster (ST22 cluster 2) comprised 5 patients. Patients 1 and 2 had prolonged direct ward contact (> 7 days in same ward) on the head and neck unit during February 2015, prior to the real-time screening exercise. One of these was positive at outpatient screening in May 2015, and the other patient was positive on their next admission screen, suggesting that both patients acquired during their inpatient stay in February.

Patients 3 and 4 had direct stays on ward 7C (acute general medicine) in June 2015 and were both subsequently positive on their next admission screening, suggesting acquisition on 7C. No contact was found between these two and patients 1 and 2.

Patient 5 was positive on admission to the Horton General Hospital in June 2015. This patient had no previous admissions recorded, although they had attended a single outpatient appointment (in urology) at the Churchill site in the previous 6 months. The source of their MRSA is, however, unclear.

The ST927 cluster comprised 4 patients. Patient 1 was positive on admission in May 2015, with the most recent admission in September 2014 (ward 5D, vascular surgery). Patient 2 became positive during a prolonged stay in a community hospital (known to be negative on admission to the unit). Patient 3 was positive on admission to EAU in June 2015, with their most recent inpatient stay being in cardiothoracic surgery. Patient 4 was positive on admission to the trauma ward in June 2015, with 2 previous admissions on SEU and SSW. The closest contact between any of the patients was at the level of same hospital, different times, suggesting a multi-factorial outbreak, again pre-dating the screening study.

5.3.4 Time to most recent common ancestor

Results from the BEAST analysis are given in table 5.2.

TABLE 5.2 DURATION -ADJUSTED TMRCA VALUES FOR CLUSTERS DETECTED BY REAL-TIME SEQUENCING

Cluster	Days between samples	TMRCA, days (95% HPD interval)	Duration adjusted TMRCA, days (95% HPD interval)
ST22 cluster 1	67	227 (129-337)	160 (62-270)
ST22 cluster 2	52	80 (52-124)	28 (0-72)
ST927	40	334 (201-460)	294 (161-420)

The duration-adjusted values for ST22 cluster 1 and the ST927 cluster are prolonged, and above the cut-off value of 156 determined in chapter 4 for predicting the involvement of a long-term carrier, while ST22 cluster 2 has a much shorter duration-adjusted TMRCA.

5.4 Discussion

5.4.1 Identification of clusters

Rapid sequencing and analysis identified 7 clusters which had not been detected by routine surveillance.

Of the larger clusters, ST22 cluster 1 provided evidence for an unsuspected ward-based outbreak of MRSA carriage, with all patients having been admitted to ward 5A or having been in direct contact with a patient recently admitted to that ward.

ST22 has previously been identified as the most common UK hospital MRSA clone and its prevalence in Oxfordshire well documented (184). Consequently, if an outbreak had been suspected on screening and epidemiological grounds, standard PFGE typing may

have lacked the resolution to show that these isolates were part of a cluster, as a conserved pulsotype may be observed in unrelated isolates of a prevalent clone (46).

The epidemiology of this outbreak (indirect ward contacts) is in keeping with the possibility of a long-term carrier – either a long stay patient or a staff member. This is supported by the duration-adjusted TMRCA result, which gave an estimated time to most recent common ancestor of 160 days, just above the threshold determined in chapter 4. However, as routine inpatient surveillance and discharge screening are not performed in the Trust, no long-term carrier was identified, and the lack of clinical cases meant a full infection control investigation was not instigated.

The ST927 cluster also has a duration-adjusted TMRCA in keeping with a long-term carrier. In this cluster, the patients were all on different wards at different times, and had no known outpatient encounters in common. If a long-term carrier was involved, it may have been due to having a procedure carried out, with the patients acquiring the strain while attending for an investigation or intervention in a department and returning with the strain to their respective wards (although excepting X-ray, none of the patients had similar procedures or departmental visits noted). Alternatively, transmission may have been via a peripatetic member of staff who sees and examines patients on different wards – for example a specialist nurse, physiotherapist or similar allied professional, or a doctor from a consult specialty. Identifying common exposures to a peripatetic member of staff is difficult without detailed analysis of all patient movements, visitors and exposures, and these staff are very rarely sampled or investigated, even during an outbreak.

ST22 cluster 2 had a duration adjusted TMRCA of 28 days, not in keeping with a long-term carrier. This may, however, be in keeping with the transfer of an un-sampled patient acquiring the strain on the head and neck ward and shortly afterwards being admitted to ward 7C where there was subsequent onward transmission. However, the source of the fifth patient's MRSA remains cryptic.

There were minimal epidemiological links for the four pairs of patients. None of them had direct ward contact, and in three pairs, the patients were in different hospitals and no obvious common epidemiological link could be found. This is most likely due to un-sampled transmissions links. Both patients and staff members, particularly doctors, may be transferred between wards and other hospitals within the trust, taking their strain with them (185), which may explain these results. Alternatively, there may be a community source such as a day care centre or community nursing exposure, although their attendance at different hospitals suggests that their community contacts are likely to be different as well.

5.4.2 Comparison with routine infection control

None of the clusters had been identified by infection control. The majority of isolates were from screening swabs taken on admission. This therefore triggered isolation and decolonisation of the individuals, but not an investigation into source as this is only performed for *S. aureus* bacteraemias or if there has been a cluster of infections noted. There was therefore insufficient epidemiological data collected to suspect increased transmission on the wards in question or to support enhanced surveillance (for example, weekly inpatient screening). In addition, in most of the clusters the majority of patients were identified on outpatient pre-admission screening or admission screening on entry to the emergency assessment unit. The actual acquisition may have pre-dated the screening sample by several weeks and, in some cases, months. Consequently, although the sequencing results were available in a timely manner, they reflected events of the preceding months and therefore did not prompt any infection control intervention.

5.4.3 Effect of screening strategy

The main drawback of this study was the sampling strategy. The aim was to assess how real-time sequencing of isolates from newly identified MRSA carriers could enhance infection control interventions, utilising isolates sent as part of routine practice. The assumption was that if WGS can replace routine laboratory methods for identification and resistance testing of isolates (discussed in chapter 6), the additional information (MLST and subsequent genetic relatedness) may be of value to infection control teams as clusters would be detected sooner, allowing interventions to prevent further spread.

However, there were no outbreaks of clinical disease identified during the study period, and most isolates were from admission screening swabs. Where clusters were detected, the majority of putative acquisition / transmission events occurred some time prior to the identification of MRSA carriage: in most cases a direct or indirect ward contact pre-dated isolation of MRSA from screening when the individual was next admitted to hospital. Because there is no ongoing surveillance of patients once admitted, acquisition of MRSA carriage during an inpatient stay is rarely identified unless from a non-screening sample. Since nasal colonisation with MRSA is a risk factor for infection (186-189), there is an argument for active inpatient screening, however previous studies using an intensive screening strategy in combination with WGS for MRSA and MSSA could only identify a transmission event in 7 out of 37 acquisitions (108). The cost-benefit of such a strategy would therefore need to be assessed.

In England, there are approximately 500 suspected *S. aureus* outbreaks referred to the Staphylococcal reference laboratory per year, across 154 acute NHS trusts (27). A single trust might therefore see only 1-3 suspected outbreaks per year, and, since not all of these are MRSA, and many are disproved on standard investigation, the absence of a disease outbreak during the study period is perhaps not surprising. A longer study may therefore capture a disease outbreak in real-time, and this work has been taken forward by a colleague within the MMM research group. Having a database of known carriers and their

sequences may enhance the investigation should a disease outbreak occur, as it could be rapidly screened for similar isolates and allow previous contacts to be investigated.

5.5 Conclusion

From a clinical or screening sample, processing time from positive growth to confirmation of MRSA can take up to 48hrs (190). Using the MiSeq sequencing platform in combination with rapid analysis tools means that this time can be matched, with the addition of MLST data allowing rapid identification of potential clusters. Further information, such as resistance prediction, can also be obtained as will be discussed in chapter 6.

Although in this study, no disease outbreaks were detected and the majority of putative transmission events pre-dated the screening sample by weeks to months, WGS was still able to identify plausible transmission links. Therefore, if it becomes tenable to use WGS routinely for identification and resistance testing of organisms, sequence data will be available for comparison with other isolates in the same time frame as it takes to confirm that an individual has MRSA, at no additional cost and without the need to send isolates to the reference laboratory for further testing, with the associated costs and delays.

6 Resistance prediction using WGS

6.1 Introduction

Current phenotypic methods for testing bacterial susceptibility to antimicrobials are based on the ability of bacteria to grow in the presence of specific concentrations of antimicrobial agent. The techniques employed (agar diffusion, agar dilution, manual or automated broth dilution) have changed little since they were originally explored by Fleming in the 1920s (191, 192).

In the UK, the disc diffusion method employed in most laboratories which do not have automated phenotyping is, in essence, the same as it was in 1947 when paper discs infused with antibiotic were first used (193), although it has been modified and standardised to allow quality control and uniformity of interpretation (190).

Automated methods based on broth dilution have been available since the 1970s (194). These are considered highly cost-effective as they do not rely on reading and interpretation by trained laboratory staff, and the results can be communicated directly to an electronic reporting system, reducing the likelihood of human / transcription error.

While there are problems for individual organism / antimicrobials reported for all phenotypic methods (195, 196), particularly for isolates with borderline or hetero-resistance, the major drawbacks are the need for an initial pure culture and the growth time to allow optimal interpretation, meaning results may take up to 48hrs to obtain. Fastidious organisms are also problematic, as they require special media and growth conditions, take longer to grow, and validated antimicrobial breakpoints may be lacking.

Rapid identification of resistance genes, from pure culture or directly from a sample, is possible using PCR or DNA hybridisation: this can reduce the time taken for inference of phenotypic resistance to 2-4hrs, but both these methods are capable of testing only for a small number of antimicrobials (197), and in routine laboratories in the UK are predominantly used purely for methicillin resistance testing.

Bacteria become resistant to antimicrobials in a number of ways: by altering the antimicrobial target (either by mutation or acquiring an insensitive target), by producing an enzyme which inactivates the antimicrobial, or by altering the ability of the antimicrobial to reach its target (e.g. altering the cell wall or up-regulating efflux pumps) (198). At the genetic level, the mechanisms arise because of a mutation or insertion/deletion in the core genome (frequently a point mutation causing an altered target protein), by acquisition of extraneous DNA from a phage or plasmid which codes for an enzyme or alternative target, or by altered expression of constitutive genes. Therefore, having the entire genome of an organism, should, at least in theory, make it possible to recover the entire complement of genes or mutations encoding resistance, analogous to *in silico* multiplex PCR.

To be used confidently in clinical practice, reliable genotypic prediction of antimicrobial resistance phenotype has to be demonstrated to the same standards as any new phenotypic method, using large diverse sets of unrelated, wild-type clinical isolates. Comprehensively validated genotypic prediction of antimicrobial resistance, ready for implementation in clinical practice, will require multiple large studies.

In this chapter, I used WGS data and locally available bioinformatic tools to develop a resistance gene prediction method for *Staphylococcus aureus* as a feasibility study. Development and refinement of the method was done on an initial "derivation set" of 501 isolates, with testing and validation performed on a further, unrelated set of 491 isolates.

6.2 Methods

6.2.1 Creation of a catalogue of antimicrobial resistance genes

A panel of antimicrobial agents (tables 6.1 and 6.2) was identified for investigation based on those used routinely for management of *S. aureus* infections in the Oxford University Hospitals (OUH) NHS Trust. To identify the genetic determinants encoding resistance to these antimicrobials, a literature search was conducted in PubMed using the MESH terms *"Staphylococcus aureus"*, "Drug resistance, microbial" and individual antimicrobial drug

names to create a catalogue of antimicrobial resistance genes and variants, using published sequences deposited in GenBank (http://www.ncbi.nlm.nih.gov/genbank/), see tables 6.1 and 6.2.

TABLE 6.1 PANEL OF GENES ASSOCIATED WITH MOBILE GENETIC ELEMENTS USED FOR BLAST QUERY

Antimicrobial	Gene	Product	Reference gene accession number
Penicillin	*bla*Z	Class A beta-lactamase	BX571856.1 (1913827..1914672)
Methicillin	*mec*A	Low affinity penicillin binding protein (PBP2)	BX571856.1 (44919..46925)
Erythromycin	*msr*A*	Erythromycin resistance protein	CP003194 (54168..55634)
Erythromycin / clindamycin	*erm*A	rRNA adenine N-6-methyltransferase	NC_002745.2 (1685864..1686595)
	*erm*B	rRNA adenine N-6-methyltransferase	AB699882.1 (4971..5708)
	*erm*C	rRNA adenine N-6-methyltransferase	HE579068 (7858..8592)
	*erm*T	23s rRNA methylase	HF583292 (11344..12078)
Tetracycline	*tet*K	MFS tetracycline effux pump	FN433596 (69118..70497)
	*tet*L	MFS tetracycline efflux pump	HF583292 (7714..9089)
	*tet*M	Ribosomal protection protein	CP002643 (427033..428952)
Vancomycin	*van*A	Low affinity pepdidoglycan precursor	AE017171.1
Fusidic acid	*fus*B	Fusidic acid detoxification	CP003193.1 (1336..1977)
	*fus*C*	Ribosome protection protein	BX571857.1 (52805..53458)
Trimethoprim	*dfr*A	Insensitive dihydrofolate reductase	CP002120 (2093303..2093788)
	*dfr*G	Insensitive dihydrofolate reductase	FN433596 (502263..502760)
Gentamicin	*aac*A-*aph*D	6'-aminoglycoside N-acetyltransferase/2''-aminoglycoside phosphotransferase	FN433596.1 (2209531..2210970)
Mupicrocin (high level resistance)	*mup*A	Isoleucyl-tRNA synthetase	HE579068 (2157..5231)
	*mup*B	Isoleucyl-tRNA synthetase	JQ231224 (91..3192)

*added in v2.0
MFS: Major facilitator superfamily
Presence of gene correlates with phenotypic resistance.

With the exception of *van*A, reference sequences were obtained from published whole genome sequences of human clinical isolates.

TABLE 6.2 PANEL OF HOUSEKEEPING GENES WITH AMINO ACID VARIANTS KNOWN TO BE ASSOCIATED WITH ANTIMICROBIAL RESISTANCE

Antimicrobial	Gene	Amino acid substitutions	Reference gene accession number
Ciprofloxacin	*gyr*A	S84L, E88K, G106D, S85P, E88G, E88L	BX571857.1 (7005..9668)
	*grl*A	S80F, S80Y, E84K, E84G, E84V, D432G, Y83N, A116E, I45M*, A48T, D79V, V41G, S108N	BX571857.1 (1386869..1389271)
	*grl*B	R470D*, E422D*, P451S*, P585S*, D443E*, R444S*	BX571857.1 (1384872..1386869)
Fusidic acid	*fus*A	A160V*, A376V, A655E, A655P*, A655V*, A67T*, A70V*, A71V*, D434N, C473S*, D189G*, D189V*, D373N*, D463G*, E233Q*, E444K, E444V*, E449K*, F441Y, F652S*, G451V, G452C, G452S, G556S, G617D, G664S, H438N, H457Q, H457Y, L430S*, L456F, L461K, L461S, M161I*, M453I, M651I, P114H, P404L, P404Q, P406L, P478S, Q115L, R464C, R464H, R464S, R659C, R659H, R659L, R659S, R76C*, S416F*, T385N, T387I*, T436I, T656K, V607I, V90A, V90I, Y654N*	BX571857.1 (577685..579766)
Rifampicin	*rpo*B	A473T*, A477D, A477T*, A477V, D471G*, D471Y, D550G, H481D, H481N, H481Y, I527F, I527L*, I527M*, ins 475H, insG475*, L466S*, M470T*, N474K*, Q456K, Q468K, Q468L, Q468R, Q565R*, R484H, S463P, S464P, S486L, S529L*	BX571857.1 (568813..572364)
Trimethoprim	*dfr*B	F99Y, F99S, F99I, H31N, L41F, H150R, L21V*, N60I*	BX571857.1 (1464014..1464493)

*reported in association with other variants. For expected effect on MIC of each variant / combination, see appendix 3.

All housekeeping gene sequences were obtained from reference genome MSSA476(199).

6.2.2 Sampling frame and phenotypic testing: derivation set

Initial development was done using 501 clinical *S. aureus* isolates which had been sequenced and phenotyped previously and whose WGS and susceptibility data were available ("derivation set"). Isolates were identified from bacteraemia and carriage collections held at the Oxford University Hospitals NHS trust and Brighton and Sussex University Hospitals NHS Trust, spanning a period of 13 years (table 6.3) (200). The collection included 159 MRSA isolates (32%). All isolates had been tested at each site by the routine clinical laboratories for resistance to a standard first line panel of antimicrobial agents (both sites: penicillin, methicillin, erythromycin, vancomycin, ciprofloxacin, tetracycline, gentamicin, fusidic acid, rifampicin; Brighton only: mupirocin, clindamycin; Oxford only: trimethoprim). In Brighton, susceptibility testing was performed using the Vitek automated system (bioMérieux, Basingstoke, UK), and in Oxford, isolates were phenotyped by disc diffusion (201). The susceptibility testing results were retrieved electronically from laboratory databases. Susceptibility to methicillin was determined using cefoxitin (Brighton) or oxacillin (Oxford).

TABLE 6.3 SOURCE OF ISOLATES AND METHOD OF SUSCEPTIBILITY TESTING

No of isolates	Source	Specimen type	Dates	Susceptibility testing method	Sequencing
Derivation set collections, n=501					
88	Brighton (all wards)	Blood	1999-2007	Automated broth dilution (Vitek), clinical laboratory	Previously sequenced
90	Brighton carriage (ITU)	nasal swab	2010-2011	Disc diffusion, clinical laboratory	Previously sequenced
323	Oxford (all wards)	Blood	2008-2011	Disc diffusion, clinical laboratory	Previously sequenced
Validation set collections, n=491					
102	Oxford carriage (ITU)	nasal swab	2009	Disc diffusion and automated broth dilution (BD Phoenix)	Previously sequenced
100	Oxford carriage (community)	nasal swab	2009	Disc diffusion and automated broth dilution (BD Phoenix)	Previously sequenced
165	Brighton (all wards)	Blood	2011-2012	Disc diffusion and automated broth dilution (BD Phoenix)	Sequenced *de novo*, susceptibility testing performed from same subculture as sequencing
124	Oxford (all wards)	Blood	2011-2012	Disc diffusion and automated broth dilution (BD Phoenix)	Sequenced *de novo*, susceptibility testing performed from same subculture as sequencing

6.2.3 Sampling frame and phenotypic testing: validation set

A further isolate collection comprising 491 isolates ("validation set") was assembled to provide a test set following optimisation of the genotypic method on the derivation set (table 6.3). 202 isolates were sourced from carriage collections (130) which had previously been sequenced but not phenotyped. These were retrieved from -80° C storage for susceptibility testing. A further 289 isolates were obtained from archived bloodstream collections at the Oxford and Brighton sites. For these, single colonies were plated on Columbia blood agar and grown at 37˚C for 18-24 hrs. All the bacteria on the plate were harvested and suspended in 1.5 ml physiological saline. 0.1ml was removed and re-plated on to Columbia blood agar and incubated overnight at 37˚C for resistance phenotyping, and the remaining suspension used to prepare DNA for WGS.

