1 **Species-wide phylogenomics of the *Staphylococcus aureus* agr operon reveals convergent**
2 **evolution of frameshift mutations**

3

4 Vishnu Raghuram¹, Ashley M. Alexander², Hui Qi Loo³, Robert A. Petit III⁴, Joanna B.
5 Goldberg⁵, Timothy D. Read^{4,*}

6

7 ¹ Microbiology and Molecular Genetics Program, Graduate Division of Biological and
8 Biomedical Sciences, Laney Graduate School, Emory University, Atlanta, Georgia, USA

9 ² Population Biology, Ecology, and Evolution Program, Graduate Division of Biological and
10 Biomedical Sciences, Laney Graduate School, Emory University, Atlanta, Georgia, USA

11 ³ Department of Biology, Emory University, Atlanta, Georgia, USA

12 ⁴ Division of Infectious Diseases, Department of Medicine, Emory University, Atlanta,
13 Georgia, USA

14 ⁵ Division of Pulmonary, Allergy and Immunology, Cystic Fibrosis, and Sleep,
15 Department of Pediatrics, Emory University School of Medicine, Atlanta, Georgia, USA.

16 * Corresponding author, Email address: tread@emory.edu

17

18

19

20

21 Abstract

22 *Staphylococcus aureus* is a prominent nosocomial pathogen that causes several life-
 23 threatening diseases such as pneumonia and bacteremia. *S. aureus* modulates expression of its
 24 arsenal of virulence factors through sensing and integrating responses to environmental
 25 signals. The *agr* (accessory gene regulator) quorum sensing (QS) system is a major regulator
 26 of virulence phenotypes in *S. aureus*. There are four *agr* specificity groups each with a
 27 different autoinducer peptide sequence (encoded by the *agrD* gene). Though *agr* is critical for
 28 expression of many toxins, paradoxically, *S. aureus* strains often have non-functional *agr*
 29 activity due to loss-of-function mutations in the four-gene *agr* operon. To understand patterns
 30 in *agr* variability across *S. aureus*, we undertook a species-wide genomic investigation. We
 31 developed a software tool (AgrVATE; <https://github.com/VishnuRaghuram94/AgrVATE>) for
 32 typing and detecting frameshift mutations in the *agr* operon. In an analysis of over 40,000 *S.*
 33 *aureus* genomes, we showed close association between *agr* type and *S. aureus* clonal
 34 complex. We also found strong linkage between *agrBDC* alleles (encoding the peptidase, the
 35 autoinducing peptide itself, and the peptide sensor respectively) but not *agrA* (encoding the -
 36 response regulator). More than five percent of genomes were found to have frameshift
 37 mutations in the *agr* operon. Though most mutations occur only once in the entire species, we
 38 observed a small number of recurring mutations evolving convergently across different clonal
 39 lineages. Phylogenetic patterns suggested that strains with *agr* frameshifts were evolutionary
 40 dead ends. Overall, genomic analysis of *agr* operon suggests evolution through multiple
 41 processes with functional consequences that are not fully understood.

43 **Importance**

44 *Staphylococcus aureus* is a globally pervasive pathogen that produces a plethora of toxic
 45 molecules that can degrade, evade, or inhibit the host immune cells. Production of these
 46 toxins is mainly controlled by an active *agr* quorum sensing system, which senses and
 47 responds to bacterial cell density. However, there are many reports of *S. aureus* strains
 48 with genetic changes leading to impaired *agr* activity, often found during chronic
 49 bloodstream infections, and may be associated with increased disease severity. We
 50 developed an open-source software called AgrVATE to type *agr* systems and identify
 51 putative frameshifts. We used AgrVATE for a species-wide genomic survey of *S. aureus*,
 52 finding that more than 5 % of strains in the public database had non-functional *agr*
 53 systems. We also provide new insights into the evolution of these genetic mutations in the
 54 *agr* system. Overall, this study contributes to our understanding of a common but
 55 relatively understudied means of virulence regulation in *S. aureus*.

Introduction

Staphylococcus aureus is a ubiquitous nosocomial pathogen that continues to plague healthcare settings and threaten public health. The CDC reported that 119,247 *S. aureus* bloodstream infections and 19,832 deaths occurred in the US in 2017 (1). *S. aureus* causes a wide range of diseases such as pneumonia, osteomyelitis, endocarditis and skin infections (2). To elicit such diverse types of infections, *S. aureus* must be able to recognize environmental cues and adapt to its microenvironment (3, 4). The *agr* (accessory gene regulator) quorum sensing (QS) system is a key switch that links environmental sensing and virulence in *S. aureus* (5). The *agr* operon comprises two divergent promoters, P2 and P3, each driving the four genes essential for QS (*agrBDCA*) and a small RNA (RNAiii), respectively (Fig 1A) (6, 7). AgrD is a precursor protein that is processed by the membrane-bound peptidase AgrB, into the autoinducer peptide (AIP). The secreted AIP is then recognized by a classical two-component regulatory system – AgrC, a histidine kinase, and AgrA, a response regulator which transcriptionally activates P2 and P3 thereby continuing the autoinduction. *S. aureus* has been found to have four *agr* specificity groups, each with a distinct autoinducer peptide sequence. Other *Staphylococcus* species have their own specific *agr* autoinducer peptides (5, 8-10). Autoinduction and activation of P2 and P3 leads to *S. aureus* upregulating a large arsenal of extracellular toxins such as phenol soluble modulins, haemolysins and leukotoxins. The *agr* system also downregulates factors that facilitate cell-cell attachment, biofilm production and immune evasion (11-15). Collectively, these sum to a cell-density dependent switch between adherent and virulent modes.

Though the *agr* system regulates many important pathogenesis-related functions, many clinical isolates with impaired *agr* activity have been reported, which in some conditions may lead to worse patient outcomes (16-21). Typically, *agr*⁺ strains produce factors that are associated with increased virulence and attenuated expression of these *agr*-mediated

81 virulence factors seemingly lead to reduced disease severity and decreased host cell damage
82 (22-24). While this appears paradoxical, there are several, non-exclusive, speculations as to
83 why strains may have evolved non-functional *agr* systems (25). QS systems are ‘public
84 goods’ that promote evolutionary cheating strategies (26-29). Impaired *agr* activity may be
85 related to intra/inter-species competition between strains with different *agr* groups, which
86 tend to suppress each other with no obvious effect on colonization ability (30-34). Defective
87 *agr* function may provide pleiotropically selected phenotypes, such as decreased
88 susceptibility to vancomycin (35, 36). *agr*- strains may have traded their ability to produce
89 energetically expensive virulence factors but may be limited in their ability to compete with
90 the host immune system (37-39). This attenuated toxicity appears to be inconsequential or
91 sometimes even beneficial in chronic diseases such as cystic fibrosis, bacteremia and
92 osteomyelitis, where *agr*- *S. aureus* may show increased persistence and higher mortality
93 rates compared to *agr*+ strains (20, 21, 24, 31, 40). However, reduced/lack of expression of
94 *agr* mediated virulence may be detrimental to colonization in some other circumstances (41).

95 It has been proposed that *agr*- strains have sacrificed long-term viability through successful
96 between-host transmissions for increased adaptation to specific environmental niches within
97 the host (40, 42-44). However, studies to date have typically focused on small numbers of
98 strains in limited clinical settings. It is important to assess whether specific patterns of
99 variation in *agr* can be observed from a genome-wide scale spanning multiple clonal lineages
100 of *S. aureus*, as it is crucial for understanding the mechanisms driving virulence regulation.

101 The amount of publicly available genome sequences of *S. aureus* has grown rapidly over the
102 past decade (45), offering the opportunity to examine the evolution and diversity of critical
103 virulence determinants in *S. aureus* from a species-wide standpoint.

104 In this study, we developed a bioinformatics pipeline for rapid identification of the *agr* group

from a given *S. aureus* genome as well as putative null mutations in the *agr* operon.
<https://github.com/VishnuRaghuram94/AgrVATE>) and used it to analyse the 42,999 *S.*
aureus genomes from the Staphopia database of consistently assembled and annotated public
genome sequences compiled in 2017 (45). We found that there was a high degree of purifying
selection for specific alleles of *agr* genes based on *agr* group and clonal complex. We
detected frameshift mutations in 5.7% of all analysed *agr* operons, most of which were
singular events. We also detected instances of identical frameshift mutations in *agr* genes of
unrelated strains across different clonal complexes, suggesting that there is a mechanism
promoting mutations at these specific sites. Overall, these results highlight the highly variable
nature of the *agr* operon and suggest conserved mechanisms for acquiring genetic changes
that may affect *agr* mediated virulence regulation.

The following are definitions for the terminology we will be using consistently in this study:
Group-1, Group-2, Group-3, and Group-4 refer to *S. aureus* strains that belong to one of the
four *agr* specificity groups. A cluster refers to a collection of an *agr* gene where either the
nucleotide or amino-acid sequence of the gene is 100% identical among all sequences within
that collection. A cluster representative is a random sequence chosen to represent each
cluster, whose sequence is identical to all other sequences within that cluster. A nucleotide
sequence cluster representative is referred to as an allele. An amino acid sequence cluster
representative is abbreviated to AACR. Frameshift mutations in the coding regions of the *agr*
operon are referred to as “putative *agr* null” mutations as the true phenotype is unknown.
Strains reported to have impaired *agr* activity are referred to as *agr*- while strains with
canonical *agr* activity are referred to as *agr*+ .

Results

AgrVATE: A tool for kmer based assignment of *agr* groups and *agr* operon frameshift detection

We designed the AgrVATE (*agr* Variant Assessment & Typing Engine) bioinformatic workflow to process *S. aureus* genome sequences to assign *agr* groups and detect frameshift mutations in the *agr* operon. Current methods for *agr* group assignment involve traditional PCRs or alignment searches against *ad hoc* databases (41, 46-48). AgrVATE was designed to be a fast, standardized workflow for assigning *agr* groups that is conveniently installable through the Conda package manager (49). AgrVATE contains a database of 4 distinct collections of kmers, where each collection corresponds exclusively to a single *agr* group. This kmer database is used to perform a BLASTn search against a given input genome to assign the *agr* group. The process of building and verifying this kmer database is outlined in the methods section. AgrVATE then extracts the *agr* operon by *in-silico* PCR using usearch (50) and performs variant calling using Snippy v4.6 (51) to detect putative *agr* null mutations such as frameshifts and early stops. As a reference for the variant calling, the cluster representative from the largest cluster for each *agr* group is used. The advantage of *in-silico* PCR over global alignment methods for extracting the *agr* operon is that if the *agr* operon contains large indels/possible novel sequences which would normally break alignments, the operon will still be extracted as we only rely on the primers “binding” to the up and downstream regions of the operon. AgrVATE analysis took < 4 seconds per whole genome assembly on a Linux server with 12 core CPU and 96GB RAM. The AgrVATE workflow is outlined in **Fig S1**.

