Evolutionary processes driving the rise and fall of *Staphylococcus aureus* ST239, a dominant hybrid pathogen

Jacqueline L. Gill, Jessica Hedge, Daniel J. Wilson* and R. Craig MacLean*

Abstract

Staphylococcus aureus ST239 has been one of the most successful epidemic MRSA strains, and one of the leading causes of healthcare-associated MRSA infections. Here we investigate the evolution of ST239 using a combination of computational and experimental approaches. ST239 is thought to have emerged by a large scale chromosomal replacement event in which an ST8 clone acquired approximately 600 kb of DNA from an ST30 clone. Analysis of large-scale genomic data sets allowed us to confirm and refine the model of the origin of ST239. Importantly, we found that ST239 originated between the 1920s and 1945, implying that this MRSA lineage evolved at least 14 years before the clinical introduction of methicillin. Molecular evolution within ST239 has been dominated by purifying selection, although we found some evidence that the acquired region of the genome has evolved rapidly as a result of relaxed selective constraints. Crucially, we found that ST239 isolates have low competitive ability relative to both ST30 and ST8, demonstrating that this hybrid lineage is characterized by low fitness. We also found evidence of positive selection in a small number of genes involved in antibiotic resistance and virulence, suggesting that ST239 has evolved towards an increasingly pathogenic lifestyle. Collectively, these results support the view that low fitness has driven the recent decline of ST239, and highlight the challenge of using evolutionary approaches to understand the dynamics of pathogenic bacteria.

Introduction

Antimicrobial resistance (AMR) in pathogenic bacteria has created a healthcare crisis by increasing the costs and mortality rates associated with bacterial infections. In many important pathogens, the rise of resistance has been driven by the epidemic spread of a small number of very successful clones, often those that have successfully acquired a range of resistance genes by horizontal gene transfer. One important challenge at the moment is to understand the evolutionary processes that give rise to these AMR superbugs[1], such as *Staphylococcus aureus* sequence type ST22[2], *Escherichia coli* ST131[3] and *Klebsiella pneumonia* ST258[4].

S. aureus is an important commensal pathogen that provides a clear illustration of the AMR crisis. Approximately 1/3 of humans show persistent asymptomatic carriage of *S. aureus*, mainly in the nares, but this bacterium is capable of causing serious invasive infections at a number of sites in the body[5]. Antibiotic use has driven the epidemic spread of waves of resistant *S. aureus* strains; in particular, epidemic strains of methicillin-resistant *S. aureus* (EMRSAs) are now a global problem in healthcare settings and a growing problem in the community. For example, ST239 (a member of clonal complex CC8) is a major EMRSA strain that has been the causative agent of multiple epidemics in healthcare settings around the globe. The first ST239 isolates were collected in the late 1970s in Australia[6]. This was followed by the first EMRSA strain, also an ST239 and known as EMRSA-1, which emerged in the UK in 1981[7]. Between 1987 – 1988, over 40% of MRSA isolates collected in England and Wales were ST239 EMRSA-1[8], and ST239 became highly prevalent around the globe. By the mid-2000s, ST239 was the predominant MRSA strain in Asia, causing up to 90% of hospital-acquired MRSA within a region accounting for over 60% of the world's population[9]. More recently, reports from around the globe show that levels of ST239 MRSA have been decreasing dramatically, but the causes of this decline remain unclear[10 11 12].

Robinson and Enright proposed that ST239 was formed as a result of a large-scale chromosomal replacement event in which an ST8 clone acquired approximately 550 kb from an ST30 clone by recombination[13]. Interestingly, five of the six known other cases of large-scale chromosomal replacement in *S. aureus* overlap to some extent with the acquired region of ST239[13 14 15 16], suggesting that this region may be a recombination hotspot[17]. Large-scale chromosomal replacements have also been detected in a number of important AMR pathogens outside of *S. aureus*, including *K. pneumonia* ST258, *Campylobacter coli* ST1150 and *Streptococcus agalactiae*[18 19 20]. In ST239, this acquired region also contains a large (60 kb) SCC*mec*-III element that confers resistance to a broad spectrum of antibiotics and heavy metals, and ST239 is thought to have spread in healthcare settings around the world as a result of the selective advantage of the resistance provided by this acquired element[21].

Although subsequent studies have provided good support for the Robinson and Enright model, important questions regarding the origin of ST239 remain unresolved[22 23 24]. All of the 2,979 ST239 genomes in the Staphopia database that are associated with a typable SCC*mec* element carry a type III SCC*mec* element, suggesting that this element was acquired early in the evolutionary history of ST239. However, the SCC*mec*-III element has not been identified in any ST30 genomes, including all 1,896 ST30 genomes in the Staphopia database. The absence of SCC*mec*-III in ST30 suggests that the true ancestor of ST239 may be a closely related lineage that has not been considered in previous studies of ST239. Secondly, the origins of ST239 remain unclear. It is often assumed that antibiotic resistant pathogens evolve in response to treatment; in this case, it has been argued that the introduction of methicillin

into clinical practice in the 1960s drove the evolution of ST239. However, recent work has shown that some other MRSA STs from CC8 pre-dated the clinical introduction of methicillin[25], raising the possibility that ST239 may also have a deeper evolutionary history.

Finally, it is unclear why the prevalence of ST239 has recently declined. Acquiring DNA is usually associated with fitness costs[26 27 28], and many studies have shown fitness costs associated with acquired resistance genes[29 30 31], including large SCC*mec* elements[32 33]. Given these costs, it is possible that the recombination event(s) that gave rise to ST239 created a low fitness 'hybrid' pathogen, although this has not previously been investigated in detail.

In this paper, we use a combination of computational and experimental techniques to investigate the underlying evolutionary processes that have driven the rise of ST239. First, we assemble a diverse collection of ST239 genome sequences to reconstruct the evolutionary history of this lineage and to test the Robinson and Enright model. We then use competition assays to test the hypothesis that ST239 has low fitness, and to explore the link between AMR and fitness. Finally, we use population genetic approaches to understand how selection has operated in ST239, and to identify key genes involved in adaptive evolution in this lineage.

Results and discussion

Reconstructing the origin of ST239

The evolutionary drivers that contributed to the global emergence of the epidemic multi-drug-resistant MRSA strain ST239 are not well understood, however it has been suggested that the introduction of methicillin into clinics in the early 1960s may have been a contributing factor[34]. To reconstruct the evolutionary history of ST239, we assembled a collection of 96 ST239 genomes from isolates that were collected from diverse geographic locations and time points, deliberately avoiding over-representing isolates from intensively sampled ST239 outbreaks (e.g.[35] and [23]) (Supplementary Table 1A).

We constructed a pan-genome of 3,337 ST239 genes, of which 1,889 were present in 99 – 100% of all strains, resulting in a core genome length of 1,754,805 bp, representing a greater number of ST239 isolates from varied locations and dates than in previous evolutionary studies of ST239[36 37 38]. After identifying and excluding sites involved in recombination[39], 3,696 core variant sites remained, from which we reconstructed a phylogeny by maximum likelihood (Figure 1A). The phylogeny had strong bootstrap support, and the long branch lengths were suggestive of a diverse set of ST239 isolates[40 34]. The clustering of isolates into geographically distinct clades is consistent with the population structure observed by Harris (2010), and by Castillo-Ramirez (2012), who identified strong geographical clustering of ST239 sequences on continental, national and city scales[37 34]. In this study, isolates from Oceania, Asia and South America formed distinct clades, with rare exceptions. In contrast, North American and European isolates were dispersed throughout the tree.

We estimated a time to the most recent common ancestor (MRCA) of these ST239 sequences by fitting four evolutionary models using BEAST (Supplementary Table 2A)[41]. The time to MRCA was consistent between all four models, with no significant difference between the models identified through Bayes Factor analysis. Therefore, the time to MRCA from the simplest model (strict molecular clock, constant population size) is recorded here, as 1940.1

(95% Highest Posterior Density (HPD) intervals: 1934.7 – 1945.5). Similar results were obtained using BactDating (Supplementary Table 2B). Although there is some uncertainty in these estimates, these models predict that the origin of ST239 predated the clinical introduction of methicillin in 1959[42] by more than 10 years.

Identifying the donor of the acquired region of the ST239 genome

Robinson and Enright initially proposed that ST30 was the closest known ancestor of the ST239 acquired-region[13], and they were able to identify potential boundaries of the recombination event that gave rise to ST239[13]. However, their conclusions were based on the partial sequencing of a small number of genes from representative isolates. To test the Robinson-Enright model using a large collection of genomes, we extracted the core genes shared by >99% of the isolates from the ST239 collection (N = 96), combined with additional collections of diverse ST8 (N = 111) and ST30 (N = 57) genomes. A total of 1,980 core genes were identified that were shared between >99% of all 264 genomes. Within each ST (ST239, ST8 and ST30), we generated a consensus sequence for each of these genes, and calculated the percentage similarity between each ST239 gene and its homolog in either ST30 or ST8 (Figure 1C).

There was a clear distinction between regions of the ST239 genome that were closely related to ST8 and ST30, allowing us to clearly differentiate between the backbone and acquired regions of the ST239 genome (Figure 1D, E). This was consistent with Holden *et al*., who compared an ST239 genome sequence with a CC30 complete genome sequence, and found that the acquired region was more closely related to CC30 by a shift of roughly 1% in DNA percentage identity compared to the backbone region[21].

Although the acquired region of the ST239 genome is similar to ST30, it is possible that the true ancestor of this region was a closely related lineage of *S. aureus* that was not considered in previous analyses of ST239. For example, all of the 2,979 ST239 isolates in the Staphopia database that are associated with a typable SCC*mec* element carry a type III SCC*mec* element, but this element has not been identified in any of the 1,896 ST30 genomes in the Staphopia database[43], suggesting that ST30 may be a sister group, rather than the true ancestor of the acquired region (Supplementary Table 3). To systematically search for the ancestor of the acquired region, we used BIGSI to screen 447,833 bacterial and viral raw-read and assembled genomes for short sequences that match those found in the acquired region of the ST239 genome[44]. Sequences from isolates that were closely related to the acquired region of the ST239 genome were mapped to the acquired region of the ST239 reference genome (NCBI accession number FN433596), and assembled into a phylogeny alongside the acquired region of the 96 ST239 genomes (Figure 2A). Crucially, we found that all ST239 isolates share a recent common ancestor with five ST30 isolates dating from the 1950s and 1960s (and one ST30 isolate of unknown origin), and this branch is well-supported by bootstrapping. These five ST30 isolates (Supplementary Table 4) were all from the penicillin-resistant *S. aureus* phage type 80/81 clone, which caused serious healthcare-associated and community-associated infections worldwide, and was largely eliminated in the 1960s as a result of the introduction of methicillin[42 45 46].

Interestingly, none of the phage type 80/81 genomes that are closely related to ST239 contain an SCC*mec* element. The ubiquity of SCC*mec*-III in ST239 genomes suggests that the SCC*mec*-III element was acquired by the ancestor of the acquired region of ST239 following

the divergence of this lineage from the phage type 80/81 lineage. However, it is possible that the SCC*mec*-III element was secondarily acquired following the chromosomal replacement that gave rise to ST239. To test the secondary acquisition hypothesis, we estimated the time to MRCA of the SCC*mec*-III elements in the collection of ST239 isolates (Supplementary Figure 1, Supplementary Table 2A). However, there was considerable uncertainty in this estimate due to the small size of the SCC*mec*-III element (60 kb) and the high rate of recombination in this region of the genome. Given, these limitations, this analysis had limited power to reject the null hypothesis that the SCC*mec*-III element was acquired as part of the initial chromosomal replacement event that gave rise to ST239.