To control for differences in phenotypic testing methods, all 491 isolates had antimicrobial susceptibility testing performed using the Phoenix Automated Microbiology System (BD Biosciences, Sparks, MD, USA) for a standard panel of antimicrobial agents (penicillin, methicillin, ciprofloxacin, erythromycin, fusidic acid, gentamicin, mupirocin, rifampicin, tetracycline, vancomycin). Isolates were also tested by disc diffusion (202), and concordant results for the two methods were used as the final phenotype. Isolates which were resistant to erythromycin were further tested for clindamycin resistance by disc diffusion and the clindamycin D test (203). Where there was discordance between disc diffusion and BD Phoenix, repeat testing was performed by the gradient diffusion method (E-test) using EUCAST breakpoints, and this value used as the phenotype.

6.2.4 Sequencing and assembly

Isolates were sequenced using the Illumina HiSeq platform, and *de novo* assembled using Velvet as described in previous chapters.

6.2.5 Genotypic interrogation using BLAST

A locally developed programme combining BLAST with interpretative thresholds ("Typewriter") was utilised. The *de novo* assembled genomes were interrogated with BLAST+ (v 2.2.28+) blastn and tblastn (118) to identify nucleotide sequences matching genes from the panel and their matching protein sequences using default parameters. Relative coverage was defined as the product of the proportion of reference allele matched and the sequence identity of the match. For the initial algorithm (v1.0), a relative coverage threshold of >80% was chosen to define gene presence with a high degree of similarity to the reference, based on pilot data on the initial few isolates (for example, 95% relative coverage may be 95% of the gene length with 100% identity, or 100% of the gene length with 95% identity). For housekeeping genes, where resistance is conferred by one or more point mutations, differences between the tblastn result and the query protein sequence were compared to the wild type protein, and compared against the catalogue of known antimicrobial resistance encoding mutations compiled above. Changes in protein sequence (at the same or different codons) which were not previously reported as conferring resistance were counted as susceptible.

To determine the diversity of the isolates tested, *in silico* prediction of multi-locus sequence type (MLST) was also performed using BLAST+. The *S. aureus* MLST alleles were extracted from assemblies based on sequence similarity to allele 1 for each locus, and the online MLST database [http://saureus.mlst.net/] was used to predict ST.

6.2.6 Comparison of phenotype and predicted genotypic resistance

The predicted susceptibilities based on mobile and chromosomal genetic elements, using the whole genome sequences of the isolates in the derivation set, were compared with the routine laboratory susceptibility testing results. Where there was a mismatch between genotypic prediction and recorded phenotype, isolates were retrieved from storage at -80° C and had repeat susceptibility tested by gradient diffusion using EUCAST breakpoints [http://www.eucast.org/clinical_breakpoints/] to resolve the phenotype. A very major error

(VME) was defined as a susceptible genotype with a resistant phenotype, and a major error (ME) was defined as a resistant genotype with a susceptible phenotype (204).

6.2.7 Optimisation using the derivation set

Using the results obtained for the derivation set, genotype-phenotype mismatches were explored to identify reasons for the mismatch and whether the accuracy of the genotype prediction could be improved. Additional hand searching of the references from the original literature search identified 2 additional genes (*msr*A and *fus*C) plus several additional variants of *fus*A, which were added to the catalogue.

A relatively high VME rate was noted for penicillin and fusidic acid. This was reduced by accepting hits for *blaZ*, *fus*B and *fus*C occurring on short or low coverage contigs, which were initially rejected by the default quality filters (see Results for details).

Sensitivity and specificity were estimated across varying thresholds for the relative coverage required for these genes to be considered present in the derivation set, and a revised algorithm (v2.0) derived which provided the best compromise between overall sensitivity and specificity. This was obtained by defining resistance as >30% relative coverage for *blaZ, fus*B and *fus*C, and as >80% relative coverage for the remaining mobile genes.

6.2.8 Blinded validation of revised prediction method v 2.0

Because repeat susceptibility testing and revision of the genotype prediction algorithm were done with *a priori* knowledge of the phenotype, v2.0 of the algorithm was applied to the validation set with no previous information regarding expected phenotype available before genotypic interrogation (i.e. blinded to phenotype).

Whole genome sequences were interrogated by BLAST+ using the revised parameters as described above, and the phenotypic and genotypic profiles were compared. Sensitivity, specificity and error rates were calculated against the concordant phenotype.

Where there was a mismatch between genotype and phenotype, isolates had 1) phenotype confirmed by E-test, 2) genotype confirmed by manual inspection of the presumed resistance encoding genes or mutations and 3) identity confirmed by comparison of known *spa*-type and *in silico* MLST, and resequencing if no explanation for the mismatch could be found. For penicillin mismatches, isolates were also tested for penicillinase production using nitrocefin discs (205).

6.3 Results

6.3.1 Derivation set

WGS and routine laboratory antimicrobial susceptibility testing results were available in the first instance for 506 isolates, with 46 different sequence types found by *in silico* MLST. Using the initial v1.0 algorithm, 439 (87%) had complete concordance between the genotype and phenotype for all antimicrobial agents tested. The remaining 67 isolates had a total of 123 discrepancies. For 5 of these, the frozen bacterial stocks were found to be contaminated with other organisms (1 with *Proteus sp*, 4 with coagulase negative staphylococci), and further confirmatory testing could not be performed. These were excluded from further analysis, leaving 501 isolates in the derivation set. Repeating the susceptibility testing for the isolates with phenotype/genotype mismatches reduced the error rate, with a total of 71 discrepancies in 49 isolates remaining, and a categorical agreement of 98.6% (5116 / 5187) of individual antimicrobial “tests” (trimethoprim, clindamycin and mupirocin testing was only done for a subset of isolates).

Compared to other antimicrobials with similar numbers of resistant phenotypes, a relatively high VME rate was noted for penicillin, with 9 (1.8%) isolates apparently phenotypically resistant to penicillin (MIC >0.12mg/L) in the absence of the *blaZ* gene, based on >80% relative coverage in WGS data. However, all 9 isolates were positive for *blaZ* by PCR. On manual inspection of the sequences, it was found that the *blaZ* gene was in fact present,

either on small sequence contigs (<300bp) or contigs with low coverage (read depth <5) which did not meet the algorithm quality standard and had therefore been excluded. Including these short or low coverage contigs in the algorithm (v2.0) improved predictions (with relative coverage as low as 30% associated with phenotypic resistance). This phenomenon was also noted for fusidic acid, where the inclusion of low quality contigs for the *fus*B and *fus*C genes improved prediction of phenotype. This was not the case for the remaining genes associated with mobile elements in the pane,l where lowering the relative coverage threshold substantially increased the false-positive rates.

Using this revised genotypic prediction method (v 2.0), there were 31 persistent discrepancies in 28 isolates across all antimicrobials in the derivation set (table 6.4). There were 22 very major errors (2.0% VME rate) and 9 major errors (0.2% ME rate). Overall sensitivity and specificity were 0.98 (95% CI 0.97-0.99) and 1.00 (95% CI 0.99-1.00) respectively.

TABLE 6.4 DERIVATION SET RESULTS

Comparison of results for individual antimicrobials for 501 carriage/bacteraemia isolates by phenotype (Vitek or disc diffusion) and predicted susceptibility using v2.0 genotypic prediction method

	Phenotype: resistant		Phenotype: susceptible		Total	Error rate		Sensitivity	Specificity
	Genotype		Genotype			Very major	Major		
Antimicrobial	Susceptible	Resistant	Susceptible	Resistant		%	%	(95% CI)	(95% CI)
Penicillin	4	438	59	0	501	0.9	0	0.99 (0.98-1.00)	1.00 (0.92-1.00)
Methicillin	1	158	341	1	501	0.6	0.3	0.99 (0.96-1.00)	1.00 (0.98-1.00)
Ciprofloxacin	7	165	326	3	501	4.1	0.9	0.96 (0.91-0.98)	0.99 (0.97-1.00)
Erythromycin	1	133	366	1	501	0.7	0.3	0.99 (0.95-1.00)	1.00 (0.98-1.00)
Clindamycin	0	88	89	0	177[b]	0	0	1.00 (0.95-1.00)	1.00 (0.95-1.00)
Tetracycline	0	28	473	0	501	0	0	1.00 (0.85-1.00)	1.00 (0.99-1.00)
Vancomycin	0	0	501	0	501	0	0	N/A	1.00 (0.99-1.00)
Fusidic acid	3[a]	38	458	2	501	7.3	0.4	0.93 (0.79-0.98)	1.00 (0.98-1.00)
Trimethoprim	5	10	308	0	323	33.3	0	0.67 (0.39-0.87)	1.00 (0.98-1.00)
Gentamicin	0	7	494	0	501	0	0	1.00 (0.60-1.00)	1.00 (0.99-1.00)
Mupirocin	0	2	174	2	178	0	1.1	1.00 (0.20-1.00)	0.99 (0.96-1.00)
Rifampicin	1	2	498	0	501	33.3	0	0.67 (0.13-0.98)	1.00 (0.99-1.00)
Overall	**22**	**1069**	**4087**	**9**	**5187**	**2.0**	**0.2**	**0.98 (0.97-0.99)**	**1.00 (0.99-1.00)**

[a]2 isolates had 2 non-synonymous mutations in *fus*A not previously described in the literature (T326I + E468V and T326I + V90I) which may be responsible for the observed phenotypes. [b]One isolate failed to grow for clindamycin testing.

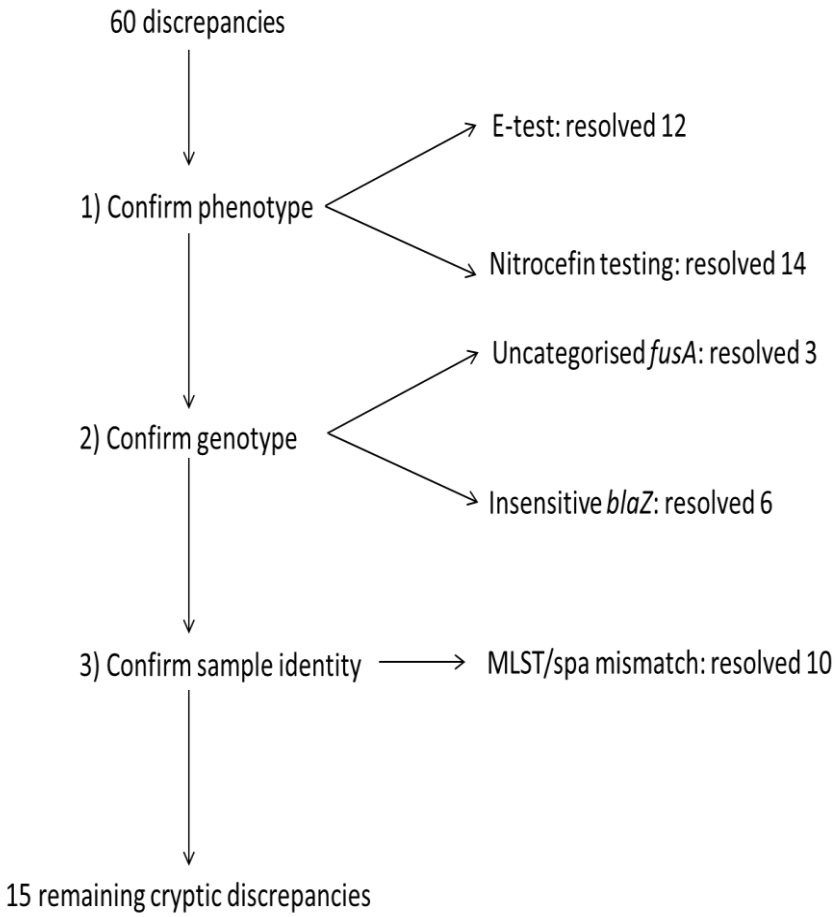
6.3.2 Blinded validation set, genotypic prediction method v2.0

For the 491 isolates in the validation set, 61 different sequence types were identified by *in silico* MLST. The most frequent sequence types were ST22 (13%) and ST30 (12%). 57 were MRSA (12%).

On first pass, there was phenotypic-genotypic agreement for 5133 / 5193 individual susceptibility tests (categorical agreement 99.7%) using v2.0 of the algorithm, giving initial overall VME and ME rates of 3.4% and 0.8% respectively.

There were 60 initial phenotype-genotype discrepancies in 49 isolates. 28 of the discrepancies (47%) were for penicillin, with 3 resolved by E-test and 14 by nitrocefin testing, indicating an initial phenotyping or transcription error. Of the non-penicillin discrepancies, 8 were resolved by E-test (i.e. initial phenotyping error), 3 were for fusidic acid and related to uncategorised *fus*A variants (see below), and 9 errors occurred in 4 isolates with *spa* / MLST mismatches, indicating that the sample sequenced was not the same as the sample stored and phenotyped. These isolates were discarded from further analysis (figure 6.1). In 2 cases of *spa* / MLST mismatches, the phenotype had matched the genotype on repeat testing and all 4 isolates had >1 phenotype-genotype discrepancy, giving further evidence for sample / labelling error. The genotypes of the isolates with unresolved mismatches were confirmed by re-sequencing which did not identify any further storage / labelling errors.

FIGURE 6.1 RESOLUTION OF PHENOTYPE / GENOTYPE DISCREPANCIES



After investigation into the discrepancies, there were 7 remaining VMEs: 2 for penicillin, 1 for ciprofloxacin, 1 for erythromycin, 1 for clindamycin, 1 for fusidic acid, and 1 for gentamicin. There were a further 8 MEs: 3 for penicillin, 2 for erythromycin, 2 for tetracycline, and 1 for trimethoprim. The final very major error rate was 0.9% and the major error rate was 0.2% (table 6.4). Overall sensitivity was 0.99 (95% CI 0.98-1.0) and specificity 1.0 (95% CI 0.99-1.0).

Further details for each of the discrepant isolates are given in tables 6.6 (penicillin) and 6.7 (other antimicrobial agents), and discussed for each antimicrobial below.

TABLE 6.5 VALIDATION SET RESULTS

Comparison of results for individual antimicrobials for 487 isolates by phenotype (combined BD Phoenix and disc diffusion) and predicted susceptibility using v2.0 genotypic predication method

	Phenotype: resistant		Phenotype: susceptible			Error Rates			
	Genotype		Genotype		Total isolates tested	VME	ME	Sensitivity (95% CI)	Specificity (95% CI)
Antimicrobial	Susceptible	Resistant	Susceptible	Resistant		(%)	(%)		
Penicillin	2	398	84	3	487	0.5	3.4	0.99 (0.98-1)	0.97 (0.90-0.99)
Methicillin	0	55	432	0	487	0	0	1 (0.92-1)	1 (0.99-1)
Ciprofloxacin	1	64	422	0	487	1.5	0	0.98 (0.91-1)	1 (0.99-1)
Erythromycin	1	80	404	2	487	1.2	0.5	0.99 (0.92-1)	0.99 (0.98-1)
Clindamycin	1	76	2	0	79	1.3	0	0.99 (0.92-1)	1 (0.20-1)
Tetracycline	0	18	467	2	487	0	0.4	1 (0.78-1)	0.99 (0.98-1)
Vancomycin	0	0	487	0	487	n/a	0	N/A	1 (0.99-1)
Fusidic acid	1	42	444	0	487	2.3	0	0.98 (0.86-1)	1 (0.99-1)
Trimethoprim	0	2	200	1	203	0	0.5	1 (0.20-1)	0.99 (0.97-1)
Gentamicin	1	2	484	0	487	33.3	0	0.67 (0.13-0.98)	1 (0.99-1)
Mupirocin	0	2	485	0	487	0	0	1 (0.20-1)	1 (0.99-1)
Rifampicin	0	5	482	0	487	0	0	1 (0.46-1)	1 (0.99-1)
Total	7	744	4393	8	5152	0.9	0.2	0.99 (0.98-1)	1 (0.99-1)

TABLE 6.6 RESULTS FOR MICS, NITROCEFIN DISC TESTING AND *BLAZ* VARIANTS FOR VALIDATION SET: INITIAL PENICILLIN DISCREPANT ISOLATES, V2.0 GENOTYPE PREDICTION METHOD

Isolate name	*In silico* MLST	Initial consensus phenotype	Repeat MIC (mg/L)*	Nitrocefin		*blaZ* variant	Discrepancy resolved?
				t =10m	t=120m		
C01124	ST45	S	0.12	-	-	InsA436	Yes
C01241	ST30	S	0.12	-	-	InsA256	Yes
C01203	ST36	S	0.06	-	-	A99-	Yes
C13228	ST30	S	0.06	-	-	A99-	Yes
C13375	ST7	S	0.12	-	-	A99-	Yes
C01144	ST30	S	0.12	-	-	A99-	Yes
C01104	ST8	S	0.023	-	-	no indels	No
C01205	ST30	S	0.06	+	+	no indels	Yes
C13249	ST45	S	0.06	-	+	no indels	Yes
C01092	ST2417	S	0.12	+	+	no indels	Yes
C01111	ST5	S	0.12	-	-	no indels	No
C01148	ST1	S	0.12	+	+	no indels	Yes
C01158	ST582	S	0.12	+	+	no indels	Yes
C01142	ST582	S	0.12	+	+	no indels	Yes
C01192	ST2417	S	0.12	+	+	no indels	Yes
C01199	ST30	S	0.12	+	+	no indels	Yes
C01080	ST22	S	0.12	+	+	no indels	Yes
C01217	ST30	S	0.12	+	+	no indels	Yes
C01231	ST22	S	0.12	+	+	no indels	Yes
C01266	ST188	S	0.12	+	+	no indels	Yes
C01277	ST2445	S	0.12	-	+	no indels	Yes
C01182	ST5	S	0.12	-	-	no indels	No
C01093	ST15	S	0.12	-	+	no indels	Yes
C01112	ST2438	S	0.25(R)	+	+	no indels	Yes
C12754	ST20	S	0.25(R)	+	+	no indels	Yes
C13232	ST7	R	0.06	-	-	absent	Yes
C12780	ST97	R	0.25 (R	-	-	absent	No
C01147	ST22	R	0.5 (R)	-	-	absent	No

* MIC testing used BSAC breakpoints, with ≤ 0.12mg/L considered susceptible.