We ran AgrVATE on 88 *S. aureus* genomes that had been typed for haemolysis activity, a phenotype that is generally associated with functional *agr* systems (16). These 88 genomes

152 included clinical samples taken from cystic fibrosis patients from the Emory Cystic Fibrosis
 153 Center (52). We found 15 genomes had putative *agr* null mutations, 14 of which (93%) tested
 154 negative for sheep blood haemolysis (**Table S1**). The one putative null that displayed
 155 haemolysis on sheep blood agar had a frameshift mutation in the C terminal end of *agrC*.
 156 The haemolysis phenotype of this strain was relatively weak (CFBR_17 – **Fig S2**). We also
 157 observed multiple samples that were negative for haemolysis despite having no frameshift
 158 mutations in the *agr* operon (**Table S1**), suggesting that other genetic factors reduced
 159 haemolysis activity and *agr* frameshifts are not the sole indicator. In patient CFBR311,
 160 AgrVATE found two different *agr* groups (group-1 and group-2) from the genome sequence
 161 of the *S. aureus* population isolated from CF sputum (CFBR_EB_Sa110 – **Table S1**),
 162 showing that this patient is colonized by *S. aureus* of heterologous *agr* groups. Genome
 163 sequences of 8 individual colonies (CFBR_EB_Sa111 to CFBR_EB_Sa118 – **Table S1**) from
 164 this patient showed an *agr* group-2 majority (6 out of 8) and an *agr* group-1 minority (2 out
 165 of 8), thereby validating the AgrVATE prediction. This illustrated an advantage of using a
 166 kmer-based approach, as AgrVATE could identify the presence of multiple *agr* groups. Each
 167 *agr* group assignment was also scored, thereby indicating the proportions of each *agr* group if
 168 more than one is present. Collectively, these findings demonstrate a potential use-case for
 169 AgrVATE in clinical settings, where we identified CF patient isolates having *agr* mutations
 170 and showed one instance where AgrVATE identified a patient colonized by *S. aureus* of
 171 heterologous *agr* groups.

172 ***agr* type distribution in Staphopia**

173 AgrVATE identified the *agr* groups for 42,491 genomes out of 42,999 in Staphopia, with
 174 25,539 group-1 (60.10%), 9639 group-2 (22.68%), 6224 group-3 (14.65%) and 1,089
 175 (2.56%) group-4 genomes (**Fig 1B**). Each clonal complex contained only one *agr* group,

except CC45 which had both group-1 and group-4 *agr*, as had been reported previously (53, 54) (Fig 1C). 1,601 out of the 42,491 genomes showed the presence of more than one *agr* group not unlike our previous finding in patient CFBR311 (Table S1). However, all of these genomes had the secondary *agr* group call on short, low coverage contigs and there were no instances of multiple *agr* group calls on the same contig. This also shows that AgrVATE can identify *agr* groups in fragmented genome assemblies where the *agr* genes are broken across multiple contigs. These genomes were considered to be contaminated by sequences from *S. aureus* of other *agr* groups and were not included for further analysis (Fig S3), leaving 40,890 genomes with high confidence *agr* group assignments. From the limited number of strains with the associated metadata, we found the distribution of *agr* groups across blood (3,755 genomes), skin (2,602 genomes) and nasal (4,257 genomes) isolates to be similar to the overall distribution of *agr* groups. Group-2 genomes (CC5) were enriched in respiratory tract isolates (1,107 genomes) (Fig 1D).

The remaining 508 genomes were reported to be low confidence/unknown *agr* group calls by AgrVATE. We used a mash sketch (55) built from all publicly available complete genomes of *Staphylococcus* species to determine if these 508 genomes were indeed *S. aureus*. We found that 312 genomes did not belong to the *S. aureus* species and were likely mis-annotated submissions in NCBI and therefore discarded from this analysis. From the remaining 196 samples we found complete *agr* deletions (63 genomes), samples with their *agr* operons fragmented across low quality contigs, leading to unreliable base calls (82 genomes), and *agr* group-1 operons which have relatively low sequence identity (<96%) to a canonical *S. aureus* *agr* group-1 (51 genomes). Upon further investigation by BLASTn, we found 35 of these 51 operons belong to *S. argenteus* species while the remaining 16 were *S. aureus* operons. The fate of all 42,999 genomes in Staphopia after processing them through AgrVATE is outlined in Fig S3.

***agr* cluster recombination within clonal complexes is rare**

39,174 complete *agr* operons were extracted by AgrVATE from 40,890 *Staphylococcus* genomes and clustered with 100% nucleotide identity, resulting in a total of 5,143 unique *agr* operon sequences. The remaining 1,716 genomes had their *agr* operon sequences fragmented across multiple contigs and therefore were not included for further analyses. 97% of all extracted operons were of length 3481 to 3484bp. While a single CC could harbour multiple *agr* operon clusters, there was no *agr* operon cluster that was shared between genomes of different CCs, as would be expected to be produced by recombination events that introduced entire *agr* operons from one CC to another. When individual genes were clustered at the 100% identity threshold, we found 1,086 unique *agrA*, 440 *agrB*, 2,544 *agrC* and 51 *agrD* alleles. As expected, there was remarkably low sequence variability in *agrD* and each allele represented a single *agr* group. All alleles of *agrB* and *agrC* were exclusive to a particular *agr* group (**Fig 2A**).

With two exceptions, there was little evidence of *agr* recombination between CCs. In the first exception, an *agrB* allele found in 1027 CC15 genomes was also found in 244 CC5 and 150 CC12 genomes. A different *agrB* allele was also found in 20 CC15 and 17 CC5 genomes. This suggests that *agr* gene alleles can be shared between CCs of the same *agr* group, though relatively rare. The second, when comparing the *agr* alleles across group-1 and group-4 CC45 genomes, we found that both *agr* groups had the same *agrA* allele but different group-specific *agrBDC* alleles. Specifically, out of 1,686 CC45 genomes in *Staphylococcus*, 1,294 were group-1 and 392 were group-4. 94% of all CC45 genomes, which included 1,217 group-1 and 384 group-4 genomes have identical *agrA* alleles. However, each *agr* group had distinct *agrBDC* alleles. This suggested that a recombination event led to stable introduction of group-4 specific *agrBDC* alleles in CC45.

Though most *agrA* alleles were CC-specific, there were multiple instances of the same *agrA* allele being found in different CCs and different *agr* groups. This lack of *agr* group specificity in *agrA* became more apparent while analysing the amino acid sequences. 83% of all AgrA amino acid sequences were identical and therefore have the same Amino-Acid sequence Cluster Representative (AACR). CC5, CC8, CC30 and CC45 exclusively had this major AACR of AgrA, encompassing all four *agr* groups (**Fig 2A - Red arrow, Fig 2B**). 9% had an alternate amino acid sequence of AgrA which differed from the major AACR by a single amino acid (K136R) (**Fig 2A-Blue arrow, Fig 2B**). This included mainly CC15 and other rare CCs. We designated the major AgrA AACR AgrA^{K136} and the minor AgrA AACR AgrA^{R136}. Upon constructing a Maximum Likelihood phylogeny using IQ-TREE (56) from a curated set of 355 genomes from Staphoplia each representing a unique ST, called Non-Redundant Diversity (NRD) set (45), we found that *S. aureus* can be broadly divided into two clades or sub-species: all genomes harbouring AgrA^{R136} exclusively belong to a single clade of *S. aureus* (**Fig 2C, blue tips**). The remaining rare AgrA amino acid sequences (6%) were variants of one of the two major AACRs. To observe the linkage between individual *agr* genes, we identified SNPs in the region comprising the *agr* operon and the 1000bp flanking regions on our filtered NRD set of 355 genomes using Snippy (51). We then measured Linkage Disequilibrium (LD) between these SNPs by calculating the Pearson coefficient (R^2) using plink (57). SNP pairs with $R^2 > 0.8$ were considered to be in LD. The resulting LD plot (**Fig 2D**) showed that SNPs in the variable region of *agrBDC* are in LD with each other but not with *agrA*, and that *agrA* is in LD with the flanking regions of the operon. Overall, this suggests that *agrBCD* coevolve and are unlinked to *agrA* or the rest of the genome, while *agrA* evolution is linked to the *S. aureus* genome.

Non-functional *agr* operons are common across diverse *S. aureus* genomes

249 In 39,174 *agr* operons, there were 405 sites that had a frameshift mutation in at least one
 250 genome. Most frameshifts occurred in only one genome, but a small minority of sites were
 251 more frequently mutated - 24 sites had frameshifts in at least 10 genomes and 5 sites had
 252 frameshifts in more than 100 genomes. At least one frameshift mutation was found in each
 253 *agr* gene (**Fig 3A**). The *agrC* gene had acquired the greatest number of different frameshift
 254 mutations, including truncation mutations in CC5, CC8, CC22 and CC30 (865 genomes) (**Fig**
 255 **3A, B**). The most frequent frameshift was the insertion of an adenine in the terminal end of
 256 *agrA*, occurring in 561 genomes (35 unique *agr* operons) across multiple CCs (p.Ile238fs –
 257 **Fig 3B**). This mutation has previously been investigated by Traber & Novick and was found
 258 to cause delayed *agr* activation (58). Another mutation toward the end of the gene occurring
 259 in a polyA tract found in *agrB*, occurred in 114 genomes (34 unique *agr* operons), mainly
 260 CC15 and CC5 (p.Phe201fs – **Fig 3B**). A frameshift that resulted in loss of the start-codon in
 261 *agrC* was observed in 86 genomes (25 unique *agr* operons) of CC8, CC30, CC22 and some
 262 other rarer CCs (p.Val? – **Fig 3B**). In relatively low frequency (69 genomes), we observed
 263 complex mutations such as collapsed repeats, tandem duplications and large (> 30bp) in-
 264 frame and out-of-frame indels. However, such mutations were mostly singular sporadic
 265 events, and no mutation was recurrent in more than three different genomes. Two of these
 266 cases were an Insertion Sequence (IS) element insertion in the *agr* operon. One being a
 267 1,326bp insertion of an IS256 family transposon sequence found commonly in
 268 *Staphylococcus* species (59), and the other a 1057bp insertion of an IS1252 transposon
 269 sequence found in *Enterococcus* species (60).

270 Frameshifts in the delta-toxin gene (*hld*), present within the RNAiii transcript (**Fig 1A**) were
 271 rare, with only 4 genomes having one of two mutations - a deletion at position 76 leading to a
 272 frameshift, or a G to A substitution in position 44 leading to a premature stop. We found
 273 3943 indels in RNAiii, 3931 of which were single nucleotide indels. The most common

274 mutation occurring 2025 times was the insertion of a T at position 406 of RNAiii, found
275 exclusively in *agr* group-3 genomes, suggesting that this might be a common variant. It is
276 unknown whether these single nucleotide indels have a functional impact on RNAiii. On the
277 other hand, we also found 8 genomes with large >30 bp indels in RNAiii which may affect
278 function. Namely, two *agr* group-1 genomes with a 42bp deletion at position 391, four *agr*
279 group-2 genomes with a 41bp insertion at position 2, and one *agr* group-3 strain with a 31bp
280 insertion at position 9.

281 We compared the number of genomes with indels in the *agr* operon to other two-component
282 regulatory systems (TCS) to determine if the frequency of potentially deleterious mutations in
283 the *agr* operon was significantly different. This was done by extracting the histidine kinase
284 (HK) and response regulator (RR) genes of TCS *arlRS*, *kdpDE*, *nreBC*, *phoPR*, *srrAB* and
285 *walKR* from Staphopia and calculating the number of genomes with HK or RR genes with
286 non-canonical gene lengths. The length of the reference gene for each HK and RR was
287 considered the canonical length and the length of each annotated gene in Staphopia best
288 matching the reference gene was identified (BLASTn). At least 42,500 hits for each HK or
289 RR were extracted out of 42,999 genomes in Staphopia. Hits for each HK and RR with length
290 not equal to their corresponding reference were considered non-canonical gene lengths due to
291 indels. Commonly occurring variant gene lengths (> 5000 genomes) were still considered
292 canonical and filtered out. We found that the *agr* TCS has a significantly greater number of
293 variable gene lengths when compared to TCS *arl*, *kdp*, *nre*, *pho*, *srr* and *wal* ($p < 0.0001$,
294 negative binomial regression). We found 5.8% of all *agrC* (HK) genes and 3.3% of all *agrA*
295 (RR) genes analysed were of variable lengths. In contrast, only $\leq 1\%$ of all other HKs and
296 RRs analysed had variable gene lengths except for *kdpD* (HK) which had 3.2% (**Fig 3C**).
297 Overall, *agrAC* had a higher percentage of non-canonical gene lengths compared to the

corresponding genes from other two-component systems, suggesting higher frequencies of indels.