As a final test of the Robinson-Enright model, we used BIGSI to identify the closest known ancestor of the backbone region of the ST239 genome. Consistent with the Robinson and Enright model, this analysis identified ST8 as the closest ancestor of the ST239 backbone region[13]. Specifically, the ancestor of ST239 was part of a diverse lineage of ST8 that has been isolated across multiple continents over the last 70 years (Figure 2B, Supplementary Table 4).

Dating the MRCA of ST239 isolates places an upper bound on the origin of ST239, but it is possible that ST239 originated prior to the MRCA of contemporary isolates. To place a lower bound on the origin of ST239, we estimated times to the MRCA of the acquired and backbone regions of ST239 with their respective ST8 and ST30 ancestors. The MRCA of the acquired region and ST30 was estimated as 1900.3 – 1926.5 (95% HPD) and MRCA of the backbone region and ST8 was estimated as 1924.5 – 1934.7 (95% HPD). The overlap of these two estimates is encouraging, and this analysis suggests that ST239 is unlikely to have originated prior to the 1920s.

Recombination

Given that recombination created the ST239 lineage, it is possible that recombination has also played an important role in the subsequent evolution of ST239. To investigate this idea, we used ClonalFrameML to identify regions of recombination within the collection of 96 ST239 genomes (Figure 3, Supplementary Figure 2). Ten isolates from the TW20-like clade (Supplementary Table 5) were excluded from this analysis due to high sequence similarity to the reference sequence that all ST239s were mapped to.

The ratio of rates of recombination and mutation (R/θ) was 0.41, and the ratio of effects of recombination and mutation (r/m) was 2.37. Therefore, recombination events occurred 2.5 times less often than mutations, however, because each recombination event introduced an average of 6.0 substitutions, recombination overall was responsible for 2.4 times more substitutions than mutations[47]. In line with previous analyses, we found evidence of recombination hotspots in regions of the genome containing mobile genetic elements[24]. Recombination was particularly frequent in the region surrounding the φSA1 prophage, suggesting that this element, which borders the acquired region, may have played a key role in the recombination event that gave rise to ST239. Notably, the SCC*mec*-III element is also a hotspot for recombination.

Evolutionary consequences of large-scale chromosomal replacement

To begin to understand the evolutionary consequences of large-scale chromosomal replacement, we calculated the evolutionary rate of the backbone and acquired regions of the ST239 genome using BEAST (Supplementary Table 2A)[38]. After removing recombination, the overall genomic substitution rate was estimated as 1.205×10^{-6} SNPs/site/year (95% HPD 1.13×10^{-6} – 1.28×10^{-6} SNPs/site/year), which is similar to the substitution rates of other EMRSAs[2 48 49 50]. Interestingly, the substitution rate of the acquired region, 1.515×10^{-6} (1.32×10^{-6} – 1.71×10^{-6} SNPs/site/year; 95% HPD), was 2% – 51% more rapid than the backbone, 1.21 (1.13×10^{-6} – 1.29×10^{-6} SNPs/site/year; 95% HPD). On the one hand, the rapid evolution of the acquired region could be a signature of rapid adaptive evolution, perhaps to overcome the costs associated with horizontal gene transfer. Alternatively, it is possible that the acquired region has evolved at a high rate due to weak selective constraints in this region of the genome.

To understand how selection has acted in the two genetic regions of the ST239 genome, we used the McDonald-Kreitman test to compare patterns of polymorphism and divergence in the backbone and acquired regions of the genome[51]. The McDonald-Kreitman net neutrality index (*N*) indicates whether selection is overall purifying ($N > 1$) or positive ($N < 1$)[52].

In this analysis, we compared the acquired region of the 96 ST239 genomes with the homologous region in the previously defined collection of 57 ST30 genomes. Additionally, we compared the backbone region of the 96 ST239 genomes with the homologous region in the previously defined collection of 111 ST8 genomes. Only "core" genes that were shared by more than 99% of all ST239, ST8 and ST30 genomes in the collections were included. The divergence between ST239 and ST30 in the acquired region (0.326 substitutions/Mb) was much greater than the divergence between ST239 and ST8 in the backbone region (0.188 substitutions/Mb), reflecting the fact that ST239 is more closely related to ST8 than ST30 in the respective regions. Divergence between STs at non-synonymous sites was low relative to levels of within-ST polymorphism, demonstrating an overall trend towards purifying selection in ST239 (Table 1; $N > 1$). However, we did not find any difference in the net neutrality index between the backbone ($N = 2.06$; 95% C.I. $N = 0.92 – 2.51$) and the acquired region ($N = 1.52$; 95% C.I. $N = 1.34 – 3.19$).

One weakness of this approach in this context is that the backbone and acquired regions of the ST239 genome have to be compared to different outgroups. Given that signatures of purifying selection should become stronger over time, it could be argued that the McDonald-Kreitman test is biased towards detecting purifying selection in the acquired region of the genome, which was compared to a more divergent outgroup. To further test the idea that selection on the acquired region has been weak, we calculated the ratio of GC → AT to AT → GC substitutions in each region of the ST239 genome (Table 2). Spontaneous mutation in *S. aureus* is biased towards GC → AT transitions, and regions of the genome that are subject to weak selective constraints are therefore expected to have a high ratio of GC → AT/AT → GC substitutions[53]. In this case, the GC → AT/AT → GC ratio in the acquired region was significantly greater than that of the backbone region (Fisher's exact test, odds ratio = 1.15, $P = 0.0303$). This analysis, which uses data on substitutions that have occurred during the diversification of ST239, provides evidence of relaxed selective constraints on the acquired region of the ST239 genome.

Fitness costs of chromosomal replacement

To understand the fitness effects of chromosomal replacement more directly, we measured the competitive ability of a collection of *S. aureus* isolates *in vitro* (Supplementary Figure 3). Our collection of isolates included divergent isolates of ST239 (N = 4), ST8 (N = 6) and ST30 (N = 5) that were deliberately chosen to avoid including clonal isolates within STs. If chromosomal replacement is costly, then we would expect ST239 isolates to have low competitive ability relative to ST8. The additional ST30 isolates provided a useful reference point to compare fitness values. Isolates were directly competed against each other in three different culture media; Tryptic Soy Broth (TSB), Brain-Heart Infusion (BHI) broth, and Porcine Serum (PS). These media impose different stresses that mimic some of the challenges encountered by *S. aureus* in clinical environments.

We deep sequenced (142 – 211x depth) each competition mixture before and after growth, and estimated changes in the relative abundance of each isolate by quantifying the relative abundance of isolate-specific SNPs during competition. Competitive ability varied between isolates, and we found a strong statistical interaction between ST and media, which reflects the fact that the average competitive ability of the three STs varied across media (Figure 4 A – C; Table 3; Supplementary Table 6A). For example, we did not find any evidence of low fitness associated with ST239 in porcine serum. In spite of this variation in fitness, we found a significant difference in competitive ability between STs. Crucially, we found that ST239 had lower competitive ability than both ST8 and ST30 (Figure 4D; post-hoc Tukey test P <0.05) The low fitness of ST239 relative to the ST8 is consistent with the idea that chromosomal replacement carries a long-term fitness cost. This hypothesis is further supported by the high fitness of ST30, which suggests that the difference in fitness between ST239 and ST8 reflects low fitness of ST239, rather than high fitness of ST8.

As a complementary approach to measure fitness, we also measured the growth rate of the individual ST239, ST8 and ST30 isolates in TSB, BHI and PS (Supplementary Figure 4). There was significant variation in growth rate between isolates, media, and ST showing that this trait is very plastic (Supplementary Table 6B). Crucially, we found that ST239 isolates had significantly lower growth rate than ST8 isolates, providing further evidence of costs associated with chromosomal replacement.

Parallel evolution

A recurring theme of studies of microbial evolution is that genes that are under strong positive selection evolve in parallel[54 55 56]. To better understand the selective pressures that have shaped the evolution of ST239, we compared the distribution of observed mutations per gene with a neutral model derived from the Poisson distribution, in which mutations are randomly distributed across genes. This analysis was carried out independently for the acquired and backbone regions of the ST239 genome to take into account the differences in substitution rate between these genomic regions. Only the 1,980 “core” genes that were shared between the previously defined ST239, ST30 and ST8 genome collections were included.

The number of mutations per gene differed from the Poisson expectation in both the acquired region (X^2 = 19.158, P = 0.0014) and backbone region (X^2 = 177.8, P <0.0001), showing that

substitutions are non-randomly distributed across the ST239 genome. A subset of genes that show more evidence of parallel evolution than expected due to chance alone were defined as those that had 9 or more substitutions per gene. Our justification for this cut-off is that the Poisson distribution predicts that 1 or 2 genes in each region of the genome should have acquired 9 mutations or more due to chance alone, given an average of 2.53 substitutions per gene in the backbone region and 2.97 substitutions per gene in the acquired region. The proportion of genes showing evidence of parallel evolution did not differ between the backbone (20/1659 = 1.21% genes) and acquired regions (6/316 = 1.90% genes), indicating that genes under positive selection are evenly distributed across the ST239 genome (Chi-squared test, $X^2 = 0.9818$, $P = 0.3218$). However, of the genes showing evidence of parallel evolution, those in the acquired region had a much larger proportion of substitutions ($N = 154$ substitutions, 15.75%) than the backbone region ($N = 284$ substitutions; 6.61%). The acquired region includes the *spa* gene ($N = 62$ substitutions) which is known to undergo strong selection mediated by the immune system. Even after excluding this gene, the remaining genes showing evidence of parallel evolution in the acquired region are significantly enriched for substitutions compared to those in the backbone (Chi-squared test, $X^2 = 13.111$, $P = 0.000294$). The high rate of substitutions in these genes suggests that the acquired region has been a hotspot for adaptive evolution in the ST239 genome. Note that this analysis is robust to the overall elevated substitution rate of the acquired region because it is based on a sub-set of genes that show a high rate of substitution compared to other genes in the region.

Interestingly, many ($N = 7$) of the genes that show evidence of parallel evolution (Supplementary Table 7) are involved in resistance to antibiotics, including vancomycin/daptomycin (*walK*), fluoroquinolones (*grlA*), β-lactams (*ponA* and *mprF*) and rifampicin (*rpoB*). To test for elevated resistance at phenotypic level, we measured the resistance of our isolates to a broad panel of antibiotics that have activity against *S. aureus*. ST239 isolates were resistant to a greater number of antibiotics (Figure 5; Supplementary Table 8; mean = 7.25; s.e. = 0.75) than either ST30 or ST8 (One-Way ANOVA followed by Hsu's test $F_{2,12} = 5.63$; $P = 0.0187$), highlighting the high levels of AMR associated with this lineage.