TABLE 6.7 DETAILS OF INITIAL NON-PENICILLIN ANTIMICROBIAL DISCREPANCIES FOR VALIDATION SET

Antimicrobial	Isolate name	Initial consensus phenotype	Repeat MIC (mg/L)	Discrepancy resolved?
Very major errors				
Methicillin	C01115*	R	0.5 (S)	Yes – susceptible on repeat testing. *Spa* / MLST mismatch noted also
	C01162*	R	16 (R)	Yes - *spa* / MLST mismatch
Ciprofloxacin	C13185	R	1.5 (R)	No
	C01162*	R	2 (R)	Yes - *spa* / MLST mismatch
	C01092*	R	2 (R)	Yes - *spa* / MLST mismatch
	C01115*	R	2 (R)	Yes - *spa* / MLST mismatch
	C01105	R	<0.125 (S)	Yes – susceptible on repeat testing
	C01109	R	<0.125 (S)	Yes – susceptible on repeat testing
Erythromycin	C01147	R	0.75 (S)	Yes – susceptible on repeat testing
	C13384	R	0.75 (S)	Yes – susceptible on repeat testing
Erythromycin/ Clindamycin	C01224	R / R	2 (R) / ND	No / No
	C01162*	R / R	3 (R) / ND	Yes – *spa* / MLST mismatch
Fusidic acid	C13212	R	R	Yes - V90I substitution in *fus*A, compatible with resistance
	C13194	R	R	Yes -V90I substitution in *fus*A, compatible with resistance
	C01259	R	R	No
	C01276	R	R	Yes -T656S substitution in *fus*A, presumed resistant variant
Trimethoprim	C01092*	R	S	Yes – *spa* / MLST mismatch
	C01235	R	S	Yes – susceptible on repeat testing
Gentamicin	C01115*	R	24 (R)	Yes – *spa* / MLST mismatch
	C13331	R	3 (R)	No
Major errors				
Methicillin	C01222	S	16 (R)	Yes – resistant on repeat testing
	C01080*	S	0.3 (S)	Yes – *spa* / MLST mismatch
Ciprofloxacin	C01080*	S	2 (R)	Yes – resistant on repeat testing, spa / MLST mismatch noted also
Erythromycin	C01249	S	0.25 (S)	No - contains *erm*A gene, no mutations
	C01189	S	0.25 (S)	No - contains *erm*C gene, no mutations
	C01080*	S	0.5 (S)	Yes – *spa* / MLST mismatch
Tetracycline	C12796	S	0.19 (S)	No – contains *tet*K gene, no mutations
	C01247	S	0.75 (S)	No – contains *tet*M gene, no mutations
Trimethoprim	C01240	S	S	No – contains H31N mutation in *dfr*B
	C01123	S	R	No – contains *dfr*G gene, no mutations

*Isolates with *spa*-MLST mismatch; ND not done

Trimethoprim / fusidic acid: repeat testing done by disc diffusion only.

Penicillin

400 isolates tested penicillin resistant (82%). Of the 3 very major errors (see table 6.5), one isolate was susceptible on repeat testing (MIC 0.06 mg/L), while the other 2 were confirmed as resistant by E-test (MICs 0.5mg/L and 0.25mg/L).

Although using the prototype genotypic prediction method using >80% relative coverage to define resistance due to *bla*Z gave a high rate of very major errors for penicillin, using the lower relative coverage (>30%) threshold to define resistance for v2.0 resulted in a high rate of major errors (5%) in the validation set, illustrating the inherent trade-off between sensitivity and specificity for any algorithm.

Of the 25 major error isolates (*bla*Z positive but susceptible by initial phenotyping), 2 were penicillin resistant when retested by E-test (MICs 0.25mg/L), and were also positive for penicillinase production on nitrocefin disc testing. Eleven were positive for pencillinase production at ten minutes, and a further 3 isolates were weakly positive (i.e. positive after 2 hours), despite being susceptible on disc and BD Phoenix testing. Nine remaining isolates with *bla*Z had no penicillinase activity on nitrocefin disc testing and had susceptible MICs by gradient diffusion (≤0.12mg/L). Manual inspection of the *bla*Z genes found single base pair insertions in 2 cases (positions 256 and 436 respectively, relative to the reference sequence), causing a frameshift in the translated protein which may explain the lack of enzymic activity. Similarly, 4 further isolates, belonging to 3 MLST sequence types, had identical deletions at position 99, again resulting in a frameshift and premature termination and correlating with a complete absence of pencillinase activity. None of these mutations were found in the *bla*Z positive, fully phenotypically resistant isolates. After revising the algorithm for *bla*Z alone to define presence (>30% coverage) of *bla*Z with these insertions/deletions defined as susceptible (designated v2.1 of the algorithm), only 3 isolates had wild type *bla*Z with no penicillinase production detected by any of the phenotypic test methods, giving a major error rate of 0.6% which is in keeping with the observed error rates for the other antimicrobial agents.

There was poor correlation between the different phenotypic testing methods for penicillin. 404/491 isolates (82%) contained the *bla*Z gene. 25 (5.1%) were *bla*Z positive but initially tested penicillin susceptible by either disc diffusion, BD Phoenix or both. Two of these were resistant by E-test, however 16/25 (64%) were positive by nitrocefin disc testing only.

Methicillin

57 isolates (12%) were MRSA by initial phenotype. Initially, there were 2 very major errors and 2 major errors for methicillin. Of the very major errors, one isolate was susceptible on repeat testing (MIC 0.5mg/L) and similarly, one of the major error isolates was resistant on repeat testing (MIC 16mg/L). Both methicillin resistance in the absence of *mec*A resulting from overexpression of penicillinase (206), and methicillin susceptibility despite the presence of *mecA* have been previously described (146, 207), which may explain the heterogeneity observed. One further VME and ME occurred in isolates with *spa*-MLST mismatched isolates and these isolates were therefore excluded from further analysis. None of the isolates contained *mec*C (157).

Ciprofloxacin

70 isolates were phenotypically resistant (14%). Initially there were 6 VMEs. Three of these occurred in *spa*-MLST mismatches and were discounted, and 2 were susceptible on repeat phenotypic testing. For the remaining discrepant isolate, the *grl*B, *gyr*B and *nor*A genes were examined but no previously described resistance associated mutations were found (208, 209). There was a single ME which was found to be resistant on repeat testing (MIC 2mg/L). Of the 64 isolates predicted to be resistant by genotype and resistant by phenotype, 60 had the amino acid substitutions S80F in *grl*A and S84L *in gyr*A. Four isolates had S84L in *grl*A only.

Erythromycin / Clindamycin

83 (17%) isolates were resistant to erythromycin by phenotype. Initially there were 4 very major errors. Two of these were susceptible on repeat testing (MICs 0.75mg/L), 1 occurred in a *spa*-MLST mismatch and 1 was confirmed as resistant with no evidence of any *erm* or *msr*A genes on BLAST. There were 3 major errors, one occurring in a *spa*-MLST mismatched and 2 confirmed as susceptible on E-test (MICs 0.25mg/L). One had *erm*C detected by BLAST, and the other had *erm*A (both with 100% relative coverage). For the concordant resistant isolates, 37 had *erm*A alone, 37 had *erm*C alone, 1 had *erm*T, 2 had *msr*A alone, 1 had *erm*C plus *msr*A, and 1 had *erm*C plus *erm*A. None of the isolates had *erm*B. Out of 81 erythromycin resistant isolates, 45 were resistant to clindamycin by disc diffusion and 36 were susceptible. Of these, 34 had inducible resistance by D-test, and contained either *erm*A, C or T as above. The 2 isolates which were resistant to erythromycin but susceptible to clindamycin contained the *msr*A gene only. There was 1 very major error (same isolate as for erythromycin VME). There was no correlation between *erm* variant and inducible clindamycin resistance. The recently described *vga*$(A)_{LC}$ gene, which confers resistance to clindamycin but not erythromycin, was not found in any of the isolates (210).

Tetracycline

18 isolates were phenotypically resistant (4%). There were no very major errors, and 2 major errors. Both were confirmed susceptible by E-test. One had *tet*K detected by BLAST (100% relative coverage, MIC 0.19mg/L) and the other had *tet*M (95% relative coverage, MIC 0.75mg/L). Of the 18 concordant resistant isolates, 16 had *tet*K, 1 had both *tet*K and *tet*L, and 1 had *tet*M.

Vancomycin

Vancomycin resistance was not identified in any of the isolates either phenotypically or genotypically. Consequently, although the specificity of the method was 1.00, its sensitivity

is not estimable due to the rarity of *van*A-mediated resistance. To investigate the possibility of intermediate resistance missed by phenotyping, the collection was also screened post hoc for recently published mutations in the *yyc*G gene (211) found to be associated with intermediate MICs in laboratory mutants, but none of these mutations were found in any of the isolates.

Fusidic acid

43 isolates were phenotypically resistant (9%). There were 4 very major errors, all confirmed by repeat phenotyping. In 2 cases, manual examination of the sequences revealed single point mutations in the chromosomal *fus*A, predicted to result in the amino acid substitution of isoleucine for valine at position 90. This substitution has been reported as a resistance conferring mutation but it has also been observed in susceptible isolates (212) and therefore its role in resistance is unresolved. One resistant isolate had a substitution of serine for threonine at position 656 (predicted sensitive as not found in the literature review). Since the substitution of lysine for threonine at the same position is associated with full phenotypic resistance (212), this may explain the discrepancy seen in this case, leaving a single unexplained VME. There were no major errors.

Trimethoprim

Phenotypic testing for trimethoprim resistance was only performed for 203 isolates, as it is not routinely tested in Brighton and it was not part of the BD Phoenix panel used. Results are therefore available for Oxford isolates, tested only by disc diffusion. Three isolates were phenotypically resistant (1.5%). Initially there were 2 very major errors and 2 major errors. Both very major error isolates were susceptible on repeat testing (performed by disc diffusion only) and 1 of the major error isolates was resistant on repeat testing. The susceptible ME isolate had a histidine to asparagine substitution at position 31 (H31N), which has previously been reported as conferring resistance (213).

Gentamicin

Four isolates were phenotypically resistant (1%). Initially there were 2 very major errors, one of which occurred in a *spa*-MLST discrepant isolate. In the remaining isolate, resistance was confirmed by repeat testing (MIC 3 mg/L). *Aac*A/*aph*D was not found in any isolate, and they were also negative for *aad*D, *aad*E and *aph*A3. The low frequency of resistance in this collection means the evaluation is under-powered and makes accurately estimating overall sensitivity impossible.

Mupirocin and rifampicin

2 (0.4%) and 5 (1%) isolates were resistant to mupirocin or rifampicin respectively. There were no errors observed, however the relatively low resistance rates make estimating sensitivity impossible. Mupirocin resistance in both cases was attributable to *mup*A.

6.4 Discussion

The final results show highly promising concordance, with a final very major error rate of 0.9% and a major error rate of 0.2% using the v2.1 algorithm. This is comparable with those for current phenotypic methodologies (table 6.8) and within the acceptable limits set by the US Food and Drug Administration (214) for marketing approval of new susceptibility testing methods (<1.5% very major error rate, <3% major error rate).

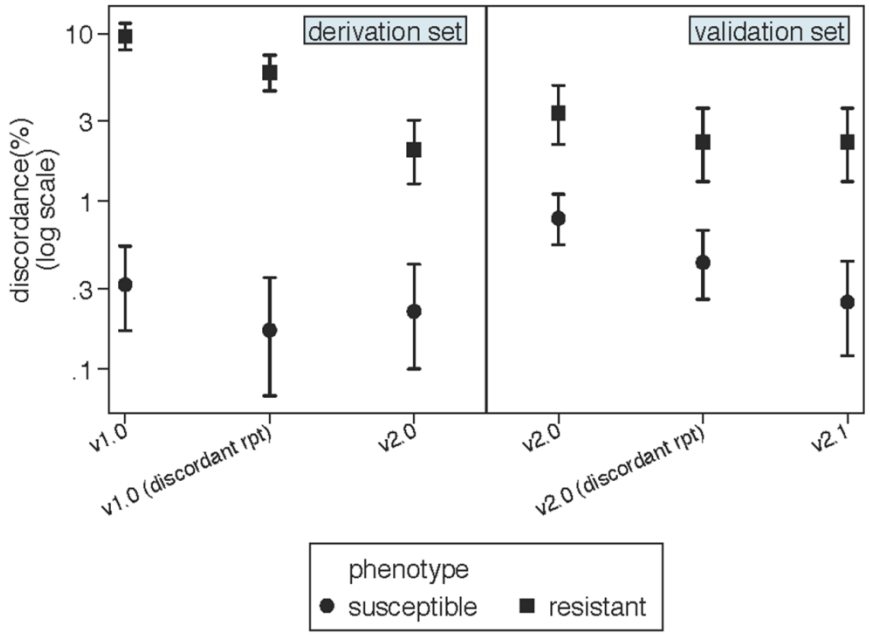
In both the derivation and validation sets, a substantial number (40%) of the initial observed errors were resolved by repeat phenotypic susceptibility testing (figure 6.2).

TABLE 6.8 ME AND VME RATES FOR STUDIES COMPARING RESISTOTYPING METHODS

Study	Comparison	No. of isolates	Categorical agreement (%)	ME rate (%)	VME rate (%)
Ligozzi 2002	Vitek 2 vs agar dilution	100	94-100	0	0
Fahr 2003	BD Phoenix vs broth dilution plus *mec*A PCR	116	97.6	1.2	1.7
Nonhoff 2005	Vitek 2 vs agar dilution	273	-	1.5	0.7
Carroll 2006	BD Phoenix vs agar dilution	232	98.2	0.3	0.4
Giani 2012	BD Phoenix vs broth dilution	95	98	1.3	2.1
Bobenchik 2014	Vitek 2 vs broth dilution	134	98.9	0.1	1.4
This study	WGS vs combined disc diffusion / BD Phoenix	491	98.8	0.2	1.1

The gold standard method used in each study is underlined.

FIGURE 6.2 COMPARISON OF % OF ERRORS FOR DERIVATION AND VALIDATION SETS



Illustrates change in error rate with repeat phenotyping and with optimised algorithm versions. Error rates for resistant and susceptible isolates are shown for each step of algorithm development: in the derivation set, the error rate was decreased by repeat susceptibility testing (discordant rpt). VME rate was reduced by adjusting the algorithm parameters (v2.0), although this resulted in a slight increase in ME rate. In the validation set, error rates were relatively low to start with and were improved further by repeat testing of the discordants (v2.0 discordant rpt) and by incorporating the novel *blaZ* mutations (v2.1).

Even when testing according to published guidelines, occasional errors in reagent and media storage, incubation conditions, and inoculum density may contribute to variation in observed phenotype, and similarly, inaccuracies in labelling, interpretation and data entry are liable to occur in any system which is not fully automated. Some of these factors may also contribute to the genotyping results, and this is supported by the four isolates where the predicted MLST type did not match the known *spa*-type, suggesting a labelling or storage error.

The most problematic antimicrobial agent was penicillin, with an unacceptably high very major error rate (1.8%) using v1.0 in the derivation set, and an unacceptably high major error rate (3.4%) using v2.0 in the in the validation set. The initial high very major error rate may be due to the variable location of *bla*Z*,* which may be on a plasmid or integrated into the chromosome (215). Isolates with chromosomally integrated *bla*Z are likely to have coverage in the sequencing reads commensurate with the rest of the sequence, while isolates with plasmid-carried copies may have very high (if multiple copies are carried) or very low (because of poor mapping to the reference) coverage in that region. As a result, these regions may be rejected as poor quality by the assembly software, because they fall outside the coverage levels of the rest of the genome. This problem may be overcome in future with longer reads or alternative methods for *de novo* assembly; however, these results highlight that for a BLAST approach, relative coverage cut-off values may need to be set individually based on gene location. Receiver operator curve (ROC) analysis would be one method of doing this. Conversely, in the validation set, most of the major errors were due to a lack of concordance between penicillinase production and disc or broth dilution testing (216). Previous authors have also raised concern about the ability of phenotypic tests to detect true penicillin resistance (217).
Whole genome sequencing readily identified previously unreported mutations in the *bla*Z gene which correlated with reversion to susceptible penicillin phenotype. Insertion / deletions caused frameshifts in the translated proteins, presumably accounting for the lack of enzyme activity. These mutations were not seen in any of the *bla*Z positive, penicillin resistant isolates. Four susceptible isolates contained the same mutation, giving a high degree of confidence

that this is causative. Similar frameshift deletions have been reported in *bla*Z positive, penicillin susceptible isolates from bovine mastitis (218).

Where isolates remain susceptible, penicillin remains the drug of choice for *S. aureus*, although widespread resistance means that some laboratories no longer test for penicillin susceptibility (219). In the validation set, 404/491 (18%) of isolates were *bla*Z negative, indicating that penicillin may still have a role in the management of some *S. aureus* infections in the UK, although susceptibility should be confirmed by e.g. nitrocefin testing or identification of a functional *bla*Z gene.

In the derivation set, the highest overall error rate was for ciprofloxacin. Staphylococcal resistance to quinolones is predominantly due to point mutations in the *grl*A and *gyr*A genes (208). Low level resistance may result from mutations in *grl*B / *gyr*B or alterations in expression of the efflux pump NorA (209). However, in most cases these are in combination with *grl*A / *gyr*A mutations, and consequently the overall phenotypic effect is difficult to predict. All isolates in this study contained the *nor*A gene, but no association was seen between quinolone resistance and the *nor*A gene or its known regulators, or the *grl*B / *gyr*B genes. Further studies may be able to elucidate the contribution of these to overall quinolone resistance in *S. aureus*.

Restricting the analysis to those antimicrobials most commonly used for severe or invasive infection (methicillin, ciprofloxacin + rifampicin, clindamycin, vancomycin) gives even more promising results, with an overall VME rate of 0.1% and an ME rate of 0%. The remaining antimicrobials are usually used as adjuvant treatment, in combination for chronic infections (e.g. bone infection), or minor skin infections and consequently there is less urgency in identifying antimicrobial susceptibility.