To estimate whether factors such as *agr* group, clonal complex, host body site and infection/colonization status can serve as predictors of null mutations, we trained models using a general linear model (GLM), random forest (RF), Extreme Gradient Boosting (XGB) and K-Nearest Neighbours (KNN) to predict the presence/absence of frameshift mutations in the *agr* operon. All 4 models had high negative predictive value and low precision which could be due to the imbalanced nature of the test dataset (See methods) (Table S2). This suggested that the likelihood of acquiring frameshift mutations in the *agr* operon cannot be predicted by the site of infection and pathogenicity status alone.

Some *agr* frameshift mutations have occurred repeatedly through convergent evolution

We noticed that certain frameshift mutations in the *agr* genes occurred frequently across different strain backgrounds. To test whether these recurrent mutations could be explained by a purely random mutational process, we simulated frameshifts in a random set of wild type operons and compared the resulting frequency distribution to the real distribution of mutation sites in the genomes carrying *agr* frameshifts (referred to as frameshift+ genomes). We chose a set of dereplicated genomes from CC8, CC22, CC5 and CC30 to reduce sampling bias affecting frameshift counts (See Methods for dereplication strategy). These are the most abundant CCs in Staphopia and carry many of all identified frameshift mutations. We found that though the total number of mutation events in the real and simulated dataset were similar (~300), the number of unique sites mutated in the simulated dataset was greater than the real dataset. This showed that the simulated distribution was significantly different from the real distribution of frameshifts in the *agr* operon (Kolmogorov-Smirnov $p < 0.01$). (Fig 4A).

Moreover, we calculated the consistency index of each frameshift site on CC specific

322 maximum likelihood trees using HomoplasyFinder (61). In short, a consistency index of one
 323 for a given site on an alignment indicates that the nucleotides at that site are consistent with
 324 phylogeny, and a consistency index of 0 indicates that the nucleotides at the site are
 325 homoplasious (evolved independent of phylogeny). We found a trend of decreasing
 326 consistency index with increasing frequency of each recurring frameshift. We observed an
 327 almost identical trend when the tips of the trees are shuffled to have the frameshift+ genomes
 328 at random positions on the tree, mimicking non-ancestral, independent acquisition of each
 329 frameshift (**Fig 4B**). Overall, this showed that there is a preference for specific sites in the *agr*
 330 operon to acquire potentially null-inducing mutations and that identical mutations can occur
 331 independently across different *S. aureus* lineages.

332

Discussion

In this study, we used Staphopia, the largest available database of consistently assembled and annotated *S. aureus* genome sequences (n = 42,999) to analyse patterns in evolution and diversity of the *agr* quorum sensing system. Our goal was to place previous work on the evolutionary genetics of *agr* in the context of the thousands of genomes currently available. We developed a bioinformatics tool, AgrVATE, for rapid genome-based classification of *agr* specificity groups and for identification of putative null mutations in the *agr* operon (<https://github.com/VishnuRaghuram94/AgrVATE>). Our findings to a large extent were consistent with previous studies but the increased scale of the analysis revealed new features. We confirmed that only the 4 previously known autoinducing peptides that define *agr* groups 1-4 were present in *S. aureus* (**Fig 1B, Fig S3**). To our knowledge, there has been no credible report of any other peptide reported, although other *Staphylococcus* species have *agr* operons encoding different cyclic peptides (62). We also found, as previously reported (10, 63), that *agrBDC* alleles are *agr* group specific while *agrA* alleles are independent of *agr* group (**Fig 2A**). In addition, we found that *agrBDC* are in linkage disequilibrium while being unlinked to *agrA* (**Fig 2D**). Moreover, with the exception of CC45, clonal complexes were exclusively linked to specific *agr* groups and specific alleles within these groups (**Fig 1C**). A third major result was that, while the *agr* operon was rarely deleted completely in any strain, ~ 5% of all analysed *agr* operons have at least one frameshift mutation in the coding regions, indicating that potential non-functional *agr* variants are relatively common (**Fig 3A, B**). *agr* defective strains have been frequently reported (16-21), and we showed through genomic comparison that *agrAC* more frequently accumulated frameshift mutations compared to other *S. aureus* two component systems (**Fig 3C**). Using the somewhat limited publicly available metadata, we could not link *agr* null strains to a particular body site or type of infection. While most of these frameshift mutations are singular evolutionary events, we found a handful of sites

across unrelated strains that have independently acquired a disproportionately high number of frameshifts (**Fig 4A**) with no evidence for phylogenetic transmission (**Fig 4B**), suggesting selection or a generative mechanism for high frequency mutations. At least one of these frameshifts has been studied functionally (58), but many of these frequent frameshifts were only detected through large-scale genomic analysis reported here.

This study re-emphasizes that *S. aureus* is not phylogenetically structured according to *agr* groups: i.e strains belonging to each *agr* group do not fall into their own monophyletic clades (**Fig 3C**) (63). This pattern can be best explained by rare homologous recombination of the *agrBCD* genes. Strikingly, all strains within a clonal complex (except the previously mentioned CC45) belonged to only a single *agr* group. While there have been reports of multiple *agr* groups within the same sequence type in *S. aureus* (53, 54), we did not observe any such instances across more than 40,000 genomes. It has been proposed that clonal complexes in *S. aureus* emerge from recombination and/or genome rearrangement events and remain stable due to the presence of barriers to recombination and HGT between CCs (64, 65). These results suggest that *agrBDC* recombination may be the impetus for formation of clonal complexes. CC45 may be in the process of CC formation after a recent switch in *agr* group and may thus be an interesting natural laboratory for understanding the evolutionary dynamics that drive this process.

We observed strong purifying selection in *agrA*. Most of the CC specific nucleotide changes in *agrA* were synonymous changes, with 83% of all AgrA sequences being identical. The only significant non-synonymous change we observed on a species-wide scale was a single amino-acid substitution at position 136 (9% of all AgrA sequences). This alternate AgrA (AgrA^{R136}) was found only within one monophyletic clade comprising CC15 and other rare CCs of *S. aureus* (**Fig 2B, C**). Overall, this shows that nucleotide differences in the *agr*

operons, even in operons belonging to the same *agr* group, can serve as a predictor of the subspecies and clonal complex, however the functional impact of these alternate alleles, if any, are unknown.

We know that *agr* function is not always essential for *S. aureus* survival, as non-functional *agr* variants are commonplace and are frequently isolated from patients (16-21). For example, there is a relatively high occurrence of nasal colonization by strains with downregulated *agr* expression in hospital settings (66). Nasal carriage is an important step for initiation of *S. aureus* infection, and it has been observed that the presence of isolates with impaired *agr* function in the bloodstream is often associated with isolates of identical *agr* function in the nasal cultures (67, 68). This suggests that complete virulence capacity is not an absolute requirement for colonization and transmission of *S. aureus* in the hospital environment (42). However, community associated transmission by *agr* defective strains is thought be curtailed and the *agr*- strains do not remain long enough to establish a circulating population outside the initial location (44, 66). This brings to light the possible evolutionary trade-off for strains that become *agr* defective, supported by our phylogenetic analyses, which show no evidence of stable lineages of putative *agr* null populations. In addition, we also found that frequency of mutations in the *agr* TCS is enriched when compared to other TCS in *S. aureus*. The relatively common occurrences of independently acquired *agr* mutations suggests that they may be adaptive convergent mutations in response to specific selective pressures. It is also important to note that this study does not investigate non-synonymous substitutions and mutations in genes outside the *agr* operon which may affect *agr* activity. A recent study (69) showed that isolates with reduced toxin production need not necessarily harbour *agr* mutations. Our haemolysis results from CF *S. aureus* strains (Table S1) also support this, as we see strains without *agr* mutations showing reduced haemolysis. This highlights the multi-faceted nature of *agr* mediated virulence and that the true frequency of phenotypically *agr*

null is likely higher than what we report in this study.

The *Staphylococcus agr* system is a central feature of virulence gene regulation that has been studied for more than forty years but much regarding the evolution and maintenance of *agr* remains poorly understood. There are two particularly interesting negative findings in this study: the absence of non-canonical “intermediate” AIPs in 42,999 strains, and the absence of any strain that has acquired *agrD* from a *Staphylococcus* outside of *S. aureus*. The evolutionary mechanism behind the diversity of *S. aureus agr* was hypothesized to be random mutation of the *agr* locus to give rise to multiple sequence configurations, followed by selection for only functional configurations leading to the four *agr* specificity groups that exist today (32). This model implies intermediate or transitional *agr* groups, which were presumably non-functional *agr* operons were driven to extinction by diversifying selection, allowing only functional *agr* systems to become successful lineages. The absence of intermediates suggests a strong selection for maintenance for four group specificities in *S. aureus* that leaves producers of novel peptides at a disadvantage. Although, the high number of frameshifts suggests that not producing any peptide at all may confer higher fitness in some environments and could be a viable transitory strategy. Similarly, while there is abundant evidence for HGT of antibiotic resistance genes from other *Staphylococcus* species (70-74), we did not find evidence of *S. aureus* acquiring *agr* genes encoding novel AIP specificities. Close relatives of *S. aureus* such as *S. argenteus* and *S. schweitzeri* share the *agr* group-1 AIP, though *S. argenteus* and *S. schweitzeri* also developed their own distinct AIPs (75). This may suggest that while a common ancestor of these three *Staphylococcal* species may have also been *agr* group-1, environmental niche selection drove the emergence of species specific *agr* groups. It may be that AIP specificity plays a role beyond just intra-species competition that we do not yet understand.

Methods

AgrVATE workflow:

AgrVATE was written in Bash and uses freely available software. AgrVATE only requires a *S. aureus* genome assembly in FASTA format as input and the outputs include the detected *agr* group, the extracted *agr* operon and a table with annotated variants if any. The installation and usage instructions as well as descriptions of all output files can be found on Github (<https://github.com/VishnuRaghuram94/AgrVATE>). It is recommended to run AgrVATE on Unix based operating systems. The methods for building AgrVATE and running it on Staphopia genomes are outlined below.

Identifying a unique set of 31mers for each *agr* group:

We first assigned *agr* groups based on the AIP amino acid sequence to all genomes in Staphopia (45) where a canonical AgrD protein was annotated by Prokka v1.14.6 (40,812 AgrDs) (76). We then extracted the *agr* operons from these 40,812 genomes by *in-silico* PCR using usearch v11.0.667_i86linux32 (50). To identify kmers unique to each *agr* group, DREME v5.1.1 (77) was used to identify 31bp kmers (31mers) that were unique to the *agr* operon of each *agr* group, resulting in four distinct groups of 31mers (evalue < 0.0001). AgrVATE uses this output of 31mers unique to each *agr* group as a database to conduct a BLASTn v2.10.1 (78) search against an assembly of a given *S. aureus* genome to identify the *agr* group. We also used AgrVATE to re-assign *agr* groups to the preliminary set of 40,812 genomes and the group assignments matched in 40,725 cases. In the 87 cases where the initially assigned *agr* group did not match the AgrVATE assignment, we found that the genome assemblies were contaminated with another *S. aureus* isolate of a different *agr* group. In these cases, AgrVATE will assign the *agr* group with the most kmer matches while also noting that more than one group was found.

agr operon and *agr* gene extraction:

In-silico PCR was performed for 43,000 *S. aureus* whole genome sequences in Staphopia database using usearch -search_pcr tool (50) with the following primers – 5'aaaaaaggccgcgagcttgggaggggctca'3 & 5'ttatatttttaacgtttctaccgatgc'3. Both primers were required to bind and 8 mismatches in total were allowed. Extracted *agr* operons were clustered with 100% identity using usearch -fastx_uniques (50) to obtain all possible unique *agr* operon configurations. This unique set of operons was annotated using Prokka v1.14.6 (76), *agr* genes were extracted and clustered again with the same parameters to obtain all possible nucleotide and amino-acid configurations of each *agr* gene.

