Comparative genome-scale modelling of *Staphylococcus aureus* strains identifies strain-specific metabolic capabilities linked to pathogenicity

Emanuele Bosi[a,1], Jonathan M. Monk[b,1], Ramy K. Aziz[c], Marco Fondi[a], Victor Nizet[d,e], and Bernhard Ø. Palsson[b,d,2]

[a]Department of Biology, University of Florence, I-50019 Sesto Fiorentino, Italy; [b]Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093-0412; [c]Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University, 11562 Cairo, Egypt; [d]Department of Pediatrics, University of California, San Diego, La Jolla, CA 92093-0760; and [e]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA 92093-0760

Edited by Sang Yup Lee, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea, and accepted by Editorial Board Member James J. Collins April 29, 2016 (received for review December 2, 2015)

Staphylococcus aureus **is a preeminent bacterial pathogen capable of colonizing diverse ecological niches within its human host. We describe here the pangenome of *S. aureus* based on analysis of genome sequences from 64 strains of *S. aureus* spanning a range of ecological niches, host types, and antibiotic resistance profiles. Based on this set, *S. aureus* is expected to have an open pangenome composed of 7,411 genes and a core genome composed of 1,441 genes. Metabolism was highly conserved in this core genome; however, differences were identified in amino acid and nucleotide biosynthesis pathways between the strains. Genome-scale models (GEMs) of metabolism were constructed for the 64 strains of *S. aureus*. These GEMs enabled a systems approach to characterizing the core metabolic and panmetabolic capabilities of the *S. aureus* species. All models were predicted to be auxotrophic for the vitamins niacin (vitamin B3) and thiamin (vitamin B1), whereas strain-specific auxotrophies were predicted for riboflavin (vitamin B2), guanosine, leucine, methionine, and cysteine, among others. GEMs were used to systematically analyze growth capabilities in more than 300 different growth-supporting environments. The results identified metabolic capabilities linked to pathogenic traits and virulence acquisitions. Such traits can be used to differentiate strains responsible for mild vs. severe infections and preference for hosts (e.g., animals vs. humans). Genome-scale analysis of multiple strains of a species can thus be used to identify metabolic determinants of virulence and increase our understanding of why certain strains of this deadly pathogen have spread rapidly throughout the world.**

systems biology | mathematical modeling | pathogenicity | core genome | pangenome

The preeminent gram-positive bacterial pathogen, *Staphylococcus aureus*, is capable of colonizing diverse ecological niches within its human host, including the respiratory tract, skin, and nasal passages. As a species it possesses several immune resistance and evasion factors, toxins, and invasiveness mechanisms. As a result, *S. aureus* is a leading cause of skin and soft tissue infections, pneumonia, sepsis, and endocarditis. In recent years, clinical management of this leading pathogen has been complicated by its continuous acquisition of resistance to front-line antibiotics (1). Despite this deadly status, broad diversity exists among strains within this species. Some strains are present as asymptomatic colonizers of the human nose (2), others cause skin and soft tissue infections, and some can cause life-threatening and severe disease (3). One strain, known as USA300, has replaced other strains of *S. aureus* to become the predominant cause of methicillin-resistant *S. aureus* (MRSA) infections in the United States (4), and its prevalence is increasing rapidly worldwide (5). Thus, it is important to understand the genetic factors that allow some strains to spread aggressively, whereas others exist asymptomatically.

The epidemiology of *S. aureus* and phenotypic diversity present in different strains is reflected in their genotypes. Therefore, it may be possible to analyze multiple *S. aureus* genomes to discover factors that could be predictors of disease phenotypes and virulence capabilities. Unfortunately, classical genotyping methodologies used today in the clinic such as pulsed-field gel electrophoresis (PFGE) and multilocus sequence tags (MLST) rely on the evaluation of highly conserved housekeeping genes representative of the vertical gene pool that do not provide sufficient resolution for prediction of disease phenotypes (6). Due to the revolution in DNA sequencing technologies, hundreds of full *S. aureus* genome sequences are now available and can be analyzed using new methods. One such method is the analysis of shared and unique genes, within a species, termed the "pangenome." A bacterial species can be effectively described by its pangenome (7), which can be divided into the core genome (genes shared by genomes of all strains in the species and that are likely to encode functions related to basic cellular biology) and the dispensable genome [genes present in some, but not all, of the representatives of a species (8)]. The dispensable genome includes functions that confer specific advantages under particular environmental conditions, such as adaptation to distinct niches, antibiotic resistance, and the ability to colonize new hosts.

Despite many studies focusing on *S. aureus* genomics, molecular epidemiology, and mechanisms of drug resistance and cytotoxicity, there are relatively few studies that examine *S. aureus* basic biochemistry and metabolic function on a genome scale. Metabolic

Significance

Comparative analysis of multiple strains within a species is a powerful way to uncover pathoadaptive genetic acquisitions. Hundreds of genome sequences are now available for the human pathogen *Staphylococcus aureus,* mostly known for its antibiotic-resistant variants that threaten the emergence of panresistant superbugs. In this study, genome-scale models of metabolism are used to analyze the shared and unique metabolic capabilities of this pathogen and its strain-specific variants. The models are used to distinguish *S. aureus* strains responsible for severe infections based solely on growth capabilities and presence of different virulence factors. The results identify metabolic similarities and differences between *S. aureus* strains that provide insights into the epidemiology of *S. aureus* and may help to combat its spread.

Author contributions: E.B., J.M.M., and B.Ø.P. designed research; E.B., J.M.M., and R.K.A. performed research; E.B., J.M.M., R.K.A., M.F., V.N., and B.Ø.P. analyzed data; and E.B., J.M.M., V.N., and B.Ø.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. S.Y.L. is a guest editor invited by the Editorial Board.

Freely available online through the PNAS open access option.

[1]E.B. and J.M.M. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: palsson@ucsd.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1523199113/-/DCSupplemental.

network reconstructions have proven to be powerful tools to probe the genetic diversity of metabolism between organisms (9) and among strains within a species (10, 11). As useful as genome annotation is, it does not provide an understanding of the integrated function of gene products to produce phenotypic states. Becker et al. provided a well-curated genome-scale model of *S. aureus* N315 (named iSB619) that represented the first biochemically, genomically, and genetically structured knowledge base for *S. aureus* metabolism, and several updates of the model for this strain have followed (12–14). However, knowledge from one strain is never sufficient to represent an entire species. Currently, there are 48 complete genome sequences available for *S. aureus* strains in the National Center for Biotechnology Information (NCBI) Genome database (15) with an additional 450 draft sequences for comparison, which provide comprehensive insight into the pangenome of the *S. aureus* species. In this study we set out to construct the pangenome of the *S. aureus* species and to build GEMs of strains that represent the breadth of genetic, phenotypic, and pathogenic characteristics within this species. Together, the genomic sequences and the GEMs provide two different and complementary ways to explore diversity within the *S. aureus* species and to compare core genomic vs. pangenomic functionality and metabolic capabilities. Our integration of genomic and biochemical data illustrates the dynamic evolution of *S. aureus* strains, and the results provide insights into the metabolic determinants of pathogenicity.

Results

Characteristics of the *S. aureus* Core Genome and Pangenome. A set of publicly available *S. aureus* genome sequences was downloaded from the NCBI including 48 completely assembled genomes (16). From this set, a phylogenetic tree was constructed using the concatenated sequence of seven conserved housekeeping genes (*arcC*, *aroE*, *glpF*, *gmk*, *pta*, *tpi*, and *yqiL*) commonly used to define clonal complexes in clinical studies of *S. aureus* using the MLST approach (Fig. 1*A*). The most distantly related strain was the Australian *S. aureus* isolate MSHR1132 belonging to the clonal complex 75 lineage (17). Next, a representative set of 64 *S. aureus*

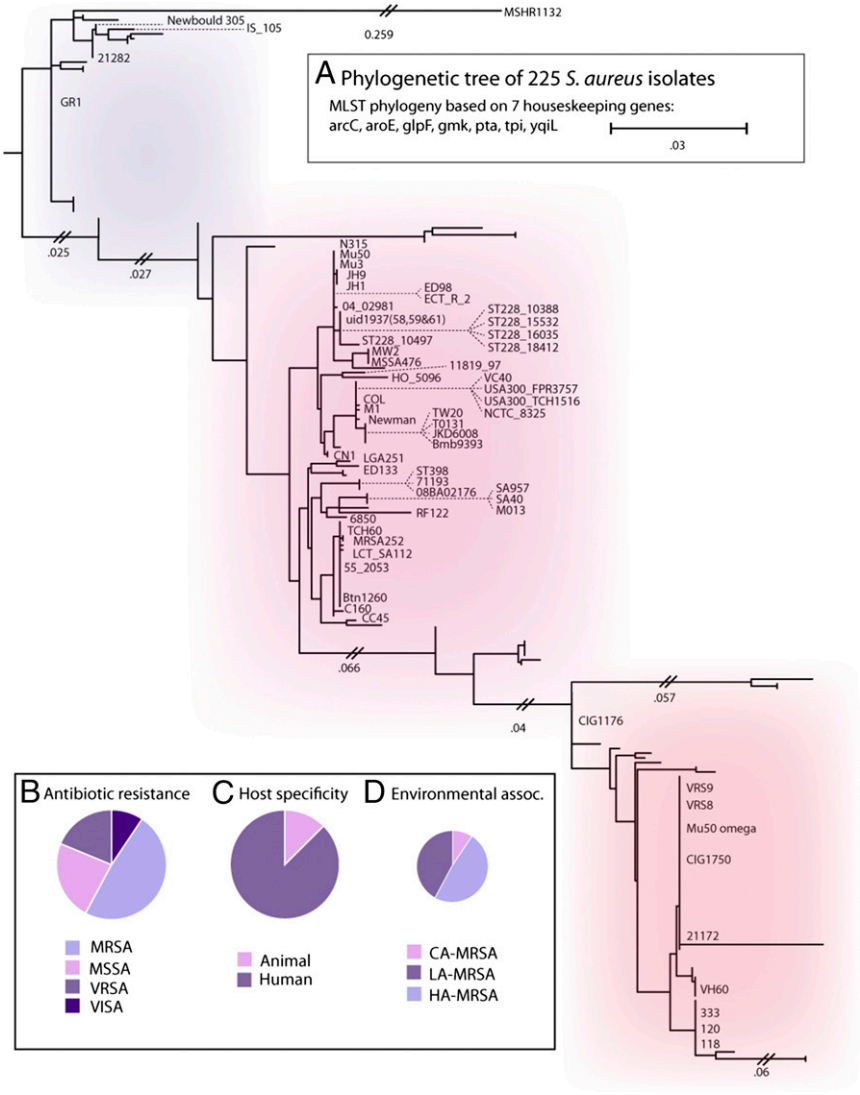


Fig. 1. *S. aureus* dataset construction. (*A*) Phylogenic tree of 225 *S. aureus* genomes based on seven housekeeping genes (*arcC*, *aroE*, *glpF*, *gmk*, *pta*, *tpi*, and *yqiL*). A set of 64 strains (labeled) were selected from this set to create a heterogeneous dataset of *S. aureus* strains based on the evolutionary distance (tree topology), as well as (*B*) drug resistance (MRSA, MSSA, VRSA, and VISA), (*C*) host specificity (human vs. animal), and (*D*) virulence/environmental association (CA-MRSA, community-associated MRSA; HA-MRSA, healthcare-acquired MRSA; LA-MRSA, livestock-associated MRSA). Evolutionary distance is based on tree topology.

strains (Dataset S1) were filtered from this larger dataset based on four criteria: (*i*) drug resistance [MRSA, methicillin-sensitive *S. aureus* (MSSA), vancomycin-resistant *S. aureus* (VRSA), and vancomycin-intermediate *S. aureus* (VISA)] (Fig. 1*B*), (*ii*) host specificity (human vs. animal) (Fig. 1*C*), and (*iii*) epidemiological source [community-associated MRSA (CA-MRSA), healthcare-acquired MRSA (HA-MRSA), and livestock-associated MRSA (LA-MRSA)] (Fig. 1*D*). Thus, the topology of the tree coupled with clinical and epidemiological phenotypic information allowed for development of a heterogeneous dataset spanning representative strains of the *S. aureus* species.

We used these 64 strains to construct the *S. aureus* pangenome that can be used to describe this bacterial species. The pangenome is divided into three sections: (*i*) the core genome (the set of genes shared by virtually all strains in a species), (*ii*) the accessory genome (the set of genes present in some, but not all, representatives of a species), and (*iii*) the unique genome (genes unique to individual members of the species). Because the pangenome is made up of thousands of genes, it offers a much higher resolution for strain typing compared with genotyping techniques such as MLST (18).

