eLife

Microbiology and Infectious Disease
Computational and Systems Biology

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

Reviewed Preprint
v2 • July 11, 2024
Revised by authors

Reviewed Preprint
v1 • November 8, 2023

Saugat Poudel, Jason Hyun, Ying Hefner, Jon Monk, Victor Nizet, Bernhard O Palsson

Department of Bioengineering, University of California San Diego • Palmona Pathogenomics • Collaborative to Halt Antibiotic-Resistant Microbes (CHARM), Department of Pediatrics, University of California San Diego • Department of Pediatrics, University of California San Diego

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 ancestral 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 Single Nucleotide Polymorphism (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.

eLife assessment

This study presents **valuable** findings on core genome mutations that might have driven the emergence of the Staphylococcus aureus lineage USA300, a frequent cause of community-acquired infections. The authors present a **solid** novel approach that combines genome-wide association studies and RNA-expression analyses, both applied to extensive publicly available datasets. This approach generated an intriguing hypothesis that should be validated experimentally. The work will interest microbiologists working in genomic epidemiology and phenotype-genotype association studies.

https://doi.org/10.7554/eLife.90668.2.sa2

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 population structure 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].

Emerging clonal complexes, 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 pyogenic skin infections, increased ability to colonize locations outside of the nasopharynx etc[1,7,8]. As these strains often emerge clonally from closely related 'ancestral 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 as their effect on phenotype is easier to determine than that of point mutations[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. Taking advantage of these modeling techniques, we use a reconstruction of the transcriptional regulatory network (TRN) of CC8 strains to find USA300 specific mutations that are associated with changes in gene regulation.

First, we used De Bruijn graph GWAS (DBGWAS) to discover enriched mutations associated with the 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 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 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 downloaded 2033 *S.* aureus genomes for analysis and excluded six of them with genome length of less than 2.5 million base pairs. 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**). The collection formed a closed pangenome, as adding new genomes did not introduce many new genes (**Figure 1a**), suggesting that our collection had a good gene level coverage of the CC8 pangenome. We confirmed the pangenome coverage with Roary (**Figure S1b**). To get a higher resolution view of these genomes, we surveyed unique alleles within the ORFs and in the 300 base-pair 5' upstream and 3' downstream sequences. 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. This is reflective of the fact that most of *S. aureus* genome sequence comprises of ORFs e.g. ~84% of TCH1516 genome is part of an ORF.

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 SCC*mec* IVa or the presence of Panton-Valentine Leukocidin (PVL) in case of methicillin sensitive *S. aureus* (MSSA).

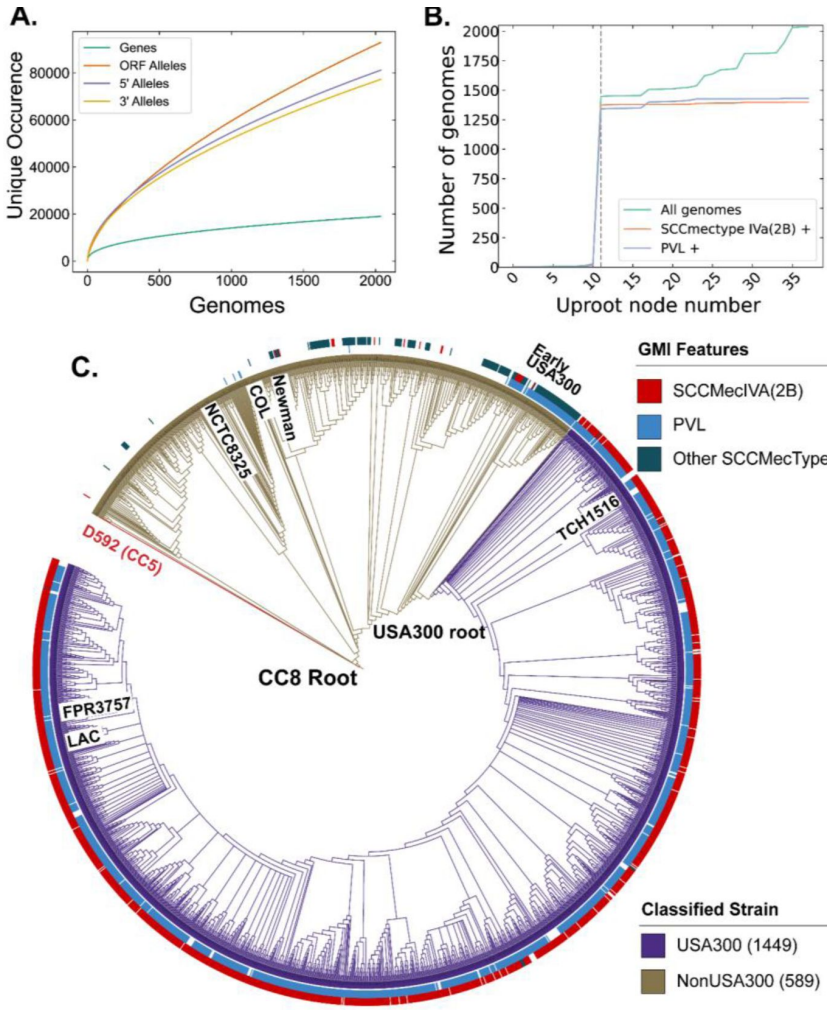


Figure 1.

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 SCC*mec* IVa, 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.

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 SCC*mec* IVa 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 SCC*mec* IVa 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 S1c**). 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 SCC*mec* types and therefore are likely to be clinically distinct from the 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]. We used 2030 genomes for this analysis; the 2027 genomes in pangenomics analysis above were "spiked" with three well known CC8 genomes-TCH1516, COL, and Newman-to help annotate the DBGWAS unitigs. 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-*SCCmec* IVA (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.

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 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**).

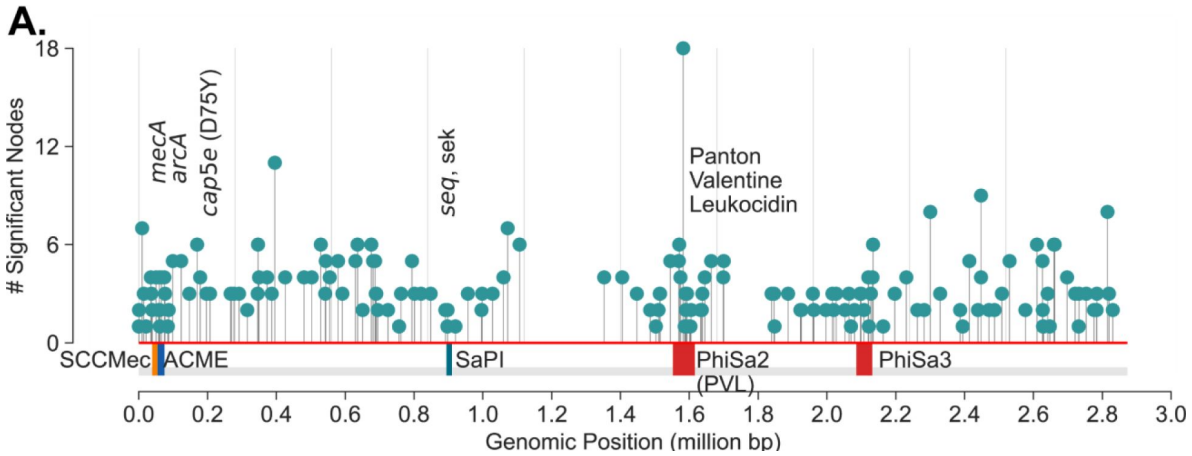


Figure 2.

USA300 strains associated mutations.

(a) DBGWAS recovers components associated with USA300 previously described markers of USA300 strains including *mecA* (SCC*mec* 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).