Conclusion

In line with previous work, our results support the hypothesis that ST239 was produced by a large scale chromosomal replacement event in which an ST8 clone acquired >600 kb of DNA from an ST30 clone[13 21]. We were able to refine this model by showing that the ancestor of the acquired region of the ST239 genome was most closely related to phage type 80/81 clones that were associated with the epidemic spread of penicillin resistance in the 1950s and 1960s[57]. The most parsimonious explanation for the presence of the SCC*mec*-III in ST239 is that this element was acquired by the ST30 ancestor of ST239, following the divergence of this lineage from phage type 80/81. However, our analysis had limited power to detect secondary acquisition of SCC*mec*-III. We estimate that ST239 originated between the 1920s and 1945, providing further evidence that MRSA pre-dated the clinical introduction of methicillin in 1959[25]. SCC*mec*-III provides resistance to first generation antibiotics that were used prior to the introduction of methicillin, such as tetracycline and erythromycin, and heavy metals, such as cadmium and mercury, that are used in disinfectants and biocides in

healthcare settings[58 59], suggesting that these resistance phenotypes may have provided ST239 with a selective advantage prior to the introduction of methicillin.

Fitness costs of laboratory-evolved antibiotic resistance have been demonstrated in many studies[28 60 61], but estimates of the cost of resistance in pathogen populations have received less attention[62 63 64]. We found extensive variation in competitive fitness between *S. aureus* isolates and culture conditions. In spite of this variation, we found a clear overall trend towards low fitness in ST239 relative to ST8, providing good evidence of a fitness cost associated with the evolution of elevated antibiotic resistance. This hypothesis is further supported by epidemiological evidence. ST8 is primarily found in the community, where antimicrobial use is low, whereas ST239 has mainly been restricted to healthcare settings where antimicrobial use is high, suggesting that the low fitness of this ST has restricted the spread of ST239 into the community[65 21 66]. The acquisition of large SCC*mec* elements tends to generate a fitness cost, suggesting that the SCC*mec*-III element (which is the largest known SCC*mec* element) contributes to the low fitness of ST239 (see also[57]). Although we found some evidence that the acquired region of the ST239 genome is subject to relaxed selective constraints, the evolution of this region is dominated by purifying selection, suggesting that the chromosomal replacement may have had costs beyond those associated with the acquisition of SCC*mec*-III.

Experimental studies have found evidence that bacteria can adapt to the cost of gene acquisition through the process of compensatory evolution[25 27]. Although there are clear examples of compensatory adaptation in pathogenic bacteria (i.e.[67 68 69]), the prevalence of the compensatory evolution is unclear[59]. Although the evolution of ST239 has been dominated by purifying selection, we found evidence of positive selection in genes that are implicated in antibiotic resistance, virulence and metabolism. Notably, many of the genes that show clear hallmarks of lineage-specific positive selection in the ST239 genome are associated with resistance to antibiotics that have been used to treat MRSA infections, such as ciprofloxacin (*grlA*), vancomycin (*walK*), and rifampicin (*rpoB*). These patterns of parallel evolution suggest that the ST239 has evolved to increase antibiotic resistance and virulence, rather than to overcome the costs associated with chromosomal replacement and SCC*mec*-III acquisition.

Bacterial recombination is typically associated with the exchange of short DNA sequences between closely related strains or species[70], and the large-scale chromosomal replacements that have been detected in pathogenic bacteria are conspicuous exceptions to this overall trend[18 19 20]. Our results support the idea that ST239 is a ‘hopeful monster’ that has declined in prevalence due to fitness costs of chromosomal replacement, and an important goal for future work will be to understand the fitness consequences of chromosomal replacement in other dominant hybrid pathogens.

Acknowledgements

This project was funded by a Wellcome Trust Grant (106918/Z/15/Z), held by RCM. JLG was supported by funding from the Biotechnology and Biological Sciences Research Council (BBSRC) (BB/M011224/1). DJW is a Sir Henry Dale Fellow, jointly funded by the Wellcome Trust and the Royal Society (101237/Z/13/B). This work was funded by a Big Data Institute Robertson Fellowship (DJW). We thank the Oxford Genomics Center (funded by Wellcome Trust Grant 203141/Z/16/Z) for the generation and initial processing of Illumina sequence data.

Methods

Genome data retrieval and processing

Sequences with an MLST profile corresponding to ST239 were identified within the Staphopia database[43]. The isolation date and geographic location were extracted from the metadata files. Additional ST239 sequences with corresponding isolation date and location metadata were identified from the NCBI database using MegaBLAST[71] and through a literature search. A total of 96 ST239 sequences were selected and downloaded from EMBL-EBI (Supplementary Table 1A). Similarly, a total of 57 ST30 and 111 ST8 genomes were identified and downloaded from EMBL-EBI (Supplementary Table 1B and 1C).

Where the sequences were downloaded as assemblies or complete genomes, raw sequence reads were simulated with dwgsim v0.1.11[72]. Illumina reads were trimmed using Trimmomatic[73], and bwa v0.7.15 and SAMtools v1.3.1[74] were used to map the ST239 sequences to the ST239 reference sequence (accession number FN433596). ST30 sequences were mapped to both the ST239 and ST30 reference sequences (accession number LN626917); similarly, the ST8 sequences were mapped to the ST239 and ST8 reference sequences (accession number CP007690).

The boundaries of the SCC*mec*-III element were identified in the ST239 reference genome, using BLAST against the reference SCC*mec*-III element with the accession number AB037671[52]. The boundaries were confirmed using SCC*mec*Finder[75].

The SCC*mec* type of all 2,979 ST239 sequences in the Staphopia database was predicted using the Staphopia API. The Staphopia database was also mined for all sequences predicted to contain SCC*mec*-III. The MLST of these sequences was identified using PubMLST[76]. The MLST was confirmed using ARIBA v2.13.3[77] and the SCC*mec* type was confirmed using SCC*mec*Finder.

Construction of ST239 phylogeny

RaxML v8.2.9[78] was used to construct a maximum likelihood phylogeny from the collection of 96 ST239 genomes that had been mapped to the ST239 reference sequence, using a GTR model with gamma correction for among site rate variation, which was replicated for 100 bootstraps, with recombination masked using Gubbins[39]. The ST8 reference sequence, mapped to the ST239 reference sequence, was used as an out-group. Genes were annotated with Prokka v1.13[79]. Gene function for genes that could be annotated by Prokka was identified in UniProt, and MegaBLAST[71] was used to identify genes where no annotation was found with Prokka. The pangenome and core genome (genes shared by >99% of the ST239 isolates) was extracted using Roary v3.12.0[80].

Estimation of time to the MRCA of the ST239 collection

The time to the MRCA for the collection of 96 ST239 genomes was initially estimated from the maximum likelihood phylogeny (with recombination masked using Gubbins[39]), using the BactDating R package linear regression function[81]. Mixed gamma, strict gamma and relaxed

gamma evolutionary models were run for 10,000,000 MCMC steps to estimate a time to the MRCA of the ST239 whole-genome sequences. The Effective Sample Size (ESS) values of all parameters were greater than 100, indicating adequate sampling of the posterior distribution.

A total of 6,819 variant sites were extracted from the ST239 sequence alignments, after recombination was masked using Gubbins[39], using snp-sites v2.3.3[82]. Bayesian phylogenetic analysis was also carried out using BEAST version 1.10.4[38], using the GTR nucleotide substitution model with all combinations of the strict and uncorrelated relaxed molecular clock models, and constant and exponential growth models. The XML file was edited to reflect the number of unchanging sites in the original alignments. For each model, three independent MCMC chains with 300,000,000 steps were run and combined, with path sampling/stepping-stone sampling every 100 steps. In all cases, the Bayes Factor showed no significant difference in the likelihood of the different models, and therefore the estimated time to the MRCA from the simplest model (strict molecular clock, constant population size) was recorded. The burn-in was set at 10%, and runs were combined using LogCombiner, with a re-sample size of 10,000. The MRCA and evolutionary rates were estimated with 95% HPD intervals.

BEAST analysis was repeated for 95 ST239 SCC*mec*-III element sequences (N = 54 variant sites), using the same clock and nucleotide substitution models as previous (one ST239 sequence was removed from the analysis due to low mapping quality of the SCC*mec*-III region).

BLAST comparison of the ST239, ST30 and ST8 core genes

A multiple sequence alignment was constructed from the 96 ST239, 57 ST30 and 111 ST8 genomes that had been mapped to the ST239 reference genome. The shared pan-genome was calculated (N = 2,962 genes), and the core genes that were shared between >99% of all 264 genomes were extracted using Roary v3.12.0[77] (N = 1,980 core genes). EMBOSS[83] was used to generate ST-specific consensus sequences from the core genes of each ST (ST239, ST8 and ST30). For each gene in the ST239 consensus sequence, the percentage identities compared to the homologous gene in the ST30 and ST8 consensus sequences were calculated using MegaBLAST[71]. The ST239 core genes were defined as “acquired” (i.e. from the ST30-like region) or “backbone” (i.e. from the ST8-like region) depending on their position in the genome and similarity to the ST30 and ST8 consensus sequences.

Identifying the closest known ancestor of the acquired and backbone regions

The consensus sequences of the 316 core genes from the acquired region of the 96 ST239 sequences were queried for similar sequences in BIGSI using a k-mer threshold of 99. This allowed for an average divergence of around 10 SNPs per gene. The most closely related sequences were identified by ST using the Staphopia API. The MLST was double checked using ARIBA, and 28 sequences were removed due to uncertainty in typing, which indicated contamination.

The consensus sequences of the 1,659 core genes from the backbone region of the 96 ST239 sequences were also queried for similar sequences in BIGSI using a k-mer threshold of 99, and Staphopia, as above. This allowed for an average divergence of around 9 SNPs per gene.

Sequences that shared at least 99% k-mer identity with over 200 of the ST239 core genes from the acquired region, and shared two or fewer MLST alleles with ST239, were downloaded from the EBI database. Only a selection of the twelve most closely related ST30, ST36 and ST39 sequences were included. These 143 sequences were mapped to the ST239 TW20 reference sequence, as above. Four sequences were removed due to poor mapping quality (>25% gaps). The acquired region was extracted (minus the SCC*mec* element) and combined into a multifasta alignment with the acquired regions from the 96 ST239 sequences and 57 ST30 sequences that were described previously.

Sequences that shared at least 99% k-mer identity with over 1,400 of the ST239 core genes from the backbone region, and were not previously identified as ST239-like, were then downloaded from the EBI database. All 32 non-ST239-like sequences were identified using the Staphopia API as ST8. These 32 ST8 sequences were mapped to the ST239 TW20 reference sequence, as above. The backbone region was extracted and combined into a multifasta alignment with the backbone regions from the 96 ST239 sequences and 111 ST8 sequences that were described previously.

All variant sites were extracted using snp-sites v2.3.3, and RaxML was used to estimate a maximum-likelihood phylogeny of all 292 acquired-region sequences, using a GTR model with gamma correction for among site rate variation and replicated for 100 bootstraps, after recombination was masked using Gubbins as previous. This was outgroup-rooted to the corresponding region of the ST8 reference sequence. A maximum-likelihood phylogeny was also estimated for the backbone region sequences, which was outgroup-rooted to the corresponding region of the ST30 reference sequence.

The acquired region of the 96 ST239 sequences and the closest related non-ST239 clade, consisting of six ST30 isolates, were combined into a multiple sequence alignment. Variant sites were extracted using snp-sites v2.3.3. To calculate the time to the MRCA, Bayesian phylogenetic analysis was carried out using BEAST version 1.10.4, as previous (for the GTR nucleotide substitution model with all combinations of the strict and uncorrelated relaxed molecular clock models, and constant and exponential growth models). This was repeated for the backbone region of the 96 ST239 sequences and the closest related non-ST239 clade, consisting of 18 ST8 isolates (one ST8 sequence was removed from the analysis, as it had no associated date of isolation). In both cases, the Bayes Factor showed no significant difference in the likelihood of the different models, and therefore the estimated time to the MRCA from the simplest model (strict molecular clock, constant population size) was recorded.

