PhD rotation project 3

Using homoplasy to study recent selection pressures in *Staphylococcus aureus*

Benjamin Y. H. Bai[†]

Supervisors: Simon R. Harris, Dorota M. Jamrozy, and Julian Parkhill

Homoplasies in phylogenetic reconstruction arise due to character similarity that is not due to common ancestry (i.e. homology). Processes that introduce homoplasy include positive selection, recombination, and movement of mobile genetic elements (MGEs)[1]. In bacterial species with relatively low rates of recombination such as *Staphylococcus aureus*[2], homoplasic sites in a phylogeny constructed from the core genome can indicate strong selective pressures, especially when convergent evolution in independent lineages is observed.

S. aureus is ubiquitous, colonising approximately 20-30% of the human population. Multiple drug resistant strains (methicillin-resistant *S. aureus*, MRSA) cause difficult-to-treat infections in both healthcare-associated (HA-MRSA) and community-associated (CA-MRSA) settings[3]. In a previous analysis of 63 *S. aureus* genomes from multi-locus sequence type (ST) 239, around a quarter of homoplasic sites identified were directly linked to antibiotic resistance[4]. Other sites occurred in intergenic regions and uncharacterised loci, suggesting auxiliary selective forces may be at play.

In this study, I looked for evidence of homoplasy in a large global *S. aureus* dataset (3209 isolates from ST22, ST8, and ST239) from the British Society for Antimicrobial Chemotherapy (BSAC)[5] and Pfizer[*] collections. Homoplasic sites were rare, and most detected homoplasies were convergences, with many close to, or within known antimicrobial resistance (AMR) genes. There was an excess of homoplasies in intergenic regions, and Gene Ontology (GO) enrichment of the downstream loci revealed functions relevant to AMR and virulence. This suggests that regulatory variants in these regions under strong convergent selection pressures may be important for the evolution of *S. aureus* pathogenicity.

Detecting homoplasic sites

For each ST, a maximum likelihood phylogenetic tree was built based on an alignment of SNP sites in the core genome. For each site, characters (SNP state) were mapped onto the tree using maximum parsimony. Both accelerated transformation (ACCTRAN) and delayed transformation (DELTRAN) character optimisation algorithms were applied to resolve ambiguous reconstructions, producing two most parsimonious reconstructions (MPRs), each containing the minimum number of changes required to explain the data[6]. The two MPRs represent extremes: ACCTRAN assigns changes as close to the root as possible (maximising reversals i.e. reversion of an ancestral change of state), whereas DELTRAN assigns changes as close to the tips as possible (maximising convergences i.e. independent change to the same state)[6].

Homoplasic sites (with at least one convergence or reversal) were distributed throughout the genome (Appendix, Fig. 4), and relatively uncommon (Table 1)–consistent with a low recombination rate. The recently emerged ST239[7] had the least amount of homoplasy overall. Of the 38 homoplasic sites previously identified in ST239[4], 34 were reproduced in this analysis.

There was no ambiguity in reconstruction for most sites, with the ACCTRAN and DELTRAN algorithms agreeing over 90% of the time (Table 1). The majority of homoplasies were due to convergent changes, suggesting convergence under

[†] Wellcome Trust Sanger Institute, Hinxton, UK. *Project duration*: Mar-Jun 2017. *Contact*: bb9@sanger.ac.uk.

[*] Unpublished collection. See Materials and Methods.

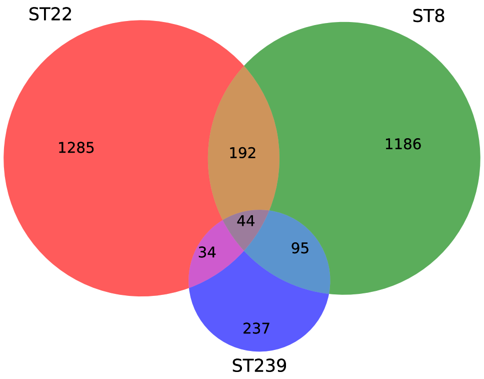

Figure 1: Numbers of sites homoplasic in multiple STs.

similar selection pressures is the dominant evolutionary force.

Homoplasic SNPs in multiple STs

Evidence for convergent evolution is strongest when homoplasies are observed at the same site in multiple STs. I identified 365 sites that were homoplasic in two or more STs and 44 sites that were homoplasic in all three STs (Fig. 1).

Sites were associated with the closest locus in the reference clone's genome annotation, then additionally annotated with a set of 129 known AMR loci. Excluding sites with a known AMR annotation, 25 sites were homoplasic in all three STs. These sites occurred near or within loci related to virulence, membrane transport, and the cell wall (Table 2). Homoplasy was also detected within tRNA genes, which are relevant antibiotic targets[10], although the high density observed suggests the signal may be spurious, perhaps caused by gene duplication/deletion within the tRNA array.