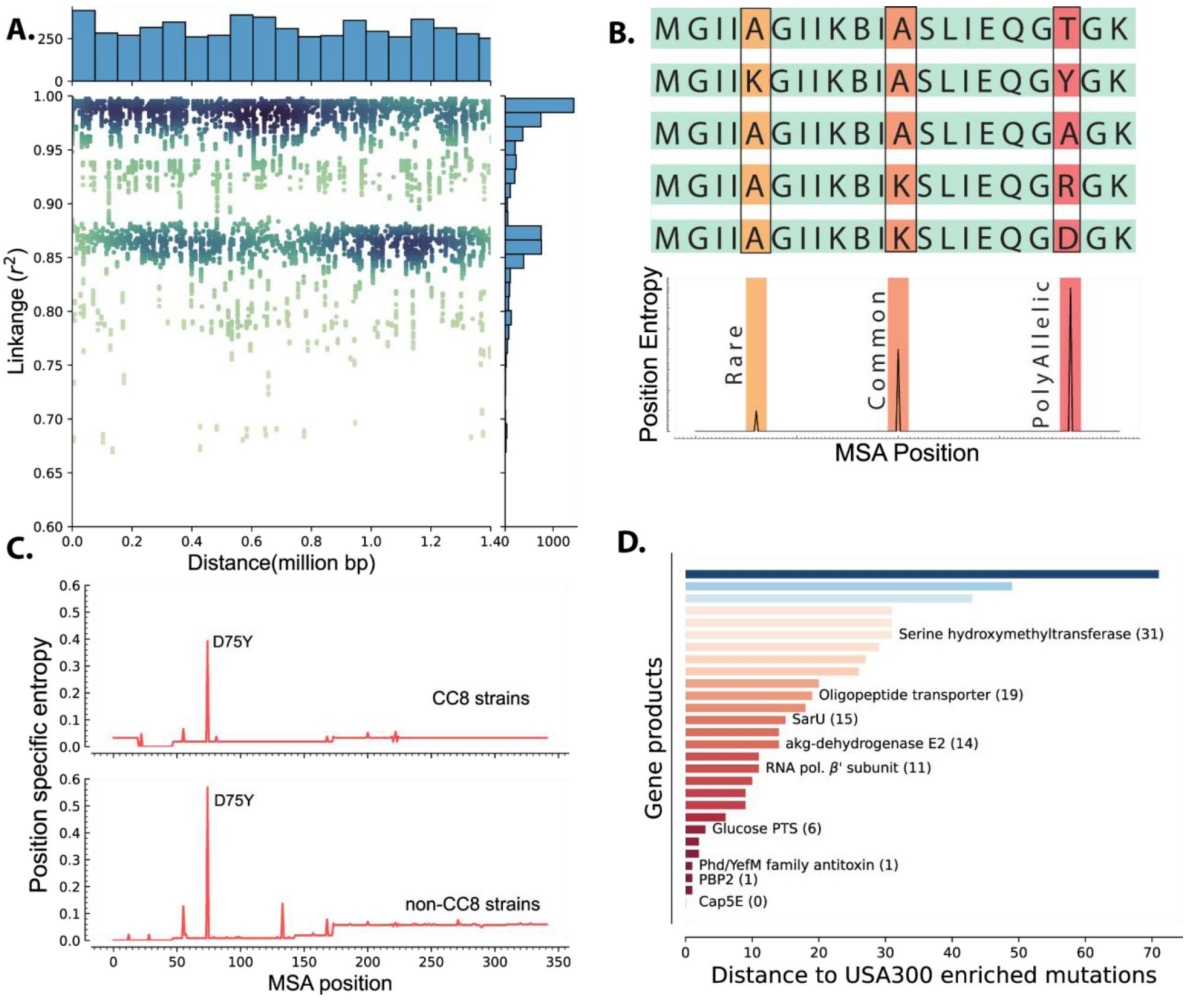


Figure 3.

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.

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.

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**).

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, SCC*mec*, 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 S4**). 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 make a headway through the impasse created by the high linkage disequilibrium and identify GWAS-enriched mutations that are also associated with measurable phenotypic changes in the TRN. From the combined RNA sequencing dataset of USA300 and non-USA300 strains, ICA calculated two iModulons 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 SCC*mec* because they have zero expression level in non-

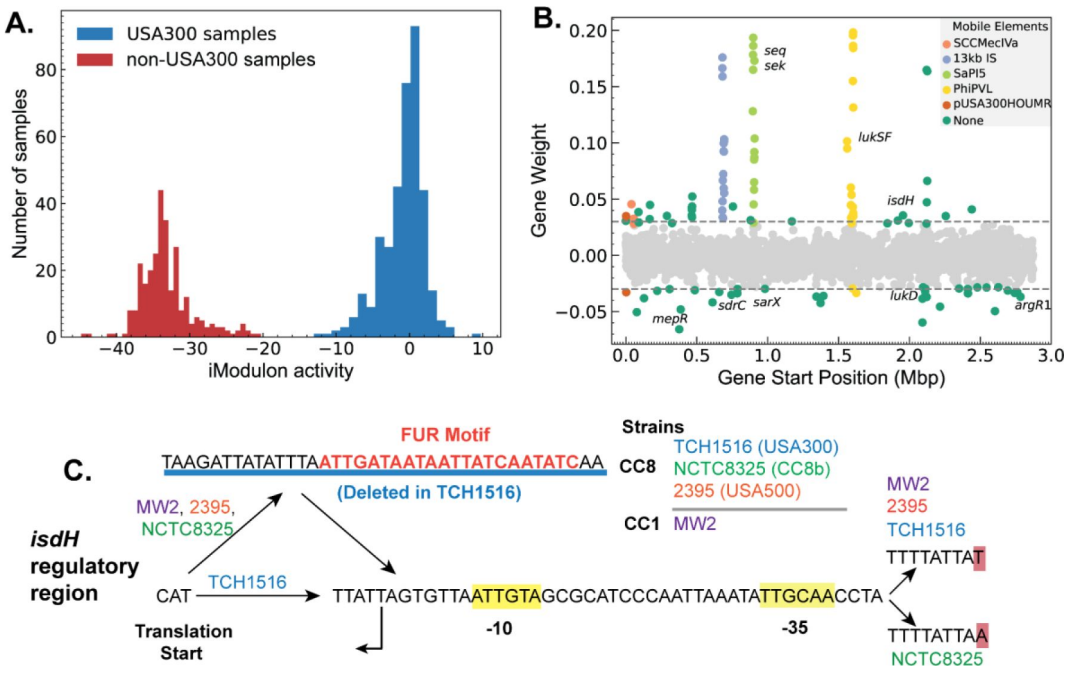


Figure 4.

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.

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. This gene encodes IsdH, a surface receptor that binds to human hemoglobin, causing it to release the heme[31]. It is part of an arsenal of *S. aureus* iron sequestration proteins including Staphyloferrins and ferrous iron transporters that it uses to compete with the host for essential iron[32]. Despite having many different pathways for obtaining iron, it has been observed that *S. aureus* prefers heme as its iron source over transferrin-bound iron[33]. Our analysis shows that preference for heme is reflected in the genomic and transcriptomic signature of USA300 as a deletion of Fur binding region upstream of *isdH* and subsequent increase in its expression. Combining GWAS with large-scale transcriptomic modeling was therefore able to predict potential causal mutations contributing 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. This misalignment points to the limit of our approach. Most other phenotypes of clinical interest such as antibiotic resistance may not separate cleanly into distinct clades. In those cases, it is not obvious which strains should be chosen as the reference strain for RNA-sequencing and subsequent TRN reconstruction. The choice of reference strain as well as the choices in the RNA-sequencing sample conditions will impact which association between mutations and changes in gene regulations are uncovered.

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 clonal 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 [34,35]. 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 (v4.6) with minimum identity (-T) and minimum alignment length (-aL) of 80% and word size of 5 (-w 5)[36]. 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[37]. 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.

Roary (v3.13.0) was used with -i 95 flag to confirm the output of our pangenome analysis[38].

Reconstructing CC8 phylogenetic tree

The phylogenetic tree was reconstructed using the standardized PHaME pipeline on the PATRIC sequences that passed the QC/QA[39]. Using the pipeline, the contigs and sequences were aligned to the reference TCH1516 genome NC_010079 and plasmids NC_012417, NC_010063[40] and 24881 core SNPs at were calculated. The core SNPs were then used to estimate the phylogenetic tree using IQ-TREE(v1.6.7) run with 1000 bootstraps and utilizing the ultrafast bootstrap[41,42]. The tree was built using the "TVMe+ASC+G4" model as suggested by the IQ-TREE ModelFinder[43]. Finally, iTOL was used to visualize, annotate and root the tree with the USA100 D592 (NZ_CP035791) from CC5 as the outgroup [44].