The 64 *S. aureus* strains examined yielded a pangenome size of 7,457 genes. Based on this dataset, the core genome is composed of 1,441 genes, the accessory genome is composed of 2,871 genes, and the unique genome is composed of 3,145 genes (Fig. 2*A*). Functional annotation of genes in the pangenome performed using the Clusters of Orthologous Groups (COG) database (19) revealed a unique distribution of functional categories among the three different pangenome sets. When genes were classified into metabolic and nonmetabolic, the largest fraction of the core genome consisted of genes with metabolic functions (58%). These included those involved in central metabolism including glycolysis/gluconeogenesis, the pentose phosphate pathway, and TCA cycle. A much lower fraction of metabolic gene content is observed in the accessory and unique genome (33% and 18%, respectively).

The size of the pangenome and its increase in size upon addition of new strains can be used to predict the future rate of discovery of novel genes in a species. We used a sampling approach to obtain a set of randomly permuted *S. aureus* pangenomes (with specific breakdowns of total genes, shared genes, and new genes). By fitting a double exponential decay function to

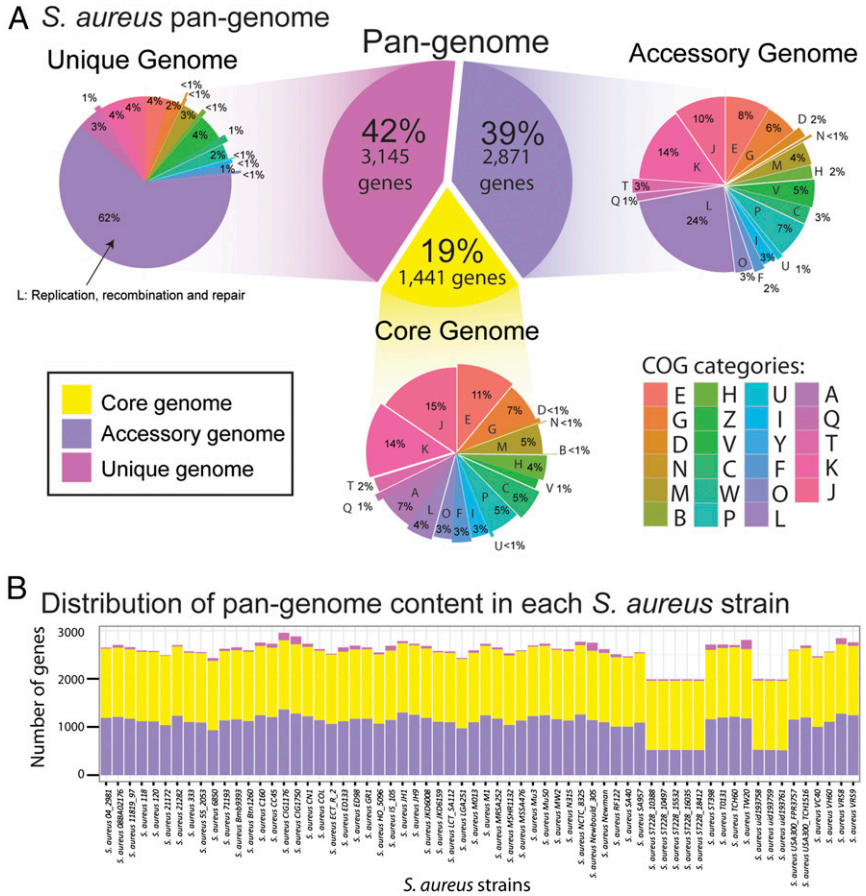


Fig. 2. *S. aureus* pangenome statistics. (*A*) The *S. aureus* pangenome can be subdivided into three categories: (*i*) the core genome (the set of genes shared by all genomes), (*ii*) the accessory genome (the set of genes present in some but not all genomes), and (*iii*) the unique genome (genes that are unique to a single genome). The function of each gene in a group is classified using COGs. COG categories are as follows: For cellular processes and signaling, D is cell cycle control, cell division, and chromosome partitioning; M is cell wall/membrane/envelope biogenesis; N is cell motility; O is posttranslational modification, protein turnover, and chaperones; T is signal transduction mechanisms; U is intracellular trafficking, secretion, and vesicular transport; V is defense mechanisms; W is extracellular structures; Y is nuclear structure; and Z is cytoskeleton. For information storage and processing, A is RNA processing and modification; B is chromatin structure and dynamics; J is translation, ribosomal structure, and biogenesis; K is transcription; and L is replication, recombination, and repair. For metabolism, C is energy production and conversion; E is amino acid transport and metabolism; F is nucleotide transport and metabolism; G is carbohydrate transport and metabolism; H is coenzyme transport and metabolism; I is lipid transport and metabolism; P is inorganic ion transport and metabolism; and Q is secondary metabolites biosynthesis, transport, and catabolism. (*B*) Distribution of the genes in the pangenome for each examined *S. aureus* strain.

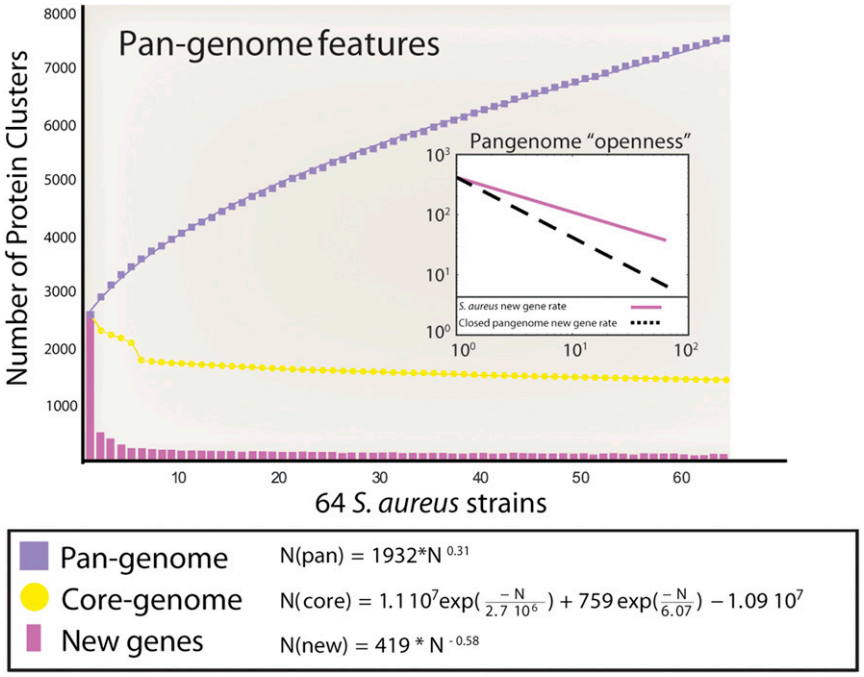


Fig. 3. Pangenome, core, and novel genes of the 64 analyzed *S. aureus* strains. Pangenome features are as follows: The purple squares denote the number of novel genes discovered with the sequential addition of new genomes. The yellow dots denote the values of the core genes as genomes are added to the pangenome. The purple bars indicate the number of new genes added to the total pangenome size as new genomes are added. Each of the values represents the median from a distribution of randomly selected genomes at each genome addition. The purple line represents the number of new genes found for each genome addition. For comparison, the same trend for a closed genome is reported as a dashed line. The equations below the graph show parameters for fits to Heap's law. Positive exponents indicate an open state and that the category is boundless so new genes are likely to be discovered continually as new genomes are sequenced.

the number of shared genes, we estimated the *S. aureus* core genome to have 1,425 genes (Fig. 3). Discovery of new genes and count of total genes were used to obtain fits to the Heap's law function, resulting in γ values of −0.58 and 0.31, respectively. The γ parameter determines the behavior of the curve. For γ values >0, the function has no asymptote, indicating that the *S. aureus* pangenome repertoire is likely to grow indefinitely as more strains are sequenced.

Size and Content of the Core Genome Provide Insight into Essential *S. aureus* Capabilities. The average *S. aureus* genome encodes 2,800 genes; therefore, the size of the core genome represents a high portion (56%, on average) of each *S. aureus* genome (Fig. 2*B*). This portion is particularly high compared with other organisms including *Clostridium difficile* (core represents 25% of the average genome) (20) and *Escherichia coli* (core represents 40% of the average genome) (21). The fact that each *S. aureus* strain has such a high portion of shared genes can be interpreted as a direct consequence of the proposed clonal structure for the species (22). The majority of the 1,441 core genes (shared by all 64 *S. aureus* strains examined) are involved in housekeeping processes. These include nonmetabolic functions such as transcription (15%), translation, ribosomal structure and biogenesis (14%), and RNA processing and modification (7%). A core set of 239 genes functionally annotated to be involved in transcription and translation was found in all *S. aureus* strains examined (Fig. S1). A detailed discussion of these genes is provided in *SI Text*. Of these 239 conserved genes, 94 and 87 are known to be experimentally essential on LB media in *E. coli* (23) and *Bacillus subtilis* (24), respectively. Therefore, inhibitors of these proteins may be good targets for antibiotic therapies against *S. aureus*. Genes involved in metabolic functions were also highly conserved. Such functions included amino acid transport and metabolism (11%), carbohydrate transport and metabolism (7%), coenzyme transport and metabolism (4%), cell wall and membrane biosynthesis (5%), and energy production and conversion (12%). Genome-scale network reconstructions for each of the 64 strains were built to perform an integrated analysis of the function of these metabolic genes (see below). Together, genes involved in metabolic, transcription, and translation processes make up 75% of the *S. aureus* core genome.

Relative to the core genome, widely different functional assignments are present for those genes in the accessory and unique genomes. The accessory genome has a large portion of genes associated with mobile genetic elements such as transposons and bacteriophages [replication, recombination, and repair (24%)] or those with nonmetabolic functions including defense mechanisms (5%). Metabolic functions were also included in the accessory genome, including those related to amino acid metabolism (8%), inorganic ion transport (7%), and carbohydrate metabolism (6%). The unique genome is heavily enriched in genes related to mobile elements (62%), with the other categories being poorly represented. This high proportion of mobile elements in the unique genome is similar to other organisms, including *E. coli*, and indicates that horizontal gene transfer (HGT) has had a large effect on *S. aureus* evolution (25–27). Indeed, the archaic MRSA clone, thought to be the ancestor strain of MRSA in Europe, obtained its methicillin resistance because of the horizontal transfer of the *mecA* gene from an unknown source (1). Even strains in the same household have been shown to engage in HGT with other *Staphylococcus* species (28). Thus, analysis of all HGT events in different strains of *S. aureus* and their putative source can shed light on the evolution of this species (see *SI Text* for analysis of atypical genes and their putative transfer source).

Ancestral Events of Gene Loss and Acquisition Affect the Pathogenicity of *S. aureus* Strains. We used the identified 1,441 core genes to produce an *S. aureus* phylogeny based on a concatenated sequence of the aligned genes. The patterns of gene presence were used with this phylogeny based on the core genome to infer events of ancestral gene gains and losses using a parsimony approach (29). This analysis allowed, for each internal node of the phylogenetic tree, us to compute (*i*) the number of gene gains and losses and (*ii*) the corresponding genetic repertoire. In particular, we focused on the evolutionary events related to the virulence factors in each strain because they have profound implications on pathogenesis of different *S. aureus* clones. The mapping of the events of gene gains and losses revealed that *S. aureus* strains have undergone extensive rearrangement of their genetic repertoire. A large number of evolutionary events occurred both ancestrally and recently (Fig. S2). For example, the ancestor of the ST228 lineage is characterized by a massive genome reduction (522 genes lost), whereas all of the LA-MRSA strains have recently acquired a large number of genes. Considering that the *S. aureus* last common ancestor (Sa-LCA) comprised 2,673 ortholog groups, the species has acquired a large number of new genes (4,784) during its evolutionary history.

