Interpreting roles of mutations associated with the emergence of *S. aureus* USA300 strains using transcriptional regulatory network reconstruction

Saugat Poudel[1], Jason Hyun[1], Ying Hefner[1], Jon Monk[2], Victor Nizet [3,4], Bernhard O. Palsson[1,3,4,*]

Affiliations:

1. Department of Bioengineering, University of California San Diego

2. Palmona Pathogenomics

3. Collaborative to Halt Antibiotic-Resistant Microbes (CHARM), Department of Pediatrics, University of California San Diego

4. Department of Pediatrics, University of California San Diego

*Correspondence: Bernhard O. Palsson, Department of Bioengineering, University of California San Diego, 9500 Gilman Drive, La Jolla, CA. Email: palsson@eng.ucsd.edu

Abstract

The *Staphylococcus aureus* clonal complex 8 (CC8) is made up of several subtypes with varying levels of clinical burden; from community-associated methicillin resistant *S. aureus* (CA-MRSA) USA300 strains to hospital-associated (HA-MRSA) USA500 strains and basal methicillin susceptible (MSSA) strains. This phenotypic distribution within a single clonal complex makes CC8 an ideal clade to study the emergence of mutations important for antibiotic resistance and community spread. Gene level analysis comparing USA300 against MSSA and HA-MRSA strains have revealed key horizontally acquired genes important for its rapid spread in the community. However, efforts to define the contributions of point mutations and indels have been confounded by strong linkage disequilibrium resulting from clonal propagation. To break down this confounding effect, we combined genetic association testing with a model of the transcriptional regulatory network (TRN) to find candidate mutations that may have led to changes in gene regulation. First, we used a De Bruijn graph genome-wide association study (DBGWAS) to enrich mutations unique to the USA300 lineages within CC8. Next, we reconstructed the TRN by using Independent Component Analysis on 670 RNA sequencing samples from USA300 and non-USA300 CC8 strains which predicted several genes with strain-specific altered expression patterns. Examination of the regulatory region of one of the genes enriched by both approaches, *isdH,* revealed a 38 base pair deletion containing a Fur binding site and a conserved SNP which likely led to the altered expression levels in USA300 strains. Taken together, our results demonstrate the utility of reconstructed TRNs to address the limits of genetic approaches when studying emerging pathogenic strains.

Introduction

Comparative genomic methods represent an important approach to understand the emergence and evolution of new strains of pathogens. In *S. aureus* alone, whole genome comparisons have enabled rapid characterization of genetic basis for antibiotic resistance, increased virulence, host specificity and altered metabolic capabilities [1–5]. However, genome-wide linkage disequilibrium and strong lineage structuring currently limits the differentiation of causative alleles from genetically linked ones. By calculating lineage level associations, methods like bugwas address these issues for single, recurring phenotypes like antibiotic resistance[6].

Endemic strains, on the other hand, exhibit multiple complex phenotypes that may contribute to their emergence and proliferation. For example, USA300 strains carry antibiotic resistance cassettes, Panton Valentine Leukocidin (PVL) associated with pyomastitis, increased ability to colonize locations outside of the nasopharynx etc[1,7,8]. As these strains often emerge clonally from closely related ‘basal strains,’ efforts to discern causal mutations that lead to their increased clinical burden is hampered by strong population-stratification and genome-wide linkage disequilibrium [9–11]. Though recombination at species level is common in *S. aureus*, within clade recombination rates tend to be lower, thus preserving the linkage between mutations [10,12–14]. Due to this limitation, studies of emerging strains often focus on gene level

analysis such as acquisition of mobile genetic elements or loss of gene function while determining the possible phenotypic effect (if any) of all enriched Single Nucleotide Polymorphisms (SNPs) remains challenging [15,16]. Computational modeling methods can help tackle these challenges by predicting phenotype differences between strains, thus acting as a sieve to filter enriched mutations with potential phenotypic effects and therefore find candidate causal mutations [17–19]. Even if experimentally intractable, the large possible phenotypic space of an organism can be explored quickly with computational models.

Here, we used De Bruijn graph GWAS (DBGWAS) to discover enriched mutations associated with the endemic USA300 strain within clonal complex 8 (CC8) [20]. Due to clonal expansion of USA300 strains from their progenitors within CC8, the enriched USA300 specific mutations were in high linkage disequilibrium. Further complicating the matter, we found that almost all mutations enriched within open reading frames (ORFs) were unique to USA300 lineage and not found in any other clonal complexes, precluding identification of potential causative mutations by homoplasy. Instead, we turned to reconstruction of a transcriptional regulatory network (TRN) to identify genes that were both associated with an enriched mutation and had altered regulation in USA300 strains. We built an Independent Component Analysis (ICA)-based reconstruction of the TRN using 670 publicly available RNA sequencing samples from both USA300 and non-USA300 CC8 strains. By factoring the RNA sequencing data into a series of signals and their activities, the ICA-based reconstruction of the TRN shows both the static gene-regulator interaction and the dynamic activity of these interactions in a sample specific manner[21]. However, ICA is a generalized signal extraction algorithm and therefore does not distinguish between biological sources of signals like regulatory elements and 'artificial' sources that can be created by sourcing data from multiple strains. Therefore, in addition to signals associated with gene regulators, ICA also outputs signals associated with strain-specific changes in gene regulation. Furthermore, by utilizing RNA sequencing data from hundreds of samples to identify genes with strain-specific expression patterns, this approach is more likely to find strain-specific differences than previous approaches that focus on specific conditions [22,23].

This analysis revealed several genes with distinct expression patterns in USA300 strains that were also associated with DBGWAS enriched mutations. One of these genes, *isdH*, which encodes a haptoglobin binding protein, showed several enriched mutations in the regulatory region of USA300 strains including the deletion of the Fur repressor binding site. Additionally, the *isdH* gene generally had higher expression levels in samples from USA300 strains, connecting the mutation enriched by DBGWAS with differences in TRN enriched by ICA. Overall, our analysis shows how the reconstruction of TRN can be used to extend the limits of current GWAS approaches when studying emerging and endemic populations of bacterial pathogens.