Identifying variants in the *agr* operon:

The most frequently occurring *agr* operon nucleotide configuration was determined for each *agr* group and used as a reference. Variant calling was performed using snippy v4.6.0 (51). Only loss of start, gain of stop and frameshift mutations occurring within the coding regions of the *agr* operon were considered possible non-functional variants. AgrVATE filters the snippy output and reports the above-mentioned mutations in tabulated format.

Staphopia metadata

Metadata associated with *S. aureus* genomes submitted to the NCBI Short Read Archive was downloaded as a table using the Run Browser tool. This was then subjected to a series of bioinformatic filters to clean up key fields such as collection date and location, host body site and host status. Additional data from supplemental tables of several published *S. aureus* genome sequencing studies was also added. The data and scripts can be accessed at https://github.com/Read-Lab-Confederation/staphopia_metadata/. The table used in this study was 'Stage3.4.csv' (commit 4548f17, 2020-08-14).

Whole genome phylogeny and Linkage Disequilibrium

The Staphopia non-redundant diversity set which contains 380 genomes was filtered to contain only genomes where the full *agr* operon was extracted by AgrVATE and all four *agr* genes as well as the genes up and downstream of the *agr* operon were annotated by Prokka v1.14.6 (76), resulting in 355 genomes. A core genome alignment was constructed using parsnip v1.5.3 (79) and this alignment was used to build a maximum likelihood phylogeny using IQ-TREE v1.6.12 (56) with the GTR+FO model and 1000 ultrafast bootstrap replicates. The resulting tree was then plotted using the R package ggtree (80). The genomic region comprising the *agr* operon and 1000bp on each side was extracted from the filtered NRD set of 355 genomes and this region was aligned using snippy-core v4.6.0 (51). The full alignment (core.full.aln) file was then converted to a vcf file using snp-sites v2.5.1 (81), and this vcf file was used to calculate Pearson's coefficient for Linkage Disequilibrium (LD) using plink v1.90b6.21 (options: --r2 inter-chr) (57). The resulting table was used to build a LD plot using R.

Comparing indel rate of *agr* to other *S. aureus* global regulators

USA 300 strain NRS384 (accession NZ_CP027476.1) was used as a reference to extract histidine kinase (HK) and response regulator (RR) genes belonging different *S. aureus* two component regulatory systems (*arlRS*, *kdpDE*, *nreBC*, *phoPR*, *srrAB*, *walKR*). A TCSs was chosen if the number of Gene Ontology enrichment hits exceeded ten in the regulon of a constitutive RR strain where all other TCSs are deleted (Rapun-Ariaz et al, 2020; Table S1 (82)). The length of the reference gene for each HK and RR was considered the canonical length and the length of each annotated gene in Staphopia best matching the reference gene was identified (BLASTn). Hits for each HK and RR with length not equal to their corresponding reference were considered non-canonical gene lengths due to indels. Commonly occurring variant gene lengths (> 5000 strains) were still considered canonical

and filtered out. We performed negative binomial regression on raw count data of frequency of variable gene lengths across the TCS offsetting for canonical gene length and total number of genes. Tukey's method was used for multiple comparisons. Statistical tests were performed using the `nb.glm()` function from the MASS R package (83) and multiple comparisons were performed using the `emmeans` R package (84).

Classifiers for predicting frameshift mutations in the *agr* operon:

R package `caret` (85) was used to train classifiers using repeated k-fold cross validation (10-fold, 3 repeats). The training dataset comprised 400 randomly sampled frameshift-positive and frameshift-negative strains each to overcome imbalanced representation of each class (24:1). Strain metadata was obtained from Staphopia and only features annotated in > 25% of all strains were included. Strains with unknown host status and host body site data were filtered out. The final dataset contained 11500 frameshift-negative and 486 frameshift positive strains.

Dereplication of Staphopia genomes:

CC5, CC8, CC22 and CC30 genomes were separated into frameshift+ and wild-type groups within each CC based on the presence/absence of frameshift mutations in the *agr* operon. For each CC, the two groups were independently clustered using a Mash (55) distance threshold of 0.0005 and a representative for each cluster was chosen at random. Each frameshift+ cluster was limited to a size of 50 genomes and each wild-type cluster was limited to a size of 200 genomes to produce evenly sized clusters and to prevent underrepresentation of frameshift+ genomes. In total, 1093 genomes represented CC5, 1110 CC8, 404 CC22 and 705 CC30.

Simulating mutant *agr* operons:

A combined total of 312 genomes from the dereplicated set of CC5, CC8, CC22 and CC30

genomes were frameshift+, which equated to 312 mutational events as each genome in the dereplicated set contained only one frameshift in the *agr* operon. To simulate a similar number of mutations, we used Mutation-Simulator v2.0.3 (<https://github.com/mkpython3/Mutation-Simulator>) to induce insertions or deletions at a rate of 0.0002 (parameters: --insert 0.0002 --deletion 0.0002) in 350 randomly selected wild-type genomes from the dereplicated set, leading to 307 simulated indels.

Calculating consistency indices:

Core genome alignments of the dereplicated genomes for each CC was performed using parsnip v1.5.3 (79) and maximum likelihood phylogenetic trees were constructed using IQ-TREE v1.6.12 (56) using the GTR+FO model with 1000 ultrafast bootstrap replicates. Java version of HomoplasyFinder was fed the phylogeny and a presence absence matrix of *agr* operon frameshift positions to obtain the consistency index for each position. The phylogenies for each CC were then imported to R using the ggtree package (80) and the tip labels were randomized 100x to produce 100 shuffled trees. The consistency indices for *agr* operon frameshift positions were calculated for all shuffled trees in the same fashion. Kolmogorov-Smirnov test was used to compare the distributions of consistency indices using the R function ks.test().

Sputum sample collection and whole genome sequencing:

The whole genome sequences for 64 out of the 88 CF strains analysed in this study are associated with a previous publication and can be found in the accession PRJNA480016 (86). The remaining 24 strains are from sputum samples provided by 3 CF patients and can be found in the accession PRJNA742745. The methods for processing these 24 strains are as follows:

Sputum samples were collected from patients at the Emory Adult Cystic Fibrosis Center and spread onto Mannitol Salt Agar (MSA) the same day. Both volumes of 10 μ L or 100 μ L of resuspended sputum were plated for each sample. Three sputum samples from three different patients were collected and processed as mentioned above in the laboratory of Dr. Stephen P. Diggle at Georgia Institute of Technology. From each sample, 6-8 single colony isolates were picked, grown overnight in Luria Broth media, and re-streaked on Staphylococcus isolation agar (SIA) for further purification before being frozen with 25% glycerol and stored at -80°C. SIA agar is composed of 30g L^{-1} Trypticase Soy Broth, 15g L^{-1} agar, and 70g L^{-1} NaCl. At least one ‘pool’ or population sample was collected per patient by scraping all remaining colonies on a single inoculation loop, resuspending the collected colonies in Luria Broth and incubating the liquid culture overnight at 37 °C. Overnight cultures of population samples were also further purified on SIA before being made into frozen stocks. Population samples are always recovered by scraping the entire plate, never as single colonies, throughout the rest of the experiments. Haemolysis phenotyping was conducted for all single colony isolates and population samples using Congo-Red agar as previously described (86).

To extract genomic DNA for sequencing, each sample was streaked on SIA. An inoculation loop was used to collect cells directly from the plate and one loop-full of cells was suspended in 50 mM EDTA. To lyse the cells, 20 μ L of freshly prepared 10mg mL^{-1} lysozyme, and 100 μ L of 5mg mL^{-1} lysostaphin were added to the cell mixtures which were then incubated for 1hr at 37°C. Genomic DNA was then extracted using the Promega®, Wizard Genomic DNA Purification Kit. All samples were sequenced at the Microbial Genome Sequencing Center (Pittsburgh, PA, USA) using the Illumina Nextera kit on the NextSeq 550 platform. Single colony isolates were sequenced at a depth of 150Mb and population samples were sequenced at a depth of 625Mb. Raw paired-end sequence files were screened for quality and minimum length using FastQC v0.11.9 (87). Raw sequence files were then fed into the

Bactopia analysis pipeline version 1.4.10 (88). Bactopia output was used to determine sequence type and clonal complex identities for each sample. Assemblies produced by Bactopia were then analysed in AgrVATE for *agr* type and frameshift status.

Data availability

Source code for AgrVATE as well as the R code and datasets for generating the figures in this study can be found in <https://github.com/VishnuRaghuram94/AgrVATE>.

Acknowledgements

Bacterial isolates and human subject samples were provided by the CF Biospecimen Repository at the Children's Healthcare of Atlanta and Emory University CF Discovery Core. We thank the Diggle Lab at Georgia institute of Technology for obtaining the Cystic Fibrosis sputum samples and isolating the *S. aureus* populations. We also thank Rachel Done, Dina Moustafa, Justin Luu and Brooke Talbot for providing constructive comments on the manuscript.

Vishnu Raghuram and Timothy D. Read were supported by grant # AI139188 from the National Institute of Health (NIH). Ashley M. Alexander was supported by the Infectious Disease Across Scales Training Program at Emory University. This work was also supported by grants from the Cystic Fibrosis Foundation (GOLDBE19P0) and the NIH (R21AI48847) to Joanna B. Goldberg.

Fig 2: AgrA evolves independently of agr group with only two major amino acid sequence configurations across *S. aureus*. (A) Scoring agr group exclusivity in clusters of unique AgrABC amino acid sequences. Extracted amino acid sequences of each agr gene were clustered with 100% identity to obtain all possible AA sequence configurations. Each cluster was then scored based on the number of agr groups the cluster sequence was found in (1 = One agr group, 4 = Four agr groups) represented by a circle. Only clusters with more than 50 sequences are shown. The colour of each circle represents the number of sequences within the cluster. The red and blue arrows indicate the major (AgrA-K) and minor cluster (AgrA-R) of AgrA AA sequences respectively. (B) Amino acid sequence alignment of the two major alleles of AgrA. (C) Maximum likelihood phylogeny (GTR+FO model, 1000 ultrafast bootstrap replicates with average bootstrap support of 97.8%) of 355 *S. aureus* strains with each tip representing a unique ST. Tip colours represent the AgrA alleles and the corresponding heatmap shows the agr group of each tip. (D) Linkage disequilibrium (LD) block plot of the agr operon and 1000bp flanking regions. Each point on the block indicates R² values of LD calculated by plink for a given pair of SNPs. Y axis indicates distance between SNP pairs.

Fig 3: Presence of putative non-functional variants of the *agr* operon. (A) Frequency of frameshift mutations in coding regions of unique *agr* operon sequences across Staphylococcus aureus. Arrows indicate *agr* genes and bars indicate number of frameshifts at the corresponding position (binwidth = 40). Bar colours represent each *agr* group. (B) Frequency (RIGHT) and Effect (LEFT) of commonly occurring frameshift mutations across unique *agr* operon sequences. Bars are coloured based on *agr* group and arrows are coloured based on *agr* gene, black outlines represent canonical protein length and red outlines represent truncated protein lengths. Labels (CENTER) indicate the amino acid change due to the frameshift mutation. (C) Percentage of samples with non-canonical Two component regulator (TCS) gene lengths. Histidine kinase (HK) and response regulator (RR) genes of TCS were extracted from Staphylococcus aureus and commonly occurring gene lengths (> 5000 genomes) were excluded. The remaining strains were considered to have non-canonical gene lengths.