Known *S. aureus* Virulence Factors Are Unequally Distributed Between Core and Accessory Genomes. To gain insight into the conservation of virulence factors across the *S. aureus* species we curated a set of known virulence factors (VFs) present in different strains based on literature and database searches (30). We identified a total of 90 different VFs that were present in at least one of the 64 different *S. aureus* strains. Of the 90 VFs, 35 were shared by all of the strains, forming a core set of VFs. Nine of the conserved VFs are cap8 genes (B, C, E, F, L, M, N, O, and P), which are involved in the synthesis of the polysaccharide capsule (PC). Other conserved VFs included five involved in the production of two different cytotoxins, namely, the Panton–Valentine leukocidin (PVL), encoded by the genes *lukS* and *lukF*, and the gamma–hemolysin, encoded by *hlgA*, *hlgB*, and *hlgC*. Four other conserved VFs encode iron-regulated proteins (*isdA*, *isdC*, *isdE*, and *isdF*) that bind to extracellular matrix components such as fibrinogen and fibronectin to promote cell adherence. Among these, the *isdA* gene plays a role in the *S. aureus* iron acquisition system, important for *S. aureus* in vivo replication and disease pathogenesis (31). Other canonical *S. aureus* virulence factors were highly conserved but were not found across all 64 strains. Protein A (binds immunoglobin G to disrupt phagocytosis, encoded for by *spa*) was found in 90% of strains. Alpha toxin (disrupts the membrane and enhances invasiveness, encoded for by *hla*) was found in 96% of strains. Biosynthesis genes for the staphyloxanthin pigment (encoded for by *crtMNPQ*) were found in all but one of the strains. This strain (MSHR1132) was specifically sequenced because of its lack of the yellow pigment (17).

Several other VFs were strain-specific. For example, the Staphylococcal complement inhibitor (SCIN) protein (encoded by *scn*) was present in 48 of the 64 strains, and Chemotaxis inhibitory protein of staphylococci (CHIPS, encoded by *chp*) was present in 27 of the 64 strains. The AGR quorum sensing system (regulates biofilm development, encoded for by *agrABCD*) was found in 25% of strains. Meanwhile, the exfoliative toxin B was found only in a single strain (*S. aureus* 11819-97). *S. aureus* MW2 was found to have the highest count of established virulence factors (79 VFs) followed by MSSA476 (74 VFs), Mu3, and Mu50 (70 VFs). In contrast, the ST288 strains had the fewest number of VFs (53 total).

Metabolic Models Facilitate Investigation into Genetic Basis of Strain-Specific Growth Capabilities. Because so much of the *S. aureus* core genome is dedicated to metabolic functions, the 64 *S. aureus* genome sequences were used to construct strain-specific genome-scale metabolic reconstructions (Dataset S2) that were used to compare gene, reaction, and metabolite content between the strains. Each reconstruction serves as a comprehensive representation of the metabolic capabilities of an *S. aureus* strain. Content shared among all reconstructions defines the core metabolic capabilities of the *S. aureus* species. Similarly, the metabolic capabilities of all strains were combined to define the full potential of metabolic capabilities for the *S. aureus* species, or its panmetabolic network (Fig. 4*A*) (10).

The core metabolic network contained 700 metabolic genes that catalyze 1,222 reactions involving 1,257 metabolites. Highly conserved metabolic subsystems across all *S. aureus* models include lipid metabolism, energy metabolism, glycan biosynthesis, and metabolism of polyketides and terpenoids. Reactions involved in these metabolic subsystems were highly represented (>95% conserved) in the core metabolic network. By contrast, only 80% of amino acid biosynthesis reactions were part of the core metabolic network. Conserved amino acid biosynthesis pathways included those for valine, alanine, and serine biosynthesis. The core metabolic network also contained known metabolic virulence determinants including catalase, an enzyme that hydrolyzes hydrogen peroxide into water and oxygen and that is used in the clinical laboratory to distinguish staphylococci from enterococci and streptococci. Catalase production and oxidant resistance have been shown to be predisposing factors for nasal colonization and subsequent infection (32).

The panmetabolic reactions are composed of the union of different metabolic reactions found in all strains and thus indicate the pool of metabolic capabilities within a species. The *S. aureus* panmetabolic network contains 1,090 metabolic genes, 1,592 reactions, and 1,519 metabolites. About 45% of reactions in carbohydrate metabolism were not present in the core metabolic network. Thus, these reactions were not shared by all strains of *S. aureus* examined. These formed the largest group of reactions present in the set difference between the core metabolic and panmetabolic networks (Fig. 4*B*). A majority of these reactions are involved in alternate carbon metabolism, including catabolic pathways for unique niche-specific nutrients. Amino acid metabolism was also disproportionately present in the panmetabolic network including biosynthesis pathways for leucine, arginine, and histidine metabolism, indicating that these capabilities may have been lost by several strains of *S. aureus*.

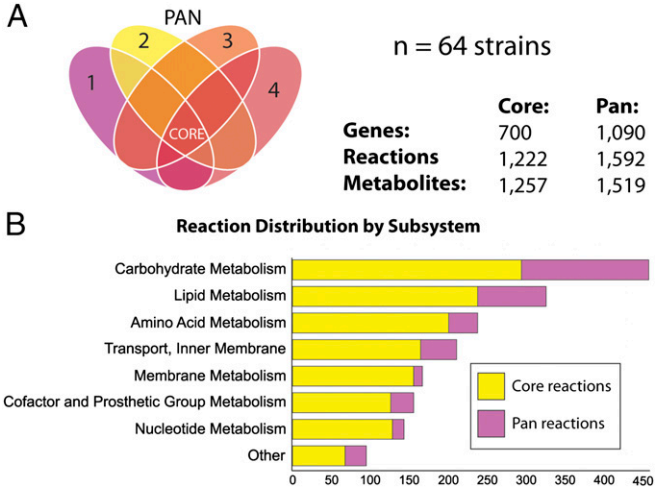


Fig. 4. Core metabolic and panmetabolic capabilities of the *S. aureus* species. The core metabolic and panmetabolic content was determined for genome-scale metabolic models (GEMs) of 64 unique *S. aureus* strains. (*A*) The core content, illustrated by the intersection of the Venn diagram, is shared with all strains. The pancontent consists of all content in any model and includes the core content. Note that the Venn diagram is not to scale and is simplified to only include the first 4 out of $n = 64$ strains. (*B*) Classification of reactions in the core reactomes and panreactomes by metabolic subsystem.

The conversion of metabolic network reconstructions into computable mathematical models enables computation of phenotypes in diverse nutrient environments based on the content of each reconstruction. Thus, the 64 *S. aureus* strain-specific reconstructed networks were converted into genome-scale metabolic models (GEMs) that were used to compute phenotypes in more than 300 unique, growth-supporting environmental conditions. A detailed biomass composition was defined (Table S1) and used to identify the distribution of metabolic fluxes leading to optimal growth. Reactions belonging to the amino acid metabolism subsystem made up the majority of reactions in set difference between the core reactomes and panreactomes (Fig. 4*B*). Thus, we hypothesized that functional differences in amino acid biosynthesis capabilities of different strains of *S. aureus* may allow different strains to adapt to different nutritional environments. To test this hypothesis, we simulated growth in silico for all 64 *S. aureus* GEMs on a variety of minimal media growth conditions, including a minimal growth media reported for *S. aureus* N315 (12). The in silico growth analysis revealed that all 64 models growing in glucose minimal media require at least vitamins B1 (thiamin) and B3 (niacin) to be added to the media. Some of the strains required more components to be added to the minimal media beyond just these (Table 1). The thiamin auxotrophy is due to a lack of the pathway that converts tyrosine to thiamin via tyrosine lyase and thiazole phosphate synthase. The niacin auxotrophy is due to a lack of nicotinate-nucleotide diphosphorylase (EC 2.4.2.19). These auxotrophies have been experimentally documented for *S. aureus* in the past (33).

There were also several additional strain-specific auxotrophies with 90% of the models (55/64) unable to grow in the in silico glucose minimal media containing thiamin and niacin (Table S2). These models lacked the ability to synthesize additional compounds including the nucleotide guanine and the vitamin riboflavin as well as the amino acids leucine, arginine, histidine, tryptophan, phenylalanine, methionine, proline, and tyrosine (Table 1). Previous work has demonstrated that some strains of *S. aureus* are able to grow in minimal media devoid of amino acids after a long period of training or weaning away from amino acid addition (33, 34). We experimentally confirmed that four strains (USA300, N315, 8325, and Mu50) could grow in minimal media supplemented with proline, serine leucine, threonine, thiamin, and niacin (Fig. S3). *S. aureus* is known to have a complicated gene regulatory structure that may inhibit its growth on minimal media, despite the presence of biosynthetic pathways for these metabolites (12, 14). Longer-term growth or training in the laboratory for these strains may allow for eventual growth in the minimal medium defined here (35).

Table 1. Percentage of the models with a predicted strain-specific auxotrophy

Compound	Percent auxotrophic
Thiamin	64/64 (100%)
Niacin	64/64 (100%)
L-leucine	38/64 (59%)
Guanine	31/64 (48%)
L-methionine	20/64 (31%)
L-cysteine	16/64 (25%)
Riboflavin	16/64 (25%)
L-proline	12/64 (19%)
L-asparagine	11/64 (17%)
L-histidine	3/64 (5%)
L-phenylalanine	3/64 (5%)
L-tyrosine	3/64 (5%)
L-arginine	2/64 (3%)
L-tryptophan	1/64 (2%)

***S. aureus* Strains Can Be Differentiated Using GEM-Predicted Metabolic Capabilities and Presence of Virulence Factors.** The 64 *S. aureus* GEMs were used to predict growth capabilities on alternative carbon, nitrogen, phosphorous, and sulfur sources by removing glucose, ammonia, sulfate, and phosphate from the in silico growth media and adding alternative sources one at a time. Over 300 alternative nutrient sources were tested in aerobic and anaerobic conditions using flux balance analysis (FBA) (36) to assess whether each S. *aureus* strain grew in silico (Fig. 5*A*). Overall, 238 nutrients were universally catabolized by all strains including glucose and glycerol as carbon sources and arginine as a nitrogen source. Other nutrients were strain-specific, including dulcose (42 strains) and inosine (13 strains) as carbon sources and uracil and thymidine as nitrogen sources (42 strains each). We specifically looked to see if the models were predicted to be able to use arginine for growth. The arginine mobile catabolic element (ACME) is present in MRSA USA300 strains and has been shown to be a key element for its successful colonization (37). We, however, found that the arginine catabolic capability was conserved across all *S. aureus* models due to the presence of a separate *arc* operon (encoded for by *arcA-arcD*) present in all strains. This finding is consistent with previous work (38). Despite this similarity, several other growth capabilities could be used to distinguish groups of strains (Fig. 5). For example, simulation results showed that the two *S. aureus* USA300 isolates were the only strains capable of using spermidine as a sole source of carbon and nitrogen due to the presence of spermidine acetyltransferase (encoded for by *speG*). Spermidine is produced at high levels in areas of keratinocyte proliferation, inflammation, and wound healing (39), conditions under which *S. aureus* invades and causes skin infection. We experimentally confirmed that only *S. aureus* USA300 could grow in TSB media supplemented with 6 mM of spermidine, whereas *S. aureus* strains Newman and NCTC8325 could not (Fig. S4).

Another example of using model predictions to distinguish strains was made possible by combining the simulated growth results with presence/absence of virulence factors in each strain. This combination allowed the formation of a classification schema that could be used to differentiate strains with a lifestyle that is livestock associated (LA) compared with those that were human associated (HA). The classifier was capable of distinguishing strains with these two host preferences based on evaluation of only three factors: presence of staphylokinase precursor (*sak*), the ability to catabolize maltotriose (mlttr), and presence of IgG binding protein A precursor (*sak*) (Fig. 5*B*). Interestingly, 100% of the *sak* positive strains were human-associated. We also observed that although many of the strains from the same clonal complexes clustered near each other, some important differences arose. For example, strains from the ST228 lineage appear to have lost several metabolic capabilities relative to others, causing two separate groups from this clonal complex to cluster apart from each other. Therefore, GEM-predicted metabolic capabilities and presence of virulence factors can be used to classify strains of *S. aureus* based on lifestyle and preferred niche. Thus, the tools presented here are capable of offering more precise classification schemes than those used routinely in the clinic to type *S. aureus* strains today.