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 SCC*mec* IVa cassette. We detected SCC*mec* types using SCC*mec*Finder (v1.2), and only those genomes where the cassette could be identified by both BLASTn and k-mer based methods were marked as positive[45]. PVL was detected using nucleotide BLAST(v2.2.31). 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 SCC*mec* IVa positive genomes that grouped separately from other USA300 strains in the phylogenetic tree were not labeled as USA300. To find the root of the USA300 strains in the phylogenetic tree, the genomes in the tree were first annotated by their PVL and SCC*mec* status. Then the tree traversed from leaf to root starting from known USA300 strains – TCH1516 and FPR3757-while keeping track of the number of descendant genomes from the current root that contained known markers SCC*mec* IVa and PVL. The node where the number of genomes with the markers started flatlining was marked as the root of USA300.

We detected SCC*mec* cassettes in 1588 genomes of which 1358 were SCC*mec* IVa 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 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 SCC*mec* IVa positive genomes and the number of PVL positive genomes in each root [46]. The root of USA300 was placed manually where the number of total genomes kept increasing while the number of PVL and SCC*mec* positive genomes plateaued. All strains in the clade represented by the USA300 root were classified as USA300 regardless of their SCC*mec* or PVL status.

DBGWAS and k-mer linkage calculations

DBGWAS (v0.5.4) was used to enrich mutations unique to USA300 strains using default k-mer size of 31 (-k 31) and neighborhood size of 5 (-nh 5). Alleles with frequency less than 0.1 were filtered (-maf 0.1) and all components enriched with q-values less than 0.05 were documented (-SFF q0.05). Genome-wide linkage was estimated by Pearson correlation (calculated with built-in Pandas function) of the presence/ absence of enriched k-mers and distance was measured based on the k-mer alignment to the reference TCH1516 genome as determined by BLASTn.

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[47]. 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 SCC*mec* 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 (v1.83) pairwise global alignments to find the SNPs or indels that differentiate the sequences from case and control[48]. If reference sequences are passed, the concatenated sequences are aligned to the reference sequences using nucleotide 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 online PATRIC website was used to find and filter the target sequences. 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[49]. 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

A detailed version of the methods for RNA-sequencing and ICA analysis is available as Supplementary Note 2. 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 (Transcripts per Million) and normalized to a single reference condition (SRX3760886, SRX3760891). This contrasts with 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 (see Supplementary Note 2)[21]. 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[36]. 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 (v5.1.0) with default settings[50]. The *S. aureus* strain NCTC8325 Fur motif from collecTF was used as a reference[51].

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*).

Supplementary Materials

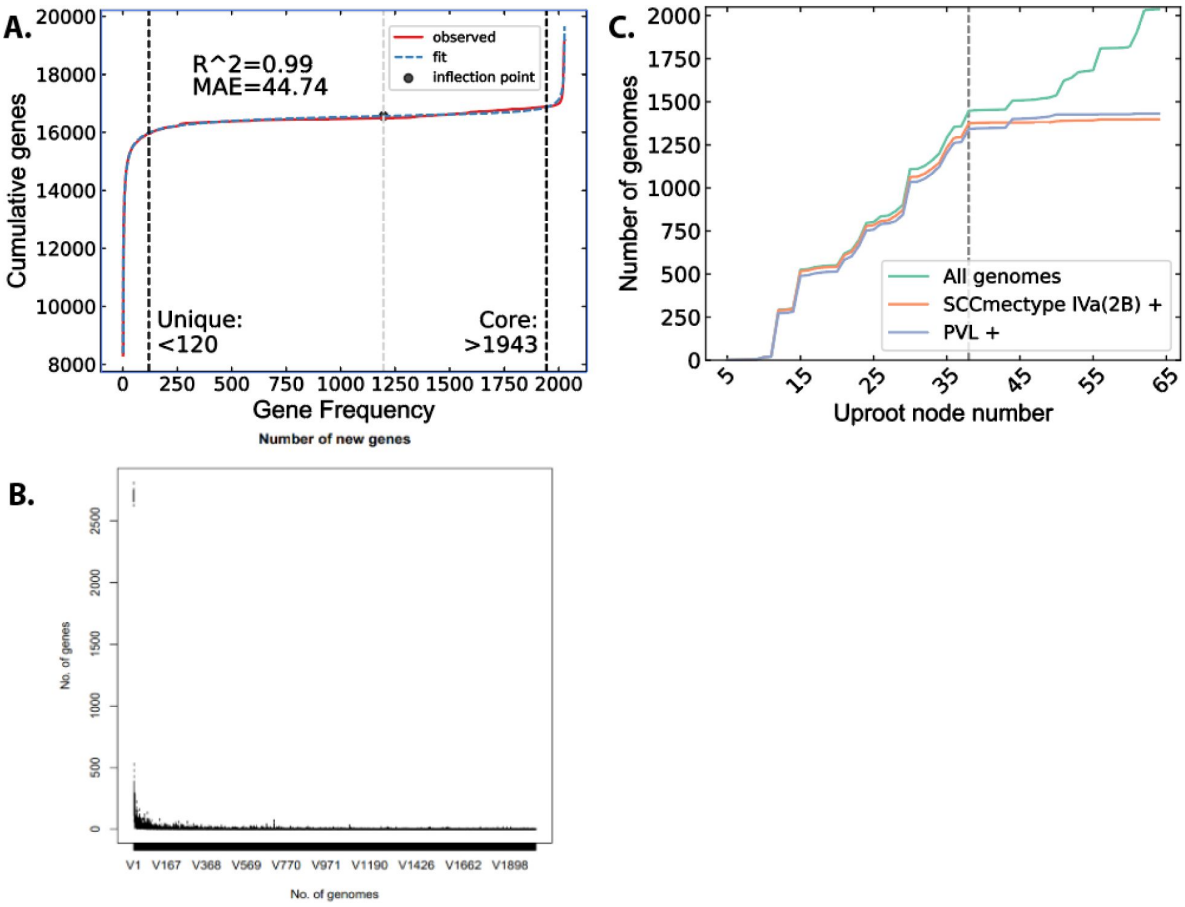


Figure S1.

Pangenome analysis and strain classification.

(a) Cumulative distribution of unique genes used to fit the pangenomic parameters. The core and unique genes threshold were calculated at 90% of the distance from the inflection point (black dot) of the curve. (b) Analysis with Roary confirmed that adding new genomes to the analysis collection were unlikely to introduce many new genes which indicates a good gene level coverage of the CC8 clade. (c) SCC*mec* and PVL distribution in the CC8 tree as it is traversed up from the FPR3757 leaf towards the root. Starting from FPR3757 gives the same delineation between USA300 and non-USA300 genomes as the search that starts from TCH1516.

Figure S2.

