Distribution Agreement

Signature: DocuSigned by: Vishnu Raghuram 069E960441C0442...

Vishnu Raghuram
Name

5/24/2023 | 10:58 AM EDT
Date

Title	From species-wide to single colony: Multi-scale analysis of the ubiquitous pathogen Staphylococcus aureus
Author	Vishnu Raghuram
Degree	Doctor of Philosophy
Program	Biological and Biomedical Sciences Microbiology and Molecular Genetics

Approved by the Committee

DocuSigned by: Timothy Read 2A4D4A4362BB4E0...	Timothy Read *Advisor*
DocuSigned by: Joanna Goldberg 001C9394541C4B8...	Joanna Goldberg *Advisor*
DocuSigned by: Cassandra Quave 30533E0F32FD4C1...	Cassandra Quave *Committee Member*
DocuSigned by: Marcin Grabowicz 477ADBD547224BC...	Marcin Grabowicz *Committee Member*
DocuSigned by: Samuel Brown 6E606157D62C4F9...	Samuel Brown *Committee Member*
DocuSigned by: Bernardo Mainou 80028705A1DA42B...	Bernardo Mainou *Committee Member*

Accepted by the Laney Graduate School

Kimberly Jacob Arriola, PhD, MPH
Dean, James T. Laney School of Graduate Studies

From species-wide to single colony: Multi-scale analysis of the ubiquitous pathogen *Staphylococcus aureus*

by

Vishnu Raghuram

B. Tech, SRM University, 2016

MS, Georgia Institute of Technology, 2018

Advisors:

Timothy D. Read, PhD

Joanna B. Goldberg, PhD

An abstract of

A dissertation submitted to the Faculty of the James T. Laney School of Graduate Studies of Emory University in partial fulfilment of the requirements for the degree of

Doctor of Philosophy in

Graduate Division of Biological and Biomedical Science

Microbiology and Molecular Genetics

2023

Abstract

From species-wide to single colony: Multi-scale analysis of the ubiquitous pathogen *Staphylococcus aureus*

By Vishnu Raghuram

Whole genome sequencing (WGS) is a powerful tool for both large- and small-scale analysis of any given species. With the increasing accessibility of WGS, bacterial pathogens have been sequenced at an explosive rate over the past decade. This abundance of sequencing data has allowed us to answer questions about pathogens in the context of human infection like never before. In this dissertation, I use collections of sequences from a ubiquitous opportunistic pathogen, *Staphylococcus aureus*, as a model system to answer questions regarding speciation, genome evolution, mutation signatures and human clinical sampling strategies. *S. aureus* is a prominent healthcare-associated pathogen that causes bloodstream, skin, and respiratory infections. *S. aureus* comprises many genomically distinct strains having abundant but selective gene exchange. Therefore, diverse sampling of *S. aureus* is key to understanding the introduction and evolution of new lineages in a given population. Here, I (1) used a dataset of > 80,000 *S. aureus* sequences to outline genomic characteristics that distinguish strains and substrains; (2) developed a software pipeline for rapid mutational analysis of specific loci and identified signatures of convergent evolution on a key *S. aureus* virulence regulator; (3) described optimal clinical sampling strategies for maximising observed genomic diversity; and finally (4) showed how strain specific microdiversity can impact polymicrobial interactions. The overall goal of this work is to describe methods and provide resources to the broader scientific community for analysis of large bacterial sequence datasets as well as sampling strategies for small-scale pathogen evolution studies.

From species-wide to single colony: Multi-scale analysis of the ubiquitous pathogen *Staphylococcus aureus*

By

Vishnu Raghuram

B. Tech, SRM University, 2016

MS, Georgia Institute of Technology, 2018

Advisors:

Timothy D. Read, PhD

Joanna B. Goldberg, PhD

A dissertation submitted to the Faculty of the James T. Laney School of Graduate Studies of Emory University in partial fulfilment of the requirements for the degree of

Doctor of Philosophy in

Graduate Division of Biological and Biomedical Science

Microbiology and Molecular Genetics

2023

Acknowledgements

To my family and friends from back home, thank you for your constant love and support.

To my labmates in the Read and Goldberg lab, thank you for always being there. I could not have done this without any of you. You are the best lab mates I could have ever asked for.

To my MMG friends, I am so glad I got to share this journey with all of you and I could not have imagined a better group.

To my committee, thank you for your advice and direction.

To Brian, thank you for helping me discover my passion for research.

To Tim and Joanna, your approach to science and your mentorship is inspiring. Your endless patience, constant availability and support, and amazing guidance is what got me through this PhD. You made me the scientist I am and for that I will be forever grateful.

Table of Contents

Chapter I – Introduction

Chapter II - A species-scale pangenomic exploration of *Staphylococcus aureus*

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

Chapter IV - Comparison of genomic diversity between single and pooled *Staphylococcus aureus* colonies isolated from human colonisation cultures

Chapter V - *Staphylococcus aureus* and *Pseudomonas aeruginosa* isolates from the same cystic fibrosis respiratory sample coexist in coculture

Chapter VI – Conclusions and future directions

Chapter I – Introduction

Using WGS to study pathogen evolution across scales

This dissertation is about estimating population diversity across different scales using the globally prevalent pathogen *Staphylococcus aureus*. An accurate estimation of population diversity requires close examination of population structure, genome evolution, mutation signatures, and sampling strategies. In this chapter (**Chapter I**), I will introduce concepts around speciation in bacteria, our current knowledge of the population structure of *S. aureus* (Sequence Types and Clonal Complexes), virulence regulation and adaptation, and pangenome ascertainment. These concepts will serve as the foundation for my PhD research.

The true spectrum of diversity and the number of lineages of *S. aureus* is unknown. In the context of pathogens, this is important as genetic diversity can lead to acquisition of new traits. Understanding the population structure and identifying genetic markers that can aid in distinguishing between different lineages is key to understanding the speciation of *S. aureus*. This can be done using complete and diverse sampling of *S. aureus* genomes and this is explored in **Chapter II**. Such large species-wide datasets allow us to examine the evolution and diversity of critical virulence determinants in *S. aureus*. *S. aureus* is a successful global pathogen largely due to its ability to produce a number of virulence factors. It is important to assess whether specific patterns of variation in virulence regulators can be observed across multiple clonal lineages and this is explored in **Chapter III**. The impact of appropriate sampling is not only significant for understanding species-wide macrodiversity, but also for understanding microdiversity in highly homogeneous environments. Clinical and within-host diversity studies repeatedly sample from the same hosts leading to clusters of highly similar isolates. However, the number of isolates required to get an accurate representation of the total diversity within a given sampling space is not clear. Sampling strategies and analysis methods for capturing and measuring microdiversity of *S. aureus* populations from clinical samples are explored in **Chapter IV**. This microdiversity in *S. aureus* can cause variations in polymicrobial interactions during co-infection scenarios. *S. aureus* colonises several sites on the human body where other microbes also reside/infect. One well studied *S. aureus* co-infection is with *Pseudomonas aeruginosa* in wounds and in the cystic fibrosis (CF) lung. Interactions between diverse *S. aureus* and *P. aeruginosa* strains co-isolated from the same CF lung samples are explored in **Chapter V**.

Bacterial species, strains and genomes

While there is no concrete definition of a 'species', biologists are in general agreement that, barring purely taxonomic reasons, a typical species is a distinct cluster in the tree of life, with forces driving cohesion within the cluster while driving separation between other clusters (1,2). These separations are differences in phenotypic and genotypic traits. A prevailing theory for the evolution of a new bacterial species suggests a combination of both slow, gradual accumulation of minor changes (phyletic gradualism) and bursts of relatively large changes between periods of stasis (punctuated equilibrium) (**Fig 1**) (3).

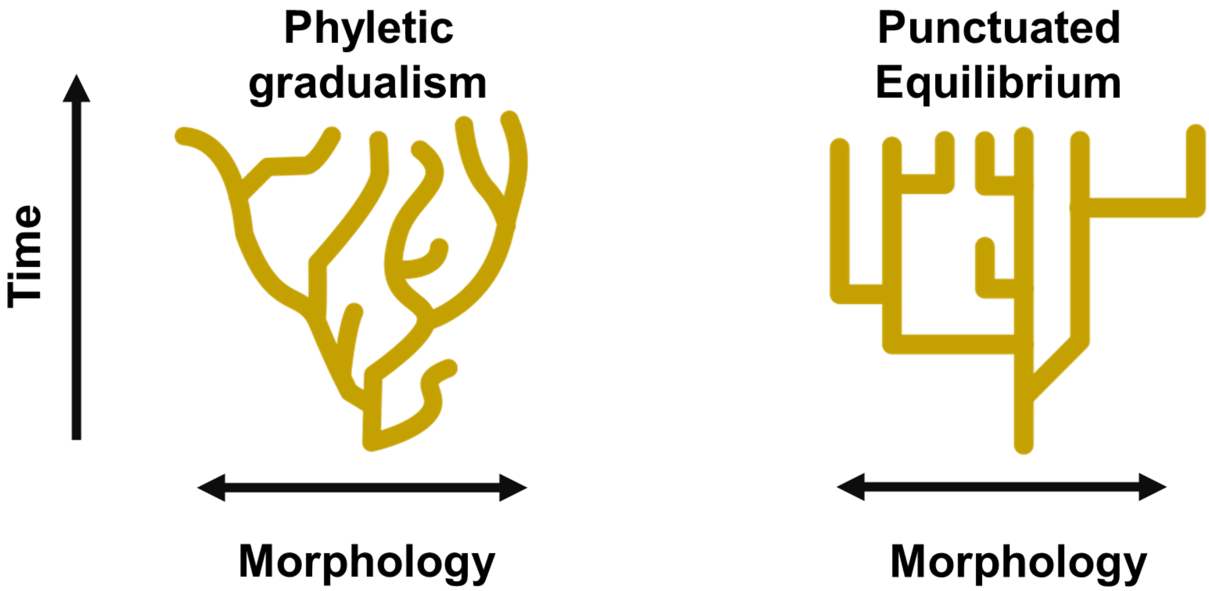

Fig 1: Simplified schematic depicting two proposed mechanisms of speciation. Figures represent change in "morphology" (phenotypic and/or genotypic traits) on the x axis over time in the y axis. Figure adapted from Wikipedia.

Relatively short generation times and large effective population sizes in combination with drift and selective forces allow bacteria to experience accumulation of both adaptive and non-adaptive mutations which may lead to acquisition of novel phenotypic traits (4–6). In addition, due to mechanisms in place for environmental gene uptake, homologous recombination and large-scale horizontal gene transfer (HGT), bacteria also experience rapid genomic restructuring leading to formation of new sub-lineages or even new species (3,7–10). Highly recombinogenic bacterial species diversify in rates that are significantly different from highly clonal species, adding another layer of complexity to the mechanisms that drive speciation (11).

Methods of assigning species/subspecies have largely involved examining phenotypes and morphologies under laboratory conditions till the early-2000s after which PCR based methods gained more popularity (12). However, phenotype-based species identification methods have in the past led to two distinct organisms assigned to the same species (13–15). 16S ribosomal RNA sequencing offered relatively accurate genus-level distinction in its early days but sequencing the full gene, which was not commonplace, was necessary for species or strain level distinction (16). With the advent of large-scale bacterial whole genome sequencing (WGS) since the 2010s, species and strain identification has become a lot more reliable as in most cases each bacterial species is a collection of genomically-distinct organisms (17). Whole genome sequencing of populations directly from environmental or clinical samples (metagenomic sequencing) circumvents the need for laboratory cultivation and culture media biases (18). In-situ metagenomics also allows characterization of several cohabiting natural populations (19,20).

With the increasing ease of access and dropping costs of WGS, the amount of publicly available sequences have more than doubled in the past 5 years (**Fig 2**). These large repositories of sequences allow examinations of population structures on a species-wide and community-wide scale like never before.

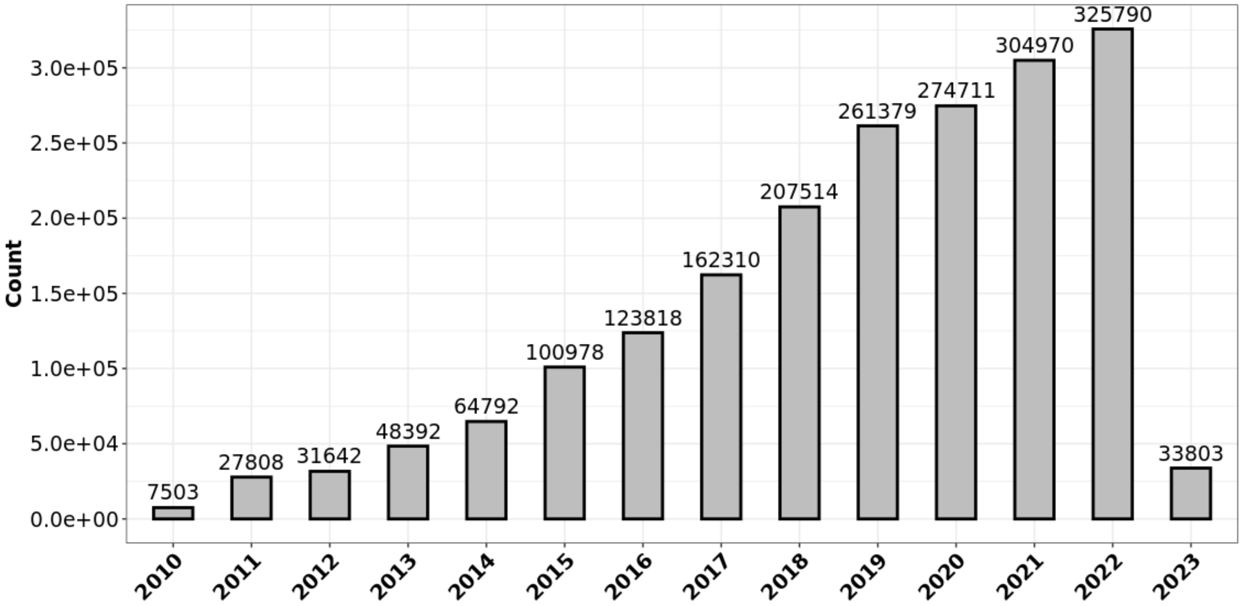

Fig 2: Bar chart showing number of bacterial short-read sequences deposited per year from January 2010 to February 2023 in NCBI.

WGS and metagenomics have augmented experimental evolution and mutation accumulation studies to identify the drivers of genetic change on a short evolutionary time-scale, while also allowing investigation of large scale gene flow events across populations on longer evolutionary time-scales. Studying the evolution of individual organisms in the context of the total population can provide insights into the emergence of potentially new lineages (21–24).

Staphylococcus aureus

General background

S. aureus is a gram-positive opportunistic human pathogen that is prevalent worldwide, causing several diseases such as bacteremia, osteomyelitis, endocarditis, pneumonia, and skin infections (25). *S. aureus* is also one of the most frequent pathogens infecting the respiratory tract of individuals with CF in the US (Cystic fibrosis foundation, 2018). *S. aureus* can contaminate medical equipment such as catheters and surgical devices, which is a major mode of transmission in hospital settings. *S. aureus* strains carrying a mobile genetic element conferring broad spectrum β-lactam resistance, called Methicillin Resistant *Staphylococcus aureus* (MRSA), are particularly problematic. According to the Center for Disease Control (CDC) 2019 antibiotic resistance threat report, MRSA caused 323,700 cases and 10,600 deaths in 2017, consuming $1.7 billion in healthcare costs. In addition, the CDC also estimated 119,247 *S. aureus* bloodstream infections and 19,832 related deaths occurred in the US in 2017 (26). Though *S. aureus* can survive on abiotic surfaces, it is largely associated with human, bovine and other mammalian hosts (27–29). 20 – 30% of humans are colonised by *S. aureus* at any given time (30). Though colonisation is typically asymptomatic, *S. aureus* has the capacity to infect any human tissue. This disease-causing capacity of *S. aureus* is closely tied to its ability to produce a plethora of extracellular toxins such as haemolysins, phenol-soluble modulins, and leukotoxins (31,32). Due to this broad spectrum of virulence factors, some of which being lineage specific, a universal vaccine for *S. aureus* has not been developed despite several efforts (33). The current protocol for treatment of *S. aureus* infections involves drainage of the infection site where possible and/or prolonged antimicrobial treatment (25). Several drug classes are available for the treatment of both MRSA and MSSA (Methicillin Sensitive *S. aureus*), however, antimicrobial resistance in *S. aureus* has been observed for every class of drug thus far and therefore it is essential to discover new drugs, drug targets and also develop alternative treatment strategies.

Sequence Types, Clonal Complexes and genomic characterization

S. aureus has a single ~2.8 Mbp chromosome with an average of 2370 genes and varying plasmid content (34,35). Multi-locus Sequence Typing (MLST) is used to assign "Sequence Types" (STs) to different strains of *S. aureus* based on alleles of seven specific housekeeping genes. Two strains with identical alleles for all seven of these housekeeping genes are considered to be the same ST. STs are then assigned to broader clonal groups or clonal complexes (CCs) if each ST shares five out of the seven alleles with another ST (**Fig 3**) (36). The MLST based typing methods were developed and widely used during the late 90s and early 2000s for phylogenetic grouping of bacterial populations (37,38).

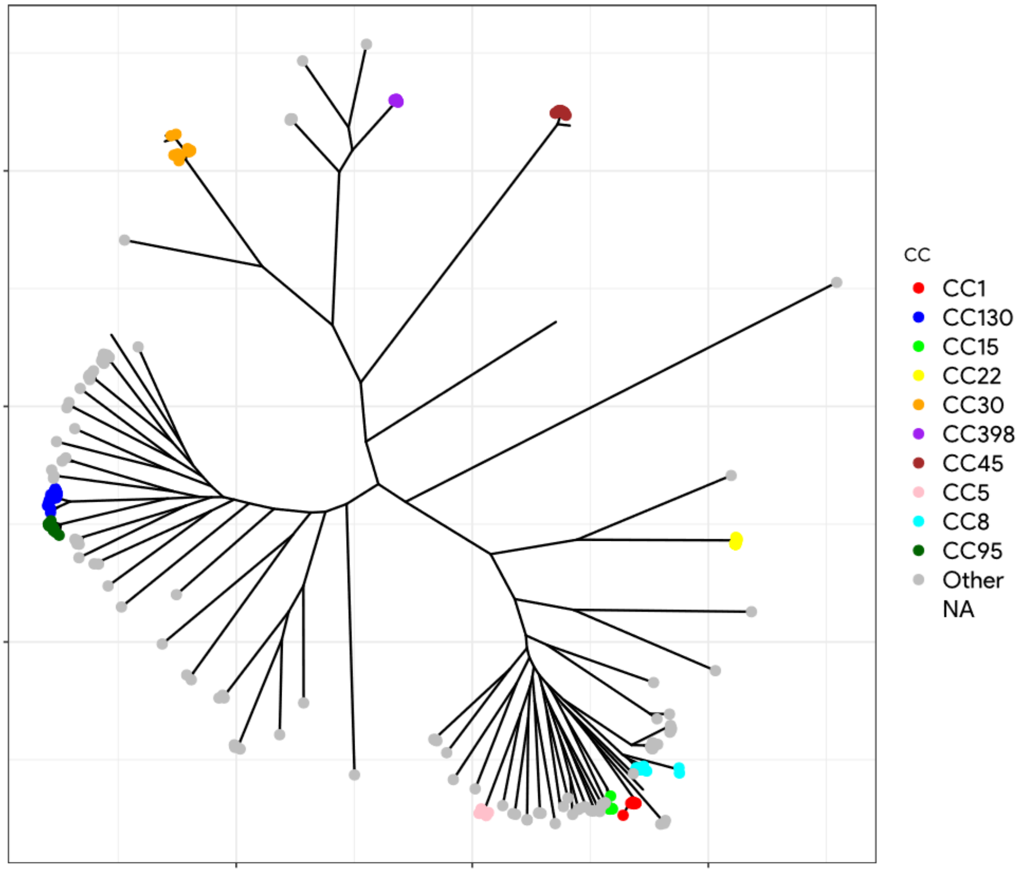

Fig 3: Core genome unrooted maximum likelihood phylogeny of 380 diverse *S. aureus* strains comprising several STs and CCs (25).
Top ten most abundant CCs in NCBI are coloured, remaining CCs are grey. Core genome alignment was built using parsnp v1.5.3 and phylogeny was constructed using IQ-TREE v1.6.12.

Though limited by the technology of the time, phylogenies constructed based on the seven MLST alleles in *S. aureus* are still mostly congruent with population structures shown by whole-genome phylogenies constructed in the 'Next Generation Sequencing' (NGS) era, and therefore, this method of strain and substrain naming is still used today (36,39). Currently, 1258 STs of *S. aureus* have been described (40). There is specific geographic distribution of different STs, with the ST8 MRSA clone – USA300 being the most dominant strain in the US. CC8 and CC5 lineages, two of the most sampled and sequenced lineages of *S. aureus*, can be found worldwide. However, it is important to acknowledge the sampling bias as most *S. aureus* genomes in NCBI are human associated and are from the USA or Europe. Moreover, only 35% of *S. aureus* sequences uploaded to NCBI had any location associated with them as of 2017 (41). Though there are 1258 described STs, with possible more undetermined/novel STs, only ten STs represent over 67% of all sequenced *S. aureus* (40). This could be due to multiple non-mutually exclusive possibilities – 1) There is a severe oversampling of specific STs and the current data do not capture the true prevalence of *S. aureus* clones, 2) The current most prevalent STs have a significant selective advantage over all other STs, 3) the evolutionary forces driving *S. aureus* favour formation of rare STs, 4) the current method of using MLST to classify lineages has limited resolution.

S. aureus and polymicrobial interactions

As *S. aureus* can colonise several body sites such as the skin, lungs, gastrointestinal tract, and nasopharynx, there are numerous opportunities for polymicrobial interactions at those sites. *S. aureus* co-infections have been documented for multiple microbial species including but not limited to *Enterococcus*, *Haemophilus*, *Streptococcus* and of course other *Staphylococci* (42–44). These interactions may be either cooperative or competitive, and in some cases, lead to worse patient outcomes (45,46). Some of the most well studied *S. aureus* co-infections are with *Pseudomonas aeruginosa* in the CF lung. There is some evidence that suggests CF afflicted individuals having both *S. aureus* and *P. aeruginosa* may experience more severe disease compared to mono-infected individuals (47–49). It has been well documented that *P. aeruginosa* outcompetes and kills *S. aureus* in laboratory settings, however, this killing is less pronounced in chronic infection models (50–53). Though *P. aeruginosa* produces several factors that can kill *S. aureus* such as LasA, pyocyanin, quinolines and phage inducing metabolites (54,55), mechanisms promoting increased *S. aureus* survival in co-infection environments have been proposed. Mucoid conversion by *P. aeruginosa* and small colony variant formation, acetoin production and pH alteration by *S. aureus* can promote *S. aureus* survival (53,56). However, the factors driving coexistence between these two species are poorly understood. Studies examining interactions between *S. aureus* and *P. aeruginosa* isolated from the same CF sample can provide more insight into drivers of coexistence.

Quorum sensing and regulation of virulence

S. aureus recognizes environmental cues and adapts to the colonised microenvironment accordingly to elicit diverse types of infections. This adaptation includes regulating immune evasion, cell-cell attachment, biofilm formation and toxin secretion, allowing *S. aureus* to switch between acute and chronic infection lifestyles (57,58). Therefore, regulation of virulence in response to the environment is a critical component of the *S. aureus* life-cycle. The key switch linking environmental sensing and virulence in *S. aureus* is the *agr* quorum sensing system.

Quorum sensing (QS) is a well-documented phenomenon in many bacterial species where there is a cell density dependent global transcriptional change in the bacterial population (59,60). The QS signalling molecule, typically a small molecule for gram-negatives and a small peptide for gram-positives, is secreted and sensed by individual bacterial cells leading to further induction of its own release (59,61). Therefore, this molecule is called an 'autoinducer'. The autoinducer generally varies according to the exact species of bacteria that produces it (62). This is especially true for gram-positive bacteria, where individual species can be subdivided based on the exact amino acid sequence of the small peptide that is produced (63–66). When the concentration of the autoinducer exceeds a certain threshold in the environment, the population reaches a "quorum", and this results in a collective behavioural change. QS controls behaviours such as bioluminescence, biofilm formation, virulence factor secretion and motility (59).

The *agr* QS system in *S. aureus* comprises four genes – *agrBDCA* and a small RNA – RNAIII. AgrD encodes a precursor protein that is processed by the membrane-bound peptidase AgrB, converting AgrD into the autoinducer peptide (AIP) which serves as the QS signal. The secreted AIP is then recognized by a two-component regulatory system (TCS) – AgrC, a histidine kinase, and AgrA, a response regulator. As mentioned earlier, some gram-positives can produce a diverse range of AIPs even within the same species. The *S. aureus agrD* is polymorphic, and can code for one of four possible AIPs. These four possible small peptides, known as *agr* specificity groups, each have a distinct amino acid sequence. As *agrB* is responsible for AIP secretion and *agrC* is an AIP sensor, both genes also have polymorphic sites that are specific to the corresponding AgrD. Due to the nature of these specific sites, there is limited cross-reactivity between the AIP of one specificity group and the *agr* system of a different specificity group. A given *S. aureus* strain can belong to only one of the four specificity groups, and *S. aureus* can be genotyped based on their *agr* groups – *agr* group 1/2/3/4 (67–72) .

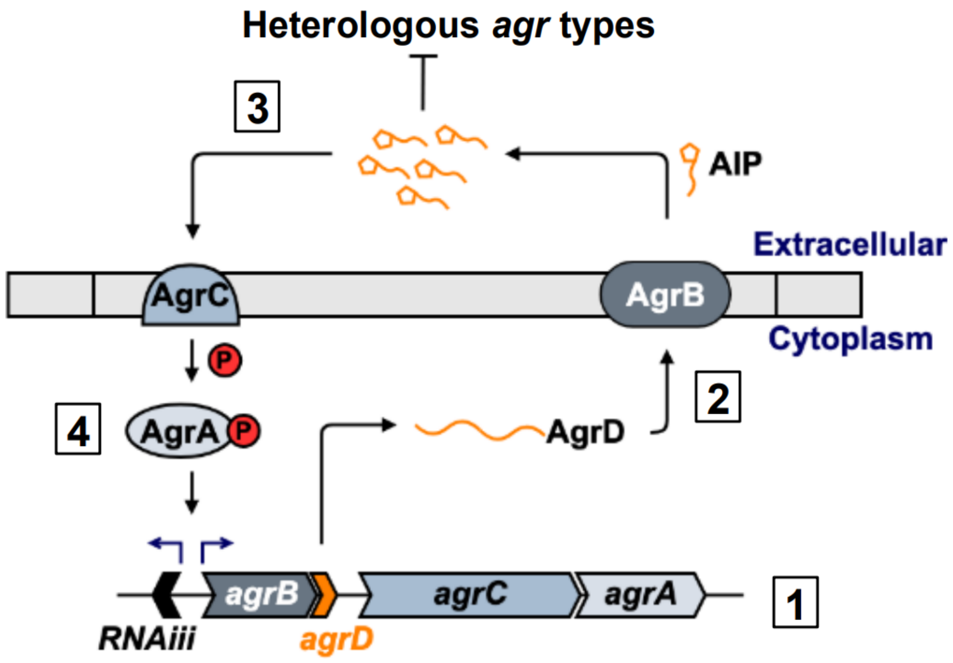

Fig 4: The *agr* (accessory gene regulator) operon is a global transcriptional regulator in *S. aureus*.
1: A schematic depiction of the agr operon showing two divergent promoters (P2 and P3) driving *agrBDCA* and small RNA RNAIII. **2**: AgrD is a precursor protein that is processed by AgrB into an AIP and subsequently released. **3**: The AIP is the QS signalling molecule that is recognized by a sensor kinase AgrC, which in turn phosphorylates AgrA – the response regulator. **4**: AgrA activates the two divergent promoters, continuing the QS cycle. AgrA and RNAIII also regulate other virulence genes elsewhere in the genome (not shown here). Figure adapted from Raghuram *et al*, 2022 (73).

In typical TCS fashion, AgrC phosphorylates AgrA, leading to AgrA transcriptionally activating two divergent promoters in the *agr* operon, P2 and P3. P2 drives the expression of *agrBDCA* while P3 drives RNAIII and a haemolysin called delta-toxin, which is encoded within RNAIII (67,69). RNAIII is the primary effector of the *agr* system, controlling virulence genes such as delta-toxin (Hld), α-haemolysin (Hla), as well as other transcriptional regulators such as MgrA, which promotes dispersal and transmission by downregulating attachment and biofilm associated proteins (74,75). AgrA also upregulates an important family of *S. aureus* pore forming toxins, phenol soluble modulins (PSMs) (76,77). PSMα is a potent cytolytic toxin that promotes neutrophil lysis and *S. aureus* survival during phagocytosis (78). PSMα expression is also closely correlated with the expression of Hla (79). Hld can cause mast-cell degranulation and induce a strong pro-inflammatory response (80). In addition to pore-forming toxins, *S. aureus* also secretes various factors that interfere with neutrophil homing, complement activation, fibrinogen formation and host antibody responses (32). Therefore, the virulence factors controlled by *agr* are important for adaptation and survival in different environments.

Though *agr* mediated virulence promotes disease progression, paradoxically many clinical isolates have defective *agr* genes. Typically, attenuated expression of *agr* mediated virulence factors would lead to reduced disease severity and decreased host cell damage (81,82). However, there may be short term evolutionary benefits to having impaired *agr* activity. *agr*$^-$ strains may have traded their ability to produce energetically expensive virulence factors in favour of a non-toxic but less immune-exposed niche (83,84). This niche adaptation may contribute to worse clinical outcomes in chronic infections such as bacteremia and osteomyelitis (85,86). Defective *agr* function is also associated with decreased susceptibility to vancomycin (87). Environmental factors also play a key role in influencing the *agr* response. The emergence of *agr*$^-$ mutants is enriched during aerobic conditions compared to hypoxic conditions (88). Global metabolic regulators *codY* and *purR* have also been shown to regulate *agr* controlled haemolysins (89,90). Overall, these data suggest that the benefits of *agr* mediated virulence are situational, and that the recurring accounts of *agr* mutants suggests there may be selection for attenuated *agr* activity in certain contexts. *S. aureus* may be actively adapting to different infection niches by modulating *agr* activity. Results from many independent studies suggest that *agr*$^-$ strains have sacrificed long-term viability through successful between-host transmissions for increased adaptation to specific environmental niches within the host. However, this has yet to be consistently shown across multiple clonal lineages. It is important to evaluate whether genome-wide patterns of variation in *agr* can be observed because it is crucial for understanding the mechanisms driving virulence regulation.

In addition to *agr* mutations, another more prominent layer of variability are the four *agr* specificity groups mentioned before. While *S. aureus* is not phylogenetically structured based on the *agr* groups, there is strong evidence to suggest that the four *agr* groups are closely tied to clonal lineages (91,92). Each CC of *S. aureus* exclusively contains a single *agr* group and the only observed exception to this rule is CC45 (92,93). This however is regardless of phylogenetic grouping of *S. aureus*, meaning, phylogenetically related CCs can have different *agr* groups and phylogenetically distant CCs can have the same *agr* group. In other words, each *agr* group does not form its own monophyletic clade. This suggests that *agr* divergence is a foundational event that preceded formation of the CCs as we see them today.

Clonal lineages and population structure

CCs define groups of related STs and are an easier way to assess population structure, with each CC forming a distinct monophyletic group with long branches. The evolutionary history of *S. aureus* suggests that individual CCs were first formed due to large recombination events followed by clonal expansion within CCs, with biological barriers such as competence systems and restriction modification systems selectively preventing between-CC recombination (39,94–96). This has resulted in genomically distinct CCs with gaps in similarity between CCs: *i.e* strains within the same CCs are very closely related but strains between CCs have 1000s of SNPs.

However, these barriers are not absolute. Though *S. aureus* is not a highly recombinogenic species, multiple studies have shown that mobile genetic elements (MGEs) are hotspots for recombination (97,98). MGEs such as phages, plasmids and pathogenicity islands are typically involved in transfer of antimicrobial resistance genes and virulence factors (99). For example, the SCCMec element (Staphylococcus Chromosomal Cassette *mec*) that contains the *mecA* gene, which is responsible for

Methicillin resistance in MRSA strains, is exchanged between different *Staphylococci* (100–103). Non-MGE genes proximal to MGEs can also be exchanged (97).

The genomically distinct nature of *S. aureus* CCs are also evidenced by niche-specific adaptations across different CCs. Correlations between specific CCs of and their host tropism have been reported. CC5, CC8, CC22, CC30 and CC45 are typically human-associated and can spread through hospital and community settings. On the other hand, lineages including CC97, CC121, CC133 and CC151 are commonly associated with livestock (96,104,105). Extensive host-switching has also been documented in the evolutionary history of *S. aureus* with several zoonosis events, with CC97 and CC398 being associated with bovine-to-human transmission (105,106). This host tropism shown by lineages is associated with host-specific adaptations in certain virulence factors/MGEs. For example, Leukotoxin *lukM-lukF-PV* genes are prevalent in bovine isolates and are associated with severe bovine mastitis (107,108) Antimicrobial resistance genes outside of beta-lactam resistance are more commonly associated with human-adapted lineages and are less prevalent in bovine-adapted isolates (109). Identifying CC specific genes, especially AMR genes and virulence factors, are critical for predicting and subverting outbreaks. Multiple reviews have outlined the different host-specific genetic correlates in *S. aureus* across a diverse array of host species and most of these genetic elements are in MGEs (pathogenicity islands or prophages). (104,110).

There is also some evidence to suggest that specific *agr* groups may be associated with disease outcomes. *agr* group 4 isolates are associated with exfoliatin production, *agr* group 3 isolates are associated with menstrual toxic shock syndrome, and *agr* group 2 isolates are associated with increased vancomycin tolerance (63,111–113).

As stated earlier, CCs are also exclusive to specific *agr* groups. Strains within a given CC comprise only one *agr* group with very limited exceptions (92,93,114). The *agr* group separation was an important event in the evolutionary history of *S. aureus* that predated but is still linked to the CC-based subspeciation (91). As *agr* activation causes global transcriptional changes, and gene content varies between CCs, *agr* may have distinct regulons across different CCs. This link between *agr* and the CC specific gene content is yet to be explored.

The Pangenome and species-wide diversity

Gene gain, loss and allelic variation are common phenomena in bacteria and are a major contributor of intra-species variation. These genome evolution events can be shaped by both adaptive and neutral forces depending on the biological context (115–117). This would lead to members of the same species having similar but not identical genomes across different environments. Therefore, observing genomes from one or few members can vastly underestimate the capacity of a species. A "Pangenome" – a collection of all genes/alleles that are found within a group of organisms, encompassing diverse members is required to fully understand the capacity of a species. Genes in the pangenome can be broadly classified into the "core" – genes present in all/most members (>95%), the "accessory" – genes only present in some members (> 95%). The "accessory" genome can be further divided into "intermediate" - genes present in many (> 10% but < 95%) and "unique/rare" – genes that are almost never present (<10%).

While *S. aureus* is considered more clonal than highly promiscuous species such as *Streptococcus pneumoniae* (36), the above outlined literature suggests the presence of large-scale recombination events in the past, on-going recombination events in MGEs, lineage specific genetic markers, and lineage specific restriction barriers preventing indiscriminate gene exchange. These selective gene transfer events can add to phylogenetic noise and uncertainty, making inferences regarding CC divergence challenging. However, with diverse sampling and identifying genomic regions that are common to almost every strain in the species – the "core", it is possible to identify the lineage specific genes that keep the CCs separate. The "accessory" genes - genes that are present in some members of the populations but never all, can also potentially predict the emergence of future lineages. However, genes designated "core" and "accessory" are context dependent and are highly influenced by the sampled population (118,119). As discussed in the previous section, host/environmental factors can affect the total gene content thereby leading to some genes designated "core" in specific environments but "accessory" in others (118,119). This core and accessory designations can also vary when considering within-CC vs between CC comparisons (120). Meaning, the more diverse genomes that are considered, the smaller the core genome and the larger the accessory genome.

As of February 2023, there are over 93,000 whole genome sequences (WGS) of *S. aureus* publicly available [**Fig 2**]. This offers the opportunity to examine *S. aureus* on a species-wide scale, estimating the total genetic makeup of the species while also serving as a natural laboratory for observing the emergence of new CCs.

The goals of this dissertation

Macrodiversity

This dissertation is primarily about estimating population diversity across different scales. In the macro-scale, the species wide diversity of *S. aureus* and the number of clonal lineages is unknown. Due to the potentially varying gene content across lineages, identifying differences in pangenome content within and between lineages can serve as markers for *S. aureus* subspeciation. To do this, a large-scale pangenomic study integrating all publicly available *S. aureus* sequences with uniform processing is required. In **Chapter II**, I used 83,000 publicly available genome sequences of *S. aureus* to build a species-wide pangenome, the largest *S. aureus* pangenome to date. I outline workflows for efficient processing of large datasets and explore the genetic makeup of diverse *S. aureus* lineages to understand how different combinations of genes can form different lineages. Such large diverse datasets can be used for examining evolution of specific, clinically relevant genes on a species wide scale.

For example, the *agr* operon, a key virulence determinant in *S. aureus*, is frequently lost in chronic infection scenarios. It is hypothesised that this loss of *agr* may provide increased fitness within a specific niche but in turn leads to lack of long-term transmission. However, the true prevalence of these *agr* mutations was unknown. In **Chapter III**, I analysed the allelic diversity of the *S. aureus agr* Quorum Sensing operon and found that more than 5% of strains in the public database had nonfunctional *agr* systems. I also provided new insights into the evolution of these genetic mutations in the *agr* system. In the process, I developed computational tools to aid in rapid analysis of allelic diversity of specific genes across large datasets.

Microdiversity

In contrast to macro-scale diversity on a species-wide scale, microdiversity within small, homogeneous populations pose a different challenge. The most common approach to sampling microbial populations within an infected or colonised host is to sequence genomes from a single colony. Here, clinically relevant genes or mutations can go undetected if they are harboured by minority subpopulations not represented in one colony. Alternative strategies are to sequence multiple single colonies, or perform metagenomic sequencing - pooling and sequencing the total population. A direct comparison between single isolate and pooled population sequences can help devise optimal sampling strategies for clinical and within-host diversity studies. In **Chapter IV**, I attempt to answer the question of how many colonies obtained from a single patient is enough to obtain an accurate picture of the total population diversity within the patient while keeping in mind time and labour costs. I compared genomic diversity between pure colonies and pooled populations of *S. aureus* obtained from skin swabs and to evaluate sampling strategies for clinical and within-host studies.

Capturing total population diversity is also important during polymicrobial infection scenarios. Different strains of *S. aureus* and their interactions with different strains of a co-infecting species can lead to different outcomes. One of the most frequent co-infections involve *S. aureus* and *P. aeruginosa* in the Cystic Fibrosis lung. In **Chapter V**, I used co-culture data from diverse *S. aureus* and *P. aeruginosa* strains sampled from CF patients to assess outcomes of competition and coexistence.

Finally, in **Chapter VI**, I summarise my contributions to the field while also discussing potential future work that can further our understanding of *S. aureus* evolution and virulence regulation.

My main goal for this dissertation was to answer four main questions – 1) How do we capture the complete species-wide diversity of an organism? 2) How do we capture complete allelic diversity of specific genes? 3) How do we capture maximum diversity in highly homogeneous environments? and 4) How does this microdiversity change properties of the species in clinical environments?

S. aureus is both a model system to answer these questions and an important pathogen where we can potentially translate what is learned into better treatments. The overall goal of this work is to provide tools and resources to the broader microbial genomics community that can aid in analysis of bacterial sequence datasets from various sources.

References

1. Cohan FM. Bacterial Speciation: Genetic Sweeps in Bacterial Species. Curr Biol. 2016 Feb 8;26(3):R112–5.
2. Waller DA. Speciation and Its Consequences. Ann Entomol Soc Am. 1991 Sep 1;84(5):567.
3. Gao Y, Wu M. Microbial genomic trait evolution is dominated by frequent and rare pulsed evolution. Sci Adv. 2022 Jul 15;8(28):eabn1916.
4. Bobay LM, Ochman H. Factors driving effective population size and pan-genome evolution in bacteria. BMC Evol Biol. 2018 Oct 12;18(1):153.
5. Kuo CH, Moran NA, Ochman H. The consequences of genetic drift for bacterial genome complexity. Genome Res. 2009 Aug 1;19(8):1450–4.
6. Wein T, Dagan T. The Effect of Population Bottleneck Size and Selective Regime on Genetic Diversity and Evolvability in Bacteria. Genome Biol Evol. 2019 Nov 1;11(11):3283–90.
7. Brochet M, Rusniok C, Couvé E, Dramsi S, Poyart C, Trieu-Cuot P, et al. Shaping a bacterial genome by large chromosomal replacements, the evolutionary history of Streptococcus agalactiae. Proc Natl Acad Sci. 2008 Oct 14;105(41):15961–6.
8. Holden MTG, Lindsay JA, Corton C, Quail MA, Cockfield JD, Pathak S, et al. Genome Sequence of a Recently Emerged, Highly Transmissible, Multi-Antibiotic- and Antiseptic-Resistant Variant of Methicillin-Resistant Staphylococcus aureus, Sequence Type 239 (TW). J Bacteriol. 2010 Feb;192(3):888–92.
9. Chen L, Mathema B, Pitout JDD, DeLeo FR, Kreiswirth BN. Epidemic Klebsiella pneumoniae ST258 is a hybrid strain. mBio. 2014 Jun 24;5(3):e01355-01314.
10. Coyle NM, Bartie KL, Bayliss SC, Bekaert M, Adams A, McMillan S, et al. A Hopeful Sea-Monster: A Very Large Homologous Recombination Event Impacting the Core Genome of the Marine Pathogen Vibrio anguillarum. Front Microbiol. 2020;11:1430.
11. Shapiro BJ. How clonal are bacteria over time? Curr Opin Microbiol. 2016 Jun 1;31:116–23.
12. Emerson D, Agulto L, Liu H, Liu L. Identifying and Characterizing Bacteria in an Era of Genomics and Proteomics. BioScience. 2008 Nov 1;58(10):925–36.
13. Yabuuchi E, Kosako Y, Oyaizu H, Yano I, Hotta H, Hashimoto Y, et al. Proposal of Burkholderia gen. nov. and Transfer of Seven Species of the Genus Pseudomonas Homology Group II to the New Genus, with the Type Species Burkholderia cepacia (Palleroni and Holmes 1981) comb. nov. Microbiol Immunol. 1992;36(12):1251–75.
14. Palleroni NJ, Bradbury JF. Stenotrophomonas, a New Bacterial Genus for Xanthomonas maltophilia (Hugh 1980) Swings et al. 1983. Int J Syst Evol Microbiol. 1993;43(3):606–9.
15. Townsend SM, Hurrell E, Caubilla-Barron J, Loc-Carrillo C, Forsythe SJ. Characterization of an extended-spectrum beta-lactamase Enterobacter hormaechei nosocomial outbreak, and other Enterobacter hormaechei misidentified as Cronobacter (Enterobacter) sakazakii. Microbiology. 2008;154(12):3659–67.
16. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome

analysis. Nat Commun. 2019 Nov 6;10(1):5029.

17. Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, et al. Microbial species delineation using whole genome sequences. Nucleic Acids Res. 2015 Aug 18;43(14):6761–71.
18. Konstantinidis K. Bypassing Cultivation To Identify Bacterial Species. In 2014 [cited 2023 Apr 20]. Available from: https://www.semanticscholar.org/paper/Bypassing-Cultivation-To-Identify-Bacterial-Species-Konstantinidis/6e69f25ed9daac65ed59836d34ccb18e6f8df141
19. Caro-Quintero A, Konstantinidis KT. Bacterial species may exist, metagenomics reveal. Environ Microbiol. 2012;14(2):347–55.
20. National Research Council (US) Committee on Metagenomics: Challenges and Functional Applications. The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet [Internet]. Washington (DC): National Academies Press (US); 2007 [cited 2023 Apr 20]. (The National Academies Collection: Reports funded by National Institutes of Health). Available from: http://www.ncbi.nlm.nih.gov/books/NBK54006/
21. Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM. The dynamics of molecular evolution over 60,000 generations. Nature. 2017 Nov;551(7678):45–50.
22. Couce A, Magnan M, Lenski RE, Tenaillon O. Predictability shifts from local to global rules during bacterial adaptation [Internet]. bioRxiv; 2023 [cited 2023 May 2]. p. 2022.05.17.492360. Available from: https://www.biorxiv.org/content/10.1101/2022.05.17.492360v2
23. Kurokawa M, Nishimura I, Ying BW. Experimental Evolution Expands the Breadth of Adaptation to an Environmental Gradient Correlated With Genome Reduction. Front Microbiol [Internet]. 2022 [cited 2023 May 2];13. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2022.826894
24. Martinez-Gutierrez CA, Aylward FO. Genome size distributions in bacteria and archaea are strongly linked to evolutionary history at broad phylogenetic scales. PLOS Genet. 2022 May 23;18(5):e1010220.
25. David MZ, Daum RS. Treatment of Staphylococcus aureus Infections. Curr Top Microbiol Immunol. 2017;409:325–83.
26. Kourtis AP, Hatfield K, Baggs J, Mu Y, See I, Epson E, et al. 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 Mar 8;68(9):214–9.
27. Heaton CJ, Gerbig GR, Sensius LD, Patel V, Smith TC. Staphylococcus aureus Epidemiology in Wildlife: A Systematic Review. Antibiot Basel Switz. 2020 Feb 18;9(2):89.
28. Desai R, Pannaraj PS, Agopian J, Sugar CA, Liu GY, Miller LG. Survival and transmission of community-associated methicillin-resistant Staphylococcus aureus from fomites. Am J Infect Control. 2011 Apr 1;39(3):219–25.
29. Howden BP, Giulieri SG, Wong Fok Lung T, Baines SL, Sharkey LK, Lee JYH, et al. Staphylococcus aureus host interactions and adaptation. Nat Rev Microbiol. 2023 Jan 27;1–16.
30. Sakr A, Brégeon F, Mège JL, Rolain JM, Blin O. Staphylococcus aureus Nasal Colonization: An Update on Mechanisms, Epidemiology, Risk Factors, and

Subsequent Infections. Front Microbiol [Internet]. 2018 [cited 2023 May 2];9. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2018.02419

31. Cheung GYC, Bae JS, Otto M. Pathogenicity and virulence of Staphylococcus aureus. Virulence. 2021 Dec 31;12(1):547–69.
32. Thammavongsa V, Kim HK, Missiakas D, Schneewind O. Staphylococcal manipulation of host immune responses. Nat Rev Microbiol. 2015 Sep;13(9):529–43.
33. Jansen KU, Girgenti DQ, Scully IL, Anderson AS. Vaccine review: "Staphyloccocus aureus vaccines: Problems and prospects." Vaccine. 2013 Jun 7;31(25):2723–30.
34. Novick RP. Staphylococcal Plasmids and Their Replication. Annu Rev Microbiol. 1989;43(1):537–63.
35. Shearer JES, Wireman J, Hostetler J, Forberger H, Borman J, Gill J, et al. Major Families of Multiresistant Plasmids from Geographically and Epidemiologically Diverse Staphylococci. G3 GenesGenomesGenetics. 2011 Dec 1;1(7):581–91.
36. Feil EJ, Cooper JE, Grundmann H, Robinson DA, Enright MC, Berendt T, et al. How clonal is Staphylococcus aureus? J Bacteriol. 2003 Jun;185(11):3307–16.
37. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A. 1998 Mar 17;95(6):3140–5.
38. Enright MC, Day NP, Davies CE, Peacock SJ, Spratt BG. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of Staphylococcus aureus. J Clin Microbiol. 2000 Mar;38(3):1008–15.
39. Planet PJ, Narechania A, Chen L, Mathema B, Boundy S, Archer G, et al. Architecture of a Species: Phylogenomics of Staphylococcus aureus. Trends Microbiol. 2017 Feb 1;25(2):153–66.
40. Staphylococcus aureus [Internet]. PubMLST. [cited 2023 Apr 18]. Available from: https://pubmlst.org/organisms/staphylococcus-aureus
41. Iii RAP, Read TD. Staphylococcus aureus viewed from the perspective of 40,000+ genomes. PeerJ. 2018 Jul 12;6:e5261.
42. Ray AJ, Pultz NJ, Bhalla A, Aron DC, Donskey CJ. Coexistence of vancomycin-resistant enterococci and Staphylococcus aureus in the intestinal tracts of hospitalized patients. Clin Infect Dis Off Publ Infect Dis Soc Am. 2003 Oct 1;37(7):875–81.
43. Margolis E, Yates A, Levin BR. The ecology of nasal colonization of Streptococcus pneumoniae, Haemophilus influenzae and Staphylococcus aureus: the role of competition and interactions with host's immune response. BMC Microbiol. 2010 Feb 23;10:59.
44. Parlet CP, Brown MM, Horswill AR. Commensal Staphylococci Influence Staphylococcus aureus Skin Colonization and Disease. Trends Microbiol. 2019 Jun 1;27(6):497–507.
45. Harriott MM, Noverr MC. Candida albicans and Staphylococcus aureus form polymicrobial biofilms: effects on antimicrobial resistance. Antimicrob Agents Chemother. 2009 Sep;53(9):3914–22.
46. Pettigrew MM, Gent JF, Revai K, Patel JA, Chonmaitree T. Microbial interactions during upper respiratory tract infections. Emerg Infect Dis. 2008

Oct;14(10):1584–91.

47. Hubert D, Réglier-Poupet H, Sermet-Gaudelus I, Ferroni A, Le Bourgeois M, Burgel PR, et al. Association between Staphylococcus aureus alone or combined with Pseudomonas aeruginosa and the clinical condition of patients with cystic fibrosis. J Cyst Fibros Off J Eur Cyst Fibros Soc. 2013 Sep;12(5):497–503.
48. Limoli DH, Yang J, Khansaheb MK, Helfman B, Peng L, Stecenko AA, et al. Staphylococcus aureus and Pseudomonas aeruginosa co-infection is associated with cystic fibrosis-related diabetes and poor clinical outcomes. Eur J Clin Microbiol Infect Dis Off Publ Eur Soc Clin Microbiol. 2016 Jun;35(6):947–53.
49. Maliniak ML, Stecenko AA, McCarty NA. A longitudinal analysis of chronic MRSA and Pseudomonas aeruginosa co-infection in cystic fibrosis: A single-center study. J Cyst Fibros Off J Eur Cyst Fibros Soc. 2016 May;15(3):350–6.
50. Bernardy EE, Petit RA, Raghuram V, Alexander AM, Read TD, Goldberg JB. Genotypic and Phenotypic Diversity of Staphylococcus aureus Isolates from Cystic Fibrosis Patient Lung Infections and Their Interactions with Pseudomonas aeruginosa. mBio. 2020 Jun 23;11(3):e00735-20.
51. Filkins LM, Graber JA, Olson DG, Dolben EL, Lynd LR, Bhuju S, et al. Coculture of Staphylococcus aureus with Pseudomonas aeruginosa Drives S. aureus towards Fermentative Metabolism and Reduced Viability in a Cystic Fibrosis Model. J Bacteriol. 2015 Jul;197(14):2252–64.
52. DeLeon S, Clinton A, Fowler H, Everett J, Horswill AR, Rumbaugh KP. Synergistic Interactions of Pseudomonas aeruginosa and Staphylococcus aureus in an In Vitro Wound Model. Infect Immun. 2014 Nov;82(11):4718–28.
53. Limoli DH, Whitfield GB, Kitao T, Ivey ML, Davis MR, Grahl N, et al. Pseudomonas aeruginosa Alginate Overproduction Promotes Coexistence with Staphylococcus aureus in a Model of Cystic Fibrosis Respiratory Infection. mBio. 2017 Mar 21;8(2):e00186-17.
54. Beaume M, Köhler T, Fontana T, Tognon M, Renzoni A, van Delden C. Metabolic pathways of Pseudomonas aeruginosa involved in competition with respiratory bacterial pathogens. Front Microbiol. 2015;6:321.
55. Jancheva M, Böttcher T. A Metabolite of Pseudomonas Triggers Prophage-Selective Lysogenic to Lytic Conversion in Staphylococcus aureus. J Am Chem Soc. 2021 Jun 9;143(22):8344–51.
56. Kvich L, Crone S, Christensen MH, Lima R, Alhede M, Alhede M, et al. Investigation of the Mechanism and Chemistry Underlying Staphylococcus aureus' Ability to Inhibit Pseudomonas aeruginosa Growth In Vitro. J Bacteriol. 2022 Oct 11;204(11):e00174-22.
57. Tuchscherr L, Löffler B. Staphylococcus aureus dynamically adapts global regulators and virulence factor expression in the course from acute to chronic infection. Curr Genet. 2016 Feb 1;62(1):15–7.
58. García-Betancur JC, Goñi-Moreno A, Horger T, Schott M, Sharan M, Eikmeier J, et al. Cell differentiation defines acute and chronic infection cell types in Staphylococcus aureus. Gilmore MS, editor. eLife. 2017 Sep 12;6:e28023.
59. Papenfort K, Bassler BL. Quorum sensing signal-response systems in Gram-negative bacteria. Nat Rev Microbiol. 2016 Aug 11;14(9):576–88.
60. Fuqua WC, Winans SC, Greenberg EP. Quorum sensing in bacteria: the LuxR-LuxI

family of cell density-responsive transcriptional regulators. J Bacteriol. 1994 Jan;176(2):269–75.

61. Novick RP, Geisinger E. Quorum sensing in staphylococci. Annu Rev Genet. 2008;42:541–64.
62. Hawver LA, Jung SA, Ng WL. Specificity and complexity in bacterial quorum-sensing systems. FEMS Microbiol Rev. 2016 Sep;40(5):738–52.
63. Ji G, Beavis R, Novick RP. Bacterial interference caused by autoinducing peptide variants. Science. 1997 Jun 27;276(5321):2027–30.
64. Whatmore AM, Barcus VA, Dowson CG. Genetic diversity of the streptococcal competence (com) gene locus. J Bacteriol. 1999 May;181(10):3144–54.
65. Tran LS, Nagai T, Itoh Y. Divergent structure of the ComQXPA quorum-sensing components: molecular basis of strain-specific communication mechanism in Bacillus subtilis. Mol Microbiol. 2000 Sep;37(5):1159–71.
66. Tortosa P, Logsdon L, Kraigher B, Itoh Y, Mandic-Mulec I, Dubnau D. Specificity and genetic polymorphism of the Bacillus competence quorum-sensing system. J Bacteriol. 2001 Jan;183(2):451–60.
67. Novick RP. Autoinduction and signal transduction in the regulation of staphylococcal virulence. Mol Microbiol. 2003 Jun;48(6):1429–49.
68. 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 Oct;12(10):3967–75.
69. Novick RP, Projan SJ, Kornblum J, Ross HF, Ji G, Kreiswirth B, et al. The agr P2 operon: an autocatalytic sensory transduction system in Staphylococcus aureus. Mol Gen Genet MGG. 1995 Aug 30;248(4):446–58.
70. 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 Dec 19;92(26):12055–9.
71. Lina G, Jarraud S, Ji G, Greenland T, Pedraza A, Etienne J, et al. Transmembrane topology and histidine protein kinase activity of AgrC, the agr signal receptor in Staphylococcus aureus. Mol Microbiol. 1998 May;28(3):655–62.
72. Dufour P, Jarraud S, Vandenesch F, Greenland T, Novick RP, Bes M, et al. High genetic variability of the agr locus in Staphylococcus species. J Bacteriol. 2002 Feb;184(4):1180–6.
73. Raghuram V, Alexander AM, Loo HQ, Petit RA, Goldberg JB, Read TD. Species-Wide Phylogenomics of the Staphylococcus aureus Agr Operon Revealed Convergent Evolution of Frameshift Mutations. Microbiol Spectr. 2022 Jan 19;10(1):e01334-21.
74. Morfeldt E, Taylor D, von Gabain A, Arvidson S. Activation of alpha-toxin translation in Staphylococcus aureus by the trans-encoded antisense RNA, RNAIII. EMBO J. 1995 Sep;14(18):4569–77.
75. Gupta RKr, Luong TT, Lee CY. RNAIII of the Staphylococcus aureus agr system activates global regulator MgrA by stabilizing mRNA. Proc Natl Acad Sci. 2015 Nov 10;112(45):14036–41.
76. Peschel A, Otto M. Phenol-soluble modulins and staphylococcal infection. Nat Rev Microbiol. 2013 Oct;11(10):667–73.
77. Queck SY, Jameson-Lee M, Villaruz AE, Bach THL, Khan BA, Sturdevant DE, et al. RNAIII-Independent Target Gene Control by the agr Quorum-Sensing System:

Insight into the Evolution of Virulence Regulation in Staphylococcus aureus. Mol Cell. 2008 Oct 10;32(1):150–8.

78. Surewaard BGJ, de Haas CJC, Vervoort F, Rigby KM, DeLeo FR, Otto M, et al. Staphylococcal alpha-phenol soluble modulins contribute to neutrophil lysis after phagocytosis. Cell Microbiol. 2013 Aug;15(8):1427–37.
79. Berube BJ, Sampedro GR, Otto M, Bubeck Wardenburg J. The psmα locus regulates production of Staphylococcus aureus alpha-toxin during infection. Infect Immun. 2014 Aug;82(8):3350–8.
80. Nakamura Y, Oscherwitz J, Cease KB, Chan SM, Muñoz-Planillo R, Hasegawa M, et al. Staphylococcus δ-toxin induces allergic skin disease by activating mast cells. Nature. 2013 Nov 21;503(7476):397–401.
81. Das S, Lindemann C, Young BC, Muller J, Österreich B, Ternette N, et al. Natural mutations in a Staphylococcus aureus virulence regulator attenuate cytotoxicity but permit bacteremia and abscess formation. Proc Natl Acad Sci U S A. 2016 May 31;113(22):E3101-3110.
82. Cheung AL, Eberhardt KJ, Chung E, Yeaman MR, Sullam PM, Ramos M, et al. Diminished virulence of a sar-/agr- mutant of Staphylococcus aureus in the rabbit model of endocarditis. J Clin Invest. 1994 Nov;94(5):1815–22.
83. Laabei M, Uhlemann AC, Lowy FD, Austin ED, Yokoyama M, Ouadi K, et al. Evolutionary Trade-Offs Underlie the Multi-faceted Virulence of Staphylococcus aureus. PLoS Biol. 2015;13(9):e1002229.
84. 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 Oct 24;8(5):e01476-17.
85. Schweizer ML, Furuno JP, Sakoulas G, Johnson JK, Harris AD, Shardell MD, et al. Increased mortality with accessory gene regulator (agr) dysfunction in Staphylococcus aureus among bacteremic patients. Antimicrob Agents Chemother. 2011 Mar;55(3):1082–7.
86. Suligoy CM, Lattar SM, Noto Llana M, González CD, Alvarez LP, Robinson DA, et al. Mutation of Agr Is Associated with the Adaptation of Staphylococcus aureus to the Host during Chronic Osteomyelitis. Front Cell Infect Microbiol. 2018;8:18.
87. Sakoulas G, Eliopoulos GM, Fowler VG, Moellering RC, Novick RP, Lucindo N, et al. 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 Jul;49(7):2687–92.
88. George SE, Hrubesch J, Breuing I, Vetter N, Korn N, Hennemann K, et al. Oxidative stress drives the selection of quorum sensing mutants in the Staphylococcus aureus population. Proc Natl Acad Sci U S A. 2019 Sep 17;116(38):19145–54.
89. Roux A, Todd DA, Velázquez JV, Cech NB, Sonenshein AL. CodY-mediated regulation of the Staphylococcus aureus Agr system integrates nutritional and population density signals. J Bacteriol. 2014 Mar;196(6):1184–96.
90. Sause WE, Balasubramanian D, Irnov I, Copin R, Sullivan MJ, Sommerfield A, et al. The purine biosynthesis regulator PurR moonlights as a virulence regulator in Staphylococcus aureus. Proc Natl Acad Sci. 2019 Jul 2;116(27):13563–72.
91. 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 Dec;187(24):8312–21.

92. Wright JS, Traber KE, Corrigan R, Benson SA, Musser JM, Novick RP. The agr radiation: an early event in the evolution of staphylococci. J Bacteriol. 2005 Aug;187(16):5585–94.

93. Peacock SJ, Moore CE, Justice A, Kantzanou M, Story L, Mackie K, et al. Virulent combinations of adhesin and toxin genes in natural populations of Staphylococcus aureus. Infect Immun. 2002 Sep;70(9):4987–96.

94. Feil EJ. Small change: keeping pace with microevolution. Nat Rev Microbiol. 2004 Jun;2(6):483–95.

95. Monk IR, Shah IM, Xu M, Tan MW, Foster TJ. Transforming the untransformable: application of direct transformation to manipulate genetically Staphylococcus aureus and Staphylococcus epidermidis. mBio. 2012;3(2):e00277-11.

96. Park S, Jung D, O'Brien B, Ruffini J, Dussault F, Dube-Duquette A, et al. Comparative genomic analysis of Staphylococcus aureus isolates associated with either bovine intramammary infections or human infections demonstrates the importance of restriction-modification systems in host adaptation. Microb Genomics. 2022 Feb;8(2):000779.

97. Everitt RG, Didelot X, Batty EM, Miller RR, Knox K, Young BC, et al. Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus. Nat Commun. 2014 May 23;5:3956.

98. Méric G, Miragaia M, de Been M, Yahara K, Pascoe B, Mageiros L, et al. Ecological Overlap and Horizontal Gene Transfer in Staphylococcus aureus and Staphylococcus epidermidis. Genome Biol Evol. 2015 May 1;7(5):1313–28.

99. Firth N, Jensen SO, Kwong SM, Skurray RA, Ramsay JP. Staphylococcal Plasmids, Transposable and Integrative Elements. Microbiol Spectr. 2018 Dec 13;6(6):6.6.06.

100. Hanssen AM, Kjeldsen G, Sollid JUE. 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 Jan;48(1):285–96.

101. Wielders CL, Vriens MR, Brisse S, de Graaf-Miltenburg LA, Troelstra A, Fleer A, et al. In-vivo transfer of mecA DNA to Staphylococcus aureus [corrected]. Lancet Lond Engl. 2001 May 26;357(9269):1674–5.

102. Hanssen AM, Ericson Sollid JU. SCCmec in staphylococci: genes on the move. FEMS Immunol Med Microbiol. 2006;46(1):8–20.

103. Bloemendaal ALA, Brouwer EC, Fluit AC. Methicillin resistance transfer from Staphylocccus epidermidis to methicillin-susceptible Staphylococcus aureus in a patient during antibiotic therapy. PloS One. 2010 Jul 29;5(7):e11841.

104. Park S, Ronholm J. Staphylococcus aureus in Agriculture: Lessons in Evolution from a Multispecies Pathogen. Clin Microbiol Rev. 2021 Mar 17;34(2):e00182-20.

105. Yebra G, Harling-Lee JD, Lycett S, Aarestrup FM, Larsen G, Cavaco LM, et al. Multiclonal human origin and global expansion of an endemic bacterial pathogen of livestock. Proc Natl Acad Sci. 2022 Dec 13;119(50):e2211217119.

106. Richardson EJ, Bacigalupe R, Harrison EM, Weinert LA, Lycett S, Vrieling M, et al. Gene exchange drives the ecological success of a multi-host bacterial pathogen. Nat Ecol Evol. 2018 Sep;2(9):1468–78.

107. Yamada T, Tochimaru N, Nakasuji S, Hata E, Kobayashi H, Eguchi M, et al.

Leukotoxin family genes in Staphylococcus aureus isolated from domestic animals and prevalence of lukM-lukF-PV genes by bacteriophages in bovine isolates. Vet Microbiol. 2005 Sep 30;110(1–2):97–103.

108. Younis A, Krifucks O, Fleminger G, Heller ED, Gollop N, Saran A, et al. Staphylococcus aureus leucocidin, a virulence factor in bovine mastitis. J Dairy Res. 2005 May;72(2):188–94.

109. Molineri AI, Camussone C, Zbrun MV, Suárez Archilla G, Cristiani M, Neder V, et al. Antimicrobial resistance of Staphylococcus aureus isolated from bovine mastitis: Systematic review and meta-analysis. Prev Vet Med. 2021 Mar 1;188:105261.

110. Matuszewska M, Murray GGR, Harrison EM, Holmes MA, Weinert LA. The Evolutionary Genomics of Host Specificity in Staphylococcus aureus. Trends Microbiol. 2020 Jun 1;28(6):465–77.

111. Jarraud S, Lyon GJ, Figueiredo AM, Lina G, Vandenesch F, Etienne J, et al. Exfoliatin-producing strains define a fourth agr specificity group in Staphylococcus aureus. J Bacteriol. 2000 Nov;182(22):6517–22.

112. Jarraud S, Mougel C, Thioulouse J, Lina G, Meugnier H, Forey F, et al. Relationships between Staphylococcus aureus Genetic Background, Virulence Factors, agr Groups (Alleles), and Human Disease. Infect Immun. 2002 Feb;70(2):631–41.

113. Sakoulas G, Eliopoulos GM, Moellering RC, Novick RP, Venkataraman L, Wennersten C, et al. Staphylococcus aureus accessory gene regulator (agr) group II: is there a relationship to the development of intermediate-level glycopeptide resistance? J Infect Dis. 2003 Mar 15;187(6):929–38.

114. Biggs SL, Jennison AV, Bergh H, Graham R, Nimmo G, Whiley D. Limited evidence of patient-to-patient transmission of Staphylococcus aureus strains between children with cystic fibrosis, Queensland, Australia. PLOS ONE. 2022 Oct 7;17(10):e0275256.

115. McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. Nat Microbiol. 2017 Mar 28;2(4):1–5.

116. Andreani NA, Hesse E, Vos M. Prokaryote genome fluidity is dependent on effective population size. ISME J. 2017 Jul;11(7):1719–21.

117. Karcagi I, Draskovits G, Umenhoffer K, Fekete G, Kovács K, Méhi O, et al. Indispensability of Horizontally Transferred Genes and Its Impact on Bacterial Genome Streamlining. Mol Biol Evol. 2016 May 1;33(5):1257–69.

118. Horesh G, Taylor-Brown A, McGimpsey S, Lassalle F, Corander J, Heinz E, et al. Different evolutionary trends form the twilight zone of the bacterial pan-genome. Microb Genomics. 2021;7(9):000670.

119. Rosconi F, Rudmann E, Li J, Surujon D, Anthony J, Frank M, et al. A bacterial pan-genome makes gene essentiality strain-dependent and evolvable. Nat Microbiol. 2022 Oct;7(10):1580–92.

120. Talbot BM, Jacko NF, Petit RA III, Pegues DA, Shumaker MJ, Read TD, et al. Unsuspected Clonal Spread of Methicillin-Resistant Staphylococcus aureus Causing Bloodstream Infections in Hospitalized Adults Detected Using Whole Genome Sequencing. Clin Infect Dis. 2022 Dec 15;75(12):2104–12.

Chapter II - A species-scale pangenomic exploration of *Staphylococcus aureus*

Vishnu Raghuram[1], Zach Karol[2], Robert A. Petit III[3], Daniel B. Weissman[2], Timothy Read[3,*]

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

[2] Department of Physics, Emory University, Atlanta, Georgia, USA

[3] Division of Infectious Diseases, Department of Medicine, Emory University, Atlanta, Georgia, USA

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

Author contributions

VR worked on study conceptualization and design, data curation, analysis, methodology, validation, visualisation, writing and editing.

ZK performed the coding for F_{ST} estimation which was used for the results in **Fig S2**, **Fig 7**, **Fig 8** and **Fig S3**.

RAP collected and ran Bactopia on the public dataset of 83,383 *S. aureus* genomes used in this study.

DBW provided guidance on data analyses.

TDR helped with study conceptualization and design, supervision, funding, resources, writing and editing.

Abstract

Staphylococcus aureus is a major causative agent of both hospital and community acquired infections in humans worldwide. Due to the high incidence of infection, *S. aureus* is also one of the most sampled and sequenced pathogens today. However, most available sequences are biassed towards the few lineages of *S. aureus* associated with human infections. In this paper, we used all publicly available genome sequences of *S. aureus* as of May 2021 (83,383 genomes) to evaluate the true spectrum of diversity of the species. We outlined strategies for dereplication to counter sampling biases as well as to decrease computational resources needed for analyses. After significant filtration and dereplication, we reduced the dataset to 8,166 isolates of diverse *S. aureus* strains and constructed a pangenome. Using this pangenome, we identified naturally occurring thresholds that separate different subspecies of *S. aureus* based on core genes, which we termed "Strain groups". We also found conserved compositions of accessory genes unique to each strain group. Using the fixation index, we identified accessory genes that were specific to one or few strain groups, and also accessory genes that were agnostic of strain groups. Understanding gene gain/loss and gene exchange between different strains of *S. aureus* can provide valuable insights into the past evolutionary history as well as future subspeciation of *S. aureus*.

Importance

We analysed the genetic diversity of *Staphylococcus aureus*, a globally prevalent bacteria that causes infections in humans. We started with a publicly available dataset of 83,383 genome sequences and rationally reduced it to 8,166 genomes by removing duplicates while still maintaining diversity. We constructed a pangenome with the reduced dataset, i.e., a union of all genes found in a population, and used this to outline relationships between the core (genes present in all or most members of the population) and the accessory (genes present only in some members) genome. This dataset captures all the diversity of *S. aureus*, making it an excellent resource for understanding genetic diversity and genome evolution of pathogens. Additionally, this study outlines strategies for processing large genomic datasets which will also benefit the greater microbial genomics community.

Introduction

The first pangenome constructed by Tettelin et al was of *Streptococcus agalactiae* using 8 genomes, and showed that only 80% of a given genome was shared by all isolates (1). This was one of the first studies to demonstrate that a significant amount of the total genomic content of a bacterial species cannot necessarily be represented by a single, or even 8 genomes. The next natural question is how many genome sequences are required to get all possible genes for a given species? Tettelin *et al* also hypothesised that the number of genomes required to discover all possible genes in *S. agalactiae* is non-existent. In other words - the pangenome is open, the more genomes that are added, the more genes will be discovered. However, we know that this cannot be true as the number of bacterial cells and genes are finite. Therefore, the terms "open" and "closed" pangenomes merely indicate the outer bounds of the sampled diversity.

The total gene content of a given species may be shaped by either selective pressures or random drift (2). While the core genome may be stable, the accessory genes are readily exchangeable and/or highly evolvable, leading to a *theoretically* infinite pangenome many times the size of the number of genes in a single genome (3,4). However, as stated earlier, this cannot be true. With the massive increase in publicly available microbial genome sequence data over the past decade and a half, a species pangenome can be now estimated with 100s to 1000s of genomes (5–8). This magnitude of data allows us to move towards closing the pangenome for several species, provided adequate diverse sampling of the species has been performed. The challenge with using publicly available genomes as the starting dataset is that there can be several steps of manual curation and clean-up before the data are usable. Taxonomic misidentifications, contamination, data redundancy and sampling biases are the major contributors. If these hurdles are overcome, the collection of all publicly available genomes of a given species (post-quality filtering) would be the best resource to capture all the (sequenced) genetic diversity.

S. aureus is a ubiquitous nosocomial pathogen responsible for more than 100,000 bloodstream infections in 2017 in the US alone (9). Several pangenome studies with *S. aureus* genomes have been performed using a range of diverse datasets for epidemiological investigations (8,10–14), vaccine candidate discovery (15,16), and evolutionary phylogenomics (5,17–19). The analysis methods used by the above mentioned studies are variable, leading to a prediction of 4000 - 21000 gene families in the *S. aureus* pangenome, with up to approximately 60 sequence types (ST) and 40 clonal complexes (CC). This large variation in the total number of gene families is due to a combination of the diversity of the dataset as well as the pangenome estimation tool used (Tools that split paralogs by default, e.g. ROARY, show larger numbers of gene families (20)). *S. aureus* lineages are discontinuous, with large genomic gaps separating the different CCs possibly due to selective barriers for genetic exchange (21). The standard system of Multi-locus Sequence typing (MLST), though useful for rapid strain typing, is outperformed by whole-genome based methods for lineage assignment (22,23).

Current literature on *S. aureus* speciation suggests that the species is shaped by large recombination events forming CCs, selective genetic barriers between CCs preventing exchange, and long periods of evolution within CCs (21,24). Both large recombination events and long-term clonal evolution manifest as long branches on phylogenetic trees. Therefore, approaches beyond phylogenetics alone may be required to provide a complete picture of the relationship between lineages.

The complete range of diversity and the number of lineages of *S. aureus* was not known at the start of the study. How many lineages are there and can they be distinguished by specific genetic markers? To answer this question, a large-scale pangenomic study integrating all publicly available *S. aureus* sequences with uniform processing is required.

In this study, we used publicly available genomes of the ubiquitous pathogen *S. aureus* to construct the most complete pangenome to date. We processed ~83,000 genome sequences of *S. aureus* using the standardised comprehensive microbial genome analysis pipeline, Bactopia to acquire consistently assembled and annotated genomes (25). We filtered this set to remove low quality, misidentified, and contaminated sequences and then dereplicated by removing identical/near-identical genomes, leaving us with a final set of 8,166 genomes that still represents all the diversity found in the initial dataset. We then used the final set to build a species-wide pangenome.

This pangenome was then used to identify a natural threshold to subdivide the *S. aureus* species into subspecies akin to conventional clonal complex assignments but with higher resolution - termed “Strain groups”. By comparing the core (> 95% of population) and accessory genome (< 95%), we identified that the accessory genome composition within a strain group was conserved for the abundant, established clones of *S. aureus*. In addition, we also found a cluster of different strain groups having similar accessory genome content, suggesting active gene exchange between these strain groups. Moreover, we also found a subset of accessory genes specific to certain strain groups as well as a subset that is randomly present throughout the population. We believe this study has laid the foundation for several upcoming projects uncovering genome evolution and subspeciation in *S. aureus*. This study is not only a resource for the *S. aureus* community, but we also believe it can also be a reference for future large-scale pangenome studies incorporating tens of thousands of sequences.

Results

Filtering

83,383 genomes were processed by the Bactopia pipeline and based on read and assembly quality, were divided into four ranks - Gold, Silver, Bronze and Exclude. Only Samples falling in 'Gold' or 'Silver' ranks were considered for further analysis. Samples flagged by Bactopia and/or CheckM as non-*S. aureus* or as having non-*S.aureus* reads were then removed, followed by samples that appear to be intraspecies mixtures based on minor-allele frequency (**Fig 1**).

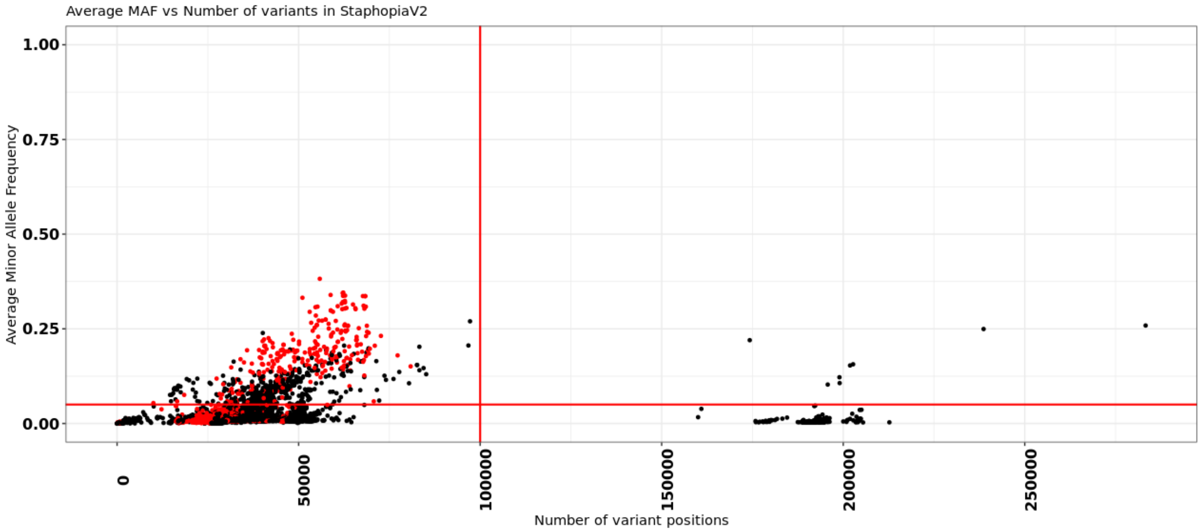

Fig 1: Samples with high average minor allele frequency (MAF), low bactopia quality or high number of variants were filtered out.
The x-axis shows the total number of variants when compared with the Bactopia auto-chosen reference, and the y-axis shows the average minor allele frequency. Each dot is one of 83,383 genomes. Red dots are samples ranked 'Bronze' or 'Exclude' by Bactopia and were discarded. Samples in the top quadrant (Average MAF > 0.05) were considered to be contaminated and were discarded. Samples in the right quadrant (> 100,000 total variants) were considered non-S. aureus and were discarded. The remaining samples in the bottom left quadrant (< 0.05 Average MAF and < 100,000 total variants) were used for further analysis.

Based on the above plot - the non-*S. aureus* samples are the points with > 100,000 variant positions (x axis) when compared to a *S. aureus* reference sequence (

variation >5% of the genome). Samples with average minor allele frequency > 0.05 (y axis) were considered intraspecies mixtures. Red points are samples ranked 'Bronze' or 'Exclude'. Leaving only black points in the bottom left quadrant. This filtered 83,383 samples down to 56,771.

Clustering

This set of 56,771 samples were grouped into Sequence Types (ST) based on PubMLST defined groupings and all-vs-all pairwise mash distances were calculated for each ST. Samples with unassigned STs were grouped together. Based on a pilot study using a dataset of 380 genomes defined by our previous Staphopia V1 study as a 'Non-redundant diversity' set (NRD), we identified that a mash distance of 0.0005 corresponds to approximately 50 SNPs (26,27). We also further explained the rationale in a blog post (28) In other words, samples < ~50 SNPs apart were clustered together. A representative was chosen from each cluster at random. This collapsed 56,771 into 7,654 genomes.

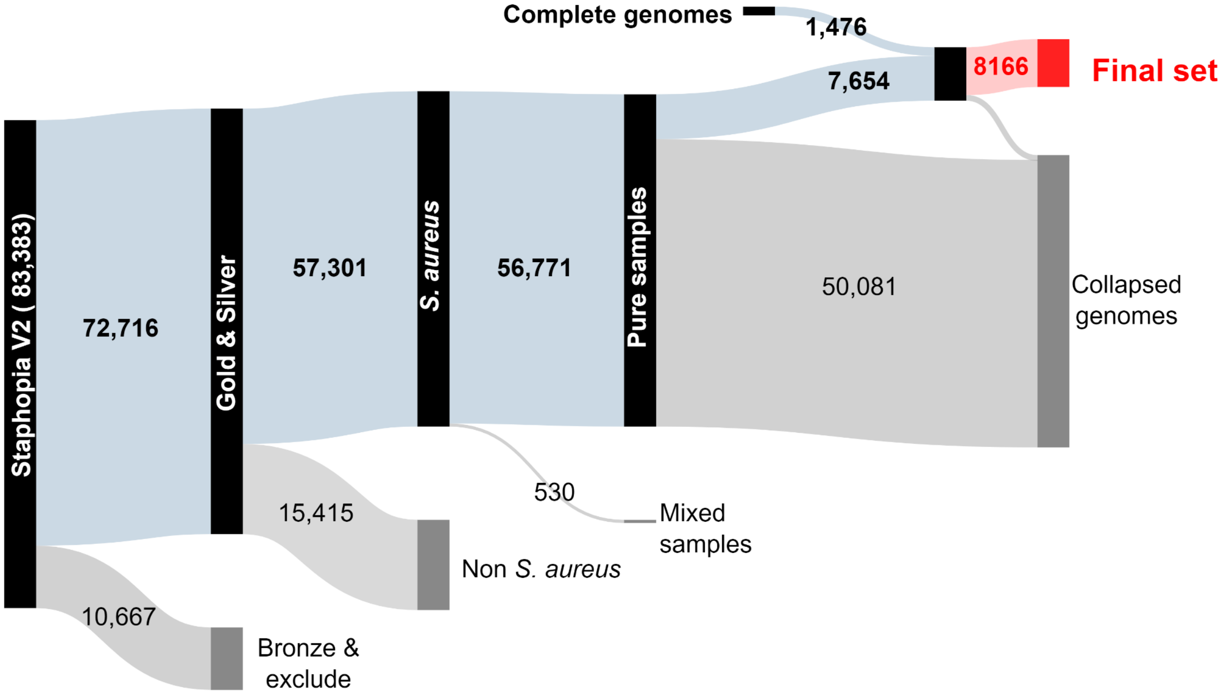

Fig 2: Sankey diagram showing the fate of 83,383 *S. aureus* genomes after processing and filtering.

As the initial set of 83,383 genomes only comprised publicly available data from short-read sequencing, we also added to this set complete *S. aureus* genomes that were assembled using long-read/hybrid methods. We added 1,476 complete *S. aureus* assemblies from NCBI to our dereplicated set of 7,654 genomes and redid the same mash-distance based clustering protocol. This led to our final set of 8,166 genomes (**Fig 2).** For the purposes of this document we will use 'the final dataset' to refer to our dereplicated set of 8,166 genomes.

Pangenome construction

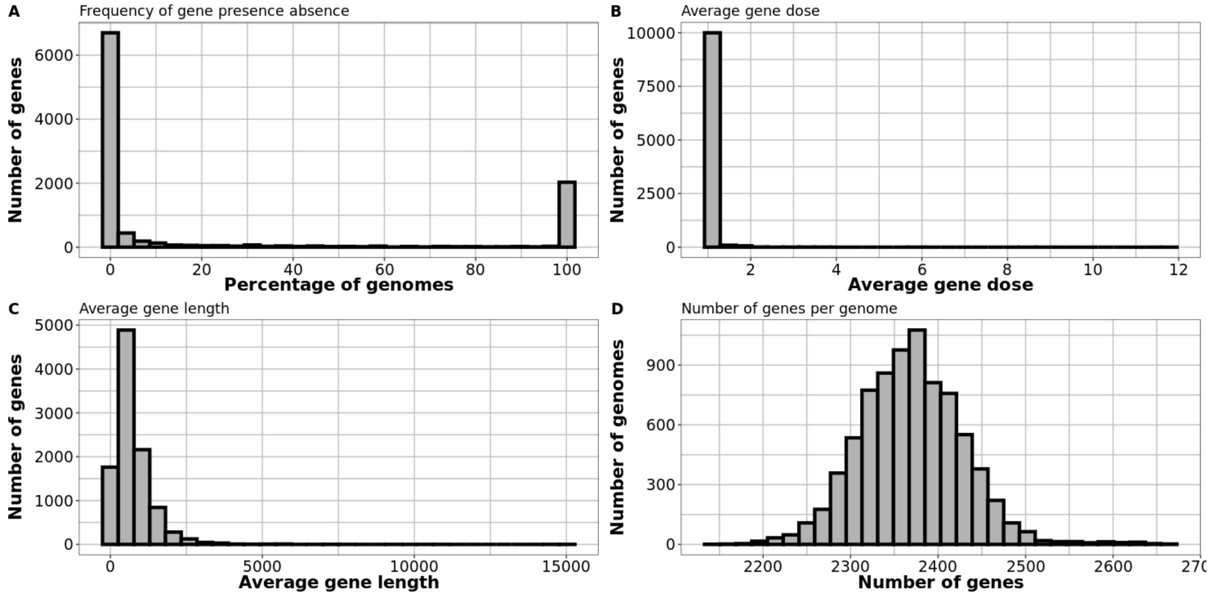

Fig 3: Description of the pangenome of *S. aureus*.

Histograms depicting the (**A**) frequency distribution of genes in our dataset, (**B**) the average dosage of each gene per genome, (**C**) the average length distribution of each gene, and (**D**) the distribution of the number of genes per genome.

We used Bakta to annotate the final dataset and ran PIRATE to build the pangenome. According to some basic metrics from PIRATE, 10,714 unique genes (including alleles) were found, out of which 20.2% (2054 genes) were found in > 95% of the dataset. These 2,054 genes were considered to be core genes. 7.3% (741 genes) were found at intermediate frequencies (more than 10% of the dataset but less than 95%) and these were considered intermediate genes. 72.5% (7,379 genes) were found in less than 10% of the dataset and these were considered rare/unique genes. The intermediate genes and rare genes collectively were considered the accessory genome. Most genes (90%) were in single copy and the average gene length was 733bp. The average number of genes per genome was 2,370 (**Fig 3**).

Lineage assignment

The current method for *S. aureus* lineage assignment uses Multi-locus sequence typing (MLST) to assign STs and STs that are identical by 5 alleles or more are considered to be in the same Clonal Complex (CC). Currently, according to PubMLST, 10 CCs of *S. aureus* are defined. In our initial filtered set of 56,771 genomes, 20% of samples belonged to the group of unassigned CCs.

While major prominent CCs have been defined, we still do not know the true spectrum of diversity in *S. aureus* - how many CCs are there? Identification of emerging/minority CCs can help establish relatedness and gene flow between CCs. To begin to answer these questions, we performed core genome alignment using PIRATE for the final dataset and calculated all vs all pairwise SNP distances using snp-dist (29).