Visualisation of recombination in ST239

ClonalFrameML v1.0-20[44] was used to estimate regions of recombination from the ST239 phylogeny, before recombination was masked. Ten isolates from the TW20-like clade (Supplementary Table 5) were excluded from this analysis due to high sequence similarity to the ST239 reference sequence.

Estimating evolutionary rates of the acquired and backbone regions

The evolutionary rates of the acquired and backbone regions of ST239 were estimated using BEAST analysis on 5,353 variant sites from the 96 ST239 backbone region sequences, and 1,449 variant sites from the ST239 acquired region sequences. The GTR nucleotide substitution model was used with all combinations of the strict and uncorrelated relaxed molecular clock models, and constant and exponential growth models, as previous.

McDonald-Kreitman comparison of ST239, ST30 and ST8 core genes

All 1,980 core genes from the 96 ST239 genomes, 111 ST8 genomes and 57 ST30 genomes were converted into amino acid sequence alignments. Each of the ST239, ST30 and ST8 consensus core gene sequences that were generated previously using EMBOSS[80] were also converted into consensus amino acid sequences.

The number of fixed synonymous and fixed non-synonymous SNPs was calculated for core genes within the whole ST239 genome, core genes within the backbone region, and core genes within the acquired region using snp-sites v2.3.3. This analysis was repeated for the 111 ST8 genomes collection and the 57 ST30 genomes collection. The McDonald-Kreitman neutrality index (N) was calculated as $N = (P_n/P_s)/(D_n/D_s)$ where N is the net neutrality index, P_n is the number of non-synonymous polymorphisms, P_s is the number of synonymous polymorphisms, D_n is the number of non-synonymous substitutions, and D_s is the number of synonymous substitutions.

Competition experiments

Cryostocks of the fifteen isolates were streaked on TSA, and incubated at 37^{O}C for 24 hours. Single colonies were incubated for 24 hours in 3 mL TSB at 37^{O}C with 225 RPM shaking. One mL of each culture was combined into a single mixture and mixed thoroughly. Genomic DNA was extracted and purified from 1 mL of the mixture, and sequencing was carried out, as previous.

Six mL of the mixture was pelleted and washed three times in PBS, and separated into six 1 mL aliquots. These were diluted 50x in either TSB, BHI or PS. A total of 3 mL of each mixed culture was incubated at 37^{O}C with 225 RPM shaking for 24 hours. After 24 hours, genomic DNA was extracted and purified, and sequencing was carried out, as previous. All competition experiments were repeated in triplicate.

The DNA sequences from before and after each competition were mapped to the consensus sequence, as previous. The number of reads supporting each unique variant site allele (for both the reference allele and the variant allele) was determined from the mapping of the raw sequence reads from each competition experiment. Any sites that were supported with a total of three reads or less, for both the variant allele and the consensus allele, were removed from the analysis, to reduce the number of incorrect alleles due to sequencing error.

For each competition experiment, the number of supporting reads for each unique variant site was recorded. From this, the average coverage of all variants that were unique to each isolate was determined, before and after being exposed to the competition conditions for 24 hours, to

determine how the proportion of each isolate changed during each competition. The limit of detection for each isolate was also calculated, as $M_Z=4/((N_Z + N_z')/V)$, where M is the minimum detection limit for isolate Z, N_Z is the total number of reads in support of isolate Z at sites unique to isolate Z, N_Z' is the total number of reads in support of non-Z isolates at sites unique to isolate Z, and V is the number of variant sites that are unique to isolate Z. Any isolate with an average coverage that was lower than the limit of detection was called at the limit of detection for that isolate. Raw competitive ability was calculated as the difference in log relative abundance of each isolate before and after competition in each replicate. Raw competition values were then re-scaled, such that the mean competitive ability in each culture medium was equal to zero. This small correction factor was used to account for the fact that the true final density of some isolates was below the minimal detection threshold.

Sequencing DNA from isolates for competition experiments

Cryostocks of the fifteen isolates were streaked on TSA, and incubated at 37^{O}C for 24 hours. Single colonies were incubated for 24 hours in 3 mL TSB at 37^{O}C with 225 RPM shaking. Genomic DNA was extracted and purified using the Qiagen DNeasy Blood and Tissue kit, and following the protocol for purification of bacterial or yeast DNA with enzymatic lysis, using QIAcube. Sequencing was carried out using an Illumina HiSeq4000 system with 150 bp paired-end reads by the Oxford Genomic Centre, Wellcome Trust Centre for Human Genetics, University of Oxford, UK.

The sequences were *de novo* assembled into contigs using SPAdes v3.13.0[84], and ordered against the ST239, ST30 or ST8 reference sequences using abacas v1.3.1[85]. Sequences were annotated using PROKKA or UniProt, as previous. Using Roary, core genes were defined as genes that were shared between all 15 isolates. Unique variant alleles were defined for each isolate, which were identified in only a single isolate, using snp-sites v2.3.379.

To verify the validity of each unique variant site, each DNA sequence was re-mapped to the consensus sequence using bwa and SAMtools[71]. The number of reads supporting each unique variant site was extracted, and only unique variant sites that could be used to accurately identify a single isolate and that were supported with 4 or more reads were included as true unique variant sites.

AMR profiles

Cryostocks of the fifteen isolates were streaked on TSA, and incubated at 37^{O}C for 24 hours. Single colonies were incubated for 24 hours in 3 mL MH2 at 37^{O}C with 225 RPM shaking. Cultures were diluted 100x in MH2 broth (to a density of ~1 x10^{6} CFUs/mL), according to the MicroNaut evaluation protocol for MicroNaut-S MRSA/GP. 100 µL was added to each well of a MicroNaut-S MRSA/GP plate and incubated for 18 hours at 37^{O}C with 225 RPM shaking, after which OD_{595} was measured. AMR breakpoints were assessed according to EUCAST[86]. This was repeated in triplicate for each isolate. Tests for ampicillin and penicillin always gave the same results, hence we considered the result from these two antibiotics as a single test score.

Figures

Figure 1A. Out-group rooted maximum-likelihood phylogeny of the 96 ST239 genomes. Distances are shown in in SNPs/Mb and bootstrap support values below 95 are shown for branches. Isolates are colour coded according to geographical origin, as displayed on panel B (Blue, Europe; Pink, Asia; Yellow, Oceania; Green, South America; Black, North America; Red, Africa). **1C.** BLAST DNA percentage identity of the ST8 (blue triangle) and ST30 (red circle) consensus core gene sequences, compared to the ST239 consensus core gene sequences (consensus sequences were formed from 96 ST239 genomes, 111 ST8 genomes and 57 ST30 genomes, for 1,980 core genes that were shared between all 264 genomes). In the ST239 genome, the acquired region spans the origin of replication, and hence is split between the beginning and the end of the linearized genome. The hybrid boundaries estimated by Castillo-Ramirez *et al*. (2011) are highlighted with vertical dashed grey lines. **1D.** Close up of the first boundary of the large chromosomal replacement event. **1E.** Close up of the second boundary of the large chromosomal replacement event. The hybridisation boundaries estimated by Castillo-Ramirez *et al*. (2011) are highlighted with vertical dashed grey lines, and the hybridisation boundary ranges used in this study are highlighted by a grey rectangle. Each data-point represents a single gene. The gene positions correspond to the genomic position within the ST239 reference genome.

Figure 2A. Outgroup-rooted maximum likelihood phylogeny of the acquired region of the ST239 genomes, and the most closely related non-ST239-like isolates from BIGSI analysis. Distances are shown in in SNPs/Mb. ST239 isolates are in cyan, ST30 isolates in red and other STs are in green. The zoom panel contains the ST239 clade and the most closely related non-ST239 isolates, with bootstrap support values above 95 shown for branches. **2B.** Outgroup-rooted maximum likelihood phylogeny of the backbone region of the ST239 genomes, and the most closely related non-ST239-like isolates from BIGSI analysis. Distances are shown in in SNPs/Mb. ST239 isolates are in cyan and ST8 isolates are in blue. The zoom panel contains the ST239 clade and the most closely related non-ST239 isolates, with bootstrap support values above 95 shown for branches.

Figure 3. Recombination plot of the ST239 genomes compared with the ST239 TW20 reference genome. Dark blue regions represent potential sites of recombination. Sites that are non-polymorphic are shown in light blue. Polymorphic sites are shown in a colour indicating their level of homoplasy; white is no homoplasy and the range from yellow to red is increasing degrees of homoplasy. The positions of known MGEs and the two inherited regions estimated by Holden *et al*. and Castillo-Ramirez *et al*. are highlighted along the ST239 reference genome beneath the plot; green, acquired region; magenta, backbone region.

Figure 4. Competitive ability. Plotted points show the competitive ability of *S. aureus* isolates (blue circles) and the mean competitive ability of each ST (red circles; +/- s.e.m; N = 4-6). The competitive ability of each isolate was measured in triplicate, and error for individual isolates was small (s.e = 0.045-0.001). ST239 isolates have reduced competitive ability relative to ST8 and ST30 in BHI and TSB, but not PS, as judged by a post-hoc Tukey test (P <0.05). The final panel shows the overall effect of ST on competitive ability (+/- s.e.) across all three media.

Figure 5A. Mean (+/- s.e.m.; N = 4 – 6) number of AMR phenotypes for ST239, ST8 and ST30 isolates. **5B**. Heat map of AMR phenotypes for ST239, ST8 and ST30 isolates. Orange (R), resistant; green (S), susceptible.

Tables

Table 1. McDonald-Kreitman net neutrality index for the two genetic regions of ST239. Fixed non-synonymous substitutions (Dn); fixed synonymous substitutions (Ds); non-synonymous polymorphisms (Pn); synonymous polymorphisms (Ps); McDonald-Kreitman net neutrality index (N).

Table 2. Total number of GC → AT and AT → GC substitutions in the acquired and backbone regions of the ST239 genome.

Table 3. Reduced ANOVA table of significant competitive ability effects.