S. aureus MLST distribution of genomes from PATRIC used in this study.

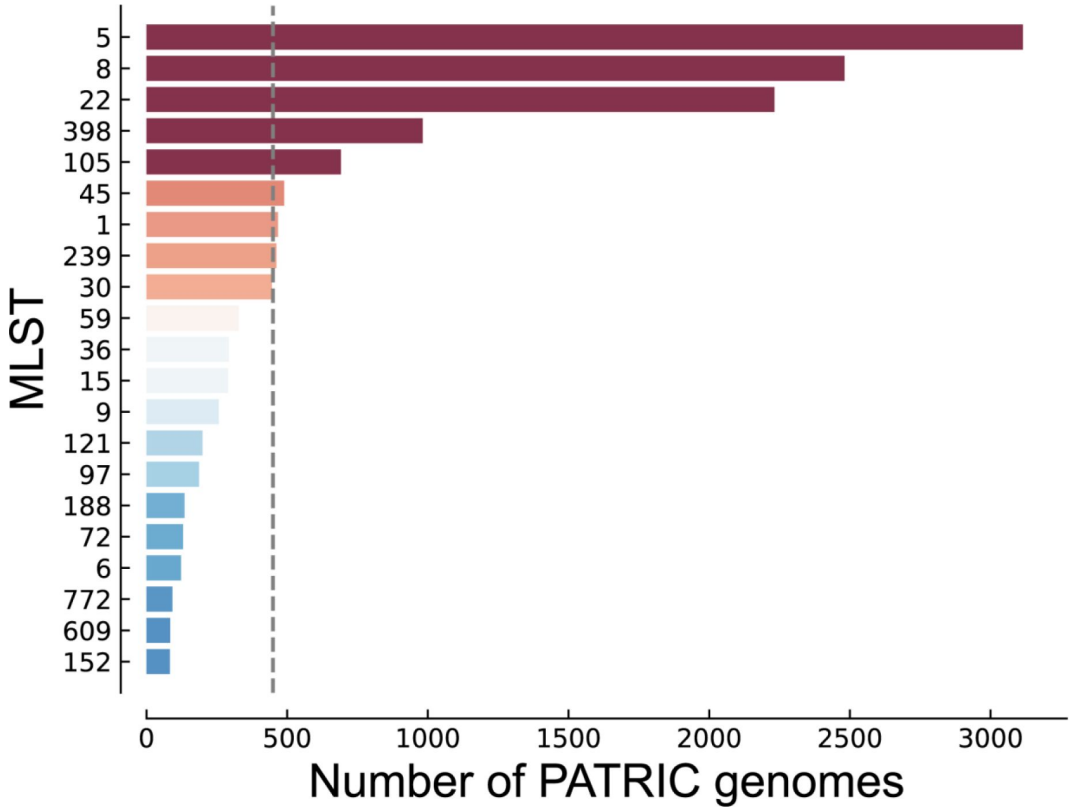
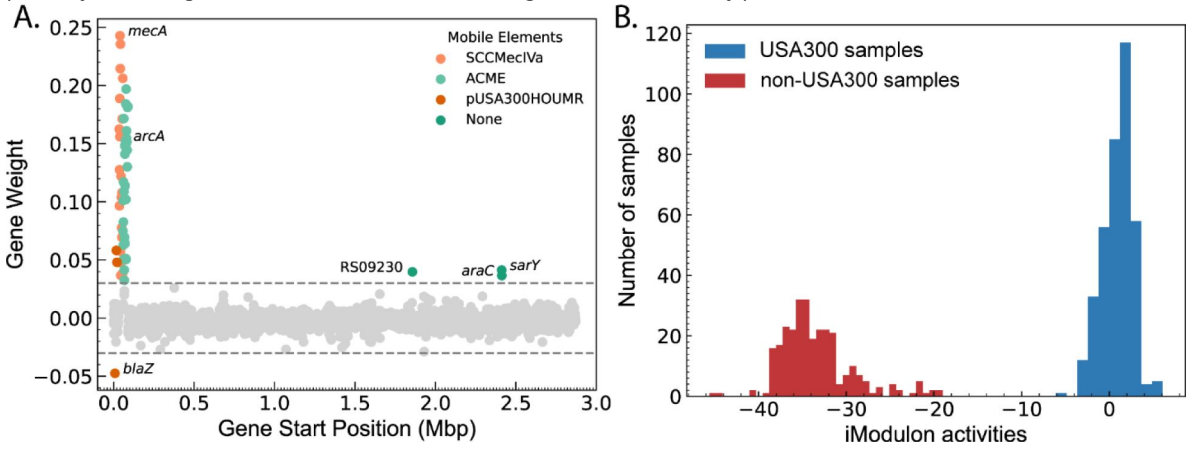


Figure S3.

SCC*mec*/ACME iModulons weighting and strain-specific activity.

(a) The activity of the SCC*mec*/ACME iModulon shows clear strain-specific separation. (b) Gene weighting for the iModulon primarily containing SCC*mec* and ACME. Genes encoding SarY and AraC family proteins were also enriched.



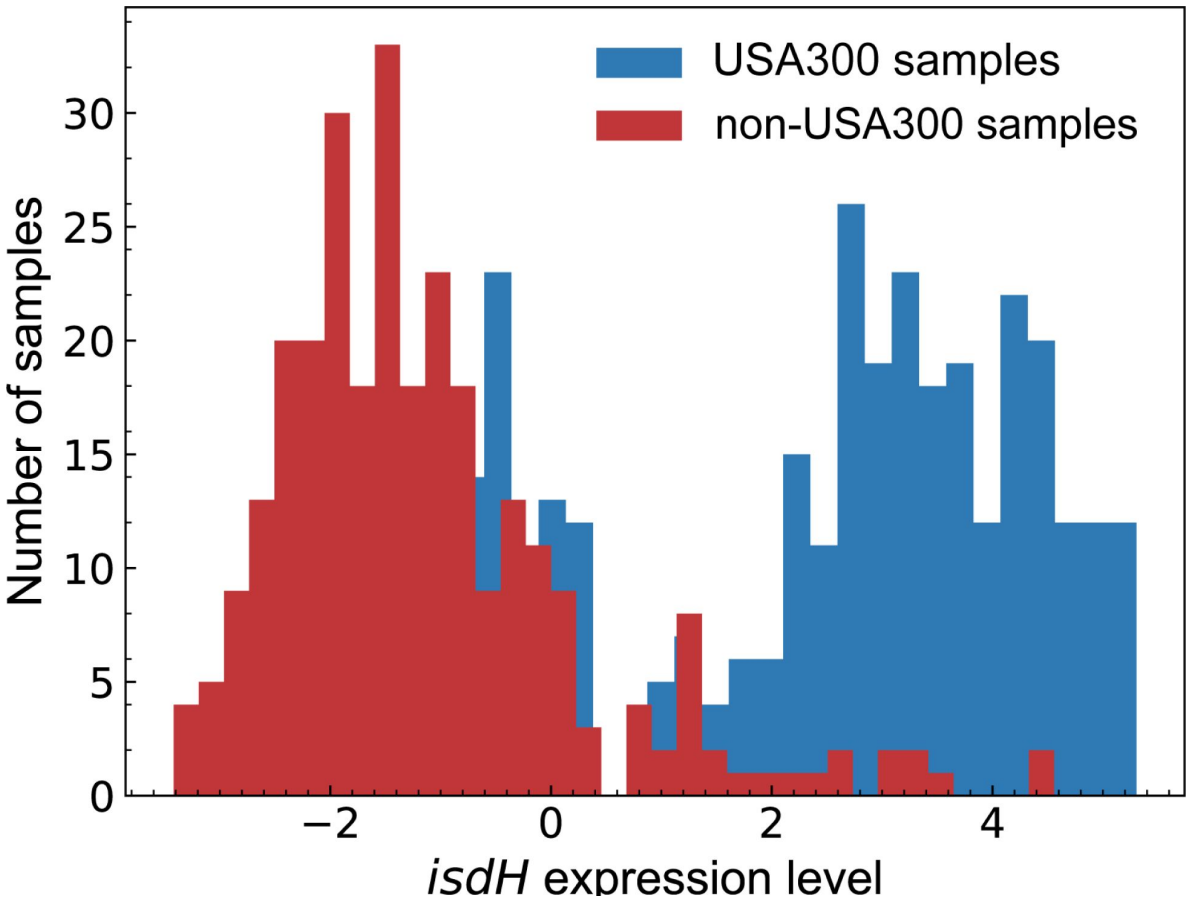


Figure S4.

***isdH* gene shows strain-specific gene expression level.**

The increased expression level in USA300 is in line with the deletion of the Fur repressor binding site. The expression levels are log-TPM centered on the expression profile from the TCH1516 strain grown in RPMI + 10%LB.

Supplementary Note 1. Converting DBGWAS enriched component graphs to interpretable mutations

Currently, DBGWAS outputs the graph consisting of the nodes with the enriched k-mers and its genetic neighborhood, but does not automatically yield the exact mutation associated with each of the significant nodes. By analyzing the structure of the component graphs with networkX, we were able to extract the exact genetic changes represented by these components [1]. Mobile genetic elements (MGEs) and large indels can be identified by a series of nodes that are all enriched in either USA300 or non-USA300 genomes (**Figure S5a**). The enrichment of multiple sequential k-mers in only one of the groups implies deletion of the sequence (or conversely insertion) in the other group. SNPs and indels smaller than the k-mer-size on the other hand form 'cycles' containing significant nodes (**Figure S5b**). Consequently, the k-mers in the nodes of each of the 'paths' around the cycle represent sequences unique to either case or control group. The enriched mutation can therefore be extracted by comparing the sequences with global alignment. Lastly, the unique sequences from each path can also be mapped to reference genomes if needed. The exact mutations used in the subsequent sections were extracted from the components using this method (**Supplementary Table 2**). The code used to automatically analyze the network has been uploaded to github (*https://github.com/SBRG/dbgwas-network*).