Discussion

The pangenome of *S. aureus* was assembled and applied to highlight the genetic, metabolic, and pathogenic diversity of the species. A particular emphasis was placed on the analysis of strain-specific metabolic capabilities and the distribution of virulence factors. Based on the content of the pangenome, GEMs of metabolism were reconstructed and deployed for 64 *S. aureus* strains to determine their functional differences by computing minimal media compositions and growth capabilities across more than 300 nutrient sources both aerobically and anaerobically. All strains were predicted to be auxotrophic for niacin and thiamin, whereas strain-specific auxotrophies were predicted for riboflavin, guanine, and

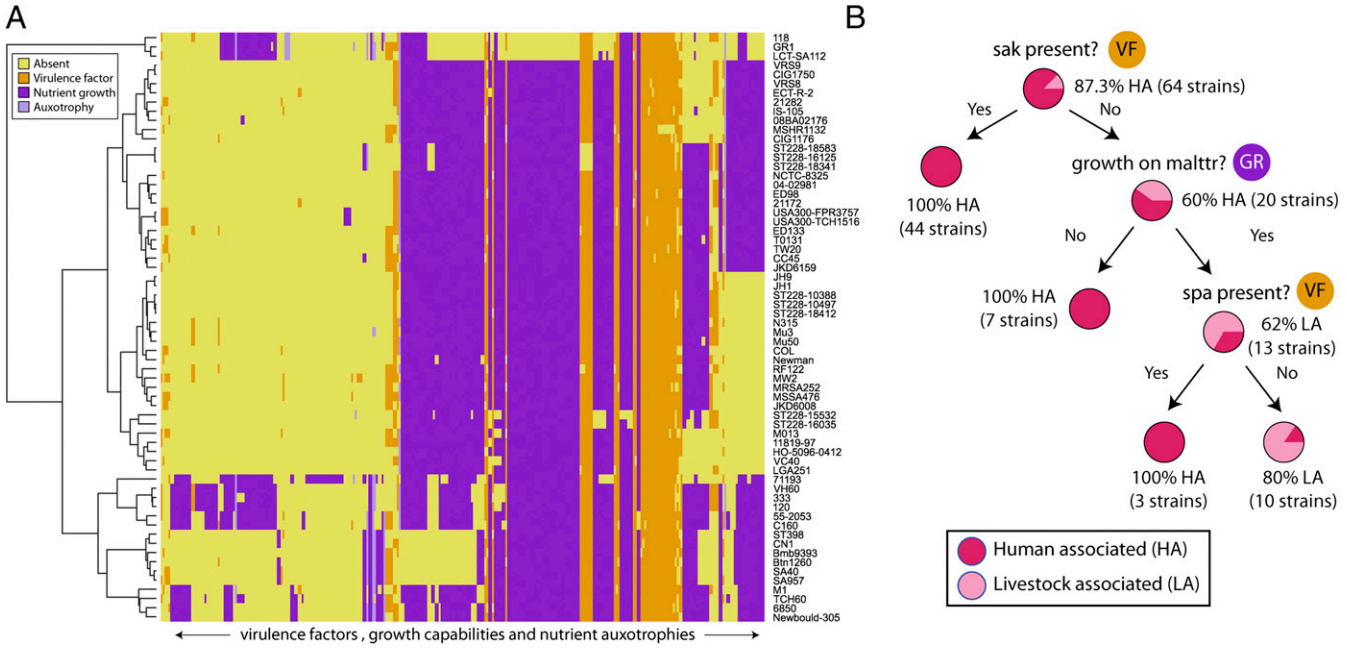


Fig. 5. *S. aureus* virulence factors and predicted metabolic capabilities. (*A*) Presence and absence of virulence factors and predicted metabolic capabilities across the 64 *S. aureus* strains examined in this study. Metabolic capabilities were predicted using the strain-specific metabolic models (dark purple, growth capability; light purple, nutrient required; yellow, no growth or nutrient is not required). The virulome consists of curated virulence factors known to be present in different strains of *S. aureus*. Orange indicates a factor is present, and yellow indicates a factor is absent. Full matrix with strains, predicted growth capabilities, and virulence factor is available in Dataset S1. Virulence factor and growth profiles can be used to classify strains. For example in *B*, a classification is constructed that separates human-associated *S. aureus* strains from livestock-associated strains using only the presence of two virulence factors and the ability to catabolize maltotriose. Abbreviations are as follows: staphylokinase precursor (sak), maltotriose (malttr), and IgG binding protein A precursor (spa).

leucine among others (Table 1). Model-predicted growth capabilities and identification of virulence factors allowed for the creation of classification schemas that could distinguish groups of strains based on relatively few traits. The results presented here demonstrate that comparisons of GEM-predicted metabolic capabilities among different strains in a species can be used to identify strain-specific biomarkers and to uncover and elucidate metabolic determinants of virulence.

Identification and characterization of the makeup of a species' pangenome is a powerful tool to analyze genomic diversity within a taxon. The *S. aureus* pangenome described here is composed of 7,457 unique genes. Of these, 1,441 genes are shared among all strains in the species, forming a core genome. Functional assignment of the core genes revealed that they are mostly associated with housekeeping functions (i.e., control of gene expression machinery and basic biochemistry). On average, 56% of genes in the average *S. aureus* genome are part of the core genome. This confirms previous observations that *S. aureus* is a clonal species (22). Recent studies have shown that even mutations in the core genome of closely related *S. aureus* isolates can have significant effects on virulence, proliferation, and persistence of *S. aureus* strains (6). Therefore, a deeper analysis of strain-specific genetic variants in this set could offer further insights into *S. aureus* pathogenicity.

The 1,441 core genes were aligned to produce a reliable phylogeny compared with those derived from a limited number of genes (Fig. S2). In general, the phylogeny is consistent with information from literature including agreement between strains' ST and clade. However, some exceptions were observed: (*i*) the ST5 group is included within a paraphyletic clade corresponding to CC5 and (*ii*) *S. aureus* COL (ST250) is placed in the ST8 clade. Although these exceptions are consistent with the evolutionary history of *S. aureus* [i.e., the paraphyletic clade includes members of the same clonal complex and ST8 is the predicted ancestor of ST250 (40)], the differences between MLST genotyping and the presented phylogeny underline the limitations of molecular typing techniques to differentiate between closely related strains. Furthermore, in regards to host specificity, the distribution of the LA-MRSA strains along the tree indicates that they have originated through different independent events. In particular, we identified a monophyletic clade of ST398 strains, a clade of strains from different lineages (ST425, ST133, and ST151), and two strains included in different human-associated clades (corresponding to ST5 and ST1).

In contrast to the core genome, the size of the pangenome was used to predict, via extrapolation, the number of genome sequences required for bounding the gene repertoire of a clade. Our regression analysis shows that the *S. aureus* pangenome is open, indicating that the gene repertoire of this species is theoretically boundless. This result is in agreement with a previous DNA microarray experiment involving 36 *S. aureus* strains (41), in which extensive genetic variability was reported. A high portion of unique genes were found to be related to mobile genetic elements (i.e., transposon, phages, and plasmids), which may drive acquisition of novel functional modules via HGT, including drug resistance and virulence (42). Beyond these evolutionary insights, the pangenome has important practical implications. The presence/absence of genes from the dispensable genome (i.e., genes unique to some strains) represents a high-resolution alternative to MLST that we have used to diagnose distinct groups of pathogenic bacteria based on specific biomarkers.

We specifically examined the presence of virulence factors in the core genomes and pangenomes. Overall, most of the *S. aureus* virulome is conserved across the 64 strains examined. There is only one VF unique to a single strain (i.e., the *etb* gene, encoding the exfoliative toxin B), whereas 54 VFs were shared by a small number of strains but were not present in all strains. Some virulence factors are redundantly present across the species to overcome herd immunity. For example, several enterotoxins are

found as unique variants in different strains to avoid detection by a dynamic host immune system. The vast majority of the virulome is composed of core and pseudocore (shared by most of the strains) genes. The presence of these genes is interesting from an evolutionary point of view because it implies that *S. aureus* has evolved a highly conserved system to carry out its infectious cycle. From an applied viewpoint, these genes may represent targets for virulence inhibitor or antibody-based therapeutic strategies (43).

To gain insights into the metabolic diversity between *S. aureus* strains, we produced 64 strain-specific GEMs and used them to simulate growth capabilities in different nutrient sources. Experimental verification of minimal growth-supporting media was partially confirmed. It is conceivable that certain infectious niches may activate latent biosynthetic pathways under the appropriate conditions, and the fact that genes in these pathways have not developed into pseudogenes indicates that they may have relevance in niche adaptation (44).

A clustering analysis based on computed metabolic phenotypes distinguished groups of strains from each other. One distinction that arose was the identification of USA300 isolate as the only strain with a capability to catabolize spermidine due to the presence of *speG* encoded spermidine acetlytransferase. In humans, spermidine is known to potentiate keratinocyte killing of *S. aureus* in conjunction with antimicrobial peptides (45) and several classes of antibiotics, especially β-lactams (46). Recently, studies have found regions of ACME (including the arginine catabolic *arc* region) in many different, non-USA300 strains (47, 48). However, the association of the *speG* locus with ACME is uncommon outside of USA300. Therefore, acquisition of *speG* has likely been helpful in the rapid spread of USA300 strains. The *speG* gene was horizontally transferred from *S. epidermis*, a commensal colonizer of human skin (49). Therefore, a future comparison of shared metabolic capabilities between *S. aureus* strains and *S. epidermis* might provide further insights. Furthermore, design of antibiotics that specifically interfere with strain-specific capabilities such as those identified here offer new ways to inhibit epidemiological spread of particularly infectious strains.

The GEMs presented here open the *S. aureus* species to a wide array of systems biology methods and techniques (50). For example, using these models, it is straightforward to predict essential genes in different conditions (we found an average of 190 essential genes for the 64 strains in LB media; Dataset S1). Furthermore, single gene essentiality screens have been extended to the study of synthetic lethality and discovery of synergistic antibiotics (51). In conclusion, the multiscale comparative approach used in this work provides insights into the diversity of the *S. aureus* species. Historical methods for classification of *S. aureus* strains have developed considerably from phage-typing to PFGE and MLST approaches (52). New high-throughput DNA sequencing technologies and analysis techniques will continue to improve our ability to track and distinguish pathogens. Computational analysis of GEMs built for multiple strains in a species provides a powerful tool that can be deployed to provide insights into metabolic determinants of virulence, to identify novel biomarkers capable of distinguishing certain strains from one another, and to discover new pharmacological targets.

Materials and Methods

Tools and methods used to identify and construct the core genome and pangenome of *S. aureus* as well as the analysis of atypical genes are presented in *SI Materials and Methods*. Briefly, InParanoid (53) was used to identify orthologous genes to construct a pangenome. The strain-specific model reconstruction procedure was performed with Simpheny (Genomatica, Inc.), and gap-filling algorithms and in silico growth simulation conditions were implemented in Constraints-Based Reconstruction and Analysis Toolbox for Python (COBRApy) (54) and are described in *SI Materials and Methods*. Heat map, phylogenetic tree, and decision tree construction are described in *SI Materials and Methods*. *S. aureus* strains NCTC8325, Newman, N315, Mu50, and USA300 were used for carbon source and growth testing. All experimental protocols are described in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank Samira Dahesh, Satoshi Uchiyama, and Jason Munguia for assistance with culture and storage of the strains used in this work, as well as for helpful insights and discussions. This work was funded by National Institutes of Health/National Institute of General Medical Sciences Grant 1R01GM098105, 1-U01-AI124316-01, and 5-U54-HD071600-05.