Results

Classifying USA300 and non-USA300 genomes based on genetic markers

We sought to compare the genetic differences between USA300 CA-MRSA strains and other subtypes within CC8 that have lower clinical and community burden. Given that both

subtypes exist within the same clonal complex, this comparison allowed us to probe the genetic basis for the success of USA300 strains with limited confounding effects of different genetic backgrounds. We analyzed 2038 *S. aureus* CC8 genomes which formed a closed pangenome, suggesting that the sampled genomes mostly captured the gene level variations within the clonal complex (**Figure 1a**). The CC8 pangenome consisted of 19176 gene clusters with 2291 core genes that were present in at least ~95% of the genomes analyzed. Among the remainder of the genes, 931 were categorized as accessory genes and 15954 were found in less than 5% of the genomes (**Figure S1a**). To get a higher resolution view of these genomes, we surveyed for unique alleles within the ORFs and in the 300 base-pair 5' upstream and 3' downstream sequences. Interestingly, we found a larger number of mutations within the ORFs, indicating the presence of greater genetic variation in the ORFs than in the neighboring regulatory regions.

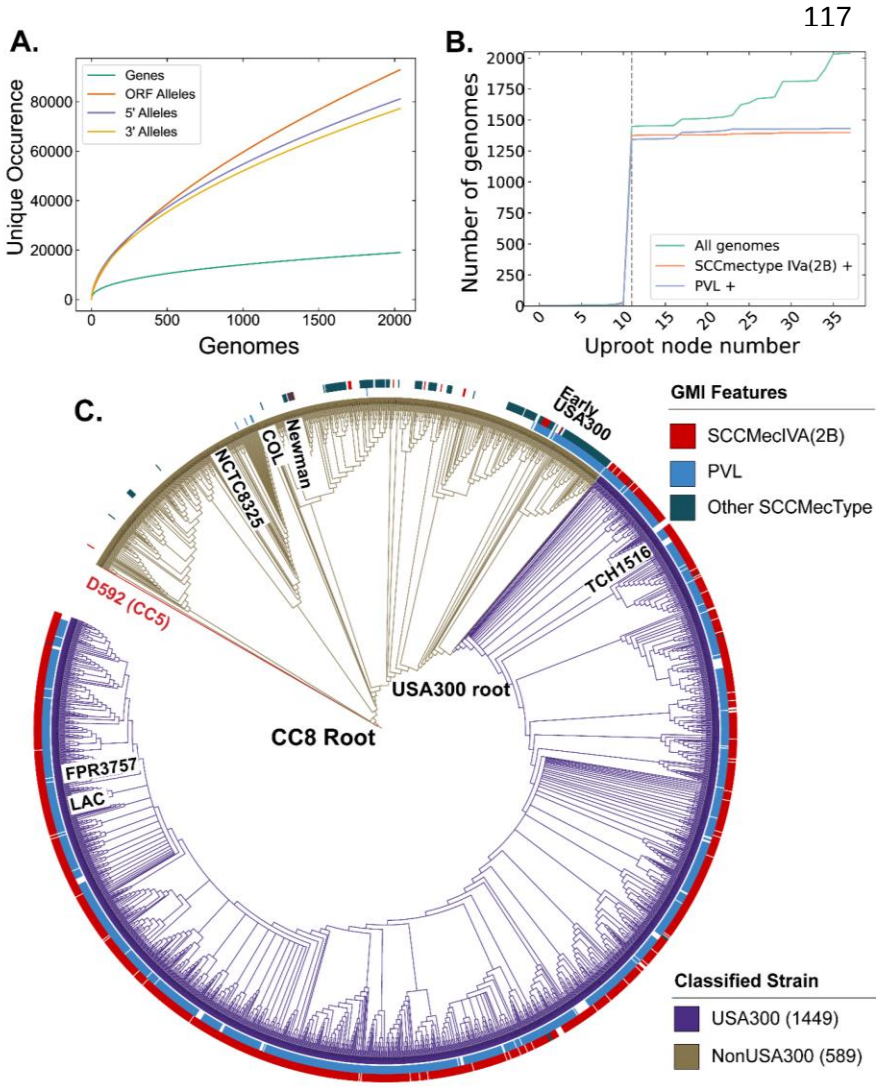

Figure 1. CC8 pangenome and phylogeny. (a) Pangenomic analysis of CC8 genomes shows the distribution of genes and mutations in ORFs and regulatory regions. (b) Prevalence of USA300 specific genetic markers, PVL and SCCMecIVa, as you traverse up the phylogenetic tree from TCH1516. The gray dashed line represents the node where the USA300 root is placed. (c) Phylogenetic tree of CC8 genomes classified into USA300 and non-USA300 strains.

Next, we classified the CC8 genomes into USA300 and non-USA300 strains using Genetic Marker Inference (GMI). GMI was previously developed to rapidly and systematically identify different subclades within inner-CC8 strictly based on genetic markers [24]. In this scheme, USA300 genomes can be differentiated from non-USA300 CC8 genomes by the presence of either SCCMecIVa or the presence of Panton-Valentine Leukocidin (PVL) in case of methicillin sensitive *S. aureus* (MSSA).

To identify the root of the USA300 clades, we first traversed up nodes of the phylogenetic tree starting from known USA300 strain TCH1516 and determined the number of strains, fraction PVL positive and fraction SCCMecIVa positive for each node during traversal. The root was placed at the last node where >90% of the strains within the subclade represented by the daughter nodes were SCCMecIVa and PVL positive **(Figure 1b)**. As phylogenetic trees are nested, root finding with this procedure is not dependent on the starting USA300 strain. Same root was identified when the procedure was initialized with another well known USA300 reference strain FPR3757 **(Figure S1b).** Combining the genetic markers with phylogenetic grouping led to the classification of 1449 genomes as USA300 and 589 genomes as non-USA300 **(Figure 1c, Supplementary Table 1)**. Strains previously identified as 'early USA300' were not part of our USA300 classification[24]. While many of these strains are PVL positive, they have variable SCCMec types and therefore are likely to be clinically distinct from the endemic USA300 strains.