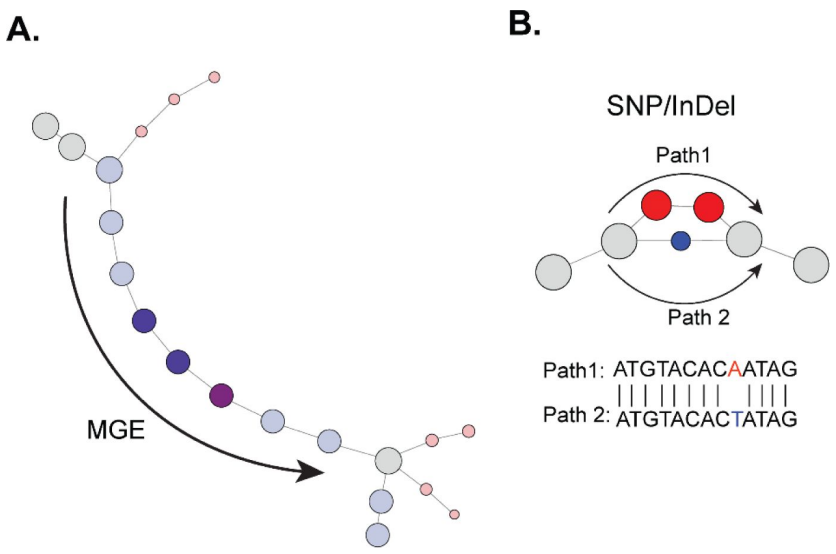


Figure S5.

Interpreting DBGWAS output.

(a) Example of components associated with Mobile genetic elements (MGE)s; components have a series of nodes that are enriched in one group (blue circles). (b) Example of components associated with SNP. Component graph contains a cycle around the mutation location with the paths from the cycle forming a sequence unique to either case or control group. Aligning the sequences reveals the enriched mutation.

Supplementary Note 2. Creating iModulons for CC8 Clade Staphylococcus aureus

We followed our previously established pipeline to generate iModulons for CC8 clade Staphylococcus aureus[2]. We began by collecting all available RNA-sequencing data and metadata for *S. aureus* strains from Sequence Read Archives (SRA) that belonged to the CC8 clade. Most of the sequences were from well-studied CC8 strains-TCH1516, FPR3757, LAC, Newman, NCTC8325. Others had less specific labels e.g. "USA300" but still belonged to CC8. The fastq files for the samples were trimmed with TrimGalore (v0.6.5) and aligned to reference TCH1516 genome (NC_010079, NC_012417, NC_010063) with bowtie2 (v1.2.3) [3,4]. The gene-specific read counts were calculated with HTSeqCount (v2.0.1) using the intersection-strict criteria. The number of mapped reads were then normalized to transcripts per million (TPM) and log-transformed (log-TPM).

Before using the data, the quality of the reads and alignment were assessed using FastQC and MultiQC (v 1.11)[5,6]. Any samples failing 'per base sequence quality', 'per sequence quality score,' 'per base n content', or 'adapter content' were dropped. Additionally, we also removed samples with less than 500,000 reads aligned to the reference. Lastly, samples that did not contain replicates or those with replicates with Pearson Correlation coefficients less than 0.9 were also excluded. We then collected additional metadata for the remaining 670 RNA-sequencing samples including growth conditions, genetic changes, associated experiment etc. The log-TPM were then centered to the reference condition of S.*aureus* TCH1516 grown in RPMI + 10 % LB. By centering data from non-USA300 strains on USA300 reference, ICA is able to pick up strain-specific regulatory changes e.g. ICA captures the activity of Fur transcription factor as a linear combination of Fur iModulon containing gene regulated by Fur and a second 'strain-specific' iModulon that captures differences between USA300 and non-USA300 strains (**Figure S3**).

We applied FastICA to the centered log-TPM to calculate the M and the A matrix which respectively describe the iModulon structure and their activities[7,8]. To find the best possible model, we first had to compute an optimal number of stable components. As FastICA is non-deterministic, each iteration yields a slightly different component weightings and activity levels. It may also yield "spurious" components that are only present in a subset of runs. To find stable components, we ran ICA 100 times with a random seed. Similar components (e.g. same component containing Fur associated iModulon) from different iterations which may have slightly different weightings were detected by clustering with DBSCAN. Only components that appear in every run were accepted. When running ICA, users must also provide the number of desired components that the data will be decomposed into. Decomposition into too few components could lead to signals from several transcription factors being combined into a single components while over decomposition leads to many unstable and "single-gene" iModulons that likely capture noise in the data. To find the optimal number of components we used a heuristic method, OptICA, which runs ICA with different numbers of input components from 10 to 340 and suggests an optimal component that minimizes single-gene iModulons while maximizing robust components. Based on this heuristic, the final model was built with 270 components as input, 148 of which were determined to be robust components.

In each component, we labeled a gene as being part of an iModulon if their weighting in that component did not fall within a Gaussian distribution as determined by D'Agostino's test. The genes in each iModulon was then compared to genomic features (e.g. regulons, phage, mobile cassettes etc; see 'TRN' object in the model), and was determined to be associated with the feature if there was significant overlap between the two groups (hypergeometric test; adjusted p-value <

0.05, precision >= 0.5 and coverage >= 0.2). We also manually curated other iModulons associated with other features e.g. iModulon where all member genes associated with translation were labeled 'Translation iModulon.'

References

1. Young B. C., et al. (2019) **Panton-Valentine leucocidin is the key determinant of Staphylococcus aureus pyomyositis in a bacterial GWAS** *Elife* **8**

2. Bosi E., et al. (2016) **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

3. Choudhary K. S., et al. (2018) **The Staphylococcus aureus Two-Component System AgrAC Displays Four Distinct Genomic Arrangements That Delineate Genomic Virulence Factor Signatures** *Front. Microbiol* **9**

4. Copin Correction for, et al. (2019) **Sequential evolution of virulence and resistance during clonal spread of community-acquired methicillin-resistant Staphylococcus aureus** *Proc. Natl. Acad. Sci. U. S. A* **116**

5. Krishna A., Holden M. T. G., Peacock S. J., Edwards A. M., Wigneshweraraj S (2018) **Naturally occurring polymorphisms in the virulence regulator Rsp modulate Staphylococcus aureus survival in blood and antibiotic susceptibility** *Microbiology* **164**:1189–1195

6. Earle S. G., et al. (2016) **Identifying lineage effects when controlling for population structure improves power in bacterial association studies** *Nat Microbiol* **1**

7. Diep B. A., et al. (2008) **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

8. Faden H., et al. (2010) **Importance of colonization site in the current epidemic of staphylococcal skin abscesses** *Pediatrics* **125**:e618–24

9. Steinig E. J., et al. (2019) **Evolution and Global Transmission of a Multidrug-Resistant, Community-Associated Methicillin-Resistant Staphylococcus aureus Lineage from the Indian Subcontinent** *MBio* **10**

10. Challagundla L., et al. (2018) **Phylogenomic Classification and the Evolution of Clonal Complex 5 Methicillin-Resistant Staphylococcus aureus in the Western Hemisphere** *Front. Microbiol* **9**

11. Bal A. M., et al. (2016) **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

12. Uhlemann A.-C., et al. (2014) **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

13. Challagundla L., et al. (2018) **Range Expansion and the Origin of USA300 North American Epidemic Methicillin-Resistant Staphylococcus aureus** *MBio* **9**

14. Everitt R. G., et al. (2014) **Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus** *Nat. Commun* **5**

15. Thurlow L. R., Joshi G. S., Richardson A. R (2012) **Virulence strategies of the dominant USA300 lineage of community-associated methicillin-resistant Staphylococcus aureus (CA-MRSA)** *FEMS Immunol. Med. Microbiol* **65**:5–22

16. Boyle-Vavra S., et al. (2015) **USA300 and USA500 clonal lineages of Staphylococcus aureus do not produce a capsular polysaccharide due to conserved mutations in the cap5 locus** *MBio* **6**

17. Nishizaki S. S., et al. (2020) **Predicting the effects of SNPs on transcription factor binding affinity** *Bioinformatics* **36**:364–372

18. Choi Y., Sims G. E., Murphy S., Miller J. R., Chan A. P (2012) **Predicting the functional effect of amino acid substitutions and indels** *PLoS One* **7**

19. Kavvas E. S., Yang L., Monk J. M., Heckmann D., Palsson B. O (2020) **A biochemically-interpretable machine learning classifier for microbial GWAS** *Nat. Commun* **11**

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

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

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

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

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

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

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

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

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

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

30. Prados J., Linder P., Redder P (2016) **TSS-EMOTE, a refined protocol for a more complete and less biased global mapping of transcription start sites in bacterial pathogens** *BMC Genomics* **17**