Fig 4: Identical *agr* mutations evolve independent of phylogeny across different clonal complexes. (A) Bars show frequency of mutations at a given site in descending order. LEFT – frequency of each mutation in a dereplicated set of CC8, CC30, CC22 and CC5 samples. RIGHT – frequency of each mutation in a set of randomly selected CC8/CC30/CC22/CC5 *agr* operons with simulated indels. (B) Consistency index vs number of occurrences of frameshift mutations. Each circle represents a position on the *agr* operon that has acquired a mutation in at least 2 samples in a dereplicated set of CC8, CC30, CC22 and CC5 sequences. The consistency index was measured at these sites for each CC from a given phylogenetic tree (GTR+FO model, 1000 ultrafast bootstrap replicates with average bootstrap support of at least 71%) Blue line follows $y=1/x$ distribution.

Fig S1: The AgrVATE pipeline is described for a single input (fasta file) genome assembly of *S. aureus*. Briefly, AgrVATE accepts a *S. aureus* genome assembly as input and performs a kmer search using an *agr* group-specific kmer database to assign the *agr* group. The *agr* operon is then extracted using *in-silico* PCR and variants are called using an *agr* group specific reference operon. If the *agr* group is untypeable by AgrVATE, nhmmer search is performed to detect the presence of *agrD*, after which *ad hoc* bioinformatic steps are performed to identify the *agr* group.

Fig S2: Strain CFBR_17 shows weak haemolysis activity on Sheep Blood Agar plate. CAMP test as described in Traber et al, 2008 (16) was performed to verify the haemolysis status of strain CFBR_17. USA300 JE2 strain was used as a haemolysis positive control, strain RN4220 was used as the vertical haemolysis negative strain and strain CFBR_15, a previously observed haemolysis negative strain (52), was used as a secondary negative control.

Fig S3: *agr* typing status of 42,999 genomes from Staphopia. All possible outcomes from AgrVATE were analysed further to determine reason(s) behind failed/uncertain *agr* group assignments. BLAST searches for individual *agr* genes and AIP sequences were performed to identify presence of *agr* genes. Contig length and coverage information were obtained from fasta header lines from SPAdes assemblies which are a part of the Staphopia analysis pipeline (45, 89). Accession AP018562 was used as a reference for *S. argenteus* ST2250 strains (90).

Biosample	Assembly	Strain ID	Patient ID	agr group	Match score	No. of agr	Frameshift	sheep blood haemolysis
SAMN09847804	GCA_003720255	BCH-SA-01	1	2	15	s	0	-
SAMN09847805	GCA_003720945	BCH-SA-02	2	2	15	s	1	-
SAMN09847806	GCA_003721095	BCH-SA-03	3	1	13	s	0	+
SAMN09847807	GCA_003721055	BCH-SA-04	4	3	15	s	0	+
SAMN09847808	GCA_003721035	BCH-SA-05	5	2	15	s	0	+
SAMN09847809	GCA_003721075	BCH-SA-06	5	2	15	s	u	+
SAMN09847810	GCA_003720995	BCH-SA-07	6	1	13	s	0	+
SAMN09847811	GCA_003721005	BCH-SA-08	7	3	15	s	u	+
SAMN09847812	GCA_003720975	BCH-SA-09	8	1	13	s	0	+
SAMN09847813	GCA_003720935	BCH-SA-10	9	1	13	s	u	-
SAMN09847814	GCA_003720915	BCH-SA-11	10	3	15	s	0	-
SAMN09847815	GCA_003720885	BCH-SA-12	11	2	15	s	0	+
SAMN09847816	GCA_003720875	BCH-SA-13	12	1	13	s	0	+
SAMN09847817	GCA_003720845	BCH-SA-14	13	1	13	s	0	+
SAMN09847818	GCA_003720835	BCH-SA-15	14	1	13	s	0	+
SAMN09847832	GCA_003720355	CFBR_01	CFBR-122	2	14	s	0	+
SAMN09847836	GCA_003720485	CFBR_02	CFBR-148	2	15	s	0	-
SAMN09847841	GCA_003720445	CFBR_03	CFBR-153	1	13	s	1	-
SAMN09847844	GCA_003720415	CFBR_04	CFBR-172	3	15	s	0	-
SAMN09847850	GCA_003720085	CFBR_05	CFBR-238	2	15	s	0	+
SAMN09847840	GCA_003720375	CFBR_06	CFBR-152	2	15	s	0	-
SAMN09847842	GCA_003720385	CFBR_07	CFBR-170	2	15	s	0	-
SAMN09847846	GCA_003720235	CFBR_08	CFBR-196	2	15	s	1	-
SAMN09847849	GCA_003720325	CFBR_09	CFBR-219	2	15	s	0	+
SAMN09847834	GCA_003720305	CFBR_10	CFBR-134	2	15	s	0	+
SAMN09847823	GCA_003720245	CFBR_11	CFBR-101	2	15	s	0	-
SAMN09847845	GCA_003719965	CFBR_12	CFBR-178	2	15	s	0	-
SAMN09847848	GCA_003719845	CFBR_13	CFBR-213	1	13	s	0	+
SAMN09847856	GCA_003720055	CFBR_14	CFBR-316	1	13	s	1	-
SAMN09847835	GCA_003720025	CFBR_15	CFBR-146	2	15	s	1	-
SAMN09847830	GCA_003720205	CFBR_16	CFBR-105	2	15	s	0	+
SAMN09847824	GCA_003720115	CFBR_17	CFBR-102	2	15	s	1	+
SAMN09847831	GCA_003720195	CFBR_18	CFBR-120	1	13	s	0	+
SAMN09847833	GCA_003719855	CFBR_19	CFBR-123	2	15	s	0	+
SAMN09847837	GCA_003720135	CFBR_20	CFBR-149	2	14	s	0	+
SAMN09847838	GCA_003720015	CFBR_21	CFBR-150	2	15	s	1	-
SAMN09847839	GCA_003719995	CFBR_22	CFBR-150	2	15	s	1	-
SAMN09847843	GCA_003719905	CFBR_23	CFBR-171	2	15	s	0	+
SAMN09847847	GCA_003719945	CFBR_24	CFBR-201	2	15	s	0	+
SAMN09847819	GCA_003720795	CFBR_25	CFBR-101	2	15	s	0	+
SAMN09847820	GCA_003720815	CFBR_26	CFBR-101	2	15	s	0	+
SAMN09847821	GCA_003719895	CFBR_27	CFBR-101	1	13	m	0	+
SAMN09847822	GCA_003720745	CFBR_28	CFBR-101	2	15	s	0	+
SAMN09847825	GCA_003431365	CFBR_29	CFBR-105	2	15	s	0	+
SAMN09847826	GCA_003720435	CFBR_30	CFBR-105	2	15	s	0	+
SAMN09847827	GCA_003720775	CFBR_31	CFBR-105	2	15	s	0	+
SAMN09847828	GCA_003720715	CFBR_32	CFBR-105	2	15	s	0	+
SAMN09847829	GCA_003720735	CFBR_33	CFBR-105	2	15	s	0	+
SAMN09847851	GCA_003720685	CFBR_34	CFBR-280	2	15	s	0	-
SAMN09847852	GCA_003720675	CFBR_35	CFBR-280	1	13	s	0	+
SAMN09847853	GCA_003720635	CFBR_36	CFBR-280	2	15	s	0	-
SAMN09847854	GCA_003720655	CFBR_37	CFBR-309	3	15	s	0	+
SAMN09847855	GCA_003720575	CFBR_38	CFBR-314	1	13	s	0	+
SAMN09847857	GCA_003720605	CFBR_39	CFBR-322	3	15	s	u	+
SAMN09847858	GCA_003720595	CFBR_40	CFBR-336	2	15	s	0	+
SAMN09847859	GCA_003720535	CFBR_41	CFBR-429	1	13	s	1	-
SAMN09847860	GCA_003720125	CFBR_42	CFBR-430	1	12	s	0	+
SAMN09847861	GCA_003720295	CFBR_43	CFBR-447	1	13	s	0	+
SAMN09847862	GCA_003720075	CFBR_44	CFBR-487	1	13	s	0	+
SAMN09847863	GCA_003721195	CFBR_45	CFBR-487	1	13	s	0	+
SAMN09847864	GCA_003720175	CFBR_46	CFBR-509	1	13	s	0	+
SAMN09847865	GCA_003720545	CFBR_47	CFBR-515	1	13	s	0	+
SAMN09847866	GCA_003720515	CFBR_48	CFBR-530	1	13	s	0	+
SAMN09847867	GCA_003720475	CFBR_49	CFBR-573	1	13	s	0	+

672

673

Continued next page

Biosample	Assembly	Strain ID	Patient ID	agr group	Match score	No. of agr	Frameshift	sheep blood haemolysis
PRJNA742745*	TBD*	CFBR_EB_Sa90	CFBR-623	1	13	s	0	+
PRJNA742745*	TBD*	CFBR_EB_Sa91	CFBR-623	1	13	s	0	+
PRJNA742745*	TBD*	CFBR_EB_Sa92	CFBR-623	1	13	s	0	+
PRJNA742745*	TBD*	CFBR_EB_Sa93	CFBR-623	1	13	s	0	+
PRJNA742745*	TBD*	CFBR_EB_Sa95	CFBR-623	1	13	s	0	+
PRJNA742745*	TBD*	CFBR_EB_Sa97	CFBR-623	1	13	s	0	+
PRJNA742745*	TBD*	CFBR_EB_Sa100	CFBR-196	2	15	s	0	+
PRJNA742745*	TBD*	CFBR_EB_Sa101	CFBR-196	2	15	s	0	+
PRJNA742745*	TBD*	CFBR_EB_Sa102	CFBR-196	2	15	s	0	+
PRJNA742745*	TBD*	CFBR_EB_Sa103	CFBR-196	2	15	s	0	+
PRJNA742745*	TBD*	CFBR_EB_Sa104	CFBR-196	2	15	s	0	+
PRJNA742745*	TBD*	CFBR_EB_Sa105	CFBR-196	2	15	s	0	+
PRJNA742745*	TBD*	CFBR_EB_Sa106	CFBR-196	2	15	s	0	+
PRJNA742745*	TBD*	CFBR_EB_Sa108	CFBR-196	2	15	s	0	+
PRJNA742745*	TBD*	CFBR_EB_Sa109	CFBR-196	2	15	s	0	+
PRJNA742745*	TBD*	CFBR_EB_Sa110	CFBR-311	2	15	m	u	+
PRJNA742745*	TBD*	CFBR_EB_Sa111	CFBR-311	2	15	s	1	-
PRJNA742745*	TBD*	CFBR_EB_Sa112	CFBR-311	2	15	s	1	-
PRJNA742745*	TBD*	CFBR_EB_Sa113	CFBR-311	2	15	s	1	-
PRJNA742745*	TBD*	CFBR_EB_Sa114	CFBR-311	2	15	s	1	-
PRJNA742745*	TBD*	CFBR_EB_Sa115	CFBR-311	2	15	s	1	-
PRJNA742745*	TBD*	CFBR_EB_Sa116	CFBR-311	2	15	s	1	-
PRJNA742745*	TBD*	CFBR_EB_Sa117	CFBR-311	1	13	s	0	+
PRJNA742745*	TBD*	CFBR_EB_Sa118	CFBR-311	1	13	s	0	+

Table S1: AgrVATE and sheep blood haemolysis results from 88 *S. aureus* samples isolated from CF patients. Columns 1 and 2 correspond to the Biosample accession number and Genbank assembly accession number respectively. Columns 3 and 4 correspond to strain identifier and patient identifier of each sample respectively. Columns 5-9 are direct outputs of AgrVATE (<https://github.com/VishnuRaguram94/AgrVATE#outputs>) and are as follows: **Column 5: agr group as designated by AgrVATE, **Column 6:** Match score of the designated agr group (can range from 0 to 15), **Column 7:** Number of agr groups found – s: single, m: multiple, **Column 8:** Presence/absence of frameshifts – 1: present, 0: absent, u: unknown. **Column 9** represents the haemolysis status on Sheep Blood Agar – + : haemolysis positive, - : haemolysis negative.**

* Assemblies will be uploaded under the Bioproject accession PRJNA742745

	GLM	RF	XGB	KNN
Sensitivity	0.561	0.617	0.607	0.645
Specificity	0.513	0.572	0.555	0.516
Pos Pred Value	0.012	0.014	0.014	0.013
Neg Pred Value	0.991	0.993	0.993	0.993
Precision	0.012	0.014	0.014	0.013
Recall	0.561	0.617	0.607	0.645
F1	0.023	0.028	0.027	0.026
Prevalence	0.01	0.01	0.01	0.01
Detection Rate	0.006	0.006	0.006	0.006
Detection Prevalence	0.487	0.43	0.447	0.486
Balanced Accuracy	0.537	0.594	0.581	0.58

Table S2: Classifiers used to predict presence/absence of *agr* frameshifts using *agr* group, clonal complex, host body site and host infection/colonization status. GLM – General Linear Model, RF – Random Forest, XGB – Extreme Gradient Boost, KNN – K nearest neighbours. Models were trained using the R package caret with a randomly sampled dataset comprising 400 frameshift+ and 400 frameshift- samples. Model summary parameters were computed using the confusionMatrix() function.