Enriching USA300 specific genes and mutations using DBGWAS

After classifying the genomes into USA300 and non-USA300 strains, we identified genes and mutations associated with each subtype by using the De Bruijn graph Genome Wide Association Study (DBGWAS)[20]. DBGWAS provides a reference genome-free method for conducting GWAS analysis in prokaryotes by building a compacted De Brujin Graph to represent the pan genome of input sequences. The nodes of the graph represent unique compacted k-mers that are joined by edges to other nodes with k-mers that appear adjacent to it in genomes. The procedure enriches unique k-mers that appear with different frequencies in each classification and outputs the enriched k-mer as well as its genetic neighborhood (called 'components') from the De Bruijn graph. Visualizing the components associated with the enriched k-mers makes it easier to interpret the k-mers and makes it easy to identify large structural variations (e.g. cassette acquisition) which are often represented by multiple enriched k-mers that fall within the same component. We took the output component graphs and automatically extracted the enriched genetic changes e.g. indels, SNPs, phage insertions etc (**see Supplementary Note 1**).

Many of the components were associated with genes and genetic elements expected to be enriched with USA300 strains- SCCMecIVA (the GMI marker), Arginine Catabolite Repressor Element (ACME), cap5E point mutation, multiple prophages etc **(Supplementary Table 2)**. In total, we found k-mers in 149 components associated with 137 unique TCH1516 genes that were enriched in this analysis, pointing to a large array of mutations that are unique to the USA300 lineage (**Figure 2a**). Significant k-mers in some components did not uniquely match to TCH1516 genes or only matched to genes in non-USA300 reference genome, NCTC8325.

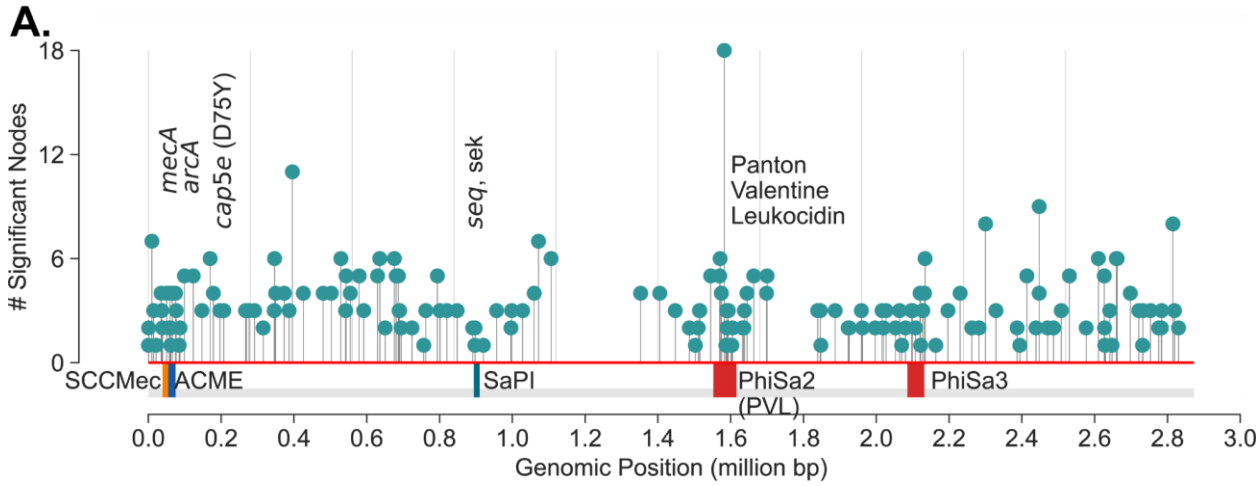

Figure 2. USA300 strains associated mutations. (a) DBGWAS recovers components associated with USA300 previously described markers of USA300 strains including *mecA* (SCCMec IVa), *arcA* (ACME), *cap5e* mutation, *seq, sek* and Phi-PVL. In addition, components with many other mutations scattered throughout the genome (NC_010079) are also enriched. Each 'significant node' represents a k-mer sequence (with minimum size of 31 nucleotides) that are associated with USA300 strains (adjusted p-value < 0.05).

Genome-wide linkage and de novo mutations obfuscate identification of causal mutations

Though these mutations were enriched in USA300 strains with DBGWAS, we could not attribute the prevalence of any particular mutation to selection due to strong genome-wide linkage. We quantified the linkage disequilibrium by calculating the square of the correlation coefficient (r^2) for each of the enriched k-mer not associated with mobile genetic elements. High correlation coefficient indicates tight co-occurrence of k-mers in the genomes and therefore high linkage disequilibrium between the sequences. There was a strong linkage between the k-mers that were enriched in USA300 strains. Surprisingly, even k-mers that were 1.4 million base pairs away (the maximum distance between two sites in the circular 2.8 million base pairs long *S. aureus* genome) still had r^2 over 0.9 (**Figure 3a**).

To differentiate potential causal mutations from genetically linked alleles, we searched for mutation hotspots by comparing the positions of USA300 mutations in open reading frames (ORFs) to mutations in other clonal complexes. Barring recombination events, presence of mutation hotspots in the same position in multiple clades could point to selection acting on the sequence. Therefore, we searched for prevalence of enriched mutations in other non-CC8 clades. We identified 61 SNPs within open reading frames (ORFs) that were enriched in USA300 strains. To identify mutational hotspots in other clades, we downloaded all the amino acid sequences belonging to the PATRIC genus protein family of each of the gene products encoded by the selected ORFs [25]. The PATRIC local protein family consists of sequences of homologous proteins within the same genus which were further filtered down to *S. aureus* species specific sequences. After filtering, each protein family comprised 2,000 to 16,000 unique sequences and the strains from which the amino acid sequences were derived spanned dozens of clades allowing for broad comparisons **(Figure S2)**. Lastly, we removed sequences

associated with ST239 as it is thought to have emerged from large-scale recombination of ST8 and ST30 strains [26].

We determined mutation hotspots by calculating position specific allelic entropy. Allelic entropy at a given amino acid position is a function of the number of unique amino acids found in that position and the frequency of the mutation [27]. Positions where all queried sequences have the same amino acid have low entropy, while positions that have frequent amino acid substitutions (hotspots) have high entropy (**Figure 3b**). This measure allows us to quickly determine the positions of mutation hotspots while accounting for multiple possible amino acid substitutions and rare mutations. Before calculating the position-specific entropy, all sequences within each of the PATRIC local protein families were aligned with Multiple Sequence Alignment (MSA). This alignment ensures proper comparison of amino acids even when there are deletions or insertions in some of the genes in the family.