31. Ellis-Guardiola K., et al. (2020) **The Staphylococcus aureus IsdH Receptor Forms a Dynamic Complex with Human Hemoglobin that Triggers Heme Release via Two Distinct Hot Spots** *J. Mol. Biol* **432**:1064–1082

32. van Dijk M. C., de Kruijff R. M., Hagedoorn P.-L (2022) **The Role of Iron in Staphylococcus aureus Infection and Human Disease: A Metal Tug of War at the Host-Microbe Interface** *Front Cell Dev Biol* **10**

33. Skaar E. P., Humayun M., Bae T., DeBord K. L., Schneewind O (2004) **Iron-source preference of Staphylococcus aureus infections** *Science* **305**:1626–1628

34. Baek M., et al. (2021) **Accurate prediction of protein structures and interactions using a three-track neural network** *Science* **373**:871–876

35. Jumper J., et al. (2021) **Highly accurate protein structure prediction with AlphaFold** *Nature* **596**:583–589

36. Fu L., Niu B., Zhu Z., Wu S., Li W (2012) **CD-HIT: accelerated for clustering the next-generation sequencing data** *Bioinformatics* **28**:3150–3152

37. Jones Eric, Oliphant Travis, Peterson Pearu (2001) **SciPy: Open Source Scientific Tools for Python**

38. Page A. J., et al. (2015) **Roary: rapid large-scale prokaryote pan genome analysis** *Bioinformatics* **31**:3691–3693

39. Shakya M., et al. (2020) **Standardized phylogenetic and molecular evolutionary analysis applied to species across the microbial tree of life** *Sci. Rep* **10**

40. Highlander S. K., et al. (2007) **Subtle genetic changes enhance virulence of methicillin resistant and sensitive Staphylococcus aureus** *BMC Microbiol* **7**

41. Minh B. Q., et al. (2020) **IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era** *Mol. Biol. Evol* **37**:1530–1534

42. Hoang D. T., Chernomor O., von Haeseler A., Minh B. Q., Vinh L. S (2018) **UFBoot2: Improving the Ultrafast Bootstrap Approximation** *Mol. Biol. Evol* **35**:518–522

43. Kalyaanamoorthy S., Minh B. Q., Wong T. K. F., von Haeseler A., Jermiin L. S (2017) **ModelFinder: fast model selection for accurate phylogenetic estimates** *Nat. Methods* **14**:587–589

44. Letunic I., Bork P (2021) **Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation** *Nucleic Acids Res* **49**:W293–W296

45. Kaya H., et al. (2018) **SCCmecFinder, a Web-Based Tool for Typing of Staphylococcal Cassette Chromosome mec in Staphylococcus aureus Using Whole-Genome Sequence Data** *mSphere* **3**

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

47. Hagberg A., Swart P., S Chult D. (2008) **Hagberg, A., Swart, P. & S Chult, D. Exploring Network Structure, Dynamics, and Function Using Networkx. https://www.osti.gov/biblio/960616 (2008).**

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

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

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

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

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

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

3. Krueger F. (2015) **Trim galore: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files**

4. Langmead B., Salzberg S. L (2012) **Fast gapped-read alignment with Bowtie 2** *Nat. Methods* **9**:357–359

5. Andrews S. (2010) **FastQC: a quality control tool for high throughput sequence data**

6. Ewels P., Magnusson M., Lundin S., Käller M (2016) **MultiQC: summarize analysis results for multiple tools and samples in a single report** *Bioinformatics* **32**:3047–3048

7. Pedregosa F., et al. (2011) **Scikit-learn: Machine Learning in Python** *J. Mach. Learn. Res* **12**:2825–2830

8. Koldovsky Z., Tichavsky P., Oja E (2006) **Efficient Variant of Algorithm FastICA for Independent Component Analysis Attaining the CramÉr-Rao Lower Bound** *IEEE Trans. Neural Netw* **17**:1265–1277

Editors

Reviewing Editor
Marisa Nicolás
Laboratório Nacional de Computação Científica, Rio de Janeiro, Brazil

Senior Editor
Aleksandra Walczak
École Normale Supérieure - PSL, Paris, France

Reviewer #1 (Public Review):

Summary:

This is large-scale genomics and transcriptomics study of the epidemic community-acquired methicillin-resistant S. aureus clone USA300, designed to identify core genome mutations that drove the emergence of the clone. It used publicly available datasets and a combination of genome-wide association studies (GWAS) and independent principal-component analysis (ICA) of RNA-seq profiles to compare USA300 versus non-USA300 within clonal complex 8. By overlapping the analyses the authors identified a 38bp deletion upstream of the iron-scavenging surface-protein gene isdH that was both significantly associated with the USA300 lineage and with a decreased transcription of the gene.

Strengths:

Several genomic studies have investigated genomic factors driving the emergence of successful S. aureus clones, in particular USA300. These studies have often focussed on acquisition of key accessory genes or have focussed on a small number of strains. This study makes a smart use of publicly available repositories to leverage the sample size of the analysis and identify new genomics markers of USA300 success.

The approach of combining large-scale genomics and transcriptomics analysis is powerful, as it allows to make some inferences on the impact of the mutations. This is particular important for mutations in intergenic regions, whose functional impact is often uncertain.

The statistical genomics approaches are elegant and state-of-the-art and can be easily applied to other contexts or pathogens.

Weaknesses:

The main weakness of this work is that these data don't allow a casual inference on the role of isdH in driving the emergence of USA300. It is of course impossible to prove which mutation or gene drove the success of the clone, however, experimental data would have strengthen the conclusions of the authors in my opinion.

Another limitation of this approach is that the approach taken here doesn't allow to make any conclusions on the adaptive role of the isdH mutation. In other words, it is still possible that the mutation is just a marker of USA300 success, due to other factors such as PVL, ACMI or the SCCmecIVa. This is because by its nature this analysis is heavy influenced by population structure. Usually, GWAS is applied to find genetic loci that are associated with a phenotype and are independent of the underlying population structure. Here, authors are using GWAS to find loci that are associated with a lineage. In other words, they are simply running a univariate analysis (likely a logistic regression) between genetic loci and the lineage without any correction for population structure, since population structure is the outcome. Therefore, this approach can't be applied to most phenotype-genotype studies where correction for population structure is critical.

Finally, the approach used is complex and not easily reproduced to another dataset. Although I like DBGWAS and find the network analysis elegant, I would be interested in seeing how a simpler GWAS tool like Pyseer would perform.

https://doi.org/10.7554/eLife.90668.2.sa1

Author response:

The following is the authors' response to the original reviews.

> ***Reviewer #1 (Recommendations For The Authors):***
>
> *(1) Line 56: replace "pyomastitis" with "pyogenic skin infections".*

Corrected.

> *(2) Line 58: replace "basal strains" with "ancestral strains".*

Corrected.

> *(3) Line 62: population structure impacts gene acquisition too, however, gene acquisitions can be easier to connect with a phenotype. For example, acquisition of mecA is thought to be adaptive rather than just linked to a successful lineage. This same reasoning applies to resistance-associated mutations such as gyrA mutations in ST22 emergence.*

We completely agree with the reviewer that population structure also impacts gene acquisition. We wanted to convey that connecting gain or loss of genes to a change in particular phenotype is much easier than doing the same for a mutation, specially in the presence of strong linkage, and therefore gene level analysis is the focus of many previous studies. We have rewritten the sentence to better convey this idea:

"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 as their effect on phenotype is easier to determine than that of point mutations."

> *(4) Line 112 this might be simply due to the smaller size of the intergenic regions chosen. I suggest to correct for the size of the genome segment considered.*

We thank the reviewer for pointing this out. The size of the intergenic was indeed the simple explanation for this observation. We have added the following sentence to the manuscript:

"This is reflective of the fact that most of *S. aureus* genome sequence comprises of ORFs e.g. ~84% of TCH1516 genome is part of an ORF."

> *(5) Line 189: please add p values to supp table 2.*

We have added the p and q values from DBGWAS into Supp table 2. It is under the 'DBGWAS Result' sheet.