Literature cited

1 MacLean RC, San Millan A. The evolution of antibiotic resistance. Science. 2019;365: 1082–1083. doi:10.1126/science.aax3879

2 Holden MTG, Hsu L-Y, Kurt K, Weinert LA, Mather AE, Harris SR, et al. A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant Staphylococcus aureus pandemic. Genome Research. 2013;23: 653–664. doi:10.1101/gr.147710.112

3 Nicolas-Chanoine M-H, Bertrand X, Madec J-Y. Escherichia coli ST131, an Intriguing Clonal Group. Clinical Microbiology Reviews. 2014;27: 543–574. doi:10.1128/cmr.00125-13

4 Wyres KL, Holt KE. Klebsiella pneumoniae Population Genomics and Antimicrobial-Resistant Clones. Trends in Microbiology. 2016;24: 944–956. doi:10.1016/j.tim.2016.09.007

5 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: 751–762. doi:10.1016/s1473-3099(05)70295-4

6 Pavillard R, Harvey K, Douglas D, Hewstone A, Andrew J, Collopy B, et al. Epidemie of hospital-acquired infection due to methicillin-resistant Staphylococcus aureus in major Victorian hospitals. Medical Journal of Australia. 1982;1: 451–454. doi:10.5694/j.1326-5377.1982.tb132413.x

7 Cookson BD, Phillips I. Epidemic methicillin-resistant Staphylococcus aureus. Journal of Antimicrobial Chemotherapy. 1988;21: 57–65. doi:10.1093/jac/21.suppl_c.57

8 Kerr S, Kerr GE, Mackintosh CA, Marples RR. A survey of methicillin-resistant Staphylococcus aureus affecting patients in England and Wales. Journal of Hospital Infection. 1990;16: 35 – 48. doi: 10.1016/0195-6701(90)90047-r

9 Feil EJ, Nickerson EK, Chantratita N, Wuthiekanun V, Srisomang P, Cousins R, et al. Rapid Detection of the Pandemic Methicillin-Resistant Staphylococcus aureus Clone ST 239, a Dominant Strain in Asian Hospitals. Journal of Clinical Microbiology. 2008;46: 1520–1522. doi:10.1128/jcm.02238-07

10 Amorim ML, Faria NA, Oliveira DC, Vasconcelos C, Cabeda JC, Mendes AC, et al. Changes in the Clonal Nature and Antibiotic Resistance Profiles of Methicillin-Resistant Staphylococcus aureus Isolates Associated with Spread of the EMRSA-15 Clone in a Tertiary Care Portuguese Hospital. Journal of Clinical Microbiology. 2007;45: 2881–2888. doi:10.1128/jcm.00603-07

11 Conceição T, Aires-de-Sousa M, Füzi M, Tóth Á, Pászti J, Ungvári E, et al. Replacement of methicillin-resistant Staphylococcus aureus clones in Hungary over time: a 10-year surveillance study. Clinical Microbiology and Infection. 2007;13: 971–979. doi:10.1111/j.1469-0691.2007.01794.x

12 Hu F, Zhu D, Wang F, Wang M. Current Status and Trends of Antibacterial Resistance in China. Clinical Infectious Diseases. 2018;67: S128–S134. doi:10.1093/cid/ciy657

13 Robinson DA, Enright MC. Evolution of Staphylococcus aureus by Large Chromosomal Replacements. JB. 2004;186: 1060–1064. doi:10.1128/jb.186.4.1060-1064.2004

14 Thomas JC, Godfrey PA, Feldgarden M, Robinson DA. Draft Genome Sequences of Staphylococcus aureus Sequence Type 34 (ST34) and ST42 Hybrids. Journal of Bacteriology. 2012;194: 2740–2741. doi:10.1128/jb.00248-12
15 Nimmo GR, Steen JA, Monecke S, Ehricht R, Slickers P, Thomas JC, et al. ST2249-MRSA-III: a second major recombinant methicillin-resistant Staphylococcus aureus clone causing healthcare infection in the 1970s. Clinical Microbiology and Infection. 2015;21: 444–450. doi:10.1016/j.cmi.2014.12.018
16 Spoor LE, Richardson E, Richards AC, Wilson GJ, Mendonca C, Gupta RKr, et al. Recombination-mediated remodelling of host–pathogen interactions during Staphylococcus aureus niche adaptation. Microbial Genomics. 2015;1. doi:10.1099/mgen.0.000036
17 Everitt RG, Didelot X, Batty EM, Miller RR, Knox K, Young BC, et al. Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus. Nat Commun. 2014;5. doi:10.1038/ncomms4956
18 Chen L, Mathema B, Pitout JDD, DeLeo FR, Kreiswirth BN. Epidemic Klebsiella pneumoniae ST258 Is a Hybrid Strain. Jacoby G, editor. mBio. 2014;5. doi:10.1128/mbio.01355-14
19 Sheppard SK, Didelot X, Jolley KA, Darling AE, Pascoe B, Meric G, et al. Progressive genome-wide introgression in agriculturalCampylobacter coli. Mol Ecol. 2012;22: 1051–1064. doi:10.1111/mec.12162
20 Brochet M, Rusniok C, Couve E, Dramsi S, Poyart C, Trieu-Cuot P, et al. Shaping a bacterial genome by large chromosomal replacements, the evolutionary history of Streptococcus agalactiae. Proceedings of the National Academy of Sciences. 2008;105: 15961–15966. doi:10.1073/pnas.0803654105
21 Monecke S, Slickers P, Gawlik D, Müller E, Reissig A, Ruppelt-Lorz A, et al. Molecular Typing of ST239-MRSA-III From Diverse Geographic Locations and the Evolution of the SCCmec III Element During Its Intercontinental Spread. Front Microbiol. 2018;9. doi:10.3389/fmicb.2018.01436
22 Holden MTG, Lindsay JA, Corton C, Quail MA, Cockfield JD, Pathak S, et al. Genome Sequence of a Recently Emerged, Highly Transmissible, Multi-Antibiotic- and Antiseptic-Resistant Variant of Methicillin-Resistant Staphylococcus aureus, Sequence Type 239 (TW). JB. 2009;192: 888–892. doi:10.1128/jb.01255-09
23 Cong Y, Chan Y, Ragan MA. A novel alignment-free method for detection of lateral genetic transfer based on TF-IDF. Sci Rep. 2016;6. doi:10.1038/srep30308
24 Castillo-Ramírez S, Harris SR, Holden MTG, He M, Parkhill J, Bentley SD, et al. The Impact of Recombination on dN/dS within Recently Emerged Bacterial Clones. Balloux F, editor. PLoS Pathog. 2011;7: e1002129. doi:10.1371/journal.ppat.1002129
25 Harkins CP, Pichon B, Doumith M, Parkhill J, Westh H, Tomasz A, et al. Methicillin-resistant Staphylococcus aureus emerged long before the introduction of methicillin into clinical practice. Genome Biol. 2017;18. doi:10.1186/s13059-017-1252-9
26 Baltrus DA. Exploring the costs of horizontal gene transfer. Trends in Ecology & Evolution. 2013;28: 489–495. doi:10.1016/j.tree.2013.04.002
27 San Millan A, MacLean RC. Fitness Costs of Plasmids: A Limit to Plasmid Transmission. Microbial Transmission. American Society of Microbiology; 2019. pp. 65–79. doi:10.1128/microbiolspec.mtbp-0016-2017
28 Brockhurst MA, Harrison E, Hall JPJ, Richards T, McNally A, MacLean C. The Ecology and Evolution of Pangenomes. Current Biology. 2019;29: R1094–R1103. doi:10.1016/j.cub.2019.08.012
29 Vogwill T, MacLean RC. The genetic basis of the fitness costs of antimicrobial resistance: a meta-analysis approach. Evol Appl. 2014;8: 284–295. doi:10.1111/eva.12202
30 Porse A, Schou TS, Munck C, Ellabaan MMH, Sommer MOA. Biochemical mechanisms determine the functional compatibility of heterologous genes. Nat Commun. 2018;9. doi:10.1038/s41467-018-02944-3
31 Yang Q, Li M, Spiller OB, Andrey DO, Hinchliffe P, Li H, et al. Balancing mcr-1 expression and bacterial survival is a delicate equilibrium between essential cellular defence mechanisms. Nat Commun. 2017;8. doi:10.1038/s41467-017-02149-0

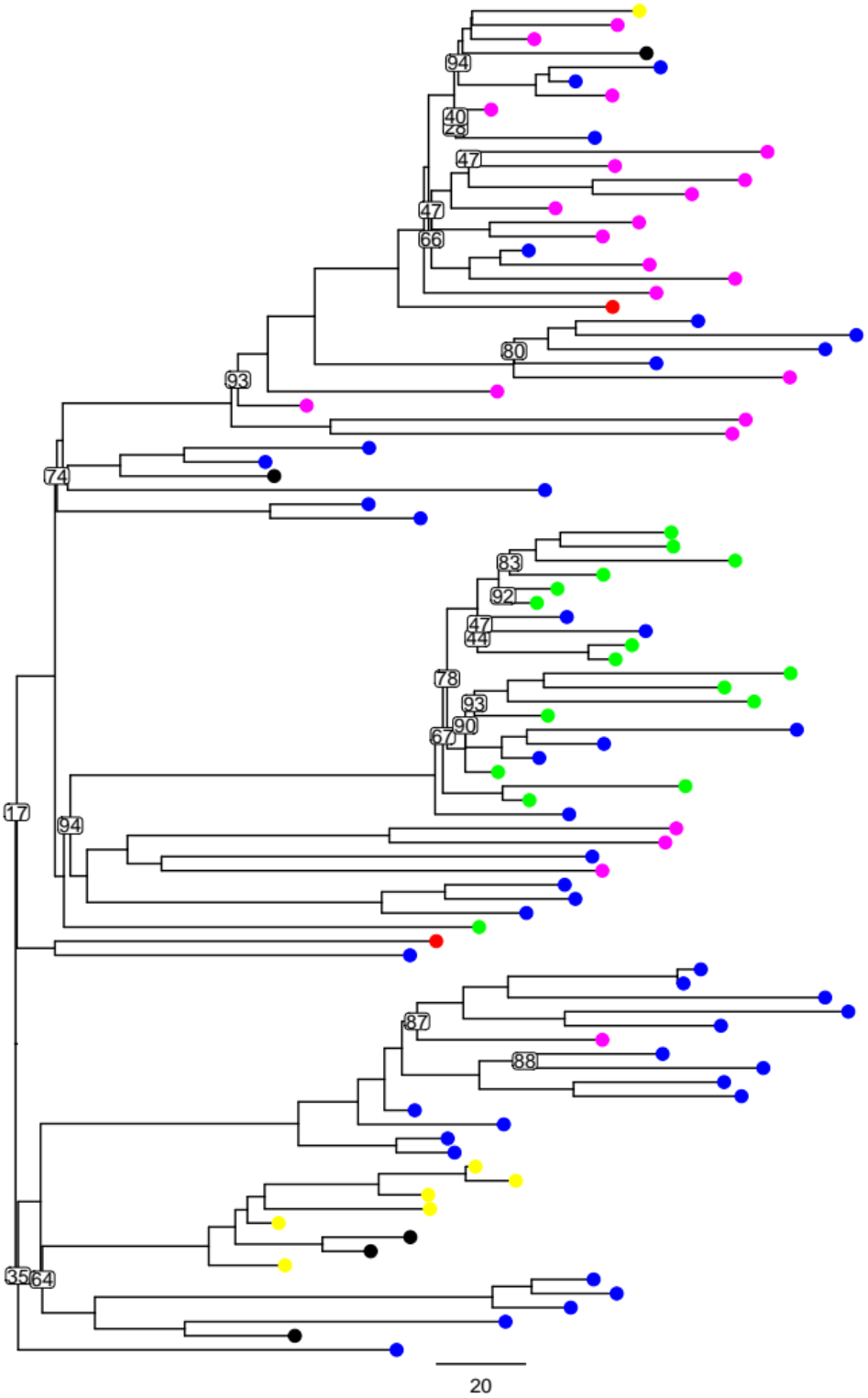
32 Lee SM, Ender M, Adhikari R, Smith JMB, Berger-Bächi B, Cook GM. Fitness Cost of Staphylococcal Cassette Chromosome mec in Methicillin-Resistant Staphylococcus aureus by Way of Continuous Culture. AAC. 2007;51: 1497–1499. doi:10.1128/aac.01239-06
33 Knight GM, Budd EL, Whitney L, Thornley A, Al-Ghusein H, Planche T, et al. Shift in dominant hospital-associated methicillin-resistant Staphylococcus aureus (HA-MRSA) clones over time. Journal of Antimicrobial Chemotherapy. 2012;67: 2514–2522. doi:10.1093/jac/dks245
34 Gray RR, Tatem AJ, Johnson JA, Alekseyenko AV, Pybus OG, Suchard MA, et al. Testing Spatiotemporal Hypothesis of Bacterial Evolution Using Methicillin-Resistant Staphylococcus aureus ST239 Genome-wide Data within a Bayesian Framework. Molecular Biology and Evolution. 2010;28: 1593–1603. doi:10.1093/molbev/msq319
35 Hsu L-Y, Harris SR, Chlebowicz MA, Lindsay JA, Koh T-H, Krishnan P, et al. Evolutionary dynamics of methicillin-resistant Staphylococcus aureus within a healthcare system. Genome Biol. 2015;16. doi:10.1186/s13059-015-0643-z
36 Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK, Chantratita N, et al. Evolution of MRSA During Hospital Transmission and Intercontinental Spread. Science. 2010;327: 469–474. doi:10.1126/science.1182395
37 Smyth DS, McDougal LK, Gran FW, Manoharan A, Enright MC, Song J-H, et al. Population Structure of a Hybrid Clonal Group of Methicillin-Resistant Staphylococcus aureus, ST239-MRSA-III. DeLeo FR, editor. PLoS ONE. 2010;5: e8582. doi:10.1371/journal.pone.0008582
38 Baines SL, Holt KE, Schultz MB, Seemann T, Howden BO, Jensen SO, et al. Convergent Adaptation in the Dominant Global Hospital Clone ST239 of Methicillin-Resistant Staphylococcus aureus. Gilligan P, Fowler V, editors. mBio. 2015;6. doi:10.1128/mbio.00080-15
39 Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Research. 2014;43: e15–e15. doi:10.1093/nar/gku1196
40 Castillo-Ramírez S, Corander J, Marttinen P, Aldeljawi M, Hanage WP, Westh H, et al. Phylogeographic variation in recombination rates within a global clone of methicillin-resistant Staphylococcus aureus. Genome Biol. 2012;13: R126. doi:10.1186/gb-2012-13-12-r126
41 Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution. 2012;29: 1969–1973. doi:10.1093/molbev/mss075
42 Knox R. A New Penicillin (BRL 1241) Active Against Penicillin-resistant Staphylococci. BMJ. 1960;2: 690–693. doi:10.1136/bmj.2.5200.690
43 Petit RA III, Read TD. Staphylococcus aureus viewed from the perspective of 40,000+ genomes. PeerJ. 2018;6: e5261. doi:10.7717/peerj.5261
44 Bradley P, den Bakker HC, Rocha EPC, McVean G, Iqbal Z. Ultrafast search of all deposited bacterial and viral genomic data. Nat Biotechnol. 2019;37: 152–159. doi:10.1038/s41587-018-0010-1
45 van Tonder AJ, Mistry S, Bray JE, Hill DMC, Cody AJ, Farmer CL, et al. Defining the Estimated Core Genome of Bacterial Populations Using a Bayesian Decision Model. Ouzounis CA, editor. PLoS Comput Biol. 2014;10: e1003788. doi:10.1371/journal.pcbi.1003788
46 Robinson DA, Kearns AM, Holmes A, Morrison D, Grundmann H, Edwards G, et al. Re-emergence of early pandemic Staphylococcus aureus as a community-acquired meticillin-resistant clone. The Lancet. 2005;365: 1256–1258. doi:10.1016/s0140-6736(05)74814-5
47 Didelot X, Wilson DJ. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. Prlic A, editor. PLoS Comput Biol. 2015;11: e1004041. doi:10.1371/journal.pcbi.1004041
48 Nübel U, Dordel J, Kurt K, Strommenger B, Westh H, Shukla SK, et al. A Timescale for Evolution, Population Expansion, and Spatial Spread of an Emerging Clone of Methicillin-Resistant Staphylococcus aureus. Koehler T, editor. PLoS Pathog. 2010;6: e1000855. doi:10.1371/journal.ppat.1000855
49 Alam MT, Read TD, Petit RA III, Boyle-Vavra S, Miller LG, Eells SJ, et al. Transmission and Microevolution of USA300 MRSA in U.S. Households: Evidence from Whole-Genome Sequencing. Pettigrew MM, editor. mBio. 2015;6. doi:10.1128/mbio.00054-15