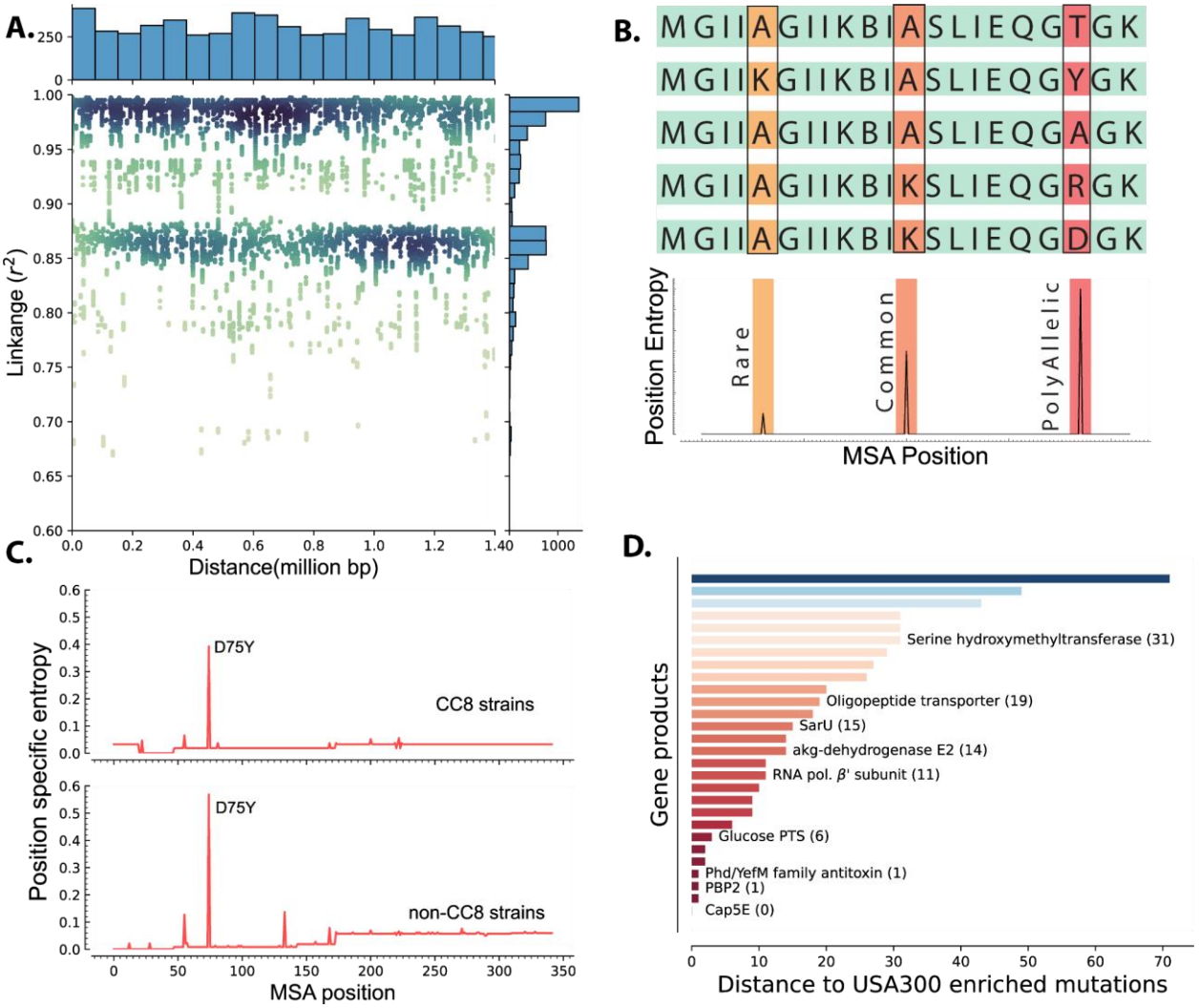

Figure 3. Linkage Disequilibrium and de novo mutations in USA300 strains. (a) Enriched k-mers showed high linkage disequilibrium, with some k-mers at 1.4 Mbp distance still having r^2 of greater than 0.98. (b) Schematic of position specific entropy analysis. Positions with heterogeneous sequences have higher calculated

entropy than more conserved sequences with fewer mutations. (c) Using position specific entropy, we only found one example of shared enriched mutation in ORFs of USA300 and non-USA300 strains. (d) Distance (in base pairs) between the position of enriched mutation in USA300 strains and the position of the nearest entropy peak in other non-CC8 strains.

Of the 36 enriched ORF mutations only the Asp75Tyr mutation in the *cap5E* gene, which was previously shown to ablate capsule production in USA300 strains, was found in other strains (**Figure 3c**) [16]. Peaks in entropy corresponding to this mutation position were present in both the CC8 and non-CC8 strains while all other mutation positions were unique to CC8. Despite not having any perfect matches outside of the *cap5E* mutation, we found that for 28 of the mutations, a peak was present in sequences from other clades within 71 MSA positions. Together, our data suggests that mutations within ORFs in USA300 strains are likely de novo mutations and are not acquired through horizontal gene transfer though many of these mutations have occurred in hotspot regions (**Figure 3d**).