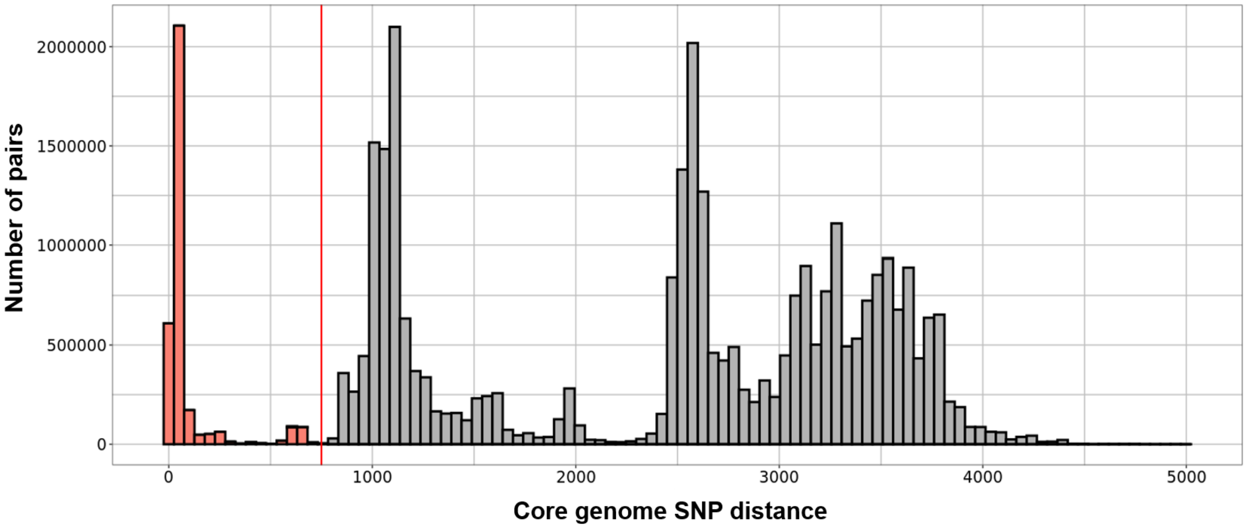

Fig 4: Natural boundaries in core genome SNP distances can be used to categorise strains. For our dataset of 8166 isolates, all-vs-all pairwise SNP distances were calculated and plotted as a histogram. Sample pairs less than 750 core genome SNPs apart were grouped into the same cluster or "strain groups" (Sample pairs to the left of the red line).

The above histogram shows that the strains fall into 3 natural groups based on SNP distances with prominent valleys in-between each group. This suggests the first group with the least SNP distances (< 750, red line) are likely comparisons of strains from the same lineage. The second and third group may correspond to strains from different but related lineages, and completely unrelated lineages respectively. Moreover, we found that strain pairs that belong to the same CC according to the current definition (>= 5/7 MLST allele match) have a core SNP distance > 750, causing them to bleed into the second group (**Fig S1**). This suggests that the natural grouping of *S. aureus* strains is better defined by core gene SNPs compared to MLST alleles alone. From hereon, we used the 750 core gene SNP threshold to assign strain groups to the final dataset by clustering strains < 750 core gene SNPs apart (Fig 4). This led to 136 clusters, these clusters will be referred to as 'strain groups'. These strain groups were also assigned to the total set of 56,771 samples prior to dereplication based on the assigned strain group of each dereplicated cluster.

Gene discovery and lineage discovery

We wanted to estimate the rate of discovery of new genes and new lineages or strain-groups using our pangenome compared to a random set of *S. aureus* genomes obtained from NCBI. This is primarily to answer the question - is a dereplicated dataset better for sampling diversity than a random dataset? To answer this, we calculated the total number of gene families discovered with an increasing number of genomes sampled (One genome to one thousand genomes). We performed this calculation for a random set of 1000 genomes from our dereplicated set and a random set of 1,000 genomes from the total set of 56,771 genomes before dereplication. We then repeated this subsampling five times. Similarly, We also calculated the total number of strain groups discovered with an increasing number of genomes (**Fig 5**).

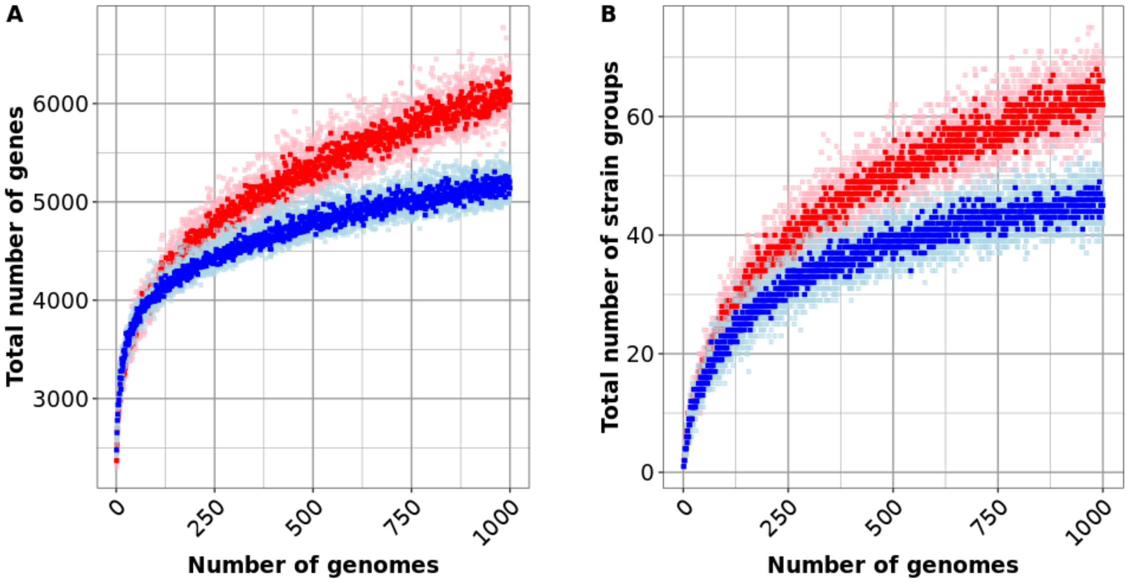

Fig 5: The dereplicated dataset provides an increased number of genes and an increased number of strain groups with the same number of genomes sampled compared to the total dataset.

The x-axis depicts the number of genomes sampled and the y-axis depicts (**A**) the total number of genes as described by PIRATE or, (**B**) the total number of strain groups as described in **Fig 4**. Red dots correspond to the dereplicated set and blue dots correspond to the non-dereplicated set. The light coloured dots represent the number of genes or number of strain groups for each iteration of the random sampling. The dark coloured dots represent the median obtained from all five iterations.

We found that although the rate of gene or strain-group discovery is similar for the first ~100 genomes, the dereplicated dataset discovered an average of 1,031 new genes (**Fig 5A**) and 20 new strain groups (**Fig 5B**) more than the non-dereplicated set after 1000 genomes. This suggests that a random dataset can sample common genes and lineages of *S. aureus* but our dereplicated dataset has a better sampling of minority/rare lineages. This highlights the importance of a diverse initial dataset for achieving pangenome completeness - a random set from ~56k genomes has less total diversity than a random set from our dereplicated ~8k genomes.

Relationship between core and accessory genome

We constructed a Maximum Likelihood phylogeny using IQ-TREE with the core genome alignment obtained from PIRATE. We then rooted the tree at an ST93 strain and coloured the tips based on SNP groups as described in Raghuram *et al* (27) (**Fig 6A**).

Another method to observe natural strain groupings would be to visualise accessory gene composition. To perform this, we created a gene presence-absence matrix for the final dataset comprising only accessory genes (present in > 10% but < 95% of samples) and used tSNE for dimensionality reduction and plotted the resulting coordinates. We then coloured each point based on the core-gene strain groups after the fact (**Fig 6B**).

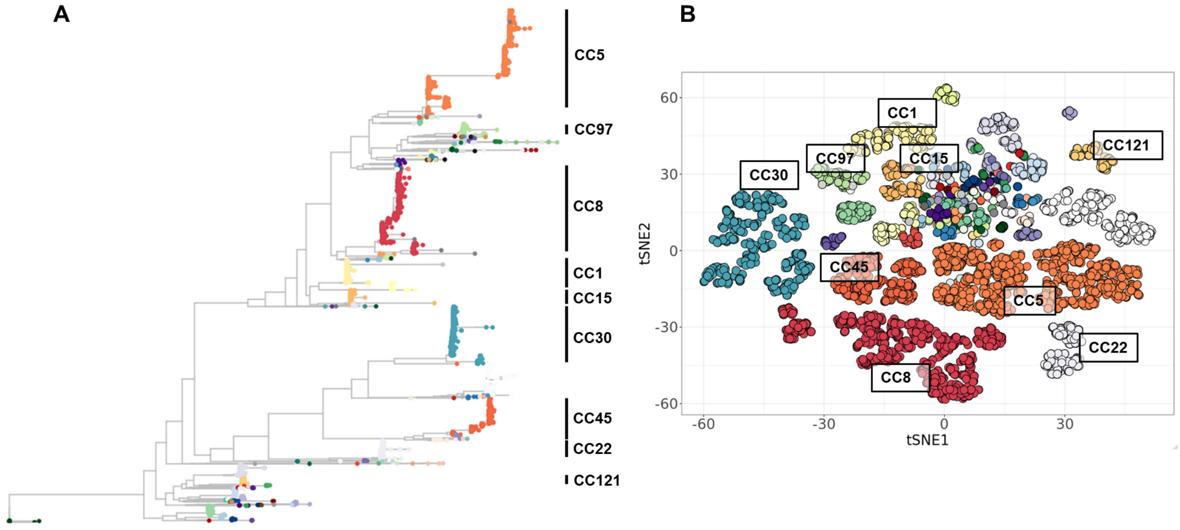

Fig 6: Prominent strain groups from their own clades on a core genome phylogeny and distinct clusters based on their accessory genome composition.

(**A**) Maximum likelihood phylogeny (GTR+FO model, 1000 ultrafast bootstrap replicates) of

8,166 *S. aureus* strains with each tip coloured by the designated strain group. Scale bar indicates the number of substitutions per site. The conventional CC mapping is stated to the right of the corresponding clades. (**B**) tSNE plot with each dot depicting one of the 8,166 genomes and its position in 2D space representing the accessory genome composition. Dots are coloured based on their strain group designations and the corresponding conventional CC assignment is also stated next to each strain group cluster.

Based on the above figure (**Fig 6**), we see that the conventional CC groupings and accessory gene composition mostly overlap - suggesting that accessory gene composition is unique to each CC at least for the major defined CCs. Our strain group assignment is also in concordance with the accessory gene clusters as well as the conventional CCs. However, our strain groups have resolved previously unassigned CCs into their own clusters, which is also in concordance with their accessory genome composition.

The large cluster of mixed SNP groups in the middle of the tSNE plot (**Fig 6B**) corresponds to two major distinct clades on the tree (**Fig 6A**). Suggesting gene exchange between these two clades has led to a distinct core genome but a similar accessory genome, making these strains an interesting case for studying gene exchange between different lineages.

In addition to grouping unassigned samples, the current method of CC assignment appears to lead to some false groupings. For example, CC1 likely needs to be split into multiple CCs based on both core gene and accessory gene groupings (**Fig S1**). This further demonstrates the need for a better system of assigning lineages. We believe our SNP groups provide better resolution for the analysis of population structure.

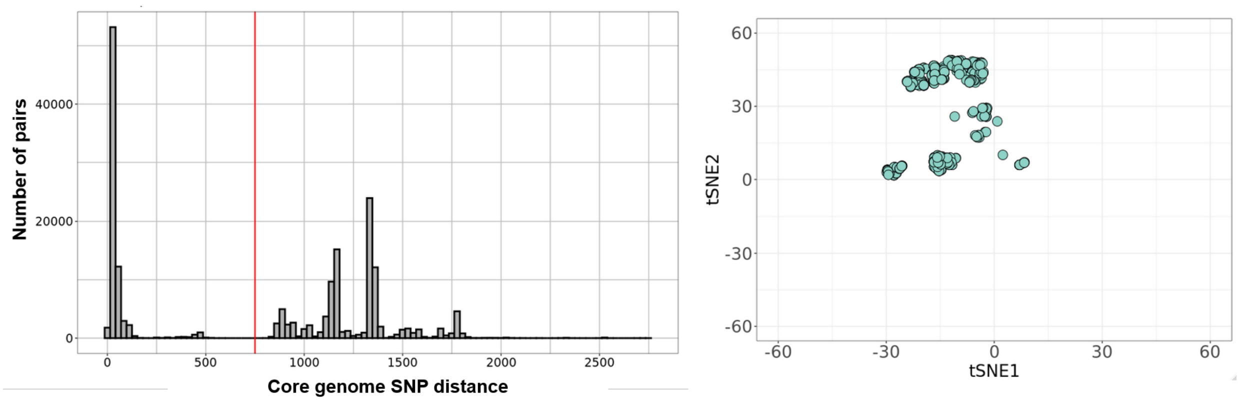

Fig S1: Pairwise core-genome SNP distance histogram (LEFT) and tSNE based on accessory genome composition (RIGHT) suggest CC1 S. aureus strains defined by MLST are likely multiple strain groups.

<u>Lineage specific genes and fixation index</u>

Next, we wanted to understand the prevalence of different intermediate genes across different strain groups. This would help us identify lineage specific markers if any. F_{ST} or fixation index is a measure of genetic segregation of a trait between different populations (30). In this context, the populations refer to our strain groups and traits refer to presence/absence of specific genes. In addition to identifying lineage or strain groups specific markers, we also wanted to know if markers with specific genomic context (Chromosomal, plasmid or phage associated) are more or less likely to be fixed in a strain group. We used geNomad to predict mobile genetic elements (MGEs) and calculated F_{ST} for each gene (31). F_{ST} of 0 indicates a gene that displays no genetic segregation, i.e it is indiscriminately found across members of different populations. In contrast, F_{ST} of 1 indicates perfect genetic segregation, i.e it is only found in specific populations.

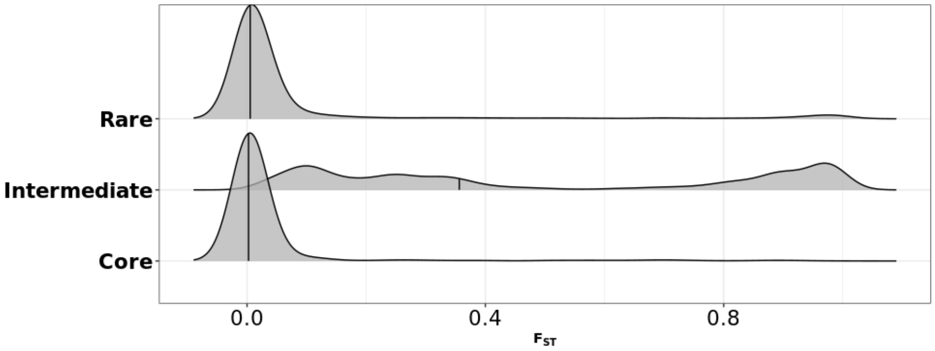

Fig S2: Intermediate F_{ST} genes show bimodal distribution of either high or low F_{ST}. Ridgeline plot depicting F_{ST} distribution for Core (> 95%), intermediate (10 - 95%) and rare (<5%) genes. Height of plot normalised for total number of genes in each group.

Upon examining the distribution of F_{ST} across core, intermediate, and rare genes we found that as expected, the core genes have low (near 0) F_{ST} as by definition they are found in all members of the population and are not strain group specific (**Fig S2**).

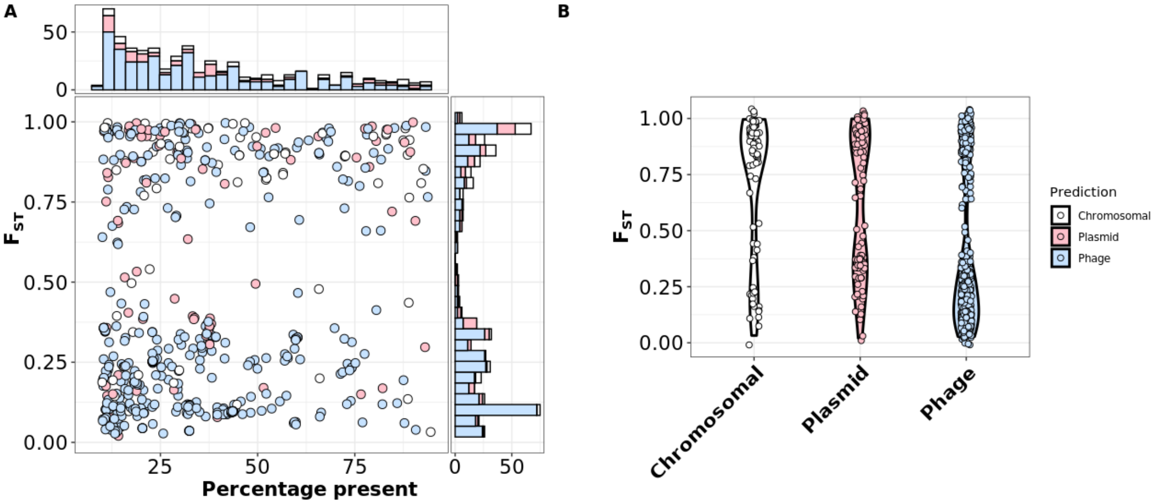

Fig 7: Mobile genetic elements were not associated with high or low F_{ST} in intermediate genes.
(**A**) Dot plot showing percentage prevalence of only intermediate genes (> 10%, < 95%) on the x-axis and the corresponding F_{ST} on the y-axis. Histograms along the x and y axis show density of dots along each axis. (**B**) Violin plot showing distribution of F_{ST} for each geNomad prediction category. There is no significant difference in the F_{ST} across the three different categories (Kruskal Wallis test, $p > 0.01$). Colour of dots show geNomad prediction (Chromosomal, plasmid, phage).

Similarly, we also found rare genes to have low F_{ST}. This is likely due to the fact that since they are present in very few members, they are not fixed in any specific strain group. Interestingly, the F_{ST} distribution across intermediate genes showed a distinct bimodal distribution compared to the core and rare genes. This suggests some intermediate genes are strain-group specific while others are not (**Fig S2, Fig 7A**). We also observed no significant difference in the F_{ST} distribution between Chromosomal, plasmid and phage associated genes (Kruskal Wallis $p > 0.01$) (**Fig 7B**). Genes with geNomad prediction probability < 0.5 were not considered for this analysis. Collectively, these results show that specific chromosomal and MGE present in intermediate frequencies in the populations are strain-group specific and can be used as markers to differentiate between *S. aureus* lineages. The F_{ST} can also be used to examine patterns of distribution of specific virulence factors across the *S. aureus* species-wide phylogeny. We selected a few well known *S. aureus* toxins - Panton-Valentine Leukocidin (PVL), Toxic Shock Syndrome toxin 1 (TSST), and different types of *Staphylococcal* Enterotoxins (SEA, SEB, SEG, SEO). PVL comprises two phage-encoded proteins, LukF-PV and LukS-PV, both acting synergistically to form pores in host-cell membranes (32). TSST and SEs are superantigens, highly potent toxins that can elicit severe inflammatory responses and other immunomodulatory effects (33). Then, We mapped the presence/absence of these toxins and their corresponding F_{ST} scores to our core genome phylogeny (**Fig 9**)

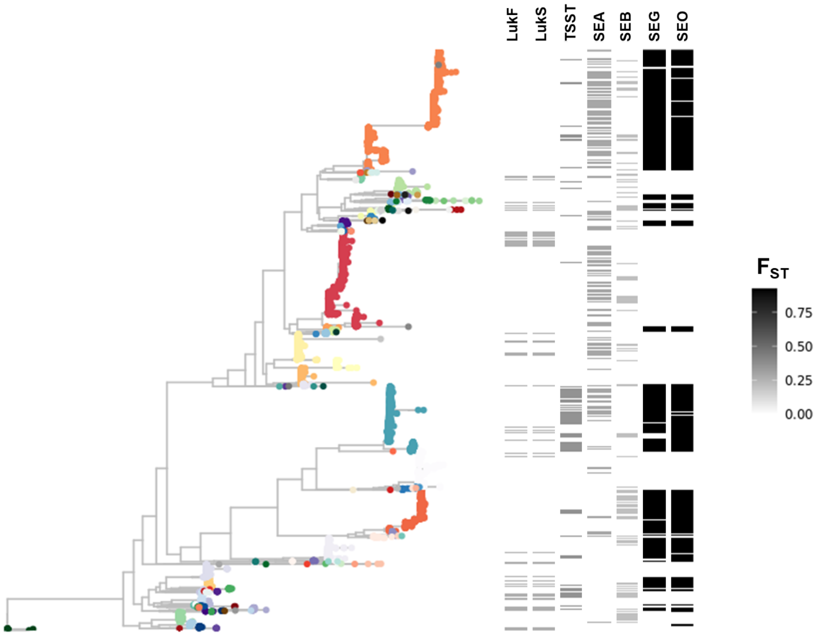

Fig 8: Strain-group specificity and co-occurrence of specific Staphylococcal toxins. Core genome phylogeny is the same as described in **Fig 6A**. Heatmap on right shows presence absence and F_{ST} of specific Staphylococcal toxins - Panton-Valentine Leukocidin (LukF and LukS), Toxic Shock Syndrome Toxin (TSST), and Staphylococcal Enterotoxins type A, B, G, O (SEA, SEB, SEG, SEO).

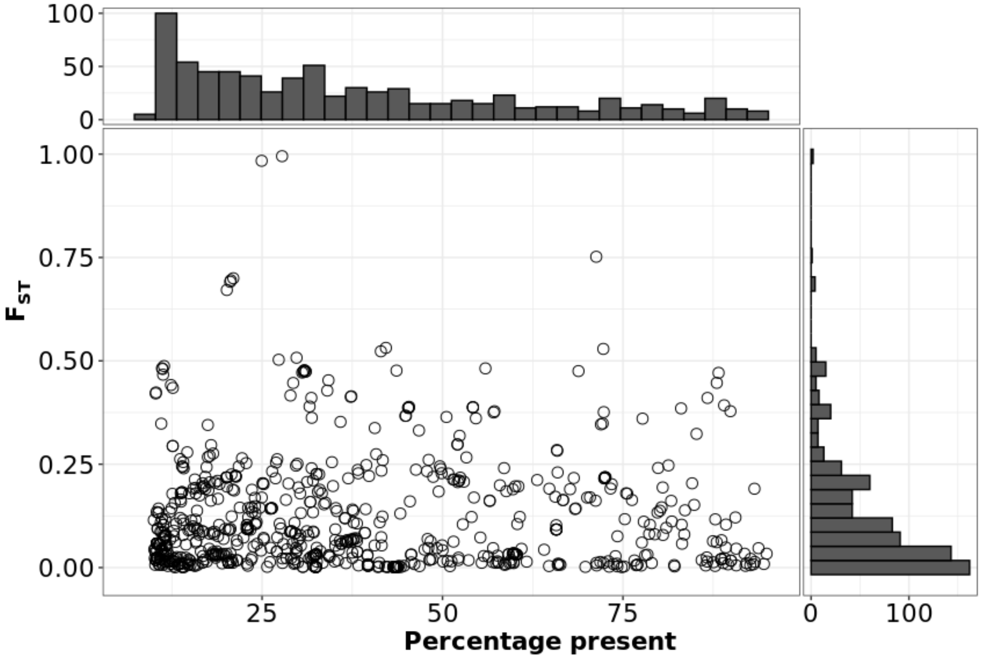

Fig S3: There are no *agr* group specific intermediate genes aside from *agrD*.
Dot plot showing percentage prevalence of only intermediate genes (> 10%, < 95%) on the

x-axis and the corresponding F_{ST} on the y-axis. F_{ST} scores calculated for *agr* type-based population segregation. The three dots > 0.75 F_{ST} correspond to the *agrD* of three out of four *agr* groups which are known to be lineage specific (27). The *agrD* of the fourth *agr* group is absent in this plot as it is present in < 10% of the population.

We found that PVL (LukF and LukS), TSST, SEA and SEB were not lineage specific and had low F_{ST} (~ 0.25). Interestingly, the enterotoxins SEG and SEO, not only had high F_{ST} (> 0.9) suggesting they are strain-group specific, they also appeared to largely co-occur. This result demonstrates that there are patterns of fixation and co-occurrence of toxins across specific lineages of *S. aureus* and understanding these patterns can provide insight into their emergence and pathogenic potential.

Discussion

Why our pangenome is "not another pangenome"

In this study, we used the publicly available collection of over 80,000 genomes labelled as *S. aureus* and reduced that dataset to approximately 1/10th the size (~8000) while still maintaining the genetic diversity. Here, we outlined an approach for uniform data processing and filtering steps to identify possible misidentifications and contaminated genomes. Out of 83,383 input short-read genome sequences, we discovered that 15,363 genomes (18%) were either not *S. aureus* or contained detectable amounts of non-*S. aureus* reads (**Fig 1**, **Fig 2**). This emphasises the importance of appropriate cleanup steps while using publicly available datasets. Moreover, using minor allele frequencies, we also discovered 530 genomes (0.6%) having intra-species mixtures, suggesting the sequencing source may not have been a pure single colony (See Chapter IV for more details).

Here, we de-duplicated the complete public dataset of all *S. aureus* genomes, leading to the largest *S. aureus* pangenome constructed to date that still represents the total sequenced diversity. Dereplication also significantly reduced data storage and computation time of downstream analysis steps while also reducing artifactual results from unequal sampling. The workflow outlined here could be used for any bacterial species. We found that our dereplicated set provided faster discovery of new genes and new strain groups compared to a random non-dereplicated set, further highlighting the importance of dereplication (Fig 5).

What did we learn about *S. aureus*?

Currently, the most common method to assign lineages/sublineages in *S. aureus* is to use MLST-based groupings. MLST assignment, despite being based on alleles of only seven core genes, still provides an accurate picture of the species structure. However, as highlighted here, there are cases where MLST alone is not sufficient to resolve differences between lineages (eg: CC1 - vastly different STs grouped into the same CCs). Though alternate MLST schema as well as core-genome and whole-genome MLST schema have been developed for *S. aureus* (34–36), they are still allele based methods and do not take into account other types of genomic variation (23). In this study, we used natural grouping of individual genomes based on core gene SNP distances to assign lineages. As has been documented before, *S. aureus* appears to be split into two broad clades, with CC groupings within each clade. Upon observing the pairwise core-gene SNP distances across our dereplicated dataset, we found three distinct distributions - within lineage (or within CC), between-lineage in the same clade, and between-clade, with clear valleys between each distribution. These valleys serve as naturally occurring thresholds which we can use to delineate the species structure of *S. aureus* (**Fig 4**). Our strain group assignment was also in concordance with the conventional CC assignment for the prominent CCs (**Fig 6A**). In addition, we was also able to resolve spurious CC assignments using our strain group SNP threshold (**Fig S1**)

It is hypothesised that the formation of *S. aureus* lineages was driven by large recombination and rearrangement events followed by clonal expansion, giving rise to the distinct lineages we see today. This process of sub-speciation may be driven by biological barriers preventing between-CC homogenization. These barriers include but are not limited to restriction modification systems, phage host range, and CRISPR interference (24,37–41). Upon observing core and accessory genome content, we observed that the accessory genome composition was largely unique to the core genome of specific lineages. In addition, previous studies show that the evolutionary history of *S. aureus* was also shaped by the *agr* quorum sensing operon, predating the CC divergence (27,42,43). Each CC is almost exclusively linked to one out of the four *agr* types with no *agr* type exchange between CCs (See **Chapter III** for more details). This suggests that the CC genetic background is a barrier to *agr* exchange. This is corroborated by the fact that we did not see gene fixation patterns specific to *agr* groups independent of the phylogeny aside from *agrD* (**Fig S3**).

From examining the strain-group specific fixation patterns of intermediate frequency accessory genes, we found that 184 out of 741 intermediate frequency genes (25%) had an $F_{ST} > 0.9$ (**Fig 7A**). This suggests that most intermediate genes are not lineage-specific, which is in stark contrast to *E. coli* according to a recent study where they found 84% of intermediate genes to be lineage specific (44). We also found that though most intermediate genes were predicted to be phage related, we saw no significant association between F_{ST} and MGEs (**Fig 7B**).

A key observation was that of a bimodal F_{ST} distribution pattern for our intermediate genes (**Fig S2**), with 558 out of 741 (75%) of genes having F_{ST} either > 0.75 or < 0.25, meaning 75% of all intermediate genes are either nearly fixed in specific lineages or are completely randomly distributed. The intermediate genes with low F_{ST} are maintained in the population yet have high turn over, i.e they are gained and lost repeatedly (**Fig 8** - LukFS, TSST, SEA, SEB). In contrast to rare genes that also have high turnover but are not maintained, i.e they are gained, lost and almost never gained again. Whether or not pangenomes are adaptive is up for debate (2,45–47). However, there is evidence to suggest that the rate of deletion of genes is greater than the rate of acquisition (48,49), and that the acquisition and maintenance of new genes is at least slightly beneficial (50–52). Collectively, our results are consistent with the hypothesis that rare genes are acquired neutrally while intermediate genes are adaptive to specific micro-niches, hence their maintainence. Movement between different niches can lead to further gain/loss of genetic material (53).

The next natural question is what does this mean for *S. aureus* strain formation? We did observe the prominent strain groups having distinct accessory genome compositions as well as several minority strain groups across the core-genome phylogeny with overlapping accessory genome compositions (**Fig 6B**). This suggests that these lineages are currently undergoing active recombination events, presumably exchanging intermediate genes, and may be in the process of lineage formation. Understanding these gene gain/loss and fixation events would serve as a natural laboratory for observing speciation in *S. aureus*. Overall, these results show that the core genome makes the species, and the accessory genome makes the strain.

What can we do further?

Understanding the species-wide mutational landscape of specific genes along with their associated lineages can provide important insights into past evolutionary history as well as predict future patterns (28). Our pangenome can be used to examine the lineage specificity and mutational signatures of specific key genes involved in virulence (PVL, *agr*), antibiotic resistance (vraS), or genes having known associations with diseases such as bacteremia (*sarZ* (54), *tcaA* (55)). Apart from individual genes/operons, this pangenome is also a resource for analysis of plasmids, phages, other mobile genetic elements, pseudogenization patterns, and gene turnover rates in *S. aureus*.

Methods

Genome collection and processing by Bactopia

Bactopia v1.7.0 was used to download and process all genomes used in this dataset. Bactopia is a software pipeline for comprehensive analysis of bacterial genomes based on Nextflow (25,56). The command `bactopia search "Staphylococcus aureus" --prefix saureus` was used to download all *S. aureus* short-read sequences available on Sequence Read Archive (SRA) as of May 2021. Then, each genome was processed using a custom nextflow wrapper script (57). In short, Bactopia used SKESA to assemble genomes, Bakta to annotate and Snippy for variant calling (58,59). Assembly quality was evaluated using QUAST and CheckM (60,61).

Filtering low quality samples:

Only samples having greater than 50× coverage, mean per-read quality greater than 20, mean read length greater than 75 bp, and an assembly with less than 200 contigs were considered for the analysis (corresponding to 'Gold' and 'Silver' ranks as designated by Bactopia. Samples that were detected as not *S. aureus* according to kmer based identification or CheckM were then removed. Coverage for all samples were capped at 100x.

Filtering mixed strain samples:

For every sample, bactopia performs variant calling using Snippy against an auto-chosen reference sequence based on the smallest MASH distance to a complete *S. aureus* genome in RefSeq (59,62). For each variant identified, the allele frequencies were calculated from the bam files using bcftools mpileup (63). Samples having average allele frequency > 0.05 were considered mixed strains and therefore removed. This process reduced 83,383 samples to 56,771.

Clustering and dereplication:

Samples were grouped by their MLST types as assigned by Bactopia and for each ST, an all vs all MASH distance estimation was run. Samples with a MASH distance < 0.0005 were clustered and a random representative was chosen. Samples with unknown STs were grouped together and treated the same. This reduced the filtered set of 56,771 to a set of 7,654. Since Bactopia collected and processed only short read *S. aureus* data, we added 1476 complete *S. aureus* genome sequences to this set and performed the MASH distance based clustering again. Cluster representatives were again chosen at random, however, where possible, we replaced the cluster representative with a randomly chosen complete genome. The resulting final dereplicated set comprised 8166 genomes and was used for pangenome construction.

Pangenome analysis:

Bakta 1.5.1 (64) with default parameters was used to annotate the dereplicated set of 8166 genomes and the resulting gff files were used for pangenome construction with PIRATE 1.0.5 (65). PIRATE was run using default parameters with the additional flags `-a` to obtain core genome alignments and `-k “--diamond”` to use DIAMOND for the amino-acid sequence comparisons (66). snp-dists v0.7.0 was run on the PIRATE core genome alignment to obtain all-vs-all pairwise SNP distances (29). The PIRATE core genome alignment was also used to construct a core genome phylogeny with IQ-TREE 1.6.12 (GTR+FO model, 1000 ultrafast bootstrap replicates) and was visualised using the R package ggtree (67,68).

geNomad v1.5 was used to predict mobile genetic elements (31). AgrVATE v1.0.5 was used to assign *agr* groups (27).

Statistical analysis and data visualisation:

All statistics and tSNE were performed in R using packages stats and rstatix (69,70). All plots were visualised using R package ggplot2 (71). Other visualisations were performed using draw.io and Sakneymatic (72,73).

References

1. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome." Proc Natl Acad Sci. 2005 Sep 27;102(39):13950–5.
2. Shapiro BJ. The population genetics of pangenomes. Nat Microbiol. 2017 Dec;2(12):1574–1574.
3. Lapierre P, Gogarten JP. Estimating the size of the bacterial pan-genome. Trends Genet. 2009 Mar 1;25(3):107–10.
4. Bobay LM, Ochman H. Factors driving effective population size and pan-genome evolution in bacteria. BMC Evol Biol. 2018 Oct 12;18(1):153.
5. Yebra G, Harling-Lee JD, Lycett S, Aarestrup FM, Larsen G, Cavaco LM, et al. Multiclonal human origin and global expansion of an endemic bacterial pathogen of livestock. Proc Natl Acad Sci. 2022 Dec 13;119(50):e2211217119.
6. Cummins EA, Hall RJ, Connor C, McInerney JO, McNally A. Distinct evolutionary trajectories in the Escherichia coli pangenome occur within sequence types. Microb Genomics. 2022 Nov 23;8(11):mgen000903.
7. Rocha J, Henriques I, Gomila M, Manaia CM. Common and distinctive genomic features of Klebsiella pneumoniae thriving in the natural environment or in clinical settings. Sci Rep. 2022 Jun 21;12:10441.
8. Jamrozy DM, Harris SR, Mohamed N, Peacock SJ, Tan CY, Parkhill J, et al. Pan-genomic perspective on the evolution of the Staphylococcus aureus USA300 epidemic. Microb Genomics. 2016;2(5):e000058.
9. Kourtis AP, Hatfield K, Baggs J, Mu Y, See I, Epson E, et al. 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 Mar 8;68(9):214–9.
10. Long DR, Wolter DJ, Lee M, Precit M, McLean K, Holmes E, et al. Polyclonality, Shared Strains, and Convergent Evolution in Chronic Cystic Fibrosis Staphylococcus aureus Airway Infection. Am J Respir Crit Care Med. 2021 May;203(9):1127–37.
11. Montelongo C, Mores CR, Putonti C, Wolfe AJ, Abouelfetouh A. Whole-Genome Sequencing of Staphylococcus aureus and Staphylococcus haemolyticus Clinical Isolates from Egypt. Microbiol Spectr. 2022 Jun 21;10(4):e02413-21.
12. Xu Z, Yuan C. Molecular Epidemiology of Staphylococcus aureus in China Reveals the Key Gene Features Involved in Epidemic Transmission and Adaptive Evolution. Microbiol Spectr. 2022 Oct 3;10(5):e01564-22.
13. Holm MKA, Jørgensen KM, Bagge K, Worning P, Pedersen M, Westh H, et al. Estimated Roles of the Carrier and the Bacterial Strain When Methicillin-Resistant Staphylococcus aureus Decolonization Fails: a Case-Control Study. Microbiol Spectr. 2022 Aug 24;10(5):e01296-22.
14. Cella E, Sutcliffe CG, Tso C, Paul E, Ritchie N, Colelay J, et al. Carriage prevalence and genomic epidemiology of Staphylococcus aureus among Native American children and adults in the Southwestern USA. Microb Genomics. 2022;8(5):000806.

15. Naz K, Ullah N, Naz A, Irum S, Dar HA, Zaheer T, et al. The Epidemiological and Pangenome Landscape of Staphylococcus aureus and Identification of Conserved Novel Candidate Vaccine Antigens. Curr Proteomics. 19(1):114–26.
16. Naz K, Naz A, Ashraf ST, Rizwan M, Ahmad J, Baumbach J, et al. PanRV: Pangenome-reverse vaccinology approach for identifications of potential vaccine candidates in microbial pangenome. BMC Bioinformatics. 2019 Mar 12;20(1):123.
17. Blaustein RA, McFarland AG, Ben Maamar S, Lopez A, Castro-Wallace S, Hartmann EM. Pangenomic Approach To Understanding Microbial Adaptations within a Model Built Environment, the International Space Station, Relative to Human Hosts and Soil. mSystems. 2019 Jan 8;4(1):e00281-18.
18. Rao RT, Sivakumar N, Jayakumar K. Analyses of Livestock-Associated Staphylococcus aureus Pan-Genomes Suggest Virulence Is Not Primary Interest in Evolution of Its Genome. OMICS J Integr Biol. 2019 Apr;23(4):224–36.
19. John J, George S, Nori SRC, Nelson-Sathi S. Phylogenomic Analysis Reveals the Evolutionary Route of Resistant Genes in Staphylococcus aureus. Genome Biol Evol. 2019 Oct 1;11(10):2917–26.
20. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015 Nov 15;31(22):3691–3.
21. Planet PJ, Narechania A, Chen L, Mathema B, Boundy S, Archer G, et al. Architecture of a Species: Phylogenomics of Staphylococcus aureus. Trends Microbiol. 2017 Feb 1;25(2):153–66.
22. Falush D. Toward the Use of Genomics to Study Microevolutionary Change in Bacteria. PLOS Genet. 2009 Oct 26;5(10):e1000627.
23. Maiden MCJ, van Rensburg MJJ, Bray JE, Earle SG, Ford SA, Jolley KA, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol. 2013 Oct;11(10):728–36.
24. Feil EJ. Small change: keeping pace with microevolution. Nat Rev Microbiol. 2004 Jun;2(6):483–95.
25. Petit RA, Read TD. Bactopia: a Flexible Pipeline for Complete Analysis of Bacterial Genomes. mSystems. 2020 Aug 4;5(4):e00190-20.
26. Iii RAP, Read TD. Staphylococcus aureus viewed from the perspective of 40,000+ genomes. PeerJ. 2018 Jul 12;6:e5261.
27. Raghuram V, Alexander AM, Loo HQ, Petit RA, Goldberg JB, Read TD. Species-Wide Phylogenomics of the Staphylococcus aureus Agr Operon Revealed Convergent Evolution of Frameshift Mutations. Microbiol Spectr. 2022 Jan 19;10(1):e01334-21.
28. Raghuram V, Read T. Help, I have too many genome sequences! 2022 Jan 21 [cited 2023 Apr 23]; Available from: https://zenodo.org/record/7278310
29. Seemann T. Source code for snp-dists software [Internet]. Zenodo; 2018 [cited 2023 Mar 15]. Available from: https://zenodo.org/record/1411986
30. Wright S. Isolation by Distance. Genetics. 1943 Mar;28(2):114–38.
31. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, et al. You can move, but you can’t hide: identification of mobile genetic elements with geNomad [Internet]. bioRxiv; 2023 [cited 2023 Apr 23]. p. 2023.03.05.531206. Available from: https://www.biorxiv.org/content/10.1101/2023.03.05.531206v1
32. Melles DC, van Leeuwen WB, Boelens HAM, Peeters JK, Verbrugh HA, van Belkum

A. Panton-Valentine Leukocidin Genes in Staphylococcus aureus. Emerg Infect Dis. 2006 Jul;12(7):1174–5.

33. Krakauer T. Staphylococcal Superantigens: Pyrogenic Toxins Induce Toxic Shock. Toxins [Internet]. 2019 Mar [cited 2023 Apr 24];11(3). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6468478/
34. Liu N, Liu D, Li K, Hu S, He Z. Pan-Genome Analysis of Staphylococcus aureus Reveals Key Factors Influencing Genomic Plasticity. Microbiol Spectr. 2022 Nov;10(6):e03117-22.
35. Jalil M, Quddos F, Anwer F, Nasir S, Rahman A, Alharbi M, et al. Comparative Pan-Genomic Analysis Revealed an Improved Multi-Locus Sequence Typing Scheme for Staphylococcus aureus. Genes. 2022 Nov;13(11):2160.
36. Leopold SR, Goering RV, Witten A, Harmsen D, Mellmann A. Bacterial Whole-Genome Sequencing Revisited: Portable, Scalable, and Standardized Analysis for Typing and Detection of Virulence and Antibiotic Resistance Genes. J Clin Microbiol. 2014 Jul;52(7):2365–70.
37. Marraffini LA, Sontheimer EJ. CRISPR Interference Limits Horizontal Gene Transfer in Staphylococci by Targeting DNA. Science. 2008 Dec 19;322(5909):1843–5.
38. Corvaglia AR, François P, Hernandez D, Perron K, Linder P, Schrenzel J. A type III-like restriction endonuclease functions as a major barrier to horizontal gene transfer in clinical Staphylococcus aureus strains. Proc Natl Acad Sci. 2010 Jun 29;107(26):11954–8.
39. Monk IR, Shah IM, Xu M, Tan MW, Foster TJ. Transforming the untransformable: application of direct transformation to manipulate genetically Staphylococcus aureus and Staphylococcus epidermidis. mBio. 2012;3(2):e00277-11.
40. Moller AG, Petit RA, Read TD. Species-Scale Genomic Analysis of Staphylococcus aureus Genes Influencing Phage Host Range and Their Relationships to Virulence and Antibiotic Resistance Genes. mSystems. 2022 Jan 18;7(1):e01083-21.
41. Feil EJ, Cooper JE, Grundmann H, Robinson DA, Enright MC, Berendt T, et al. How clonal is Staphylococcus aureus? J Bacteriol. 2003 Jun;185(11):3307–16.
42. 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 Dec;187(24):8312–21.
43. Wright JS, Traber KE, Corrigan R, Benson SA, Musser JM, Novick RP. The agr radiation: an early event in the evolution of staphylococci. J Bacteriol. 2005 Aug;187(16):5585–94.
44. Horesh G, Taylor-Brown A, McGimpsey S, Lassalle F, Corander J, Heinz E, et al. Different evolutionary trends form the twilight zone of the bacterial pan-genome. Microb Genomics. 2021;7(9):000670.
45. Andreani NA, Hesse E, Vos M. Prokaryote genome fluidity is dependent on effective population size. ISME J. 2017 Jul;11(7):1719–21.
46. McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. Nat Microbiol. 2017 Mar 28;2(4):1–5.
47. Vos M, Eyre-Walker A. Are pangenomes adaptive or not? Nat Microbiol. 2017 Dec;2(12):1576–1576.
48. Kuo CH, Ochman H. Deletional Bias across the Three Domains of Life. Genome Biol

Evol. 2009 Jan 1;1:145–52.