There remains a very small subset of 15 isolates where the genetic basis for an antimicrobial phenotype of the organism was not clear (excluding presumed phenotyping errors or identified novel variants). If not due to human error, these may be due to sequencing or assembly error, for example a miscalled base at a critical position. However, comparisons using the MRSA

252 reference genome, which is included routinely on all *S. aureus* HiSeq plates run in Oxford, have estimated an in-house false positive rate (detection of a spurious variant) for the bioinformatics pipeline to be 2.5×10^{-9} per nucleotide, i.e. 0.0075 per genome (107). Consequently, although an incorrect susceptibility prediction may in theory occur due to a sequencing error, in practice it is anticipated that the impact of this will be extremely small.

A more likely explanation for the discrepancies may be as yet unidentified alterations in regulatory regions or alternative resistance mechanisms. This highlights a major challenge for *in silico* resistotyping, since a query-based BLAST method cannot recognise novel variants which are not in the relevant database. Furthermore, gene expression is the result of complex interactions between transcription promoters, repressors and other regulatory molecules, which may be remotely located from the gene itself. Phenotypic assays address these challenges by measuring the overall combined impact of all these mechanisms, although it is important to recognise that factors resulting in delayed transcription may cause isolates to be falsely identified as susceptible, as found for penicillin.

It is unlikely therefore that WGS will be able to replace phenotypic methods entirely and some form of phenotypic surveillance will need to be maintained, for example based on clusters of treatment failures. However, the need for routine phenotyping should diminish as examination of WGS and phenotyping data for isolates with apparently novel mutations or with non-functioning resistance genes elucidate the contribution of the underlying genetics. This new knowledge can then be added to resistance determinant databases and absorbed into WGS investigations, as demonstrated by the improved specificity resulting from the incorporation of the novel *bla*Z variants above into the algorithm. The huge potential of WGS data lies in its completeness: once sequenced, a genome can be accessed repeatedly to query for novel genes of interest as these arise (e.g. for recently published *yyc*G gene mutations (211) which were not included in the original gene catalogue).

Current turnaround times are directly comparable, with next generation sequencers able to deliver results in 27 hours (146). Newer nanopore technologies are reducing this to a matter of hours (220). Similarly, advances in software are allowing for much more rapid processing times: the data generated in this chapter has also been used to develop and test a method for extracting resistance information from raw reads, with error rates similar to the BLAST based approach (183). Combined with nanopore sequencing, this has the potential to provide clinically useful results in less than 24 hours. The same methods can also be used to ascertain the presence / absence of virulence factors such as exfoliative toxin, PVL or toxic shock syndrome toxin (114).

The potential for full automation of WGS may also reduce human error, and reporting of results should be as easily integrated into electronic laboratory systems as automated broth dilution methods.

6.5 Conclusion

The results demonstrate that WGS can predict phenotypic behaviour with a high degree of confidence. The advances provided by WGS, combined with robust clinical outcome data, should greatly enhance our understanding of the genetic basis and clinical impact of antimicrobial resistance, with the potential for identifying new antimicrobial drug targets. The consequent promise of improved drug discovery in the face of current global concern regarding emerging antimicrobial resistance makes the prospect of routine use of WGS increasingly attractive.

7 Conclusion and future work

In recent years, the emergence of multi-drug resistant Gram negative pathogens such as ESBL and carbapenemase producing *E. coli* and *Klebsiella pneumoniae* has led to a shift in focus away from *S. aureus*. Combined with the steady decline in MRSA bacteraemia rates over the past decade, both in the UK and abroad (84, 221-223) , it would be tempting to think that the "war" against MRSA has been won, and to concentrate public health and infection control efforts at the containment of emerging MDR organisms.

The reduction in MRSA bacteraemia follows sustained, concerted efforts over several years, with interventions aimed at improving vascular line care, decolonisation of carriers, and measures to prevent infections such as surgical site infections which may lead to bacteraemia. Most of these interventions are part of "care bundles" comprising multiple factors, and determining which elements are most effective is difficult (224-226). In some cases, a reduction in bacteraemia rates predated intensification of infection control measures, suggesting that there may also be strain specific factors influencing overall epidemic behaviour (184).

In addition, surveillance of colonisation or non-bacteraemia infections is not mandated in the UK, so the overall disease burden remains unknown. More worryingly, rates of MSSA bacteraemia in England have shown a year on year increase over the past decade (154), despite the assumption that interventions which reduce transmission and infection due to MRSA would have the same effect on MSSA.

In spite of encouraging progress, morbidity and mortality for *S. aureus* bacteraemia remain high, and place a considerable burden on healthcare due to the need for prolonged intravenous antibiotics and investigations necessary to exclude seeding to or from deep infections such as bone / joint infections, abscesses or endocarditis (94, 227).

In the US, the USA300 MRSA strain is an ongoing public health concern (228), and there is potential for a similar rapid clonal expansion in the UK, as demonstrated by the t008 MRSA

outbreak described in chapter 4, where there was evidence of dissemination of the strain into the wider community. There is, therefore, no room for complacency, and active surveillance and control efforts for *S. aureus* should not be forgotten in the efforts to contain MDR Enterobacteriaceae. The apparent success of efforts to reduce MRSA has not resulted in a reduction of MSSA, and therefore ongoing effort is still required to understand, treat and control *S. aureus*.

The aim of my thesis has been to examine the use of a new technology, whole genome sequencing, to enhance the knowledge of the *S. aureus* transmission and resistance, particularly in the context of outbreaks. I have adopted a stepwise approach, developing, calibrating, and testing WGS against current standard methods for outbreak typing and antimicrobial susceptibility testing.

Chapter 2 established what happens in acquisition of nasal carriage, and demonstrated minimal diversity in early nasal carriage, aiming to allay concern over interpretation of sequencing data. The results provide evidence for the use of single colony sequencing in outbreaks, allowing a simplified approach to outbreak investigation and helping to minimise costs.

In chapter 3, I used well-characterised outbreaks to calibrate the WGS approach. The results are broadly comparable with routine PFGE results in outbreaks involving a common clone, but also demonstrate enhanced resolution allowing exclusion of some isolates from outbreaks. The WGS approach also gave more robust evidence for putative transmission links. Since the majority of the outbreaks were sequenced retrospectively, the true impact of WGS is hard to assess, however overall the infection control teams responded favourably and stated that the information was useful in confirming the outbreak cases with greater certainty.

The SNV thresholds observed in these outbreaks were in keeping with the small number of outbreaks already described in the literature, with highly conserved core genomes again providing reassurance that single colony sequencing can provide useful information.

Chapter 4 examined 5 outbreaks where the epidemiological links were unclear, and again demonstrated the greater resolution of WGS compared with PFGE. Analysis of the time to most recent common ancestor also demonstrated that, as well as greater resolution, WGS may give additional information and can be used to direct the infection control response, for example providing support for wider screening for a long-term carrier. This follows on from the results from chapters 2 and 3, where it is shown that recent acquisition is associated with lower diversity. This seems intuitive in retrospect, but this is the first piece of work to demonstrate this consistently, and is borne out by the surprising clonality of all the outbreaks. Rather than confounding the investigation, these results show how within-host diversity may be utilised to improve the understanding of outbreak dynamics.

In chapter, I applied WGS in real time. WGS did in fact identify 3 clusters undetected by routine surveillance, however it was clear from the epidemiological data that the actual transmissions predated isolation of the MRSA strains, in many cases by several months. The utility of WGS was hindered by the sampling strategy, as, in the absence of routine inpatient surveillance or an outbreak of invasive disease, real-time sequence was not accompanied by real-time sampling. Real-time screening of MRSA is only likely to be effective in a higher prevalence setting and where there is ongoing inpatient surveillance: in the UK at present this is only carried out in intensive care units or in the context of enhanced surveillance as part of an outbreak investigation. Overall, this chapter has highlighted some of the issues with current infection control surveillance practices, for examples excluding isolates on the basis of antibiograms, incomplete sampling frames due to poor compliance with screening, and the lack of surveillance screening in between clusters of disease. However, the cost-effectiveness of increasing screening frequency in conjunction with WGS will require a formal economic costing to determine viability, something that is outside the scope of this thesis but which is being carried forward by the MMM group for future work.

In chapter 6, I demonstrated that WGS can predict phenotypic antimicrobial resistance behaviour with a high degree of confidence. The fact that many of the apparent discrepancies were resolved by repeat phenotyping shows that current routine tests are subject to significant errors, and it may be that by removing the need for human interpretation, WGS may ultimately be more accurate for antimicrobial susceptibility testing. This chapter demonstrated the capability of WGS as a highly useful tool for investigating the epidemiology of resistance, with the rapid ability to investigate new resistance factors and the capacity for studying the spread of resistant elements (e.g. phage or plasmid borne resistance), as well as the spread of the organism itself. This has been demonstrated elsewhere for determining the molecular epidemiology of multi-drug resistant plasmids in outbreaks of multiply resistant Enterobacteriaceae (229) .

Genetic testing for MRSA via rapid PCR has been available for many years (230, 231) and has been used successfully in screening programmes to reduce turn-around times and allow timely intervention to prevent onward transmission (232). However, this is a single, organism-specific platform which tests for one gene only (*mec*A), making it an expensive outlay for a laboratory. In contrast, next generation WGS platforms such as the MiSeq demonstrate the advances in sequencing technology, and are rapidly approaching the cost and turn-around of PCR-based technologies. WGS methods have the advantage of being organism-agnostic, so that a single platform can be used across samples and organisms. It seems inevitable that in a few years, benchtop WGS will be the technology of choice in many laboratories.

The work for this thesis was undertaken as part of a collaboration with Public Health England, and work is ongoing to enhance the resistance work (e.g. expanding the catalogue and comparing different methods). While PFGE remains the accepted gold standard for outbreak investigation, WGS is now being used by the reference laboratory in selected outbreaks as part of a transition to its routine adoption.

While this thesis outlines an approach for using WGS for *S. aureus*, this was chosen as an exemplar organism of public health importance and the methodology should be broadly applicable to other organisms. Colleagues have used similar approaches to investigate *Mycobacterium tuberculosis*, norovirus, and multi-drug resistant *E. coli*, all presenting different complexities both in sequencing and epidemiology. It is hoped that this work will be formative in establishing the routine use of WGS both for *S. aureus* and other organisms in the ongoing challenge to understand these pathogens.

o0o

8 Glossary of terms

Bayesian: incorporation of prior knowledge into probability statements about a model based on evidence from observed data.

De Bruijn graph: sequences are split into shorter lengths (k-mers), which are assembled based on k-1 common nucleotides to give a topology which can be traversed to deduce the full sequence

De novo assembly: assembly of short sequence data into large contigs without using a reference genome. Programmes include e.g. Velvet, BWA, BowTie

HKY substitution model: substitution model which distinguishes between transitions (purine to pyramidine or vice versa), substitutions, and transversions (i.e. gives a higher probability to a purine-purine substitution), and allows for unequal base frequencies.

Jukes Cantor (JC69) substitution model: Simplest substitution model which assumes equal base frequencies and equal mutation rates (e.g. base A is equally likely, following a mutation event, to be replaced by T, C, G or even A.

Maximum clade credibility tree: summarises the result of Bayesian inference: from the possible trees which could be constructed from a set of sequences; gives the tree with the maximum product of the posterior clade probabilities

MCMC (Markov Chain Monte Carlo) method: derived from casino statistical methods. A random set of variables is generated for X and an "energy" E calculated where E= E(X). X is then modified to give X', and E(X') calculated. If this gives a better valued for E, the process is repeated from X'. This is repeated a specified number of times to achieve an optimal value for X.

Sanger sequencing (dideoxy chain-termination method): method of DNA sequencing where a complementary strand is synthesised in the presence of labelled (radio or fluorescent) chain-terminating nucleotides. Multiple strands of different lengths are generated with a labelled terminal nucleotide. These can be separated according to molecular weight by running on a gel or by capillary separation, to determine the order of nucleotides.

9 References

1. Leewenhoeck A. An Abstract of a Letter from Mr. Anthony Leewenhoeck at Delft, Dated Sep. 17. 1683. Containing Some Microcopical Observations, about Animals in the Scurf of the Teeth, the Substance Call'd Worms in the Nose, the Cuticula Consisting of Scales. Philosophical Transactions. 1684;14:568-74.
2. Ewart JC. On the Life History of Bacillus Anthracis. Quarterly Journal of Microscopical Science. 1878;s2-18:161-70.
3. Geddes PE, J. C. On the Life-History of Spirillum. Proc R Soc London. 1878;27(185-189):481-5.
4. Wright J. Reports of Societies: Aberdeen, Banf, Kincardine Branch, Wednesday February 4th, 1880. BMJ. 1880;1(1003):442-3.
5. Kocher T. Zur Aetiologie der acuten Entzündungen Von Prof Dr Theodor Kocher. Archiv für klinische Chirurgie. 1879;23:101.
6. Cheyne WW. On the relation of organisms to antiseptic dressings. Transactions of the Pathological Society of London. 1879;30:557-82.
7. Ogston A. Report upon Micro-Organisms in Surgical Diseases. British medical journal. 1881;1(1054):369 b2-75.
8. Bilroth T. Untersuchungen über die Vegetationsformen von Coccobacteria septica. Berlin: C. Reimer; 1874.
9. Lister J. On the Antiseptic Principle in the Practice of Surgery. British medical journal. 1867;2(351):246-8.
10. Rosenbach AJ. Mikro-Organismen bei den Wund-Infections-Krankheiten des Menschen. Wiesbaden: J. F. Bergmann; 1884.
11. Landy JJ, Havens I, Clarke JS, Benham RS. An outbreak of wound infections due to antibiotic-resistant *Staphylococcus aureus*. Surgical forum. 1955;5:814-7.
12. Bartlett PC, Martin RJ, Cahill BR. Furunculosis in a high school football team. The American journal of sports medicine. 1982;10(6):371-4.
13. Dammann TA, Wiens RM, Taylor GD. Methicillin resistant *Staphylococcus aureus*: identification of a community outbreak by monitoring of hospital isolates. Canadian journal of public health. 1988;79(5):312-4.
14. Sosin DM, Gunn RA, Ford WL, Skaggs JW. An outbreak of furunculosis among high school athletes. The American journal of sports medicine. 1989;17(6):828-32.
15. Chira S, Miller LG. *Staphylococcus aureus* is the most common identified cause of cellulitis: a systematic review. Epidemiology and infection. 2010;138(3):313-7.
16. Levy PY, Ollivier M, Drancourt M, Raoult D, Argenson JN. Relation between nasal carriage of *Staphylococcus aureus* and surgical site infection in orthopedic surgery: the role of nasal contamination. A systematic literature review and meta-analysis. Orthopaedics & traumatology, surgery & research. 2013;99(6):645-51.
17. Ulug M, Ayaz C, Celen MK. A case report and literature review: osteomyelitis caused by community-associated methicillin resistant *Staphylococcus aureus*. Journal of infection in developing countries. 2011;5(12):896-900.
18. Bowles D, Perrin K. A retrospective case series of 44 patients with community-acquired *Staphylococcus aureus* pneumonia. The New Zealand medical journal. 2014;127(1391):74-83.
19. Rello J, Torres A, Ricart M, Valles J, Gonzalez J, Artigas A, et al. Ventilator-associated pneumonia by *Staphylococcus aureus*. Comparison of methicillin-resistant and methicillin-sensitive episodes. American journal of respiratory and critical care medicine. 1994;150(6 Pt 1):1545-9.
20. Espersen F, Frimodt-Moller N. *Staphylococcus aureus* endocarditis. A review of 119 cases. Archives of internal medicine. 1986;146(6):1118-21.
21. Brandt CM, Sistrunk WW, Duffy MC, Hanssen AD, Steckelberg JM, Ilstrup DM, et al. *Staphylococcus aureus* prosthetic joint infection treated with debridement and prosthesis retention. Clinical infectious diseases. 1997;24(5):914-9.

22. Wyllie DH, Crook DW, Peto TE. Mortality after *Staphylococcus aureus* bacteraemia in two hospitals in Oxfordshire, 1997-2003: cohort study. BMJ. 2006;333(7562):281.
23. Johnson AP, Aucken HM, Cavendish S, Ganner M, Wale MC, Warner M, et al. Dominance of EMRSA-15 and -16 among MRSA causing nosocomial bacteraemia in the UK: analysis of isolates from the European Antimicrobial Resistance Surveillance System (EARSS). The Journal of antimicrobial chemotherapy. 2001;48(1):143-4.
24. Moore PC, Lindsay JA. Molecular characterisation of the dominant UK methicillin-resistant *Staphylococcus aureus* strains, EMRSA-15 and EMRSA-16. Journal of medical microbiology. 2002;51(6):516-21.
25. Tenover FC, Goering RV. Methicillin-resistant *Staphylococcus aureus* strain USA300: origin and epidemiology. The Journal of antimicrobial chemotherapy. 2009;64(3):441-6.
26. Kluytmans JA, Wertheim HF. Nasal carriage of *Staphylococcus aureus* and prevention of nosocomial infections. Infection. 2005;33(1):3-8.
27. Kearns A. MRSA outbreaks referred to PHE Staphylococcus reference laboratory, personal communication, 2012.
28. Cohen ML, Wong ES, Falkow S. Common R-plasmids in *Staphylococcus aureus* and *Staphylococcus* epidermidis during a nosocomial *Staphylococcus aureus* outbreak. Antimicrobial agents and chemotherapy. 1982;21(2):210-5.
29. Archer GL, Mayhall CG. Comparison of epidemiological markers used in the investigation of an outbreak of methicillin-resistant *Staphylococcus aureus* infections. Journal of clinical microbiology. 1983;18(2):395-9.
30. Collins JK, Smith JS, Kelly MT. Comparison of phage typing, plasmid mapping, and antibiotic resistance patterns as epidemiologic markers in a nosocomial outbreak of methicillin-resistant *Staphylococcus aureus* infections. Diagnostic microbiology and infectious disease. 1984;2(3):233-45.
31. Venezia RA, Harris V, Miller C, Peck H, San Antonio M. Investigation of an outbreak of methicillin-resistant *Staphylococcus aureus* in patients with skin disease using DNA restriction patterns. Infection control and hospital epidemiology. 1992;13(8):472-6.
32. Stephenson JR, Crook SJ, Tabaqchali S. New method for typing *Staphylococcus aureus* resistant to methicillin based on sulphur-35 methionine labelled proteins: its application in an outbreak. Br Med J (Clin Res Ed). 1986;293(6547):581-3.
33. Pekkala DH, Low DE, Wyper PA, Rose D, Sturman D, Pritchard MF, et al. The utility of restriction endonuclease analysis and phage typing in the epidemiologic investigation of a *Staphylococcus aureus* outbreak in a neonatal nursery. Diagnostic microbiology and infectious disease. 1992;15(4):307-11.
34. Maslow JN, Mulligan ME, Arbeit RD. Molecular epidemiology: application of contemporary techniques to the typing of microorganisms. Clinical infectious diseases. 1993;17(2):153-62
35. Tenover FC, Arbeit R, Archer G, Biddle J, Byrne S, Goering R, et al. Comparison of traditional and molecular methods of typing isolates of *Staphylococcus aureus*. Journal of clinical microbiology. 1994;32(2):407-15.
36. Schwartz DC, Cantor CR. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. Cell. 1984;37(1):67-75.
37. Tenover FC, Arbeit RD, Goering RV, Mickelsen PA, Murray BE, Persing DH, et al. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. Journal of clinical microbiology. 1995;33(9):2233-9.
38. Bannerman TL, Hancock GA, Tenover FC, Miller JM. Pulsed-field gel electrophoresis as a replacement for bacteriophage typing of *Staphylococcus aureus*. Journal of clinical microbiology. 1995;33(3):551-5.
39. Leonard RB, Mayer J, Sasser M, Woods ML, Mooney BR, Brinton BG, et al. Comparison of MIDI Sherlock system and pulsed-field gel electrophoresis in characterizing strains of methicillin-resistant *Staphylococcus aureus* from a recent hospital outbreak. Journal of clinical microbiology. 1995;33(10):2723-7.