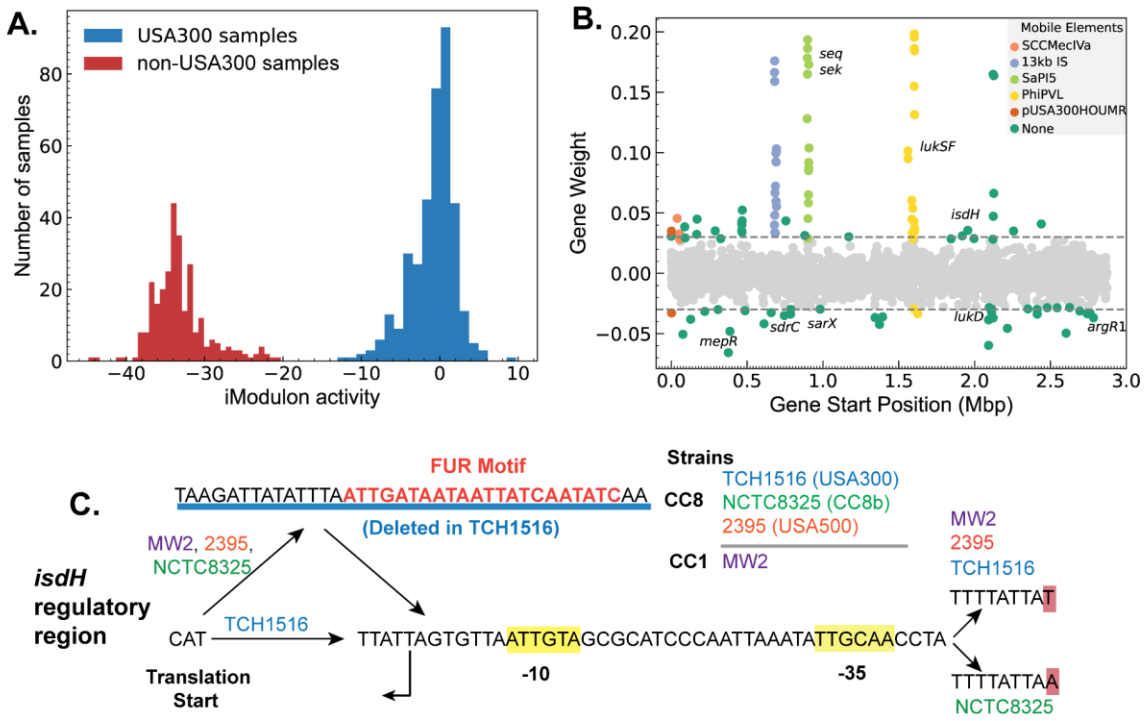

FIgure 4. Strain-specific regulatory changes in the CC8 clade. (a) ICA analysis of USA300 and non-USA300 RNA-sequencing data identified an iModulon with strain specific activity.(b) The strain-specific iModulon contained various horizontally acquired elements (e.g. ACME, PhiPVL) that are prevalent in USA300 lineage as well as conserved genes with strain-specific expression patterns. (c) Comparing the 5' regulatory region of the gene *isdH* from various *S. aureus* strains revealed a unique deletion containing Fur binding site in USA300 reference strain TCH1516.

iModulon in the CC8 TRN points to mutations associated with differential regulation

The presence of genome-wide linkage and de novo mutations in ORFs severely limited the ability to distinguish causal SNPs contributing to increased pathogenesis in USA300 strains.

The effect of some mutations, especially in ORFs, has been successfully linked to distinct phenotypes such as the absence of a capsule in USA300 and USA500 strains [16]. However, the effect of mutations associated with changes in gene regulations can be much more difficult to assess [15]. To look for mutations that may be associated with changes in transcriptional regulation, we used ICA to model gene-regulation in CC8 strains which can predict strain specific differences in expression patterns.

We collected CC8 strains associated RNA-sequencing data from Sequence Read Archive (SRA). After stringent QC/QA and curation, 291 non-USA300 strains (e.g. Newman, NCTC8325) and 379 USA300 (e.g. LAC, TCH1516) samples were used to create a single reconstruction of TRN across CC8 using ICA [21,28] (**Supplementary Table 3)** ICA calculates independently modulated sets of genes, iModulons, and the activities of those gene sets in each sample. iModulons calculated by ICA represent distinct sources of signals in the RNA-sequencing data. While most of the signals can be associated with different regulatory elements, iModulons associated with other biological features such as mobile genetic elements, genomic backgrounds are also enriched. In *Escherichia coli and Salmonella enterica Typhimurium*, multi-strain ICA has been used to calculate strain-specific iModulons that represent differences in gene expression [21,29].

In our reconstruction, two iModulons captured a large number of genes with different expression levels in the non-USA300 and USA300 strains (**Figure 4a; Figure S3a**). Most of the genes in the strain-specific iModulons belonged to mobile elements associated with USA300 strains such as ACME, SCCMec, Phi-PVL etc. However, the iModulons also contained core genes that are present in both strains, pointing to possible differences in gene regulation (**Figure 4b; Figure S3b**).

We mapped the enriched mutations from DBGWAS onto the core genes enriched in the strain-specific iModulon. 10 core genes- including *isdH, argR1 and araC* family regulator- with mutations in the ORF or in the regulatory region were also enriched in the strain-specific iModulon (**Supplementary Table 4**). Of these genes, gene *isdH*, encoding a heme scavenger molecule showed distinct strain-specific expression levels and had enriched k-mers that are mapped to the upstream regulatory region. Therefore, we compared the upstream regulatory region of several reference strains including TCH1516 (USA300), NCTC8325 (CC8b), 2395 (USA500). Additionally, we included MW2 (CC1 CA-MRSA) as the transcription start site (TSS) in the region has been experimentally confirmed in this strain[30]. Comparisons showed a 38 base-pair deletion in the 5' untranslated region containing a transcription factor Fur binding site (q-val=0.033e-4) (**Figure 4c**).

This deletion was detected in all of the 1385 USA300 genomes, but only present in 95 of the 589 non-USA300 genomes. As Fur is a repressor that blocks expression in presence of iron concentration, this deletion in the Fur binding site may be responsible for the general increase in *isdH* expression observed in USA300 samples (**Figure S4a**). We also found a second mutation upstream of the predicted -35 binding site that was also enriched in USA300. Interestingly, while the MW2 strain did not have the 38 bp deletion, it contained the exact upstream A->T mutation. All other base-pairs in the region were perfectly matched in between all the reference genomes. The combination of evidence from genetic and transcriptomic analysis suggests that regulation of *isdH* is altered in USA300 strains compared to its non-USA300 progenitors.

As with many point mutations detected in our analysis, the absence of Fur binding site

upstream of *isdH* gene is prevalent only in the USA300 lineage. We searched for Fur binding motif in the 100 base-pairs upstream regions from 3515 non-CC8 strains spanning multiple clonal complexes (**Figure S2**). We detected the binding motif in all but 21 strains. Of the 21 strains with no detectable Fur binding sites, 6 belonged to ST72 (out of 28 total from this type) and 6 had uncharacterized MLST type. The rest were distributed among types 121, 1750, 375, 1, 3317, 15, 7, 398, and 4803, with one positive strain per type.