> *(6) Line 227: high entropy indicates that this site is polymorph, not necessarily that there is selective pressure. In the extreme, this might actually point to a neutral position, since any amino-acid could be equally present (see for example https://www.nature.com/articles/s41467-022-31643-3#Sec10).*

We agree that high entropy by itself may point to a position with neutral selection leading to some false positives. However, we were focused on positions that were mostly biallelic in CC8, and with differential prevalence in USA300 vs non-USA300 (albeit in the presence of strong linkage disequilibrium) in addition to having high entropy in non-CC8 strains. This helps us filter some of the positions that were mostly monoallelic or with rare mutations

while preserving other sites of interest. The approach was able to find *cap5E* mutation which has been associated with disruption of capsule production.

> *(7) Line 271: show USA500 on the tree.*

Our current study is mostly focused on differences between USA300 and non-USA300 strains and we want to highlight those differences in the tree.

> *(8) Line 327: still not possible to infer causality.*

We have changed the language to remove mentions of causality and instead talk about the association of GWAS enriched genes with measured transcriptional changes. The revised sentence now reads:

"Here, we demonstrated how a model of transcriptional regulation with iModulons can be used to make a headway through the impasse created by the high linkage disequilibrium and identify GWAS-enriched mutations that are also associated with measurable phenotypic changes in the TRN."

> *(9) Line 324: subclades reference.*

We are unsure what this means.

> *(10) Line 366: the authors seem to have used a bespoke pan-genome analysis approach. Would they be able to validate it using established tools such as Roary, Pirate or Panaroo? Panaroo in particular appears to have superior accuracy thanks to its pan-genome graph approach (https://github.com/gtonkinhill/panaroo).*

We have added the results of Roary to our analysis (Figure S1b). The roary results largely agree with our biggest take away from pangenomics which is that our collection of genomes have a good coverage of the CC8 clade at the gene level.

> *(11) Line 397: what was the size of the core genome?*

There were 24881 core sites. We have added the number to the manuscript.

> *(12) Line 407: please add citation or website for SCCmecFinder.*

The citation of SCCmecFinder (45) is at the end of the sentence.

> *(13) Line 421: I was not able to find the code used for this analysis in the github repository provided.*

The code can be found in "notebook/02_Preprocess_DBGWAS.ipynb" within the repo.

> *(14) Line 427: this is a very complex analysis for a simple univariate comparison between USA300-vs-non USA300 strains with no correction for population structure. The authors should compare their results with a more established pipeline like Pyseer or Gemma that can handle kmers and show the added value of their approach.*

We wanted to take advantage of DBGWAS's ability to collapse kmers into unitigs and further collapse significant unitigs within a genetic neighborhood into components. Unfortunately, we found that in many cases, it became difficult to determine the exact mutation that was being enriched e.g. (T234G) without doing lots of manual work. Our network analysis simply

parses the DBGWAS graph to automatically extract these mutations, making the results more interpretable. It does not do any additional hypothesis testing.

We also attempted to pass kmer data into GEMMA but without the compaction provided by DBGWAS the memory required (>168 GB) exceeded what we had available.

> *(15) DBGWAS: please indicate DBGWAS version and the options used for kmer size and number of neighbour nodes retained in the subgraph. Also, I assume that no correction for population structure was applied.*

We have added the version and parameters for DBGWAS. The method section now reads:

"DBGWAS (v0.5.4) was used to enrich mutations unique to USA300 strains using default kmer size of 31 (-k 31) and neighborhood size of 5 (-nh 5). Alleles with frequency less than 0.1 were filtered (-maf 0.1) and all components enriched with q-values less than 0.05 were documented (-SFF q0.05)."

> *(16) Could the authors provide the DBGWAS output for the most significant unitings in graph format? This would help readers understand the findings.*

The outputs are available in the github repo. The link to this specific data is (https://github.com/sapoudel/USA300GWASPUB/tree/master/data/dbgwas/dbgwas_output/visualisations)

The text format of the output is part of Supplementary Table 2 under "DBGWAS Result" sheet.

> *(17) Line 469: please provide more details on iModulons, it is not enough to simply reference the paper: specific QC criteria, mapping algorithm and parameters, ICA algorithm.*

We have now added a new Supplementary Note 2 section with more details about building iModulons.

> *(18) Line 474: what is log-TPM?*

Log-Transcripts per Million. We have added the description in the text.

> *(19) Line 479: not sure what "Chapter 3" refers to.*

Thank you for correcting the mistake. The reference has been corrected.

> ***Reviewer #2 (Recommendations For The Authors):***
>
> *Line 45. The introduction is not well-structured, and there is a lack of coherence among the topics pertinent to the research objective. I would recommend rewriting this section addressing the following topics: the challenge of distinguishing lineages within the CC8, especially the CA-MRSA USA300 strains; discussing the state-of-the-art GWAS methodologies, elucidating the main confounding factors in the application of GWAS to bacterial studies, and finally, exploring how current methods aim to address these concerns.*

We would like to thank the reviewer for the suggestions. The main innovation of the paper is using iModulons to find phenotype associated mutations from a set of linked mutations. The challenge of distinguishing CC8 subclades has been largely resolved thanks to efforts by Bowers et al. (PMID: 29720527). We have made some revisions to address the GWAS methodologies (bugwas and DBGWAS), the effect of linkage disequilibrium in interpreting the

output of these methods and how combining the results of these association tests with modeling of TRN with iModulons can lead to finding candidate mutations of interest that are linked to specific changes in gene regulation.

> *Line 56. Replace "pyomastitis" with "pyomyositis".*

Corrected to “pyogenic skin infections.”

> *Lines 71. What do the authors mean by "endemic USA300 strain"?*

We have removed references to endemic strains.

> *Line 106. Please verify the number of genomes used in the DBGWAS analysis. In the text, the authors mention that 2038 genomes were utilized. However, in Supplementary Table 1, only 2030 genomes are listed.*

Thank you for catching the discrepancy. We started the analysis with 2037 genomes, including four “spiked-in” reference genomes- USA100 D592 (CC5 strain used for rooting the CC8 tree), TCH1516 (same accession number as the one used for ICA), COL and Newman. Before further analysis, we removed 6 genomes for being smaller than 2.5 million base-pairs (see preprocessing.ipynb) and the USA100 D592 strain as it is not part of CC8. This resulted in 2030 genomes being used for DBGWAS. We kept the other 3 spiked CC8 genomes to help annotate the unitigs from DBGWAS. Lastly, we removed the other three CC8 clade spiked genomes for pangenomic analysis. To clarify this, we have made the following changes to the text:

(1) Changed line 106: We downloaded 2033 *S.* aureus genomes for analysis and excluded six of them with genome length of less than 2.5 million base pairs. The remaining 2027 *S. aureus* CC8 genomes formed a closed pangenome, suggesting that the sampled genomes mostly captured the gene level variations within the clonal complex (Figure 1a).

(2) DBGWAS section Line 177: We used 2030 genomes for this analysis; the 2027 genomes in pangenomics analysis above were “spiked” with three well known CC8 genomes- TCH1516, COL, and Newman- to help annotate the DBGWAS unitigs.

> *Line 108. Could the authors provide a table with the genes that constitute the core, accessory genome, and unique genes for each of the strains?*

The genes presence absence tables are very large files and therefore we have only added them to our github repo. The results can be found in following files:

Pangenomics: data/pangenome/Pangenomics/CC8_strain_by_gene.pickle.gz

> *Lines 112 and 315. On what basis did the authors decide on the size of the upstream regulatory region? In the search for mutations, they extracted segments of 300 base pairs, whereas, in the search for the Fur binding motif, only 100 base pairs were considered. The RegPrecise database contains regulons for Staphylococcus aureus N315 (https://regprecise.lbl.gov/genome.jsp?genome_id=26), including the Fur regulon with multiple Transcription Factor Binding Sites (TFBSs) that extend beyond the 100 base-pair sequence. I would recommend reconsidering the search within the standardized upstream region of -400 base pairs. In the case of the Fur binding motif search, it might be beneficial to include the TFBSs available in the RegPrecise database.*

For Fur motif search, we chose 100 base-pairs because the Fur motif in non-USA300 strains were within ~20 base-pairs of *isdH* translation start site (Figure 4C). In our search of Fur

motif in this analysis, we were not looking to see if any exists, we were simply looking to see if the one proximal to the translation start site exists as our DBGWAS analysis suggested that specific region was deleted in USA300 strains.