40. Schmitz FJ, Steiert M, Tichy HV, Hofmann B, Verhoef J, Heinz HP, et al. Typing of methicillin-resistant *Staphylococcus aureus* isolates from Dusseldorf by six genotypic methods. Journal of medical microbiology. 1998;47(4):341-51.
41. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proceedings of the National Academy of Sciences of the United States of America. 1998;95(6):3140-5.
42. Frenay HM, Bunschoten AE, Schouls LM, van Leeuwen WJ, Vandenbroucke-Grauls CM, Verhoef J, et al. Molecular typing of methicillin-resistant *Staphylococcus aureus* on the basis of protein A gene polymorphism. European journal of clinical microbiology & infectious diseases. 1996;15(1):60-4.
43. Harmsen D, Claus H, Witte W, Rothganger J, Claus H, Turnwald D, et al. Typing of methicillin-resistant *Staphylococcus aureus* in a university hospital setting by using novel software for spa repeat determination and database management. Journal of clinical microbiology. 2003;41(12):5442-8.
44. Faria NA, Carrico JA, Oliveira DC, Ramirez M, de Lencastre H. Analysis of typing methods for epidemiological surveillance of both methicillin-resistant and methicillin-susceptible *Staphylococcus aureus* strains. Journal of clinical microbiology. 2008;46(1):136-44.
45. Stefani S, Chung DR, Lindsay JA, Friedrich AW, Kearns AM, Westh H, et al. Meticillin-resistant *Staphylococcus aureus* (MRSA): global epidemiology and harmonisation of typing methods. International journal of antimicrobial agents. 2012;39(4):273-82.
46. Kazakova SV, Hageman JC, Matava M, Srinivasan A, Phelan L, Garfinkel B, et al. A clone of methicillin-resistant *Staphylococcus aureus* among professional football players. The New England journal of medicine. 2005;352(5):468-75.
47. Rao N, Jacobs S, Joyce L. Cost-effective eradication of an outbreak of methicillin-resistant *Staphylococcus aureus* in a community teaching hospital. Infection control and hospital epidemiology. 1988;9(6):255-60.
48. Mehtar S, Drabu YJ, Mayet F. Expenses incurred during a 5-week epidemic methicillin-resistant *Staphylococcus aureus* outbreak. The Journal of hospital infection. 1989;13(2):199-200.
49. Wertheim HF, Melles DC, Vos MC, van Leeuwen W, van Belkum A, Verbrugh HA, et al. The role of nasal carriage in *Staphylococcus aureus* infections. The Lancet Infectious diseases. 2005;5(12):751-62.
50. Safdar N, Bradley EA. The risk of infection after nasal colonization with *Staphylococcus aureus*. The American journal of medicine. 2008;121(4):310-5.
51. Vigil DI, Harden WD, Hines AE, Hosokawa PW, Henderson WG, Bessesen MT. Risk of MRSA Infection in Patients with Intermittent versus Persistent MRSA Nares Colonization. Infection control and hospital epidemiology. 2015:1-6.
52. Coello R, Glenister H, Fereres J, Bartlett C, Leigh D, Sedgwick J, et al. The cost of infection in surgical patients: a case-control study. The Journal of hospital infection. 1993;25(4):239-50.
53. Kanerva M, Blom M, Tuominen U, Kolho E, Anttila VJ, Vaara M, et al. Costs of an outbreak of meticillin-resistant *Staphylococcus aureus*. The Journal of hospital infection. 2007;66(1):22-8.
54. Plowman R, Graves N, Griffin MA, Roberts JA, Swan AV, Cookson B, et al. The socio-economic burden of hospital acquired infection: executive summary. 1999. Department of Health, London, UK.
55. Rampling A, Wiseman S, Davis L, Hyett AP, Walbridge AN, Payne GC, et al. Evidence that hospital hygiene is important in the control of methicillin-resistant *Staphylococcus aureus*. The Journal of hospital infection. 2001;49(2):109-16.
56. Lee SC, Chen KS, Tsai CJ, Lee CC, Chang HY, See LC, et al. An outbreak of methicillin-resistant *Staphylococcus aureus* infections related to central venous catheters for hemodialysis. Infection control and hospital epidemiology. 2004;25(8):678-84.
57. Batty D. Actor Ash gets £5m payout over hospital superbug. The Guardian. 17 January 2008.

58. Wiese-Posselt M, Heuck D, Draeger A, Mielke M, Witte W, Ammon A, et al. Successful termination of a furunculosis outbreak due to lukS-lukF-positive, methicillin-susceptible *Staphylococcus aureus* in a German village by stringent decolonization, 2002-2005. Clinical infectious diseases. 2007;44(11):e88-95.
59. Higashiyama M, Ito T, Han X, Nishiyama J, Tanno A, Wada T, et al. Trial to control an outbreak of Panton-Valentine leukocidin-positive methicillin-resistant *Staphylococcus aureus* at a boarding school in Japan. American journal of infection control. 2011;39(10):858-65.
60. Yamamoto T, Takano T, Yabe S, Higuchi W, Iwao Y, Isobe H, et al. Super-sticky familial infections caused by Panton-Valentine leukocidin-positive ST22 community-acquired methicillin-resistant *Staphylococcus aureus* in Japan. Journal of infection and chemotherapy. 2012;18(2):187-98.
61. Khojasteh VJ, Edwards-Jones V, Childs C, Foster HA. Prevalence of toxin producing strains of *Staphylococcus aureus* in a pediatric burns unit. Burns : journal of the International Society for Burn Injuries. 2007;33(3):334-40.
62. Lindenmayer JM, Schoenfeld S, O'Grady R, Carney JK. Methicillin-resistant *Staphylococcus aureus* in a high school wrestling team and the surrounding community. Archives of internal medicine. 1998;158(8):895-9.
63. Elias AF, Chaussee MS, McDowell EJ, Huntington MK. Community-based intervention to manage an outbreak of MRSA skin infections in a county jail. Journal of correctional health care. 2010;16(3):205-15.
64. Fanoy E, Helmhout LC, van der Vaart WL, Weijdema K, van Santen-Verheuvel MG, Thijsen SF, et al. An outbreak of non-typeable MRSA within a residential care facility. EuroSurveillance. 2009;14(1).
65. Gasch O, Hornero A, Dominguez MA, Fernandez A, Suarez C, Gomez S, et al. Methicillin-susceptible *Staphylococcus aureus* clone related to the early pandemic phage type 80/81 causing an outbreak among residents of three occupational centres in Barcelona, Spain. Clinical microbiology and infection. 2012;18(7):662-7.
66. Jensen JU, Jensen ET, Larsen AR, Meyer M, Junker L, Ronne T, et al. Control of a methicillin-resistant *Staphylococcus aureus* (MRSA) outbreak in a day-care institution. The Journal of hospital infection. 2006;63(1):84-92.
67. Chan P, Dipper A, Kelsey P, Harrison J. Newspaper reporting of meticillin-resistant *Staphylococcus aureus* and 'the dirty hospital'. The Journal of hospital infection. 2010;75(4):318-22.
68. Boyce JM. Should we vigorously try to contain and control methicillin-resistant *Staphylococcus aureus*? Infection control and hospital epidemiology. 1991;12(1):46-54.
69. Stein M, Navon-Venezia S, Chmelnitsky I, Kohelet D, Schwartz O, Agmon O, et al. An outbreak of new, nonmultidrug-resistant, methicillin-resistant *Staphylococcus aureus* strain (sccmec type iiia variant-1) in the neonatal intensive care unit transmitted by a staff member. The Pediatric infectious disease journal. 2006;25(6):557-9.
70. Park C, Pearce J. A major outbreak of methicillin resistant *Staphylococcus aureus* among patients and staff at Johannesburg Hospital during 1986-1987. Nursing RSA. 1989;4(6):37-9.
71. Ramsing BG, Arpi M, Andersen EA, Knabe N, Mogensen D, Buhl D, et al. First outbreak with MRSA in a Danish neonatal intensive care unit: risk factors and control procedures. PloS one. 2013;8(6):e66904.
72. PHLS. First Report of the Department of Health's Mandatory MRSA bacteraemia surveillance scheme in acute NHS Trusts in England: April to September 2001. CDR Weekly. 2002;12:1-15.
73. Towards cleaner hospitals and lower rates of infection: a summary of action. Department of Health, London, UK. 2004.
74. MRSA/HCAI/Cleanliness Improvement Programme. Department of Health, London, UK. 2005.
75. Clean hands help to save lives. Patient safety alert 4. In: Agency NPS, editor. London2004.
76. Bloodborne MRSA infection rates to be halved by 2008 - Press notice. Department of Health, London, UK. 2004.

77. Saving Lives - a delivering programme to reduce healthcare associated infection, including MRSA. Department of Health, London, UK. 2005.
78. Chief Medical Officer. Hospital pharmacy initiative for promoting prudent use of antimicrobials in hospitals. Department of Health, London, UK. 2003.
79. Pemberton M. No wonder dress codes have doctors up in (bare) arms. The Telegraph. 9 June 2014.
80. Jacob G. Uniforms and Workwear: An evidence base for developing local policy. Department of Health, London, UK. 2007.
81. Fuller C, Robotham J, Savage J, Hopkins S, Deeny SR, Stone S, et al. The national one week prevalence audit of universal meticillin-resistant *Staphylococcus aureus* (MRSA) admission screening 2012. PloS one. 2013;8(9):e74219.
82. Cairns S, Packer S, Reilly J, Leanord A. Targeted MRSA screening can be as effective as universal screening. BMJ. 2014;349:g5075.
83. Dueden B, Fry C, Johnson AP, Wilcox MH. The Control of Methicillin-Resistant *Staphylococcus aureus* Blood Stream Infections in England. Open Forum Infect Dis. 2015;2(2).
84. Newitt S, Myles PR, Birkin JA, Maskell V, Slack RC, Nguyen-Van-Tam JS, et al. Impact of infection control interventions on rates of *Staphylococcus aureus* bacteraemia in National Health Service acute hospitals, East Midlands, UK, using interrupted time-series analysis. The Journal of hospital infection. 2015;90(1):28-37.
85. Stone SP, Fuller C, Savage J, Cookson B, Hayward A, Cooper B, et al. Evaluation of the national Cleanyourhands campaign to reduce *Staphylococcus aureus* bacteraemia and Clostridium difficile infection in hospitals in England and Wales by improved hand hygiene: four year, prospective, ecological, interrupted time series study. BMJ. 2012;344:e3005.
86. Sroka S, Gastmeier P, Meyer E. Impact of alcohol hand-rub use on meticillin-resistant *Staphylococcus aureus*: an analysis of the literature. The Journal of hospital infection. 2010;74(3):204-11.
87. Abraham EPC, E.; Fletcher, C. M.; Florey, H. W.; Gardner, A. D.; Heatley, N. G.; Jennings, M. A. Further Observations on Penicillin. Lancet 1941(2):177-89.
88. Kirby WM. Extraction of a Highly Potent Penicillin Inactivator from Penicillin Resistant Staphylococci. Science. 1944;99(2579):452-3.
89. Jevons MP. "Celbenin" resistant Staphylococci. BMJ. 1961;1(5219):124.
90. Barber M. Methicillin-resistant Staphylococci. Journal of clinical pathology. 1961;14:385-93.
91. Kim SH, Kim KH, Kim HB, Kim NJ, Kim EC, Oh MD, et al. Outcome of vancomycin treatment in patients with methicillin-susceptible *Staphylococcus aureus* bacteremia. Antimicrobial agents and chemotherapy. 2008;52(1):192-7.
92. Schweizer ML, Furuno JP, Harris AD, Johnson JK, Shardell MD, McGregor JC, et al. Comparative effectiveness of nafcillin or cefazolin versus vancomycin in methicillin-susceptible *Staphylococcus aureus* bacteremia. BMC infectious diseases. 2011;11:279.
93. FDA Drug Safety Communication: Eosinophilic pneumonai associated with the use of Cubicin (daptomycin). 2010.
94. Holland TL, Arnold C, Fowler VG, Jr. Clinical management of *Staphylococcus aureus* bacteremia: a review. JAMA. 2014;312(13):1330-41.
95. Usery JB, Vo NH, Finch CK, Cleveland KO, Gelfand MS, Self TH. Evaluation of the treatment of methicillin-resistant *Staphylococcus aureus* bacteremia. Am J Med Sci. 2015;349(1):36-41.
96. Dellinger RP, Carlet JM, Masur H, Gerlach H, Calandra T, Cohen J, et al. Surviving Sepsis Campaign guidelines for management of severe sepsis and septic shock. Intensive care medicine. 2004;30(4):536-55.
97. Funk DJ, Kumar A. Antimicrobial therapy for life-threatening infections: speed is life. Critical care clinics. 2011;27(1):53-76.

98. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America. 1977;74(12):5463-7.
99. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome research. 2011;21(6):936-9.
100. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research. 2008;18(5):821-9.
101. Kilianski A, Haas JL, Corriveau EJ, Liem AT, Willis KL, Kadavy DR, et al. Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. GigaScience. 2015;4:12.
102. Monecke S, Coombs G, Shore AC, Coleman DC, Akpaka P, Borg M, et al. A field guide to pandemic, epidemic and sporadic clones of methicillin-resistant *Staphylococcus aureus*. PloS one. 2011;6(4):e17936.
103. Votintseva AA, Miller RR, Fung R, Knox K, Godwin H, Peto TE, et al. Multiple-strain colonization in nasal carriers of *Staphylococcus aureus*. Journal of clinical microbiology. 2014;52(4):1192-200.
104. Lim MS, Marshall CL, Spelman D. Carriage of multiple subtypes of methicillin-resistant *Staphylococcus aureus* by intensive care unit patients. Infection control and hospital epidemiology. 2006;27(10):1063-7.
105. Mongkolrattanothai K, Gray BM, Mankin P, Stanfill AB, Pearl RH, Wallace LJ, et al. Simultaneous carriage of multiple genotypes of *Staphylococcus aureus* in children. Journal of medical microbiology. 2011;60(Pt 3):317-22.
106. Bloemendaal AL, Fluit AC, Jansen WT, Vriens MR, Ferry T, Amorim JM, et al. Colonization with multiple *Staphylococcus aureus* strains among patients in European intensive care units. Infection control and hospital epidemiology. 2009;30(9):918-20.
107. Golubchik T, Batty EM, Miller RR, Farr H, Young BC, Larner-Svensson H, et al. Within-host evolution of *Staphylococcus aureus* during asymptomatic carriage. PloS one. 2013;8(5):e61319.
108. Price JR, Golubchik T, Cole K, Wilson DJ, Crook DW, Thwaites GE, et al. Whole-genome sequencing shows that patient-to-patient transmission rarely accounts for acquisition of *Staphylococcus aureus* in an intensive care unit. Clinical infectious diseases. 2014;58(5):609-18.
109. Sherertz RJ, Reagan DR, Hampton KD, Robertson KL, Streed SA, Hoen HM, et al. A cloud adult: the *Staphylococcus aureus*-virus interaction revisited. Annals of internal medicine. 1996;124(6):539-47.
110. Sherertz RJ, Bassetti S, Bassetti-Wyss B. "Cloud" health-care workers. Emerging infectious diseases. 2001;7(2):241-4.
111. Worby CJ, Lipsitch M, Hanage WP. Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. PLoS computational biology. 2014;10(3):e1003549.
112. Paterson GK, Harrison EM, Murray GG, Welch JJ, Warland JH, Holden MT, et al. Capturing the cloud of diversity reveals complexity and heterogeneity of MRSA carriage, infection and transmission. Nature communications. 2015;6:6560.
113. Koser CU, Holden MT, Ellington MJ, Cartwright EJ, Brown NM, Ogilvy-Stuart AL, et al. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. The New England journal of medicine. 2012;366(24):2267-75.
114. Harris SR, Cartwright EJ, Torok ME, Holden MT, Brown NM, Ogilvy-Stuart AL, et al. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. The Lancet Infectious diseases. 2013;13(2):130-6.
115. Miller RR, Walker AS, Godwin H, Fung R, Votintseva A, Bowden R, et al. Dynamics of acquisition and loss of carriage of *Staphylococcus aureus* strains in the community: the effect of clonal complex. The Journal of infection. 2014;68(5):426-39.

116. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
117. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic biology. 2003;52(5):696-704.
118. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC bioinformatics. 2009;10:421.
119. Gordon NC, Price JR, Cole K, Everitt R, Morgan M, Finney J, et al. Prediction of *Staphylococcus aureus* antimicrobial resistance by whole-genome sequencing. Journal of clinical microbiology. 2014;52(4):1182-91.
120. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. PHAST: a fast phage search tool. Nucleic acids research. 2011;39(Web Server issue):W347-52.
121. Carattoli A, Zankari E, Garcia-Fernandez A, Voldby Larsen M, Lund O, Villa L, et al. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrobial agents and chemotherapy. 2014;58(7):3895-903.
122. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular biology and evolution. 2012;29(8):1969-73.
123. Strommenger B, Kettlitz C, Weniger T, Harmsen D, Friedrich AW, Witte W. Assignment of *Staphylococcus* isolates to groups by spa typing, SmaI macrorestriction analysis, and multilocus sequence typing. Journal of clinical microbiology. 2006;44(7):2533-40.
124. Mellmann A, Weniger T, Berssenbrugge C, Rothganger J, Sammeth M, Stoye J, et al. Based Upon Repeat Pattern (BURP): an algorithm to characterize the long-term evolution of *Staphylococcus aureus* populations based on spa polymorphisms. BMC microbiology. 2007;7:98.
125. Chambers HF, Deleo FR. Waves of resistance: *Staphylococcus aureus* in the antibiotic era. Nature reviews Microbiology. 2009;7(9):629-41.
126. Kulhankova K, King J, Salgado-Pabon W. Staphylococcal toxic shock syndrome: superantigen-mediated enhancement of endotoxin shock and adaptive immune suppression. Immunologic research. 2014;59(1-3):182-7.
127. 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. The Lancet Infectious diseases. 2013;13(1):43-54.
128. Verkaik NJ, Benard M, Boelens HA, de Vogel CP, Nouwen JL, Verbrugh HA, et al. Immune evasion cluster-positive bacteriophages are highly prevalent among human *Staphylococcus aureus* strains, but they are not essential in the first stages of nasal colonization. Clinical microbiology and infection. 2011;17(3):343-8.
129. De Maio N, Wu CH, O'Reilly KM, Wilson D. New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. PLoS genetics. 2015;11(8):e1005421.
130. Young BC, Golubchik T, Batty EM, Fung R, Larner-Svensson H, Votintseva AA, et al. Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. Proceedings of the National Academy of Sciences of the United States of America. 2012;109(12):4550-5.
131. Rocha EP, Smith JM, Hurst LD, Holden MT, Cooper JE, Smith NH, et al. Comparisons of dN/dS are time dependent for closely related bacterial genomes. Journal of theoretical biology. 2006;239(2):226-35.
132. Tong SY, Holden MT, Nickerson EK, Cooper BS, Koser CU, Cori A, et al. Genome sequencing defines phylogeny and spread of methicillin-resistant *Staphylococcus aureus* in a high transmission setting. Genome research. 2015;25(1):111-8.
133. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science. 2010;327(5964):469-74.
134. Argudin MA, Mendoza MC, Mendez FJ, Martin MC, Guerra B, Rodicio MR. Clonal complexes and diversity of exotoxin gene profiles in methicillin-resistant and methicillin-susceptible

Staphylococcus aureus isolates from patients in a Spanish hospital. Journal of clinical microbiology. 2009;47(7):2097-105.
135. Imani Fooladi AA, Ashrafi E, Tazandareh SG, Koosha RZ, Rad HS, Amin M, et al. The distribution of pathogenic and toxigenic genes among MRSA and MSSA clinical isolates. Microbial pathogenesis. 2015;81:60-6.
136. Durand G, Bes M, Meugnier H, Enright MC, Forey F, Liassine N, et al. Detection of new methicillin-resistant *Staphylococcus aureus* clones containing the toxic shock syndrome toxin 1 gene responsible for hospital- and community-acquired infections in France. Journal of clinical microbiology. 2006;44(3):847-53.
137. Parsonnet J, Goering RV, Hansmann MA, Jones MB, Ohtagaki K, Davis CC, et al. Prevalence of toxic shock syndrome toxin 1 (TSST-1)-producing strains of *Staphylococcus aureus* and antibody to TSST-1 among healthy Japanese women. Journal of clinical microbiology. 2008;46(8):2731-8.
138. Park JY, Kim JS, Woo H. Prevalence of antibody to toxic shock syndrome toxin-1 in burn patients. Annals of laboratory medicine. 2015;35(1):89-93.
139. Nhan TX, Leclercq R, Cattoir V. Prevalence of toxin genes in consecutive clinical isolates of *Staphylococcus aureus* and clinical impact. European journal of clinical microbiology & infectious diseases. 2011;30(6):719-25.
140. Deurenberg RH, Rijnders MI, Sebastian S, Welling MA, Beisser PS, Stobberingh EE. The *Staphylococcus aureus* lineage-specific markers collagen adhesin and toxic shock syndrome toxin 1 distinguish multilocus sequence typing clonal complexes within spa clonal complexes. Diagnostic microbiology and infectious disease. 2009;65(2):116-22.
141. Nienaber JJ, Sharma Kuinkel BK, Clarke-Pearson M, Lamlertthon S, Park L, Rude TH, et al. Methicillin-susceptible *Staphylococcus aureus* endocarditis isolates are associated with clonal complex 30 genotype and a distinct repertoire of enterotoxins and adhesins. The Journal of infectious diseases. 2011;204(5):704-13.
142. Otokunefor K, Sloan T, Kearns AM, James R. Molecular characterization and panton-valentine leucocidin typing of community-acquired methicillin-sensitive *Staphylococcus aureus* clinical isolates. Journal of clinical microbiology. 2012;50(9):3069-72.
143. Shallcross LJ, Williams K, Hopkins S, Aldridge RW, Johnson AM, Hayward AC. Panton-Valentine leukocidin associated staphylococcal disease: a cross-sectional study at a London hospital, England. Clinical microbiology and infection. 2010;16(11):1644-8.
144. Stanczak-Mrozek KI, Manne A, Knight GM, Gould K, Witney AA, Lindsay JA. Within-host diversity of MRSA antimicrobial resistances. The Journal of antimicrobial chemotherapy. 2015.
145. McCarthy AJ, Loeffler A, Witney AA, Gould KA, Lloyd DH, Lindsay JA. Extensive horizontal gene transfer during *Staphylococcus aureus* co-colonization in vivo. Genome biology and evolution. 2014;6(10):2697-708.
146. Eyre DW, Golubchik T, Gordon NC, Bowden R, Piazza P, Batty EM, et al. A pilot study of rapid benchtop sequencing of *Staphylococcus aureus* and Clostridium difficile for outbreak detection and surveillance. BMJ open. 2012;2(3).
147. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLoS computational biology. 2014;10(1):e1003457.
148. Orendi JM, Coetzee N, Ellington MJ, Boakes E, Cookson BD, Hardy KJ, et al. Community and nosocomial transmission of Panton-Valentine leucocidin-positive community-associated meticillin-resistant *Staphylococcus aureus*: implications for healthcare. The Journal of hospital infection. 2010;75(4):258-64.
149. Ellington MJ, Hope R, Livermore DM, Kearns AM, Henderson K, Cookson BD, et al. Decline of EMRSA-16 amongst methicillin-resistant *Staphylococcus aureus* causing bacteraemias in the UK between 2001 and 2007. The Journal of antimicrobial chemotherapy. 2010;65(3):446-8.
150. Lozano C, Garcia-Migura L, Aspiroz C, Zarazaga M, Torres C, Aarestrup FM. Expansion of a plasmid classification system for Gram-positive bacteria and determination of the diversity of

plasmids in *Staphylococcus aureus* strains of human, animal, and food origins. Applied and environmental microbiology. 2012;78(16):5948-55.
151. Reingold A. Outbreak investigation--a perspective. Epidemiological bulletin. 2000;21(2):1-7.
152. Screening for Meticillin-resistant *Staphylococcus aureus*(MRSA) colonisation: a strategy for NHS trusts. Department of Health, London, UK. 2006.
153. Surveillance of Surgical Site Infections in NHS Hospitals in England 2013/2014. Public Health England, 2014.
154. Annual Epidemiological Commentary: Mandatory MRSA, MSSA and *E. coli* bacteraemia and *C. difficile* infection data, 2013/14. Public Health England, 2014.
155. Blanc DS, Struelens MJ, Deplano A, De Ryck R, Hauser PM, Petignat C, et al. Epidemiological validation of pulsed-field gel electrophoresis patterns for methicillin-resistant *Staphylococcus aureus*. Journal of clinical microbiology. 2001;39(10):3442-5.
156. Harrison EM, Paterson GK, Holden MT, Larsen J, Stegger M, Larsen AR, et al. Whole genome sequencing identifies zoonotic transmission of MRSA isolates with the novel mecA homologue mecC. EMBO molecular medicine. 2013;5(4):509-15.
157. Garcia-Alvarez L, Holden MT, Lindsay H, Webb CR, Brown DF, Curran MD, et al. Meticillin-resistant *Staphylococcus aureus* with a novel mecA homologue in human and bovine populations in the UK and Denmark: a descriptive study. The Lancet Infectious diseases. 2011;11(8):595-603.
158. Altman DR, Sebra R, Hand J, Attie O, Deikus G, Carpini KW, et al. Transmission of methicillin-resistant *Staphylococcus aureus* via deceased donor liver transplantation confirmed by whole genome sequencing. American journal of transplantation. 2014;14(11):2640-4.
159. Wendt JM, Kaul D, Limbago BM, Ramesh M, Cohle S, Denison AM, et al. Transmission of methicillin-resistant *Staphylococcus aureus* infection through solid organ transplantation: confirmation via whole genome sequencing. American journal of transplantation. 2014;14(11):2633-9.
160. 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;5(5):e01692-14.
161. Uhlemann AC, Dordel J, Knox JR, Raven KE, Parkhill J, Holden MT, et al. Molecular tracing of the emergence, diversification, and transmission of *S. aureus* sequence type 8 in a New York community. Proceedings of the National Academy of Sciences of the United States of America. 2014;111(18):6738-43.
162. Alam MT, Read TD, Petit RA, 3rd, Boyle-Vavra S, Miller LG, Eells SJ, et al. Transmission and microevolution of USA300 MRSA in U.S. households: evidence from whole-genome sequencing. mBio. 2015;6(2):e00054.
163. Azarian T, Cook RL, Johnson JA, Guzman N, McCarter YS, Gomez N, et al. Whole-Genome Sequencing for Outbreak Investigations of Methicillin-Resistant *Staphylococcus aureus* in the Neonatal Intensive Care Unit: Time for Routine Practice? Infection control and hospital epidemiology. 2015:1-9.
164. Anderson DJ, Harris SR, Godofsky E, Toriscelli T, Rude TH, Elder K, et al. Whole Genome Sequencing of a Methicillin-Resistant *Staphylococcus aureus* Pseudo-Outbreak in a Professional Football Team. Open Forum Infect Dis. 2014;1(3):ofu096.
165. Sutton SS, Stacy JJ, Mensch J, Torres-McGehee T, Bennett CL. Tackling community-acquired methicillin-resistant *Staphylococcus aureus* in collegiate football players following implementation of an anti-MRSA programme. British journal of sports medicine. 2014;48(4):284-5.
166. Lear A, McCord G, Peiffer J, Watkins RR, Parikh A, Warrington S. Incidence of *Staphylococcus aureus* nasal colonization and soft tissue infection among high school football players. Journal of the American Board of Family Medicine : JABFM. 2011;24(4):429-35.
167. Garza D, Sungar G, Johnston T, Rolston B, Ferguson JD, Matheson GO. Ineffectiveness of surveillance to control community-acquired methicillin-resistant *Staphylococcus aureus* in a professional football team. Clinical journal of sport medicine. 2009;19(6):498-501.

168. Ginsburg S. Former kicker sues NFL's Buccaneers over MRSA infection. Reuters. Apr 7, 2015.
169. Romano-Bertrand S, Filleron A, Mesnage R, Lotthe A, Didelot MN, Burgel L, et al. *Staphylococcus aureus* in a neonatal care center: methicillin-susceptible strains should be a main concern. Antimicrobial resistance and infection control. 2014;3:21.
170. van der Mee-Marquet N, Francois P, Domelier AS, Arnault L, Girard N, Schrenzel J, et al. Variable-number tandem repeat analysis and multilocus sequence typing data confirm the epidemiological changes observed with *Staphylococcus aureus* strains isolated from bloodstream infections. Journal of clinical microbiology. 2009;47(9):2863-71.
171. PVL sub-group of the Steering Group on Healthcare Associated Infection nE. Guidance on the diagnosis and management of PVL-associated *Staphylococcus aureus* infections (PVL-SA) in England. Public Health England, 2008.
172. Uhlemann AC, Kennedy AD, Martens C, Porcella SF, Deleo FR, Lowy FD. Toward an understanding of the evolution of *Staphylococcus aureus* strain USA300 during colonization in community households. Genome biology and evolution. 2012;4(12):1275-85.
173. Reilly JS, Stewart S, Christie P, Allardice G, Smith A, Masterton R, et al. Universal screening for meticillin-resistant *Staphylococcus aureus*: interim results from the NHS Scotland pathfinder project. The Journal of hospital infection. 2010;74(1):35-41.
174. Vogel V, Falquet L, Calderon-Copete SP, Basset P, Blanc DS. Short term evolution of a highly transmissible methicillin-resistant *Staphylococcus aureus* clone (ST228) in a tertiary care hospital. PloS one. 2012;7(6):e38969.
175. Basset P, Senn L, Prod'hom G, Bille J, Francioli P, Zanetti G, et al. Usefulness of double locus sequence typing (DLST) for regional and international epidemiological surveillance of methicilin-resistant *Staphylococcus aureus*. Clinical microbiology and infection. 2010;16(8):1289-96.
176. Antonio M, McFerran N, Pallen MJ. Mutations affecting the Rossman fold of isoleucyl-tRNA synthetase are correlated with low-level mupirocin resistance in *Staphylococcus aureus*. Antimicrobial agents and chemotherapy. 2002;46(2):438-42.
177. Williamson DA, Coombs GW, Nimmo GR. *Staphylococcus aureus* 'Down Under': contemporary epidemiology of *S. aureus* in Australia, New Zealand, and the South West Pacific. Clinical microbiology and infection. 2014;20(7):597-604.
178. Ali H, Nash JQ, Kearns AM, Pichon B, Vasu V, Nixon Z, et al. Outbreak of a South West Pacific clone Panton-Valentine leucocidin-positive meticillin-resistant *Staphylococcus aureus* infection in a UK neonatal intensive care unit. The Journal of hospital infection. 2012;80(4):293-8.
179. Nelson MU, Gallagher PG. Methicillin-resistant *Staphylococcus aureus* in the neonatal intensive care unit. Semin Perinatol. 2012;36(6):424-30.
180. Nubel U, Nachtnebel M, Falkenhorst G, Benzler J, Hecht J, Kube M, et al. MRSA transmission on a neonatal intensive care unit: epidemiological and genome-based phylogenetic analyses. PloS one. 2013;8(1):e54898.
181. Ferry T, Bes M, Dauwalder O, Meugnier H, Lina G, Forey F, et al. Toxin gene content of the Lyon methicillin-resistant *Staphylococcus aureus* clone compared with that of other pandemic clones. Journal of clinical microbiology. 2006;44(7):2642-4.
182. Vergnano S. Decolonization and decontamination: what's their role in infection control? Current opinion in infectious diseases. 2015;28(3):207-14.
183. Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, et al. Rapid antibiotic resistance predictions from genome sequence data for *S. aureus* and *M. tuberculosis*. Nature communications. 2015. 6:10063.
184. Wyllie DH, Walker AS, Miller R, Moore C, Williamson SR, Schlackow I, et al. Decline of meticillin-resistant *Staphylococcus aureus* in Oxfordshire hospitals is strain-specific and preceded infection-control intensification. BMJ open. 2011;1(1):e000160.
185. Cimolai N. The role of healthcare personnel in the maintenance and spread of methicillin-resistant *Staphylococcus aureus*. Journal of infection and public health. 2008;1(2):78-100.

186. Huang SS, Platt R. Risk of methicillin-resistant *Staphylococcus aureus* infection after previous infection or colonization. Clinical infectious diseases. 2003;36(3):281-5.
187. Davis KA, Stewart JJ, Crouch HK, Florez CE, Hospenthal DR. Methicillin-resistant *Staphylococcus aureus* (MRSA) nares colonization at hospital admission and its effect on subsequent MRSA infection. Clinical infectious diseases. 2004;39(6):776-82.
188. Patel M, Weinheimer JD, Waites KB, Baddley JW. Active surveillance to determine the impact of methicillin-resistant *Staphylococcus aureus* colonization on patients in intensive care units of a Veterans Affairs Medical Center. Infection control and hospital epidemiology. 2008;29(6):503-9.
189. Datta R, Huang SS. Risk of infection and death due to methicillin-resistant *Staphylococcus aureus* in long-term carriers. Clinical infectious diseases. 2008;47(2):176-81.
190. Andrews JM. The development of the BSAC standardized method of disc diffusion testing. The Journal of antimicrobial chemotherapy. 2001;48 Suppl 1:29-42.
191. Wheat PF. History and development of antimicrobial susceptibility testing methodology. The Journal of antimicrobial chemotherapy. 2001;48 Suppl 1:1-4.
192. Fleming A. A Comparison of the Activities of Antiseptics on Bacteria and on Leucocytes. Proceedings of the Royal Society of London, Series B. 1924;96:171-80.
193. Bondi A, Jr., Spaulding EH, et al. A routine method for the rapid determination of susceptibility to penicillin and other antibiotics. Am J Med Sci. 1947;213(2):221-5.
194. Aldridge C, Jones PW, Gibson S, Lanham J, Meyer M, Vannest R, et al. Automated microbiological detection/identification system. Journal of clinical microbiology. 1977;6(4):406-13.
195. Juretschko S, Labombardi VJ, Lerner SA, Schreckenberger PC. Accuracies of beta-lactam susceptibility test results for *Pseudomonas aeruginosa* with four automated systems (BD Phoenix, MicroScan WalkAway, Vitek, and Vitek 2). Journal of clinical microbiology. 2007;45(4):1339-42.
196. Nadarajah R, Post LR, Liu C, Miller SA, Sahm DF, Brooks GF. Detection of vancomycin-intermediate *Staphylococcus aureus* with the updated Trek-Sensititre System and the MicroScan System. Comparison with results from the conventional Etest and CLSI standardized MIC methods. American journal of clinical pathology. 2010;133(6):844-8.
197. Bergeron MG, Ouellette M. Preventing antibiotic resistance through rapid genotypic identification of bacteria and of their antibiotic resistance genes in the clinical microbiology laboratory. Journal of clinical microbiology. 1998;36(8):2169-72.
198. Berger-Bachi B. Resistance mechanisms of gram-positive bacteria. International journal of medical microbiology. 2002;292(1):27-35.
199. Holden MT, Feil EJ, Lindsay JA, Peacock SJ, Day NP, Enright MC, et al. Complete genomes of two clinical *Staphylococcus aureus* strains: evidence for the rapid evolution of virulence and drug resistance. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(26):9786-91.
200. Price J, Baker G, Heath I, Walker-Bone K, Cubbon M, Curtis S, et al. Clinical and Microbiological Determinants of Outcome in *Staphylococcus aureus* Bacteraemia. International journal of microbiology. 2010;2010:654858.
201. Gosden PE, Andrews JM, Bowker KE, Holt HA, MacGowan AP, Reeves DS, et al. Comparison of the modified Stokes' method of susceptibility testing with results obtained using MIC methods and British Society of Antimicrobial Chemotherapy breakpoints. The Journal of antimicrobial chemotherapy. 1998;42(2):161-9.
202. Howe RA, Andrews JM, Testing BWPoS. BSAC standardized disc susceptibility testing method (version 11). The Journal of antimicrobial chemotherapy. 2012;67(12):2783-4.
203. Yilmaz G, Aydin K, Iskender S, Caylan R, Koksal I. Detection and prevalence of inducible clindamycin resistance in staphylococci. Journal of medical microbiology. 2007;56(Pt 3):342-5.
204. Jorgensen JH, Ferraro MJ. Antimicrobial susceptibility testing: a review of general principles and contemporary practices. Clinical infectious diseases. 2009;49(11):1749-55.