Discussion

Emergence of CA-MRSA USA300 strains from HA-MRSA USA500 progenitors presents a natural experiment to probe the genetic basis for the establishment of the USA300 lineage. However, in studying these groups, genetic methods like GWAS were limited in finding causal mutations due to genome-wide linkage disequilibrium and presence of an unexpectedly large number of de novo mutations unique to the USA300 lineage. Here, we demonstrated how a model of transcriptional regulation with iModulons can be used to break through the impasse created by the high linkage disequilibrium and predict candidate causal mutations. From the combined RNA sequencing dataset of USA300 and non-USA300 strains, ICA calculated a single iModulon that captured strain specific variation in gene expression. As expected, most genes in the iModulons were part of mobile genetic elements such as ACME and SCCMec because they have zero expression level in non-USA300 samples. However, the iModulon also contained several core genes that are present in both groups but are differentially regulated. A deeper analysis of the regulatory region of one of these genes with enriched mutation, *isdH*, revealed a deletion of a DNA segment containing the binding site of the Fur repressor. In congruence with this observation we also found that USA300 strains with the deleted Fur binding site showed general increase in *isdH* expression level. Combining GWAS with large-scale transcriptomic modeling was therefore able to predict potential causal mutations that led to the increased clinical burden of the USA300 lineage.

The current analysis utilized the available DNA and RNA sequencing data and the methods used here are scalable to the rapidly growing number of data in the public repositories. Indeed, with the greater scale, we can get more granular insight into subclade specific differences. The transcriptomic analysis consisted of samples primarily from the USA300 (CC8e and CC8f) clades, the CC8a clade represented by Newman and the CC8b clade represented by NCTC8325 and its derivatives. However, the CC8b and CC8a clades are currently undersampled due to its minimal clinical burden compared to USA300. We therefore combined strains from all non-USA300 clades into a single group for GWAS. The misalignment of RNA sequencing samples from GWAS samples may explain the low number of hits that were enriched by both methods when many other unique gene expression patterns have been observed in USA300 strains. The scalability of the methods used herein will enable granular and more in-depth analysis as these sequencing databases continue to expand rapidly.

With time, the scaling of databases may be able to resolve the issue of imbalanced sampling. On the other hand, resolving the confounding effect of linkage disequilibrium inherent in emerging and endemic strains will require a new generation of modeling methods[11]. Our current approach focuses on modeling the changes in gene regulation at the transcriptional level, but causal mutations can have any number of effects on the phenotype of the organism.

New modeling methods that can systematically predict these other phenotypes are now rapidly emerging. Our recent work with *Mycobacterium tuberculosis* utilized a metabolic allele classifier (MACs) which combines genome scale metabolic models with machine learning to estimate biochemical effects of alleles thus mapping mutations to changes in metabolic fluxes [19]. Similarly, advances in protein structure prediction with AlphaFold2 and RosettaFold puts us at the cusp of predicting the effects of mutations on protein folding [31,32]. Combination of these modeling techniques may therefore prove to be the breakthrough required to advance solutions to the current challenges in population genetics of emerging pathogens.

Materials and Methods

Pangenomic analysis

The pangenome analysis was run as described in detail before [27]. Briefly, “complete” or “WGS” samples from CC8/ST8 were downloaded from the PATRIC database[25]. Sequences with lengths that were not within 3 standard deviations of the mean length or those with more than 100 contigs were filtered out. A non-redundant list of CDSs from all genomes was created and clustered by protein sequence using CD-HIT with minimum identity and minimum alignment length of 80%[33]. To get the 5’ and 3’ sequences, non-redundant 300 nucleotide upstream and downstream sequences from the CDS were extracted for each gene.

The CDSs were divided into core, accessory and unique genes based on the frequency of genes as previously described[27]. To calculate the frequency thresholds for each category, P(x), the number of genes with frequency x and its integral F(x), the cumulative frequency less than or equal to x were calculated. The multimodal gene distribution can be estimated by sum of two power laws as:

$$P(x) = c_1 x^{-\alpha_1} + c_2 (N + 1 - x)^{-\alpha_2} \qquad x = 1,2,\ldots,N$$

where N is the total number of genomes, x is the gene frequency and (c_1, c_2, $-\alpha_1$, $-\alpha_2$) are parameters fit based on the data. The cumulative distribution is then the integral of P(x) with additional parameter k:

$$F(x) = k + \frac{C_1}{1 - \alpha_1} x^{1-\alpha_1} - \frac{C_2}{1 - \alpha_2} (N + 1 - x)^{1-\alpha_2}$$

The parameters (c_1, c_2, $-\alpha_1$, $-\alpha_2$, and k) were fitted based on the data using nonlinear least squares regression from scipy[34]. The frequency threshold of core genomes was defined as greater than 0.9 N + 0.1 x* and the threshold for unique genome was defined as 0.1x*, where x* represents the inflection point of the fitted cumulative distribution.

Reconstructing CC8 phylogenetic tree

The phylogenetic tree was reconstructed using the standardized PHaME pipeline on the PATRIC sequences that passed the QC/QA[35]. Using the pipeline, the contigs and sequences were aligned to the reference TCH1516 genome NC_010079 and plasmids NC_012417, NC_010063 [36] and core SNPs were calculated. The core SNPs were then used to estimate the phylogenetic tree using IQ-TREE run with 1000 bootstraps and utilizing the ultrafast bootstrap[37,38]. The tree was built using the “TVMe+ASC+G4" model as suggested by the IQ-TREE ModelFinder[39]. Finally, iTOL was used to visualize, annotate and root the tree with the USA100 D592 (NZ_CP035791) from CC5 as the outgroup [40].

Classification of USA300 and non-USA300 strains

The USA300 and non-USA300 strains were classified based on a previously proposed and validated CC8 subtyping scheme[24]. In this scheme, USA300 strains can be identified from the whole genome if they are PVL positive MSSA or MRSA with SCCMec IVa cassette. We detected SCCMec types using SCCMecFinder, and only those genomes where the cassette could be identified by both BLAST and k-mer based methods were marked as positive[41]. PVL was detected using protein BLAST. To find the root of the USA300 strains in the phylogenetic tree, the genomes in the tree were annotated by their PVL and SCCMec status. We added additional criteria that all genomes identified as USA300 by GMI form a distinct subclade before they are labeled as USA300 i.e. PVL or SCCMECIVa positive genomes that grouped separately from other USA300 strains in the phylogenetic tree were not labeled as USA300.