Example: homoplasies in a convergent intergenic region

Figure 2 shows the region around position 2462688 in ST22 (a site homoplasic in all three STs, from Table 2). The site is 227 bp downstream of SAEMRSA1522820 (putative exported protein). The distribution of homoplasic changes on the tree is sporadic, showing nine convergent G to T mutations in independent lineages (Appendix, Fig. 5)–unlikely to be due to recombination. The region is repetitive, and GC-rich in all three frames; short, GC-rich repeat elements can have regulatory significance in *S. aureus*[19].

Excess homoplasy in intergenic regions

A substantial number of homoplasic sites occurred in intergenic regions. The location of homoplasic SNP sites relative to the nearest annotated locus is shown in Figure 3. There was an excess of homoplasy in intergenic regions, both for sites homoplasic in at least one ST (Fisher's exact test, p $= 1.846 \times 10^{-167}$), and for sites homoplasic in more than one ST (p $= 3.440 \times 10^{-56}$).

Homoplasic sites upstream of loci may fall within cis-regulatory regions. To characterise the loci potentially regulated by these sites, known AMR loci were excluded (enriching for other selective pressures), and GO enrichment was performed on the remaining loci that had a homoplasic site within 200 bp upstream (Table 3)–transcriptional regulators can bind over 100 bp upstream[21].

Some residual enrichment for AMR remained (e.g. drug transport, drug transmembrane transporter activity), confirming that antimicrobials are a dominating selection pressure. Functions relating to adhesion and the cell surface (cell adhesion, cell wall) are important for pathogenesis, and are subject to selection pressures imposed by the host immune response[22]. Other terms related to DNA repair (base-excision repair, alkylbase DNA N-glycosylase activity) may be associated with resilience to oxidative stress[23]. Some enriched terms seem to themselves be related to regulation of both transcription (regulation of transcription, DNA-templated; sigma factor activity) and translation (translation, structural constituent of ribosome). Further investigation of the specific loci within these ontology terms could generate more specific hypotheses about the selection pressures present, especially if these loci are downstream of homoplasic sites that lie within known regulatory motifs.

When conducting such investigations of individual candidate loci, factors that cause homoplasy unrelated to selection should be considered. Although recombination is relatively rare in *S. aureus*, any recombination that does exist will produce a strong homoplasic signal. In coding regions, clusters of homoplasic sites containing a mix of synonymous and non-synonymous changes is a pattern more likely due to recombination than selection. Another indication is homoplasic changes being distributed in single clades, rather than in independent lineages around the tree (e.g.

Table 1: Summary of homoplasic SNP sites in each ST. *Agreement*: sites for which ACCTRAN and DELTRAN MPRs assigned the same number of convergences and reversals. *Rescaled consistency index (RCI)*: varies between zero (maximum homoplasy) and one (no homoplasy).

ST	Reference clone	Num. isolates	SNPs	Homoplasic	Convergences ACCTRAN, DELTRAN	Reversals ACCTRAN, DELTRAN	Agreement (%)	RCI
ST22	UK-EMRSA-15[8]	1744	28315	1555	1463, 1520	156, 61	1457 (93.70)	0.8215
ST8	ST8-MRSA-IV USA300[9]	758	28887	1517	1432, 1485	229, 86	1366 (90.05)	0.8579
ST239	ST239-MRSA-III TW20[7]	707	14884	410	405, 408	36, 13	380 (92.68)	0.9393

Table 2: Sites homoplasic in all three STs, with distance to nearest locus (negative distance indicates upstream) and locus annotations shown for each ST. Only the ST22 annotation contained non-coding RNA (ncRNA) annotations–most sRNAs themselves are present in all three STs.