205. Pitkala A, Salmikivi L, Bredbacka P, Myllyniemi AL, Koskinen MT. Comparison of tests for detection of beta-lactamase-producing staphylococci. Journal of clinical microbiology. 2007;45(6):2031-3.
206. Kolbert CP, Connolly JE, Lee MJ, Persing DH. Detection of the Staphylococcal mecA gene by chemiluminescent DNA hybridization. Journal of clinical microbiology. 1995;33(8):2179-82.
207. Blanc DS, Petignat C, Moreillon P, Entenza JM, Eisenring M, Kleiber H, et al. Unusual spread of a penicillin-susceptible methicillin-resistant *Staphylococcus aureus* clone in a geographic area of low incidence. Clinical infectious diseases. 1999;29(6):1512-8.
208. Tanaka M, Wang T, Onodera Y, Uchida Y, Sato K. Mechanism of quinolone resistance in *Staphylococcus aureus*. Journal of infection and chemotherapy. 2000;6(3):131-9.
209. Kaatz GW, Seo SM. Mechanisms of fluoroquinolone resistance in genetically related strains of *Staphylococcus aureus*. Antimicrobial agents and chemotherapy. 1997;41(12):2733-7.
210. Qin X, Poon B, Kwong J, Niles D, Schmidt BZ, Rajagopal L, et al. Two paediatric cases of skin and soft-tissue infections due to clindamycin-resistant *Staphylococcus aureus* carrying a plasmid-encoded vga(A) allelic variant for a putative efflux pump. International journal of antimicrobial agents. 2011;38(1):81-3.
211. Vidaillac C, Gardete S, Tewhey R, Sakoulas G, Kaatz GW, Rose WE, et al. Alternative mutational pathways to intermediate resistance to vancomycin in methicillin-resistant *Staphylococcus aureus*. The Journal of infectious diseases. 2013;208(1):67-74.
212. McLaws FB, Larsen AR, Skov RL, Chopra I, O'Neill AJ. Distribution of fusidic acid resistance determinants in methicillin-resistant *Staphylococcus aureus*. Antimicrobial agents and chemotherapy. 2011;55(3):1173-6.
213. Frey KM, Viswanathan K, Wright DL, Anderson AC. Prospective screening of novel antibacterial inhibitors of dihydrofolate reductase for mutational resistance. Antimicrobial agents and chemotherapy. 2012;56(7):3556-62.
214. Guidance for Industry and FDA. Class II Special Controls Guidance Document: Antimicrobial Susceptibility Test (AST) Systems. US Department of Health and Human Services. 2009.
215. Olsen JE, Christensen H, Aarestrup FM. Diversity and evolution of *blaZ* from *Staphylococcus aureus* and coagulase-negative staphylococci. The Journal of antimicrobial chemotherapy. 2006;57(3):450-60.
216. Haveri M, Suominen S, Rantala L, Honkanen-Buzalski T, Pyorala S. Comparison of phenotypic and genotypic detection of penicillin G resistance of *Staphylococcus aureus* isolated from bovine intramammary infection. Veterinary microbiology. 2005;106(1-2):97-102.
217. El Feghaly RE, Stamm JE, Fritz SA, Burnham CA. Presence of the bla(Z) beta-lactamase gene in isolates of *Staphylococcus aureus* that appear penicillin susceptible by conventional phenotypic methods. Diagnostic microbiology and infectious disease. 2012;74(4):388-93.
218. Tasara T, Cernela N, Stephan R. Function impairing mutations in *blaZ* and blaR genes of penicillin susceptible *Staphylococcus aureus* strains isolated from bovine mastitis. Schweizer Archiv fur Tierheilkunde. 2013;155(6):359-63.
219. Nissen JL, Skov R, Knudsen JD, Ostergaard C, Schonheyder HC, Frimodt-Moller N, et al. Effectiveness of penicillin, dicloxacillin and cefuroxime for penicillin-susceptible *Staphylococcus aureus* bacteraemia: a retrospective, propensity-score-adjusted case-control and cohort analysis. The Journal of antimicrobial chemotherapy. 2013;68(8):1894-900.
220. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nature methods. 2015;12(8):733-5.
221. Lawes T, Edwards B, Lopez-Lozano JM, Gould I. Trends in *Staphylococcus aureus* bacteraemia and impacts of infection control practices including universal MRSA admission screening in a hospital in Scotland, 2006-2010: retrospective cohort study and time-series intervention analysis. BMJ open. 2012;2(3).
222. Chalfine A, Kitzis MD, Bezie Y, Benali A, Perniceni L, Nguyen JC, et al. Ten-year decrease of acquired methicillin-resistant *Staphylococcus aureus* (MRSA) bacteremia at a single institution: the

result of a multifaceted program combining cross-transmission prevention and antimicrobial stewardship. Antimicrobial resistance and infection control. 2012;1(1):18.
223. Mitchell BG, Collignon PJ, McCann R, Wilkinson IJ, Wells A. A major reduction in hospital-onset *Staphylococcus aureus* bacteremia in Australia-12 years of progress: an observational study. Clinical infectious diseases. 2014;59(7):969-75.
224. Awad SS, Palacio CH, Subramanian A, Byers PA, Abraham P, Lewis DA, et al. Implementation of a methicillin-resistant *Staphylococcus aureus* (MRSA) prevention bundle results in decreased MRSA surgical site infections. Am J Surg. 2009;198(5):607-10.
225. Widmer AF, Lakatos B, Frei R. Strict infection control leads to low incidence of methicillin-resistant *Staphylococcus aureus* bloodstream infection over 20 years. Infection control and hospital epidemiology. 2015;36(6):702-9.
226. Chien CY, Lin CH, Hsu RB. Care bundle to prevent methicillin-resistant *Staphylococcus aureus* sternal wound infection after off-pump coronary artery bypass. American journal of infection control. 2014;42(5):562-4.
227. Vogel M, Schmitz RP, Hagel S, Pletz MW, Gagelmann N, Scherag A, et al. Infectious disease consultation for *Staphylococcus aureus* bacteremia - A systematic review and meta-analysis. The Journal of infection. 2016;72(1):19-28.
228. Carrel M, Perencevich EN, David MZ. USA300 Methicillin-Resistant *Staphylococcus aureus*, United States, 2000-2013. Emerging infectious diseases. 2015;21(11):1973-80.
229. Mathers AJ, Stoesser N, Sheppard AE, Pankhurst L, Giess A, Yeh AJ, et al. *Klebsiella pneumoniae* carbapenemase (KPC)-producing *K. pneumoniae* at a single institution: insights into endemicity from whole-genome sequencing. Antimicrobial agents and chemotherapy. 2015;59(3):1656-63.
230. Zboromyrska Y, De la Calle C, Soto M, Sampietro-Colom L, Soriano A, Alvarez-Martinez MJ, et al. Rapid Diagnosis of Staphylococcal Catheter-Related Bacteraemia in Direct Blood Samples by Real-Time PCR. PloS one. 2016;11(8):e0161684.
231. Emonet S, Charles PG, Harbarth S, Stewardson AJ, Renzi G, Uckay I, et al. Rapid molecular determination of methicillin resistance in staphylococcal bacteraemia improves early targeted antibiotic prescribing: a randomized clinical trial. Clinical microbiology and infection. 2016.
232. Hos NJ, Wiegel P, Fischer J, Plum G. Comparative evaluation of two fully-automated real-time PCR methods for MRSA admission screening in a tertiary-care hospital. European journal of clinical microbiology & infectious diseases. 2016;35(9):1475-8.

10 Appendices

10.1 Appendix 1: hospital proposal for involvement in WGS outbreak study

Whole genome sequencing as an adjunct to outbreak investigation

Background

The UK-CRC/ HCIF funded Medical Microbiology Consortium is involved in several projects investigating the use of whole genome sequencing (WGS) in outbreak investigation (www.modmedmicro.ac.uk). For *S. aureus*, we are also working with Public Health England in service development to identify appropriate sampling frames, so that data obtained by WGS can be interpreted in the correct context.

To do this, we are initially investigating "known" outbreaks which have been referred to the PHE Staphylococcus reference laboratory, and comparing results obtained by whole genome sequencing with those obtained by the standard outbreak investigation (*spa*-type / PFGE plus epidemiology). By looking for single nucleotide variants (SNVs) across the whole genome, it is possible that whole genome sequencing may be able to distinguish between isolates to a greater degree than standard typing. We hope that this improved resolution will help to clarify the nature of an outbreak, by identifying the relative contributions of staff-patient transmission, patient-patient transmission, and community acquisition. Ultimately this will be of value in determining the most effective infection control measures.

Plan of investigation

For each outbreak, we will sequence the "definite" outbreak isolates plus approximately 20 "background" isolates.

1) **determination of expected SNV diversity in *S. aureus* in background isolates from the local population**

To put the results into the correct context we will compare the outbreak samples with apparently unrelated background samples. We will estimate background diversity by whole genome sequencing approximately 20 *S. aureus* isolates, not thought to be part of the outbreak, from the same hospital and preferably over the same time period. The type of sample will vary for each outbreak, for example consecutive MRSA screening isolates for in-hospital MRSA outbreaks, or *S. aureus* from wounds / pus samples for non-MRSA outbreaks. If it is not possible to send these isolates, we will obtain proxy background samples of the same *spa*-type from the reference laboratory.

2) **whole genome sequencing of all outbreak isolates** to determine the degree of relatedness to each other. These will be obtained from the reference laboratory.

3) **matching with epidemiological information**

To evaluate whether WGS can give better information for outbreak management than current typing, we will compare WGS results with the conclusions from the standard outbreak investigation (epidemiology plus *spa*/PFGE results). We do not anticipate that any further patient information will be required beyond that gathered as part of the routine investigation.

Results, timing and feedback

At present, the current turnaround time for sequencing and analysis is approximately 6 months. A written report or presentation will be provided once this is complete.
"Ownership" of the outbreak will remain with the referring hospital and any resulting papers / abstracts etc will be submitted only with their approval.

Ethical approval

This study is part of service development with Public Health England. No additional samples or patient information other than that obtained for routine analysis are required. Data will be anonymised as per information governance guidelines. Ethical approval for sequencing isolates from routine clinical samples and linkage to patient data without individual patient consent has been obtained from Berkshire Ethics Committee (10/H0505/83) and the UK National Information Governance Board (8-05(e)/2010). Public Health England has Patient Information Advisory Group approval to hold and analyse surveillance data for public health purposes under Section 60 of the Health and Social Care Act 2001.

Further information

Please contact

Claire Gordon

MRC clinical research fellow, Oxford University Hospitals NHS Trust

claire.gordon@ndm.ox.ac.uk

10.2 Appendix 2: example of outbreak feedback form

Interim whole genome sequencing summary for HST 5502: outbreak of MSSA scalded skin syndrome on a neonatal unit

Outbreak code (Oxford code)	HST 5502 (ESS)
Total no of isolates sequenced to date	95
Outbreak isolates	9
ST/spa	ST2434/ t346
Background isolates	All isolates from unit during outbreak period, including staff screening

Outbreak summary

The outbreak was identified in the first week of Jan 2013, with an initial cluster of 4 cases. A further 2 cases occurred in February, and there was a single case at the end of March. All cases were spa t346 and had identical PFGE profiles. 5 isolates were resistant to tetracycline and were positive for exfoliative toxins A and B by PCR. The 6th isolate was tetracycline resistant and positive for exfoliative toxin A (*eta*) only, and the 7th was tetracycline susceptible and positive for exfoliative toxin B (*etb*)

Cases were separated by time / space and there was minimal evidence for direct patient to patient transmission.

A further sporadic case was identified in Dec 2013, positive for toxins A and B.

Infection control

Extensive staff screening was carried out (65 staff members) plus enhanced screening of babies on the unit. A healthcare worker with regular contact with the unit was found to carry the outbreak strain. This individual was known to have chronic skin problems.

Repeated attempts at decolonisation of the healthcare worker (including screening / decolonisation of family / household contacts) were unsuccessful, and the worker was redeployed to an administrative role.

The HCW has had multiple isolates over time (first is C22780 / H131320194, March 2013). These have been a mixture of toxin positive / negative, with the later isolates positive for *eta* only. The most recent isolate is from July 2013. Family members have had toxin AB positive strains.

Sequencing summary

All isolates under PHE code HST5502 were retrieved from frozen stocks, and cultured overnight on CBA. DNA was extracted and sequenced using the Illumina HiSeq platform to generate 150bp reads. Reads were mapped to reference sequence MRSA 252 using using Stampy v1.0.18. Mapped reads were used to generate a distance matrix and maximum likelihood trees using PhyML. Reads were then *de novo* assembled using Velvet v1.0.18, and the assembled genomes interrogated with BLAST+ (v 2.2.28+) blastn and tblastn to identify resistance determinants.

Quality assessment

All isolates passed QC checks.

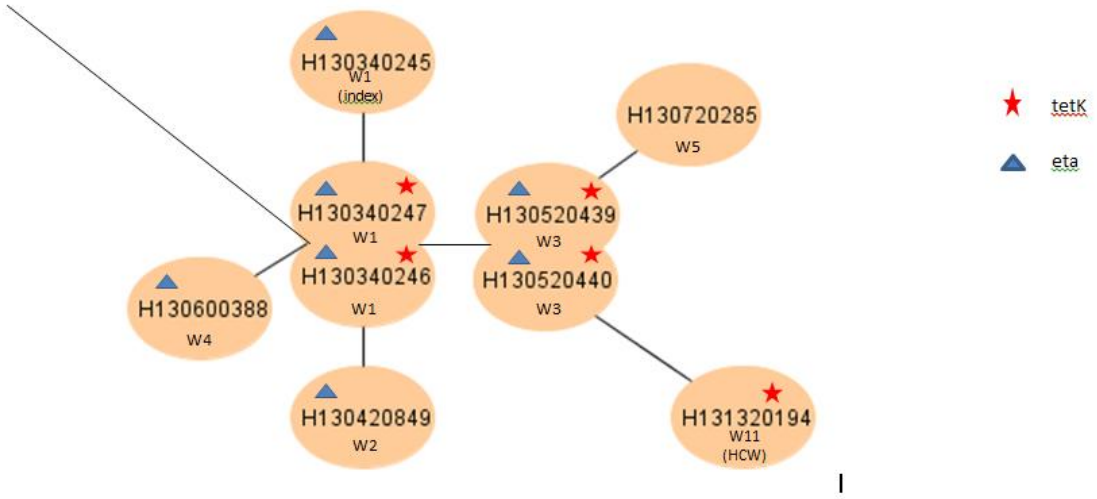
Phylogenetic analysis

9 isolates clustered very closely, confirmed as those from the outbreak, and more than 400 SNVs from their nearest neighbour. *In silico* predicted ST of these isolates was ST2434, and no other strains of this sequence type were identified in the background screening isolates.

Phylogenetic tree showing cluster of nine ST2434 isolates.

W1: week 1; HCW: healthcare worker

Asterisk / triangle indicates presence of tetracycline resistance gene / exfoliative toxin A.



There were also several incidental clusters found on sequencing of the background isolates. A phylogenetic tree showing only isolates which were within 0-5snvs of one or more samples is shown in appendix 1, along with inferred MLST sequence type. This suggests several other incidents of *S. aureus* transmission which did not cause disease or larger scale outbreaks: further epidemiological information would be useful to inform the interpretation of these clusters.

Possible interpretations

1) single, point source transmission or single introduction with rapid person to person spread: this does not fit epidemiological data as in some instances there was no overlap between cases

2) multiple introductions from e.g. colonised HCW. This would fit the epidemiology. However, the low level of variation between cases may be evidence that the HCW had recently acquired the strain, possibly from one of the early cases, and then progressed to long-term carriage (and potentially with onward transmission). This cannot be confirmed at this stage as we do not have early isolates from the healthcare worker, and only have a single colony so cannot estimate duration of carriage.

Resistance / virulence profiles

Variation was found in the presence / absence of *tetK*, which confers resistance to tetracycline, and *eta* (exfoliative toxin A). Exfoliative toxin B (*etb*) was not found in any of the isolates, differing from the previous findings. This may be due to population variation in phage integration: as we only sequenced single colonies, it is possible that the phage was not present in those particular colonies.

Planned follow-up

Further isolates have been submitted for sequencing and results are awaited, including follow-up isolates from the healthcare worker and their family members. This may help to clarify the role of the healthcare worker, as we should be able to assess the evolution / diversity of their strain which may correlate with duration of carriage.

We would also be grateful for further epidemiological information from the cluster samples, as outlined below.

Feedback

1) Does the WGS agree with the conclusion of the outbreak based on standard typing / epidemiology?

Yes / no

Comments

2) Does the WGS differ from the conclusion of the outbreak based on standard typing / epidemiology?

Yes / no

Comments

3) Does the WGS result give any additional / unexpected information?