We detected SCCMec cassettes in 1588 genomes of which 1358 were SCCMecIVa positive. We also found 1431 PVL positive genomes using BLASTn search with PVL encoding genes from USA300 TCH1516 (USA300HOU_RS07645, USA300HOU_RS07650) as reference (**Supplementary Table 5**). Lastly, we reconstructed the CC8 phylogenetic tree based on core Single Nucleotide Polymorphisms (SNPs) and rooted the tree using strain D592 (CC5) as an outgroup. The tree was then traversed from reference strain TCH1516 to the CC8 root using ete3, while tracking the total number of genomes, the total number of SCCMec IVa positive genomes and the number of PVL positive genomes in each root [42]. The root of USA300 was placed manually where the number of total genomes kept increasing while the number of PVL and SCCMec positive genomes plateaued. All strains in the clade represented by the USA300 root were classified as USA300 regardless of their SCCMec or PVL status.

DBGWAS and k-mer linkage calculations

DBGWAS was used to enrich mutations unique to USA300 strains. Alleles with frequency less than 0.05 were filtered (-maf 0.05) and all components enriched with q-values less than 0.05 were documented (-SFF q0.05). Genome-wide linkage was estimated by Pearson correlation of the presence/ absence of enriched k-mers and distance was measured based on the k-mer alignment to the reference TCH1516 genome.

To determine the enriched ‘genetic event’ (e.g. SNP, indel, mobile genetic element etc), the graph output from DBGWAS was first loaded onto a networkX model[43]. All nodes in the graph with frequency lower than 0.05 were discarded. MGEs were identified if all significant nodes from DBGWAS had higher frequencies in one strain, e.g. all nodes associated with SCCMec had higher frequencies in USA300 strains. To find SNPs and smaller indel events, the

networkx was used to find cycles in the graph, which results from bifurcation and eventual re-collapse of Debruijn graphs around mutations. For each cycle, the 'end nodes' representing the start and end of the bifurcation were identified by finding the nodes in the cycle with highest frequency across all samples. As 'end nodes' are present in both case and control samples, they will have higher frequency than other nodes in the cycle which are specific to either case or control. Once the end nodes are identified, the two paths around the bifurcations representing the case and control specific sequences were identified using the shortest path algorithm in networkx. The sequences from nodes of each path were concatenated, changing the sequences to reverse complements and removing overlaps in sequences when required. The concatenated sequences from each path were then compared using BioPython pairwise global alignments to find the SNPs or indels that differentiate the sequences from case and control[44]. If reference sequences are passed, the concatenated sequences are aligned to the reference sequences using BLAST and mutation positions were converted from k-mer positions to positions in the reference genomes. The code used for this analysis can be found in https://github.com/SBRG/dbgwas_network_analysis.

Mapping mutation hotspots with position specific Shannon entropy

For each of the CDS with enriched mutations, the PATRIC local protein family (PLfam) was identified based on the reference TCH1516 genome. All available protein sequences for each CDS PLfam were downloaded and filtered for *S. aureus* sequences. The multilocus sequence type (MLST) of the source genome of each downloaded sequence was mapped using the PATRIC database. The sequences were divided into ST8 and non ST8 and ST239 sequences were filtered. MAFFT was used for multiple alignment and position-specific Shannon entropy was calculated on the aligned file[45]. The entropy is calculated as:

$$H(X) = -\sum_{i=1}^{n} P(x_i) log_2 P(x_i)$$

where n is the total number of unique amino acids in the position and $P(x_i)$ is the probability of finding the given amino acid.

Calculating strain-specific iModulons with independent component analysis

ICA of RNA sequencing data was performed using the pymodulon package [28]. Using the package, all available RNA sequencing data for non-USA300 and USA300 strains were downloaded, run through the QC/QA pipeline, manually curated for metadata and aligned to the TCH1516 genome (NC_010079, NC_012417, NC_010063). The combined data was then transformed into log-TPM and normalized to a single reference condition (SRX3760886, SRX3760891). This is in contrast to other ICA models that normalize the data to project specific reference conditions to reduce batch effects. However, normalizing to project specific control conditions also erases the strain specific information as almost all BioProjects contain data from only one isolate (e.g. NCTC8325, TCH1516, LAC etc). ICA was then run as previously

described in chapter 3. The activities of the output iModulons were manually parsed to look for iModulons with the largest strain specific differences.

Fur box motif search

isdH genes in all the genomes were first clustered using CD-HIT with identity and coverage minimum of 0.8[33]. All annotated *isdH* genes fell within a single cluster. For each genome, the 100 bp upstream region was then extracted and used for motif search. Motif search for the Fur box was conducted using the FIMO package from the MEME suite with default settings[46]. The *S. aureus* strain NCTC8325 Fur motif from collecTF was used as a reference[47].

Code and Data Availability

The genome sequences and the RNA-sequencing data used in this study are publicly available (See Supplementary Table 1 and 3 respectively). The code used for analysis, the intermediate files and models are available on github (https://github.com/sapoudel/USA300GWASPUB).