Site pos. in ref. clone (ST22, ST8, ST239)	Dist. to nearest locus (ST22, ST8, ST239)	Name	Description (ST22, ST8, ST239)	Comments
(14600, 14619, 14583)	(-130, -130, -130)	-	(putative membrane protein, putative membrane protein, putative membrane protein)	BLASTP hit: putative azaleucine resistance protein AzlC.
(98571, 128488, 133735)	(0, 0, 0)	*spa*	(immunoglobulin G binding protein A precursor (pseudogene), immunoglobulin G binding protein A precursor, immunoglobulin G binding protein A precursor)	Protein A, virulence-related[11].
(697113, 735191, 805532) (697117, 735195, 805536)	(-52, -52, -52) (-48, -48, -48)	-	(putative sugar efflux transporter, sugar efflux transporter, putative sugar efflux transporter)	Pfam hit: PF07690, Major Facilitator Superfamily, small solute transporter.
(1040617, 992397, 1111443) (1040674, 992454, 1111500)	(-65, -182, -65) (-122, -239, -122)	*sspA*	(V8 protease, glutamyl endopeptidase precursor, glutamyl endopeptidase)	serine protease, virulence-related[12].
(1202630, 1249109, 1321491)	(-46, -13, -46)	*lytN*	(putative cell wall hydrolase, cell wall hydrolase, putative cell wall hydrolase (pseudogene))	murein hydrolase, cell wall remodelling[13].
(1945349, 1995298, 2019055) (1945354, 1995303, 2019060) (1945363, 1995312, 2019069) (1945366, 1995315, 2019072) (1945373, 1995322, 2019079) (1945381, 1995330, 2019087)	(0, 0, -72) (0, 0, -77) (0, 0, -86) (0, 0, -89) (0, 0, -96) (0, 0, -104)	tRNA-Leu	(tRNA_47, tRNA-OTHER, transfer RNA-Leu)	tRNA-Leu.
(1945424, 1995373, 2019130) (1945437, 1995386, 2019143) (1945442, 1995391, 2019148) (1945451, 1995400, 2019157) (1945454, 1995403, 2019160) (1945461, 1995410, 2019167) (1945469, 1995418, 2019175)	(0, -34, 73) (0, -47, 60) (0, -52, 55) (0, 43, 46) (0, 40, 43) (0, 33, 36) (0, 25, 28)	tRNA-Gly	(tRNA_50, tRNA-OTHER, transfer RNA-Gly)	tRNA-Gly.
(1945514, 1995463, 2019220)	(0, 0, 0)	-	(transfer RNA-Gly, tRNA-Gly, transfer RNA-Gly)	tRNA-Gly.
(1979551, 2029561, 2053467)	(0, -238, -290)	*map*	(*rsaOF*, methionine aminopeptidase, putative membrane protein)	*rsaOF* ncRNA[14] (only annotated in ST22). Downstream gene: *map*, potential host immunomodulator[15].
(2004964, 2055031, 2078819)	(-51, -51, -51)	*scpA*	(staphopain protease, staphopain A, staphopain protease)	Staphopain A, virulence-related[16].
(2106494, 2146330, 2171699)	(0, 242, 242)	hld	(*RNAIII*, delta-hemolysin precursor, delta-hemolysin precursor)	*RNAIII* ncRNA, regulation of virulence factors[17] (only annotated in ST22). Upstream gene: *hld*, delta-hemolysin, virulence-related[18].
(2462688, 2504739, 2657370)	(227, 193, 193)	-	(putative exported protein, conserved hypothetical protein, putative exported protein)	BLASTP hit: DUF4889 domain-containing protein.

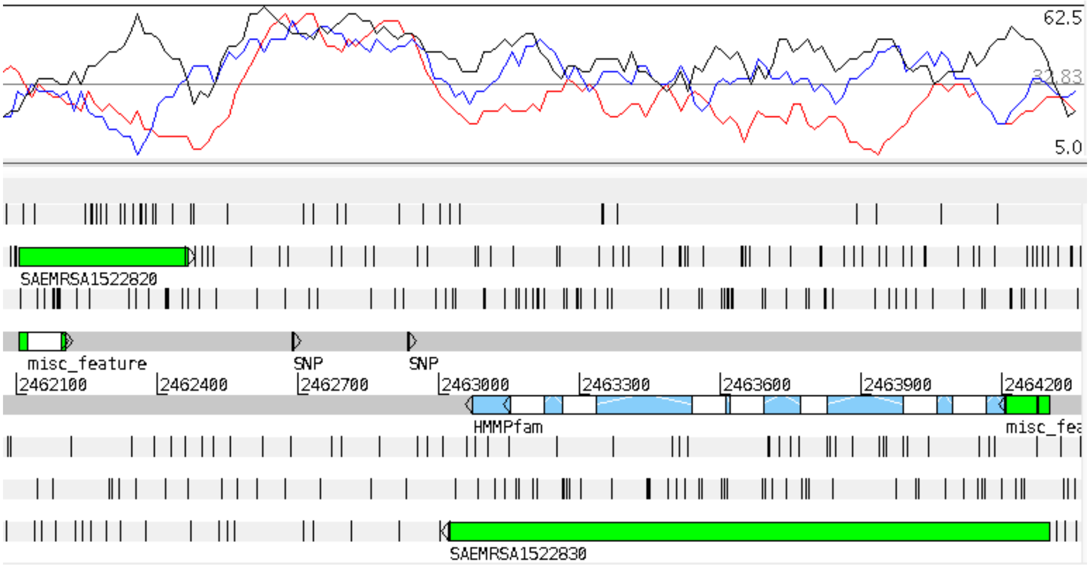

Figure 2: Genomic region around position 2462688 (left SNP, dark grey track), a homoplasic site in all three STs. Annotations from ST22 are shown. A second site, 89 bp downstream of SAEMRSA1522830 (putative proton/sodium-glutamate symport protein), is homoplasic in ST22 and ST8 only. The uppermost plot shows GC content in a sliding 120 bp window in each reading frames. Other features: predicted signal peptides (green), transmembrane helices (white) sodium dicarboxylate symporter Pfam hit (blue). Screenshot taken from Artemis[20].