1. Chambers HF, Deleo FR (2009) Waves of resistance: Staphylococcus aureus in the antibiotic era. *Nat Rev Microbiol* 7(9):629–641.
2. van Belkum A, et al. (2009) Reclassification of Staphylococcus aureus nasal carriage types. *J Infect Dis* 199(12):1820–1826.
3. Hota B, et al.; CDC Prevention Epicenters (2011) Predictors of clinical virulence in community-onset methicillin-resistant Staphylococcus aureus infections: The importance of USA300 and pneumonia. *Clin Infect Dis* 53(8):757–765.
4. Klevens RM, et al.; Active Bacterial Core surveillance (ABCs) MRSA Investigators (2007) Invasive methicillin-resistant Staphylococcus aureus infections in the United States. *JAMA* 298(15):1763–1771.
5. Otter JA, French GL (2010) Molecular epidemiology of community-associated meticillin-resistant Staphylococcus aureus in Europe. *Lancet Infect Dis* 10(4):227–239.
6. Kennedy AD, et al. (2008) Epidemic community-associated methicillin-resistant Staphylococcus aureus: Recent clonal expansion and diversification. *Proc Natl Acad Sci USA* 105(4): 1327–1332.
7. Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: The bacterial pan-genome. *Curr Opin Microbiol* 11(5):472–477.
8. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pan-genome. *Curr Opin Genet Dev* 15(6):589–594.
9. Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BO (2009) Reconstruction of biochemical networks in microorganisms. *Nat Rev Microbiol* 7(2):129–143.
10. Monk JM, et al. (2013) Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. *Proc Natl Acad Sci USA* 110(50):20338–20343.
11. Ong WK, et al. (2014) Comparisons of Shewanella strains based on genome annotations, modeling, and experiments. *BMC Syst Biol* 8:31.
12. Becker SA, Palsson BO (2005) Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: An initial draft to the two-dimensional annotation. *BMC Microbiol* 5:8.
13. Heinemann M, Kümmel A, Ruinatscha R, Panke S (2005) In silico genome-scale reconstruction and validation of the Staphylococcus aureus metabolic network. *Biotechnol Bioeng* 92(7):850–864.
14. Lee DS, et al. (2009) Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple Staphylococcus aureus genomes identify novel antimicrobial drug targets. *J Bacteriol* 191(12):4015–4024.
15. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2009) GenBank. *Nucleic Acids Res* 37(Database issue):D26–D31.
16. Geer LY, et al. (2010) The NCBI BioSystems database. *Nucleic Acids Res* 38(Database issue):D492–D496.
17. Holt DC, et al. (2011) A very early-branching Staphylococcus aureus lineage lacking the carotenoid pigment staphyloxanthin. *Genome Biol Evol* 3:881–895.
18. Hall BG, Ehrlich GD, Hu FZ (2010) Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. *Microbiology* 156(Pt 4):1060–1068.
19. Tatusov RL, et al. (2003) The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* 4:41.
20. Scaria J, et al. (2010) Analysis of ultra low genome conservation in Clostridium difficile. *PLoS One* 5(12):e15147.
21. Rasko DA, et al. (2008) The pangenome structure of Escherichia coli: Comparative genomic analysis of E. coli commensal and pathogenic isolates. *J Bacteriol* 190(20): 6881–6893.
22. Feil EJ, et al. (2003) How clonal is Staphylococcus aureus? *J Bacteriol* 185(11):3307–3316.
23. Baba T, et al. (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: The Keio collection. *Mol Syst Biol* 2:2006.0008.
24. Kobayashi K, et al. (2003) Essential Bacillus subtilis genes. *Proc Natl Acad Sci USA* 100(8):4678–4683.
25. Price MN, Dehal PS, Arkin AP (2008) Horizontal gene transfer and the evolution of transcriptional regulation in Escherichia coli. *Genome Biol* 9(1):R4.
26. Gordienko EN, Kazanov MD, Gelfand MS (2013) Evolution of pan-genomes of Escherichia coli, Shigella spp., and Salmonella enterica. *J Bacteriol* 195(12):2786–2792.
27. Kaas RS, Friis C, Ussery DW, Aarestrup FM (2012) Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse Escherichia coli genomes. *BMC Genomics* 13:577.
28. Alam MT, et al. (2015) Transmission and microevolution of USA300 MRSA in U.S. households: Evidence from whole-genome sequencing. *MBio* 6(2):e00054.
29. Mirkin BG, Fenner TI, Galperin MY, Koonin EV (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol* 3:2.
30. Chen L, Xiong Z, Sun L, Yang J, Jin Q (2012) VFDB 2012 update: Toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Res* 40(Database issue):D641–D645.

31. Kehl-Fie TE, Skaar EP (2010) Nutritional immunity beyond iron: A role for manganese and zinc. *Curr Opin Chem Biol* 14(2):218–224.
32. Park B, Nizet V, Liu GY (2008) Role of Staphylococcus aureus catalase in niche competition against Streptococcus pneumoniae. *J Bacteriol* 190(7):2275–2278.
33. Knight BC (1937) The nutrition of Staphylococcus aureus; nicotinic acid and vitamin B(1). *Biochem J* 31(5):731–737.
34. Gladstone GP (1937) The nutrition of Staphylococcus aureus; Nitrogen requirements. *Br J Exp Pathol* 18(4):322–333.
35. LaCroix RA, et al. (2015) Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. *Appl Environ Microbiol* 81(1):17–30.
36. Orth JD, Thiele I, Palsson BO (2010) What is flux balance analysis? *Nat Biotechnol* 28(3):245–248.
37. Thurlow LR, et al. (2013) Functional modularity of the arginine catabolic mobile element contributes to the success of USA300 methicillin-resistant Staphylococcus aureus. *Cell Host Microbe* 13(1):100–107.
38. Zhu Y, et al. (2007) Staphylococcus aureus biofilm metabolism and the influence of arginine on polysaccharide intercellular adhesin synthesis, biofilm formation, and pathogenesis. *Infect Immun* 75(9):4219–4226.
39. Seiler N, Atanassov CL (1994) The natural polyamines and the immune system. *Prog Drug Res* 43:87–141.
40. Enright MC, et al. (2002) The evolutionary history of methicillin-resistant Staphylococcus aureus (MRSA). *Proc Natl Acad Sci USA* 99(11):7687–7692.
41. Fitzgerald JR, Sturdevant DE, Mackie SM, Gill SR, Musser JM (2001) Evolutionary genomics of Staphylococcus aureus: Insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic. *Proc Natl Acad Sci USA* 98(15): 8821–8826.
42. McCarthy AJ, Lindsay JA (2012) The distribution of plasmids that carry virulence and resistance genes in Staphylococcus aureus is lineage associated. *BMC Microbiol* 12:104.
43. Morrison C (2015) Antibacterial antibodies gain traction. *Nat Rev Drug Discov* 14(11): 737–738.
44. McGavin MJ, Arsic B, Nickerson NN (2012) Evolutionary blueprint for host- and niche-adaptation in Staphylococcus aureus clonal complex CC30. *Front Cell Infect Microbiol* 2:48.
45. Planet PJ, et al. (2013) Emergence of the epidemic methicillin-resistant Staphylococcus aureus strain USA300 coincides with horizontal transfer of the arginine catabolic mobile element and speG-mediated adaptations for survival on skin. *MBio* 4(6): e00889–e13.
46. Kwon DH, Lu CD (2007) Polyamine effects on antibiotic susceptibility in bacteria. *Antimicrob Agents Chemother* 51(6):2070–2077.
47. Goering RV, et al. (2007) Epidemiologic distribution of the arginine catabolic mobile element among selected methicillin-resistant and methicillin-susceptible Staphylococcus aureus isolates. *J Clin Microbiol* 45(6):1981–1984.
48. Sabat AJ, et al. (2013) Novel organization of the arginine catabolic mobile element and staphylococcal cassette chromosome mec composite island and its horizontal transfer between distinct Staphylococcus aureus genotypes. *Antimicrob Agents Chemother* 57(11):5774–5777.
49. Barbier F, et al. (2011) High prevalence of the arginine catabolic mobile element in carriage isolates of methicillin-resistant Staphylococcus epidermidis. *J Antimicrob Chemother* 66(1):29–36.
50. Lewis NE, Nagarajan H, Palsson BO (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat Rev Microbiol* 10(4):291–305.
51. Aziz RK, et al. (2015) Model-driven discovery of synergistic inhibitors against E. coli and S. enterica serovar Typhimurium targeting a novel synthetic lethal pair, aldA and prpC. *Front Microbiol* 6:958.
52. DeLeo FR, et al. (2011) Molecular differentiation of historic phage-type 80/81 and contemporary epidemic Staphylococcus aureus. *Proc Natl Acad Sci USA* 108(44): 18091–18096.
53. Sonnhammer EL, Östlund G (2015) InParanoid 8: Orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res* 43(Database issue):D234–D239.
54. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR (2013) COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst Biol* 7:74.
55. Thompson JD, Gibson TJ, Higgins DG (2002) Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics* Chapter 2:Unit 2.3.
56. Guindon S, Delsuc F, Dufayard JF, Gascuel O (2009) Estimating maximum likelihood phylogenies with PhyML. *Methods Mol Biol* 537:113–137.
57. O'Brien KP, Remm M, Sonnhammer EL (2005) Inparanoid: A comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* 33(Database issue):D476–D480.
58. Galardini M, et al. (2014) DuctApe: A suite for the analysis and correlation of genomic and OmniLog™ Phenotype Microarray data. *Genomics* 103(1):1–10.
59. Heaps HS (1978) *Information Retrieval: Computational and Theoretical Aspects* (Academic, New York).
60. Waack S, et al. (2006) Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics* 7:142.
61. Merkl R (2004) SIGI: Score-based identification of genomic islands. *BMC Bioinformatics* 5: 22.
62. Langille MG, Hsiao WW, Brinkman FS (2008) Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinformatics* 9:329.
63. Henry CS, et al. (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol* 28(9):977–982.
64. Oh YK, Palsson BO, Park SM, Schilling CH, Mahadevan R (2007) Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data. *J Biol Chem* 282(39):28791–28799.
65. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 40(Database issue): D109–D114.
66. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28(1):27–30.
67. Caspi R, et al. (2008) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* 36(Database issue):D623–D631.
68. Thiele I, Palsson BO (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc* 5(1):93–121.
69. Ganter M, Bernard T, Moretti S, Stelling J, Pagni M (2013) MetaNetX.org: A website and repository for accessing, analysing and manipulating metabolic networks. *Bioinformatics* 29(6):815–816.
70. Schellenberger J, Park JO, Conrad TM, Palsson BO (2010) BiGG: A Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* 11:213.
71. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2005) GenBank. *Nucleic Acids Res* 33(Database issue):D34–D38.
72. Theodore TS, Panos C (1973) Protein and fatty acid composition of mesosomal vesicles and plasma membranes of Staphylococcus aureus. *J Bacteriol* 116(2):571–576.
73. Varma A, Palsson BO (1994) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *Appl Environ Microbiol* 60(10):3724–3731.
74. Reed JL, et al. (2006) Systems approach to refining genome annotation. *Proc Natl Acad Sci USA* 103(46):17480–17484.
75. Nizet V (2007) Understanding how leading bacterial pathogens subvert innate immunity to reveal novel therapeutic targets. *J Allergy Clin Immunol* 120(1):13–22.
76. Demsar J, et al. (2013) Orange: Data Mining Toolbox in Python. *J Mach Learn Res* 14(Aug):2349–2353.
77. Noble WC, Virani Z, Cree RG (1992) Co-transfer of vancomycin and other resistance genes from Enterococcus faecalis NCTC 12201 to Staphylococcus aureus. *FEMS Microbiol Lett* 72(2):195–198.
78. Corvaglia AR, et al. (2010) A type III-like restriction endonuclease functions as a major barrier to horizontal gene transfer in clinical Staphylococcus aureus strains. *Proc Natl Acad Sci USA* 107(26):11954–11958.
79. Sung JM, Lindsay JA (2007) Staphylococcus aureus strains that are hypersusceptible to resistance gene transfer from enterococci. *Antimicrob Agents Chemother* 51(6):2189–2191.
80. Weigel LM, et al. (2003) Genetic analysis of a high-level vancomycin-resistant isolate of Staphylococcus aureus. *Science* 302(5650):1569–1571.
81. Grosjean H, et al. (2014) Predicting the minimal translation apparatus: Lessons from the reductive evolution of mollicutes. *PLoS Genet* 10(5):e1004363.
82. Fang G, Rocha E, Danchin A (2005) How essential are nonessential genes? *Mol Biol Evol* 22(11):2147–2156.
83. Koonin EV (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol* 1(2):127–136.
84. Mushegian A (2008) Gene content of LUCA, the last universal common ancestor. *Front Biosci* 13:4657–4666.
85. Ciccarelli FD, et al. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* 311(5765):1283–1287.
86. Yutin N, Puigbò P, Koonin EV, Wolf YI (2012) Phylogenomics of prokaryotic ribosomal proteins. *PLoS One* 7(5):e36972.
87. Byrgazov K, Vesper O, Moll I (2013) Ribosome heterogeneity: Another level of complexity in bacterial translation regulation. *Curr Opin Microbiol* 16(2):133–139.
88. Lecompte O, Ripp R, Thierry JC, Moras D, Poch O (2002) Comparative analysis of ribosomal proteins in complete genomes: An example of reductive evolution at the domain scale. *Nucleic Acids Res* 30(24):5382–5390.
89. Akanuma G, et al. (2012) Inactivation of ribosomal protein genes in Bacillus subtilis reveals importance of each ribosomal protein for cell proliferation and cell differentiation. *J Bacteriol* 194(22):6282–6291.
90. Bubunenko M, Baker T, Court DL (2007) Essentiality of ribosomal and transcription antitermination proteins analyzed by systematic gene replacement in Escherichia coli. *J Bacteriol* 189(7):2844–2853.
91. Shoji S, Dambacher CM, Shajani Z, Williamson JR, Schultz PG (2011) Systematic chromosomal deletion of bacterial ribosomal protein genes. *J Mol Biol* 413(4):751–761.
92. Chadani Y, et al. (2010) Ribosome rescue by Escherichia coli ArfA (YhdL) in the absence of trans-translation system. *Mol Microbiol* 78(4):796–808.
93. Condon C (2003) RNA processing and degradation in Bacillus subtilis. *Microbiol Mol Biol Rev* 67(2):157–174.
94. Ehrenreich A, Forchhammer K, Tormay P, Veprek B, Böck A (1992) Selenoprotein synthesis in E. coli. Purification and characterisation of the enzyme catalysing selenium activation. *Eur J Biochem* 206(3):767–773.
95. Chaudhuri RR, et al. (2009) Comprehensive identification of essential Staphylococcus aureus genes using Transposon-Mediated Differential Hybridisation (TMDH). *BMC Genomics* 10:291.
96. Orth JD, et al. (2011) A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011. *Mol Syst Biol* 7:535.