49. Sela I, Wolf YI, Koonin EV. Theory of prokaryotic genome evolution. Proc Natl Acad Sci U S A. 2016 Oct 11;113(41):11399–407.
50. Karcagi I, Draskovits G, Umenhoffer K, Fekete G, Kovács K, Méhi O, et al. Indispensability of Horizontally Transferred Genes and Its Impact on Bacterial Genome Streamlining. Mol Biol Evol. 2016 May 1;33(5):1257–69.
51. Nakamura Y, Itoh T, Matsuda H, Gojobori T. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet. 2004 Jul;36(7):760–6.
52. Niehus R, Mitri S, Fletcher AG, Foster KR. Migration and horizontal gene transfer divide microbial genomes into multiple niches. Nat Commun. 2015 Nov 23;6:8924.
53. McInerney JO. Prokaryotic Pangenomes Act as Evolving Ecosystems. Mol Biol Evol. 2023 Jan 1;40(1):msac232.
54. Dyzenhaus S, Sullivan MJ, Alburquerque B, Boff D, Guchte A van de, Chung M, et al. MRSA lineage USA300 isolated from bloodstream infections exhibit altered virulence regulation. Cell Host Microbe. 2023 Feb 8;31(2):228-242.e8.
55. Douglas EJA, Palk N, Brignoli T, Altwiley D, Boura M, Laabei M, et al. Extensive re-modelling of the cell wall during the development of Staphylococcus aureus bacteraemia [Internet]. bioRxiv; 2023 [cited 2023 Apr 25]. p. 2023.02.23.529713. Available from: https://www.biorxiv.org/content/10.1101/2023.02.23.529713v2
56. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017 Apr;35(4):316–9.
57. 262588213843476. bactopia-wrapper-nextflow [Internet]. Gist. [cited 2023 Apr 23]. Available from: https://gist.github.com/rpetit3/fe1f5428be135852ec90bfb63aa32c93
58. Souvorov A, Agarwala R, Lipman DJ. SKESA: strategic k-mer extension for scrupulous assemblies. Genome Biol. 2018 Oct 4;19(1):153.
59. Seemann T. Snippy [Internet]. 2023 [cited 2023 Apr 2]. Available from: https://github.com/tseemann/snippy
60. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013 Apr 15;29(8):1072–5.
61. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015 May 14;gr.186072.114.
62. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016 Jun 20;17(1):132.
63. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinforma Oxf Engl. 2011 Nov 1;27(21):2987–93.
64. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microb Genomics. 2021;7(11):000685.
65. Bayliss SC, Thorpe HA, Coyle NM, Sheppard SK, Feil EJ. PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. GigaScience.

2019 Oct 1;8(10):giz119.

66. Buchfink B, Reuter K, Drost HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021 Apr;18(4):366–8.
67. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017;8(1):28–36.
68. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol. 2015 Jan 1;32(1):268–74.
69. R: The R Project for Statistical Computing [Internet]. [cited 2023 Apr 3]. Available from: https://www.r-project.org/
70. Kassambara A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests [Internet]. 2023 [cited 2023 Apr 3]. Available from: https://CRAN.R-project.org/package=rstatix
71. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; 2016. Available from: https://ggplot2.tidyverse.org
72. jgraph/drawio [Internet]. JGraph; 2023 [cited 2023 Mar 30]. Available from: https://github.com/jgraph/drawio
73. Bogart S. SankeyMATIC [Internet]. 2023 [cited 2023 Apr 25]. Available from: https://github.com/nowthis/sankeymatic

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

Vishnu Raghuram[1], Ashley M. Alexander[2], Hui Qi Loo[3], Robert A. Petit III[4], Joanna B. Goldberg[5], Timothy D. Read[4,*]

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

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

[3] Department of Biology, Emory University, Atlanta, Georgia, USA

[4] Division of Infectious Diseases, Department of Medicine, Emory University, Atlanta, Georgia, USA

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

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

Published in *Microbiology Spectrum*, Jan 19th 2022

Author contributions

VR worked on study conceptualization and design, data curation, analysis, methodology, software, validation, visualisation, writing, editing and answering reviewer comments.

AMA and HQL performed the analysis and the writing of the results and methods for **Fig S2** and **Table S1**.

RAP aided in troubleshooting and publishing the software tool AgrVATE (https://github.com/VishnuRaghuram94/AgrVATE)

JBG and TDR helped with study conceptualization and design, supervision, funding, resources, writing, editing and answering reviewer comments.

All authors read and provided constructive comments on the manuscript.

Abstract

Staphylococcus aureus is a prominent nosocomial pathogen that causes several life-threatening diseases such as pneumonia and bacteremia. *S. aureus* modulates expression of its arsenal of virulence factors through sensing and integrating responses to environmental signals. The *agr* (accessory gene regulator) quorum sensing (QS) system is a major regulator of virulence phenotypes in *S. aureus*. There are four agr specificity groups each with a different autoinducer peptide sequence (encoded by the *agrD* gene). Though *agr* is critical for expression of many toxins, paradoxically, *S. aureus* strains often have non-functional *agr* activity due to loss-of-function mutations in the four-gene *agr* operon. To understand patterns in *agr* variability across *S. aureus*, we undertook a species-wide genomic investigation. We developed a software tool (AgrVATE; https://github.com/VishnuRaghuram94/AgrVATE) for typing and detecting frameshift mutations in the *agr* operon. In an analysis of over 40,000 *S. aureus* genomes, we showed close association between agr type and *S. aureus* clonal complex. We also found strong linkage between *agrBDC* alleles (encoding the peptidase, the autoinducing peptide itself, and the peptide sensor respectively) but not *agrA* (encoding the response regulator). More than five percent of genomes were found to have frameshift mutations in the agr operon. While 52% of these frameshifts occur only once in the entire species, we observed cases where the recurring mutations evolve convergently across different clonal lineages with no evidence of long-term phylogenetic transmission, suggesting that strains with *agr* frameshifts were evolutionarily short lived. Overall, genomic analysis of *agr* operon suggests evolution

through multiple processes with functional consequences that are not fully understood.

Importance

Staphylococcus aureus is a globally pervasive pathogen that produces a plethora of toxic molecules that can harm host immune cells. Production of these toxins is mainly controlled by an active *agr* quorum sensing system, which senses and responds to bacterial cell density. However, there are many reports of *S. aureus* strains with genetic changes leading to impaired *agr* activity, often found during chronic bloodstream infections, and may be associated with increased disease severity. We developed an open-source software called AgrVATE to type *agr* systems and identify mutations. We used AgrVATE for a species-wide genomic survey of *S. aureus*, finding that more than 5 % of strains in the public database had non-functional *agr* systems. We also provide new insights into the evolution of these genetic mutations in the *agr* system. Overall, this study contributes to our understanding of a common but 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 virulence factors seemingly lead to reduced disease severity and decreased host cell damage (22-24). While this appears paradoxical, there are several, non-exclusive, speculations as to why strains may have evolved non-functional *agr* systems (25). QS systems are 'public goods' that promote evolutionary cheating strategies (26-29). Impaired *agr* activity may be related to intra/inter-species competition between strains with different *agr* groups, which tend to suppress each other with no obvious effect on colonisation ability (30-34). Defective agr function may provide pleiotropically selected phenotypes, such as decreased susceptibility to vancomycin (35, 36). *agr*$^-$ strains may have traded their ability to produce energetically expensive virulence factors but may be limited in their ability to compete with the host immune system (37-39). This attenuated toxicity appears to be inconsequential or sometimes even beneficial in chronic diseases such as cystic fibrosis, bacteremia and osteomyelitis, where *agr*$^-$ *S. aureus* may show increased persistence and higher mortality rates compared to agr$^+$ strains (20, 21, 24, 31, 40). Phase variable *agr*$^-$ mutants may exist in environments that fluctuate between selection for toxicity and persistence (41). However, reduced/lack of expression of agr mediated virulence may be detrimental to colonisation in some other circumstances (42).

It has been proposed that *agr*$^-$ strains have sacrificed long-term viability through successful between-host transmissions for increased adaptation to specific environmental niches within the host (40, 43-45). However, studies to date have typically focused on small numbers of strains in limited clinical settings. It is important to assess whether specific patterns of variation in agr can be observed from a genome-wide scale spanning multiple clonal lineages of S. aureus, as it is crucial for understanding the mechanisms driving virulence regulation. The amount of publicly available genome sequences of S. aureus has grown rapidly over the past decade (46), offering the opportunity to examine the evolution and diversity of critical virulence determinants in S. aureus from a species-wide standpoint.

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 (46). 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 (42, 47-49). AgrVATE was designed to be a fast, standardised workflow for assigning *agr* groups that is conveniently installable through the Conda package manager (50). 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 (51) and performs variant calling using Snippy v4.6 (52) 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 91 *S. aureus* genomes that had been typed for haemolysis activity, a phenotype that is generally associated with functional *agr* systems (16). These 91 genomes included clinical samples taken from cystic fibrosis patients from the Emory Cystic Fibrosis Center (53). We found 15 genomes had putative *agr* null mutations, 14 of which tested negative for sheep blood haemolysis (Table S1). The one putative null that displayed haemolysis on sheep blood agar had a frameshift mutation in the C terminal end of *agrC.* The haemolysis phenotype of this strain was relatively weak (CFBR_17 – **Fig S2**). We also observed 10 samples that were negative for haemolysis despite having no frameshift mutations in the *agr* operon (Table S1), suggesting that other genetic factors reduced haemolysis activity and *agr* frameshifts are not the sole indicator. In patient CFBR311, AgrVATE found two different *agr* groups (group-1 and group-2) from the genome sequence of the *S. aureus* population isolated from CF sputum (CFBR_EB_Sa110 – Table S1), showing that this patient is colonised by *S. aureus* of heterologous *agr* groups. Genome sequences of 8 individual colonies (CFBR_EB_Sa111 to CFBR_EB_Sa118 – Table S1) from this patient showed an *agr* group-2 majority (6 out of 8) and an *agr* group-1 minority (2 out of 8), thereby validating the AgrVATE prediction. This illustrated an advantage of using a kmer-based approach, as AgrVATE could identify the presence of multiple *agr* groups. Each *agr* group assignment was also scored, thereby indicating the proportions of each *agr* group if more than one is present. Collectively, these findings demonstrate a potential use-case for AgrVATE in clinical settings, where we identified CF patient isolates having *agr* mutations and showed one instance where AgrVATE identified a patient colonised by *S. aureus* of heterologous *agr* groups.

agr type distribution in the Staphopia database

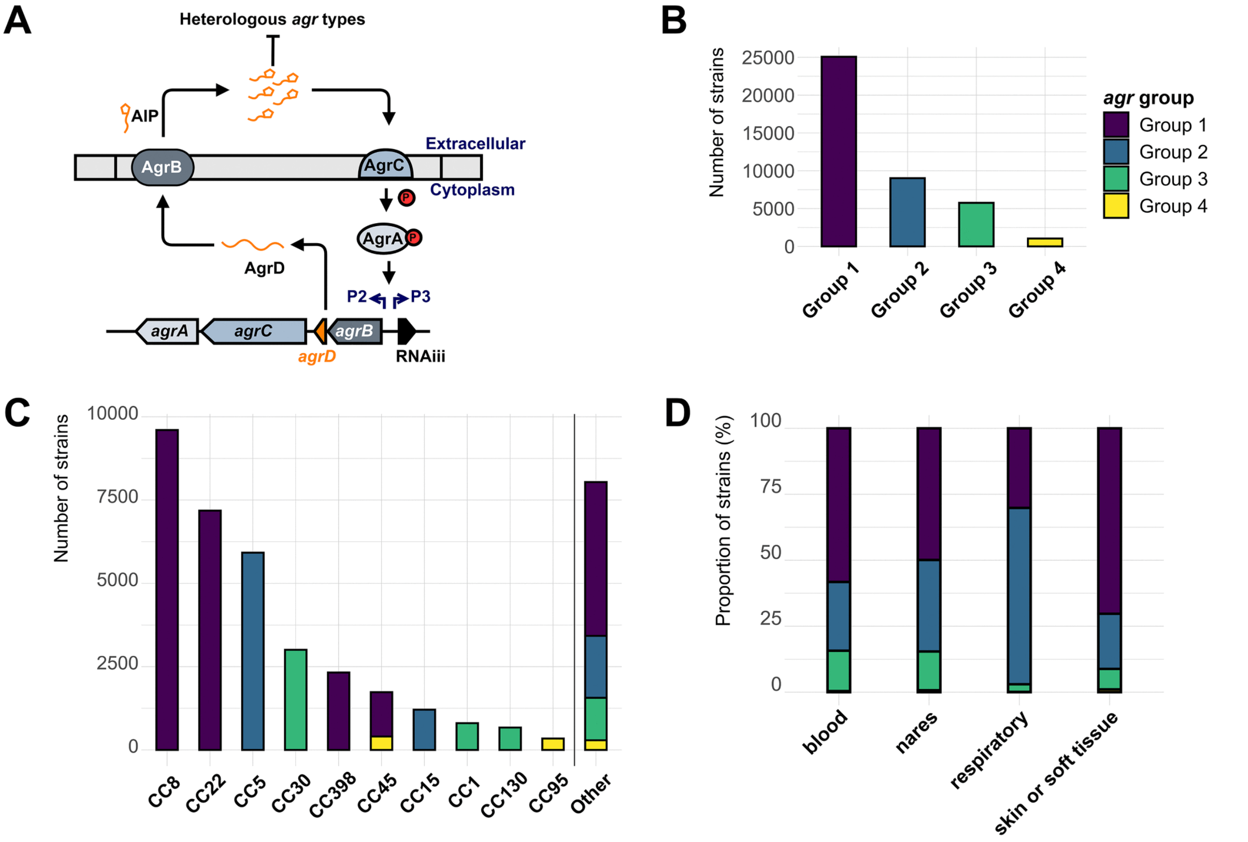

Fig 1: Distribution of *agr* groups across 40,890 *S. aureus* genomes from the Staphopia database.

AgrVATE was used to assign the *agr* groups and genomes with unknown *agr* groups were filtered out. (A): A schematic depiction of the *agr* operon showing two divergent promoters (P2 and P3) driving *agrBDCA* and small RNA RNAiii. (B): Frequency of each *agr* group in the Staphopia database. (C): Frequency of *agr* groups across the major clonal complexes (CC) of *S. aureus*. (D): Relative proportions of *agr* groups from *S. aureus* isolated from different body sites in percentage.

AgrVATE identified the *agr* groups for 42,491 genomes out of 42,999 in the Staphopia database, with 25,539 group-1 (60.10%), 9639 group-2 (22.68%), 6224 group-3 (14.65%) and 1,089 (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 (54, 55) (**Fig 1C**). 1,601 out of the 42,491 genomes showed the presence of more than one *agr* group (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, Chi-squared $p < 0.01$).

The remaining 508 genomes were reported to be low confidence/unknown *agr* group calls by AgrVATE. We used a mash sketch (56) 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 the Staphopia database 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 genomes in the Staphopia database 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). Across the 5143 unique *agr* operon sequences, we observed an average within-*agr* group SNP distance of 15 and between-*agr* group SNP distance of 167.

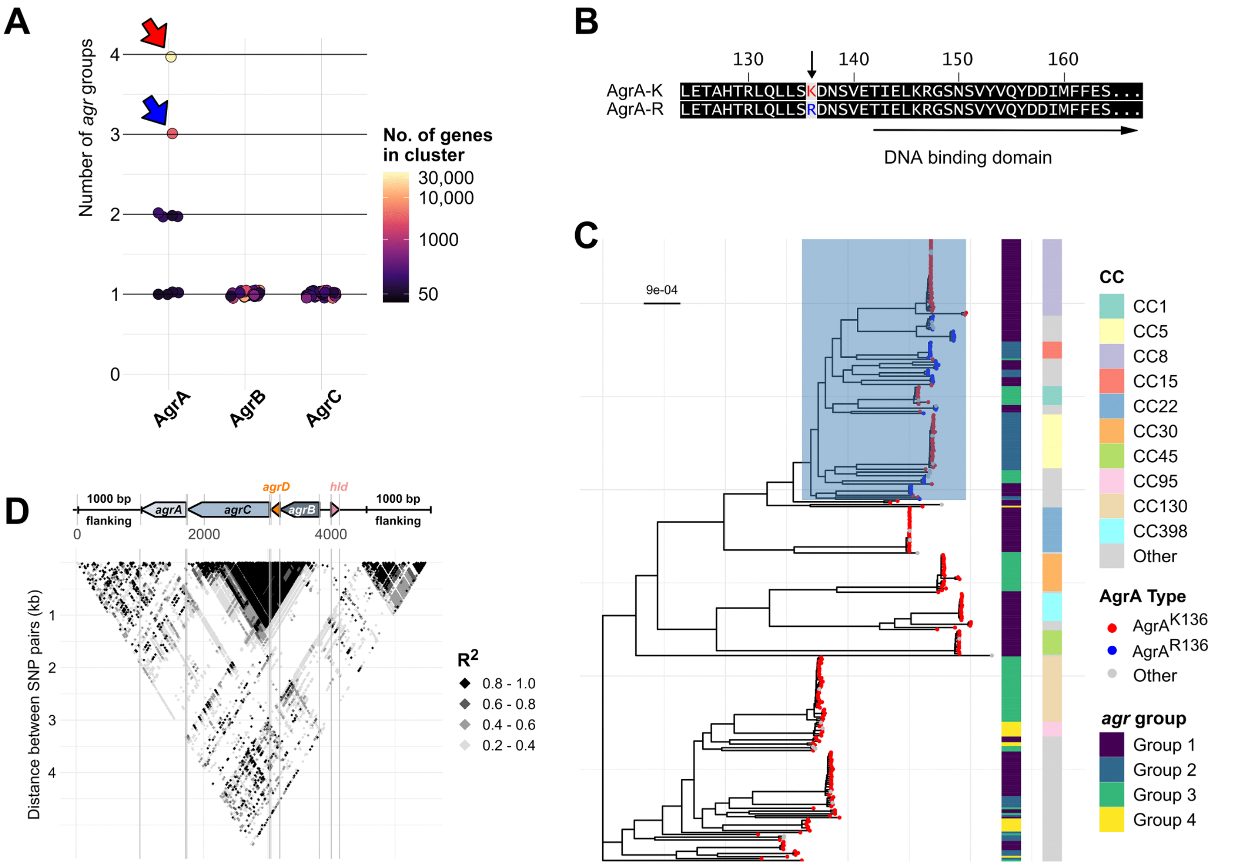

Fig 2: AgrA evolves independently of *agr* group with only two major amino acid sequence configurations across *S. aureus*.

(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^{K136}) and minor cluster (AgrA^{R136}) 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 334 *S. aureus* strains with each tip representing a unique ST. Tip colours represent the AgrA alleles and the corresponding heatmaps show the *agr* group and clonal complex of each tip. Scale bar indicates number of substitutions per site. All tips representing the AgrA^{R136} allele are confined to the clade highlighted in blue. (D) Linkage disequilibrium (LD) block plot of the *agr* operon and 1000bp flanking regions. Each point on the block indicates R^2 values of LD calculated by plink for a given pair of SNPs. The y axis indicates distance between SNP pairs.

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 the Staphopia database, 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, differing by an average of 179 SNPs. 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 (57) from a curated set of 334 genomes from the Staphopia database each representing a unique ST, called Non-Redundant Diversity (NRD) set

(46), we found that *S. aureus* can be broadly divided into two clades or sub-species: all genomes harbouring AgrA^{R136} are limited to one clade of *S. aureus* (**Fig 2C**, blue tips, blue highlighted clade). 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 334 genomes using Snippy (52). We then measured Linkage Disequilibrium (LD) between these SNPs by calculating the Pearson coefficient (R^2) using plink (58). 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

In 39,174 *agr* operons, there were 405 sites that had a frameshift mutation in at least one genome. 52% of frameshifts (210 sites) occurred in only one genome, but a small minority of sites were more frequently mutated - 24 sites had frameshifts in at least 10 genomes and 5 sites had frameshifts in more than 100 genomes. At least one frameshift mutation was found in each *agr* gene (**Fig 3A**). We observed a total of 2,997 *agr* operons with at least one frameshift mutation, only 91 of which had two. We did not observe any *agr* operons with more than two frameshift mutations. The rate of mutations in the *agr* operon follows the expected Poisson distribution with a mean of 0.0765 (**Fig S4**). The *agrC* gene had acquired the greatest number of different frameshift mutations, including truncation mutations in CC5, CC8, CC22 and CC30 (865 genomes) (**Fig 3A, B**). The most frequent frameshift was the insertion of an adenine in the terminal end of *agrA,* occurring in 561 genomes (35 unique *agr* operons) across multiple CCs (p.Ile238fs – **Fig 3B**). This mutation has previously been investigated by Traber & Novick and was found to cause delayed *agr* activation (59). Another mutation toward the end of the gene occurring in a polyA tract found in *agrB*, occurred in 114 genomes (34 unique *agr* operons), mainly CC15 and CC5 (p.Phe201fs – **Fig 3B**). A frameshift that resulted in loss of the start-codon in *agrC* was observed in 86 genomes (25 unique *agr* operons) of CC8, CC30, CC22 and some other rarer CCs (p.Val? – **Fig 3B**). In relatively low frequency (69 genomes), we observed complex mutations such as collapsed repeats, tandem duplications and large (> 30bp) in-frame and out-of-frame indels. However, such mutations were mostly singular sporadic events, and no mutation was recurrent in more than three different genomes. Two of these cases were an Insertion Sequence (IS) element insertion in the *agr* operon. One being a 1,326bp insertion of an IS256 family transposon sequence found commonly in *Staphylococcus* species (60), and the other a 1057bp insertion of an IS1252 transposon sequence found in *Enterococcus* species (61).

Frameshifts in the delta-toxin gene (*hld*), present within the RNAiii transcript (**Fig 1A**) were rare, with only 4 genomes having one of two mutations - a deletion at position 76 leading to a frameshift, or a G to A substitution in position 44 leading to a premature stop. We found 3943 indels in RNAiii, 3931 of which were single nucleotide indels. The most common mutation occurring 2025 times was the insertion of a T at position 406 of RNAiii, found exclusively in *agr* group-3 genomes, suggesting that this might be a common variant. It is unknown whether these single nucleotide indels have a functional impact on RNAiii. On the other hand, we also found 8 genomes with large >30 bp indels in RNAiii which may affect function. Namely, two *agr* group-1 genomes with a 42bp deletion at position 391, four *agr* group-2 genomes with a 41bp insertion at position 2, and one *agr* group-3 strain with a 31bp insertion at position 9.

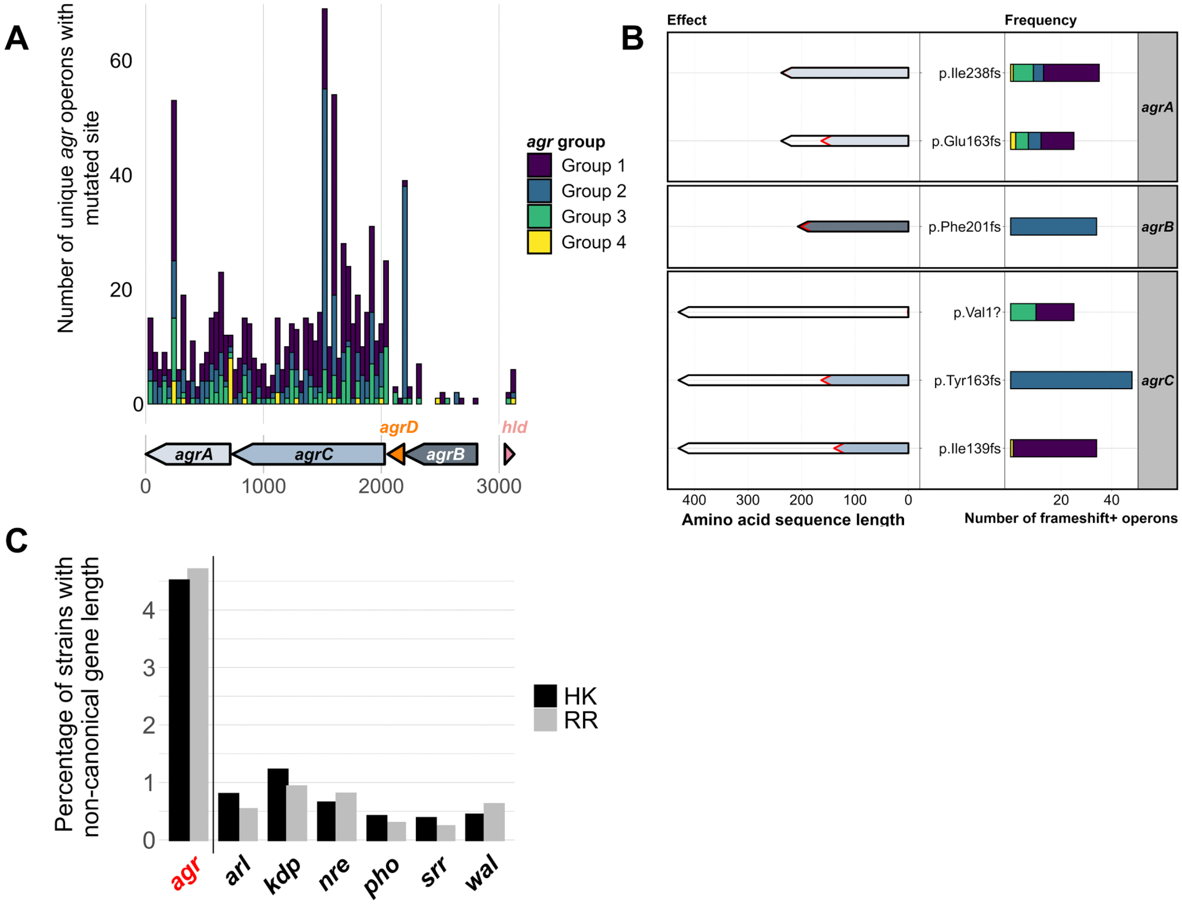

Fig 3: Presence of putative non-functional variants of the *agr* operon.

(A) Frequency of frameshift mutations in coding regions of unique *agr* operon sequences across the Staphopia database. 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) Normalised 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 the Staphopia database and commonly occurring gene lengths (> 5000 genomes) were excluded. The remaining strains were considered to have non-canonical gene lengths.

We compared the number of genomes with indels in the *agr* operon to other two-component regulatory systems (TCS) to determine if the frequency of potentially deleterious mutations in the *agr* operon was significantly different. This was done by extracting the histidine kinase (HK) and response regulator (RR) genes of TCS *arlRS, kdpDE, nreBC, phoPR, srrAB* and *walKR* from the Staphopia database and calculating the number of genomes with HK or RR genes with non-canonical gene lengths. The length of the reference gene for each HK and RR was considered the canonical length and the length of each annotated gene in the Staphopia database best matching the reference gene was identified (BLASTn). At least 42,500 hits for each HK or RR were extracted out of 42,999 genomes in the Staphopia database. Hits for each HK and RR with length not equal to their corresponding reference were considered non-canonical gene lengths due to indels. Frequently occurring alternate gene lengths (observed in > 5000 genomes) were considered common alleles and not mutated variants. We found that the *agr* TCS has a significantly greater number of variable gene lengths when compared to TCS *arl*, *kdp*, *nre*, *pho*, *srr* and *wal* ($p < 0.0001$, negative binomial regression). When normalised to 1kb, we found ~4.5% of all *agrC* (HK) and *agrA* (RR) genes analysed were of variable lengths. In contrast, only ≤1.5% of all other HKs and RRs analysed had variable gene lengths (**Fig 3C**). 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/colonisation 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 the Staphopia database 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 maximum likelihood trees using HomoplasyFinder (62). In short, a consistency index of one for a given site on an alignment indicates that the nucleotides at that site are consistent with phylogeny, and a consistency index of 0 indicates that the nucleotides at the site are homoplasious (evolved independent of phylogeny).

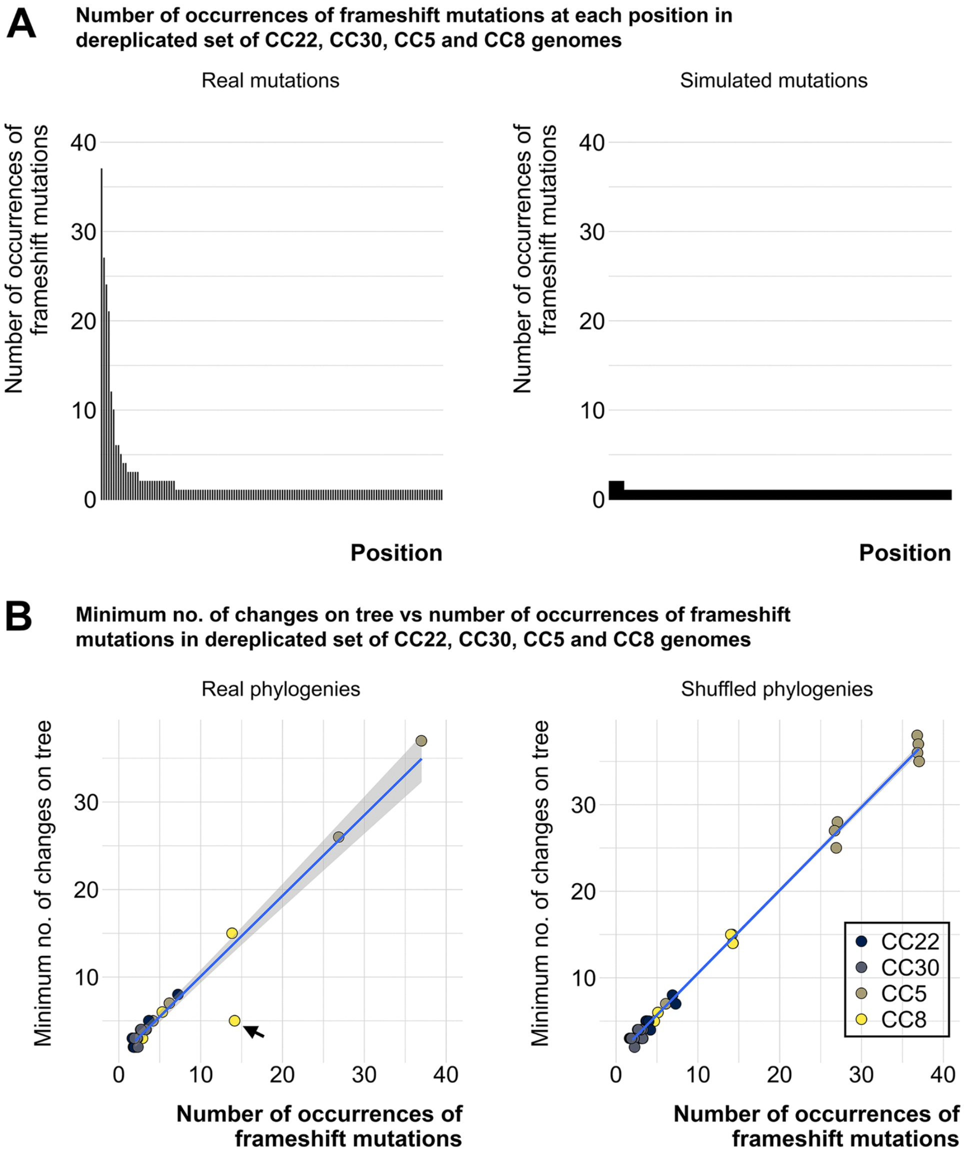

Fig 4: Identical *agr* mutations evolve independent of phylogeny across different clonal complexes.

(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) Minimum number of changes on tree vs number of occurrences of frameshift mutations. LEFT- 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 x axis represents the number of times the position has acquired a frameshift mutation. The consistency index and minimum number of changes on tree was measured at these sites for each CC from the respective phylogenetic tree (GTR+FO model, 1000 ultrafast bootstrap replicates with average bootstrap support of at least 71%) Blue line follows *y=x* distribution. The outlier CC8 point (black arrow) corresponds to a previously characterised *agrA* mutation that is not a true *agr* null (59). RIGHT – The consistency index and minimum number of changes on the tree were measured for phylogenies for the respective CCs where the tree tips were randomly shuffled. 100 shuffled trees were generated per CC.

We found a trend of decreasing consistency index with increasing frequency of each recurring frameshift. We observed an almost identical trend when the tips of the trees are shuffled to have the frameshift+ genomes at random positions on the tree, mimicking non-ancestral, independent acquisition of each frameshift. In other words, the minimum number of changes on the phylogeny equal the number of occurrences of frameshift mutations. As these mutations were being acquired repeatedly, independent of phylogeny, they are reflected in the number of changes on the tree (**Fig 4B**). The outlier mutation (**Fig 4B**, black arrow) that appeared to have transmitted to multiple isolates in CC8 was an *agrA* mutation at the 3' end found to cause delayed *agr* activation by Traber & Novick and therefore is not a true *agr* null (59). Overall, this showed that there is a preference for specific sites in the *agr* operon to acquire potentially null-inducing mutations and that identical mutations can occur independently across different *S. aureus* lineages.

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 (63). We also found, as previously reported (10, 64), 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 of long-term phylogenetic transmission (**Fig 4B**), suggesting selection or a generative mechanism for high frequency mutations. At least one of these frameshifts has been studied functionally (59), 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 2C**) (64). 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* (54, 55), 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 (65, 66). 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 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 colonisation by strains with downregulated *agr* expression in hospital settings (67). 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 (68, 69). This suggests that complete virulence capacity is not an absolute requirement for colonisation and transmission of *S. aureus* in the hospital environment (43). 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 (45, 67). This brings to light the possible evolutionary trade-off for strains that become *agr* defective. Though some short-term transmission may have occurred, our phylogenetic analyses show no evidence of stable lineages of putative *agr* null populations (**Fig 4B**). In addition, we also found that frequency of mutations in the *agr* TCS is enriched when compared to other TCS in *S. aureus* (**Fig 3C**). 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 (70) 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 hypothesised 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. 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 (71-75), 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 (76). 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 the Staphopia database (46) where a canonical AgrD protein was annotated by Prokka v1.14.6 (40,812 AgrDs) (77). We then extracted the *agr* operons from these 40,812 genomes by *in-silico* PCR using usearch v11.0.667_i86linux32 (51). To identify kmers unique to each *agr* group, DREME v5.1.1 (78) 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 (79) 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 the Staphopia database using usearch -search_pcr tool (51) with the following primers – 5’aaaaaaggccgcgagcttgggaggggctca’3 & 5’ttatattttttaacgtttctcaccgatgc’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 (51) to obtain all possible unique *agr* operon configurations. This unique set of operons was annotated using Prokka v1.14.6 (77), *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 (52). 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 database non-redundant diversity (NRD) 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 (77), resulting in 355 genomes. These 355 genomes were further filtered to include only genomes where the *agr* group prediction was unambiguous, leading to 334 genomes. A core genome alignment was constructed for these 334 using parsnp v1.5.3 (80) and this alignment was used to build a maximum likelihood phylogeny using IQ-TREE v1.6.12 (57) with the GTR+FO model and 1000 ultrafast bootstrap replicates using *S. argenteus* as the outgroup (GenBank accession: AP018562.1). The outgroup was then removed, and the tree was reconstructed with the tip closest to the outgroup (ST93) as the root. The resulting tree was then plotted using the R package ggtree (81).

The genomic region comprising the *agr* operon and 1000bp on each side was extracted from the initial filtered NRD set of 355 genomes and this region was aligned using snippy-core v4.6.0 (52). The full alignment (core.full.aln) file was then converted to a vcf file using snp-sites v2.5.1 (82), and this vcf file was used to calculate Pearson's coefficient for Linkage Disequilibrium (LD) using plink v1.90b6.21 (options: --r2 inter-chr) (58). 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 (83)). The length of the reference gene for each HK and RR was considered the canonical length and the length of each annotated gene in the Staphopia database 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 1kb normalised 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 (84) and multiple comparisons were performed using the emmeans R package (85).

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

R package caret (86) 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 the Staphopia database 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 database 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 (56) distance threshold of 0.0005 and a representative for each cluster was chosen at random. This Mash distance threshold was chosen empirically based on a comparison of pairwise Mash distances and pairwise SNP distances within the Staphopia database NRD set. Mash distances < 0.0005 represent a median SNP distance of 47 with a maximum of 282 (**Fig S5**). SNP distances were calculated from parsnp v1.5.3 (80) core genome alignments using snp-dists v0.7.0 (87). 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 parsnp v1.5.3 (80) and maximum likelihood phylogenetic trees were constructed using IQ-TREE v1.6.12 (57) 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 (81) and the tip labels were randomised 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 91 CF strains analysed in this study are associated with a previous publication and can be found in the accession PRJNA480016 (88). The remaining 27 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 10µL or 100µL 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, 4-8 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 30gL^{-1} Trypticase Soy Broth, 15gL^{-1} agar, and 70gL^{-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 (88).

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, 20μL of freshly prepared 10mgmL^{-1} lysozyme, and 100μL of 5mgmL^{-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 (89). Raw sequence files were then fed into the Bactopia analysis pipeline version 1.4.10 (90). 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, supplemental information and datasets for generating the figures in this study can be found in https://github.com/VishnuRaghuram94/AgrVATE . CF isolate genome sequences used in this study can be found under BioProject accessions PRJNA480016 and PRJNA742745. The accessions, *agr* groups, sequence type and clonal complex, and frameshift information for 40,890 *S. aureus* genomes used in this study can be found in supplemental dataset S1.

Supplementals

Supplemental tables and figures for this chapter can be found in the manuscript https://journals.asm.org/doi/10.1128/spectrum.01334-21

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, Cristian Crisan, Katrina Hofstetter and Brooke Talbot for providing constructive comments on the manuscript.

VR and TDR were supported by grant # AI139188 from the National Institutes of

Health (NIH). AMA 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 JBG.

References

1. Kourtis AP, Hatfield K, Baggs J, Mu Y, See I, Epson 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. Tuchscherr L, Loffler B. Staphylococcus aureus dynamically adapts global regulators and virulence factor expression in the course from acute to chronic infection. Curr Genet. 2016;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 psmalpha 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:

29456969; PMCID: PMC5801681.
22. Cheung AL, Eberhardt KJ, Chung E, Yeaman MR, Sullam PM, Ramos M, Bayer AS. Diminished virulence of a sar-/agr- mutant of Staphylococcus aureus in the rabbit model of endocarditis. J Clin Invest. 1994;94(5):1815-22. Epub 1994/11/01. doi: 10.1172/JCI117530. PubMed PMID: 7962526; PMCID: PMC294579.
23. Gillaspy AF, Hickmon SG, Skinner RA, Thomas JR, Nelson CL, Smeltzer MS. Role of the accessory gene regulator (agr) in pathogenesis of staphylococcal osteomyelitis. Infect Immun. 1995;63(9):3373-80. Epub 1995/09/01. doi: 10.1128/IAI.63.9.3373-3380.1995. PubMed PMID: 7642265; PMCID: PMC173464.
24. Das S, Lindemann C, Young BC, Muller J, Osterreich B, Ternette N, Winkler AC, Paprotka K, Reinhardt R, Forstner KU, Allen E, Flaxman A, Yamaguchi Y, Rollier CS, van Diemen P, Blattner S, Remmele CW, Selle M, Dittrich M, Muller T, Vogel J, Ohlsen K, Crook DW, Massey R, Wilson DJ, Rudel T, Wyllie DH, Fraunholz MJ. Natural mutations in a Staphylococcus aureus virulence regulator attenuate cytotoxicity but permit bacteremia and abscess formation. Proc Natl Acad Sci U S A. 2016;113(22):E3101-10. Epub 2016/05/18. doi: 10.1073/pnas.1520255113. PubMed PMID: 27185949; PMCID: PMC4896717.
25. Shopsin B, Copin R. Staphylococcus aureus Adaptation During Infection. In: Fong IW, Shlaes D, Drlica K, editors. Antimicrobial Resistance in the 21st Century. Cham: Springer International Publishing; 2018. p. 431-59.
26. Eldar A. Social conflict drives the evolutionary divergence of quorum sensing. Proc Natl Acad Sci U S A. 2011;108(33):13635-40. Epub 2011/08/03. doi: 10.1073/pnas.1102923108. PubMed PMID: 21807995; PMCID: PMC3158151.
27. Diard M, Garcia V, Maier L, Remus-Emsermann MN, Regoes RR, Ackermann M, Hardt WD. Stabilization of cooperative virulence by the expression of an avirulent phenotype. Nature. 2013;494(7437):353-6. Epub 2013/02/22. doi: 10.1038/nature11913. PubMed PMID: 23426324.
28. Chuang JS, Rivoire O, Leibler S. Simpson's paradox in a synthetic microbial system. Science. 2009;323(5911):272-5. Epub 2009/01/10. doi: 10.1126/science.1166739. PubMed PMID: 19131632.
29. Czaran T, Hoekstra RF. Microbial communication, cooperation and cheating: quorum sensing drives the evolution of cooperation in bacteria. PLoS One. 2009;4(8):e6655. Epub 2009/08/18. doi: 10.1371/journal.pone.0006655. PubMed PMID: 19684853; PMCID: PMC2722019.
30. Kahl BC, Becker K, Friedrich AW, Clasen J, Sinha B, Von Eiff C, Peters G. agr-dependent bacterial interference has no impact on long-term colonization of Staphylococcus aureus during persistent airway infection of cystic fibrosis patients. J Clin Microbiol. 2003;41(11):5199-201. Epub 2003/11/08. doi: 10.1128/jcm.41.11.5199-5201.2003. PubMed PMID: 14605162; PMCID: PMC262511.
31. Goerke C, Kummel M, Dietz K, Wolz C. Evaluation of intraspecies interference due to agr polymorphism in Staphylococcus aureus during infection and colonization. J Infect Dis. 2003;188(2):250-6. Epub 2003/07/11. doi: 10.1086/376450. PubMed PMID: 12854080.
32. Ji G, Beavis R, Novick RP. Bacterial interference caused by autoinducing peptide 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 modulin 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. Gor V, Takemura AJ, Nishitani M, Higashide M, Medrano Romero V, Ohniwa RL, Morikawa K. Finding of Agr Phase Variants in Staphylococcus aureus. mBio. 2019;10(4). Epub 2019/08/08. doi: 10.1128/mBio.00796-19. PubMed PMID: 31387900; PMCID: PMC6686034.
42. 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 Transl Med. 2020;12(551). Epub 2020/07/10. doi: 10.1126/scitranslmed.aay4068. PubMed PMID: 32641488; PMCID: PMC7426015.

43. Shopsin B, Drlica-Wagner A, Mathema B, Adhikari RP, Kreiswirth BN, Novick RP. Prevalence of agr dysfunction among colonizing Staphylococcus aureus strains. J Infect Dis. 2008;198(8):1171-4. Epub 2008/08/30. doi: 10.1086/592051. PubMed PMID: 18752431.

44. Sakoulas G, Moise PA, Rybak MJ. Accessory gene regulator dysfunction: an advantage for Staphylococcus aureus in health-care settings? J Infect Dis. 2009;199(10):1558-9. Epub 2009/04/28. doi: 10.1086/598607. PubMed PMID: 19392634.

45. Shopsin B, Eaton C, Wasserman GA, Mathema B, Adhikari RP, Agolory S, Altman DR, Holzman RS, Kreiswirth BN, Novick RP. Mutations in agr do not persist in natural populations of methicillin-resistant Staphylococcus aureus. J Infect Dis. 2010;202(10):1593-9. Epub 2010/10/15. doi: 10.1086/656915. PubMed PMID: 20942648.

46. Petit RA, 3rd, Read TD. Staphylococcus aureus viewed from the perspective of 40,000+ genomes. PeerJ. 2018;6:e5261. Epub 2018/07/18. doi: 10.7717/peerj.5261. PubMed PMID: 30013858; PMCID: PMC6046195.