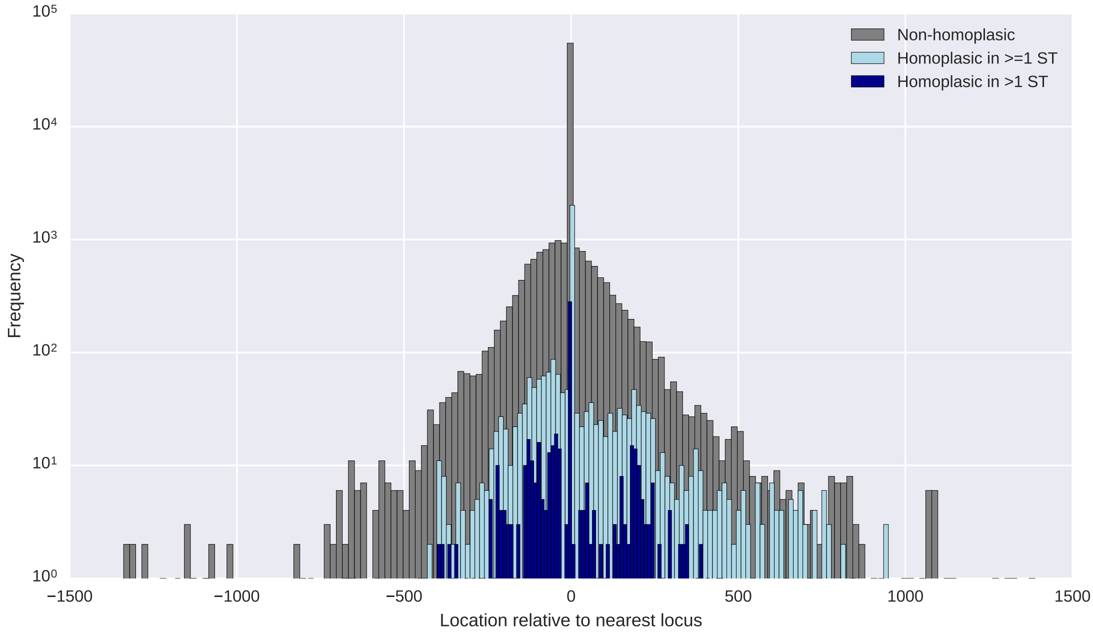

Figure 3: Distribution of SNP sites by relative distance to the nearest locus (bp), stratified by homoplasy status. Peak at zero contains intragenic sites.

Table 3: GO enrichment for loci that are not in a set of 129 known AMR-associated loci, with a homoplasic site up to 200 bp upstream in any ST. Terms significant at $p < 0.05$ are shown for each ontology (BP = biological process, CC = cellular component, MF = molecular function).

Ontology	GO.ID	Term	Annotated	Significant	Expected	p-value
Rank in BP						
1	GO:0007155	cell adhesion	33	9	0.31	9.40×10^{-12}
2	GO:0015893	drug transport	8	4	0.08	4.80×10^{-7}
3	GO:0006412	translation	358	12	3.38	0.00018
4	GO:0006351	transcription, DNA-templated	488	14	4.61	0.0006
5	GO:1901264	carbohydrate derivative transport	8	2	0.08	0.00236
6	GO:0006355	regulation of transcription, DNA-templated	467	11	4.41	0.00359
7	GO:0042254	ribosome biogenesis	56	3	0.53	0.01551
8	GO:0006352	DNA-templated transcription, initiation	24	2	0.23	0.02117
9	GO:0006284	base-excision repair	28	2	0.26	0.02831
10	GO:0006545	glycine biosynthetic process	5	1	0.05	0.04638
Rank in CC						
1	GO:0005618	cell wall	33	9	0.52	6.00×10^{-10}
2	GO:0005840	ribosome	231	10	3.63	0.0037
Rank in MF						
1	GO:0015238	drug transmembrane transporter activity	8	4	0.06	1.60×10^{-7}
2	GO:0003735	structural constituent of ribosome	227	10	1.64	3.40×10^{-6}
3	GO:0032549	ribonucleoside binding	886	9	6.41	0.00029
4	GO:0003899	DNA-directed RNA polymerase activity	20	3	0.14	0.00037
5	GO:0019843	rRNA binding	24	3	0.17	0.00064
6	GO:0003905	alkylbase DNA N-glycosylase activity	8	2	0.06	0.00139
7	GO:0003700	transcription factor activity, sequence-specific DNA binding	240	7	1.74	0.00148
8	GO:1901505	carbohydrate derivative transporter activity	16	2	0.12	0.00575
9	GO:0022884	macromolecule transmembrane transporter activity	20	2	0.14	0.00893
10	GO:0002161	aminoacyl-tRNA editing activity	24	2	0.17	0.01274
11	GO:0016987	sigma factor activity	24	2	0.17	0.01274
12	GO:0000049	tRNA binding	28	2	0.2	0.01714
13	GO:0004146	dihydrofolate reductase activity	5	1	0.04	0.03565
14	GO:0003677	DNA binding	760	10	5.5	0.04091
15	GO:0042626	ATPase activity, coupled to transmembrane movement of substances	103	3	0.74	0.04552

Fig. 5). Recombinant regions can be identified and masked using software such as Gubbins[24].