Supporting Information

Bosi et al. 10.1073/pnas.1523199113

SI Materials and Methods

Constructing a Representative Dataset of *S. aureus* Species. The genomic sequences for all *S. aureus* strains were downloaded from the ftp site of NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/). To establish the phylogenetic relationship existing between *S. aureus* representatives, extensive phylogenetic analysis was conducted by using a concatamer of a set of seven housekeeping genes (*arcC*, *aroE*, *glpF*, *gmk*, *pta*, *tpi*, and *yqiL*) to construct a phylogenetic tree (Fig. 1). Multiple sequence alignments were performed with ClustalW (55). Maximum likelihood (ML) analysis was carried out using Phyml tool (56), with a Whelan and Goldman model of amino acid substitution. Statistical support at nodes was obtained by nonparametric bootstrapping on 100 resampled datasets. A set of 64 strains was selected from the dendrogram to create a heterogeneous dataset in terms of (*i*) drug resistance (MRSA, MSSA, VRSA, and VISA), (*ii*) host specificity (human vs. animal), (*iii*) virulence/environmental association (CA-MRSA, HA-MRSA, and LA-MRSA), and (*iv*) tree topology. We wanted to sample strains from different clusters of the tree. Therefore, the final dataset has been designed to be, as far as possible, an unbiased representation of the whole *S. aureus* species.

Orthologous Gene Identification and Pangenome Construction. The homology relationships between genes of each different strain were assessed using InParanoid (57). Considering that the dataset was composed of 64 *S. aureus* strains, the systematic use of InParanoid for all of the pairwise combinations resulted in a total of 2,016 InParanoid tables as output, representing the orthologous genes for each genome pair. These were also used for mapping the genomic contents of all of the *S. aureus* strains onto the reference GEM of *S. aureus* N315 (see below). Representative genes for each cluster were functionally annotated using the COG database (19), allowing us to assign a functional category to each cluster of orthologous genes.

Estimation of Core Genomes and Pangenomes. To compute the pangenome of the *S. aureus* species, the dataset comprising each *S. aureus* representative was analyzed using the dgenome module of the Ductape suite [Galardini et al. (58)], which allowed for a fast computation of the pangenome using a pairwise best bidirectional hit approach. The estimation of generalized pangenome metrics, such as core genome and pangenome size, was performed by simulating random pangenomes with number of genomes (N) ranging from 2 to 64. For each N, a total of 10 random combinations of organisms were sampled. From each of these pangenomes, the total number of genes and the number of conserved and new genes were determined, corresponding to pangenome, core, and unique genome sizes, respectively. Also, these numbers were used to the estimate the pangenome parameters (see below). As reported by Tettelin et al. (7), Heap's law, an empirical law originally used in the field of information retrieval (59), can be used to describe the *S. aureus* pangenome size and the number of new genes. These power laws, $N(\text{pan}) = k_1 N^{\gamma}$ and $N(\text{new}) = k_2 N^{-\alpha}$, were fitted to the number of total genes and new genes, respectively, to find the parameters giving the best fit, using the curve_fit function of the Python package Scipy. Similarly, to estimate the *S. aureus core* genome the following double exponential decay function was fit on the number of core genes: $N(\text{core}) = k_1 \exp(-N/\tau_1) + k_2 \exp(-N/\tau_2) + \Theta$. To obtain a reliable estimation, the free parameters to be optimized (k_1, k_2, τ_1, τ_2, and Θ) have been optimized with the curve_fit function with the following initial guesses: 1, 1, 0.1, 0.1, and 1,400, respectively.

Identification of Atypical *S. aureus* Genes. The GC content and codon bias of all core genes was used to detect genes with atypical compositional features. For each strain, the core GC distribution has been obtained by computing the average GC content of each core gene, which represents genes that have not been transferred; therefore, their GC distribution reflects the GC value of not-transferred genes. This has been used to estimate a lower and upper threshold value, corresponding to the values composing 99% of the distribution. The genes with GC values not included between these values were reported as atypical genes. The codon use differences between core genes and pangenes were determined using the Hidden Markov Model (HMM)-based tool (SIGI-HMM) (60, 61) that detects genomic islands based on statistical analysis of codon use with high precision (62).

Strain-Specific Model Reconstruction. Genome-scale metabolic reconstructions were available for some *S. aureus* strains (12–14), including a previously published model from our group named *iSB619*. This model consists of 615 genes that catalyze 742 reactions. We first combined this model with other models of *S. aureus* to form a reference model for *S. aureus* N315. Then, we extended this model by adding content from metabolic databases including Kyoto Encyclopedia of Genes and Genomes (KEGG), SEED (63), and MetaCyc. Finally, the reconstruction was manually curated to form an updated reconstruction of *S. aureus* N315 that represents an improved genome-scale representation of the metabolism of this strain, containing 1,475 reactions, 1,232 metabolites, and 850 genes. This model was then used with the genomic sequence of *S. aureus* N315 as a reference to map the genomic content of other *S. aureus* strains to obtain a set of shared genes and reactions present in all genomes. All genomes were reannotated using the RAST server (63). Next, other models of *S. aureus* strains and related organisms were examined for strain-specific metabolic content. These included the metabolic models of 13 *S. aureus* strains (Mu50, MW2, COL, EMRSA-16 strain 252, methicillin-sensitive strain 476, JH1, JH9, RF122, USA300, USA300 TCH1516, Newman, and N315) (14) and the curated model of *B. subtilis* 168 (*iBsu1103*) (64). Additional reaction content was added from ModelSEED (63), KEGG (65, 66), and BIOCYC (67). All reactions added were manually curated according to published protocol (68). MetaNetX (69) was used to standardize metabolites and reactions to Systems Biology Research Group (70) abbreviations. All genome sequences were downloaded from GenBank (71) on July 16, 2015. Gene names conform to the NCBI locus name according to the original annotation in GenBank.

Biomass Composition. A detailed culture-based biomass composition for *S. aureus* is not available in literature, so we combined biomass compositions from three previous models of *S. aurues* (12, 13) to form the biomass composition for the strain-specific models presented here. These models also assumed that the production rates of metabolites required for cellular growth were similar to those of the related gram-positive organism *B. subtilis* (64). We adjusted the fatty acid composition of the biomass function based on *S. aureus*-specific data (72). The detailed biomass composition is provided in Dataset S1, worksheet 8 and Table S1.

In Silico Growth Simulations. Each of the 64 *S. aureus* models is available as an SBML file (Dataset S1). The models were loaded into the COBRApy toolbox (54) to perform flux balance analysis (FBA) (73). M9 minimal media was simulated by setting a lower bound of −1,000 (allowing unlimited uptake) on the exchange reactions for Ca^{2+}, Cl^{-}, CO_2, Co^{2+}, Cu^{2+}, Fe^{2+}, Fe^{3+}, H^{+}, H_2O,

K^+, Mg^{2+}, Mn^{2+}, MoO_4^{2-}, Na^+, Ni^{2+}, SeO_4^{2-}, SeO_3^{2}, and Zn^{2+}. The default carbon source was glucose with a lower bound of −10, the default nitrogen source was NH_4^- with a lower bound of −1,000, the default phosphorous source was HPO_4^2 with a default bound of −1,000, and the default sulfur source was SO_4^{2-} with a default bound of −1,000 (Dataset S1, worksheet 5). To identify sole growth-supporting carbon, nitrogen, phosphorous, and sulfur sources, each of these default compounds were removed from the media (lower bound set to 0) one at a time, and different compounds were added to determine if they supported growth. For aerobic simulations, O_2 was added with a lower bound of −20 and to 0 for anaerobic simulations. For models with identified auxotrophies, the compound for which a strain was auxotrophic (Dataset S1, worksheet 6) was also added to the M9 minimal media for each simulation with a lower bound of −1. Model growth phenotypes were determined using FBA one at a time on each condition with the core biomass reaction as the objective. Nutrient sources with growth rates above zero were classified as growth supporting, whereas nutrient sources with growth rates of zero were classified as non-growth supporting. The Gurobi 6.0 linear programming solver (Gurobi Optimization, Inc.) was used to perform FBA.

Gap Filling. The Constraints-Based Reconstruction and Analysis implementation of the SMILEY algorithm (growMatch) (74) was used to predict sets of exchange and gap-filling reactions for models that were unable to simulate biomass in silico on M9 minimal media with glucose aerobically using FBA. The universal set of reactions used to fill gaps was based on MetaNetX. The Gurobi 5.0.0 mixed-integer linear programming solver was used (Gurobi Optimization Inc.) to implement SMILEY. When adding content to enable the strains to grow, exchange reactions indicating strain-specific auxotrophies were prioritized over adding new reactions without genetic evidence. Glucose (carbon source), phosphate, sulfate, nicotinamide, and thiamine were both experimentally used and computationally verified. However, other substrates, such as the nucleosides cytidine and uridine, were predicted not to be required in their metabolic model.

Prediction of All Growth-Supporting Carbon, Nitrogen, Phosphorus, and Sulfur Sources. The possible growth-supporting carbon, nitrogen, phosphorus, and sulfur sources of each *S. aureus* strain were identified using FBA. First, all exchange reactions for extracellular metabolites containing the four elements were identified from the metabolite formulas. Every extracellular compound containing carbon was considered a potential carbon source, for example. Next, to determine possible growth-supporting carbon sources, the lower bound of the glucose exchange reaction was constrained to zero. Then the lower bound of each carbon exchange reaction was set, one at a time, to −10 $mmol{\cdot}gDW^{-1}{\cdot}h^{-1}$ (a typical uptake rate for growth-supporting substrates), and growth was maximized by FBA using the biomass reaction. The target substrate was considered growth supporting if the predicted growth rate was above zero. While identifying carbon sources, the default nitrogen, phosphorus, and sulfur sources were ammonium (nh4), inorganic phosphate (pi), and inorganic sulfate (so4), respectively. Prediction of growth-supporting sources of these other three elements was performed in the same manner as growth on carbon, with glucose as the default carbon source.

Prediction of Gene Essentiality. To simulate the effects of gene knockouts, each *S. aureus* model was used with its default constraints and biomass reaction objective. For growth on glucose, the lower bound of the glucose exchange reaction was set to −10 $mmol{\cdot}gDW^{-1}{\cdot}h^{-1}$. All genes in each model were knocked out one a time, and growth was simulated by FBA. This assessment can be performed by using FBA with the additional constraint of setting the fluxes through all of the reactions that cannot occur without a given gene (i.e., when isozymes independently catalyzing a same reaction are not present) for which the essentiality is being tested equal to zero. Gene knockout strains with a growth rate above zero were considered nonessential.

Virulome Construction. A curated set of virulence factors was obtained from the virulence factor database and literature references (30, 75). The presence of these VFs within the *S. aureus* genomes was determined using a BLAST search on the curated database using the blastp algorithm with following thresholds: e-value <1e-20, sequence similarity >70%, and alignment length >60%.

Heat Map and Phylogenetic Tree Construction. The binary results from the growth/no growth simulations for each strain were used to compute a correlation matrix based on dissimilarity indices calculated using the Jaccard method in the vegdist function of the Vegan R package. Ward's agglomerative clustering of the matrix of correlations was used to cluster the species using the hclust function of the Vegan R package and used to form a dendrogram. The heat map was visualized using the gplots R package with values aligned based on the calculated dendrogram.

Decision Tree Construction. A decision tree (Fig. 5) was calculated based on growth/no growth values for each strain classified into their major pathotypes: InPec, ExPec, or commensal. The classification tree tool, part of the Orange Canvas software package (76), was used to calculate and display the decision tree using a Gini Index attribute selection criteria with no binarization, two minimum leaves for prepruning, and $m = 2$ estimate for postpruning with leaves of the same majority class being recursively merged.

SI Text

Analysis of Atypical *S. aureus* Genes. To investigate the effect of HGT events on the *S. aureus* species, we searched for atypical genes within each strain by computing DNA composition (GC content) and codon use differences between core genes and pangenes (59). A total of 4,277 and 1,788 atypical genes were identified with the average genome having 49 and 28 unique genes present based on GC content and codon bias, respectively. Most of the identified atypical genes have no known COG functional class. The phylogenetic origin of each atypical *S. aureus* gene was traced using the nr-database. For each gene the best non-*S. aureus* hit was taken as the putative transfer source (PTS). We hypothesized that more recent HGT events were indicated by genes with fewer *S. aureus* hits, i.e., those more similar to the query gene than the PTS. We used this number as an index to estimate the ancestrality of the HGT events for each PTS, a measurement we term the "ancestrality index" (AI). It was only possible to find a significant PTS for 25% of the atypical genes we identified. Those PTS with an AI >2 mostly corresponded to HGT events that occurred at the genus level, whereas the majority of the more recent PTS (an AI ≤2) corresponded to HGT events at higher taxonomical levels (order, class, and phylum). In this group we found a high representation of proteins such as hypothetical proteins, mobilization proteins, metal and antibiotic resistance genes, and virulence factors.