47. Choudhary KS, Mih N, Monk J, Kavvas E, Yurkovich JT, Sakoulas G, Palsson BO. The Staphylococcus aureus Two-Component System AgrAC Displays Four Distinct Genomic Arrangements That Delineate Genomic Virulence Factor Signatures. Front Microbiol. 2018;9:1082. Epub 2018/06/12. doi: 10.3389/fmicb.2018.01082. PubMed PMID: 29887846; PMCID: PMC5981134.

48. Shopsin B, Mathema B, Alcabes P, Said-Salim B, Lina G, Matsuka A, Martinez J, Kreiswirth BN. Prevalence of agr specificity groups among Staphylococcus aureus strains colonizing children and their guardians. J Clin Microbiol. 2003;41(1):456-9. Epub 2003/01/09. doi: 10.1128/jcm.41.1.456-459.2003. PubMed PMID: 12517893; PMCID: PMC149583.

49. Francois P, Koessler T, Huyghe A, Harbarth S, Bento M, Lew D, Etienne J, Pittet D, Schrenzel J. Rapid Staphylococcus aureus agr type determination by a novel multiplex real-time quantitative PCR assay. J Clin Microbiol. 2006;44(5):1892-5. Epub 2006/05/05. doi: 10.1128/JCM.44.5.1892-1895.2006. PubMed PMID: 16672433; PMCID: PMC1479209.

50. Gruning B, Dale R, Sjodin A, Chapman BA, Rowe J, Tomkins-Tinch CH, Valieris R, Koster J, Bioconda T. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods. 2018;15(7):475-6. Epub 2018/07/04. doi: 10.1038/s41592-018-0046-7. PubMed PMID: 29967506.

51. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460-1. Epub 2010/08/17. doi: 10.1093/bioinformatics/btq461. PubMed PMID: 20709691.

52. Seemann T. snippy: fast bacterial variant calling from NGS reads. 4.6 ed. https://github.com/tseemann/snippy: Github; 2015.

53. Bernardy EE, Petit RA, 3rd, Raghuram V, Alexander AM, Read TD, Goldberg JB.

Genotypic and Phenotypic Diversity of Staphylococcus aureus Isolates from Cystic Fibrosis Patient Lung Infections and Their Interactions with Pseudomonas aeruginosa. mBio. 2020;11(3). Epub 2020/06/25. doi: 10.1128/mBio.00735-20. PubMed PMID: 32576671; PMCID: PMC7315118.

54. 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.

55. 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.

56. 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.

57. 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.

58. 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.

59. 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.

60. 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.

61. 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.

62. 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.

63. 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.

64. 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.

65. 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.

66. 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.

67. 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.

68. 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.

69. 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.

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

71. 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.

72. 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.

73. 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.

74. Bloemendaal AL, Brouwer EC, Fluit AC. Methicillin resistance transfer from Staphylocccus 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.
75. Berglund C, Soderquist B. The origin of a methicillin-resistant Staphylococcus aureus isolate at a neonatal ward in Sweden-possible horizontal transfer of a staphylococcal cassette chromosome mec between methicillin-resistant Staphylococcus haemolyticus and Staphylococcus aureus. Clin Microbiol Infect. 2008;14(11):1048-56. Epub 2008/12/02. doi: 10.1111/j.1469-0691.2008.02090.x. PubMed PMID: 19040477.
76. Zhang DF, Zhi XY, Zhang J, Paoli GC, Cui Y, Shi C, Shi X. Preliminary comparative genomics revealed pathogenic potential and international spread of Staphylococcus argenteus. BMC Genomics. 2017;18(1):808. Epub 2017/10/24. doi: 10.1186/s12864-017-4149-9. PubMed PMID: 29058585; PMCID: PMC5651615.
77. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068-9. Epub 2014/03/20. doi: 10.1093/bioinformatics/btu153. PubMed PMID: 24642063.
78. Bailey TL. DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics. 2011;27(12):1653-9. Epub 2011/05/06. doi: 10.1093/bioinformatics/btr261. PubMed PMID: 21543442; PMCID: PMC3106199.
79. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421. Epub 2009/12/17. doi: 10.1186/1471-2105-10-421. PubMed PMID: 20003500; PMCID: PMC2803857.
80. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol. 2014;15(11):524. Epub 2014/11/21. doi: 10.1186/s13059-014-0524-x. PubMed PMID: 25410596; PMCID: PMC4262987.
81. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution. 2017;8(1):28-36. doi: https://doi.org/10.1111/2041-210X.12628.
82. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom. 2016;2(4):e000056. Epub 2017/03/30. doi: 10.1099/mgen.0.000056. PubMed PMID: 28348851; PMCID: PMC5320690.
83. Rapun-Araiz B, Haag AF, De Cesare V, Gil C, Dorado-Morales P, Penades JR, Lasa I. Systematic Reconstruction of the Complete Two-Component Sensorial Network in Staphylococcus aureus. mSystems. 2020;5(4). Epub 2020/08/21. doi: 10.1128/mSystems.00511-20. PubMed PMID: 32817385; PMCID: PMC7438023.
84. Ripley WNVBD. Modern Applied Statistics with S. Fourth ed: Springer.
85. Lenth RV. emmeans: Estimated Marginal Means, aka Least-Squares Means. https://cran.r-project.org/web/packages/emmeans/index.html: Github; 2021.
86. Kuhn M. Building Predictive Models in R Using the caret Package. 2008. 2008;28(5):26. Epub 2008-09-23. doi: 10.18637/jss.v028.i05.
87. Seemann T. snp-dists: Pairwise SNP distance matrix from a FASTA sequence alignment. Github: Github; 2017.
88. Bernardy EE, Petit RA, 3rd, Moller AG, Blumenthal JA, McAdam AJ, Priebe GP, Chande AT, Rishishwar L, Jordan IK, Read TD, Goldberg JB. Whole-Genome Sequences

of Staphylococcus aureus Isolates from Cystic Fibrosis Lung Infections. Microbiol Resour Announc. 2019;8(3). Epub 2019/01/29. doi: 10.1128/MRA.01564-18. PubMed PMID: 30687841; PMCID: PMC6346173.

89. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/2010.

90. 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.

Chapter IV - Comparison of genomic diversity between single and pooled *Staphylococcus aureus* colonies isolated from human colonisation cultures

Vishnu Raghuram[1], Jessica J. Gunoskey[3], Katrina S. Hofstetter[2], Natasia F. Jacko[3], Margot J. Shumaker[3], Yi-Juan Hu[4], Timothy D. Read[2,*], Michael Z. David[3,*]

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

[2] Division of Infectious Diseases, Department of Medicine, Emory University, Atlanta, Georgia, USA

[3] Division of Infectious Diseases, Department of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[4] Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia, USA

* Corresponding authors, Email address:

Timothy D. Read - tread@emory.edu

Michael Z. David - michdav@pennmedicine.upenn.edu

Author contributions

VR worked on study conceptualization and design, data curation, analysis, methodology, validation, visualisation, writing and editing.

JJG, NFJ and MJS worked on all the sample collection, culturing, storage, and sequencing preparation.

KSH performed data curation and processing using Bactopia for all genome sequences.

YH provided guidance on statistical analyses

TDR and MZD helped with study conceptualization and design, supervision, funding, resources, writing and editing.

All authors read and provided constructive comments on the manuscript.

Abstract

As pathogenic bacteria go through cycles of growth and adaptation within a host, the genetic makeup of the population changes. The most common approach to sampling microbial populations within an infected or colonised host is to sequence genomes from a single colony obtained from a culture plate. However, it is recognized that this method may not capture the complete genetic diversity in the population. An alternative is to sequencing a mixture containing multiple colonies ("pool-seq"), but this has the disadvantage that it is a non-homogeneous sample, making it difficult to perform specific experiments. We compared differences in measures of genetic diversity between eight single-colony isolates (singles) and pool-seq on a set of 2286 *S. aureus* culture samples. The samples were obtained by swabbing three body sites on human participants quarterly for a year, who initially presented with a methicillin-resistant *S. aureus* skin and soft-tissue infection (SSTI). We compared parameters such as sequence quality, contamination, allele frequency, nucleotide diversity and pangenome diversity in each pool to the corresponding singles. Comparing singles from the same culture plate, we found that 17% of pooled samples contained mixtures of sequence types (STs). We showed that pool-seq data alone could predict the presence of multi-ST populations with 95% accuracy. We also showed that pool-seq could be used to estimate the number of polymorphic sites in the population. Additionally, we found that the pool may contain clinically relevant genes such as antimicrobial resistance markers that may be missed when only examining singles. These results highlight the potential advantage of analysing

genome sequences of total populations obtained from clinical cultures rather than single colonies.

Data summary

Genomes sequenced for this study are available under accession PRJNA918392. Raw data and code for analysis are available at https://github.com/VishnuRaghuram94/GASP.

Importance

While pooled population sequencing has been employed to study within-host diversity, the differences in attainable information between single and pooled sequences are not clear. A direct comparison between single isolate and pooled population sequences can help devise optimal sampling strategies for clinical and within-host diversity studies. In this study, we attempt to answer the question of how many colonies obtained from a single patient is enough to obtain a representation of the total population diversity within the patient while keeping in mind time and labour costs. These findings have implications for using whole genome sequencing (WGS) in the clinical microbiology laboratory to identify and speciate pathogens and to determine their antimicrobial susceptibilities.

Introduction

Large-scale WGS of bacterial pathogen species offers hope for more accurate infectious disease surveillance and a better understanding of within-host evolution during human disease and asymptomatic colonisation (1–3). Typically, bacterial genomics-based surveillance studies sequence DNA isolated from single bacterial colonies cultured on selective media from each clinical sample tested (1,4–6). The single-colony bacterial culture can be further tested in the laboratory for phenotypes such as antibiotic resistance and toxicity.

However, individual colonies do not provide insight into the genetic diversity of the population of the species in the sample as there could be multiple strains present in the sample (7–9). Even if only one strain is present, there will be accumulated microdiversity between individual isolates that is roughly proportional to the duration between initial colonisation and time of sampling, assuming the absence of bottlenecks (5,10).

A few studies have undertaken sequencing multiple single colonies from clinical samples (9,11,12). This strategy can allow the comparison of phenotypes associated with intra-sample genetic variation and the construction of phylogenetic trees to trace the relationship between samples. However, costs rise linearly with each additional colony sequenced per sample, necessitating a cost-benefit analysis of how many colonies to sequence. Sequencing colonies from isolation plates of a single sample is an alternative approach (12–15). This is sometimes called "sweep sequencing", "population sequencing" or "pool-seq", and the latter term will be used here. In pool-seq, multiple colonies from the same species can be sampled genetically at the same economic cost as sequencing an individual isolate. Pool-seq has generally been found to be reliable in measuring sequence variation and allele frequency (11,16). However, the disadvantages of this method are the perceived complexity of the bioinformatic analysis and the complications of assessing phenotypic characteristics of the population of bacterial clones, such as antibiotic resistance, when these assays typically require clonally purified single colonies.

The single-isolate sequencing is a convenient sampling strategy based on the assumption that strain mixtures are rare and capturing within-strain microdiversity is not worth the additional expense of sequencing. While the cost of raw sequence production has steadily declined, costs of labour and infrastructure, such as DNA extraction, library preparation, physical sample storage, and bioinformatic analysis, have not scaled down at the same rate (17). There has also been little analysis of what the increased sequencing and storage costs from sampling multiple colonies or pools yield over single colonies. However, pool-seq can still provide insights into the natural history of even well-studied pathogens, and inform us about the fate of adaptations that enhance virulence and antibiotic-resistance (1,18–21). Therefore, optimising sampling strategies and genomic workflow design is essential to minimise the number of samples processed while maximising the information obtained from each clinical sample.

In this work, we use samples from an ongoing study of S*taphylococcus aureus* colonisation on humans to compare the three strategies outlined above: single-isolate sequencing, sequencing collections of multiple single colonies, and pool-seq. *S. aureus* is a ubiquitous nosocomial pathogen prevalent worldwide, causing invasive disease syndromes such as bacteremia, endocarditis and osteomyelitis (22,23). Like other prominent pathogens, WGS has significantly improved *S. aureus* epidemiologic studies, and our ability to track the spread of antibiotic resistance and virulence across populations (2,24–28). Here, we used samples from human participants who had an index methicillin-resistant *S. aureus* (MRSA) skin and soft tissue infection (SSTI) as part of an ongoing study, SEMAPHORE (Study of the Evolution of MRSA, Antibiotics and Persistence Having the Outcome of Recurrence). The study was designed to examine clinical and demographic characteristics of the participants, and the genomes of the colonising *S. aureus* to identify factors associated with recurrent skin infections. However, for this paper, we focused on the relationship between the pool-seq and collections of single isolate genome sequencing. We first quantified the amount of variation within the collections of single genome and pool-seq and then investigated three specific questions: 1) Could the pool-seq data identify clonal *S. aureus* populations (comprising a single ST) from mixtures of diverse lineages?; 2) Could pool-seq data be used to estimate the number of sites within single-ST populations undergoing polymorphisms?; 3) Was pool-seq more sensitive in detecting antimicrobial resistance (AMR) genes than sequencing single clones?

Results

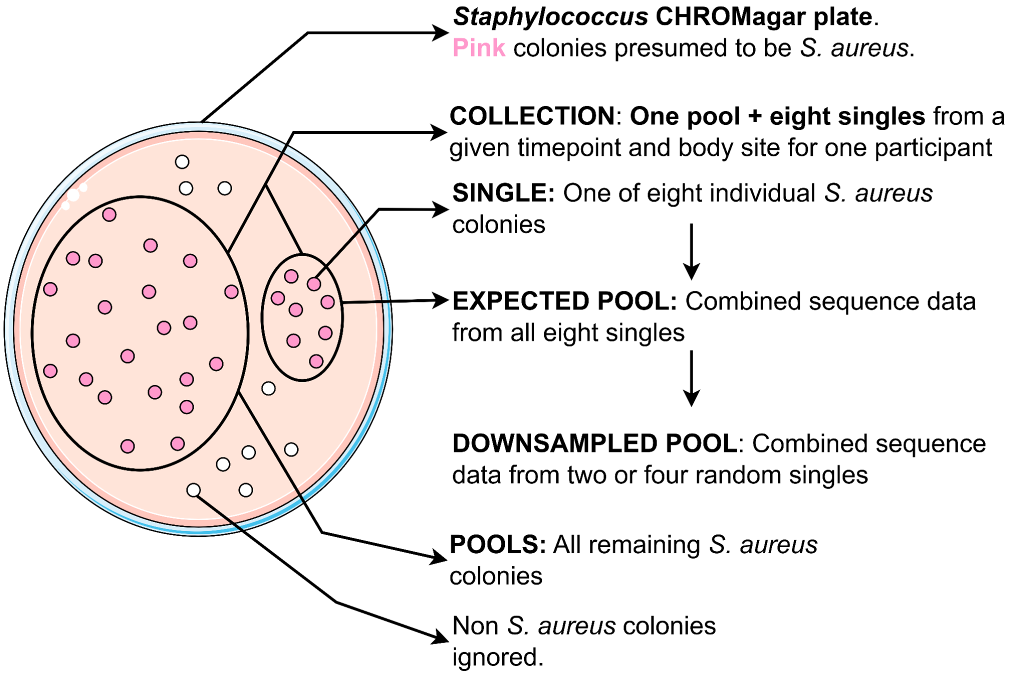
Staphylococcus CHROMagar plate.
Pink colonies presumed to be *S. aureus*.
COLLECTION: One pool + eight singles from a given timepoint and body site for one participant
SINGLE: One of eight individual *S. aureus* colonies
EXPECTED POOL: Combined sequence data from all eight singles
DOWNSAMPLED POOL: Combined sequence data from two or four random singles
POOLS: All remaining *S. aureus* colonies
Non *S. aureus* colonies ignored.

Fig 1: Schematic representation of colony collection strategy, names, and descriptions of isolate groups analysed in this study.

Samples from the SEMAPHORE study were plated on CHROMAgar *Staphylococcus aureus* and a "collection" of eight individual *S. aureus* colonies ("singles") was obtained (Fig 1). The remaining *S. aureus* colonies on the plate were pooled and sequenced, hereon referred to as "pools" or "pool-seq". The collective sequencing data obtained from all eight singles for each pool were referred to as "expected pools". Similarly, sequencing data sampled from two random singles and four random singles were combined to generate "downsampled pools". This study had 85 participants with 254 samples (254 pools and 254 collections of 8 singles - 2032 singles total). All FASTQ files (pool-seq and singles) were capped to 100x *S. aureus* genome coverage.

82% of collections (eight single genomes) and their corresponding pools have only one Multilocus Sequence Type (MLST).

We found a wide range of SNP distances between singles from the same collection, with a minimum of 0 and a maximum of 15,315. For 241 out of 254 collections (~95%), the maximum SNP distance between any two pairs of isolates was < 100 (**Fig 2A**), suggesting that most collections comprised only closely related isolates. However, 12 collections (5%) showed clear signs of mixed infections, with a maximum SNP distance > 4000. When we compared the MLST amongst the singles for each collection, all 12 of these collections had at least one isolate that was a different ST from the remaining. This showed that comparing STs and pairwise SNP distances between single isolates within collections could identify potential mixed infections, as single ST collections had lower maximum pairwise SNP distances.

In 209/254 collections (~82%) the ST types for the eight singles and the pool were identical, suggesting they were single ST samples. In the remaining 45 collections (~18%) either at least one single or the pool had a different ST (**Fig 2B, C**). For 37 out of these 45 collections, the ST of the pool was untypeable either due to presence of multiple alleles for the same gene or an unknown/undetectable allele. While 59 STs were identified, 51% of singles (1051/2032) belonged to ST8 and ST5 (**Fig 2B,D**). We observed no significant differences in the occurrence of multi-ST pools across the different timepoints, body sites and culturing methods (Chi-squared test, p>0.01). These data suggested that a given collection usually had a low level of *S. aureus* diversity, and that we can find ST mixtures by comparing SNP distances and ST types within collections of single colonies.

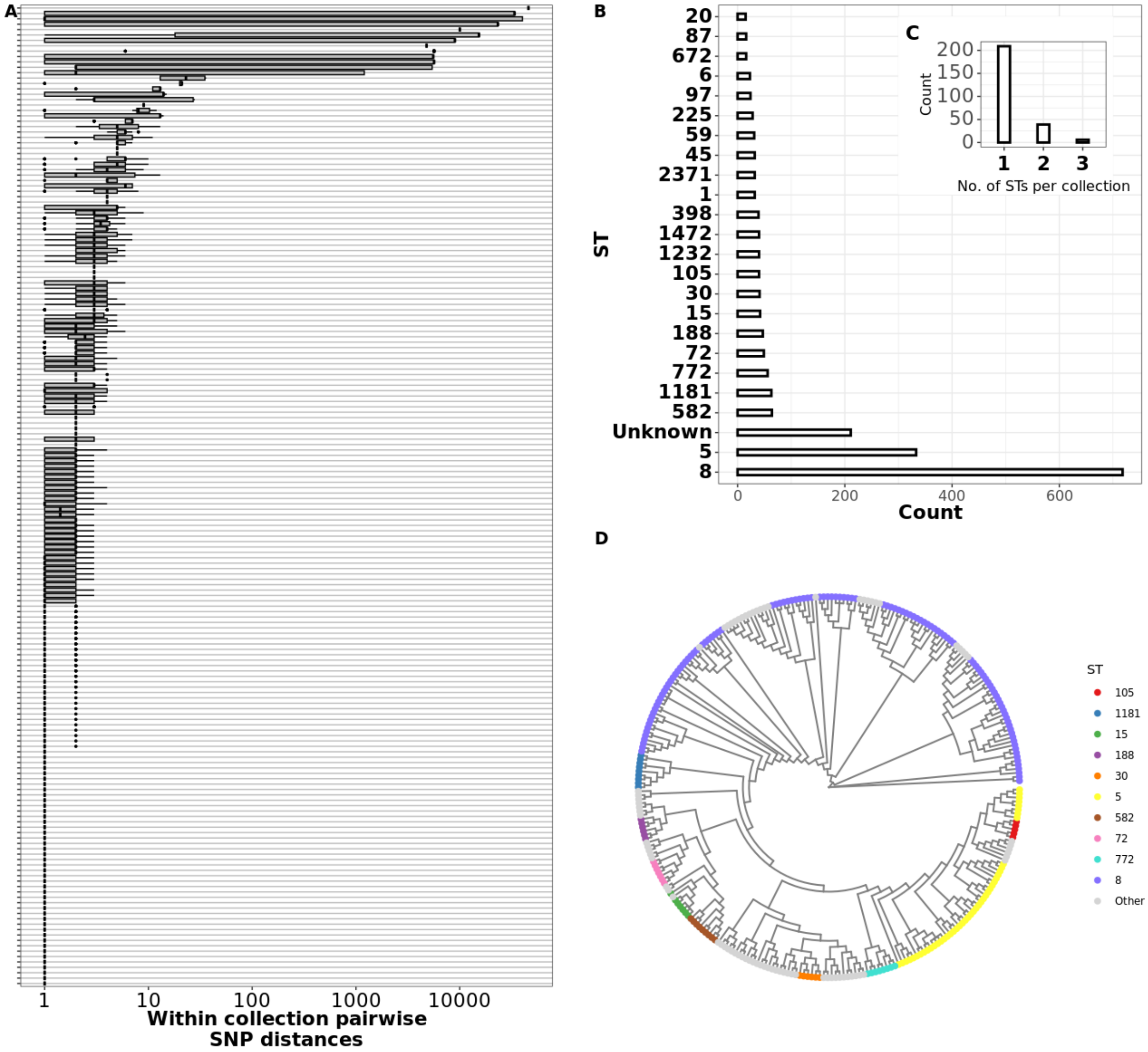

Fig 2: Pairwise SNP distance between and within collections.

(**A**) Boxplots showing per-collection SNP distance distributions. For each collection shown in the y-axis, the x-axis shows the corresponding distribution of core genome SNP distances in log scale. Black vertical lines show the median SNP distances and boxes show the interquartile range. Whiskers represent values up to 1.5 times the first or third quartile. Black dots represent outliers beyond the whiskers range. (B) Barplot showing number of genomes per Sequence Type (ST). Multilocus Sequence Typing (MLST) was performed by the software tool mlst (see methods). x-axis shows the number of isolates assigned to the corresponding ST shown in the y-axis. (C) Bar plot showing number of STs detected per participant. MLST typing was performed for all eight singles from a participant and the number of unique STs detected per participant was plotted. (D) Maximum likelihood phylogeny representing at least one isolate from all collections. All non-identical genomes from each collection were aligned by snippy and a core genome phylogeny was constructed using fasttree (see methods). Tree tips are coloured by ST, only top 10 most frequent STs are shown, and remaining are grouped into "Other".

Pool-seq samples with elevated average minor allele frequency, elevated number of contigs, higher nucleotide diversity and untypable MLST were associated with strain mixtures

We examined what features of the pool-seq data could be used to assess whether there was a single clonal ST genotype present, or mixture of genotypes, focusing on five measures: MLST, assembly quality, nucleotide diversity, gene number and minor allele frequency (MAF).

We found that we could use MLST software to determine the ST of the pool-seq data from 185 out of 224 samples. 183/185 (99%) of the typable samples were associated with single ST collections. In the remaining 39 untypable pool-seq samples, 17 (45%) were associated with ST mixtures.

Sequencing reads from both singles and pool-seq were processed identically using the Bactopia pipeline with the same quality control parameters (29). Both single and pool-seq reads had a final average quality score of 36.3 (Welch's t-test $p > 0.01$). We expected the genome assemblies (generated using the SKESA assembler (30)) from single colonies to be higher quality than pool-seq, as the latter may contain multiple *S. aureus* strains and possibly contaminating species from the culture plate. We evaluated assembly quality using CheckM and QUAST (31,32), observing that, while most pools and singles had comparable coverage (**Fig 3A**, Wilcoxon $p>0.01$, effect size=0.052), pools had higher number of contigs (**Fig 3B**, Wilcoxon $p < 0.01$, effect size = 0.20), higher heterogeneity (**Fig 3C**, Wilcoxon $p < 0.01$, effect size 0.347), and contamination scores (**Fig 3C**, Wilcoxon $p < 0.01$, effect size = 0.239). 32 out of 224 pools (14%) had more than 200 contigs in contrast to only 5 out of 1792 singles

(0.2%). The CheckM heterogeneity score indicates source of the contamination – a heterogeneity score < 50% indicates that the source of contamination is phylogenetically distant and vice versa (30). While all singles had contamination and heterogeneity scores of 0, the pools ranged from low heterogeneity contamination to high heterogeneity contamination (**Fig 3C**). 7 pools (3%) had a heterogeneity score > 50 with a contamination score >10, suggesting they were contaminated by phylogenetically similar sources. 15 pools (6%) had a heterogeneity score < 50 with a contamination score >10, suggesting they are contaminated by phylogenetically distant sources. Overall, these results suggested that genome assembly quality can be useful for assessing population heterogeneity.

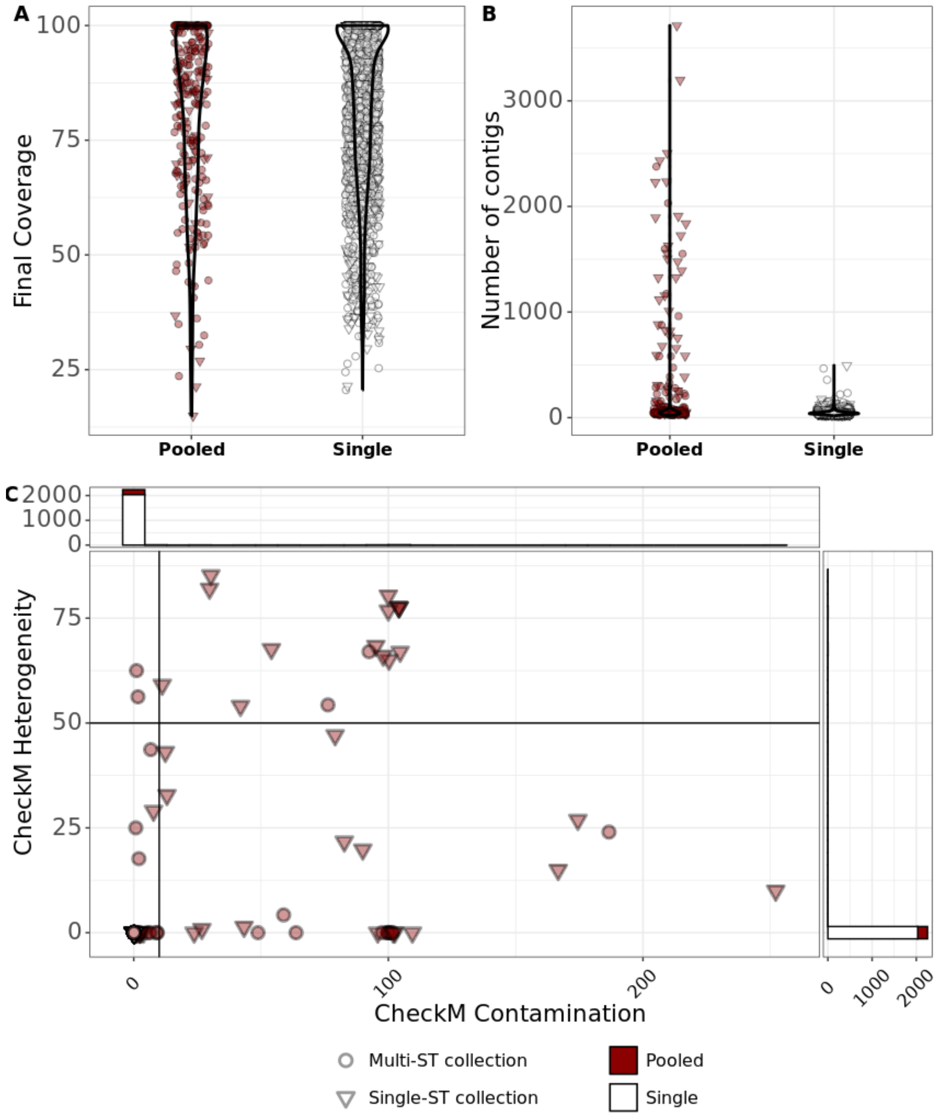

Fig 3: Assembly quality can be used to assess population heterogeneity.

(A) There was no significant difference in the assembly coverage between pools and singles. Violin plot showing distribution of assembly coverage between pools and singles. Assembly coverage for each pool and single was calculated by Bactopia against an auto-chosen reference (see methods). Circles indicate single ST collections and triangles indicate multi-ST collections. **(B) Pool assemblies were more likely to have a higher number of contigs than single assemblies.** Violin plot showing distribution of number of assembly contigs in pools and singles. Pooled samples were processed identically to singles with Bactopia using SPAdes. Circles indicate single ST collections and triangles indicate multi-ST collections. **(C) Pooled samples have varying sources of contamination while singles are pure**. CheckM contamination and heterogeneity scores showed that all single colonies have no contamination while 6% of pools are contaminated by phylogenetically distant sources and 3% of pools are contaminated by phylogenetically similar sources. The blue line marks a heterogeneity score of 50 below which the source of contamination is considered phylogenetically distant and vice versa. Circles indicate single ST collections and triangles indicate multi-ST collections.

Allele frequency (AF) - i.e., the fraction of reads piled up over a particular variant position is a useful metric of genomic heterogeneity. We mapped the pooled sequences against the closest complete *S. aureus* reference genome (see methods) to obtain the variant sites. The eight singles were also aligned to the same reference as the corresponding pool. For each sample in a collection, we calculated the average minor allele frequency, i.e., the total number of variant sites divided by the sum of all minor allele frequencies (MAF). If all reads mapped to only the reference or only the alternate alleles, the average MAF would always be 0, as would be expected from ideal single pure cultures. We plotted the average MAF against the total number of variant sites for 254 pools (**Fig 4A**). We split the plot into four quadrants based on two parameters - the number of variants cutoff of 2800 sites (or 0.1% of the *S. aureus* genome) suggesting only few variant sites, and average MAF cutoff of 0.05 below which we deemed the sample as having no minor alleles. 223 pools (~88%) had a total number of variant sites less than 0.1 % of the *S. aureus* genome (**Fig 4A left quadrant**). 55 out of these 223 had an average MAF < 0.05 (**Fig 4A bottom left quadrant**) suggesting highly homogeneous samples. In contrast, there were 14 pool-seqs with more than 2800 variants (~0.1%) of MAF > 0.05 (top right quadrant of **Fig 4A)**. Pool-seqs in the bottom right quadrant (average MAF < 0.05 with > 2800 variants) show samples that are distant from their reference sequence though still homogeneous. To assign a diversity score based on MAFs, we used the product of the total number of variants and the average MAF, which we termed the "MAF Index". The MAF Index was higher for samples with both a large number of variants compared to reference as well as a high average MAF (i.e., top right quadrant – **Fig 4A**). The MAF index of single-ST pools were not significantly greater than the MAF index of singles (Welch's t test, $p>0.01$). However, within the pools, the multi-ST pools had significantly greater MAF index than the single ST pools (Welch's t test, $p< 0.01$).

AFs are measured based on whether or not a given position in a read maps to the reference, but our calculations did not take into account the possibility of multiple alternate alleles (which we assumed to be very rare given that the number of variants was only a small percentage of the chromosome). Therefore, we also calculated intra-sample nucleotide diversity between the true pooled sequences and our expected pools as an analogous method for measuring genomic heterogeneity. Using a software called InStrain (33), we estimated Nucleotide diversity (π), which is for each position, 1 minus the sum of the frequency of each base squared. This value was then averaged across the whole genome. We used InStrain to measure the average π across our singles, downsampled pools (four and two colony pools), expected pools and true pools. The π value was significantly greater in pools compared to singles (Welch's test, $p < 0.01$). However, in 216 pools (96%), the diversity observed in the true pools were less than the expected diversity observed from an in-silico mixture of two *S. aureus* isolates 30,000 SNPs apart in a 99:1 ratio (**Fig 4B LEFT** black horizontal line - see methods). This analysis suggested most pools comprised only single strains. Moreover, similar to average MAF, π of multi-ST pools was significantly greater than π of single-ST pools (Welch's test, $p < 0.01$). In cases where there was an increased diversity value for our expected pools or downsampled pools compared to our true pools, we may have overestimated the diversity of our true pool by assuming the single colonies were present in equal abundance. Alternatively, since we pooled only the remaining colonies on the plate after picking singles, we may have negated some diversity from the true pool. In the 12 cases where the true pools had a diversity value greater than the corresponding expected pools (5%), the 8 colonies sampled did not capture the entire diversity of the pool.

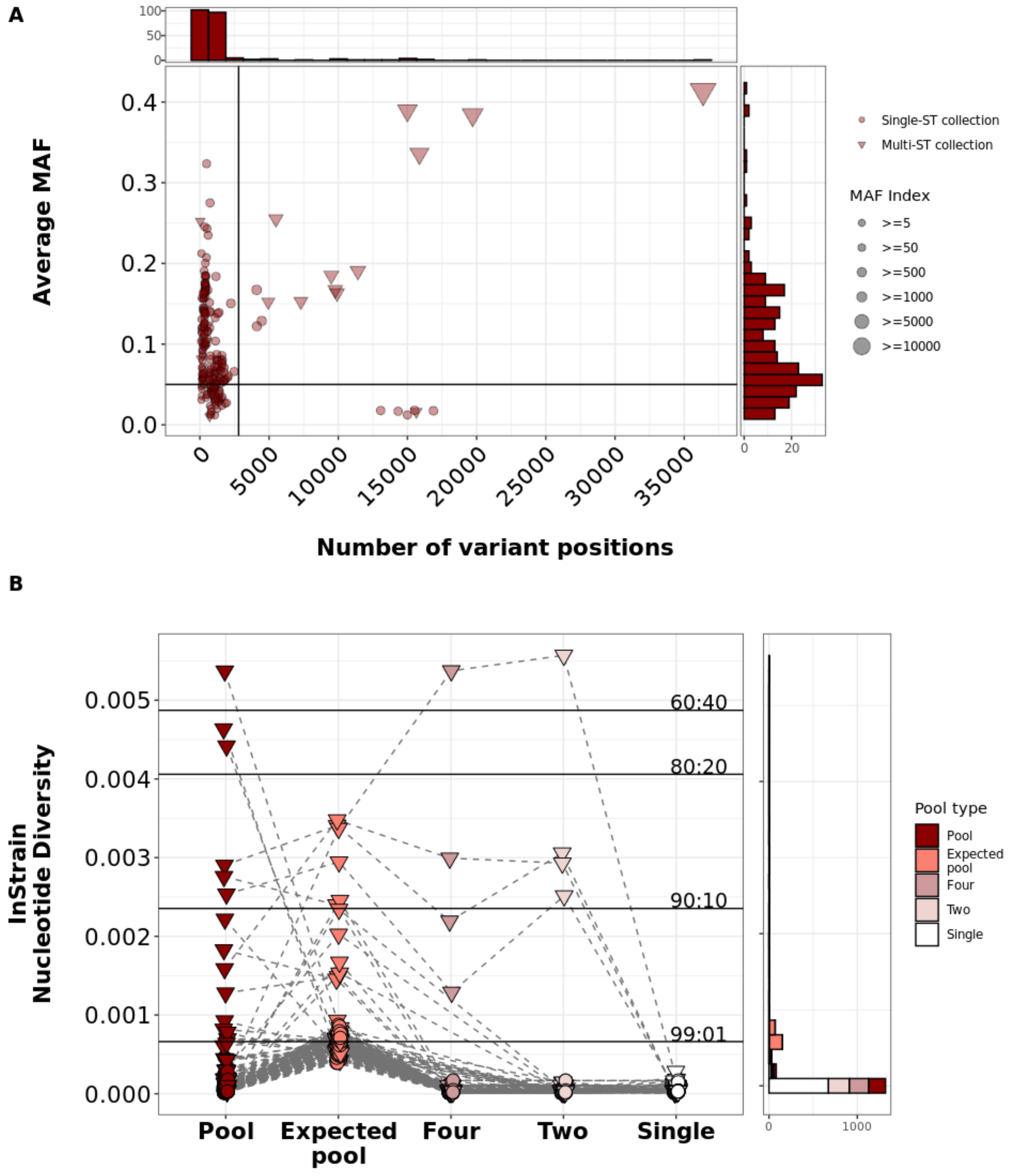

Fig 4: Average MAF and average π can be used to detect multi-ST pools

(A) The MAF index could be used to assess multi-ST pools. Dot plot depicting the number of variant positions and the average MAF for single-ST (circles) and multi-ST (triangle) pools. The x-axis indicates the number of variant positions compared to a reference. The y-axis indicates the average minor allele frequency (MAF). The average MAF was calculated by summing the MAFs of all intermediate alleles and dividing by the total number of variant positions. Red dots correspond to single ST pools and triangles correspond to multi-st pools. The black horizontal line indicates an average MAF of 0.1. The black vertical line indicates 0.1% of the *S. aureus* genome (2800 sites). The frequency of the dots at their corresponding x and y positions are indicated by the histogram above the x and y axis respectively. (**B**) **Average nucleotide diversity suggested most pools comprise single strains. 94% of pools had nucleotide diversity less than a theoretical 99:1 mixture of two strains.** LEFT: Dots and colours indicate average nucleotide diversity value for each pool, expected pool (reads from eight singles combined in equal proportions), downsampled pools (reads from four and two random singles combined in equal proportions) and singles. Grey dashed lines connect corresponding samples. Black solid horizontal lines indicate the average nucleotide diversity value for in-silico mixtures of two *S. aureus* genomes 30,000 SNPs apart. The ratio of each mixture is indicated over each solid black line. The frequency of the dots at their corresponding x positions are indicated by the histogram to the right.

So far, we have shown that the number of contigs, contamination, minor allele frequencies, and nucleotide diversity are significantly different between pools and singles (**Fig3**, **Fig4**). Next, we wanted to measure the magnitude of these parameters' contribution to the variability between pools and singles.

We performed principal component analysis (PCA) for five different parameters we measured (CheckM contamination, CheckM heterogeneity, MAF Index, Nucleotide diversity, and number of contigs) . We found that PC1 and PC2 explained ~78% of the total variance (**Fig S1A**). All five parameters had positive loadings in PC1 (>0.4) and the CheckM contamination score had positive loadings in PC2 (>0.6) (**Table S1**). This result suggested that the deviation of some pools from the singles were mainly due to contamination (PC1) and allelic variation (PC2).

We also performed an all vs all Pearson's correlation across the 5 different parameters mentioned above (**Fig S1B**). We found CheckM contamination, number of contigs, and CheckM heterogeneity were positively correlated with each other, suggesting that contamination reduced the assembly quality (larger number of contigs). However, these three parameters did not have high positive correlation with the MAF index nor with Nucleotide diversity. This showed that contamination and allelic diversity can independently drive heterogeneity in the pool, and that pooling multiple colonies may impact sequencing and assembly quality regardless of intra-species diversity.

From our analysis thus far, we have shown that there are pools that behave like singles, and there are true mixtures. As we mentioned earlier (**Fig 2**), detecting multiple MLST types in the pool or measuring pairwise SNP distances between singles from within a collection are a reliable way to ascertain true mixtures. However, when the MLST calls are unreliable (unassigned types/undetectable alleles) or hypothetically if we did not have single colonies, alternative methods would be required. Therefore, we wanted to test whether the above mentioned parameters (Number of contigs, CheckM contamination, CheckM heterogeneity, MAF index, Nucleotide diversity) could serve as predictors for mixed pools and homogeneous pools.

We performed a logistic regression using the number of contigs, CheckM contamination and heterogeneity, the MAF index, and nucleotide diversity as predictor variables to calculate the probability that a given pool is mixed (See methods). Here, we defined a mixed pool or multi-ST pool as a pool with multiple ST calls, or a pool corresponding to a collection of singles where the maximum pairwise SNP distance is > 2800 (0.1% of the *S. aureus* genome). Our logistic regression model showed strong predictive ability with a McFadden R^2 of 0.59, sensitivity of 1, specificity of 0.94, and a receiver operating characteristic (ROC) curve with an area

0.86 (**Fig S1C**). The maximum variance inflation factor (VIF) across our predictor variables was < 2.3 indicating low multicollinearity. Overall, these results show that by using information from multiple statistics, the pool-seq data alone was sufficient to predict the presence of multi-strain populations with high accuracy.

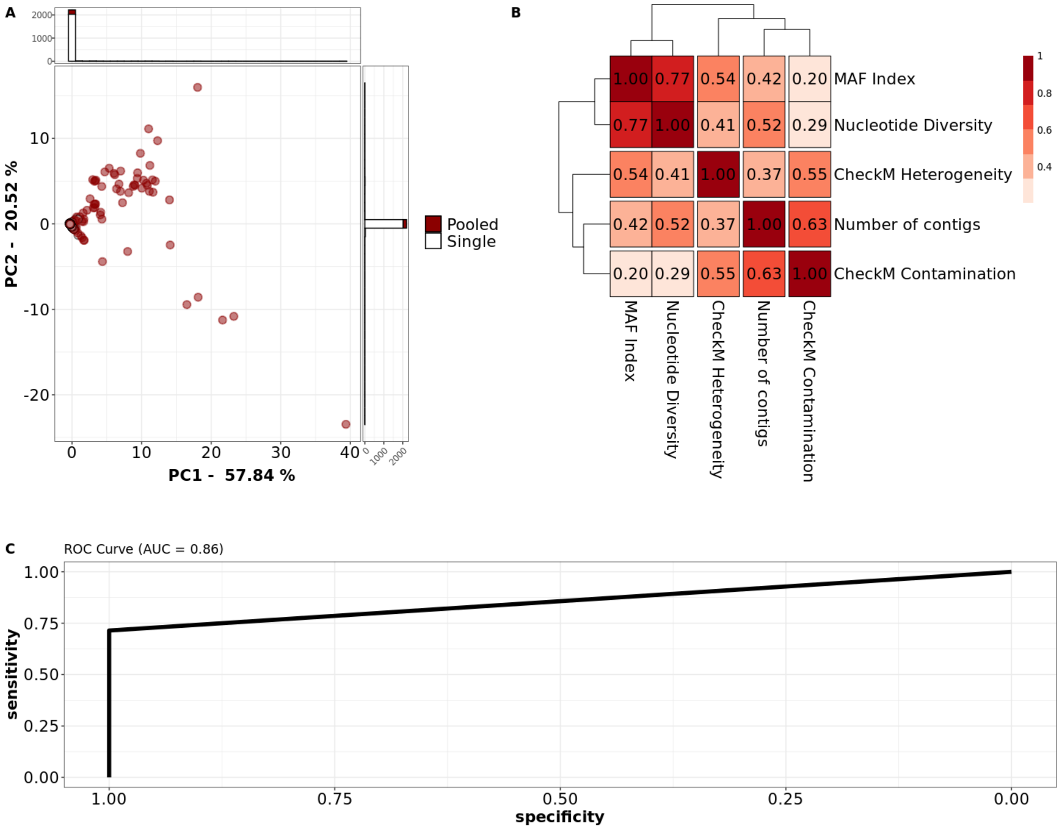

Fig S1: Variation in pools was primarily driven by contamination and allelic diversity. (**A**) PCA loading plot for principal components (PC) 1 (x-axis) and 2 (y-axis) explaining 70% of the total variance. White dots represent singles and red dots represent pools. 256 pools and 2032 singles were used for PCA. The density of the dots at their corresponding x and y positions are indicated by the histogram above and to the right of the plot respectively. The variance explained by each PC is indicated in the corresponding axis labels. (**B**) Pearson's correlation coefficient matrix across five different diversity metrics. Each square indicates the Pearson *r* for comparing the corresponding parameters as labelled in the x and y axis. Scale indicates Pearson's *r* (Darker = higher *r*). (**C**) Receiver operating characteristic (ROC) curve of the logistic model predicting multi-ST pools from parameters in **A** and **B**. Area under the curve (AUC) = 0.86.