Highly repetitive regions should be checked for poor read mapping and SNP calls. Unfortunately, the same surface proteins likely to be under immune system selection are often repetitive[25], so careful quality control will be required.

Conclusions

This study presents a snapshot of the selective forces acting on the core genome of *S. aureus* in a global context. Homoplasies were rare (consistent with low recombination levels) and mostly convergences (reflecting convergent selection pressures). Sites homoplasic in multiple STs were related to AMR, virulence, and the cell surface. There was an excess of homoplasic sites in intergenic regions, suggesting regulatory sequences are under strong selection. Potentially-regulated loci downstream of these sites were enriched in AMR, membrane transport, adhesion, resistance to oxidative stress, and regulation of transcription and translation–functions important for survivability and virulence.

Materials and Methods

Isolates

S. aureus isolates ($n = 3209$) of three different STs were sourced from the BSAC[5] and Pfizer[†] collections. For each ST, a representative clone with an annotated reference genome available was selected.

Description of STs: *ST22* (1744 isolates), a pandemic MRSA strain from clonal complex (CC) 22. The representative clone is UK-EMRSA-15[8], which is dominant in the UK, responsible for approximately 85% of MRSA bloodstream infections in 2007[5]. *ST8* (707 isolates), a pandemic MRSA that has given rise to multiple lineages, including the representative clone ST8-MRSA-IV USA300, which is responsible for the majority of

[†] Unpublished collection. *Description*: "*S. aureus* isolates were collected as part of the Tigecycline Evaluation and Surveillance Trial [...] between 2004 and 2012, and derived from 28 countries. [...] Culture identification and data management were coordinated by a single reference laboratory (Laboratories International for Microbiology Studies, International Health Management Associates [IHMA], Schaumburg, IL)".

CA-MRSA in the USA[9]. *ST239* (758 isolates), a pandemic MRSA strain, derived from integration of a CC30 genomic fragment into a CC8 parent strain[5]. The three main clades are distributed in Europe, North America, and Asia (China, Turkey, Thailand)[4]. The representative clone is ST239-MRSA-III TW20[7].

Phylogenetic reconstruction

Reads for the isolates of each ST were mapped to their respective representative reference genomes using SMALT[‡]. Regions containing known MGEs (e.g. SCC*mec*; insertion sequences, transposable elements, prophage elements, and genomic islands) were masked, leaving the core genome. A maximum likelihood tree was built for each ST based on an alignment of core genome SNPs, using RAxML 8.2.8[26], under the GTRGAMMA model of rate heterogeneity. The best-scoring tree was selected and midpoint rooted. Consistency (CI) and retention indices (RI) were calculated using phangorn[27], and rescaled consistency indices (RCI)[28] were calculated as CI × RI.

Ancestral reconstruction

An in-house script was used to map characters (SNP sites) onto the trees using the maximum parsimony criterion, applying the character optimisation algorithms ACCTRAN and DELTRAN[6] to obtain two MPRs for each SNP site. For each site, the number of convergences and reversals in the set of changes for each MPR was tallied. A change that was the reverse of another ancestral change along the path to the root was counted as a reversal. For each subset of changes that resulted in the same tip state, all but one of the changes was counted as a convergence, as a set with a single change would not be classified as a convergence. Note that a single change can contribute to both tallies, if multiple reversals resulted in the same tip state. A site was classified as homoplasic if at least one convergence or reversal was counted.

Identification and annotation of sites homoplasic in multiple STs

EMBL format annotations for each ST were sourced from European Nucleotide Archive and an in-house annotation pipeline. Each site was associated with the closest EMBL coding sequence (CDS), rRNA, tRNA, or ncRNA feature. For CDSs, BLASTP (nr database) and InterProScan[29] (default analyses and associated GO terms) were run on the translated peptide sequence.

A whole-genome alignment of the representative references was constructed using mauveAligner[30], and homoplasic sites from different STs that coincided at the same alignment position were noted. Similarity of the associated annotations for these sites was used as secondary confirmation of alignment correctness.

For assignment of known *S. aureus* AMR-associated loci, a set of 129 such loci was found by combining an existing database[§] with a literature search[4,31–35].

Intergenic sites and GO enrichment

Sites in repetitive regions within loci (*ebh*, *spa*, *clfA*, *clfB*) and sites within MGEs that escaped masking (phage terminase family protein, transposase) were excluded. Fisher's exact test was applied to detect association between a site's homoplasic and intergenic status.

topGO[¶] was used for GO enrichment (min. node size of five, Fisher's exact test statistic, weight01 algorithm), where the gene universe was all coding sequences that were assigned a GO annotation by InterProScan.