We assessed the effect of HGT events in shaping the diversity within this species. Analysis of atypical genes showed how these are mostly derived from taxonomically related donors (i.e., representatives of the same species/genus), with a minor portion of genes coming from host-associated bacteria. This finding suggests the presence of a taxonomical barrier limiting the amount of HGT in this species. The HGT events may be a major source of newly acquired antibiotic resistances. Therefore, the number of unique genes per genome may identify strains more prone to HGT and hence more likely to acquire new functionalities. Because virulence and antibiotic resistance genes are often involved in HGT events (42, 77), strains with a higher portion of unique genes (and thus more predisposed to exchange of exogenous DNA) may represent a major public health threat because

these can develop virulent and/or multidrug-resistant phenotypes by horizontally acquiring corresponding gene cassettes (78–80); however, we did not see any correlation between number of atypical genes and the number of identified virulence factors.

E-Gene Conservation. Most antibiotics not only target an organism's metabolic functions but also target the transcriptional and translational machinery of the target organism. We compared the conservation of transcriptional and translational machinery within *S. aureus* strains across the species. Genes involved in these processes were curated from *E. coli* and *B. subtilis* because *E. coli* is the organism for which almost all components of transcription and translation machinery have been identified and experimentally characterized. However, because *S. aureus* is phylogenetically closer to *B. subtilis* (both are Firmicutes), additional proteins from *B. subtilis* were used to assemble this dataset. Although *B. subtilis* homologs exist for most of the *E. coli* genes involved in transcription and translation, a few *B. subtilis* genes exist for which no homologs are found in the *E. coli* genome and vice versa. Altogether, we selected 289 query genes of which 192 are shared between *E. coli* and *B. subtilis*, whereas 59 are unique to *E. coli* and 42 are unique to *B. subtilis* (Fig. S3). The set included genes that encode functions for transcription, ribosome biogenesis, tRNA maturation and aminoacylation, and proteins and cofactors required for mRNA translation and RNA decay. Using these experimentally validated genes as input (81), genes encoding proteins of the core transcription and translation machinery were predicted in the 64 *S. aureus* strains.

A core set of 239 genes involved in transcription and translation was found in all *S. aureus* strains examined. The majority of genes coding for ribosomal proteins, aminoacyl-tRNA synthetases, translation factors, and several ribosome biogenesis/maturation enzymes are universally conserved in *S. aureus* strains. These same genes are essential in both *E. coli* and *B. subtilis* (82–84) and other bacterial organisms (85, 86). Conversely, 47 of the 289 genes were absent in all *S. aureus* strains examined. Of genes absent in all *S. aureus* strains examined, 7 are present in both *E. coli* and *B. subtilis*, 33 are unique to *E. coli*, and 7 are unique to *B. subtilis*. Most of these missing genes encode functions in transcription, tRNA modifications, rRNA modifications, and RNA processing.

Next we examined genes present in some *S. aureus* strains but not others. These included the ribonuclease RNE involved in RNA processing and the 50s ribosomal protein L33 (RPMGA). These two proteins were conserved in 28 of the *S. aureus* strains examined. RMPGA was missing in 19 strains, and RNE was missing in 10 strains. Although most of the *S. aureus* strains retained the two genes encoding L33a (RpmGa) and L33b (RpmGb), L33a was lost in 19 of the 64 strains examined. L33 is responsible for cellular ribosome heterogeneity and may generate specialized ribosomes in response to stress conditions and environmental changes (87). These proteins could have evolved to fulfill specific nonessential innovation (88) and hence easily be lost in reductive evolutions. The L33 gene is also nonessential in *E. coli* and *B. subtilis* (23, 89–91). Our analysis defines the minimal and conserved set of genes needed to encode functions that sustain protein synthesis in various *S. aureus* strains.

A core set of 239 genes involved in transcription and translation were found in all *S. aureus* strains examined (Fig. S1). Of these, 183 were present in both *E. coli* and *B. subtilis*, whereas 23 were unique to *E. coli*, and 33 were unique to *B. subtilis*.

Ribosomal Proteins, Aminoacyl-tRNA Synthetases, Translation Factors, and Several Ribosome Biogenesis/Maturation Enzymes Are Highly Conserved in *S. aureus*. Only one translation factor, ArfA, the alternative ribosome rescue factor A, was missing from all *S. aureus* strains. It was shown recently that ribosomes stalled on nonstop mRNAs are rescued by dual mechanisms in *E. coli*: tmRNA mediated transtranslation [encoded for by SsrA and ArfA-mediated peptidyl-tRNA hydrolysis (92)]. ArfA functions by recruiting release factor 2 (RF2) to release tRNA, and the presence of ArfA is essential in the absence of Ssra. In addition to all of the *S. aureus* strains, ArfA is also missing in *B. subtilis*. Thus, because they are missing the ArfA backup function, the SsrA function is essential in these organisms.

Certain drugs like tetracycline prevent the aminoacyl-tRNA from binding to the ribosomal subunit in prokaryotes. All *S. aureus* genomes analyzed encoded the complete set of aminoacyl-tRNA synthetases and protein cofactors required to charge all 20 canonical amino acids. Of the 33 aa-tRNA synthetases examined, 31 were conserved in all *S. aureus* strains. The two missing genes include SelA and SelD. SelA encodes selenocysteine synthase that catalyzes the conversion of serine to selenocysteine on serine-charged tRNASec. SelD encodes selenide, and water dikinase catalyzes the reaction that produces selenophosphate, the selenium donor for the biosynthesis of selenocysteine and modification of thiouridine to selenouridine in certain tRNAs (89). Both genes are missing in *B. subtilis* and all of the *S. aureus* strains examined.

Genes Coding for Enzymes Involved in Transcription, rRNA and Protein Processing, RNA or Protein Modification, and Ribosome Maturation RNases Are Less Conserved in *S. aureus*. Of the 29 genes involved in transcription, 12 were absent in all *S. aureus* strains. Of the 12 genes missing, 7 are also absent from *B. subtilis* including RpoE, RpoH, and RpoZ as well as FecL, DmsD, Mfd, and TorD. RpoN was present in both *E. coli* and *B. subtilis* but lacking in all *S. aureus* strains examined. Finally, SigVWXZ were missing in all *S. aureus* strains, despite being present in *B. subtilis* indicating different transcriptional control strategies.

Out of the total 33 genes coding for rRNA modification enzymes in both *E. coli* and *B. subtilis*, 7 are absent in all strains of *S. aureus* examined. Of these, 4 ribosomal RNA large subunit methyltransferases are also missing in *B. subtilis* (RlmE, RlmF, RlmJ, and RlmM). RlmA and RsmJ are found in both *E. coli* and *B. subtilis* but not *S. aureus*. RsmJ mutants in *E. coli* are cold sensitive and show a growth defect at 16 °C. Of the 39 genes examined in ribosome assembly, 33 are present in all *S. aureus* strains. Of the genes missing, all five are also all absent in *B. subtilis*.

Of the 27 genes coding for RNases and related proteins, 9 were absent in all *S. aureus* strains. Eighteen were found in all of the strains analyzed, including two genes unique to *E. coli* (absent from *B. subtilis*, RNB and RNT). RNases generally harbor broad, sometimes overlapping specificity with other RNases, making it difficult to determine their intrinsic essentiality. Also, compared with the other categories of proteins analyzed above, the set of RNases in gram-negative and gram-positive bacteria are quite different, and some RNases are essential in one organism but not in the other (89, 93). Furthermore, one gene was found to be differentially distributed in *S. aureus* strains, the RNA degradosome binding protein RnE. RnE is found in *E. coli* but not *B. subtilis* and was found in 43 of the 64 *S. aureus* genomes analyzed. RnE is an endonuclease that cleaves single stranded regions of pre-RNA transcripts. Its function is similar to the conserved RnjA and RnjB, two enzymes found only in gram-positive bacteria.

Ribosomal Proteins. The major function of a ribosomal protein is stabilization of the rRNA structure. Of the 60 genes encoding for r-proteins present in ribosomes of *E. coli* and or *B. subtilis*, 58 are present in the 53 *S. aureus* strains examined. Only one gene (RpsV/sra) encoding the ribosome associated protein S22/RpsV was missing in all *S. aureus* strains examined. This gene is also absent in *B. subtilis*. It is nonessential in *E. coli* and codes for the substoichiometric component of the 30S ribosomal subunit.

Most of the strains tend to retain the two genes encoding L33a (RpmGa) and L33b (RpmGb), but L33a was lost in 19 of the 53 strains examined. L33 is known to be responsible for cellular ribosome heterogeneity, probably generating specialized ribosomes in response to stress conditions and environmental changes. These

proteins could have evolved to fulfill specific nonessential innovations and hence could easily be lost in reductive evolutions. This gene is also nonessential in *E. coli* and *B. subtilis*.

Translation Factors. In addition to the core ribosomal components, protein synthesis requires several translation factors that ensure the speed and fidelity of translation as well as the functionality of the nascent polypeptide. Most of them are found in all *S. aureus* strains examined, illustrating the conservation of this bacterial apparatus in the bacterial world. Only one translation factor, ArfA, the alternative ribosome rescue factor A, was missing from all *S. aureus* strains. It was shown recently that ribosomes stalled on nonstop mRNAs are rescued by dual mechanisms in *E. coli*: tmRNA mediated transtranslation (encoded for by *ssrA*) and ArfA-mediated peptidyl-tRNA hydrolysis (55). ArfA functions by recruiting release factor 2 (RF2) to release tRNA and the presence of ArfA is essential in the absence of Ssra. In addition to all of the *S. aureus* strains, ArfA is also missing in *B. subtilis*, making the SsrA function essential in these organisms.

Aminoacyl-tRNA Synthetases. Aminoacyl-tRNA (aa-tRNA or charged tRNA) is tRNA to which its cognate amino acid is chemically bonded (charged). Certain drugs like tetracycline prevent the aminoacyl-tRNA from binding to the ribosomal subunit in prokaryotes. All *S. aureus* genomes analyzed encoded the complete set of aminoacyl-tRNA synthetases and protein cofactors required to charge all 20 canonical amino acids. Of the 33 aa-tRNA synthetases examined, 29 were conserved in all *S. aureus* strains.

SelA encodes selenocysteine synthase that catalyzes the conversion of serine to selenocysteine on serine-charged tRNASec. *SelD* encodes selenide, and water dikinase catalyzes the reaction that produces selenophosphate, the selenium donor for the biosynthesis of selenocysteine and modification of thiouridine to selenouridine in certain tRNAs (94). Both genes are missing in *B. subtilis* and all of the *S. aureus* strains examined.

Two other genes, *glyQ* and *glyS*, are present in *B. subtilis* but not conserved at the sequence level with their homologs in *S. aureus* strains. It was determined that *glyS* and *glyQ* are actually present in *S. aureus* at an equivalent position within a syntenic region that encodes a class-II glycyl-tRNA synthetase, similar to that encoded by *Bacillus* cereus (31). These genes are also essential in *S. aureus* (95).

Transfer RNA Modification Enzymes. Transfer RNA (tRNA) precursors are subject to enzymatic posttranscriptional modification at many positions of the base or ribose moieties. These modifications stabilize the tRNA tertiary structure, introduce recognition determined and antideterminants toward RNA-interacting macromolecules, and fine-tune the decoding process at the level of both efficiency and fidelity. Of the 44 tRNA modification enzymes examined, 35 of the 44 are present in all *S. aureus* strains. All of the nine missing genes (*cmoA*, *cmoB*, *mnmC*, *selU*, *tmcA*, *trmJ*, *truD*, *tsaA*, and *ttcA*) all are also absent in *B. subtilis*. However, of the 35 genes present in all *S. aureus* strains, 6 are missing in *B. subtilis* but present in *E. coli*. These genes are *dusA*, *dusC*/DUS2, *gluQ*, *trmA*, *trmH*, and *truC* and were possibly acquired in lateral gene transfer.

Ribosomal RNA Modification Enzymes. Many bases and riboses of rRNAs are posttranscriptionally modified like in tRNAs. Most modifications are introduced during pre-rRNA maturation and ribosome assembly, and just a few are formed at the level of the 30S and 50S subparticles or of the entire 70S ribosome. The conservation and clustering of modifications in the decoding center of the 30S subunit and in the peptidyl-transferase center of the 50S subunit attests their important roles in the translation process.