	PC1	PC2	PC3	PC4	PC5
MAF Index	0.462809	-0.5268234	0.1029145	-0.186782	-0.6802838
Number of contigs	0.4541359	0.288277	-0.5902644	-0.5812948	0.1560169
CheckM Contamination	0.4020166	0.6543029	0.0179086	0.5201767	-0.3733174
CheckM Heterogeneity	0.4418238	0.1308287	0.7578247	-0.2563222	0.3842866
Nucleotide Diversity	0.4719563	-0.4405963	-0.2576384	0.5393737	0.4752164

Table S1: Summary of all five principal components (PC1 - PC5) for five parameters used in Fig S1. All 254 pools and 2032 singles were used for principal component analysis.

Numbers of variants in pool-seq and eight singles from the same sample are correlated but pool-seq had greater number

One of the advantages of pool-seq over groups of singles is the potential to discover mutant subpopulations that may be missing in samples of individual clones. We measured the number of variant positions that were shared between the pool and at least one of the eight singles. For each collection, we calculated the number of variant positions seen both in the pool and in at least one of the corresponding eight singles as a fraction of the total number of variant positions observed. For analysing variants in the singles, and the expected and downsampled pools that were built from singles, we only considered sites with an AF > 0.95. We found that 152 collections out of 254 (~60%) had shared variant fraction >0.5, meaning, more than half the variants found in each pool and the corresponding singles were identical for 60% of our samples **(Fig S2A)**. Curiously, we observed 30 collections (~12%) having a shared fraction < 0.05.

This is what would be expected if these singles and pool-seq were not from the same sample **(Fig S2B)**. These collections may have been mis-sampled and we opted not to use them for further comparisons of pools and singles within the same sample. This brought down our total number of collections from 254 to 224. Out of the 224, 204 were cases where the pool-seq and all eight singles had the same sequence type.

We found that the number of variants found in the pools was greater than the combined number of variants from the eight singles in 178 out of 224 samples(~79%). This was as expected as the pools should more often contain more individual isolates than the collections.

To illustrate this point further we compared the number of variants detected in the pools against eight singles combined (expected pool), four random singles and two random singles combined (downsampled pools) and one random single. This was done to answer the question - How many variants would we have seen if we had sampled only eight colonies/only four/only two/only one? We considered a variant present in the expected or downsampled pools if it was present in at least one of the sampled singles at an AF > 0.95.

We found that 198 pools (~88%) captured more than 75% of all the variants in a collection (**Fig 5A**). This was significantly greater than the number of expected pools (129 pools or 56%) that captured a fraction of variants > 0.75 (Kolmogorov Smirnov $p < 0.01$). If we had sampled only one single colony for each collection, only 39% of the singles would have captured a fraction of variants > 0.75 (**Fig 5A - "one colony")**.

Though the number of variants observed in the pools were usually greater than in the singles, we found that the more singles a variant was present in, the more likely we were to detect the same variant in the pool. We counted the number of singles each variant was present in and plotted it against the AF of the same variant in the pool and found a strong positive correlation (Pearson *r* = 0.83) (**Fig 5B**).

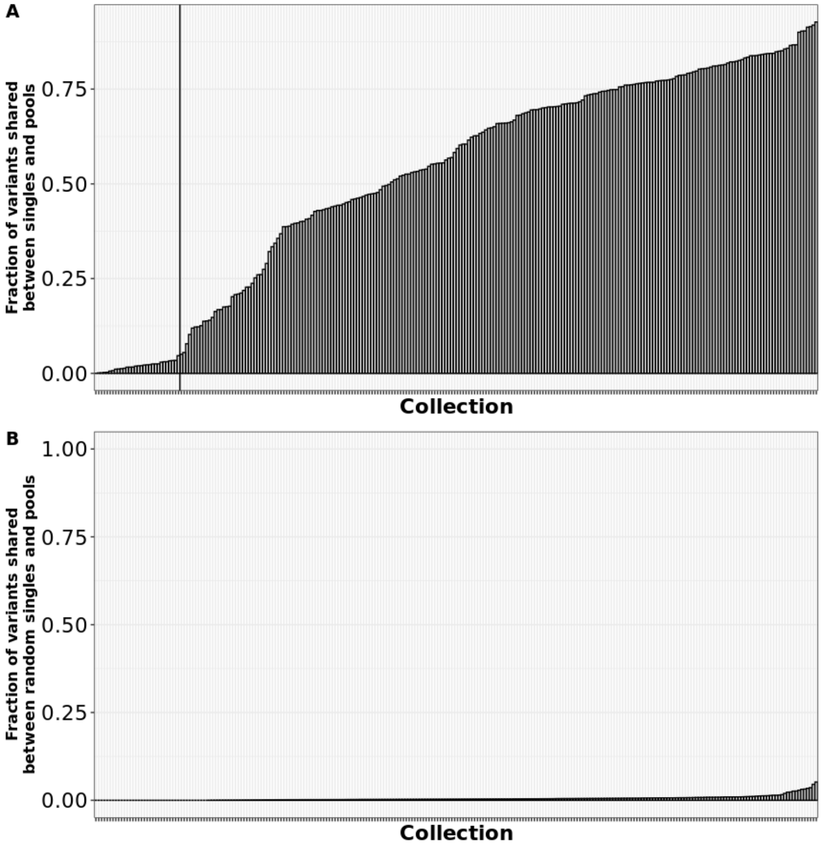

Figure S2: Collections with <5% of their total variants shared between pools and singles were discarded.

(**A**) **Number of shared allelic sites revealed differences in the amount of diversity captured by single colonies and pools.** Each bar indicates a collection, and the height of the bar indicates the fraction of variants shared between the pools and at least one of the eight corresponding singles. Black vertical line indicates the threshold for shared fraction below which the singles and pools are not from the same sample (< 5% of variants shared) (**B**) **Expected fraction of allelic sites shared between a pool and a random collection of eight singles**. Each bar indicates a collection and the height of the bar indicates the fraction of variants shared between the pool and at least one of eight singles from a random other collection. The maximum observed fraction did not exceed ~5% after 10 repetitions.

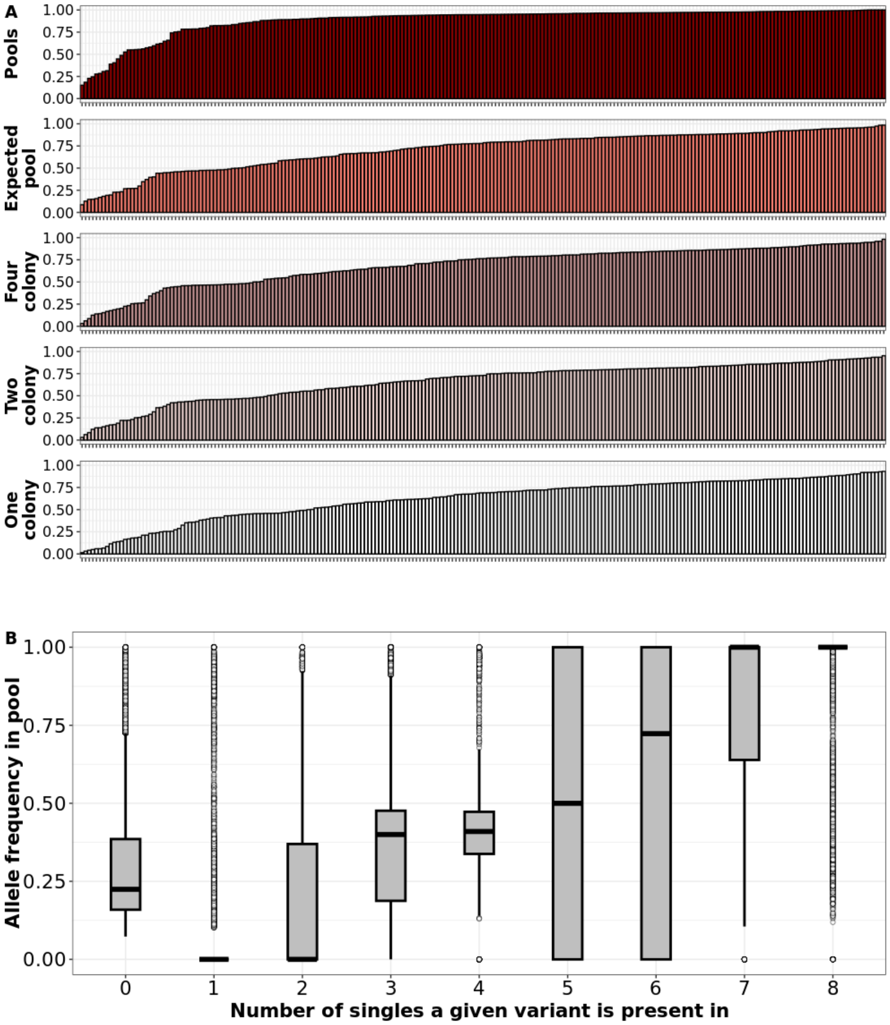

Fig 5:Pools were a better representation of the total number of variants in the population (A) Pools captured more variants than eight single colonies combined. Each bar indicates a collection and the height of the bar indicates the fraction of variants found in the corresponding sample group (Pools, expected pools, four colony pools, two colony pools, single colony) to the total number of variants found in all samples in the collection (Pool plus all eight singles). For example, a bar with height 0.25 in the fifth row (**One colony**) shows that if one random single colony was examined from the specific collection corresponding to the bar, we would find 50% of the total number of variants found in the collection (Pool plus all eight singles). Bars for each sample group are ordered by lowest to highest. A value of one indicates 100% of the variants found in both the pools and all eight singles combined are represented in the sample group. **(B) Allele frequencies in the pool were proportional to the number of singles the variant was detected in.** Boxplots showing allele frequencies of variants detected in zero singles up to eight singles. Allele frequency of each variant found in the pool increased as the variant was found in more colonies in the corresponding singles. Boxes show the interquartile range and whiskers represent values up to 1.5 times the first or third quartile. White dots represent outliers beyond the whiskers range. Black horizontal line in each boxplot indicates the mean.

Numbers of segregating sites in pools and singles from the same sample are positively correlated

As a bacterial population expands from an introduction event, mutations accumulate as a function of time (34,35). Subpopulations can segregate from the parent population by accumulating nucleotide variants at different sites across the genome and the total diversity across different subpopulations can be altered by bottlenecks and selective sweeps (8). The number of segregating sites (or within-population polymorphic sites) can therefore be an important indicator of the demographic history of the population, and it would be useful to know how well the pool-seq data could be used to estimate this value. Because we compared pools and singles to a common reference, a certain number of variants were likely fixed in the ancestor of the population. We expected these to have an AF of 1 or close to 1 (0.95 or greater) in both the pools and singles and filtered them out. We also filtered out samples where the ST of the pools and collection did not match the ST of the auto-chosen reference, as this would lead to an elevated number of variants. We found a moderate positive correlation between the number of segregation sites in the true pools and expected pools of the 198 samples that had matched ST across the pools, singles and the reference (**Fig 6A**; Pearson *r* = 0.352). The number of segregating sites in the collection ranged from 8 to 1658. While the number of segregating sites was comparable between the true and expected pools, we also wanted to measure whether the proportion of the variants in the singles could reliably predict the proportion of the same variants in the pools. For each collection of sequences, we plotted the AF of the segregating sites observed in both the expected pool sequences and the true pooled sequences and calculated Pearson's coefficient (*r*). If the AF of a variant present in an expected pool was equal to the AF of the same variant present in the true pool, we inferred that the proportion of the variant in the two populations was comparable. In other words, the variant frequencies in the eight singles combined (for example, if variant present in seven out of eight singles, AF = 0.875, if variant present in six out of eight singles, AF = 0.75... and so on) was equivalent to the variant frequencies in the pool. In contrast, if the AF of variants between the expected and true pools were not comparable, the pool-seq was significantly different from the expected pools. The distribution of *r* values indicated only 82 collections (41%) with $r > 0.5$ (**Fig 6B**). This result showed that in only less than half of our collections with matched ST, the proportion of the variants in the singles are positive predictors of the proportion of the same variant in the pool.

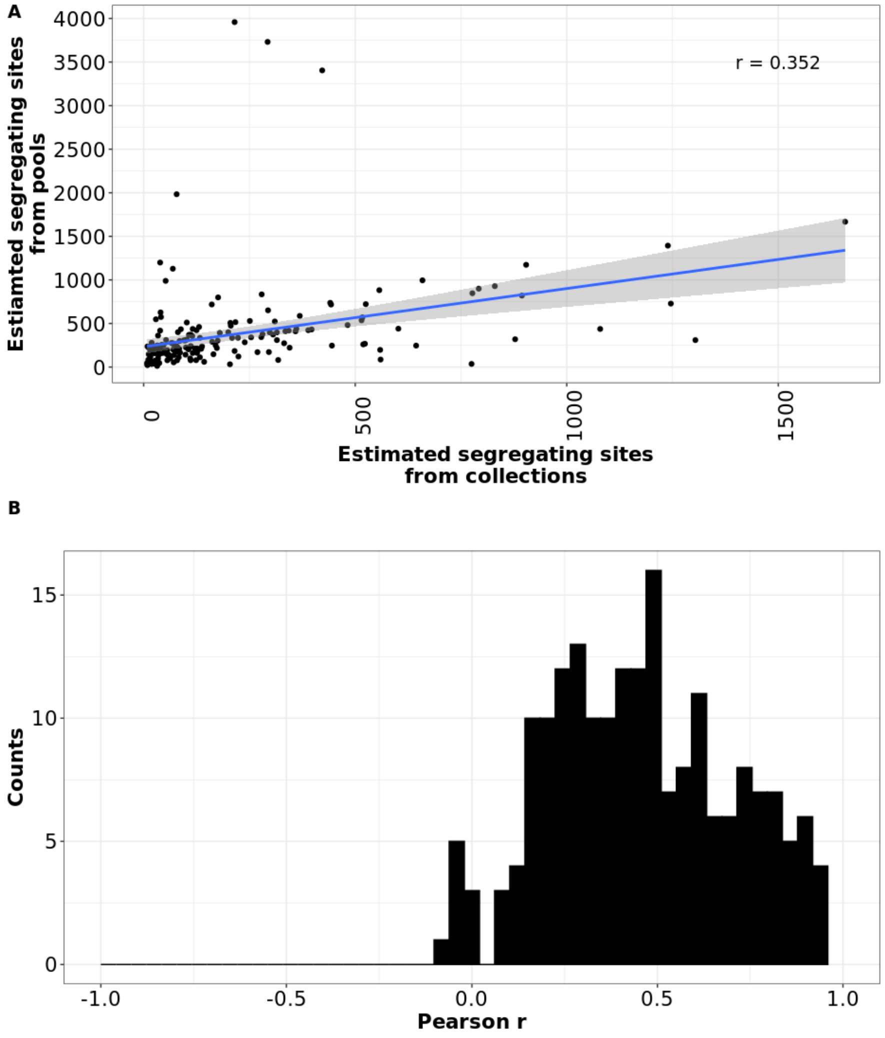

Fig 6 legend: Allelic variation in pools and singles from the same sample were positively correlated

(A) The number of segregating sites in the true pools were proportional to the number of segregating sites in the expected pool. For single-ST collections (collections where all eight singles, the pool and the auto-chosen reference were called the same ST), the number of sites with allelic variation was comparable between the true pools (y-axis) and the expected pool (x-axis) (eight singles combined). If the same site was fixed in all eight singles and in the pool, it was not included. Blue regression line depicts a linear relationship with a Pearson's *r* of 0.352. **(B) AFs of variants in the expected pool did not reliably predict the AFs of the same variants in the true pool**. Frequency distribution plot showing Pearson's *r* for all 198 single ST collections. x-axis depicts Pearson's *r* and y-axis depicts number of collections.

A median of one more AMR gene was detected in the pools compared to singles

Finally, we wanted to know if the pools could harbour subpopulations with clinically relevant genes that may be missing in singles. We annotated AMR genes using AMRFinderPlus and counted the number of antimicrobial drug classes for which resistance determinants were found in our pools, individual singles, and in the pangenome of our expected (all genes eight singles combined) and downsampled (all genes from two or four random singles combined) pools (36). In 177 collections (79%), the number of AMR classes was identical in pools and the expected pool. This group represented the bulk of the low-diversity samples in the study. However, overall, we observed a median of one additional AMR class in our true pools compared to the expected/downsampled pools and singles (**Fig 7, black vertical line**). This showed that additional genes could be detected in the pool that are absent in the pangenome of the singles and that these genes can be of clinical relevance. We would like to note that in all cases where we found *mecA* in the pools (134 out of 226 pools), we found *mecA* in at least one of the eight corresponding singles. A summary file with all detected resistant determinants for all collections is reported in the supplemental file '**Supplemental_dataset_1.xlsx**' available in the github https://github.com/VishnuRaghuram94/GASP.

We initially used the total number of genes and the number of AMR classes in the pools as predictor variables in our logistic regression model in **Fig S1**. We found that the number of genes were highly multicollinear with CheckM contamination (VIF > 5) and the number of AMR classes had no predictive power (near identical AUC, Accuracy, sensitivity and specificity with or without the AMR class parameter), and therefore did not include them in our final analysis.

Next, we wanted to measure the abundances of AMR genes present in both pools and the pangenome of the singles compared to AMR genes present in the pools alone. We hypothesised that in cases where an AMR gene was found in the pools but absent in the singles, the AMR gene was present at low abundances. To test this, we used Salmon (37) to estimate the abundance of AMR genes found in both pools and singles, and compared the abundances to when it was found in the pools alone [**Fig S3**]. We found that the mean copy number of genes belonging to a particular class of antibiotic was significantly lower when it was found only in the pools for eight out of nine AMR classes (Welch's t test with Bonferroni correction).

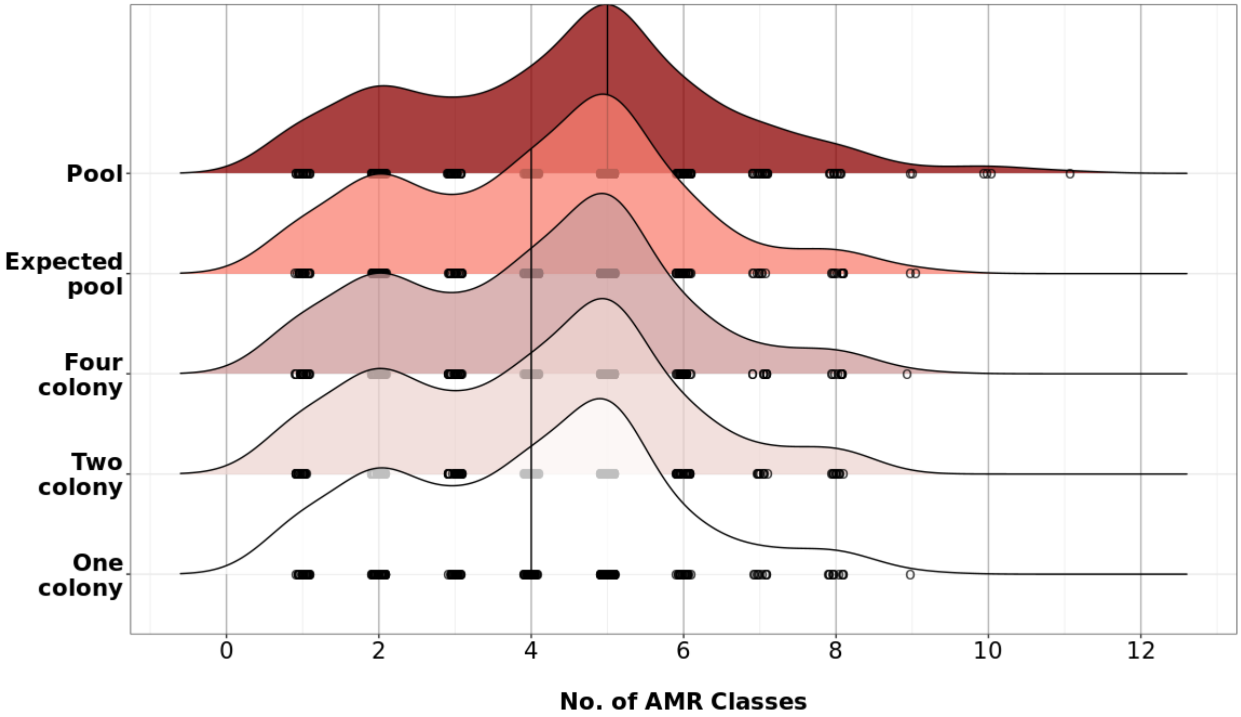

Fig 7: A median of one additional AMR class can be observed in the pools compared to singles. Ridgeline plot showing number of AMR gene classes detected in pools, the pangenome of expected and downsampled pools (pangenome of eight, four and two singles combined), and a random single colony. The x-axis shows the number of AMR classes detected in the sample by AMRFinder and the y-axis shows the corresponding sample. Black vertical line shows the median number of AMR classes detected for each sample group. White circles under each ridgeline represent individual collections and the number of AMR classes detected.

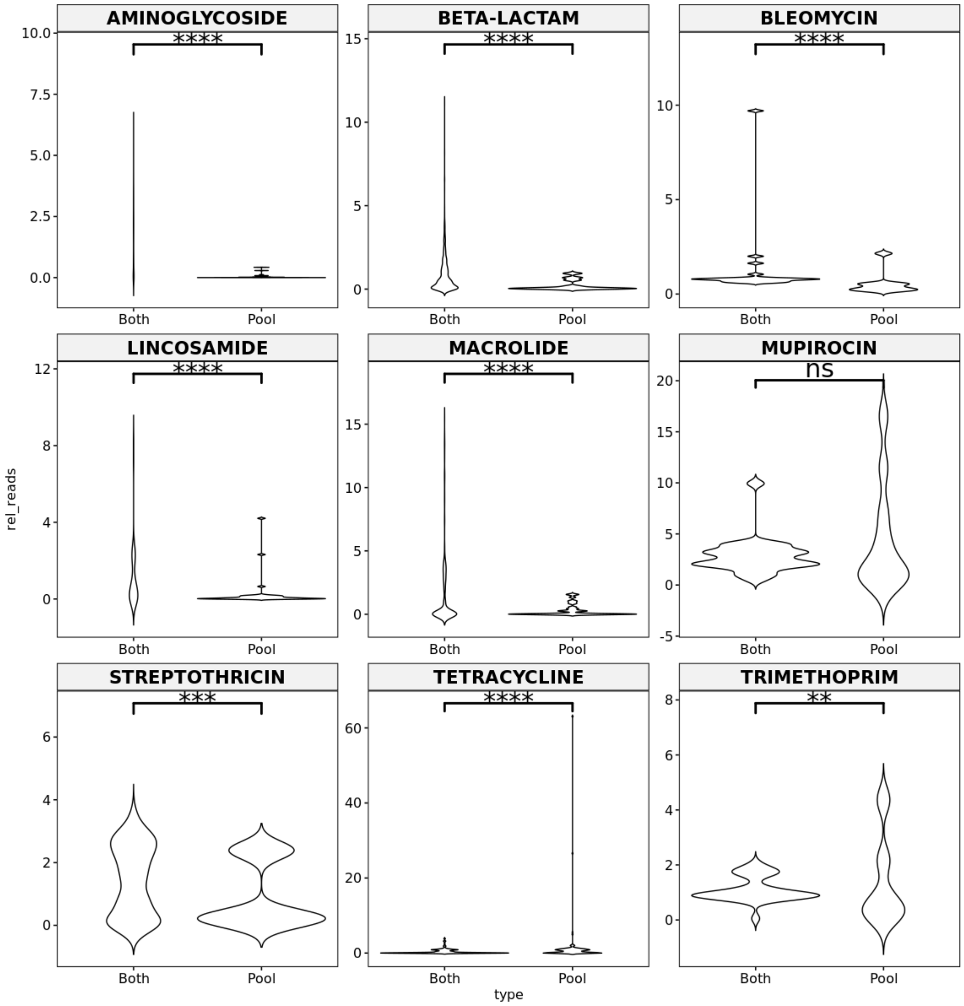

Fig S3: Mean read abundance is lower for AMR genes present only in the pools compared to AMR genes present in both pools and singles.

For each class of AMR, we estimated the number of reads mapped to each AMR gene in the AMRFinder database relative to the number of reads mapped to *rpoD* (relative copy number). All genes were normalised to 1 kb. For each AMR class, the relative copy number of genes found in both the pool and the corresponding single ("Both") were compared against genes for the same AMR class found only in the pool ("Pool") using Wilcoxon rank sum test with Bonferroni correction. ns = p > 0.01; ** = p < 0.001; *** = p < 0.0001; **** = p < 0.00001.

Discussion

From this study, we derived insights into strategies for sampling genomic diversity of *S. aureus* asymptomatically colonised human skin and mucosal surfaces. We found that in most cases (83%), *S. aureus* populations were clonal, representing only one ST. Interestingly, there were no significant differences in the incidence of multi-ST populations across the three anatomic sites sampled (anterior nares, oropharynx (throat), inguinal skin), four different timepoints (at participant enrollment, three months, six months, nine months and twelve months after enrollment), nor across different culturing methods (direct vs enrichment, see methods). Many of the conclusions learned from this study could be applied by sampling other bacterial pathogens (or *S. aureus* on other hosts/ anatomic sites). Still, as genetic diversity in the populations may be greater or lesser, different cost benefit tradeoffs may apply.

The primary question that we sought to address was: how many sampled colonies are sufficient to capture the total intra-species diversity within a host? To answer this question, we compared pure single colonies, *in-silico* single-colony mixtures (expected pools and downsampled pools), as well as total pools of 10s – 100s of colonies (**Fig 1**). One significant finding was that in our samples, the *S. aureus* within-host diversity for a given body site and time point was relatively low – most (83%) collections were of a single clonal lineage (**Fig 2**). In these cases, there was not a significant amount of additional information that could be obtained from the pooled samples (**Fig 5**).

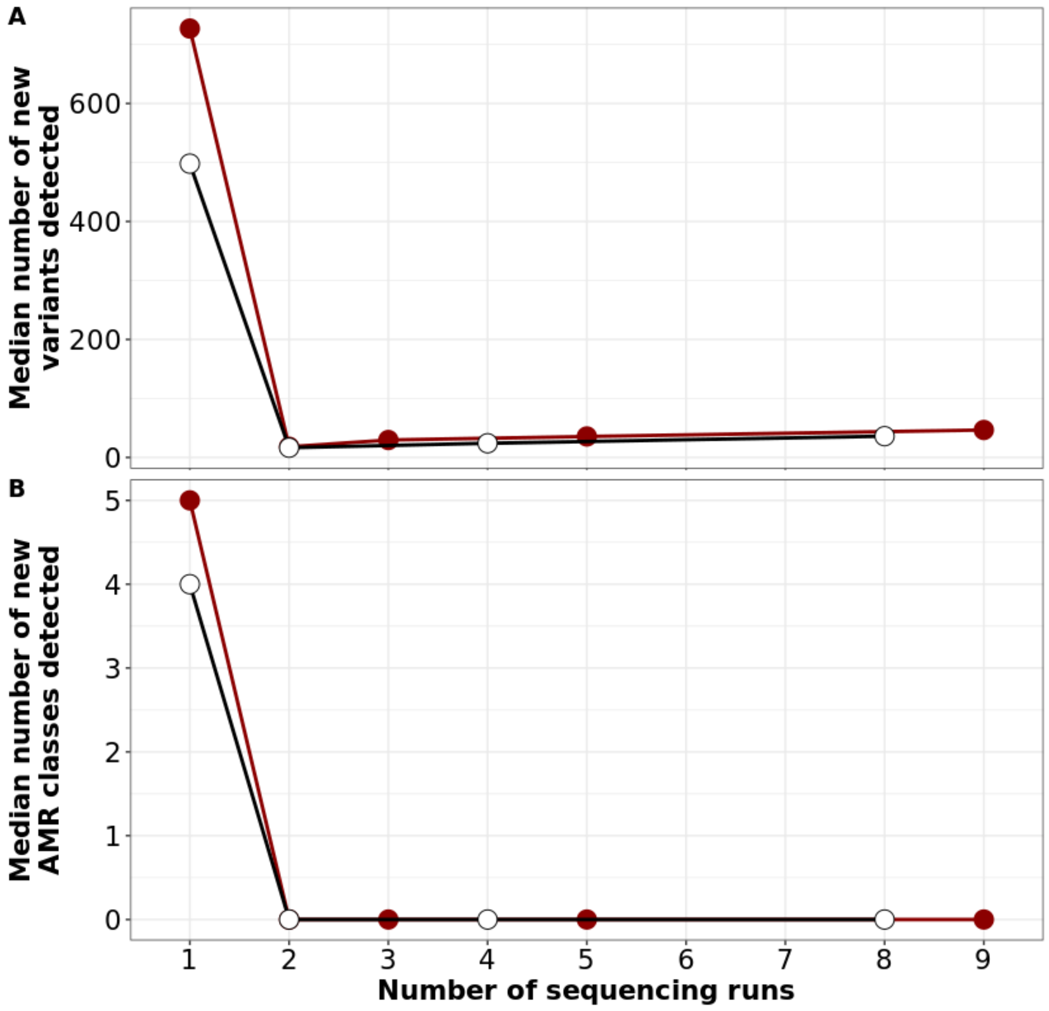

Fig 8 legend: Number of new variants or new AMR genes observed with the addition of more sequencing runs.

Dot plot depicting number of new variants (A) or new AMR genes (B) observed for additional sequencing runs. Red dots depict the first sequencing run being the pool, and the additional runs being single colonies (1 = Pool, 2 = Pool + one single, 3 = Pool + two singles…). White dots depict only singles (1 = one single, 2 = two singles, …)

Assuming sequencing one pool incurs one unit cost (cost of time, labour and resources for sample preparation, storage, sequencing, and analysis), every single colony added on top of the pool would incur an additional unit cost. For our dataset, one additional unit cost over the pool (pool + one single) yielded a median of 19 new variants and 0 new AMR gene classes (**Fig 8**). Moreover, we also showed that the pool alone is sufficient to predict the presence of a multi-ST population (**Fig 3,4, S1**).

The study had some limitations. First, we negated diversity from the true pool by picking singles prior to pooling the remaining colonies. Our assumption was that the population in a single colony may be present multiple times on a plate with 100s of colonies, but this assumption is less likely to hold true in scenarios where there were < 15 - 20 colonies left after picking singles. Ideally, the number of colonies in a pool should be relatively constant across all samples but in some cases very few colonies appeared on the selective agar. In other cases, the density of colonies was too great to measure colony counts accurately. Moreover, laboratory culture media could also cause biases in population growth. Re-dilution and re-plating were not realistic in a high-throughput setting, so the colony count of these pools was higher.

Second, our sampling space is narrow - all our samples are from one geographical location, comprising only nares, throat, and skin swab-acquired *S. aureus*. Therefore, the amount of diversity we measured across singles and pools from a given sample may not apply to cultures from other clinical contexts of *S. aureus* or other *S. aureus* strain types typically colonising people worldwide. Despite these drawbacks, the data in this study allowed us to compare three strategies for sampling: individual single colonies, collections of up to eight colonies and pool-seq.

Sampling single colonies in pure culture is the traditional approach for assessing the genotypic and phenotypic characteristics of bacterial pathogens. Only one sequencing library is required and the bioinformatic analysis methods are straightforward. However, sampling only one colony will result in missing multi-strain infections (17% of the time in our case) and therefore provide an overly simplistic view of the population structure of the pathogen.

The more single colonies sequenced, the better the estimation of true population diversity (assuming there are no systematic sampling biases in how colonies are picked from the culture plate). Having collections of single colonies also allows the construction of within-host phylogenies, observing gene gain/loss events and inferring demographic changes over longitudinal sampling. However, there is still no guarantee that the total diversity in the population is represented in the sample subset. Moreover, the cost of processing and physical and data storage scales linearly with the number of independent colonies sampled. Deciding the number of colonies to sample will be a complex calculus of budget and *a priori* estimation of the population diversity of the pathogen aligned with goals of the study.

Ideally, pool-seq would provide the best estimation diversity in the population and many more single colonies can be aggregated than sequenced individually. After pooling colonies, the stock can be treated as a single sample for storage, sequencing, and analysis therefore the cost is equal to one single colony, making pool-seq the best value for identifying variation. Here, we have shown that pool-seq can be used to accurately estimate the presence of multi-ST infections, can measure segregating sites within single-ST populations, and are most sensitive at finding AMR genes. The analysis of pool-seq data, which are effectively single-species metagenomes, are more complex than single colony sequencing due to variation, especially in multi-ST samples, and the possibility of contamination. This heterogeneity also leads to unreliable phenotypic ascertainment. However, predicted phenotypes could be validated by replating the pool to pick singles.

Based on our analysis, we recommend that many studies may benefit from using a **one plus pool** design, that is sequencing one single colony + pooling and sequencing all remaining colonies. Using the sample diversity measurements we show in this study, many of which are obtained from the default outputs of Bactopia, a streamlined beginner-friendly analysis pipeline, we can ascertain whether significant differences exist between the single colony and the pool (**Fig 3,4)**. This can aid in deciding whether sequencing additional colonies from the pool is required. We believe this approach provides more information than a single colony while demanding little additional time and labour for sample collection, storage, and analysis. A disadvantage of processing pooled samples is the reduced sequence quality and increased likelihood of contamination (**Fig 3**). However, with the additional sequencing of at least one pure single colony, this disadvantage can be mitigated.

Methods

Strain sampling

Participants were enrolled into the SEMAPHORE study after presenting with a *S. aureus* positive SSTI. Up to four timepoints (every three months for one year) and up to three body-sites (Anterior nares, oropharynx, and inguinal skin) were sampled for each participant. Each swab was streaked out onto BBL™ CHROMAgar™ *Staphylococcus aureus* (SACA). Suspected *S. aureus* colonies were verified with a catalase and Staphaurex™ Latex Agglutination tests. Then, eight individual colonies (singles) were subcultured onto blood agar and sequenced. The remaining colonies were pooled, sub-cultured onto blood agar, then sequenced. In cases where there was no growth from directly plating the swabs, each swab was enriched for growth in tryptic soy broth (TSB) overnight. These overnight cultures were then plated, and colonies were picked as mentioned above. 181 pools and 1448 singles were directly plated on SACA (Direct cultures). The remaining 73 pools and 584 singles did not show growth upon direct plating and therefore were enriched in TSB and then plated (Enrichment cultures).

In this study, we only analysed swabs from which eight singles and a pool were obtained. Swabs where fewer than eight singles were obtained were not considered. In total, we obtained 254 pools from 85 participants and eight singles corresponding to each pool giving us a total of 2286 genome sequences.

Library preparation and sequencing

Genomic DNA extractions were performed using Qiagen kits. Library preparation and whole genome sequencing were performed by the Children's Hospital of Pennsylvania High Throughput Sequencing core using Illumina MiSeq or Hiseq platforms.

Genome assembly, annotation and variant calling using Bactopia

All obtained sequences were processed using the Bactopia analysis pipeline (29). Bactopia performed adapter trimming using BBTools (38), genome assembly using SKESA (30) and the assembly quality was assessed using QUAST and CheckM (31,32). Genome annotation was done using Prokka (39) and AMR genes were annotated using AMRFinderPlus (36). Variant calling was performed by Snippy (40) using an automatically selected reference sequence based on the closest MASH (41) distance to a complete *S. aureus* genome sequence in NCBI RefSeq. MLST types were identified using MLST (42).

Pairwise SNP distance calculation, dereplication, and phylogeny

For each group of eight singles in a collection, we used Parsnp v1.5.3 to align the single colony isolated genomes and used snp-dists v0.7.0 to calculate pairwise SNP distances (43). To dereplicate singles, isolates with SNP distance < 10 were collapsed into clusters and a random isolate was chosen as the cluster representative using Assembly-Dereplicator (44). The final set comprised 294 singles, where each collection is represented at least once. This resulting set of singles was aligned again using parsnp and a core genome phylogeny was constructed using FastTree (45,46). Phylogeny was visualised using ggtree (47)

Number of variants, segregating sites, and allele frequency calculation

We calculated allele frequencies from bam files generated by Bactopia using bcftools mpileup (48). The reference for each pool was auto-chosen by Bactopia based on the closest complete *S. aureus* genome in terms of MASH distance (41). All singles were then aligned to the same reference as their corresponding pool. Only variants with a QUAL score > 50 and with at least a read depth of 25 were considered for the analysis. For each collection, we calculated the allele frequencies for every position across the genome where there was at least one read piled up with a base call differing from the reference allele. Variants with frequencies < 0.05 were considered 0 (absence) and > 0.95 were considered 1 (fixed). The allele frequencies for the expected and downsampled pools were calculated based on the number of singles the variant was fixed in. For eg: if a variant was present in one out of the eight singles at a frequency > 0.95 (fixed), its allele frequency in the expected pool would be ⅛ or 0.125. Two or four random singles out of the eight were selected to measure the allele frequencies in the downsampled pools. Variants with intermediate frequencies in the singles (>0.05 & < 0.95 were not considered).

To calculate the number of segregating sites across the true and expected pools, we wanted to exclude variants that occurred simply as a result of alignment against a specific reference. If a given variant was fixed in the expected pool (present in eight out of eight singles at an AF > 0.95) and also in the true pool (AF > 0.95), we considered these variants to be ancestral and did not count them as segregating sites. All remaining sites with AF > 0.05 were counted.

We calculated nucleotide diversity (π) using InStrain (33) using the auto-chosen reference and the alignment bam file from Bactopia. Expected pools (eight colonies) and downsampled pools (two and four colonies) were generated by combining equal proportions of reads from all eight, two or four colonies. For each collection, we used reformat.sh from the bbtools suite (38) to sample reads 12.5% from all eight colonies for the expected pool, 50% of reads from two randomly selected colonies for the two-colony downsampled pool, and 25% of reads from four randomly selected colonies for the four-colony downsampled pool. All artificial pools (expected and downsampled) contained 1 million reads.

Logistic regression

Logistic regression was performed in R using the glm function (49). 70% of 254 pool-seq samples were used as the training set and the remaining 30% was used as the test set. A pool was considered multi-ST if the MLST alleles in the pool and the corresponding eight singles were not identical. Continuous probabilities from the logistic regression model were converted to binary using a cutoff of 0.89 (If probability > 0.89, the prediction was considered to be multi-ST). This cutoff was estimated using the optimalCutoff function from the R package InformationValue (50). McFadden R^2 was calculated using the pR2 function from the R package pscl (51). Variance Inflation Factor was calculated using the vif function from the R package car (52).

Statistical analyses and data visualization

All statistics and PCA were performed in R using packages stats and rstatix (49,53). All plots were visualised using R package ggplot2 (54). Other graphics were created using bioicons and draw.io (55,56).

Data availability

All code and raw data are available at https://github.com/VishnuRaghuram94/GASP. All genome sequences used in this study are available under PRJNA918392.

Acknowledgements

TDR and MZD acknowledge funding from NIH NIAID 1 R01 AI158452-01A1. VR was supported by NIH AI139188 and by Emory University and the Infectious Disease Across Scales Training Program (IDASTP). We thank Dr. Joanna Goldberg for providing constructive comments on the manuscript.

References

1. Giulieri SG, Guérillot R, Duchene S, Hachani A, Daniel D, Seemann T, et al. Niche-specific genome degradation and convergent evolution shaping Staphylococcus aureus adaptation during severe infections. Kana BD, Van Tyne D, Zheng M, editors. eLife. 2022 Jun 14;11:e77195.
2. Talbot BM, Jacko NF, Petit RA III, Pegues DA, Shumaker MJ, Read TD, et al. Unsuspected Clonal Spread of Methicillin-Resistant Staphylococcus aureus Causing Bloodstream Infections in Hospitalized Adults Detected Using Whole Genome Sequencing. Clin Infect Dis. 2022 Dec 15;75(12):2104–12.
3. Armstrong GL, MacCannell DR, Taylor J, Carleton HA, Neuhaus EB, Bradbury RS, et al. Pathogen Genomics in Public Health. N Engl J Med. 2019 Dec 26;381(26):2569–80.
4. Chaguza C, Senghore M, Bojang E, Gladstone RA, Lo SW, Tientcheu PE, et al. Within-host microevolution of Streptococcus pneumoniae is rapid and adaptive during natural colonisation. Nat Commun. 2020 Jul 10;11:3442.
5. Golubchik T, Batty EM, Miller RR, Farr H, Young BC, Larner-Svensson H, et al. Within-Host Evolution of Staphylococcus aureus during Asymptomatic Carriage. PLOS ONE. 2013 May 1;8(5):e61319.
6. Marvig RL, Sommer LM, Molin S, Johansen HK. Convergent evolution and adaptation of Pseudomonas aeruginosa within patients with cystic fibrosis. Nat Genet. 2015 Jan;47(1):57–64.
7. Azimi S, Roberts AEL, Peng S, Weitz JS, McNally A, Brown SP, et al. Allelic polymorphism shapes community function in evolving Pseudomonas aeruginosa populations. ISME J. 2020 Aug;14(8):1929–42.
8. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. Within-host evolution of bacterial pathogens. Nat Rev Microbiol. 2016 Mar;14(3):150–62.
9. Paterson GK, Harrison EM, Murray GGR, Welch JJ, Warland JH, Holden MTG, et al. Capturing the cloud of diversity reveals complexity and heterogeneity of MRSA carriage, infection and transmission. Nat Commun. 2015 Mar 27;6(1):6560.
10. Markussen T, Marvig RL, Gómez-Lozano M, Aanæs K, Burleigh AE, Høiby N, et al. Environmental Heterogeneity Drives Within-Host Diversification and Evolution of Pseudomonas aeruginosa. mBio. 2014 Sep 16;5(5):e01592-14.
11. Lieberman TD, Flett KB, Yelin I, Martin TR, McAdam AJ, Priebe GP, et al. Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. Nat Genet. 2014 Jan;46(1):82–7.
12. Tonkin-Hill G, Ling C, Chaguza C, Salter SJ, Hinfonthong P, Nikolaou E, et al. Pneumococcal within-host diversity during colonization, transmission and treatment. Nat Microbiol. 2022 Nov;7(11):1791–804.
13. Bryant JM, Brown KP, Burbaud S, Everall I, Belardinelli JM, Rodriguez-Rincon D, et al. Stepwise pathogenic evolution of Mycobacterium abscessus. Science. 2021 Apr 30;372(6541):eabb8699.
14. Wilkinson DJ, Dickins B, Robinson K, Winter JA. Genomic diversity of Helicobacter pylori populations from different regions of the human stomach. Gut Microbes. 2022 Dec 31;14(1):2152306.