Acknowledgements

I would like to acknowledge the following people for their contributions to this project: Dorota Jamrozy (ST annotations, alignments, and RaXML trees), Simon Harris (in-house scripts), Matthew Holden (ST22 sRNA annotations), and Li Yang Hsu (*S. aureus* AMR database).

[‡] SMALT is Copyright (C) 2010 - 2015 Genome Research Ltd. http://www.sanger.ac.uk/science/tools/smalt-0

[§] Li Yang Hsu, *S. aureus* AMR database, personal communication via Simon Harris.

[¶] Alexa A and Rahnenfuhrer J (2016). *topGO: Enrichment Analysis for Gene Ontology*. R package version 2.28.0.

References

1. Read, T. D. & Massey, R. C. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. *Genome Medicine* **6,** 109 (2014).
2. Feil, E. J. *et al.* How clonal is Staphylococcus aureus? *Journal of Bacteriology* **185,** 3307–16 (2003).
3. Monecke, S. *et al.* A field guide to pandemic, epidemic and sporadic clones of methicillin-resistant <i>Staphylococcus aureus</i>. *PLoS ONE* **6** (2011).
4. Harris, S. R. *et al.* Evolution of MRSA During Hospital Transmission and Intercontinental Spread. *Science* **327,** 469 LP –474 (2010).
5. Reuter, S. *et al.* Building a genomic framework for prospective MRSA surveillance in the United Kingdom and the Republic of Ireland. *Genome Research* **26,** 263–270 (2016).
6. Agnarsson, I. & Miller, J. A. Is ACCTRAN better than DELTRAN? *Cladistics* **24,** 1032–1038 (2008).
7. Holden, M. T. *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). *Journal of Bacteriology* **192,** 888–892 (2010).
8. Holden, M. T. G. *et al.* A genomic portrait of the emergences, evolution, and global spread of methicillin-resistant Staphylococcus aureus. *Genome research* **23,** 653–664 (2013).
9. Diep, B. A. *et al.* Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant Staphylococcus aureus. *Lancet* **367,** 731–739 (2006).
10. Chopra, S. & Reader, J. tRNAs as antibiotic targets. *International Journal of Molecular Sciences* **16,** 321–349 (2015).
11. Falugi, F., Kim, H. K. & Missiakas, D. M. Role of Protein A in the Evasion of Host Adaptive Immune Responses. *mBio* **4,** 1–9 (2013).
12. Shaw, L., Golonka, E., Potempa, J. & Foster, S. J. The role and regulation of the extracellular proteases of Staphylococcus aureus. *Microbiology* **150,** 217–228 (2004).
13. Frankel, M. B., Hendrickx, A. P. A., Missiakas, D. M. & Schneewind, O. LytN, a murein hydrolase in the cross-wall compartment of Staphylococcus aureus, is involved in proper bacterial growth and envelope assembly. *Journal of Biological Chemistry* **286,** 32593–32605 (2011).
14. Marchais, A. *et al.* Single-pass classification of all noncoding sequences in a bacterial genome using phylogenetic profiles Single-pass classification of all noncoding sequences in a bacterial genome using phylogenetic profiles. *Genome Research,* 1084–1092 (2009).
15. Lee, L. Y. *et al.* The Staphylococcus aureus Map protein is an immunomodulator that interferes with T cell-mediated responses. *Journal of Clinical Investigation* **110,** 1461–1471 (2002).
16. Laarman, A. J. *et al.* Staphylococcus aureus Staphopain A inhibits CXCR2-dependent neutrophil activation and chemotaxis. *The EMBO Journal* **31,** 3607–3619 (2012).
17. Howden, B. P. *et al.* Analysis of the small RNA transcriptional response in multidrug-resistant staphylococcus aureus after antimicrobial exposure. *Antimicrobial Agents and Chemotherapy* **57,** 3864–3874 (2013).
18. Baba, T., Bae, T., Schneewind, O., Takeuchi, F. & Hiramatsu, K. Genome sequence of Staphylococcus aureus strain newman and comparative analysis of staphylococcal genomes: Polymorphism and evolution of two major pathogenicity islands. *Journal of Bacteriology* **190,** 300–310 (2008).
19. Purves, J. *et al.* Variation in the genomic locations and sequence conservation of STAR elements among staphylococcal species provides insight into DNA repeat evolution. *BMC Genomics* **13,** 515 (2012).
20. Carver, T., Harris, S. R., Berriman, M., Parkhill, J. & McQuillan, J. A. Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **28,** 464–469 (2012).
21. Barnard, A., Wolfe, A. & Busby, S. Regulation at complex bacterial promoters: How bacteria use different promoter organizations to produce different regulatory outcomes. *Current Opinion in Microbiology* **7,** 102–108 (2004).
22. Liu, G. Y. Molecular Pathogenesis of Staphylococcus aureus Infection. *Pediatr Res* **65,** 71–77 (2009).
23. Gaupp, R., Ledala, N. & Somerville, G. A. Staphylococcal response to oxidative stress. *Frontiers in Cellular and Infection Microbiology* **2,** 1–19 (2012).
24. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research* **43,** e15 (2015).