References

1. Kourtis AP, Hatfield K, Baggs J, Mu Y, See I, Epton E, Nadle J, Kainer MA, Dumyati G, Petit S, Ray SM, Emerging Infections Program Mag, Ham D, Capers C, Ewing H, Coffin N, McDonald LC, Jernigan J, Cardo D. Vital Signs: Epidemiology and Recent Trends in Methicillin-Resistant and in Methicillin-Susceptible *Staphylococcus aureus* Bloodstream Infections - United States. MMWR Morb Mortal Wkly Rep. 2019;68(9):214-9. Epub 2019/03/08. doi: 10.15585/mmwr.mm6809e1. PubMed PMID: 30845118; PMCID: PMC6421967 potential conflicts of interest. No potential conflicts of interest were disclosed.
2. David MZ, Daum RS. Treatment of *Staphylococcus aureus* Infections. Curr Top Microbiol Immunol. 2017;409:325-83. Epub 2017/09/14. doi: 10.1007/82_2017_42. PubMed PMID: 28900682.
3. Tuchscher L, Löffler B. *Staphylococcus aureus* dynamically adapts global regulators and virulence factor expression in the course from acute to chronic infection. Curr Genet. 2016;62(1):15-7. Epub 2015/07/01. doi: 10.1007/s00294-015-0503-0. PubMed PMID: 26123224.
4. Garcia-Betancur JC, Goni-Moreno A, Horger T, Schott M, Sharan M, Eikmeier J, Wohlmuth B, Zernecke A, Ohlsen K, Kuttler C, Lopez D. Cell differentiation defines acute and chronic infection cell types in *Staphylococcus aureus*. Elife. 2017;6. Epub 2017/09/13. doi: 10.7554/eLife.28023. PubMed PMID: 28893374; PMCID: PMC5595439.
5. Novick RP. Autoinduction and signal transduction in the regulation of staphylococcal virulence. Mol Microbiol. 2003;48(6):1429-49. Epub 2003/06/07. doi: 10.1046/j.1365-2958.2003.03526.x. PubMed PMID: 12791129.
6. Novick RP, Ross HF, Projan SJ, Kornblum J, Kreiswirth B, Moghazeh S. Synthesis of staphylococcal virulence factors is controlled by a regulatory RNA molecule. EMBO J. 1993;12(10):3967-75. Epub 1993/10/01. PubMed PMID: 7691599; PMCID: PMC413679.
7. Novick RP, Projan SJ, Kornblum J, Ross HF, Ji G, Kreiswirth B, Vandenesch F, Moghazeh S. The agr P2 operon: an autocatalytic sensory transduction system in *Staphylococcus aureus*. Mol Gen Genet. 1995;248(4):446-58. Epub 1995/08/30. doi: 10.1007/BF02191645. PubMed PMID: 7565609.
8. Ji G, Beavis RC, Novick RP. Cell density control of staphylococcal virulence mediated by an octapeptide pheromone. Proc Natl Acad Sci U S A. 1995;92(26):12055-9. Epub 1995/12/19. doi: 10.1073/pnas.92.26.12055. PubMed PMID: 8618843; PMCID: PMC40295.
9. Lina G, Jarraud S, Ji G, Greenland T, Pedraza A, Etienne J, Novick RP, Vandenesch F. Transmembrane topology and histidine protein kinase activity of AgrC, the agr signal receptor in *Staphylococcus aureus*. Mol Microbiol. 1998;28(3):655-62. Epub 1998/06/19. doi: 10.1046/j.1365-2958.1998.00830.x. PubMed PMID: 9632266.
10. Dufour P, Jarraud S, Vandenesch F, Greenland T, Novick RP, Bes M, Etienne J, Lina G. High genetic variability of the agr locus in *Staphylococcus* species. J Bacteriol. 2002;184(4):1180-6. Epub 2002/01/25. doi: 10.1128/jb.184.4.1180-1186.2002. PubMed PMID: 11807079; PMCID: PMC134794.

11. Surewaard BG, de Haas CJ, Vervoort F, Rigby KM, DeLeo FR, Otto M, van Strijp JA, Nijland R. Staphylococcal alpha-phenol soluble modulins contribute to neutrophil lysis after phagocytosis. *Cell Microbiol.* 2013;15(8):1427-37. Epub 2013/03/09. doi: 10.1111/cmi.12130. PubMed PMID: 23470014; PMCID: PMC4784422.
12. Li M, Cheung GY, Hu J, Wang D, Joo HS, Deleo FR, Otto M. Comparative analysis of virulence and toxin expression of global community-associated methicillin-resistant *Staphylococcus aureus* strains. *J Infect Dis.* 2010;202(12):1866-76. Epub 2010/11/06. doi: 10.1086/657419. PubMed PMID: 21050125; PMCID: PMC3058913.
13. Bhakdi S, Tranum-Jensen J. Alpha-toxin of *Staphylococcus aureus*. *Microbiol Rev.* 1991;55(4):733-51. Epub 1991/12/01. PubMed PMID: 1779933; PMCID: PMC372845.
14. Berube BJ, Sampedro GR, Otto M, Bubeck Wardenburg J. The *psm* locus regulates production of *Staphylococcus aureus* alpha-toxin during infection. *Infect Immun.* 2014;82(8):3350-8. Epub 2014/05/29. doi: 10.1128/IAI.00089-14. PubMed PMID: 24866799; PMCID: PMC4136214.
15. Thammavongsa V, Kim HK, Missiakas D, Schneewind O. Staphylococcal manipulation of host immune responses. *Nat Rev Microbiol.* 2015;13(9):529-43. Epub 2015/08/15. doi: 10.1038/nrmicro3521. PubMed PMID: 26272408; PMCID: PMC4625792.
16. Traber KE, Lee E, Benson S, Corrigan R, Cantera M, Shopsin B, Novick RP. *agr* function in clinical *Staphylococcus aureus* isolates. *Microbiology (Reading).* 2008;154(Pt 8):2265-74. Epub 2008/08/01. doi: 10.1099/mic.0.2007/011874-0. PubMed PMID: 18667559; PMCID: PMC4904715.
17. Painter KL, Krishna A, Wigneshweraraj S, Edwards AM. What role does the quorum-sensing accessory gene regulator system play during *Staphylococcus aureus* bacteremia? *Trends Microbiol.* 2014;22(12):676-85. Epub 2014/10/11. doi: 10.1016/j.tim.2014.09.002. PubMed PMID: 25300477.
18. Chong YP, Kim ES, Park SJ, Park KH, Kim T, Kim MN, Kim SH, Lee SO, Choi SH, Woo JH, Jeong JY, Kim YS. Accessory gene regulator (*agr*) dysfunction in *Staphylococcus aureus* bloodstream isolates from South Korean patients. *Antimicrob Agents Chemother.* 2013;57(3):1509-12. Epub 2012/12/21. doi: 10.1128/AAC.01260-12. PubMed PMID: 23254438; PMCID: PMC3591919.
19. Lee SO, Lee S, Lee JE, Song KH, Kang CK, Wi YM, San-Juan R, Lopez-Cortes LE, Lacoma A, Prat C, Jang HC, Kim ES, Kim HB, Lee SH. Dysfunctional accessory gene regulator (*agr*) as a prognostic factor in invasive *Staphylococcus aureus* infection: a systematic review and meta-analysis. *Sci Rep.* 2020;10(1):20697. Epub 2020/11/28. doi: 10.1038/s41598-020-77729-0. PubMed PMID: 33244173; PMCID: PMC7691521.
20. Schweizer ML, Furuno JP, Sakoulas G, Johnson JK, Harris AD, Shardell MD, McGregor JC, Thom KA, Perencevich EN. Increased mortality with accessory gene regulator (*agr*) dysfunction in *Staphylococcus aureus* among bacteremic patients. *Antimicrob Agents Chemother.* 2011;55(3):1082-7. Epub 2010/12/22. doi: 10.1128/AAC.00918-10. PubMed PMID: 21173172; PMCID: PMC3067101.
21. Suligoy CM, Lattar SM, Noto Llana M, Gonzalez CD, Alvarez LP, Robinson DA, Gomez MI, Buzzola FR, Sordelli DO. Mutation of *Agr* Is Associated with the Adaptation of *Staphylococcus aureus* to the Host during Chronic Osteomyelitis. *Front Cell Infect Microbiol.* 2018;8:18. Epub 2018/02/20. doi: 10.3389/fcimb.2018.00018. PubMed PMID:

788 29456969; PMCID: PMC5801681.

789 22. Cheung AL, Eberhardt KJ, Chung E, Yeaman MR, Sullam PM, Ramos M, Bayer
790 AS. Diminished virulence of a sar-/agr- mutant of Staphylococcus aureus in the rabbit
791 model of endocarditis. J Clin Invest. 1994;94(5):1815-22. Epub 1994/11/01. doi:
792 10.1172/JCI117530. PubMed PMID: 7962526; PMCID: PMC294579.

793 23. Gillaspay AF, Hickmon SG, Skinner RA, Thomas JR, Nelson CL, Smeltzer MS.
794 Role of the accessory gene regulator (agr) in pathogenesis of staphylococcal osteomyelitis.
795 Infect Immun. 1995;63(9):3373-80. Epub 1995/09/01. doi: 10.1128/IAI.63.9.3373-
796 3380.1995. PubMed PMID: 7642265; PMCID: PMC173464.

797 24. Das S, Lindemann C, Young BC, Muller J, Osterreich B, Ternette N, Winkler AC,
798 Paprotka K, Reinhardt R, Forstner KU, Allen E, Flaxman A, Yamaguchi Y, Rollier CS,
799 van Diemen P, Blattner S, Remmele CW, Selle M, Dittrich M, Muller T, Vogel J, Ohlsen
800 K, Crook DW, Massey R, Wilson DJ, Rudel T, Wyllie DH, Fraunholz MJ. Natural
801 mutations in a Staphylococcus aureus virulence regulator attenuate cytotoxicity but permit
802 bacteremia and abscess formation. Proc Natl Acad Sci U S A. 2016;113(22):E3101-10.
803 Epub 2016/05/18. doi: 10.1073/pnas.1520255113. PubMed PMID: 27185949; PMCID:
804 PMC4896717.

805 25. Shopsis B, Copin R. Staphylococcus aureus Adaptation During Infection. In: Fong
806 IW, Shlaes D, Drlica K, editors. Antimicrobial Resistance in the 21st Century. Cham:
807 Springer International Publishing; 2018. p. 431-59.

808 26. Eldar A. Social conflict drives the evolutionary divergence of quorum sensing.
809 Proc Natl Acad Sci U S A. 2011;108(33):13635-40. Epub 2011/08/03. doi:
810 10.1073/pnas.1102923108. PubMed PMID: 21807995; PMCID: PMC3158151.