50 McAdam PR, Templeton KE, Edwards GF, Holden MTG, Feil EJ, Aanensen DM, et al. Molecular tracing of the emergence, adaptation, and transmission of hospital-associated methicillin-resistant Staphylococcus aureus. Proceedings of the National Academy of Sciences. 2012;109: 9107–9112. doi:10.1073/pnas.1202869109
51 McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 1991;351: 652–654. doi:10.1038/351652a0
52 Rand DM, Kann LM. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. Molecular Biology and Evolution. 1996;13: 735–748. doi:10.1093/oxfordjournals.molbev.a025634
53 Hershberg R, Petrov DA. Evidence That Mutation Is Universally Biased towards AT in Bacteria. Nachman MW, editor. PLoS Genet. 2010;6: e1001115. doi:10.1371/journal.pgen.1001115
54 Marvig RL, Sommer LM, Molin S, Johansen HK. Convergent evolution and adaptation of Pseudomonas aeruginosa within patients with cystic fibrosis. Nat Genet. 2014;47: 57–64. doi:10.1038/ng.3148
55 Woods R, Schneider D, Winkworth CL, Riley MA, Lenski RE. Tests of parallel molecular evolution in a long-term experiment with Escherichia coli. Proceedings of the National Academy of Sciences. 2006;103: 9107–9112. doi:10.1073/pnas.0602917103
56 Gifford DR, Furió V, Papkou A, Vogwill T, Oliver A, MacLean RC. Identifying and exploiting genes that potentiate the evolution of antibiotic resistance. Nat Ecol Evol. 2018;2: 1033–1039. doi:10.1038/s41559-018-0547-x
57 Jevons MP, Parker MT. The evolution of new hospital strains of Staphylococcus aureus. Journal of Clinical Pathology. 1964;17: 243–250. doi:10.1136/jcp.17.3.243
58 Ito T, Katayama Y, Asada K, Mori N, Tsutsumimoto K, Tiensasitorn C, et al. Structural Comparison of Three Types of Staphylococcal Cassette Chromosome mec Integrated in the Chromosome in Methicillin-Resistant Staphylococcus aureus. Antimicrob Agents Chemother. 2001;45: 1323–1336. doi:10.1128/aac.45.5.1323-1336.2001
59 Wales A, Davies R. Co-Selection of Resistance to Antibiotics, Biocides and Heavy Metals, and Its Relevance to Foodborne Pathogens. Antibiotics. 2015;4: 567–604. doi:10.3390/antibiotics4040567
60 Andersson DI, Hughes D. Antibiotic resistance and its cost: is it possible to reverse resistance? Nat Rev Microbiol. 2010;8: 260–271. doi:10.1038/nrmicro2319
61 Melnyk AH, Wong A, Kassen R. The fitness costs of antibiotic resistance mutations. Evol Appl. 2014;8: 273–283. doi:10.1111/eva.12196
62 Nielsen KL, Pedersen TM, Udekwu KI, Petersen A, Skov RL, Hansen LH, et al. Fitness cost: a bacteriological explanation for the demise of the first international methicillin-resistant Staphylococcus aureus epidemic. Journal of Antimicrobial Chemotherapy. 2012;67: 1325–1332. doi:10.1093/jac/dks051
63 Allen RC, Angst DC, Hall AR. Resistance Gene Carriage Predicts Growth of Natural and ClinicalEscherichia coliIsolates in the Absence of Antibiotics. Liu S-J, editor. Appl Environ Microbiol. 2018;85. doi:10.1128/aem.02111-18
64 MacLean RC, Vogwill T. Limits to compensatory adaptation and the persistence of antibiotic resistance in pathogenic bacteria. Evolution, Medicine, and Public Health. 2014;2015: 4–12. doi:10.1093/emph/eou032
65 Planet PJ. Life After USA300: The Rise and Fall of a Superbug. The Journal of Infectious Diseases. 2017;215: S71–S77. doi:10.1093/infdis/jiw444
66 Song J-H, Hsueh P-R, Chung DR, Ko KS, Kang C-I, Peck KR, et al. Spread of methicillin-resistant Staphylococcus aureus between the community and the hospitals in Asian countries: an ANSORP study. Journal of Antimicrobial Chemotherapy. 2011;66: 1061–1069. doi:10.1093/jac/dkr024
67 McNally A, Oren Y, Kelly D, Pascoe B, Dunn S, Sreecharan T, et al. Combined Analysis of Variation in Core, Accessory and Regulatory Genome Regions Provides a Super-Resolution View into the Evolution of Bacterial Populations. Hughes D, editor. PLoS Genet. 2016;12: e1006280. doi:10.1371/journal.pgen.1006280

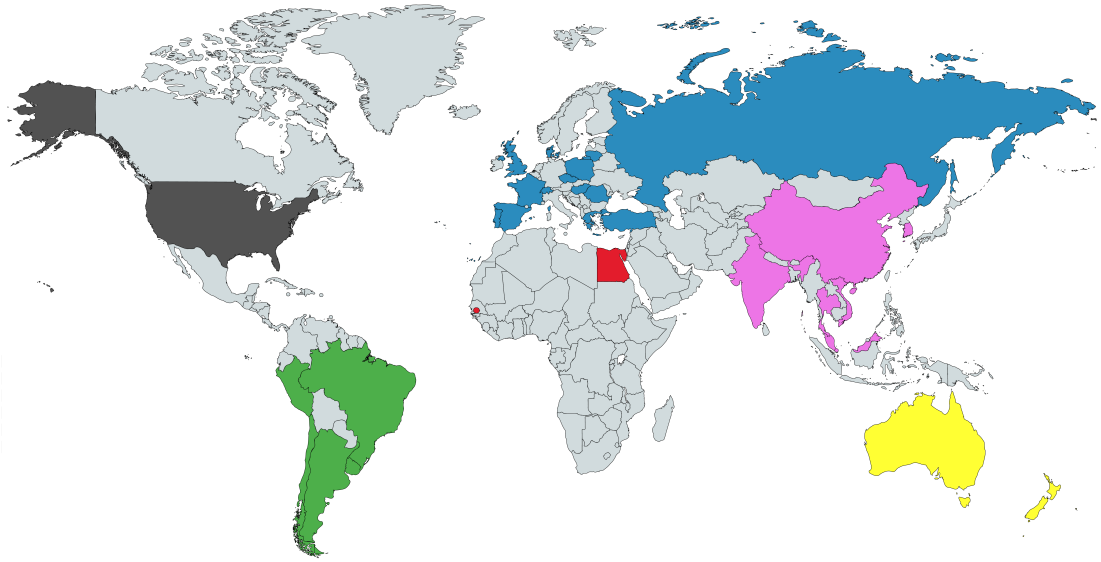
68 Nagaev I, Bjorkman J, Andersson DI, Hughes D. Biological cost and compensatory evolution in fusidic acid-resistant Staphylococcus aureus. Mol Microbiol. 2001;40: 433–439. doi:10.1046/j.1365-2958.2001.02389.x
69 Comas I, Borrell S, Roetzer A, Rose G, Malla B, Kato-Maeda M, et al. Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. Nat Genet. 2011;44: 106–110. doi:10.1038/ng.1038
70 Didelot X, Maiden MCJ. Impact of recombination on bacterial evolution. Trends in Microbiology. 2010;18: 315–322. doi:10.1016/j.tim.2010.04.002
71 Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA. Database indexing for production MegaBLAST searches. Bioinformatics. 2008;24: 1757–1764. doi:10.1093/bioinformatics/btn322
72 Homer N. DWGSIM. 2011. Available from: https://github.com/nh13/DWGSIM/wiki
73 Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30: 2114–2120. doi:10.1093/bioinformatics/btu170
74 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25: 1754–1760. doi:10.1093/bioinformatics/btp324
75 Kaya H, Hasman H, Larsen J, Stegger M, Johannesen TB, Allesøe RL, et al. SCCmecFinder, a Web-Based Tool for Typing of Staphylococcal Cassette Chromosome mec in Staphylococcus aureus Using Whole-Genome Sequence Data. Limbago BM, editor. mSphere. 2018;3. doi:10.1128/msphere.00612-17
76 Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res. 2018;3: 124. doi:10.12688/wellcomeopenres.14826.1
77 Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, Keane JA, et al. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. Microbial Genomics. 2017;3. doi:10.1099/mgen.0.000131
78 Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30: 1312–1313. doi:10.1093/bioinformatics/btu033
79 Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30: 2068–2069. doi:10.1093/bioinformatics/btu153
80 Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31: 3691–3693. doi:10.1093/bioinformatics/btv421
81 Didelot X, Croucher NJ, Bentley SD, Harris SR, Wilson DJ. Bayesian inference of ancestral dates on bacterial phylogenetic trees. Nucleic Acids Research. 2018;46: e134–e134. doi:10.1093/nar/gky783
82 Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microbial Genomics. 2016;2. doi:10.1099/mgen.0.000056
83 Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics. 2000;16: 276–277. doi:10.1016/s0168-9525(00)02024-2
84 Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology. 2012;19: 455–477. doi:10.1089/cmb.2012.0021
85 Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics. 2009;25: 1968–1969. doi:10.1093/bioinformatics/btp347
86 The European Committee on Antimicrobial Susceptibility Testing. Breakpoint tables for interpretation of MICs and zone diameters version 4.0. 2014. Available from: http://www.eucast.org