25. Foster, T. J., Geoghegan, J. A., Ganesh, V. K. & Höök, M. Adhesion, invasion and evasion: the many functions of the surface proteins of Staphylococcus aureus. *Nature Reviews Microbiology* **12,** 49–62 (2013).

26. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).

27. Schliep, K. P. phangorn: Phylogenetic analysis in R. *Bioinformatics* **27,** 592–593 (2011).

28. Farris, J. S. The retention index and the rescaled consistency index. *Cladistics* **5,** 417–419 (1989).

29. Jones, P. *et al.* InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30,** 1236–1240 (2014).

30. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Research,* 1394–1403 (2004).

31. Takahashi, H. *et al.* Characterization of gyrA, gyrB, grlA and grlB mutations in fluoroquinolone-resistant clinical isolates of Staphylococcus aureus. *Journal of Antimicrobial Chemotherapy* **41,** 49–57 (1998).

32. Piddock, L. J. V. Clinically relevant chromosomally encoded multidrug resistance efflux pumps in bacteria. *Clinical Microbiology Reviews* **19,** 382–402 (2006).

33. O'Neill, A. J., McLaws, F., Kahlmeter, G., Henriksen, A. S. & Chopra, I. Genetic basis of resistance to fusidic acid in staphylococci. *Antimicrob Agents Chemother* **51,** 1737–1740 (2007).

34. Jensen, S. & Lyon, B. Genetics of antimicrobial resistance in Staphylococcus aureus. *Future Microbiology,* .565–582. (2009).

35. Sanfilippo, C. M., Hesje, C. K., Haas, W. & Morris, T. W. 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* **57,** 363–371 (2012).

Appendix

This page intentionally left blank.