References

1. Young, B. C. *et al.* Panton-Valentine leucocidin is the key determinant of Staphylococcus aureus pyomyositis in a bacterial GWAS. *Elife* **8**, (2019).
2. Bosi, E. *et al.* Comparative genome-scale modelling of Staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E3801–9 (2016).
3. Choudhary, K. S. *et al.* The Staphylococcus aureus Two-Component System AgrAC Displays Four Distinct Genomic Arrangements That Delineate Genomic Virulence Factor Signatures. *Front. Microbiol.* **9**, 1082 (2018).
4. Correction for Copin et al., Sequential evolution of virulence and resistance during clonal spread of community-acquired methicillin-resistant Staphylococcus aureus. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 4747 (2019).
5. Krishna, A., Holden, M. T. G., Peacock, S. J., Edwards, A. M. & Wigneshweraraj, S. Naturally occurring polymorphisms in the virulence regulator Rsp modulate Staphylococcus aureus survival in blood and antibiotic susceptibility. *Microbiology* **164**, 1189–1195 (2018).
6. Earle, S. G. *et al.* Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat Microbiol* **1**, 16041 (2016).
7. Diep, B. A. *et al.* The arginine catabolic mobile element and staphylococcal chromosomal cassette mec linkage: convergence of virulence and resistance in the USA300 clone of methicillin-resistant Staphylococcus aureus. *J. Infect. Dis.* **197**, 1523–1530 (2008).
8. Faden, H. *et al.* Importance of colonization site in the current epidemic of staphylococcal skin abscesses. *Pediatrics* **125**, e618–24 (2010).
9. Steinig, E. J. *et al.* Evolution and Global Transmission of a Multidrug-Resistant, Community-Associated Methicillin-Resistant Staphylococcus aureus Lineage from the Indian Subcontinent. *MBio* **10**, (2019).

10. Challagundla, L. *et al.* Phylogenomic Classification and the Evolution of Clonal Complex 5 Methicillin-Resistant Staphylococcus aureus in the Western Hemisphere. *Front. Microbiol.* **9**, 1901 (2018).
11. Bal, A. M. *et al.* Genomic insights into the emergence and spread of international clones of healthcare-, community-and livestock-associated meticillin-resistant Staphylococcus aureus: blurring of the traditional definitions. *Journal of Global Antimicrobial Resistance* **6**, 95–101 (2016).
12. Uhlemann, A.-C. *et al.* Molecular tracing of the emergence, diversification, and transmission of S. aureus sequence type 8 in a New York community. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 6738–6743 (2014).
13. Challagundla, L. *et al.* Range Expansion and the Origin of USA300 North American Epidemic Methicillin-Resistant Staphylococcus aureus. *MBio* **9**, (2018).
14. Everitt, R. G. *et al.* Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus. *Nat. Commun.* **5**, 3956 (2014).
15. Thurlow, L. R., Joshi, G. S. & Richardson, A. R. Virulence strategies of the dominant USA300 lineage of community-associated methicillin-resistant Staphylococcus aureus (CA-MRSA). *FEMS Immunol. Med. Microbiol.* **65**, 5–22 (2012).
16. Boyle-Vavra, S. *et al.* USA300 and USA500 clonal lineages of Staphylococcus aureus do not produce a capsular polysaccharide due to conserved mutations in the cap5 locus. *MBio* **6**, (2015).
17. Nishizaki, S. S. *et al.* Predicting the effects of SNPs on transcription factor binding affinity. *Bioinformatics* **36**, 364–372 (2020).
18. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7**, e46688 (2012).
19. Kavvas, E. S., Yang, L., Monk, J. M., Heckmann, D. & Palsson, B. O. A biochemically-interpretable machine learning classifier for microbial GWAS. *Nat. Commun.* **11**, 2580

(2020).

20. Jaillard, M. *et al.* A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *PLoS Genet.* **14**, e1007758 (2018).

21. Sastry, A. V. *et al.* The Escherichia coli transcriptome mostly consists of independently regulated modules. *Nat. Commun.* **10**, 5536 (2019).

22. Jones, M. B. *et al.* Genomic and transcriptomic differences in community acquired methicillin resistant Staphylococcus aureus USA300 and USA400 strains. *BMC Genomics* **15**, 1145 (2014).

23. Iqbal, Z. *et al.* Comparative virulence studies and transcriptome analysis of Staphylococcus aureus strains isolated from animals. *Sci. Rep.* **6**, 35442 (2016).

24. Bowers, J. R. *et al.* Improved Subtyping of Staphylococcus aureus Clonal Complex 8 Strains Based on Whole-Genome Phylogenetic Analysis. *mSphere* **3**, (2018).

25. Wattam, A. R. *et al.* PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.* **42**, D581–91 (2014).

26. Robinson, D. A. & Enright, M. C. Evolution of Staphylococcus aureus by large chromosomal replacements. *J. Bacteriol.* **186**, 1060–1064 (2004).

27. Hyun, J. C., Monk, J. M. & Palsson, B. O. Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity. *BMC Genomics* **23**, 7 (2022).

28. Sastry, A. V. *et al.* Mining all publicly available expression data to compute dynamic microbial transcriptional regulatory networks. *bioRxiv* 2021.07.01.450581 (2021) doi:10.1101/2021.07.01.450581.

29. Yuan, Y. *et al.* Pan-genomic analysis of transcriptional modules across Salmonella Typhimurium reveals the regulatory landscape of different strains. *bioRxiv* 2022.01.11.475931 (2022) doi:10.1101/2022.01.11.475931.

30. Prados, J., Linder, P. & Redder, P. TSS-EMOTE, a refined protocol for a more complete

and less biased global mapping of transcription start sites in bacterial pathogens. *BMC Genomics* **17**, 849 (2016).

31. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
32. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
33. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
34. Eric Jones, Travis Oliphant, Pearu Peterson and others. *SciPy: Open Source Scientific Tools for Python*. (2001-).
35. Shakya, M. *et al.* Standardized phylogenetic and molecular evolutionary analysis applied to species across the microbial tree of life. *Sci. Rep.* **10**, 1723 (2020).
36. Highlander, S. K. *et al.* Subtle genetic changes enhance virulence of methicillin resistant and sensitive Staphylococcus aureus. *BMC Microbiol.* **7**, 99 (2007).
37. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
38. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
39. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
40. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
41. Kaya, H. *et al.* SCCmecFinder, a Web-Based Tool for Typing of Staphylococcal Cassette Chromosome mec in Staphylococcus aureus Using Whole-Genome Sequence Data. *mSphere* **3**, (2018).

42. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).

43. Hagberg, A., Swart, P. & S Chult, D. *Exploring network structure, dynamics, and function using networkx*. https://www.osti.gov/biblio/960616 (2008).

44. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

45. Katoh, K., Misawa, K., Kuma, K.-I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).

46. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–8 (2009).

47. Kılıç, S., White, E. R., Sagitova, D. M., Cornish, J. P. & Erill, I. CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria. *Nucleic Acids Res.* **42**, D156–D160 (2013).