> *Line 175. This work aimed to identify potential mutations associated with the success of a specific lineage rather than a phenotype, where correction for population structure effects is necessary. Would the implementation of the bugwas method in DBGWAS for controlling bacterial population structure not potentially impact the results? How was this issue addressed in your analysis? Would it not be pertinent to run a program without population structure correction to enable a comparison of results?*

We initially tried to use Linear Mixed Models to find kmers that were only enriched in USA300 strains. These efforts were hampered by extreme linkage disequilibrium which led to high collinearity between kmer abundance making it extremely difficult to get a good estimate of the coefficients. We also tried to run chi-squared tests individually on each kmer which led to unmanageable number (>100k) kmers that were significantly different. DBGWAS on the other hand was able to compress unbranched kmers in the De Bruijn into unitigs and further reduce the number of tests by testing at pattern level instead of unitig level. We found no straight forward way to run DBGWAS (or GEMMA) without population structure correction. Therefore, it is likely we may be underestimating the number of significant unitigs with this approach.

> *Line 189. Please italicize the gene name cap5E.*

Corrected.

> *Line 277. Please clarify the QC/QA criteria and curation process employed for the selection of RNA-seq experiments, as this constitutes a crucial step in the reconstruction of the network.*

We have now added a new supplementary material section, Supplementary Note 2 titled “Creating iModulons for CC8 Clade Staphylococcus aureus” with details of QC/QA.

> *Line 279. In Supplementary Table 3, please label the first column and standardize the use of either the experiment ID or the run ID. Furthermore, verify the experiment identifiers from rows 19 to 26, as I could not locate them in the SRA database.*

We have changed all accession to experiment ID including rows 19 to 26.

> *Lines 290, 330, 424, and 437. Please correct "SCCMec" to "SCCmec IVa" (italicize "mec").*

Corrected.

> *Line 298. What is the size of the upstream regulatory region considered for this analysis? It is important to standardize this value for all analyses involving the upstream regulatory region. In this regard, I recommend maintaining a consistent size of -400 base pairs.*

For Fur motif search we chose 100 base-pairs because the Fur motif in non-USA300 strains were within ~20 base-pairs of *isdH* translation start site (Figure 4C). In our search of Fur motif in this analysis, we were not looking to see if any exists, we were simply looking to see if the one proximal to the translation start site exists as our DBGWAS analysis suggested that specific region was deleted in USA300 strains. In our usual analysis, we use -300 base pairs.

> *Line 321. The discussion is rather concise and lacks an in-depth comparative perspective with relevant literature on any of the obtained results, whether concerning the proposed methodology or the potential new markers associated with the success of the USA300 lineage. The authors must underscore the method is not applicable to all GWAS analyses, due to the issue of correction for population structure.*

We have now added sections talking about the importance of *isdH* in *S. aureus* infection and a section addressing the limitation of the current approach when applied to other GWAS type study.

> *Line 366. The authors employed the methodology described in the article by Hyun et al. 2022 (https://doi.org/10.1186/s12864-021-08223-8) to construct the pangenome. However, this methodology was designed for comparative analysis of pangenomes across various species, which does not align with the objective of this study, focusing solely on S. aureus genomes. Consequently, it remains unclear to me why the authors made this particular choice and, more importantly, what advantages it offers over well-established tools for individual pangenomes, such as Roary. I would strongly recommend validating the results using at least one established tool.*

With our analysis, we can determine proper thresholds for core/accessory/unique genes based on the observed data (Supplementary Figure 1a). However, we agree that it would be proper to include a more established pangenome package. We have added the results of Roary to our analysis. The Roary results largely agree with our biggest take away from pangenomics which is that our collection of genomes have a good coverage of the CC8 clade at the gene level.

> *Line 370. Please include the version of CD-HIT that was utilized.*

Added. CD-HIT version 4.6 was used for the analysis.

> *Line 372. What tool did the authors use to extract these regions?*

The list of CDS, 5' and 3' sequences can be extracted easily with a combination of fasta file and gff file. The gff file was used to find the position of each of these sequences and the sequences were extracted from the fasta file with python scripts.

> *Line 395. What were the QC/QA criteria used to select the sequences?*

The QC/QA criteria for the sequences are mentioned in the beginning of the Pangnomic analysis subsection and is as follows:

"Briefly, "complete" or "WGS" samples from CC8/ST8 were downloaded from the PATRIC database. Sequences with lengths that were not within 3 standard deviations of the mean length or those with more than 100 contigs were filtered out."

> *Line 407. Please correct the tool name to "SCCmecFinder" (italicize "mec").*

The name has been corrected.

> *Line 409. I believe BLASTp was run locally, so please specify the version used and the search parameters.*

As corrected further down, we used BLASTn not BLASTp. The version v2.2.31 has been added to the methods section.

> *Line 416. There is conflicting information with line 409, which mentions that PVL was identified through a protein BLAST, but right below, it states it was a BLASTn. Please verify which information is correct and consider the previous comment to specify the version and parameters.*

Thank you catching the discrepancy. We have corrected the text:

"PVL was detected using nucleotide BLAST."

> *Line 418. Please provide the column identifiers for the Supplementary Table 5 (PVL worksheet).*

Column names are added.

> *Line 418. Please remove the repeated word "and" in Supplementary Table 5 (mecA worksheet) and italicize the gene names in this table.*

Corrected

> *Line 419. You can use the abbreviation "SNPs" since it was introduced in line 65.*

Corrected.

> *Line 420. In my view, this analysis could benefit from a more detailed and clearer explanation.*

We have added to the explanation. The section now reads:

"To find the root of the USA300 strains in the phylogenetic tree, the genomes in the tree were first annotated by their PVL and SCC_mec_ status. Then the tree traversed from leaf to root starting from known USA300 strains – TCH1516 and FPR3757- while keeping track of the number of descendant genomes from the current root that contained known markers SCC_mec_ IVa and PVL. The node where the number of genomes with the markers started flatlining was marked as the root of USA300."

> *Line 428. Specify the version and parameters used in the analysis with DBGWAS.*

Added. The text now reads:

"DBGWAS (v0.5.4) was used to enrich mutations unique to USA300 strains using default kmer size of 31 (-k 31) and neighborhood size of 5 (-nh 5). Alleles with frequency less than 0.1 were filtered (-maf 0.1) and all components enriched with q-values less than 0.05 were documented (-SFF q0.05)."

> *Line 431. What tools were employed to calculate Pearson correlation and distances relative to the reference genome?*

Added. The text now reads:

"Genome-wide linkage was estimated by Pearson correlation (calculated with built-in Pandas function) of the presence/ absence of enriched kmers and distance was measured based on the kmer alignment to the reference TCH1516 genome as determined by BLASTn."

> *Line 450. What type of BLAST was used?*

Added. Nucleotide blast was used for all kmer analysis.

> *Line 452. I didn't quite understand the reason for making this analysis available in a separate repository. It would be easier for readers looking to reproduce the work if all the codes were in a single repository.*

We kept the repository separate in case we wanted to further develop the network analysis code in the future. We have added the link to the network analysis repository in the README of the publication repo.

> *Line 460. Please specify the version and parameters, if run locally, or indicate if a web page was used.*

Corrected to indicate that we used the PATRIC website for this

> *Line 470. Specify the version and provide a detailed account of all parameters used, along with the QC/QA criteria and curation methods applied.*

We have added Supplementary Note 2 with all the details about packages and parameters used to calculate the iModulons.

> *Line 479. The phrase "ICA was then run as previously described in chapter 3" does not make sense. Please clarify.*

We have corrected the mistake and added a new supplementary note with details about our ICA run. The line now reads:

“A detailed version of the methods for RNA-sequencing and ICA analysis is available as Supplementary Note 2. ICA of RNA sequencing data was performed using the pymodulon package.”

> *Line 484. Specify the version of CD-HIT.*

Added. The version used was v4.6.

> *Line 494. To enable reproducibility, the repository should be better organized, especially the directory containing the code. Numbering each script in the order it was run would assist the reader in comprehending the overall analysis flow and adapting it to their needs. If creating a manual for method usage is not feasible, the code could be more extensively commented on to explain the parameters, choices made, and how these could be modified. The "Data" folder seems to contain some test files, such as those in the "isdh_fimo" folder, so removing test files would aid the understanding of the reader.*

Thank you for the suggestions. We have now numbered the notebooks that generate the figures, we have added more comments to the code, removed testing code and test datasets.

> *Throughout the article, please correct "SCCMec" to "SCCmec" (italicize "mec").*

Corrected.

https://doi.org/10.7554/eLife.90668.2.sa0