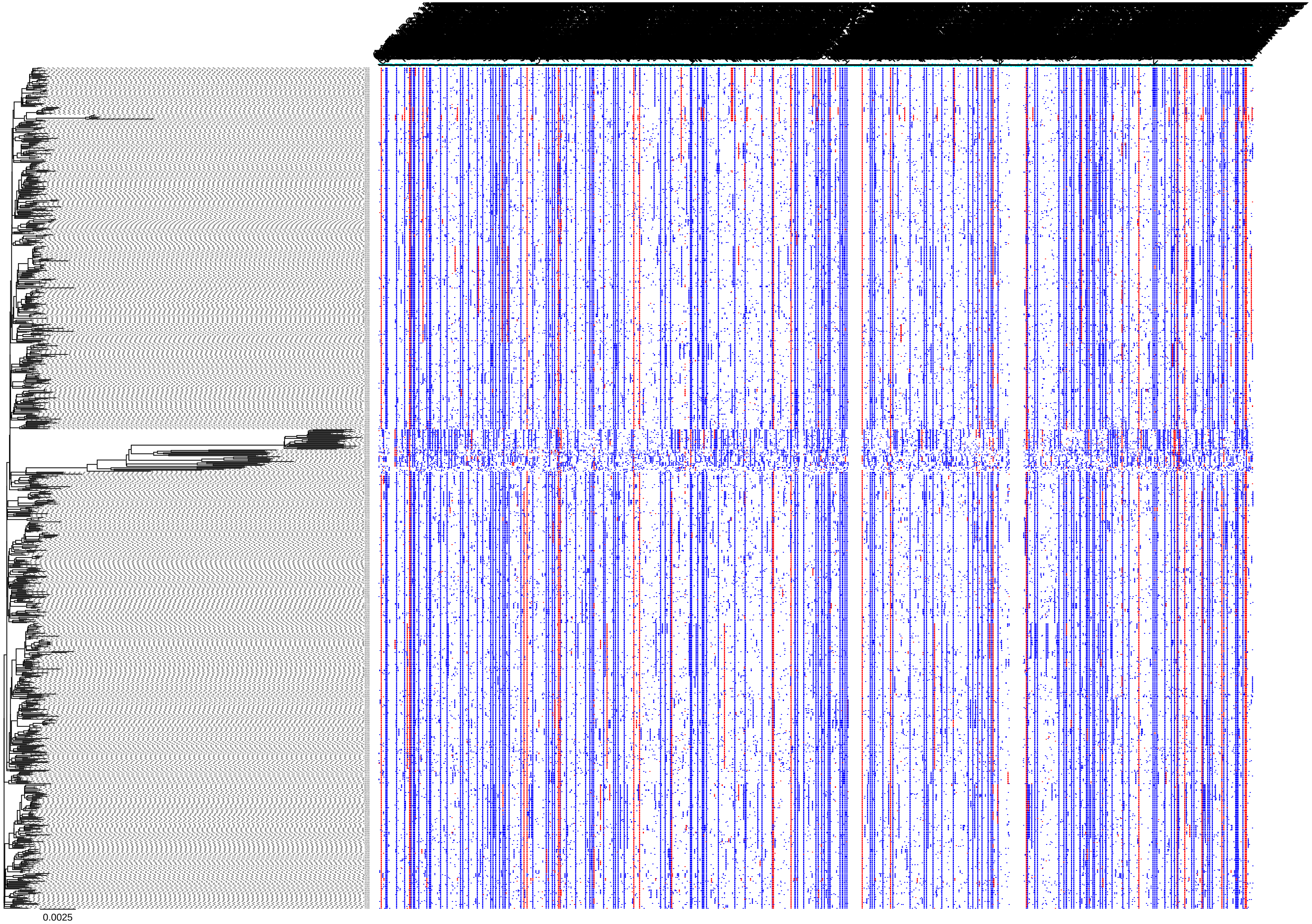

Figure 4: Distribution of non-homoplasic (blue) and homoplasic (red) SNP sites in the ST22 genome (horizontal axis) among isolates (vertical axis, RAxML tree shown).

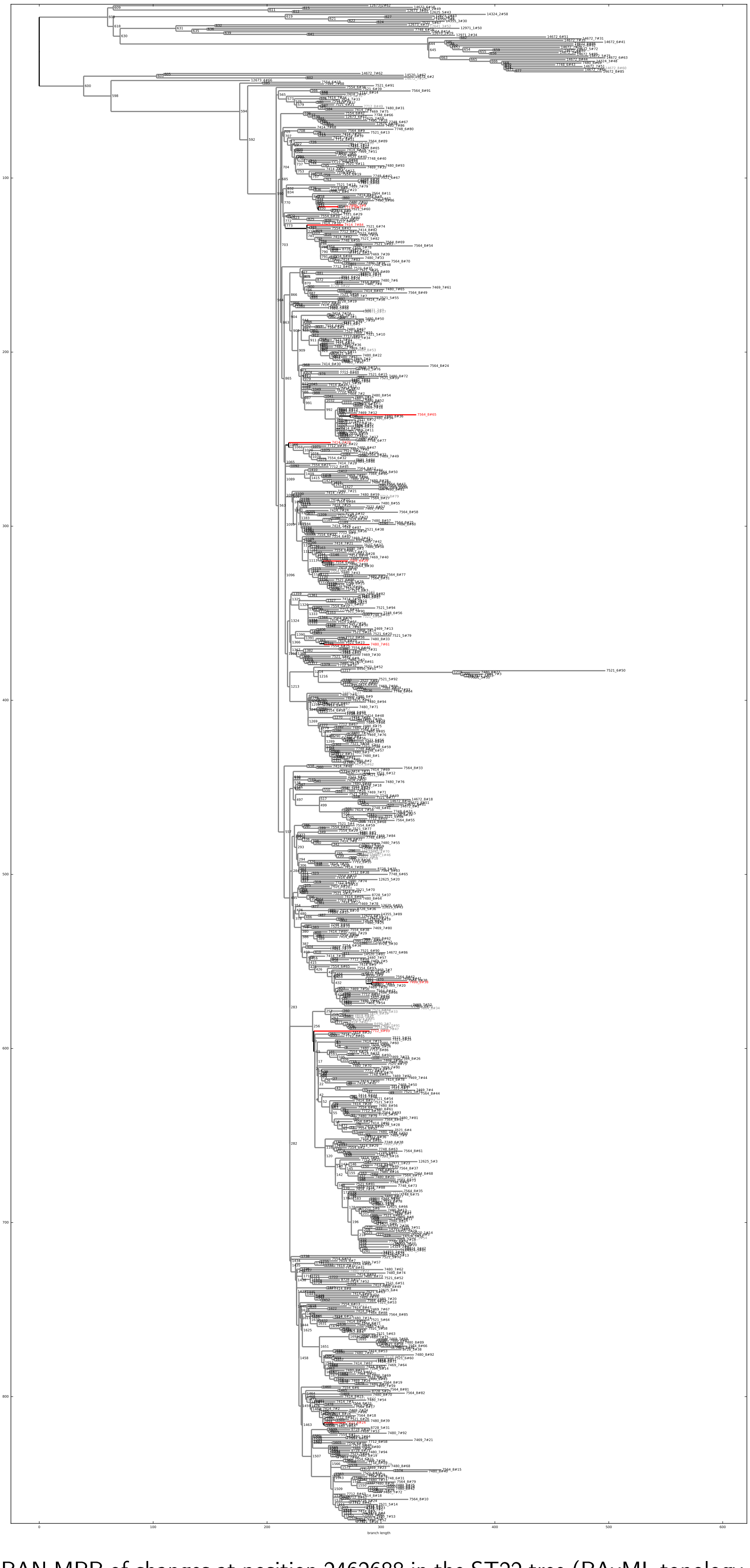

Figure 5: ACCTRAN MPR of changes at position 2462688 in the ST22 tree (RAxML topology, branch lengths from ACCTRAN). Red branches show homoplasic changes (G to T convergence) in nine lineages.