811 27. Diard M, Garcia V, Maier L, Remus-Emsermann MN, Regoes RR, Ackermann M,
812 Hardt WD. Stabilization of cooperative virulence by the expression of an avirulent
813 phenotype. Nature. 2013;494(7437):353-6. Epub 2013/02/22. doi: 10.1038/nature11913.
814 PubMed PMID: 23426324.

815 28. Chuang JS, Rivoire O, Leibler S. Simpson's paradox in a synthetic microbial
816 system. Science. 2009;323(5911):272-5. Epub 2009/01/10. doi: 10.1126/science.1166739.
817 PubMed PMID: 19131632.

818 29. Czarán T, Hoekstra RF. Microbial communication, cooperation and cheating:
819 quorum sensing drives the evolution of cooperation in bacteria. PLoS One.
820 2009;4(8):e6655. Epub 2009/08/18. doi: 10.1371/journal.pone.0006655. PubMed PMID:
821 19684853; PMCID: PMC2722019.

822 30. Kahl BC, Becker K, Friedrich AW, Clasen J, Sinha B, Von Eiff C, Peters G. agr-
823 dependent bacterial interference has no impact on long-term colonization of
824 Staphylococcus aureus during persistent airway infection of cystic fibrosis patients. J Clin
825 Microbiol. 2003;41(11):5199-201. Epub 2003/11/08. doi: 10.1128/jcm.41.11.5199-
826 5201.2003. PubMed PMID: 14605162; PMCID: PMC262511.

827 31. Goerke C, Kummel M, Dietz K, Wolz C. Evaluation of intraspecies interference
828 due to agr polymorphism in Staphylococcus aureus during infection and colonization. J
829 Infect Dis. 2003;188(2):250-6. Epub 2003/07/11. doi: 10.1086/376450. PubMed PMID:
830 12854080.

831 32. Ji G, Beavis R, Novick RP. Bacterial interference caused by autoinducing peptide
832 variants. Science. 1997;276(5321):2027-30. Epub 1997/06/27. doi:

- 10.1126/science.276.5321.2027. PubMed PMID: 9197262.
33. Canovas J, Baldry M, Bojer MS, Andersen PS, Grzeskowiak PK, Stegger M, Damborg P, Olsen CA, Ingmer H. Cross-Talk between *Staphylococcus aureus* and Other *Staphylococcal* Species via the *agr* Quorum Sensing System. *Front Microbiol.* 2016;7:1733. Epub 2016/11/24. doi: 10.3389/fmicb.2016.01733. PubMed PMID: 27877157; PMCID: PMC5099252.
34. Otto M, Echner H, Voelter W, Gotz F. Pheromone cross-inhibition between *Staphylococcus aureus* and *Staphylococcus epidermidis*. *Infect Immun.* 2001;69(3):1957-60. Epub 2001/02/17. doi: 10.1128/IAI.69.3.1957-1960.2001. PubMed PMID: 11179383; PMCID: PMC98112.
35. Sakoulas G, Eliopoulos GM, Fowler VG, Jr., Moellering RC, Jr., Novick RP, Lucindo N, Yeaman MR, Bayer AS. Reduced susceptibility of *Staphylococcus aureus* to vancomycin and platelet microbicidal protein correlates with defective autolysis and loss of accessory gene regulator (*agr*) function. *Antimicrob Agents Chemother.* 2005;49(7):2687-92. Epub 2005/06/28. doi: 10.1128/AAC.49.7.2687-2692.2005. PubMed PMID: 15980337; PMCID: PMC1168700.
36. Sieradzki K, Tomasz A. Alterations of cell wall structure and metabolism accompany reduced susceptibility to vancomycin in an isogenic series of clinical isolates of *Staphylococcus aureus*. *J Bacteriol.* 2003;185(24):7103-10. Epub 2003/12/04. doi: 10.1128/jb.185.24.7103-7110.2003. PubMed PMID: 14645269; PMCID: PMC296238.
37. Cheung GY, Kretschmer D, Duong AC, Yeh AJ, Ho TV, Chen Y, Joo HS, Kreiswirth BN, Peschel A, Otto M. Production of an attenuated phenol-soluble modulins variant unique to the MRSA clonal complex 30 increases severity of bloodstream infection. *PLoS Pathog.* 2014;10(8):e1004298. Epub 2014/08/22. doi: 10.1371/journal.ppat.1004298. PubMed PMID: 25144687; PMCID: PMC4140855.
38. Laabei M, Uhlemann AC, Lowy FD, Austin ED, Yokoyama M, Ouadi K, Feil E, Thorpe HA, Williams B, Perkins M, Peacock SJ, Clarke SR, Dordel J, Holden M, Votintseva AA, Bowden R, Crook DW, Young BC, Wilson DJ, Recker M, Massey RC. Evolutionary Trade-Offs Underlie the Multi-faceted Virulence of *Staphylococcus aureus*. *PLoS Biol.* 2015;13(9):e1002229. Epub 2015/09/04. doi: 10.1371/journal.pbio.1002229. PubMed PMID: 26331877; PMCID: PMC4558032.
39. Kumar K, Chen J, Drlica K, Shopsin B. Tuning of the Lethal Response to Multiple Stressors with a Single-Site Mutation during Clinical Infection by *Staphylococcus aureus*. *mBio.* 2017;8(5). Epub 2017/10/27. doi: 10.1128/mBio.01476-17. PubMed PMID: 29066545; PMCID: PMC5654930.
40. Fowler VG, Jr., Sakoulas G, McIntyre LM, Meka VG, Arbeit RD, Cabell CH, Stryjewski ME, Eliopoulos GM, Reller LB, Corey GR, Jones T, Lucindo N, Yeaman MR, Bayer AS. Persistent bacteremia due to methicillin-resistant *Staphylococcus aureus* infection is associated with *agr* dysfunction and low-level in vitro resistance to thrombin-induced platelet microbicidal protein. *J Infect Dis.* 2004;190(6):1140-9. Epub 2004/08/21. doi: 10.1086/423145. PubMed PMID: 15319865.
41. Nakamura Y, Takahashi H, Takaya A, Inoue Y, Katayama Y, Kusuya Y, Shoji T, Takada S, Nakagawa S, Oguma R, Saito N, Ozawa N, Nakano T, Yamaide F, Dissanayake E, Suzuki S, Villaruz A, Varadarajan S, Matsumoto M, Kobayashi T, Kono M, Sato Y, Akiyama M, Otto M, Matsue H, Nunez G, Shimojo N. *Staphylococcus Agr* virulence is critical for epidermal colonization and associates with atopic dermatitis development. *Sci*

- 879 Transl Med. 2020;12(551). Epub 2020/07/10. doi: 10.1126/scitranslmed.aay4068.
880 PubMed PMID: 32641488; PMCID: PMC7426015.
- 881 42. Shopsin B, Drlica-Wagner A, Mathema B, Adhikari RP, Kreiswirth BN, Novick
882 RP. Prevalence of agr dysfunction among colonizing Staphylococcus aureus strains. J
883 Infect Dis. 2008;198(8):1171-4. Epub 2008/08/30. doi: 10.1086/592051. PubMed PMID:
884 18752431.
- 885 43. Sakoulas G, Moise PA, Rybak MJ. Accessory gene regulator dysfunction: an
886 advantage for Staphylococcus aureus in health-care settings? J Infect Dis.
887 2009;199(10):1558-9. Epub 2009/04/28. doi: 10.1086/598607. PubMed PMID: 19392634.
- 888 44. Shopsin B, Eaton C, Wasserman GA, Mathema B, Adhikari RP, Agolory S,
889 Altman DR, Holzman RS, Kreiswirth BN, Novick RP. Mutations in agr do not persist in
890 natural populations of methicillin-resistant Staphylococcus aureus. J Infect Dis.
891 2010;202(10):1593-9. Epub 2010/10/15. doi: 10.1086/656915. PubMed PMID: 20942648.
- 892 45. Petit RA, 3rd, Read TD. Staphylococcus aureus viewed from the perspective of
893 40,000+ genomes. PeerJ. 2018;6:e5261. Epub 2018/07/18. doi: 10.7717/peerj.5261.
894 PubMed PMID: 30013858; PMCID: PMC6046195.
- 895 46. Choudhary KS, Mih N, Monk J, Kavvas E, Yurkovich JT, Sakoulas G, Palsson
896 BO. The Staphylococcus aureus Two-Component System AgrAC Displays Four Distinct
897 Genomic Arrangements That Delineate Genomic Virulence Factor Signatures. Front
898 Microbiol. 2018;9:1082. Epub 2018/06/12. doi: 10.3389/fmicb.2018.01082. PubMed
899 PMID: 29887846; PMCID: PMC5981134.
- 900 47. Shopsin B, Mathema B, Alcabes P, Said-Salim B, Lina G, Matsuka A, Martinez J,
901 Kreiswirth BN. Prevalence of agr specificity groups among Staphylococcus aureus strains
902 colonizing children and their guardians. J Clin Microbiol. 2003;41(1):456-9. Epub
903 2003/01/09. doi: 10.1128/jcm.41.1.456-459.2003. PubMed PMID: 12517893; PMCID:
904 PMC149583.
- 905 48. Francois P, Koessler T, Huyghe A, Harbarth S, Bento M, Lew D, Etienne J, Pittet
906 D, Schrenzel J. Rapid Staphylococcus aureus agr type determination by a novel multiplex
907 real-time quantitative PCR assay. J Clin Microbiol. 2006;44(5):1892-5. Epub 2006/05/05.
908 doi: 10.1128/JCM.44.5.1892-1895.2006. PubMed PMID: 16672433; PMCID:
909 PMC1479209.
- 910 49. Gruning B, Dale R, Sjodin A, Chapman BA, Rowe J, Tomkins-Tinch CH, Valieris
911 R, Koster J, Bioconda T. Bioconda: sustainable and comprehensive software distribution
912 for the life sciences. Nat Methods. 2018;15(7):475-6. Epub 2018/07/04. doi:
913 10.1038/s41592-018-0046-7. PubMed PMID: 29967506.
- 914 50. Edgar RC. Search and clustering orders of magnitude faster than BLAST.
915 Bioinformatics. 2010;26(19):2460-1. Epub 2010/08/17. doi:
916 10.1093/bioinformatics/btq461. PubMed PMID: 20709691.
- 917 51. Seemann T. snippy: fast bacterial variant calling from NGS reads. 4.6 ed.
918 <https://github.com/tseemann/snippy>: Github; 2015.
- 919 52. Bernardy EE, Petit RA, 3rd, Raghuram V, Alexander AM, Read TD, Goldberg JB.
920 Genotypic and Phenotypic Diversity of Staphylococcus aureus Isolates from Cystic
921 Fibrosis Patient Lung Infections and Their Interactions with Pseudomonas aeruginosa.
922 mBio. 2020;11(3). Epub 2020/06/25. doi: 10.1128/mBio.00735-20. PubMed PMID:
923 32576671; PMCID: PMC7315118.