15. Liu Q, Via LE, Luo T, Liang L, Liu X, Wu S, et al. Within patient microevolution of Mycobacterium tuberculosis correlates with heterogeneous responses to treatment. Sci Rep. 2015 Dec 1;5(1):17507.
16. Holt KE, Teo YY, Li H, Nair S, Dougan G, Wain J, et al. Detecting SNPs and estimating allele frequencies in clonal bacterial populations by sequencing pooled DNA. Bioinformatics. 2009 Aug 15;25(16):2074–5.
17. Rossen JWA, Friedrich AW, Moran-Gilad J. Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. Clin Microbiol Infect. 2018 Apr 1;24(4):355–60.
18. Brodrick HJ, Raven KE, Kallonen T, Jamrozy D, Blane B, Brown NM, et al. Longitudinal genomic surveillance of multidrug-resistant Escherichia coli carriage in a long-term care facility in the United Kingdom. Genome Med. 2017 Jul 25;9(1):70.
19. Jorth P, Durfey S, Rezayat A, Garudathri J, Ratjen A, Staudinger BJ, et al. Cystic Fibrosis Lung Function Decline after Within-Host Evolution Increases Virulence of Infecting Pseudomonas aeruginosa. Am J Respir Crit Care Med. 2021 Mar;203(5):637–40.
20. Joseph SJ, Bommana S, Ziklo N, Kama M, Dean D, Read TD. Patterns of within-host spread of Chlamydia trachomatis between vagina, endocervix and rectum revealed by comparative genomic analysis [Internet]. bioRxiv; 2023 [cited 2023 Mar 13]. p. 2023.01.25.525576. Available from: https://www.biorxiv.org/content/10.1101/2023.01.25.525576v1
21. Octavia S, Wang Q, Tanaka MM, Sintchenko V, Lan R. Genomic heterogeneity of Salmonella enterica serovar Typhimurium bacteriuria from chronic infection. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis. 2017 Jul;51:17–20.
22. David MZ, Daum RS. Treatment of Staphylococcus aureus Infections. Curr Top Microbiol Immunol. 2017;409:325–83.
23. Kourtis AP, Hatfield K, Baggs J, Mu Y, See I, Epson E, et al. 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 Mar 8;68(9):214–9.
24. Coll F, Harrison EM, Toleman MS, Reuter S, Raven KE, Blane B, et al. Longitudinal genomic surveillance of MRSA in the UK reveals transmission patterns in hospitals and the community. Sci Transl Med. 2017 Oct 25;9(413):eaak9745.
25. Giulieri SG, Baines SL, Guerillot R, Seemann T, Gonçalves da Silva A, Schultz M, et al. Genomic exploration of sequential clinical isolates reveals a distinctive molecular signature of persistent Staphylococcus aureus bacteraemia. Genome Med. 2018 Aug 23;10(1):65.
26. Sabat AJ, Hermelijn SM, Akkerboom V, Juliana A, Degener JE, Grundmann H, et al. Complete-genome sequencing elucidates outbreak dynamics of CA-MRSA USA300 (ST8-spa t008) in an academic hospital of Paramaribo, Republic of Suriname. Sci Rep. 2017 Jan 20;7:41050.
27. Xu Z, Yuan C. Molecular Epidemiology of Staphylococcus aureus in China Reveals the Key Gene Features Involved in Epidemic Transmission and Adaptive Evolution. Microbiol Spectr. 2022 Oct 3;10(5):e01564-22.
28. Yebra G, Harling-Lee JD, Lycett S, Aarestrup FM, Larsen G, Cavaco LM, et al.

Multiclonal human origin and global expansion of an endemic bacterial pathogen of livestock. Proc Natl Acad Sci. 2022 Dec 13;119(50):e2211217119.

29. Petit RA, Read TD. Bactopia: a Flexible Pipeline for Complete Analysis of Bacterial Genomes. mSystems. 2020 Aug 4;5(4):e00190-20.
30. Souvorov A, Agarwala R, Lipman DJ. SKESA: strategic k-mer extension for scrupulous assemblies. Genome Biol. 2018 Oct 4;19(1):153.
31. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015 May 14;gr.186072.114.
32. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013 Apr 15;29(8):1072–5.
33. Olm MR, Crits-Christoph A, Bouma-Gregson K, Firek BA, Morowitz MJ, Banfield JF. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat Biotechnol. 2021 Jun;39(6):727–36.
34. Bosshard L, Peischl S, Ackermann M, Excoffier L. Dissection of the mutation accumulation process during bacterial range expansions. BMC Genomics. 2020 Mar 23;21(1):253.
35. Biek R, Pybus OG, Lloyd-Smith JO, Didelot X. Measurably evolving pathogens in the genomic era. Trends Ecol Evol. 2015 Jun;30(6):306–13.
36. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep. 2021 Jun 16;11(1):12728.
37. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017 Apr;14(4):417–9.
38. BBMap [Internet]. SourceForge. 2022 [cited 2023 Apr 2]. Available from: https://sourceforge.net/projects/bbmap/
39. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinforma Oxf Engl. 2014 Jul 15;30(14):2068–9.
40. Seemann T. Snippy [Internet]. 2023 [cited 2023 Apr 2]. Available from: https://github.com/tseemann/snippy
41. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016 Jun 20;17(1):132.
42. Seemann T. mlst [Internet]. 2023 [cited 2023 Apr 2]. Available from: https://github.com/tseemann/mlst
43. Seemann T. Source code for snp-dists software [Internet]. Zenodo; 2018 [cited 2023 Mar 15]. Available from: https://zenodo.org/record/1411986
44. Wick RR, Holt KE. rrwick/Assembly-Dereplicator: Assembly Dereplicator v0.1.0 [Internet]. Zenodo; 2019 [cited 2023 Apr 2]. Available from: https://zenodo.org/record/3365572
45. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol. 2014 Nov 19;15(11):524.

46. Price MN, Dehal PS, Arkin AP. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. Mol Biol Evol. 2009 Jul 1;26(7):1641–50.
47. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017;8(1):28–36.
48. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinforma Oxf Engl. 2011 Nov 1;27(21):2987–93.
49. R: The R Project for Statistical Computing [Internet]. [cited 2023 Apr 3]. Available from: https://www.r-project.org/
50. InformationValue: Performance Analysis and Companion Functions for Binary Classification Models version 1.2.3 from CRAN [Internet]. [cited 2023 Apr 25]. Available from: https://rdrr.io/cran/InformationValue/
51. Tahk A. Political Science Computational Laboratory [Internet]. 2023 [cited 2023 Apr 3]. Available from: https://github.com/atahk/pscl
52. Fox J, Weisberg S. An R Companion to Applied Regression [Internet]. Third. Thousand Oaks, CA: Sage; 2019. Available from: https://socialsciences.mcmaster.ca/jfox/Books/Companion/
53. Kassambara A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests [Internet]. 2023 [cited 2023 Apr 3]. Available from: https://CRAN.R-project.org/package=rstatix
54. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; 2016. Available from: https://ggplot2.tidyverse.org
55. Duerr S. Bioicons [Internet]. 2023 [cited 2023 Mar 30]. Available from: https://github.com/duerrsimon/bioicons
56. jgraph/drawio [Internet]. JGraph; 2023 [cited 2023 Mar 30]. Available from: https://github.com/jgraph/drawio

Chapter V - *Staphylococcus aureus* and *Pseudomonas aeruginosa* isolates from the same cystic fibrosis respiratory sample coexist in coculture

Eryn E. Bernardy[1,2#], Vishnu Raghuram[2,3#], and Joanna B. Goldberg[2*]

[1]Department of Biology, Elon University, Elon, North Carolina, USA

[2] Division of Pulmonary, Asthma, Cystic Fibrosis, and Sleep, Department of Pediatrics, Emory University School of Medicine, Atlanta, Georgia, USA

[3] Microbiology and Molecular Genetics Program, Graduate Division of Biological and Biomedical Sciences, Laney Graduate School, Emory University, Atlanta, Georgia, USA

[#] **Authors contributed equally. The order was determined by seniority**.

[*]Corresponding author: joanna.goldberg@emory.edu

Published in *Microbiology Spectrum*, Jul 18th 2022

Author contributions

EEB worked on study conceptualization and design, methodology, writing, editing and answering reviewer comments.

VR worked on data curation, analysis, validation, visualisation, writing, editing and answering reviewer comments.

JBG helped with study conceptualization and design, supervision, funding, resources, writing, editing and answering reviewer comments.

Abstract

Respiratory infections with bacterial pathogens remain the major cause of morbidity in individuals with the genetic disease, cystic fibrosis (CF). Some studies have shown that CF patients that harbour both *Staphylococcus aureus* and *Pseudomonas aeruginosa* in their lungs are at even greater risk for more severe and complicated respiratory infections and earlier death. However, the drivers for this worse clinical condition are not well understood. To investigate the interactions between these two microbes that might be responsible for their increased pathogenic potential, we obtained 28 pairs of *S. aureus* and *P. aeruginosa* from the same respiratory samples from 18 individuals with CF. We compared the survival of each *S. aureus* CF isolate cocultured with its corresponding co-infecting CF *P. aeruginosa* to when it was cocultured with non-CF laboratory strains of *P. aeruginosa*. We found that the *S. aureus* survival was significantly higher in the presence of their co-infecting *P. aeruginosa* compared to laboratory *P. aeruginosa* strains, regardless of whether the co-infecting isolate was mucoid or nonmucoid. We also tested how a non-CF *S. aureus* strain, JE2, behaved with each *P. aeruginosa* CF isolate and found that its interaction was similar to how the CF *S. aureus* isolate interacted with its co-infecting *P. aeruginosa* pair. Altogether, our work suggests that interactions between *S. aureus* and *P. aeruginosa* that promote coexistence in the CF lung are isolate-dependent and that this interaction appears to be driven mainly by *P. aeruginosa*.

Importance

Previous studies have shown that in laboratory settings, *Pseudomonas aeruginosa* generally kills *Staphylococcus aureus*. However, these bacteria are often found co-infecting the lungs of cystic fibrosis (CF) patients, which has been associated with worse patient outcomes. To investigate the interactions between these two bacteria, we competed 28 co-infection pairs obtained from the same lung samples of 18 different CF patients. We compared these results to those we previously reported of each CF *S. aureus* isolate against a non-CF laboratory strain of *P. aeruginosa*. We found that *S. aureus* survival against its corresponding co-infection *P. aeruginosa* was higher than its survival against the laboratory strain of *P. aeruginosa*. These results suggest that there may be selection for coexistence of these microbes in the CF lung environment. Further understanding of the interactions between *P. aeruginosa* and *S. aureus* will provide insights into the drivers of coexistence and their impact on the host.

Introduction

The majority of the mortality in the inherited disease cystic fibrosis (CF) is due to bacterial lung infections. It is now appreciated that these respiratory infections are polymicrobial. The most common pathogens identified by culture methods include *Pseudomonas aeruginosa*, *Staphylococcus aureus*, *Haemophilus influenzae*, *Stenotrophomonas maltophilia*, *Achromobacter* species, and the *Burkholderia cepacia* complex. Of these, *S. aureus* has taken over as the microbe most commonly isolated, while *P. aeruginosa* remains associated with the majority of the morbidity and mortality in people living with CF (1).

Studies from our group and others have shown that CF patients that have lung infections with both *S. aureus* and *P. aeruginosa* are at greater risk for more severe disease and complicated respiratory infections than those infected with either *S. aureus* or *P. aeruginosa* alone (2-4), while other studies have shown no difference in the clinical outcomes between CF patients infected with *P. aeruginosa* alone vs. those co-infected with *P. aeruginosa* and *S. aureus* (2, 5, 6). Differences in the patient cohorts as well as the nature of the isolates themselves have been suggested as potential reasons for these disparate findings. However, it remains poorly understood how these species can coexist (i.e., survive together in the same environment) in the CF lung despite studies from our lab and many others showing that *S. aureus* is typically killed when cocultured with *P. aeruginosa in vitro* (7-10).

To begin to address this question, we examined a collection of *S. aureus* isolates from respiratory samples obtained from the Emory Cystic Fibrosis Biospecimen Registry.

We previously reported the outcomes of competition between these CF *S. aureus* isolates and isogenic nonmucoid and mucoid variants of the laboratory *P. aeruginosa* strain PAO1 using a coculture assay developed in our laboratory. We categorised these CF *S. aureus* isolates based on the competition outcomes: Killed by nonmucoid PAO1 but not mucoid PAO1, killed by both, or killed by neither. However, it is not known how these CF *S. aureus* fare against *P. aeruginosa* isolates that were present in the same CF respiratory sample – hereon referred to as "co-infection pairs".

In this study, we competed 28 co-infection pairs of *S. aureus* and *P. aeruginosa* against each other. These isolates were obtained from the respiratory samples of 18 different CF patients. We also compared the survival of the co-infection pairs in competition against the previously reported outcomes of each CF *S. aureus* isolate against mucoid and nonmucoid PAO1. We found that *S. aureus* survival against its corresponding co-infection *P. aeruginosa* pair was higher than its survival against a non-CF laboratory *P. aeruginosa*. This was true regardless of the *P. aeruginosa* mucoid status, suggesting possible adaptation between these microbes in the CF lung environment. Moreover, we found that survival of non-CF *S. aureus* strain JE2 was comparable to that of CF *S. aureus* when competed against CF *P. aeruginosa*. This suggests that *P. aeruginosa* primarily drives the coexistence of these two microbes. These findings set the stage for future studies that will dissect the mechanisms that allow both microbes to survive together in the CF lung.

Materials and Methods

Bacterial strains

All bacterial isolates used in this study were obtained from patients enrolled in the Emory Cystic Fibrosis Biospecimen Registry (CFBR) (**Table 1**). The *S. aureus* isolates have been previously described, sequenced (11), and characterised (10); their previous reported interaction with mucoid and nonmucoid *P. aeruginosa* PAO1 is included in **Table S1**. *S. aureus* JE2 is a USA300 derivative (12). *P. aeruginosa* isolates were obtained from the same clinical samples. The mucoid phenotype of *P. aeruginosa* was assessed by visualisation after overnight growth on Lysogeny broth (LB) agar and *Pseudomonas* Isolation Agar (PIA; BD Difco) at 37°C. The *P. aeruginosa* co-infection isolates have also been sequenced; the draft assemblies and the raw Illumina reads have been deposited in NCBI and are available under BioProject accession number PRJNA776003.

Coculture assay

We performed a quantitative coculture assay previously described in detail (10). We grew isolates of interest overnight at 37°C in LB from single colonies, taken from PIA for *P. aeruginosa* and *Staphylococcus* Isolation Agar (SIA; TSA BD BBL with 7.5% NaCl) for *S. aureus*. These cultures were back-diluted to an optical density of 0.05 and mixed in a 1:1 ratio, or with sterile LB as monoculture controls; 10 µL of each mixture was placed onto a 0.45 µm Millipore filter (Millipore-MM_NF-HAWP02500) on a TSA plate (BD BBL) and incubated at 37°C for 24 hours. After incubation, filters were removed using sterile forceps and the bacteria were resuspended in 1.5 mL of sterile LB before serial dilution in LB and plating onto PIA and SIA. After incubation at 37°C overnight, colonies were counted and colony forming units (CFU) per mL was calculated. The fold change of *S. aureus* CFU/mL was calculated by dividing the CFU/mL of *S. aureus* (either CF isolate or JE2 control) grown with *P. aeruginosa* (either CF isolate or nonmucoid/mucoid PAO1) over the CFU/mL of each *S. aureus* isolate grown in monoculture (**Figure S1**). The fold change of *P. aeruginosa* CFU/mL was calculated by dividing the CFU/mL of *P. aeruginosa* (either CF isolate or nonmucoid/mucoid PAO1 control) grown with *S. aureus* (either CF isolate or JE2 control) over the CFU/mL of each *P. aeruginosa* isolate grown in monoculture. All coculture experiments were performed in technical duplicates and at least three biological replicates. Average CFU/mL for each biological replicate was calculated from the two technical replicates and this average was used to calculate the CFU/mL fold change for each biological replicate. Average CFU/mL fold change was calculated across all biological replicates for each coculture group and these data are represented in boxplots. To ensure consistency, *S. aureus* JE2 paired with PAO1 (both mucoid and nonmucoid) was included as a control in each assay. We observed a JE2 CFU/mL fold change of ~10^{-1} when cocultured with mucoid PAO1 and ~10^{-3} to 10^{-4} when cocultured with nonmucoid PAO1 with high reproducibility.

Statistical analysis

The CFU/mL fold change values for the groups of co-infection pairs were tested for normality using the Shapiro-wilk test. p values < 0.05 were considered non-normal distributions. The CFU/mL fold change values were then statistically compared using the Welch’s t-test or the Wilcoxon rank sum test depending on whether or not the data were normally distributed and p values <0.05 were considered statistically significant. Statistical tests were performed using the shapiro.test, t.test and wilcox.test function in R. Welch's t-test with false discovery rate correction was used to compare all individual co-infection pairs using the pairwise_t_test function from the rstatix package (**Table S2**).

Results

S. aureus survives better with its co-infecting CF *P. aeruginosa*

To determine the interaction between co-infection pairs, we performed coculture experiments on *S. aureus* isolates with *P. aeruginosa* isolates that were obtained from the same respiratory sample. We calculated the CFU/mL fold change for *S. aureus* grown in the presence of its co-infecting *P. aeruginosa* isolate compared to *S. aureus* in monoculture (**Table 1**). We then compared this data to what we had previously obtained for these same *S. aureus* isolates in the presence of *P. aeruginosa* strain PAO1 (10). Since our previous studies had determined that *S. aureus* survived better in the presence of mucoid *P. aeruginosa* compared to nonmucoid *P. aeruginosa* (10), we separated our analysis depending on whether the co-infecting *P. aeruginosa* isolate was mucoid or nonmucoid.

We compared the CFU/mL fold change of CF *S. aureus* cocultured with their mucoid co-infection partner *P. aeruginosa* ("CF Sa vs CF Pa") to the CFU/mL fold change of the same CF *S. aureus* cocultured with the non-CF mucoid PAO1 ("CF Sa vs. Mucoid PAO1") (**Figure 1A**, left "mucoid" panel**,** p=5.089e-11). Similarly, we compared the CFU/mL fold change of CF *S. aureus* cocultured with their nonmucoid co-infection partner *P. aeruginosa* ("CF Sa vs. CF Pa") to the CFU/mL fold change of the same CF *S. aureus* cocultured with the non-CF nonmucoid PAO1 ("CF Sa vs Nonmucoid PAO1") (**Figure 1A**, right "nonmucoid" panel**,** p=1.847e-05). As seen in each panel in **Figure 1A**, the data showed the "CF Sa vs. CF Pa" survival was significantly higher than the "CF Sa vs. Mucoid/Nonmucoid PAO1" survival, indicating that the CF *S. aureus* isolates survived better when cocultured with their co-infecting *P. aeruginosa* (overall $p < 0.05$).

Patient Information		*S. aureus*	*P. aeruginosa*		CFU/mL fold change of Sa with Pa
Patient ID	Date of collection	Isolate name	Isolate name	Mucoidy	
102	4/24/2012	Sa_CFBR_17	CFBR102_Pae_20120424_S_Pa38	mucoid	7.47E-01
105	10/25/2011	Sa_CFBR_29	CFBR105_Pae_20111025_S_EBPa06	mucoid	9.08E-01
			CFBR105_Pae_20111025_S_EBPa07	mucoid	9.13E-01
	1/17/2012	Sa_CFBR_30	CFBR105_Pae_20120117_S_EBPa09	mucoid	7.50E-01
	4/16/2012	Sa_CFBR_31	CFBR105_Pae_20120416_S_EBPa11	mucoid	7.93E-01
	6/27/2012	Sa_CFBR_32	CFBR105_Pae_20120627_S_EBPa13	mucoid	9.16E-01
	8/2/2012	Sa_CFBR_33	CFBR105_Pae_20120802_S_EBPa15	mucoid	7.45E-01
120	6/27/2012	Sa_CFBR_18	CFBR120_Pae_20120627_S_Pa41	nonmucoid	1.01E+00
123	2/22/2012	Sa_CFBR_19	CFBR123_Pae_20120222_S_Pa44	nonmucoid	9.74E-03
			CFBR123_Pae_20120222_S_Pa43	mucoid	3.61E-01
134	3/26/2012	Sa_CFBR_10	CFBR134_Pae_20120326_S_Pa20	nonmucoid	5.13E-01
			CFBR134_Pae_20120326_S_Pa19	mucoid	1.16E+00
149	6/27/2012	Sa_CFBR_20	CFBR149_Pae_20120627_S_Pa45	mucoid	5.97E-01
152	1/25/2012	Sa_CFBR_06	CFBR152_Pae_20120125_S_Pa14	mucoid	3.27E-01
170	2/1/2012	Sa_CFBR_07	CFBR170_Pae_20120201_S_Pa15	mucoid	1.04E+00
171	2/8/2012	Sa_CFBR_23	CFBR171_Pae_20120208_S_Pa84	nonmucoid	1.08E+00
196	2/21/2012	Sa_CFBR_08	CFBR196_Pae_20120221_S_Pa17	mucoid	9.64E-01
201	1/17/2012	Sa_CFBR_24	CFBR201_Pae_20120117_S_Pa80	nonmucoid	4.15E-01
			CFBR201_Pae_20120117_S_Pa81	nonmucoid	6.15E-01
			CFBR201_Pae_20120117_S_Pa82	mucoid	5.04E-01
219	5/29/2012	Sa_CFBR_09	CFBR219_Pae_20120529_S_Pa18	mucoid	6.47E-01
309	5/10/2017	Sa_CFBR_37	CFBR309_Pae_20170510_S_EBPa20	nonmucoid	3.66E-03
336	4/5/2017	SA_CFBR_08	CFBR336_Pae_20170405_S_EBPa24	mucoid	2.54E-01
447	4/5/2017	Sa_CFBR_43	CFBR447_Pae_20170405_S_EBPa28	mucoid	1.55E-04
509	5/25/2017	Sa_CFBR_46	CFBR509_Pae_20170525_S_EBPa32	nonmucoid	2.85E-04
515	2/17/2017	Sa_CFBR_47	CFBR515_Pae_20170217_S_EBPa34	nonmucoid	1.47E+00
530	4/5/2017	Sa_CFBR_48	CFBR530_Pae_20170405_S_EBPa36	nonmucoid	1.13E+00
			CFBR530_Pae_20170405_S_EBPa37	mucoid	2.02E+00

Table 1: Survival of *S. aureus* (Sa) isolates when cocultured with concurrently isolated *P. aeruginosa* (Pa), grouped by patient ID.

Fold change was calculated as described in Materials and Methods. Date of sample isolation and mucoid status of *P. aeruginosa* isolate is also indicated.

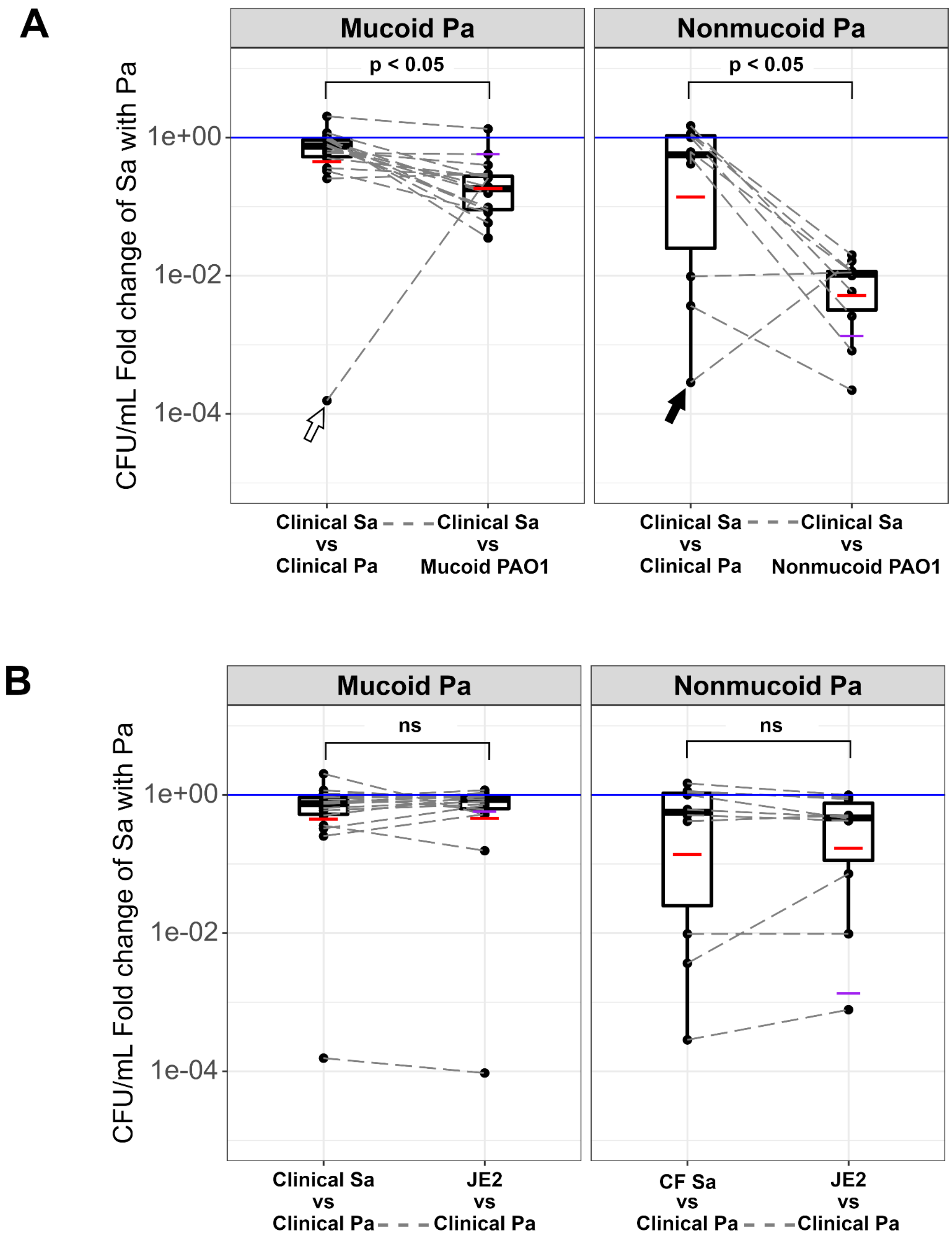

Figure 1: ***S. aureus*** **(Sa) survives better with its co-infecting cystic fibrosis (CF)** ***P. aeruginosa*** **(Pa).**
CFU/mL fold change of *S. aureus* when cocultured with *P. aeruginosa* was determined as described in Materials and Methods. Purple horizontal line shows the CFU/mL fold change of the reference *S. aureus* strain JE2 when cocultured with mucoid *P. aeruginosa* PAO1 (left panels in A and B) or nonmucoid *P. aeruginosa* PAO1 (right panels in A and B). Black horizontal line inside the boxplot shows the median and red horizontal line shows the mean. The white boxes represent the interquartile range (IQR) and the whiskers represent values up to 1.5x the first or third quartile. Values larger or smaller than 1.5 x IQR are represented by black dots. Blue solid line shows a fold change of 1 suggesting no change when grown with *P. aeruginosa* compared to monoculture. (**A**): Boxplot of CFU/mL fold change of cystic fibrosis (CF) *S. aureus* cocultured with its concurrently isolated CF *P. aeruginosa* or mucoid/nonmucoid PAO1. Dots represent average CFU/mL fold change of each *S. aureus* isolate and the grey dashed lines connect dots that correspond to the same *S. aureus* isolate. Wilcoxon signed rank test showed significant difference between the mean CFU/mL fold change of CF *S. aureus* when cocultured with CF *P. aeruginosa* compared to the mean CFU/mL fold change of CF *S. aureus* when cocultured with mucoid (p=5.089e-11, Shapiro-wilk p=0.001) or nonmucoid PAO1 (p=1.847e-05, Shapiro-wilk p=3.648e-05). Arrows represent outliers, as described in text. (**B**) Boxplot of CFU/mL fold change of CF *S. aureus* or reference strain JE2 cocultured with its concurrently isolated CF mucoid/nonmucoid *P. aeruginosa*. Dots represent average CFU/mL fold change of each *S. aureus* isolate and the grey dashed lines connect dots that correspond to the same *P. aeruginosa* isolate. Wilcoxon signed rank test/Welch's t test showed no significant difference between the mean CFU/mL fold change of CF *S. aureus* when cocultured with CF *P. aeruginosa* compared to the mean CFU/mL fold change of reference strain JE2 when cocultured with CF *P. aeruginosa* (p =0.26/0.25, for mucoid/nonmucoid, respectively, Shapiro-wilk p=0.013/0.078, ns = not significant). Average fold change was calculated from at least three biological replicates (see Table S1 for raw data).

To distinguish whether the increase in *S. aureus* survival was due to reduced killing by *P. aeruginosa* or increased resistance by *S. aureus*, we measured the survival of non-CF *S. aureus* JE2 against each CF *P. aeruginosa* isolate. We calculated the CFU/mL fold change of JE2 in coculture with CF *P. aeruginosa*, as described above. We then compared the survival of JE2 against CF *P. aeruginosa* with the survival of the co-infecting CF *S. aureus* against the same CF *P. aeruginosa*. We found no significant difference in the response shown by the CF and non-CF *S. aureus* to the CF *P. aeruginosa*. This was true regardless of whether the *S. aureus* strains were tested against mucoid or nonmucoid *P. aeruginosa* (**Figure 1B,** p=0.26 for "mucoid", p=0.25 for "nonmucoid"). These results suggested that the increased survival of CF *S. aureus* may be driven by reduced killing by *P. aeruginosa*, as the CF-adapted *P. aeruginosa* showed reduced killing of even a non-CF *S. aureus* strain.

Mucoid and nonmucoid *P. aeruginosa* isolates were collected concurrently from Patients 123, 134, 201, and 530 (**Table 1**). Previous studies had noted that mucoid *P. aeruginosa* strains were more permissive than nonmucoid isolates to *S. aureus* (13). Interestingly, we only found this to be the case for *P. aeruginosa* isolates from Patient 123: as expected the mucoid isolate from this patient was more permissive than the nonmucoid isolate when cocultured with their co-infecting *S. aureus* isolate. On the other hand, mucoid and nonmucoid isolates that were collected concurrently from Patients 134, 201, and 530 seemed to show similar results to one another; all seemed to promote coexistence (**Table 1**).

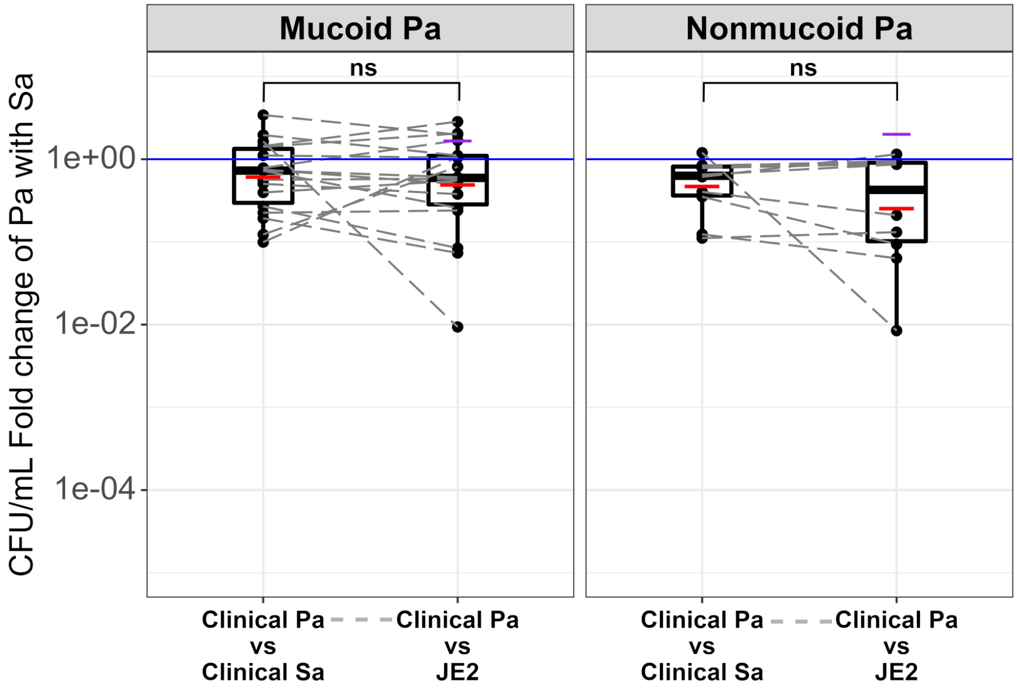

Figure 2: ***P. aeruginosa*** **(Pa) survives similarly with its co-infecting cystic fibrosis (CF)** ***S. aureus*** **(Sa) and JE2.**

CFU/mL fold change of *P. aeruginosa* when cocultured with *S. aureus* was determined as described in Materials and Methods. Purple horizontal line shows the CFU/mL fold change of mucoid *P. aeruginosa* PAO1 or nonmucoid PAO1 when cocultured with *S. aureus* JE2. Boxplot of CFU/mL fold change of mucoid and nonmucoid CF *P. aeruginosa* cocultured with its concurrently isolated CF *S. aureus* or reference strain JE2. Black horizontal line inside the boxplot shows the median and the red horizontal line shows the mean. The white boxes represent the interquartile range (IQR) and the whiskers represent values up to 1.5x the first or third quartile. Values larger or smaller than 1.5 x IQR are represented by black dots. Blue solid line shows a fold change of 1. Dots represent average CFU/mL fold change of each *P. aeruginosa* isolate, and the grey dashed lines connect dots that correspond to the same *P. aeruginosa* isolate. Wilcoxon signed rank test showed no significant difference between the mean CFU/mL fold change of CF *P. aeruginosa* when cocultured with its concurrently isolated CF *S. aureus* compared to the mean CFU/mL fold change of CF *P. aeruginosa* when cocultured with reference strain JE2 (p=0.88/p=0.19, for mucoid/nonmucoid, respectively. Shapiro-wilk p=4.533e-13/ 0.0001228, ns = not significant).

We did observe a few outliers in **Figure 1A**. In the left panel, the white arrow highlights the data related to Patient 447: Sa_CFBR_43 vs. CFBR447_Pae_20170405_EBPa28. In the right panel, the black arrow highlights the data related to Patient 509: Sa_CFBR_46 vs. CFBR509_Pae_20170525_EBPa32. Both these *S. aureus* isolates were killed more readily by their co-infecting pair. The two *P. aeruginosa* isolates were also able to readily kill the reference *S. aureus* strain JE2 (comparing **Figure 1A** and **Figure 1B** and **Figure S1**). These isolates are being investigated further.

To determine whether *P. aeruginosa* and *S. aureus* co-infecting isolates were specifically coevolving together to promote coexistence, we performed coculture experiments with non-co-infecting isolates. We chose 3 *P. aeruginosa* isolates (2 nonmucoid and 1 mucoid) and cocultured them with 4 different *S. aureus* isolates from different patients and calculated the CFU/mL fold change of *S. aureus*. For these studies we did not choose any of the outlier *P. aeruginosa* or *S. aureus* isolates (**Figure 1A**, white or black arrow). We found that the two nonmucoid strains (**Figure S2, panel A and B**) showed the same level of killing of the non-co-infecting *S. aureus* as they did with their co-infecting isolate. Interestingly, this was independent of whether the non-co-infecting *S. aureus* was killed by its own co-infection isolate. On the other hand, we noted that the mucoid *P. aeruginosa* isolate (**Figure S2, panel C**) was able to kill non-co-infecting *S. aureus* isolates, even though these *S. aureus* isolates coexisted with their respective co-infection isolates, as did the *S. aureus* isolate co-infecting with this mucoid *P. aeruginosa*. (**Figure S2, Table S3**). This suggests that coexistence may also be affected by specific isolate-dependent interactions.

P. aeruginosa survives similarly with its co-infecting CF *S. aureus* as it does with JE2

While *P. aeruginosa* has not been previously found to be negatively impacted by *S. aureus*, we also tested the survival of *P. aeruginosa* with its co-infecting *S. aureus* as well as with JE2 (**Table S1**). As seen in Figure 2, most *P. aeruginosa* isolates survived similarly in the presence of their co-infecting *S. aureus* isolate compared to their survival in the presence of JE2. This happened regardless of whether the *P. aeruginosa* was mucoid (left hand panel; **Figure 2**; p=0.88) or nonmucoid (right hand panel; **Figure 2**; p=0.19). This indicated that there was little effect on survival of *P. aeruginosa* by coculture of the *S. aureus* under the conditions of this assay.

Discussion

Multiple studies have shown that CF patients co-infected with both *S. aureus* and *P. aeruginosa* are at greater risk for more severe and complicated respiratory infections (2-4, 6); however, the mechanisms responsible for these outcomes are not well understood. To uncover the reason for the worsening clinical manifestation, the processes allowing these two microbes to survive together need to be better understood. Various studies have shown different stages of growth and environmental conditions including media and planktonic vs. biofilm modes of growth can promote the coexistence of *S. aureus* and *P. aeruginosa* (9, 14, 15). In some other cases, it has been found that bacterial segregation promotes survival (15). On the other hand, many *in vitro* studies have shown *P. aeruginosa* itself or *P. aeruginosa* factors, such as secreted LasA and rhamnolipids, can lyse or kill *S. aureu*s (7, 16-19). We and others have previously observed decreased expression of some of these factors in the context of mucoid conversion of *P. aeruginosa* promotes coexistence with *S. aureus* (13, 20). Some other studies have noted the physiological conditions that allow *S. aureus* and *P. aeruginosa* to survive and grow together (8, 21, 22). Many of the studies to uncover the mechanism of competition or coexistence have utilised laboratory isolates, however more recently investigations have been performed with *S. aureus* and *P. aeruginosa* clinical isolates (23-26).

Our goal here was to add to this growing list of studies by investigating pairs of clinical isolates of these bacteria obtained from the same patient sample on the same day. By studying paired, particularly longitudinal isolates, we hoped to glean insights into novel mechanisms of interactions between these two pathogens. We examined 28 pairs of isolates obtained from 18 CF individuals; 5 of these people provided multiple samples longitudinally. We hypothesised that isolates of *S. aureus* would survive better with *P. aeruginosa* obtained concurrently compared to a typical *P. aeruginosa* laboratory strain. And any *P. aeruginosa* or *S. aureus* that behaved differently could be a source for future comparative studies to identify potential mechanisms of coexistence.

Overall, our data generally supported our hypothesis: we showed that CF *S. aureus* isolates survive better with their co-infecting *P. aeruginosa* isolates compared to *P. aeruginosa* PAO1. We also separated our data based on the mucoid status of *P. aeruginosa* isolates in this study (mucoid or nonmucoid) since we know that this phenotype impacts the interaction with *S. aureus* (13). We noted that the difference in survival was more pronounced when comparing the interaction between *S. aureus* and the nonmucoid *P. aeruginosa* isolates vs. *S. aureus* and the mucoid *P. aeruginosa* isolates (**Figure 1A**). This suggests, as has been previously shown, that mucoidy itself is already an adaptation that facilitates coexistence (13). We also observed no difference in the interaction of these co-infecting pairs in our longitudinal samples (all coexisted). Interestingly, when the *S. aureus* reference strain JE2 was cocultured with these *P. aeruginosa* CF isolates, it showed equivalent susceptibility to *P. aeruginosa* killing as the co-infecting *S. aureus* isolate (**Figure 1B**). Thus, these results are not perfectly aligned with our original hypothesis as the reference *S. aureus* strain was not from co-infection, which has led us to conclude that *P. aeruginosa* is the main driver of this coexistence, as has been suggested by previous studies from our lab and others (13, 27). Moreover, we competed non-co-infecting CF isolates of *P. aeruginosa* and *S.*

aureus, we found that *S. aureus* can either be killed by or coexist with *P. aeruginosa* regardless of whether or not the two isolates are co-infection pairs (**Figure S2**). This suggests that coexistence is isolate-dependent and while *P. aeruginosa* may be the main driver of coexistence, *S. aureus* also plays a role.

The two observed outlier *S. aureus* and *P. aeruginosa* co-infection pairs in **Figure 1A** (white and black arrow) are currently being investigated. The fact that these two *P. aeruginosa* strains are able to kill both their corresponding co-infection *S. aureus* partner as well as JE2 supports the idea that *P. aeruginosa* drives the interaction. In addition, one of these outlier *S. aureus* isolates (SA_CFBR_43) may have *P. aeruginosa* strain PAO1-specific resistance mechanisms according to our previous study (10).

We are aware that our study has its limitations. While the *S. aureus* and *P. aeruginosa* were obtained from the same clinical sample, the interactions we are examining are all *in vitro* and our assay, by design, promotes the interaction between these two different species. Also, we only examined individual isolates that had been retrieved by the clinical microbiology laboratory. We know that *P. aeruginosa* is phenotypically and genotypically heterogeneous in this environment (28, 29) and some recent studies have also suggested that *S. aureus* may be similarly heterogenous (6, 26, 30, 31). Thus, the single isolates that we examined may only represent a subset of the genotypes/phenotypes present in the respiratory sample. Currently we are obtaining panels and pools of isolates from clinical CF samples to determine the genotypic and phenotypic variability and their impact on coexistence. Thus, whether and how these genotypes/phenotypes correlate with the clinical status of a person with CF at the time the sample was collected will be an important area for future investigations.

It is also the case that *S. aureus* and *P. aeruginosa* are not the only inhabitants in the CF lung, and that other microbes might impact the interactions of these two bacteria. However even with these recognized shortcomings, our study supports the hypothesis that *S. aureus* and *P. aeruginosa* isolated from the same CF respiratory sample have adapted to promote their coexistence within the CF lung. And since co-infection is a more deadly situation for people living with CF, understanding what drives *S. aureus-P. aeruginosa* coexistence could allow us to devise ways of disrupting this interaction to improve patients' prognosis.

Supplementals

Supplemental tables and figures for this chapter can be found in the manuscript https://journals.asm.org/doi/10.1128/spectrum.00976-22

Acknowledgments

Bacterial isolates were provided by the CF Biospecimen Repository (CFBR) at the Children's Healthcare of Atlanta and Emory University CF Discovery Core. We thank Dr. Arlene Stecenko in her role as Director of the CFBR and Katy Clemmer as Microbiology Research Coordinator of the CFBR for their assistance identifying appropriate bacterial isolates. We thank Anika Sinka for technical assistance. We thank Dr. Timothy Read, Ashley Alexander, Hui Qi Loo, Dr. Dina Moustafa, Justin Luu, Dr. Cristian Crisan, Dr. Rebecca Duncan, Rachel Done, Christine Knox, Emily Wissel, Dr. Robert Petit, Dr. Andrei Bombin and Brianna Bixler for providing constructive comments on the manuscript. We also thank the NIH IRACDA Fellowships in Research and Science Teaching (FIRST) program at Emory for financial support to E.E.B.

(K12-GM000680). This work was supported by grants from the Cystic Fibrosis Foundation (GOLDBE19P0) and the NIH (R21-AI48847) to J.B.G.