Figure 1

A



B



C

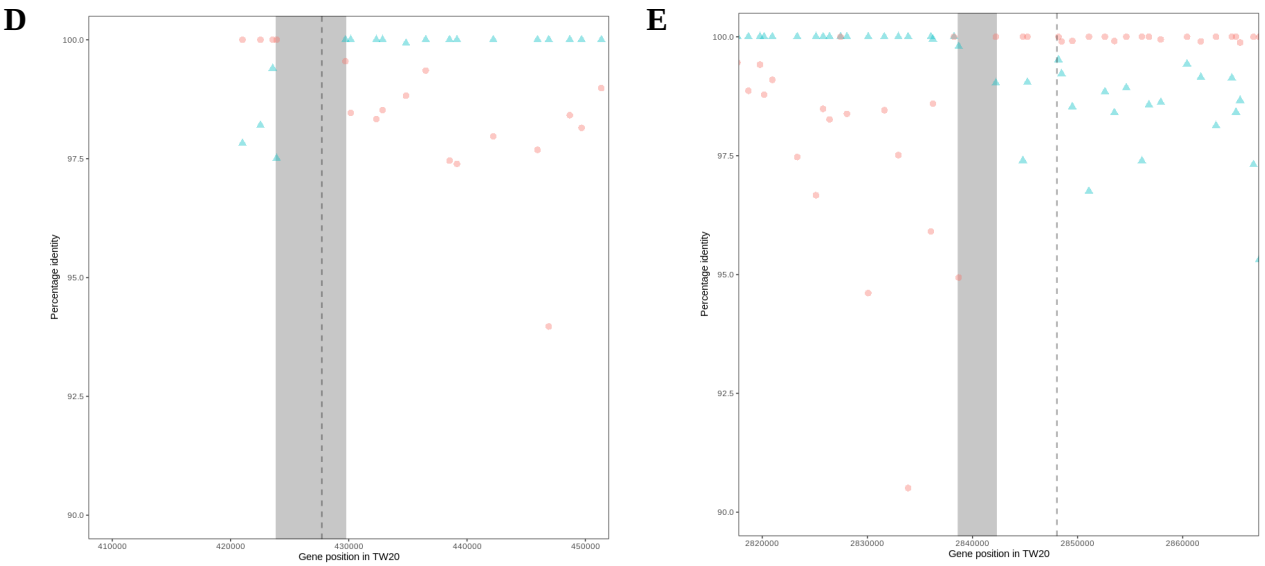
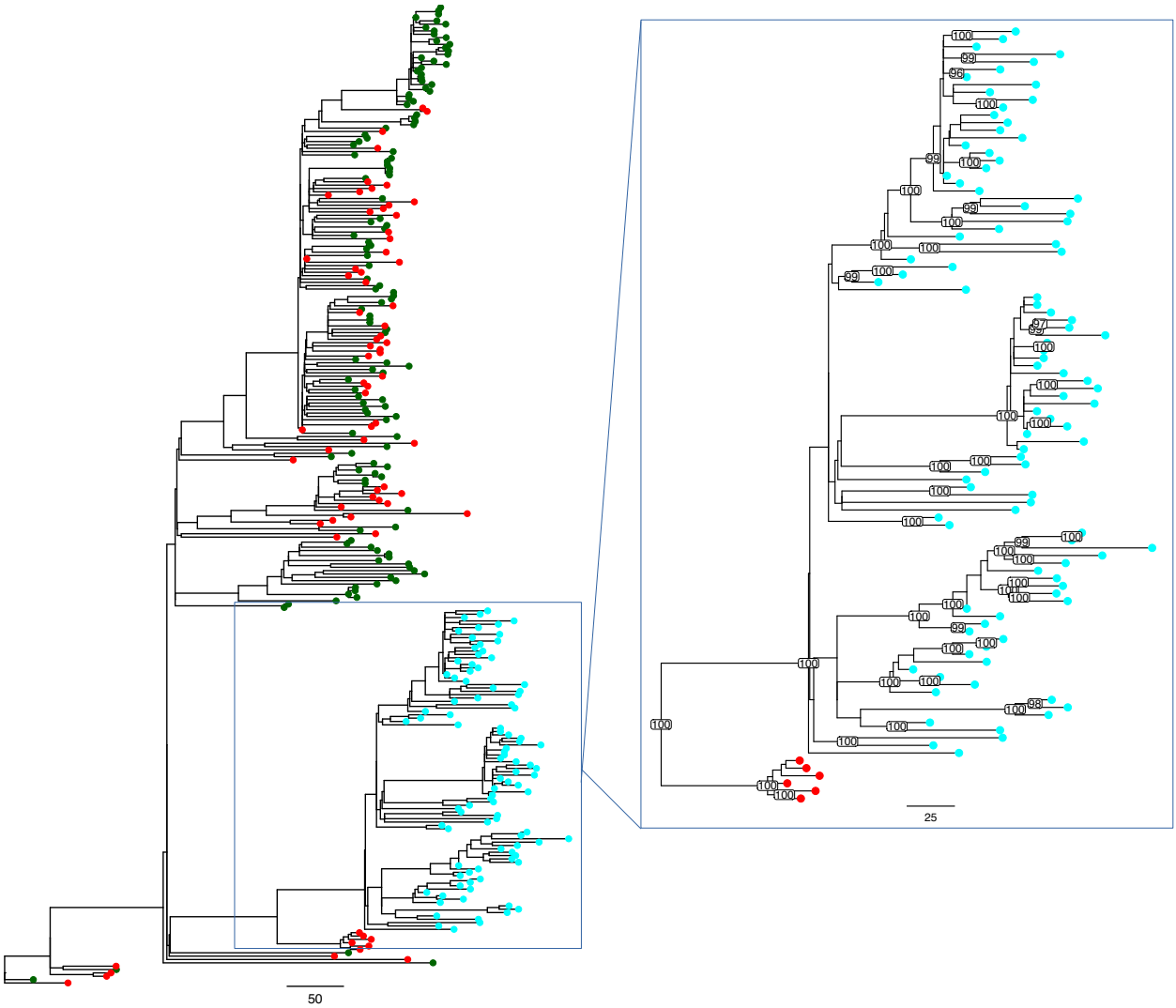