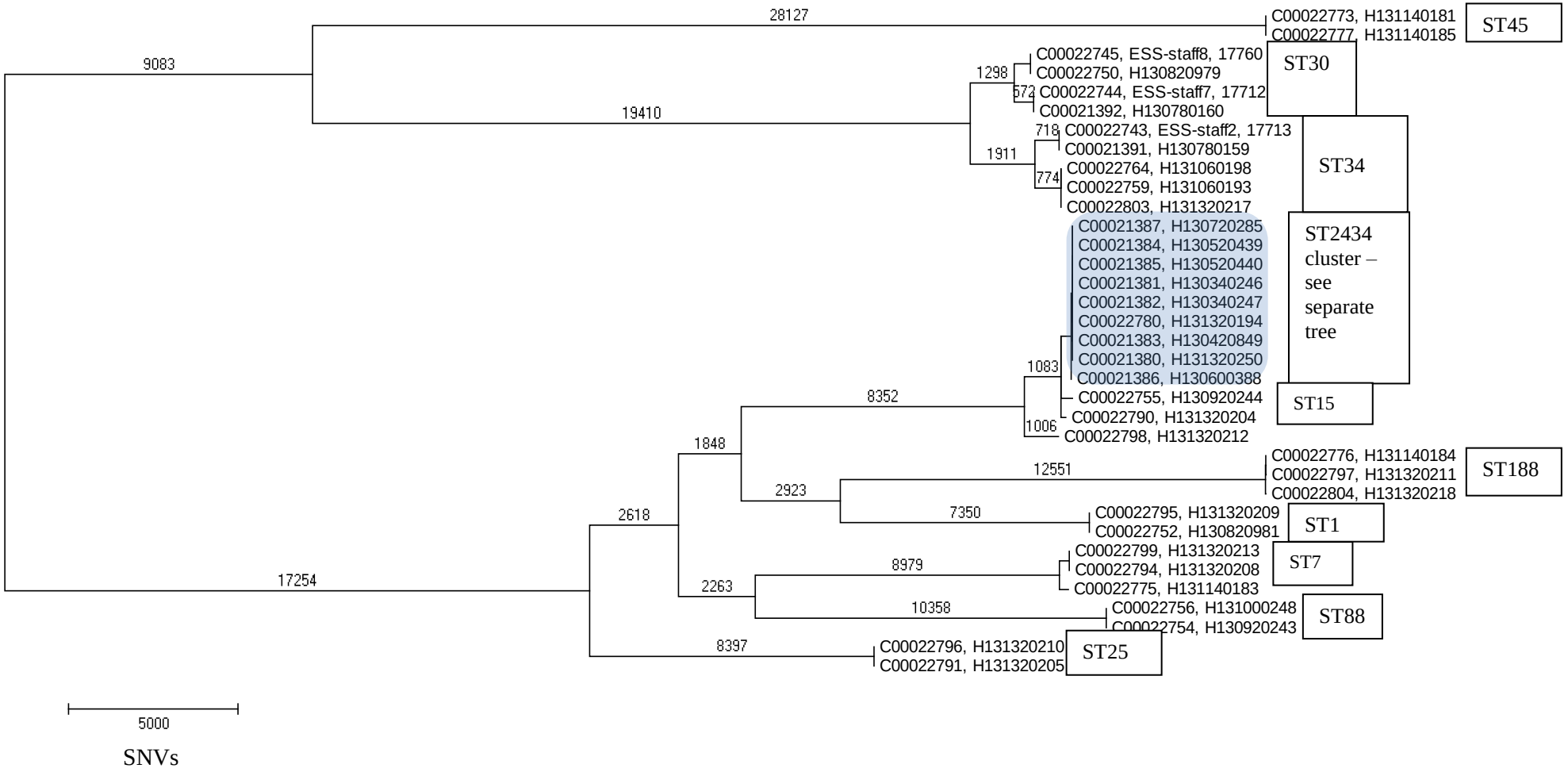
Yes / no

Please give further details

If possible, please supply details of admission / discharge dates and ward locations for the following isolates, which were shown to be closely related on whole genome sequencings

PHE reference	Oxford reference	Ward (s)	Admission date	Discharge date	Any further details? (eg bed number / bay / known close contacts)
H130340245	C00021380				
H130340246	C00021381				
H130340247	C00021382				
H130420849	C00021383				
H130520439	C00021384				

Phylogenetic tree of all isolates which are less than 5SNVs from their nearest neighbour



10.3 Appendix 3: housekeeping gene variants associated with reduced susceptibility to antimicrobials in *S. aureus*

10.3.1 Table S1: fusA variants associated with fusidic acid resistance

Amino acid substitutions	Reported MIC (mg/L)	References
B434N	>128	1
L461K	>128	2, 3, 4, 5
L461K+H457Q	>128	4
L461K+C473S	>128	4
H457Y+A67T	>128	3, 4
H457Y+A70V	>128	4
H457Y+R76C	>128	4
L461K+H457Q+A655V+V90I	>128	2
L461F+A376V +A655P+ D463G	>128	6
H457Q+L461F	128	4
V90I+H457Q+L461K	128	5
H457Y	48-96	1, 2, 3, 4, 5, 7, 8
L461S+P404L+A71V	96	3
L461S+V90I	16-64	7
H457Y +M161I	16-64	4
H457Y+S416F	64	2
H557Y +D373N	64	7
Q115L	48	1
T436I	8-48	1, 2
G452S	8-32	1, 2, 3
L456F+A376V	32	4
L461F+E444V	32	4

E444K	16	4
G556S	16	8
R659L	16	4
P404L	4-16	1, 2, 4, 7
G451V	3-16	3, 4
H457Y+R659L	16	8
H457Y+G556S	2-16	8
L461S+E233Q+V90I	16	7
A655E	12	1
G452C	12	1
P114H	12	1
R659C	12	1
R659S	12	1
H438N	8	1
P478S	8	4
R659H	8	1
A67T+P406L	8	2
A71V +P404L	8	4
F652S+Y654N	8	2
A71V + D189G+P406L	8	6
L456F	4-8	1, 2
M453I	4-8	4, 6
L461S	2-8	4, 5, 6, 7
R464C	2-8	1, 2
V90I	1-8	5, 6, 7
P406L	6	1
R464S	6	1
R464H	6	1
G617D	4	1

G664S	4	1
H457Q	4	7, 9
T385N	4	1
T656K	4	7
M651I	2	4
P404Q	2	7
Mutations where effect on MIC is not described / is less than EUCAST breakpoint		
A376V	Not described	5
F441Y	Not described	5
A70V+A160V+H457Y	Not described	5
V607I	0.19	3
V90A	Not described	5
T387I+E449K	Not described	5
D189V+L430S	Not described	5

*10.3.2 Table S2: dfr***B** *variants associated with resistance to trimethoprim*

Amino acid substitutions	Reported MIC (mg/L)	Reference
F99Y+H31N	64	10
F99Y+H150R	64	10
F99Y+L21V+N60I	64	10
H31N	20	11
F99Y	16	10, 11, 12
H150R	8	12
F99S	4	12

L41F	4	12
F99I	2.5	11

*10.3.3 Table S3: rpo*B *variants associated with resistance to rifampicin*

Amino acid substitution	Reported MIC (mg/L)	References
Q468K	>1024	13, 14
D471Y+S486L	>1024	13
S486L	1024	13, 15
H481N+A473T+A477T	256-1024	13, 14, 16
H481N+I527M	>512	13, 14, 17
Q468L	512	13
H481N+Q565R+S529L	512	13, 14
H481N+L466S	256-512	13, 14
H481Y	128-512	13, 15, 16, 17, 18
H481N+S529L	128-512	13, 14, 16
H481N+I527M	>256	17
H481D+Q468K	>256	19
H481D+I527L	>256	19
H481D+S529L	>256	19
Q468R	256	18
Q456K	128-256	16
R484H	128-256	16, 18
H481D	2-256	16, 19
A477D	128	18
D550G	128	18

H481D+A477T	128	19
ins 475H	>32	15
ins475G	16	16
M470T+D471G	12	15
S463P	8	16
D471Y	2-6	13, 15, 18
S464P	4	18
I527F	4	16
H481N	1-4	13, 14, 16
A477V	1	18

*10.3.4 Table S4: grl*A, *gyr*A *and grl*B *variants associated with resistance to quinolones*

*grl*A amino acid substitution	*gyr*A amino acid substitution	*grl*B amino acid substitution	Reported MICs (mg/L)	References
S80F+E84V	E88K		>1024	20
S80Y+E84G	S84L+S85P		512-1024	21
	S84L		>512	22, 23, 24
	S84A		>512	23
S80F+E84L	S84L+S85P		512*	25
S80Y+E84G	S84L+E88K		256-512	21
S80Y+E84K	S84L+S85P		256-512*	25
	S84L+S85P		>256	22, 24
S80Y+E84G	S84L		256	21
S80F+S108N	S84L+S85P		256*	25
S80Y+E84G	S84L+E88L		256*	25

S80Y+E84K	S84L		128-256	26, 27
S80F+E84K	S84L		32-256	27, 28, 29
S80F+A48T	S84L		32-256	20, 30
S80F+E84K	S84L+S85P		32-256	27, 29
	S84L+E88K		>128	24
S80F+E84K			>128	24
S80F+E84G	S84L		128	28
S80Y	S84L		32-128*	26, 27
S80F	S84L		8-128*	21, 25, 27, 28, 29, 30, 31, 32
S80Y	E88K		8-128*	26
E84K	S84L		8-128*	25, 27, 28, 33
S80F			0.5-128	20, 24, 27
S80F	S84L+E88K		>100	33
S80F	S84L	P451S	>64*	26, 29
S80F+E84V	S84L		>64	29
S80F+Y83N	S84L		>64	29
S80F+E84K	S84L+E88G		>64	29
	S85P		64	22
S80F	S84L+S85P		64	29
S80F	S84L+S85P	D432V	64	29
S80F	S84L+S85P	D432N or H	64	21
A116E	S84L		32	32
S80F	S84L	P585S	32	29
	E88K		16-32	27
S80F	E88G		16-32	27
S80F	E88K		8-32*	20, 26, 27, 28, 33
S80F+D432G	S84L		16	20
D79V+S80Y	S84L		16*	26

S80F+V41G+I45M	S84L		16	20
S80F+V41G	S84L	E422D	16	20
S80F	S84L+G106D		16	20
S80Y	E88G		8-12.5	27, 33
E84K	S84L	D443E+R444S	8*	25
D79V+S80F	E88K		8*	26
	S84L	R470D	4	32
S80Y			0.5-4	24
S80F+I45M		E422D	2	20

*MICs are for levofloxacin

10.3.5 Table S5: Frequency of variants across all study isolates

Gene	Variant	Frequency
*grl*A	S80F	241
	S80Y	3
	E84G	2
	E84K	2
	I45M	52
*grl*B	E422D	421
	P451S	26
	D432V	4
	P585S	1
*gyr*A	S84L	238
	S84A	1
*rpo*B	H481Y	6
	A477D	1
	S464P	1

*dfr*B	F99Y	69
	H150R	6
	L21V	1
	L41F	1
	H31N	1
*fus*A	A67T	4
	L461K	8
	L461S	3
	P406L	4
	V90I	9
	H457Y	4
	H457Q	1
	G452S	2
	L461S	2
	L461F	1
	P404Q	1
	T656S	1

10.4 Appendix references

1. Norström T, Lannergård J, Hughes D. Genetic and phenotypic identification of fusidic acid-resistant mutants with the small-colony-variant phenotype in *Staphylococcus aureus*. Antimicrob Agents Chemother 2007; 51:4438-46
2. Besier S, Ludwig A, Brade V, Wichelhaus TA. Molecular analysis of fusidic acid resistance in *Staphylococcus aureus*. Mol Microbiol 2003; 47:463-9
3. O'Neill AJ, McLaws F, Kahlmeter G et al. Genetic basis of resistance to fusidic acid in staphylococci. Antimicrob Agents Chemother 2007; 51:1737-40
4. Chen HJ, Hung WC, Tseng SP et al. Fusidic acid resistance determinants in *Staphylococcus aureus* clinical isolates. Antimicrob Agents Chemother 2010; 54:4985-91
5. Castanheira M, Watters AA, Bell JM et al. Fusidic acid resistance rates and prevalence of resistance mechanisms among *Staphylococcus* spp. isolated in North America and Australia, 2007-2008. Antimicrob Agents Chemother 2010; 54:3614-7
6. Lannergård J, Norström T, Hughes D. Genetic determinants of resistance to fusidic acid among clinical bacteremia isolates of *Staphylococcus aureus*. Antimicrob Agents Chemother 2009; 53:2059-65
7. Chen CM, Huang M, Chen HF, Ke SC, Li CR, Wang JH, Wu LT. Fusidic acid resistance among clinical isolates of methicillin-resistant *Staphylococcus aureus* in a Taiwanese hospital. BMC Microbiol 2011; 11:98
8. McLaws FB, Larsen AR, Skov RL et al. Distribution of fusidic acid resistance determinants in methicillin-resistant *Staphylococcus aureus*. Antimicrob Agents Chemother 2011; 55:1173-6
9. Castanheira M, Watters AA, Mendes RE et al. Occurrence and molecular characterization of fusidic acid resistance mechanisms among *Staphylococcus* spp. from European countries (2008). J Antimicrob Chemother 2010; 65:1353-8
10. Dale GE, Broger C, D'Arcy A et al. A single amino acid substitution in *Staphylococcus aureus* dihydrofolate reductase determines trimethoprim resistance. J Mol Biol 1997; 266:23-30
11. Frey KM, Viswanathan K, Wright DL, Anderson AC. Prospective screening of novel antibacterial inhibitors of dihydrofolate reductase for mutational resistance. Antimicrob Agents Chemother 2012; 56:3556-62
12. Vickers AA, Potter NJ, Fishwick CW et al. Analysis of mutational resistance to trimethoprim in *Staphylococcus aureus* by genetic and structural modelling techniques. J Antimicrob Agents Chemother 2009; 63:1112-1117

13. Wichelhaus T, Schäfer V, Brade V. Differential effect of rpoB mutations on antibacterial activities of rifampicin and KRM-1648 against *Staphylococcus aureus*. J Antimicrob Agents Chemother 2001; 47:153-6
14. Wichelhaus T, Schäfer V, Brade V. Molecular characterization of rpoB mutations conferring cross-resistance to rifamycins on Methicillin-resistant *Staphylococcus aureus*. Antimicrob Agents Chemother 1999; 43:2813-6
15. Hellmark B, Söderquist B, Unemo M. Simultaneous species identification and detection of rifampicin resistance in staphylococci by sequencing of the rpoB gene. Eur J Clin Microbiol Infect Dis 2009; 28:183-90
16. Villar M, Marimón JM, García-Arenzana JM et al. Epidemiological and molecular aspects of rifampicin-resistant *Staphylococcus aureus* isolated from wounds, blood and respiratory samples. J Antimicrob Chemother 2011; 66:997-1000
17. Van Rensburg MJ, Whitelaw AC, Elisha BG. Genetic basis of rifampicin resistance in Methicillin-resistant *Staphylococcus aureus* suggests clonal expansion in hospitals in Cape Town, South Africa. BMC Microbiol 2012; 12:46
18. Aubry-Damon H, Soussy CJ, Courvilin P. Characterization of mutations in the rpoB gene that confer rifampin resistance in *Staphylococcus aureus*. Antimicrob Agents Chemother 1998; 42:2590-4
19. Mick V, Domínguez MA, Tubau F. Molecular characterization of resistance to Rifampicin in an emerging hospital-associated Methicillin-resistant *Staphylococcus aureus* clone ST228, Spain. BMC Microbiol 2010; 10:68
20. Schmitz FJ, Jones ME, Hofmann B et al. Characterization of grlA, grlB, gyrA, and gyrB mutations in 116 unrelated isolates of *Staphylococcus aureus* and effects of mutations on ciprofloxacin MIC. Antimicrob Agents Chemother 1998; 42:1249-52
21. Sanfilippo CM, Hesje CK, Haas W, Morris TW. Topoisomerase mutations that are associated with high-level resistance to earlier fluoroquinolones in *Staphylococcus aureus* have less effect on the antibacterial activity of besifloxacin. Chemotherapy 2011; 57):363-71
22. Sreedharan S, Oram M, Jensen B et al. DNA gyrAse gyrA mutations in ciprofloxacin-resistant strains of *Staphylococcus aureus*: close similarity with quinolone resistance mutations in *Escherichia coli*. J Bacteriol 1990; 172:7260-2
23. Peterson LR, Willard KE, Sinn LM et al. GyrA sequence analysis of *Staphylococcus aureus* and methicillin-resistant *S. aureus* strains selected, in vitro, for high-level ciprofloxacin resistance. Diagn Microbiol Infect Dis 1993; 17:97-101
24. Takahashi H, Kikuchi T, Shoji S et al. Characterization of gyrA, gyrB, grlA and grlB mutations in fluoroquinolone-resistant clinical isolates of *Staphylococcus aureus*. J Antimicrob Agents Chemother 1998; 41:49-57

25. Emerging multiple mutations and high-level fluoroquinolone resistance in methicillin-resistant *Staphylococcus aureus* isolated from ocular infections. Iihara H, Suzuki T, Kawamura Y, Ohkusu K, Inoue Y, Zhang W, Monir Shah M, Katagiri Y, Ohashi Y, Ezaki T. Diagn Microbiol Infect Dis 2006; 56:297-303
26. Horii T, Suzuki Y, Monji A et al. Detection of mutations in quinolone resistance-determining regions in levofloxacin- and methicillin-resistant *Staphylococcus aureus*: effects of the mutations on fluoroquinolone MICs. Diagn Microbiol Infect Dis 2003; 46:139-45
27. Noguchi N, Okihara T, Namiki Y et al. Susceptibility and resistance genes to fluoroquinolones in methicillin-resistant *Staphylococcus aureus* isolated in 2002. Int J Antimicrob Agents Chemother 2005; 25:374-9
28. Costa SS, Falcão C, Viveiros M et al. Exploring the contribution of efflux on the resistance to fluoroquinolones in clinical isolates of *Staphylococcus aureus*. BMC Microbiol 2011; 11:241
29. Yoon EJ, Lee CY, Shim MJ et al. Extended spectrum of quinolone resistance, even to a potential latter third-generation agent, as a result of a minimum of two GrlA and two GyrA alterations in quinolone-resistant *Staphylococcus aureus*. Chemotherapy 2010; 56:153-7
30. Schmitz FJ, Fluit AC, Hafner D et al. Development of resistance to ciprofloxacin, rifampicin, and mupirocin in methicillin-susceptible and -resistant *Staphylococcus aureus* isolates. Antimicrob Agents Chemother 2000; 44:3229-31
31. Schmitz FJ, Hofmann B, Hansen B et al. Relationship between ciprofloxacin, ofloxacin, levofloxacin, sparfloxacin and moxifloxacin (BAY 12-8039) MICs and mutations in grlA, grlB, gyrA and gyrB in 116 unrelated clinical isolates of *Staphylococcus aureus*. J Antimicrob Chemother 1998; 41:481-4
32. Ince D, Hooper DC. Mechanisms and frequency of resistance to premafloxacin in *Staphylococcus aureus*: novel mutations suggest novel drug-target interactions. Antimicrob Agents Chemother 2000; 44:3344-50
33. Takahata M, Yonezawa M, Kurose et al. Mutations in the gyrA and grlA genes of quinolone-resistant clinical isolates of methicillin-resistant *Staphylococcus aureus*. J Antimicrob Agents Chemother 1996; 38:543-6