53. Wright JS, 3rd, Traber KE, Corrigan R, Benson SA, Musser JM, Novick RP. The agr radiation: an early event in the evolution of staphylococci. *J Bacteriol.* 2005;187(16):5585-94. Epub 2005/08/04. doi: 10.1128/JB.187.16.5585-5594.2005. PubMed PMID: 16077103; PMCID: PMC1196086.
54. Peacock SJ, Moore CE, Justice A, Kantzanou M, Story L, Mackie K, O'Neill G, Day NP. Virulent combinations of adhesin and toxin genes in natural populations of *Staphylococcus aureus*. *Infect Immun.* 2002;70(9):4987-96. Epub 2002/08/17. doi: 10.1128/IAI.70.9.4987-4996.2002. PubMed PMID: 12183545; PMCID: PMC128268.
55. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 2016;17(1):132. Epub 2016/06/22. doi: 10.1186/s13059-016-0997-x. PubMed PMID: 27323842; PMCID: PMC4915045.
56. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol.* 2020;37(5):1530-4. Epub 2020/02/06. doi: 10.1093/molbev/msaa015. PubMed PMID: 32011700; PMCID: PMC7182206.
57. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81(3):559-75. Epub 2007/08/19. doi: 10.1086/519795. PubMed PMID: 17701901; PMCID: PMC1950838.
58. Traber K, Novick R. A slipped-mispairing mutation in AgrA of laboratory strains and clinical isolates results in delayed activation of agr and failure to translate delta- and alpha-haemolysins. *Mol Microbiol.* 2006;59(5):1519-30. Epub 2006/02/14. doi: 10.1111/j.1365-2958.2006.04986.x. PubMed PMID: 16468992.
59. Lyon BR, Gillespie MT, Skurray RA. Detection and characterization of IS256, an insertion sequence in *Staphylococcus aureus*. *J Gen Microbiol.* 1987;133(11):3031-8. Epub 1987/11/01. doi: 10.1099/00221287-133-11-3031. PubMed PMID: 2833560.
60. Heaton MP, Discotto LF, Pucci MJ, Handwerger S. Mobilization of vancomycin resistance by transposon-mediated fusion of a VanA plasmid with an *Enterococcus faecium* sex pheromone-response plasmid. *Gene.* 1996;171(1):9-17. Epub 1996/05/24. doi: 10.1016/0378-1119(96)00022-4. PubMed PMID: 8675038.
61. Crispell J, Balaz D, Gordon SV. HomoplasyFinder: a simple tool to identify homoplasies on a phylogeny. *Microb Genom.* 2019;5(1). Epub 2019/01/22. doi: 10.1099/mgen.0.000245. PubMed PMID: 30663960; PMCID: PMC6412054.
62. Novick RP, Geisinger E. Quorum sensing in staphylococci. *Annu Rev Genet.* 2008;42:541-64. Epub 2008/08/21. doi: 10.1146/annurev.genet.42.110807.091640. PubMed PMID: 18713030.
63. Robinson DA, Monk AB, Cooper JE, Feil EJ, Enright MC. Evolutionary genetics of the accessory gene regulator (agr) locus in *Staphylococcus aureus*. *J Bacteriol.* 2005;187(24):8312-21. Epub 2005/12/03. doi: 10.1128/JB.187.24.8312-8321.2005. PubMed PMID: 16321935; PMCID: PMC1317016.
64. Planet PJ, Narechania A, Chen L, Mathema B, Boundy S, Archer G, Kreiswirth B. Architecture of a Species: Phylogenomics of *Staphylococcus aureus*. *Trends Microbiol.* 2017;25(2):153-66. Epub 2016/10/19. doi: 10.1016/j.tim.2016.09.009. PubMed PMID:

27751626.

65. Feil EJ. Small change: keeping pace with microevolution. *Nat Rev Microbiol.* 2004;2(6):483-95. Epub 2004/05/21. doi: 10.1038/nrmicro904. PubMed PMID: 15152204.

66. Tsuji BT, Rybak MJ, Cheung CM, Amjad M, Kaatz GW. Community- and health care-associated methicillin-resistant *Staphylococcus aureus*: a comparison of molecular epidemiology and antimicrobial activities of various agents. *Diagn Microbiol Infect Dis.* 2007;58(1):41-7. Epub 2007/02/16. doi: 10.1016/j.diagmicrobio.2006.10.021. PubMed PMID: 17300912.

67. Smyth DS, Kafer JM, Wasserman GA, Velickovic L, Mathema B, Holzman RS, Knipe TA, Becker K, von Eiff C, Peters G, Chen L, Kreiswirth BN, Novick RP, Shopsin B. Nasal carriage as a source of agr-defective *Staphylococcus aureus* bacteremia. *J Infect Dis.* 2012;206(8):1168-77. Epub 2012/08/04. doi: 10.1093/infdis/jis483. PubMed PMID: 22859823; PMCID: PMC3448967.

68. Wertheim HF, Melles DC, Vos MC, van Leeuwen W, van Belkum A, Verbrugh HA, Nouwen JL. The role of nasal carriage in *Staphylococcus aureus* infections. *Lancet Infect Dis.* 2005;5(12):751-62. Epub 2005/11/29. doi: 10.1016/S1473-3099(05)70295-4. PubMed PMID: 16310147.

69. Laabei M, Peacock SJ, Blane B, Baines SL, Howden BP, Stinear TP, Massey RC. Significant Variability exists in the Toxicity of Global Methicillin-resistant *Staphylococcus aureus* Lineages. *bioRxiv.* 2021:2021.07.16.452633. doi: 10.1101/2021.07.16.452633.

70. Hanssen AM, Kjeldsen G, Sollid JU. Local variants of *Staphylococcal cassette chromosome mec* in sporadic methicillin-resistant *Staphylococcus aureus* and methicillin-resistant coagulase-negative *Staphylococci*: evidence of horizontal gene transfer? *Antimicrob Agents Chemother.* 2004;48(1):285-96. Epub 2003/12/25. doi: 10.1128/AAC.48.1.285-296.2004. PubMed PMID: 14693553; PMCID: PMC310173.

71. Wielders CL, Vriens MR, Brisse S, de Graaf-Miltenburg LA, Troelstra A, Fleer A, Schmitz FJ, Verhoef J, Fluit AC. In-vivo transfer of *mecA* DNA to *Staphylococcus aureus* [corrected]. *Lancet.* 2001;357(9269):1674-5. Epub 2001/06/27. doi: 10.1016/S0140-6736(00)04832-7. PubMed PMID: 11425376.

72. Gill SR, Fouts DE, Archer GL, Mongodin EF, Deboy RT, Ravel J, Paulsen IT, Kolonay JF, Brinkac L, Beanan M, Dodson RJ, Daugherty SC, Madupu R, Angiuoli SV, Durkin AS, Haft DH, Vamathevan J, Khouri H, Utterback T, Lee C, Dimitrov G, Jiang L, Qin H, Weidman J, Tran K, Kang K, Hance IR, Nelson KE, Fraser CM. Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant *Staphylococcus aureus* strain and a biofilm-producing methicillin-resistant *Staphylococcus epidermidis* strain. *J Bacteriol.* 2005;187(7):2426-38. Epub 2005/03/19. doi: 10.1128/JB.187.7.2426-2438.2005. PubMed PMID: 15774886; PMCID: PMC1065214.

73. Bloemendaal AL, Brouwer EC, Fluit AC. Methicillin resistance transfer from *Staphylococcus epidermidis* to methicillin-susceptible *Staphylococcus aureus* in a patient during antibiotic therapy. *PLoS One.* 2010;5(7):e11841. Epub 2010/08/06. doi: 10.1371/journal.pone.0011841. PubMed PMID: 20686601; PMCID: PMC2912275.

74. Berglund C, Soderquist B. The origin of a methicillin-resistant *Staphylococcus aureus* isolate at a neonatal ward in Sweden-possible horizontal transfer of a

- 1015 staphylococcal cassette chromosome mec between methicillin-resistant *Staphylococcus*
1016 *haemolyticus* and *Staphylococcus aureus*. *Clin Microbiol Infect*. 2008;14(11):1048-56.
1017 Epub 2008/12/02. doi: 10.1111/j.1469-0691.2008.02090.x. PubMed PMID: 19040477.
- 1018 75. Zhang DF, Zhi XY, Zhang J, Paoli GC, Cui Y, Shi C, Shi X. Preliminary
1019 comparative genomics revealed pathogenic potential and international spread of
1020 *Staphylococcus argenteus*. *BMC Genomics*. 2017;18(1):808. Epub 2017/10/24. doi:
1021 10.1186/s12864-017-4149-9. PubMed PMID: 29058585; PMCID: PMC5651615.
- 1022 76. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*.
1023 2014;30(14):2068-9. Epub 2014/03/20. doi: 10.1093/bioinformatics/btu153. PubMed
1024 PMID: 24642063.
- 1025 77. Bailey TL. DREME: motif discovery in transcription factor ChIP-seq data.
1026 *Bioinformatics*. 2011;27(12):1653-9. Epub 2011/05/06. doi:
1027 10.1093/bioinformatics/btr261. PubMed PMID: 21543442; PMCID: PMC3106199.
- 1028 78. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden
1029 TL. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421. Epub
1030 2009/12/17. doi: 10.1186/1471-2105-10-421. PubMed PMID: 20003500; PMCID:
1031 PMC2803857.
- 1032 79. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid core-
1033 genome alignment and visualization of thousands of intraspecific microbial genomes.
1034 *Genome Biol*. 2014;15(11):524. Epub 2014/11/21. doi: 10.1186/s13059-014-0524-x.
1035 PubMed PMID: 25410596; PMCID: PMC4262987.
- 1036 80. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: an r package for
1037 visualization and annotation of phylogenetic trees with their covariates and other
1038 associated data. *Methods in Ecology and Evolution*. 2017;8(1):28-36. doi:
1039 <https://doi.org/10.1111/2041-210X.12628>.
- 1040 81. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. SNP-
1041 sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom*.
1042 2016;2(4):e000056. Epub 2017/03/30. doi: 10.1099/mgen.0.000056. PubMed PMID:
1043 28348851; PMCID: PMC5320690.
- 1044 82. Rapun-Araiz B, Haag AF, De Cesare V, Gil C, Dorado-Morales P, Penades JR,
1045 Lasa I. Systematic Reconstruction of the Complete Two-Component Sensorial Network in
1046 *Staphylococcus aureus*. *mSystems*. 2020;5(4). Epub 2020/08/21. doi:
1047 10.1128/mSystems.00511-20. PubMed PMID: 32817385; PMCID: PMC7438023.
- 1048 83. Ripley WNVBD. *Modern Applied Statistics with S*. Fourth ed: Springer.
- 1049 84. Lenth RV. emmeans: Estimated Marginal Means, aka Least-Squares Means.
1050 <https://cran.r-project.org/web/packages/emmeans/index.html>; Github; 2021.
- 1051 85. Kuhn M. *Building Predictive Models in R Using the caret Package*. 2008.
1052 2008;28(5):26. Epub 2008-09-23. doi: 10.18637/jss.v028.i05.
- 1053 86. Bernardy EE, Petit RA, 3rd, Moller AG, Blumenthal JA, McAdam AJ, Priebe GP,
1054 Chande AT, Rishishwar L, Jordan IK, Read TD, Goldberg JB. Whole-Genome Sequences
1055 of *Staphylococcus aureus* Isolates from Cystic Fibrosis Lung Infections. *Microbiol Resour*
1056 *Announc*. 2019;8(3). Epub 2019/01/29. doi: 10.1128/MRA.01564-18. PubMed PMID:
1057 30687841; PMCID: PMC6346173.
- 1058 87. Andrews S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*.

<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/2010>.

88. Petit RA, 3rd, Read TD. Bactopia: a Flexible Pipeline for Complete Analysis of Bacterial Genomes. *mSystems*. 2020;5(4). Epub 2020/08/06. doi: 10.1128/mSystems.00190-20. PubMed PMID: 32753501; PMCID: PMC7406220.

89. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;19(5):455-77. Epub 2012/04/18. doi: 10.1089/cmb.2012.0021. PubMed PMID: 22506599; PMCID: PMC3342519.

90. Thaipadungpanit J, Amornchai P, Nickerson EK, Wongsuvan G, Wuthiekanun V, Limmathurotsakul D, Peacock SJ. Clinical and molecular epidemiology of *Staphylococcus argenteus* infections in Thailand. *J Clin Microbiol*. 2015;53(3):1005-8. Epub 2015/01/09. doi: 10.1128/JCM.03049-14. PubMed PMID: 25568440; PMCID: PMC4390622.