Figure 2

A



B

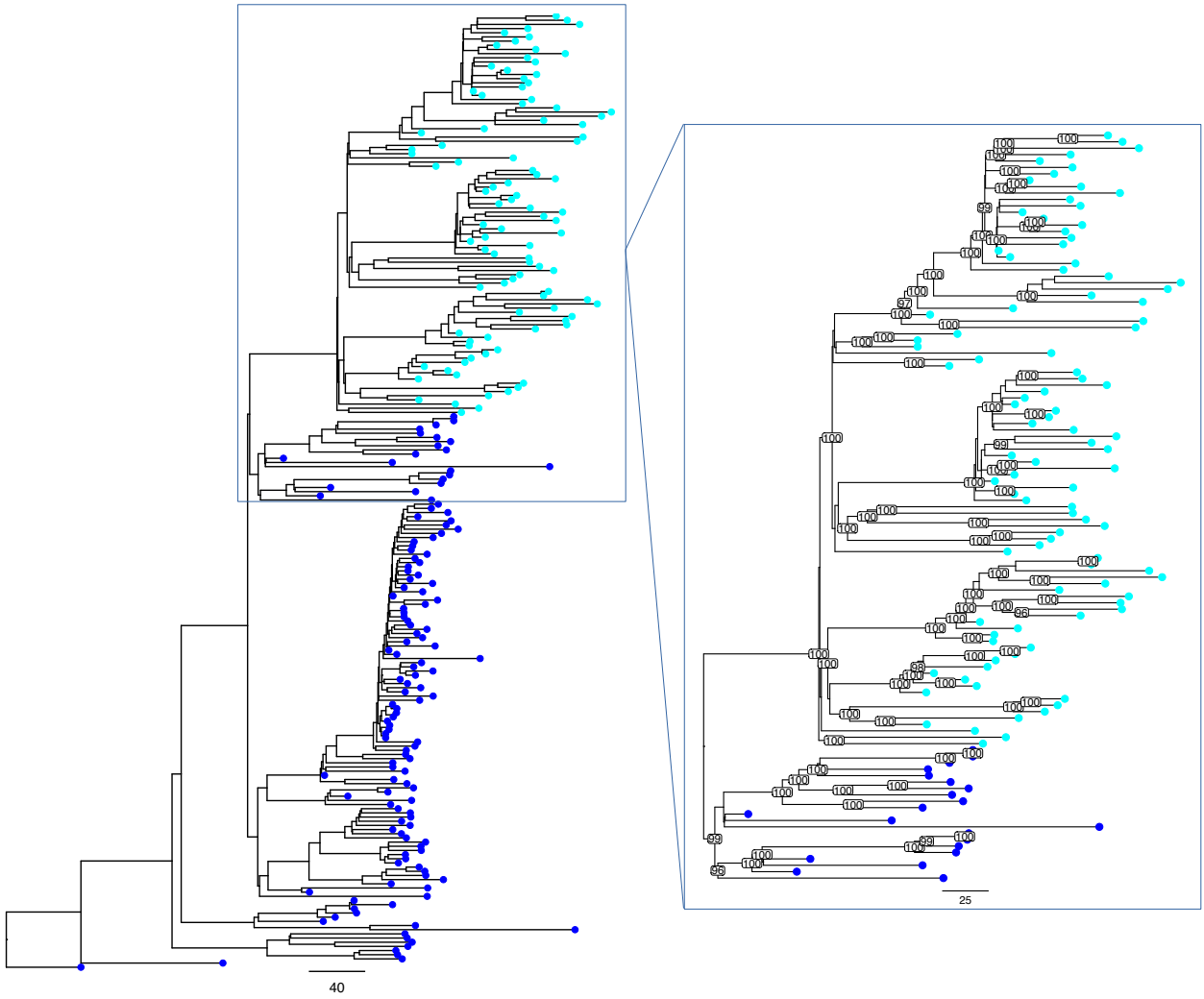


Figure 3

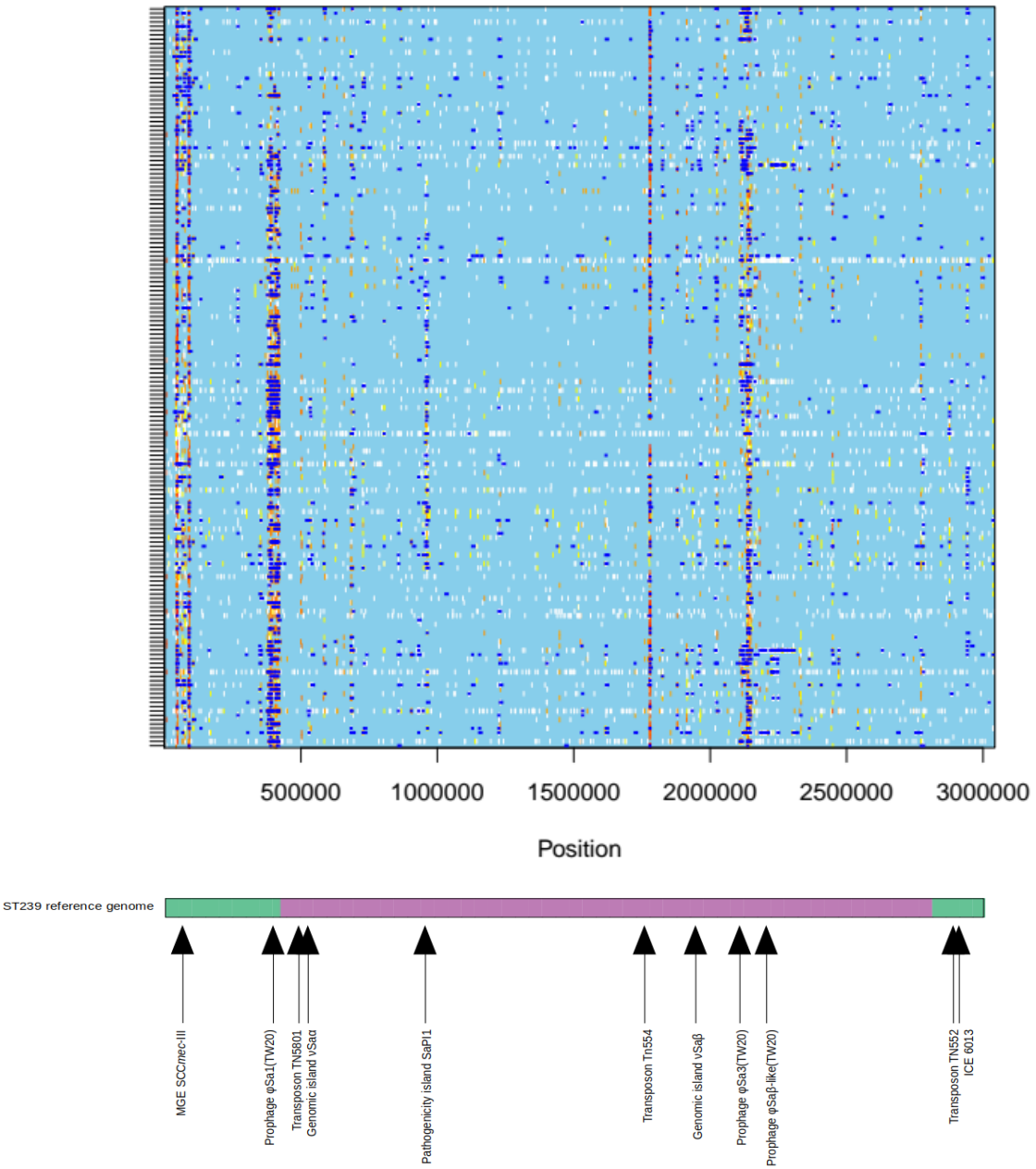


Figure 4

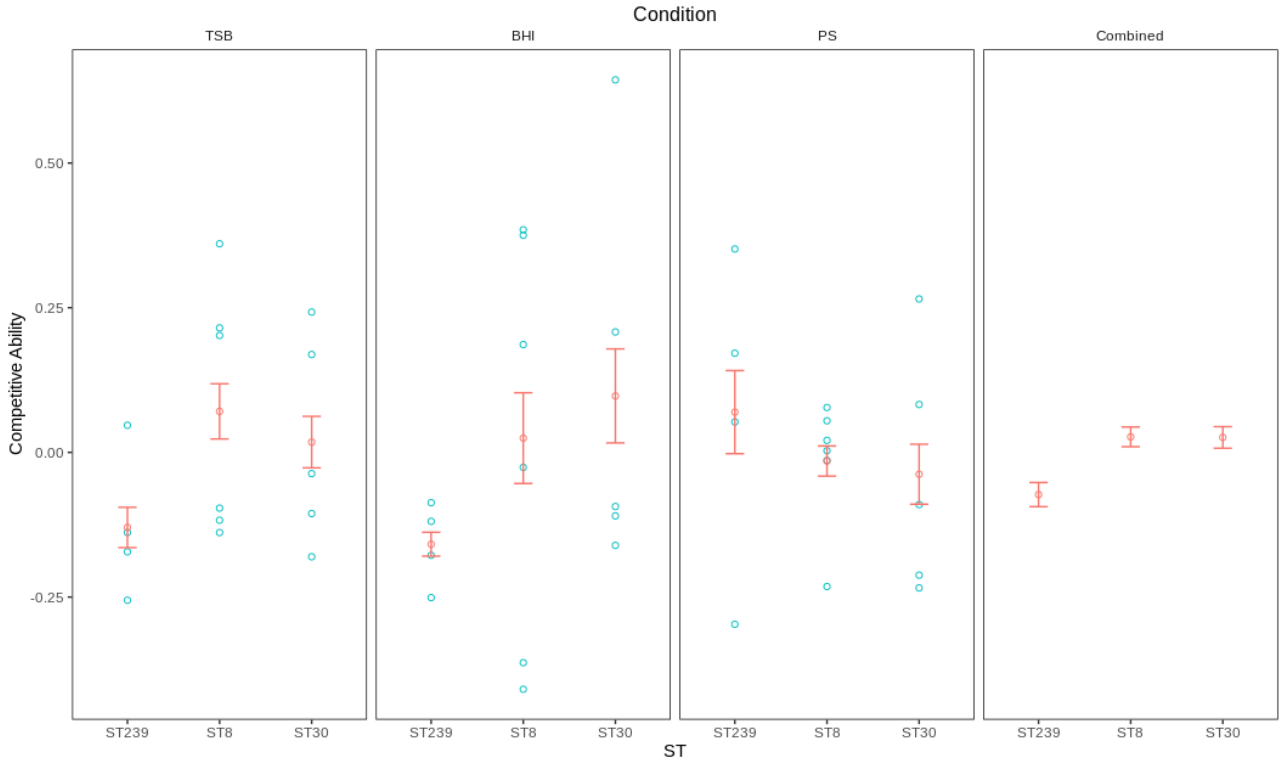
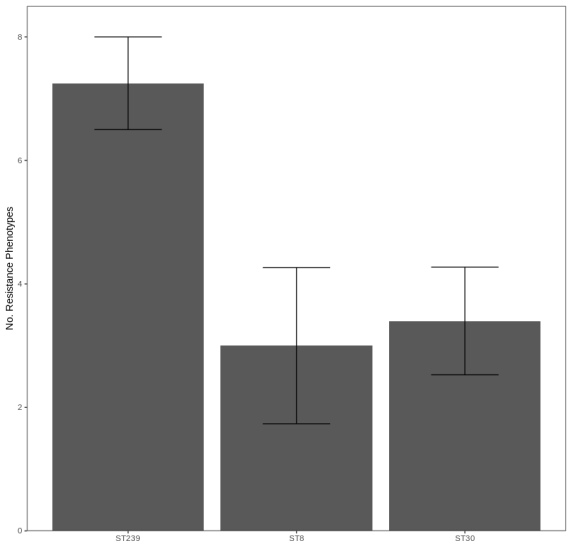


Figure 5

A



B

Isolate	Aminoglycoside (gentamycin)	β-lactam (ampicillin/penicillin)	β-lactam (oxacillin)	Carboxylic Acid (mupirocin)	Cephalosporin (ceftarolin)	Fosfomycin	Fusidic Acid	Glycopeptide (teicoplanin)	Glycopeptide (vancomycin)	Glycylcycline (tigecyclin)	Macrolide (erythromycin)	Oxazolidinone (linezolid)	Quinolone (moxifloxacin)	Rfamycin (rifampicin)	Streptogramin (synercid)	Diaminopyrimidine (trim/sulf)
ST239A	R	R	R	R	S	R	S	S	S	S	R	S	R	R	S	R
ST239B	R	R	R	S	S	R	S	S	S	S	R	S	R	R	S	R
ST239C	R	R	R	S	S	S	S	S	S	S	R	S	R	S	S	R
ST239D	R	R	R	S	S	S	S	S	S	S	R	S	R	S	S	R
ST8A	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
ST8B	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
ST8C	S	R	R	S	S	S	S	S	S	S	S	S	R	S	S	S
ST8D	S	R	R	S	S	S	S	S	S	S	R	S	R	S	S	R
ST8E	R	R	R	S	S	S	S	R	R	S	R	S	R	S	S	R
ST8F	S	R	S	S	S	S	R	S	S	S	S	S	S	S	S	S
ST30A	S	R	R	S	S	S	S	S	S	S	S	S	S	S	S	S
ST30B	S	R	S	S	S	S	S	S	S	S	S	S	S	S	S	S
ST30C	S	R	R	S	S	S	S	S	S	S	R	S	R	S	S	S
ST30D	S	R	R	S	S	S	S	S	S	S	R	S	R	S	S	S
ST30E	R	R	R	S	S	S	S	S	S	S	R	S	R	S	S	R

Table 1

Region	Comparison	D_n	D_s	D_n/D_s	P_n	P_s	P_n/P_s	*N*	Selection
Acquired	ST239/ST30	40	32	1.25	1406	740	1.90	1.52	Negative
Backbone	ST239/ST8	42	50	0.84	6080	3508	1.73	2.06	Negative

Table 2

	Total number of substitutions	
	Acquired region	**Backbone region**
GC → AT	795	2,710
AT → GC	474	1,864

Table 3

Effect	df	Sum of squares	Mean squares	F-value	*p*-value
ST	2	0.260	0.130	8.33	0.0004
Media x ST	4	0.565	0.141	9.05	<0.0001
Isolate[ST]	12	4.24	0.354	22.6	<0.0001
Error	116	1.81	0.0156	-	-