References

1. Cystic Fibrosis Foundation Patient Registry 2020 Annual Data Report Bethesda, MD 2021.
2. Hubert D, Reglier-Poupet H, Sermet-Gaudelus I, Ferroni A, Le Bourgeois M, Burgel PR, Serreau R, Dusser D, Poyart C, Coste J. Association between *Staphylococcus aureus* alone or combined with *Pseudomonas aeruginosa* and the clinical condition of patients with cystic fibrosis. J Cyst Fibros. 2013;12(5):497-503. Epub 2013/01/08. doi: 10.1016/j.jcf.2012.12.003. PubMed PMID: 23291443.
3. Limoli DH, Yang J, Khansaheb MK, Helfman B, Peng L, Stecenko AA, Goldberg JB. *Staphylococcus aureus* and *Pseudomonas aeruginosa* co-infection is associated with cystic fibrosis-related diabetes and poor clinical outcomes. Eur J Clin Microbiol Infect Dis. 2016;35(6):947-53. Epub 2016/03/20. doi: 10.1007/s10096-016-2621-0. PubMed PMID: 26993289.
4. Maliniak ML, Stecenko AA, McCarty NA. A longitudinal analysis of chronic MRSA and *Pseudomonas aeruginosa* co-infection in cystic fibrosis: A single-center study. J Cyst Fibros. 2016;15(3):350-6. Epub 2015/11/28. doi: 10.1016/j.jcf.2015.10.014. PubMed PMID: 26610860.
5. Ahlgren HG, Benedetti A, Landry JS, Bernier J, Matouk E, Radzioch D, Lands LC, Rousseau S, Nguyen D. Clinical outcomes associated with *Staphylococcus aureus* and *Pseudomonas aeruginosa* airway infections in adult cystic fibrosis patients. BMC Pulm Med. 2015;15:67. Epub 2015/06/22. doi: 10.1186/s12890-015-0062-7. PubMed PMID: 26093634; PMCID: PMC4475617.
6. Briaud P, Bastien S, Camus L, Boyadjian M, Reix P, Mainguy C, Vandenesch F, Doleans-Jordheim A, Moreau K. Impact of coexistence phenotype between *Staphylococcus aureus* and *Pseudomonas aeruginos*a isolates on clinical outcomes among cystic fibrosis patients. Front Cell Infect Microbiol. 2020;10:266. Epub 2020/06/26. doi: 10.3389/fcimb.2020.00266. PubMed PMID: 32582568; PMCID: PMC7285626.
7. Nguyen AT, Jones JW, Ruge MA, Kane MA, Oglesby-Sherrouse AG. Iron depletion enhances production of antimicrobials by *Pseudomonas aeruginosa*. J Bacteriol. 2015;197(14):2265-75. Epub 2015/04/29. doi: 10.1128/JB.00072-15. PubMed PMID: 25917911; PMCID: PMC4524187.
8. Filkins LM, Graber JA, Olson DG, Dolben EL, Lynd LR, Bhuju S, O'Toole GA. Coculture of *Staphylococcus aureus* with *Pseudomonas aeruginosa* drives *S. aureus* towards fermentative metabolism and reduced viability in a cystic fibrosis model. J Bacteriol. 2015;197(14):2252-64. Epub 2015/04/29. doi: 10.1128/JB.00059-15. PubMed PMID: 25917910; PMCID: PMC4524177.
9. Tognon M, Kohler T, Luscher A, van Delden C. Transcriptional profiling of *Pseudomonas aeruginosa* and *Staphylococcus aureus* during *in vitro* co-culture. BMC Genomics. 2019;20(1):30. Epub 2019/01/12. doi: 10.1186/s12864-018-5398-y. PubMed PMID: 30630428; PMCID: PMC6327441.
10. Bernardy EE, Petit RA, 3rd, Raghuram V, Alexander AM, Read TD, Goldberg JB. Genotypic and phenotypic diversity of *Staphylococcus aureus* isolates from cystic

fibrosis patient lung infections and their interactions with *Pseudomonas aeruginosa*. mBio. 2020;11(3). Epub 2020/06/25. doi: 10.1128/mBio.00735-20. PubMed PMID: 32576671; PMCID: PMC7315118.

11. Bernardy EE, Petit RA, 3rd, Moller AG, Blumenthal JA, McAdam AJ, Priebe GP, Chande AT, Rishishwar L, Jordan IK, Read TD, Goldberg JB. Whole-genome sequences of *Staphylococcus aureu*s isolates from cystic fibrosis lung infections. Microbiol Resour Announc. 2019;8(3). Epub 2019/01/29. doi: 10.1128/MRA.01564-18. PubMed PMID: 30687841; PMCID: PMC6346173.
12. Fey PD, Endres JL, Yajjala VK, Widhelm TJ, Boissy RJ, Bose JL, Bayles KW. A genetic resource for rapid and comprehensive phenotype screening of nonessential *Staphylococcus aureus* genes. mBio. 2013;4(1):e00537-12. Epub 2013/02/14. doi: 10.1128/mBio.00537-12. PubMed PMID: 23404398; PMCID: PMC3573662.
13. Limoli DH, Whitfield GB, Kitao T, Ivey ML, Davis MR, Jr., Grahl N, Hogan DA, Rahme LG, Howell PL, O'Toole GA, Goldberg JB. *Pseudomonas aeruginosa* alginate overproduction promotes coexistence with *Staphylococcus aureus* in a model of cystic fibrosis respiratory infection. MBio. 2017;8(2). Epub 2017/03/23. doi: 10.1128/mBio.00186-17. PubMed PMID: 28325763; PMCID: PMC5362032.
14. Cendra MDM, Blanco-Cabra N, Pedraz L, Torrents E. Optimal environmental and culture conditions allow the *in vitro* coexistence of *Pseudomonas aeruginosa* and *Staphylococcus aureus* in stable biofilms. Sci Rep. 2019;9(1):16284. Epub 2019/11/11. doi: 10.1038/s41598-019-52726-0. PubMed PMID: 31705015; PMCID: PMC6841682.
15. Barraza JP, Whiteley M. A Pseudomonas aeruginosa antimicrobial affects the biogeography but not Fitness of *Staphylococcus aureus* during coculture. mBio. 2021;12(2). Epub 2021/04/01. doi: 10.1128/mBio.00047-21. PubMed PMID: 33785630; PMCID: PMC8092195.
16. Machan ZA, Taylor GW, Pitt TL, Cole PJ, Wilson R. 2-Heptyl-4-hydroxyquinoline N-oxide, an antistaphylococcal agent produced by *Pseudomonas aeruginosa*. J Antimicrob Chemother. 1992;30(5):615-23. Epub 1992/11/01. PubMed PMID: 1493979.
17. Kessler E, Safrin M, Olson JC, Ohman DE. Secreted LasA of *Pseudomonas aeruginosa* is a staphylolytic protease. J Biol Chem. 1993;268(10):7503-8. Epub 1993/04/05. PubMed PMID: 8463280.
18. Haba E, Pinazo A, Jauregui O, Espuny MJ, Infante MR, Manresa A. Physicochemical characterization and antimicrobial properties of rhamnolipids produced by *Pseudomonas aeruginosa* 47T2 NCBIM 40044. Biotechnol Bioeng. 2003;81(3):316-22. Epub 2002/12/11. doi: 10.1002/bit.10474. PubMed PMID: 12474254.
19. Hotterbeekx A, Kumar-Singh S, Goossens H, Malhotra-Kumar S. *in vivo* and *in vitro* interactions between *Pseudomonas aeruginosa* and *Staphylococcus* spp. Front Cell Infect Microbiol. 2017;7:106. Epub 2017/04/20. doi: 10.3389/fcimb.2017.00106. PubMed PMID: 28421166; PMCID: PMC5376567.
20. Price CE, Brown DG, Limoli DH, Phelan VV, O'Toole GA. Exogenous alginate protects *Staphylococcus aureus* from killing by *Pseudomonas aeruginosa*. J Bacteriol. 2020;202(8). Epub 2019/12/04. doi: 10.1128/JB.00559-19. PubMed PMID: 31792010; PMCID: PMC7099135.

21. Hoffman LR, Deziel E, D'Argenio DA, Lepine F, Emerson J, McNamara S, Gibson RL, Ramsey BW, Miller SI. Selection for *Staphylococcus aureus* small-colony variants due to growth in the presence of *Pseudomonas aeruginosa*. Proc Natl Acad Sci U S A. 2006;103(52):19890-5. Epub 2006/12/19. doi: 10.1073/pnas.0606756104. PubMed PMID: 17172450; PMCID: PMC1750898.
22. Camus L, Briaud P, Vandenesch F, Moreau K. How Bacterial Adaptation to Cystic fibrosis environment shapes interactions between *Pseudomonas aeruginosa* and *Staphylococcus aureus*. Front Microbiol. 2021;12:617784. Epub 2021/03/23. doi: 10.3389/fmicb.2021.617784. PubMed PMID: 33746915; PMCID: PMC7966511.
23. Baldan R, Cigana C, Testa F, Bianconi I, De Simone M, Pellin D, Di Serio C, Bragonzi A, Cirillo DM. Adaptation of *Pseudomonas aeruginosa* in cystic fibrosis airways influences virulence of *Staphylococcus aureus in vitro* and murine models of co-infection. PLoS One. 2014;9(3):e89614. Epub 2014/03/08. doi: 10.1371/journal.pone.0089614. PubMed PMID: 24603807; PMCID: PMC3945726.
24. Briaud P, Camus L, Bastien S, Doleans-Jordheim A, Vandenesch F, Moreau K. Coexistence with *Pseudomonas aeruginosa* alters *Staphylococcus aureus* transcriptome, antibiotic resistance and internalization into epithelial cells. Sci Rep. 2019;9(1):16564. Epub 2019/11/14. doi: 10.1038/s41598-019-52975-z. PubMed PMID: 31719577; PMCID: PMC6851120.
25. Camus L, Briaud P, Bastien S, Elsen S, Doleans-Jordheim A, Vandenesch F, Moreau K. Trophic cooperation promotes bacterial survival of *Staphylococcus aureus* and *Pseudomonas aeruginosa*. ISME J. 2020;14(12):3093-105. Epub 2020/08/21. doi: 10.1038/s41396-020-00741-9. PubMed PMID: 32814867; PMCID: PMC7784975.
26. Wieneke MK, Dach F, Neumann C, Gorlich D, Kaese L, Thissen T, Dubbers A, Kessler C, Grosse-Onnebrink J, Kuster P, Schultingkemper H, Schwartbeck B, Roth J, Nofer JR, Treffon J, Posdorfer J, Boecken JM, Strake M, Abdo M, Westhues S, Kahl BC. Association of diverse *Staphylococcus aureus* populations with *Pseudomonas aeruginosa* coinfection and inflammation in cystic fibrosis airway infection. mSphere. 2021;6(3):e0035821. Epub 2021/06/24. doi: 10.1128/mSphere.00358-21. PubMed PMID: 34160233.
27. Millette G, Langlois JP, Brouillette E, Frost EH, Cantin AM, Malouin F. Despite antagonism *in vitro*, *Pseudomonas aeruginosa* enhances *Staphylococcus aureus* colonization in a murine lung infection model. Front Microbiol. 2019;10:2880. Epub 2020/01/11. doi: 10.3389/fmicb.2019.02880. PubMed PMID: 31921058; PMCID: PMC6923662.
28. Mowat E, Paterson S, Fothergill JL, Wright EA, Ledson MJ, Walshaw MJ, Brockhurst MA, Winstanley C. *Pseudomonas aeruginosa* population diversity and turnover in cystic fibrosis chronic infections. Am J Respir Crit Care Med. 2011;183(12):1674-9. Epub 2011/02/08. doi: 10.1164/rccm.201009-1430OC. PubMed PMID: 21297072.
29. Darch SE, McNally A, Harrison F, Corander J, Barr HL, Paszkiewicz K, Holden S, Fogarty A, Crusz SA, Diggle SP. Recombination is a key driver of genomic and phenotypic diversity in a *Pseudomonas aeruginosa* population during cystic fibrosis infection. Sci Rep. 2015;5:7649. Epub 2015/01/13. doi: 10.1038/srep07649. PubMed PMID: 25578031; PMCID: PMC4289893.

30. Azarian T, Ridgway JP, Yin Z, David MZ. Long-term intrahost evolution of methicillin resistant *Staphylococcus aureus* among cystic fibrosis patients with respiratory carriage. Front Genet. 2019;10:546. Epub 2019/06/28. doi: 10.3389/fgene.2019.00546. PubMed PMID: 31244886; PMCID: PMC6581716.
31. Fischer AJ, Singh SB, LaMarche MM, Maakestad LJ, Kienenberger ZE, Pena TA, Stoltz DA, Limoli DH. Sustained coinfections with *Staphylococcus aureus* and *Pseudomonas aeruginosa* in cystic fibrosis. Am J Respir Crit Care Med. 2021;203(3):328-38. Epub 2020/08/05. doi: 10.1164/rccm.202004-1322OC. PubMed PMID: 32750253; PMCID: PMC7874317.

Chapter VI – Conclusions and future directions

In this dissertation, I examined the genetic diversity in *S. aureus* at the species-wide scale (Macrodiversity) as well as within clouds of clonal isolates from one human body site (Microdiversity). My results provided new insights into the complex genetic landscape of this important bacterial species. I highlighted the significance of understanding subspeciation in *S. aureus*, as well as that of sampling strategies and strain diversity in accurately characterising genetic variation. The implications of my research extend to both clinical and evolutionary microbiology, providing valuable insights for future studies in these fields. I will summarise and discuss the key findings of my research in this final chapter.

Macrodiversity

As I performed the analysis for the first results chapter (**Chapter II**) I learned about the technical aspects of how to sample genome level diversity of a bacterial species. First, we started with all publicly available *S. aureus* genome sequences in NCBI as of May 2021 - ~80,000 genomes. With a series of filtration steps, we reduced the dataset down to ~8000 genomes while still maintaining as much of the total diversity as possible. We outlined an approach for uniform data processing, filtering and deduplication steps to identify possible misidentifications, contaminated genomes, and to reduce redundancy. We then used the resulting curated dataset of ~8000 genomes to build a species-wide *S. aureus* pangenome. We believe the pangenome construction approach described in **Chapter II** can be a resource for the greater microbial genomics community.

Our goals with this pangenome were to 1) Assign lineages to all isolates and identify the total number of *S. aureus* CCs or lineages that have been sampled; 2) Delineate the species wide core and accessory genome, as well as identify lineage specific genes and genes undergoing horizontal exchange between lineages; and 3) Build a complete representative dataset that can be used for answering several questions regarding *S. aureus* genome evolution and subspeciation.

Subspecies formation in *S. aureus* is hypothesised to be driven by a combination of cohesive forces driving homogenization within lineages and strong barriers driving separation between lineages. Restriction modification systems, phage host range, and CRISPR interference to name a few (1–6). We found that the species fall into natural clusters separated by genetic identity. We termed these clusters "strain groups" and found accessory genes specific to these strain groups, as well as accessory genes that were found indiscriminately across multiple strain groups. We observed that the accessory genome composition was specific to the core genome for the prevalent, established strain groups (Strain groups comprising CC8, CC5, CC30, CC45, CC22). Conversely, we also observed smaller strain groups across the phylogeny with similar accessory genome compositions. This indicated that they may be actively experiencing recombination events. We then used the 'Fixation index' or F_{ST} to estimate which genes are undergoing rapid turnover. We found specific accessory genes to be randomly distributed across the phylogeny as well as other accessory genes to be highly strain group specific, suggesting that they are fixed or nearly fixed within specific clades. Some of these strain group specific accessory genes included the *Staphylococcal* enterotoxin type G and type O (SEG &

SEO). Moreover, we also found that SEG and SEO are highly co-occurring. This is a significant finding as SEG and SEO are potent but relatively understudied superantigens (7–10). Their co-occurrence in specific strain groups suggests mechanisms driving co-selection and potentially functional dependence, making this an interesting avenue of research to follow up on. Strain groups specificity of other *Staphylococcal* toxins as well as genes associated with severe disease states such as CF, bacteremia and SSTIs can also be examined using our dataset. In addition, analysis of genome-wide co-occurrences can provide more insight into gene interactions and patterns of bacterial genome evolution (11). Furthermore, our pangenome can also serve as a database for studying the prevalence and transmission of phages, plasmids, and other mobile genetic elements (MGE). MGEs are a major source of toxins and antibiotic resistance markers in *S. aureus* and being able to detect and characterise the lineage specificity (or lack thereof) of different MGEs can serve as a powerful tool for epidemiological investigations. Finally, estimating genome-wide dN/dS ratios, pseudogenization, and gene turnover rates can identify hotspots in the genome undergoing selection, provide evidence for adaptive changes, and highlight genes contributing to long-term fitness (12–14). Overall, I believe this study has established the groundwork for further research into uncovering the past and future evolutionary trajectories of *S. aureus*.

In **Chapter III**, we found more evidence that the *agr* quorum sensing operon may have played a key role in determining the evolutionary trajectory of *S. aureus*. We analysed over 40,000 *S. aureus* genome sequences and their *agr* operons and observed strong linkage between clonal complex (CC) and the *agr* group. We found that in all cases but one, isolates of given CC contained only one *agr* group. The only exception being CC45, having two *agr* groups. We also found that there are unique *agr* gene alleles for each CC, including cases where two CCs have the same *agr* group. In addition, we found that the hypervariable region of the *agr* operon forms a distinct haplotype block with conserved flanking regions. This suggests the hypervariable region is a strong candidate for recombination, however we only observed one case of potential between CC-exchange of *agr*, in CC45, as mentioned before. Though there is evidence for transfer of many genes under strong selection (e.g antibiotic resistance genes) between *S. aureus* CCs or from other *Staphylococcus* species, we did not find evidence of *S. aureus* CCs having non-cognate AIPs, nor did we find novel AIPs outside the four defined *agr* groups (15–19). This suggests strong selection for maintaining four specific *agr* groups. We did find that close relatives of *S. aureus* - *S. argenteus* and *S. schweitzeri* share the *agr* group-1 AIP though the other *Staphylococci* have their own distinct AIPs as well (20). This suggests that diversifying selection may have also driven emergence of species specific *agr* groups. Moreover, in **Chapter II**, we did not observe any accessory genes that were unique to *agr* groups (except *agrD*) independent of phylogeny. This implies that the genetic background of the CCs may prevent exchange of *agr*. This result makes the fact that we found CC45 to be the only CC having multiple *agr* groups all the more interesting, as CC45 may be in the midst of divergence/new-CC formation.

When non-cognate *agr* recombinants were engineered by Tan et al, 2022, the native *agr* dependent phenotypes (haemolysis, pigment, exoprotein secretion) were not observed, suggesting dysregulation of *agr* regulated genes (21). This dysregulation is similar to one observed in a Δ*agr* isolate, or an *agr* mutant isolate (22). Altman et al, 2018 and Giulieri et al, 2022 suggested that *agr* mutants have increased signatures of genome-wide positive selection compared to *agr*$^+$ isolates of the same genetic background (12,22). This suggests that the engineered *agr* recombinants, due to the *agr* dysregulation, may also undergo a similar level of increased positive selection. This explains the lack of naturally occurring *agr* recombinants as the transmission capabilities of *agr*$^-$ isolates are reduced. Based on these results, I hypothesise that *agr* dysregulation mediated by either non-cognate *agr* recombination or mutations, can lead to increased genome-wide mutation rates. This hypothesis can be tested by conducting an evolution experiment comparing the mutation rates of a Wild-type strain and an isogenic mutant strain having non-cognate *agr* group. (CC8 strain with CC5 *agr*).

Apart from the evolutionary implications, my work on the *agr* operon also has clinical relevance. We found that ~5% of all *agr* operons have frameshift mutations, likely rendering them non-functional. This was significantly higher than the number of frameshift mutations observed in other core genes similar to *agr*. Showing that these mutations are indeed frequent across a large set of diverse strains. This result is significant because the attenuated toxicity mediated by loss of *agr* may lead to increased persistence in chronic infection scenarios, altering antibiotic resistance and causing worse outcomes (23–29). Moreover, we found that ~50% of these *agr* frameshifts were identical mutations occurring across unrelated clonal lineages with no evidence of *agr* gene exchange. This indicated convergent evolution of frameshift mutations, suggesting they may indeed be adaptive. In addition, we found that these *agr* mutants did not survive long-term to establish a stable lineage circulating across populations. This suggested they may have traded in their ability to initiate new infections and spread in favour of increased niche-specific adaptation. This result highlights the importance of using genomics as a surveillance or diagnostic tool for detecting mutants that can potentially alter treatment strategies.

Microdiversity

Typically, genomic surveillance efforts for pathogenic bacteria use sequencing data from single or multiple independently sequenced individual colonies. However, individual colonies may not encapsulate the complete population diversity found in the host. Capturing total diversity is important as mutant subpopulations or strains with different AMR profiles may be missed while only sampling one or few individual colonies.

First, it is key to understand the factors that impact sequence diversity and our ability to detect them. To illustrate this, I simulated a mixed population by using sequencing reads from an artificial ancestral state reconstruction of *S. aureus* (Staph-ASR). I conducted *in-silico* mutagenesis of our Staph-ASR genome by randomly inducing mutations at the rate of 1/100bp, 1/1000bp and 1/10000bp. I mixed our WT Staph-ASR sequencing reads with each of the 3 mutant genome reads in different proportions. I then calculated the minor allele frequency (MAF) at each position on the mixed genome and plotted a histogram of MAF frequencies.

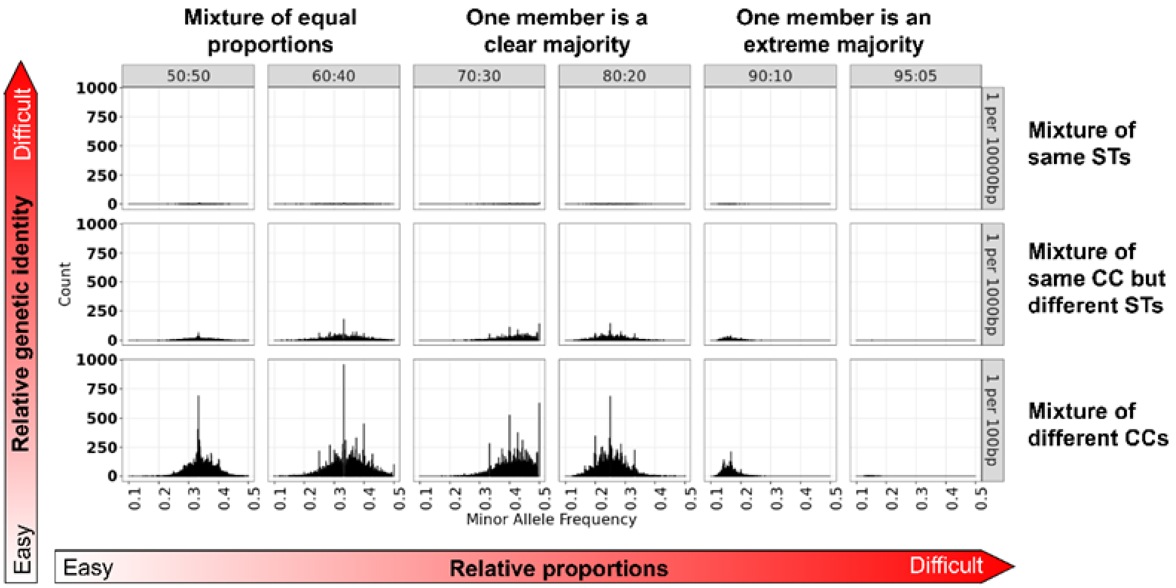

Fig 1: Relative proportions and relative genetic identity are major determinants of our ability to detect sequence mixtures

Each box contains a histogram depicting distribution of minor allele frequencies. Reads from two simulated *S. aureus* genomes were mixed in varying proportions. Both genomes are identical except one contains either 1 mutation every 10,000 bp/every 1000 bp/ every 100 bp. Random reads from the two genomes were mixed in ratios of 50:50, 60:40, 70:30, 80:20, 90:10 and 95:5 . Each row shows a varying number of mutations and each column shows the mixture proportions as mentioned in the facet headings (top or right).

This simulation highlights the difficulty scale associated with detecting population heterogeneity. Two factors – relative genetic identity between the mixed population and the ratio in which the individual members of the population are mixed, are major determinants of our ability to detect population heterogeneity. The higher the relative genetic identity between the individual members of the population, the harder it is to detect, as the number of intermediate allele frequencies decreases with increased genetic identity. This means a mixture of two isolates of the same Sequence Type (ST) will be significantly harder to detect compared to a mixture of two isolates of different Clonal Complexes (CC). Similarly, populations where one member is an extreme majority (95:5) are also difficult to detect, as opposed to an equal mixture (50:50) .

In Chapter IV, we compared pure single colonies with the total population they were isolated from to evaluate the effectiveness of different sampling strategies. We analysed 254 pooled populations (pools) of *S. aureus* and eight pure single colonies from each of the 254 pools. These samples were obtained from skin swabs across different body-sites and timepoints from 85 human participants. We set out to answer 3 main questions - 1) Could the pool-seq data identify clonal *S. aureus* populations (comprising a single ST) from mixtures of diverse lineages? - Yes, based on assembly quality, allele frequencies, nucleotide diversity and contamination metrics between clonal populations and mixed populations, we were able to distinguish them with 95% accuracy. 2) Could pool-seq data be used to estimate the number of variant sites within single-ST populations? - Yes, the number of variant sites and their allele frequencies were approximately proportional between the pools and singles. However, the exact allele frequencies for each variant in the pools did not have strong correlation with the allele frequencies of the same variant in the singles. 3) Was pool-seq more sensitive in detecting AMR genes than sequencing single clones? - Yes, overall, we detected more AMR genes in the pools compared to the singles. However, we do not know whether this genotypic AMR would translate to altered minimum inhibitory concentrations (MICs) in the lab. Due to the heterogeneous nature of pools, phenotypic tests may not be reproducible. This makes estimating the true MIC of the population to a given antibiotic a challenge. Developing high-throughput workflows to predict AMR from population sequences and to measure MICs from mixed or metagenomic samples would greatly benefit clinical microbiology.

Collectively, pooled sampling provided more information than a single colony, making it the best value for estimation of diversity. Once a pool is collected, it can effectively be treated as a single stock for culturing, storage, sequencing and analysis. The pool can also be restreaked to obtain pure colonies for downstream experiments if required. **Chapter IV** demonstrates the potential for heterogeneity in highly homogeneous environments, as well as strategies for detecting this heterogeneity. Strain specific variation should be an important consideration even while monitoring conspecific microbial communities, especially in clinical settings. The concept of a "Strain" extends beyond simply naming conventions or descriptive details. Highly related isolates of the same species can still exhibit distinct behaviours (30).

The dynamics of polymicrobial interactions are also subject to strain specific variations. *S. aureus* (*Sa*) and *P. aeruginosa* (*Pa*) are a well studied microbial community as they are both ubiquitous nosocomial pathogens capable of causing chronic infections. *Sa* - *Pa* strain specific interactions have been documented with many recent studies using clinical isolates to investigate this interspecies crosstalk (31–34). However the precise genetic determinants that alter outcomes of *Sa* - *Pa* interactions are not well studied. In **Chapter V**, we quantified changes in *Sa* survival against *P*a obtained from CF lung samples and compared them to the survival of the same *Sa* against lab adapted *Pa*. We used 28 pairs of *Sa*-*Pa* with varying strain backgrounds and we found that *Sa*-*Pa* pairs from chronic infections coexist. We also found that this coexistence is mainly driven by *Pa*, as lab *Sa*, a strain that CF *Pa* has never encountered before, was also not antagonised by the CF *Pa*. These results are in line with other similar studies examining *Sa* - *Pa* interactions using clinical isolates. (31,32,35–37). However, most interestingly, we found outliers that defy these trends. Two *Sa* isolates were greatly susceptible to their respective co-infection *Pa* compared to lab *Pa*. Finding genomic determinants, if any, of *Sa* susceptibility to *Pa*, as well as increased antagonism displayed by *Pa* can help elucidate novel mechanisms of interactions between these two bacteria. Studies investigating the outlier *Sa*-*Pa* pairs are currently underway in the Goldberg lab. The impact of other *S. aureus* strain-background related factors such as the CC and *agr* group on interaction with *P. aeruginosa* have not been well studied. Understanding the effect of *P. aeruginosa* on *agr*-mediated virulence by *S. aureus*

Chapter V further emphasises that specific genotypic or phenotypic characteristics related to the strain can be drivers of heterogeneity within an ecosystem. Therefore, it is important to devise sampling and analysis strategies that incorporate capturing this microdiversity.

Infectious disease microbiology in the era of big data

Living organisms experience their surrounding environment on a unique range of scales. If one were to study any given organism, the scale at which research is conducted is a deliberate choice. ‘Scale’ here simply refers to context, and the context determines the nature of the research questions that can be answered. The scale at which a phenomenon is observed is not the only scale that is impacted by said observation, nor is it the only phenomenon that is occurring. Therefore, connecting observations made at different scales of space and time is the key to understanding patterns in any given ecosystem. These were some of the key points put forth by Dr. Simon Levin in the 1992 paper ‘The problem of pattern and scale in ecology’ (38). With better sample collection and monitoring technologies, and a better understanding of the importance of scale, it has become significantly easier to develop models that analyse and describe biological patterns. As the amount of data available is growing exponentially, it is feasible now more than ever to aggregate data across vastly different scales, giving multiple perspectives on the same phenomena.

Over the last two decades, multidisciplinary science has become the norm, allowing us to find appropriate bridges between different scales. Vast improvements in technology, especially with regards to high-throughput whole genome sequencing has completely changed our approach to infectious disease research.

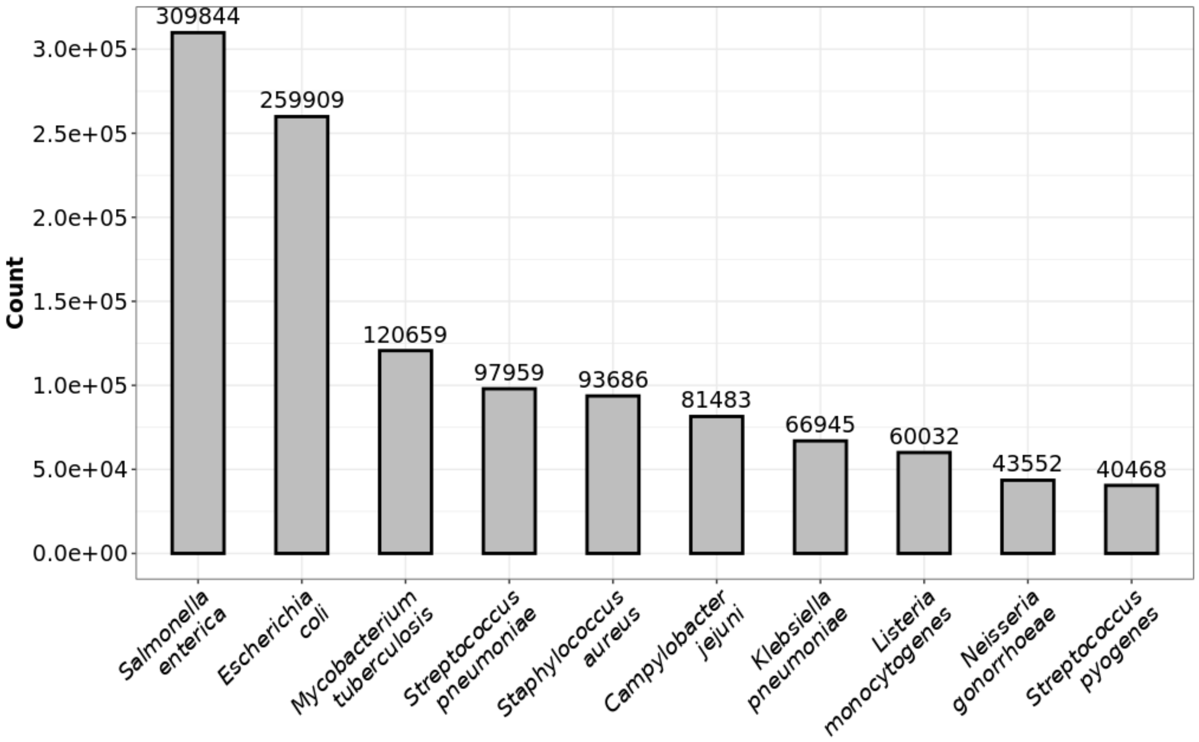

Fig 2: Bar chart showing total number of short-read sequences available per species from January 2010 to February 2023 in NCBI.

The amount of publicly available bacterial sequences have rapidly increased over the last 10 years, and most of the sequencing efforts are focused on pathogenic bacteria, *S. aureus* being one of them (**Fig 2**). This offers the opportunity to examine these pathogens from a species-wide scale spanning several decades, as well as from a smaller scale of person-to-person/person-to-environment contact required for transmission. Both scales are necessary to understand the evolution of this pathogen and provide much needed insights on the different paths to bacterial infection and survival. This PhD dissertation is but a short example of how combining data across scales, multidisciplinary science, and technological advancements can come together to provide new insights into infectious disease microbiology.

References

1. Feil EJ, Cooper JE, Grundmann H, Robinson DA, Enright MC, Berendt T, et al. How clonal is Staphylococcus aureus? J Bacteriol. 2003 Jun;185(11):3307–16.
2. Feil EJ. Small change: keeping pace with microevolution. Nat Rev Microbiol. 2004 Jun;2(6):483–95.
3. Marraffini LA, Sontheimer EJ. CRISPR Interference Limits Horizontal Gene Transfer in Staphylococci by Targeting DNA. Science. 2008 Dec 19;322(5909):1843–5.
4. Monk IR, Tree JJ, Howden BP, Stinear TP, Foster TJ. Complete Bypass of Restriction Systems for Major Staphylococcus aureus Lineages. mBio. 2015 May 26;6(3):e00308-15.
5. Moller AG, Petit RA, Read TD. Species-Scale Genomic Analysis of Staphylococcus aureus Genes Influencing Phage Host Range and Their Relationships to Virulence and Antibiotic Resistance Genes. mSystems. 2022 Jan 18;7(1):e01083-21.
6. Corvaglia AR, François P, Hernandez D, Perron K, Linder P, Schrenzel J. A type III-like restriction endonuclease functions as a major barrier to horizontal gene transfer in clinical Staphylococcus aureus strains. Proc Natl Acad Sci. 2010 Jun 29;107(26):11954–8.
7. Jarraud S, Cozon G, Vandenesch F, Bes M, Etienne J, Lina G. Involvement of Enterotoxins G and I in Staphylococcal Toxic Shock Syndrome and Staphylococcal Scarlet Fever. J Clin Microbiol. 1999 Aug;37(8):2446–9.
8. Naik S, Smith F, Ho J, Croft NM, Domizio P, Price E, et al. Staphylococcal Enterotoxins G and I, a Cause of Severe but Reversible Neonatal Enteropathy. Clin Gastroenterol Hepatol. 2008 Feb 1;6(2):251–4.
9. Ono HK, Hirose S, Naito I, Sato'o Y, Asano K, Hu DL, et al. The emetic activity of staphylococcal enterotoxins, SEK, SEL, SEM, SEN and SEO in a small emetic animal model, the house musk shrew. Microbiol Immunol. 2017;61(1):12–6.
10. Hodille E, Alekseeva L, Berkova N, Serrier A, Badiou C, Gilquin B, et al. Staphylococcal Enterotoxin O Exhibits Cell Cycle Modulating Activity. Front Microbiol [Internet]. 2016 [cited 2023 Apr 28];7. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2016.00441
11. Mehta RS, Petit RA, Read TD, Weissman DB. Detecting patterns of accessory genome coevolution in bacterial species using data from thousands of bacterial genomes [Internet]. bioRxiv; 2022 [cited 2023 Apr 28]. p. 2022.03.14.484367. Available from: https://www.biorxiv.org/content/10.1101/2022.03.14.484367v1
12. Giulieri SG, Guérillot R, Duchene S, Hachani A, Daniel D, Seemann T, et al. Niche-specific genome degradation and convergent evolution shaping Staphylococcus aureus adaptation during severe infections. Kana BD, Van Tyne D, Zheng M, editors. eLife. 2022 Jun 14;11:e77195.
13. Thomas JC, Godfrey PA, Feldgarden M, Robinson DA. Candidate Targets of Balancing Selection in the Genome of Staphylococcus aureus. Mol Biol Evol. 2012 Apr;29(4):1175–86.
14. Wilson DJ, The CRyPTIC Consortium. GenomegaMap: Within-Species Genome-Wide dN/dS Estimation from over 10,000 Genomes. Mol Biol Evol. 2020 Aug 1;37(8):2450–60.

15. Hanssen AM, Kjeldsen G, Sollid JUE. 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 Jan;48(1):285–96.
16. Wielders CL, Vriens MR, Brisse S, de Graaf-Miltenburg LA, Troelstra A, Fleer A, et al. In-vivo transfer of mecA DNA to Staphylococcus aureus [corrected]. Lancet Lond Engl. 2001 May 26;357(9269):1674–5.
17. Gill SR, Fouts DE, Archer GL, Mongodin EF, Deboy RT, Ravel J, et al. Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant Staphylococcus aureus strain and a biofilm-producing methicillin-resistant Staphylococcus epidermidis strain. J Bacteriol. 2005 Apr;187(7):2426–38.
18. Bloemendaal ALA, Brouwer EC, Fluit AC. Methicillin resistance transfer from Staphylocccus epidermidis to methicillin-susceptible Staphylococcus aureus in a patient during antibiotic therapy. PloS One. 2010 Jul 29;5(7):e11841.
19. Berglund C, Söderquist B. The origin of a methicillin-resistant Staphylococcus aureus isolate at a neonatal ward in Sweden-possible horizontal transfer of a staphylococcal cassette chromosome mec between methicillin-resistant Staphylococcus haemolyticus and Staphylococcus aureus. Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis. 2008 Nov;14(11):1048–56.
20. Zhang DF, Zhi XY, Zhang J, Paoli GC, Cui Y, Shi C, et al. Preliminary comparative genomics revealed pathogenic potential and international spread of Staphylococcus argenteus. BMC Genomics. 2017 Oct 23;18(1):808.
21. Tan L, Huang Y, Shang W, Yang Y, Peng H, Hu Z, et al. Accessory Gene Regulator (agr) Allelic Variants in Cognate Staphylococcus aureus Strain Display Similar Phenotypes. Front Microbiol [Internet]. 2022 [cited 2023 Apr 17];13. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2022.700894
22. Altman DR, Sullivan MJ, Chacko KI, Balasubramanian D, Pak TR, Sause WE, et al. Genome plasticity of agr-defective staphylococcus aureus during clinical infection. Infect Immun. 2018 Oct 1;86(10):e00331.
23. Schweizer ML, Furuno JP, Sakoulas G, Johnson JK, Harris AD, Shardell MD, et al. Increased mortality with accessory gene regulator (agr) dysfunction in Staphylococcus aureus among bacteremic patients. Antimicrob Agents Chemother. 2011 Mar;55(3):1082–7.
24. Suligoy CM, Lattar SM, Noto Llana M, González CD, Alvarez LP, Robinson DA, et al. Mutation of Agr Is Associated with the Adaptation of Staphylococcus aureus to the Host during Chronic Osteomyelitis. Front Cell Infect Microbiol. 2018;8:18.
25. Das S, Lindemann C, Young BC, Muller J, Österreich B, Ternette N, et al. Natural mutations in a Staphylococcus aureus virulence regulator attenuate cytotoxicity but permit bacteremia and abscess formation. Proc Natl Acad Sci U S A. 2016 May 31;113(22):E3101-3110.
26. Goerke C, Kümmel M, Dietz K, Wolz C. Evaluation of intraspecies interference due to agr polymorphism in Staphylococcus aureus during infection and colonization. J Infect Dis. 2003 Jul 15;188(2):250–6.
27. Fowler VG, Sakoulas G, McIntyre LM, Meka VG, Arbeit RD, Cabell CH, et al. Persistent bacteremia due to methicillin-resistant Staphylococcus aureus

infection is associated with agr dysfunction and low-level in vitro resistance to thrombin-induced platelet microbicidal protein. J Infect Dis. 2004 Sep 15;190(6):1140–9.

28. Sakoulas G, Eliopoulos GM, Fowler VG, Moellering RC, Novick RP, Lucindo N, et al. 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 Jul;49(7):2687–92.
29. Pader V, Hakim S, Painter KL, Wigneshweraraj S, Clarke TB, Edwards AM. Staphylococcus aureus inactivates daptomycin by releasing membrane phospholipids. Nat Microbiol. 2016 Oct 24;2:16194.
30. Goyal A, Bittleston LS, Leventhal GE, Lu L, Cordero OX. Interactions between strains govern the eco-evolutionary dynamics of microbial communities. Weigel D, editor. eLife. 2022 Feb 4;11:e74987.
31. Limoli DH, Yang J, Khansaheb MK, Helfman B, Peng L, Stecenko AA, et al. Staphylococcus aureus and Pseudomonas aeruginosa co-infection is associated with cystic fibrosis-related diabetes and poor clinical outcomes. Eur J Clin Microbiol Infect Dis Off Publ Eur Soc Clin Microbiol. 2016 Jun;35(6):947–53.
32. Bernardy EE, Petit RA, Raghuram V, Alexander AM, Read TD, Goldberg JB. Genotypic and Phenotypic Diversity of Staphylococcus aureus Isolates from Cystic Fibrosis Patient Lung Infections and Their Interactions with Pseudomonas aeruginosa. mBio. 2020 Jun 23;11(3):e00735-20.
33. Gomes-Fernandes M, Gomez AC, Bravo M, Huedo P, Coves X, Prat-Aymerich C, et al. Strain-specific interspecies interactions between co-isolated pairs of Staphylococcus aureus and Pseudomonas aeruginosa from patients with tracheobronchitis or bronchial colonization. Sci Rep. 2022 Mar 1;12(1):3374.
34. Wieneke MK, Dach F, Neumann C, Görlich D, Kaese L, Thißen T, et al. Association of Diverse Staphylococcus aureus Populations with Pseudomonas aeruginosa Coinfection and Inflammation in Cystic Fibrosis Airway Infection. mSphere. 2021 Jun 30;6(3):e0035821.
35. Filkins LM, Graber JA, Olson DG, Dolben EL, Lynd LR, Bhuju S, et al. Coculture of Staphylococcus aureus with Pseudomonas aeruginosa Drives S. aureus towards Fermentative Metabolism and Reduced Viability in a Cystic Fibrosis Model. J Bacteriol. 2015 Jul;197(14):2252–64.
36. DeLeon S, Clinton A, Fowler H, Everett J, Horswill AR, Rumbaugh KP. Synergistic Interactions of Pseudomonas aeruginosa and Staphylococcus aureus in an In Vitro Wound Model. Infect Immun. 2014 Nov;82(11):4718–28.
37. Millette G, Langlois JP, Brouillette E, Frost EH, Cantin AM, Malouin F. Despite Antagonism in vitro, Pseudomonas aeruginosa Enhances Staphylococcus aureus Colonization in a Murine Lung Infection Model. Front Microbiol. 2019;10:2880.
38. The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture - Levin - 1992 - Ecology - Wiley Online Library [Internet]. [cited 2023 Apr 17]. Available from: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.2307/1941447