Out of the total 33 genes coding for rRNA modification enzymes in both *E. coli* and *B. subtilis*, only 7 are absent in all strains of *S. aureus* examined. Of these, 4 ribosomal RNA large subunit methyltransferases are also missing in B. subtilius (RlmE, RlmF, RlmJ, and RlmM). RlmA and RsmJ are found in both *E. coli* and *B. subtilis* but not *S. aureus*. RsmJ mutants in *E. coli* are cold sensitive and show a growth defect at 16C. YqxC is found in *B. subtilis* alone but not *S. aureus*; this gene encodes a FtsJ/Spb1/SPOUT-like 2′-O-ribose RNA methyltransferase (81).

Ribosome Assembly, Protein Chaperones, Helicases, and Protein Modifications. In bacteria, the assembly of r-proteins onto precursor rRNA scaffolds to form functional 30S and 50S subunits requires over a dozen assembly/stability factors as well as post-translational protein modifications. Of the 39 genes examined, 33 are present in all *S. aureus* strains. Of the missing genes, all five are also all absent in *B. subtilis*. These included *rimF*, *rimK*, *ycaO*, *yibL*, and *yjgA*. In *E. coli*, RimK is an L-glutamate ligase that catalyzes post-translational addition of up to four C-terminal glutamate residues to 30S ribosomal subunit protein S6, and a mutation in *rimK* was found to confer neomycin and kanamycin resistance (nek).

RNA Processing/Ribonucleases. The various RNA components of the bacterial translation machinery are synthesized as longer precursor molecules that require subsequent processing steps, sizing, and 5′ or 3′ end trimming by a combination of endonucleases and exonucleases. These ribonucleases also play an important role in controlling the activity and quality of the translation machinery and the regulation of gene expression by RNA turnover. RNases generally harbor broad, sometimes overlapping specificity with other RNases, making difficult to determine their intrinsic essentiality. Also, at variance with the six other categories of proteins analyzed above, the set of RNases in gram-negative and gram-positive bacteria are quite different; some RNases are essential in one organism but not in the other. Of the 27 genes coding for RNases and related proteins we analyzed, 18 were found in all of the strains analyzed, including 2 genes unique to *E. coli* (absent from *S. aureus*) (RNB and RNT).

Seven genes were absent from all *S. aureus* strains examined: *bsn*, *nrnB*, *orn*, *rna*, *rnd*, and *rng*; of these, only *rnhA*, *bsn*, and *nrna* are found in *B. subtilis*, and the others are unique to *E. coli*. Bsn (YurI in *B. subtilis*) is an RNase of gram-positive bacteria and hydrolyzes RNA nonspecifically into oligonucleotides with 5′-phosphate therefore likely playing a role in nutrient cycling.

Whereas *E. coli* and other gram-negative bacteria possess only one essential oligoribonuclease (nano-RNase, Orn) for degrading oligoribonucleotides of two to five residues sin length, *B. subtilis* possesses two nonorthologous nano-RNases with redundant specificity: NrnA (Ytql) and NrnB (YngD). All *S. aureus* strains lack the *nrnA* gene. RnhA is one of the multivariant Ribonucleases H (HI = RnHA, HII = RnHB, and HIII = RnHC) that cleave RNA of RNA–DNA hybrids. Their primary function is to prevent aberrant DNA replication at sites other than oriC. All *S. aureus* strains examined lost RnHA (HI), but the function is redundant with the other two genes present in all strains.

Furthermore, one gene was found to be unevenly distributed in *S. aureus* strains: *rnE*. This gene is present in *E. coli* but not *B. subtilis* and was found in 43 of the 53 genomes analyzed. It encodes an endonuclease that cleaves single-stranded regions of pre-RNA transcripts, with a similar function to the conserved RnjA and RnjB, two enzymes found only in gram-positive bacteria.

Transcription. Transcription is the first step of gene expression, in which a particular segment of DNA is copied into RNA by the enzyme RNA polymerase. Of the 29 genes involved in transcription, 17 were conserved in all *S. aureus* strains, including *rpoA-D* and *rpoS*, *nusABG*, *greAB*, *rho* and *sigBEFH*, *fliA*, and *rhiB*. Of the 12 genes missing, 7 are also absent from *B. subtilis*, including *rpoE*, *H*, and *Z* as well as *fecL*, *dmsD*, *mfd*, and *torD*. *rpoN* was present in both *E. coli* and *B. subtilis* but lacking in all *S. aureus* strains examined. Finally, *sigVWXZ* were missing in all *S. aureus* strains, despite being present in *B. subtilis*, indicating different transcriptional control strategies.

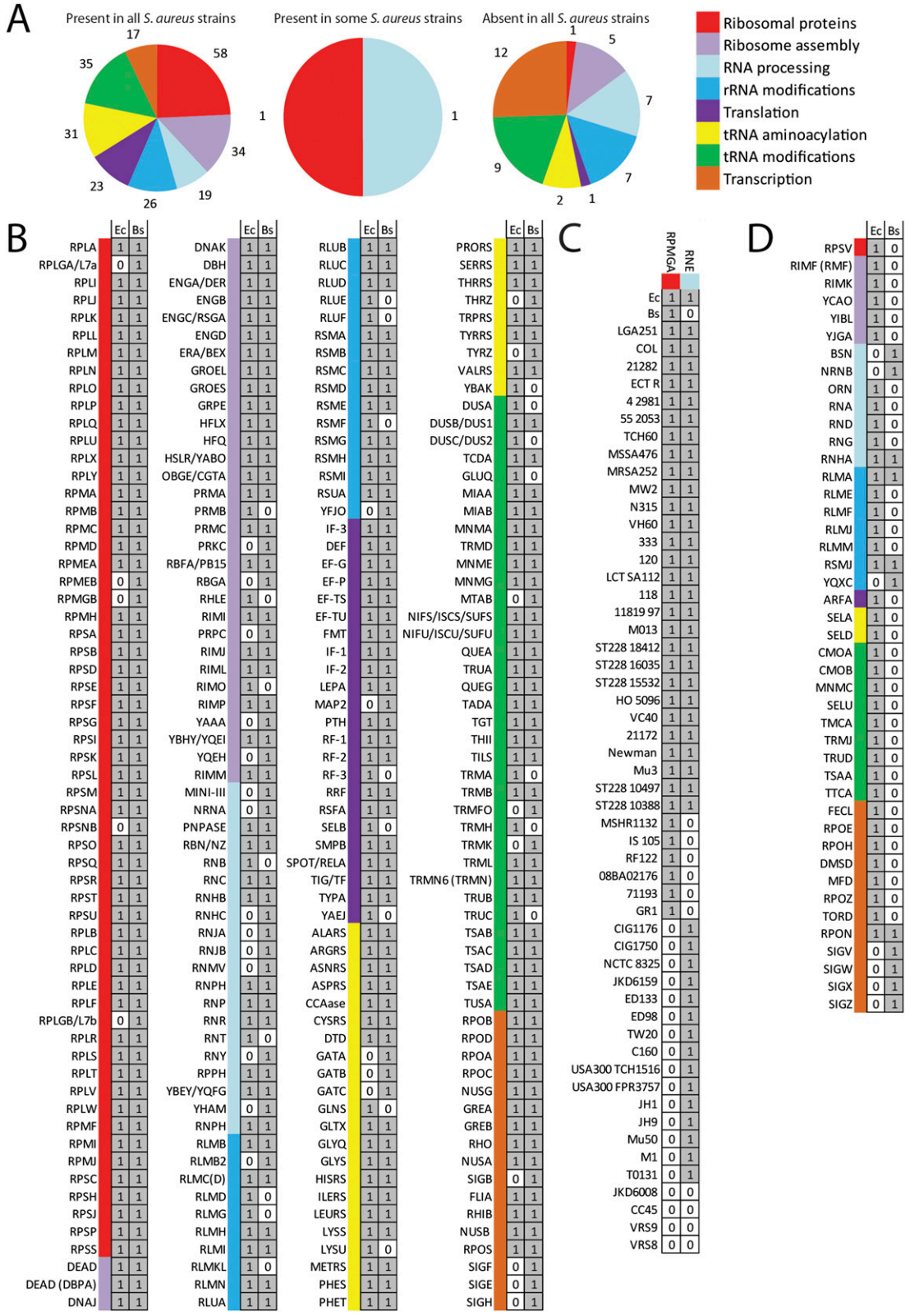


Fig. S1. Conservation of translation machinery in *S. aureus* strains. Genes involved in transcription and translation were selected from *E. coli* (Ec) and *B. subtilis* (Bs) to search for the presence of homologous proteins in 64 *S. aureus* strains. (*A*) The total number of genes in each category. The results were grouped into three panels: (*B*) conserved core genes involved in transcription and translation, (*C*) genes lost in some strains only (strains aligned vertically and genes horizontally), and (*D*) genes absent in all strains. The query gene acronyms correspond to gene names given in Dataset S1 and are ordered from top to bottom, according to the eight protein categories: ribosomal proteins, tRNA aminoacylation, rRNA modifications, tRNA modifications, ribosome assembly, transcription, translation, and RNA processing according to coding next to *A*.

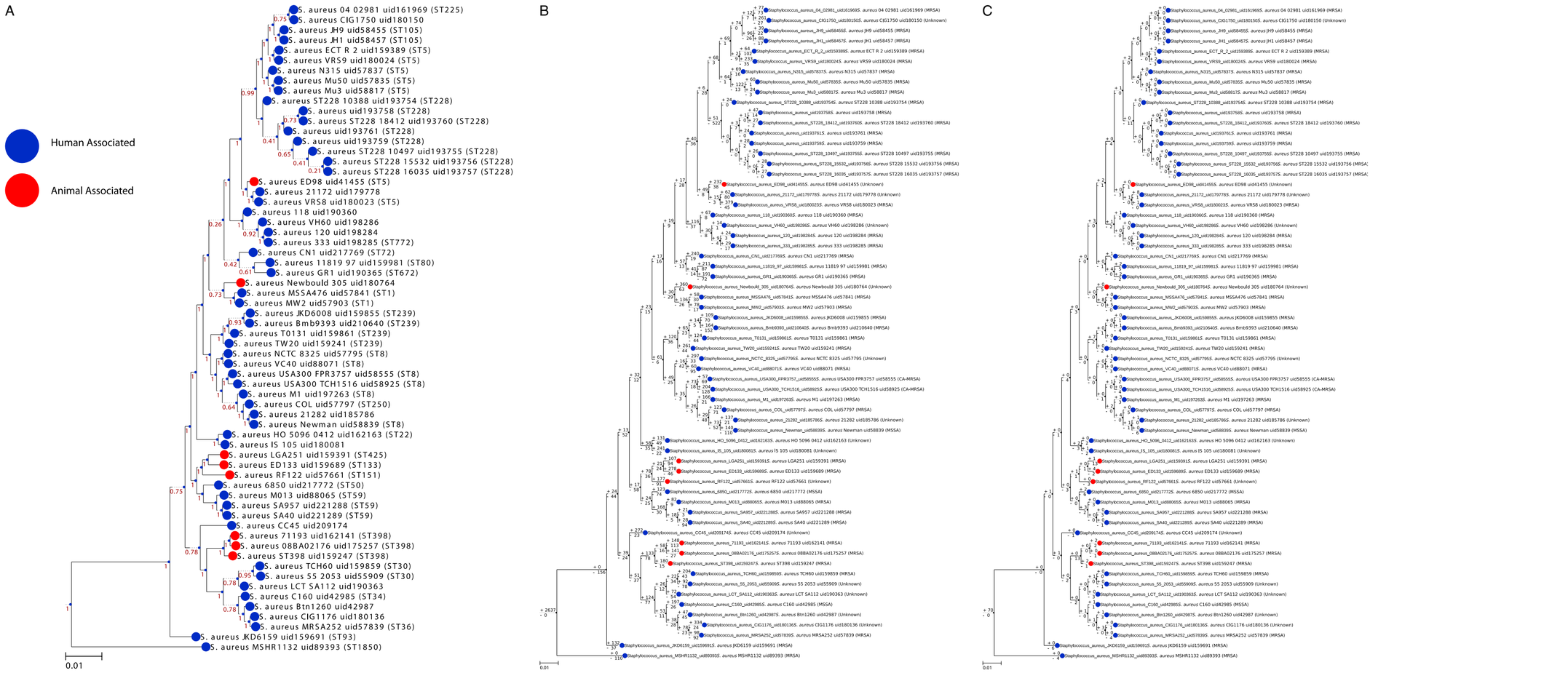
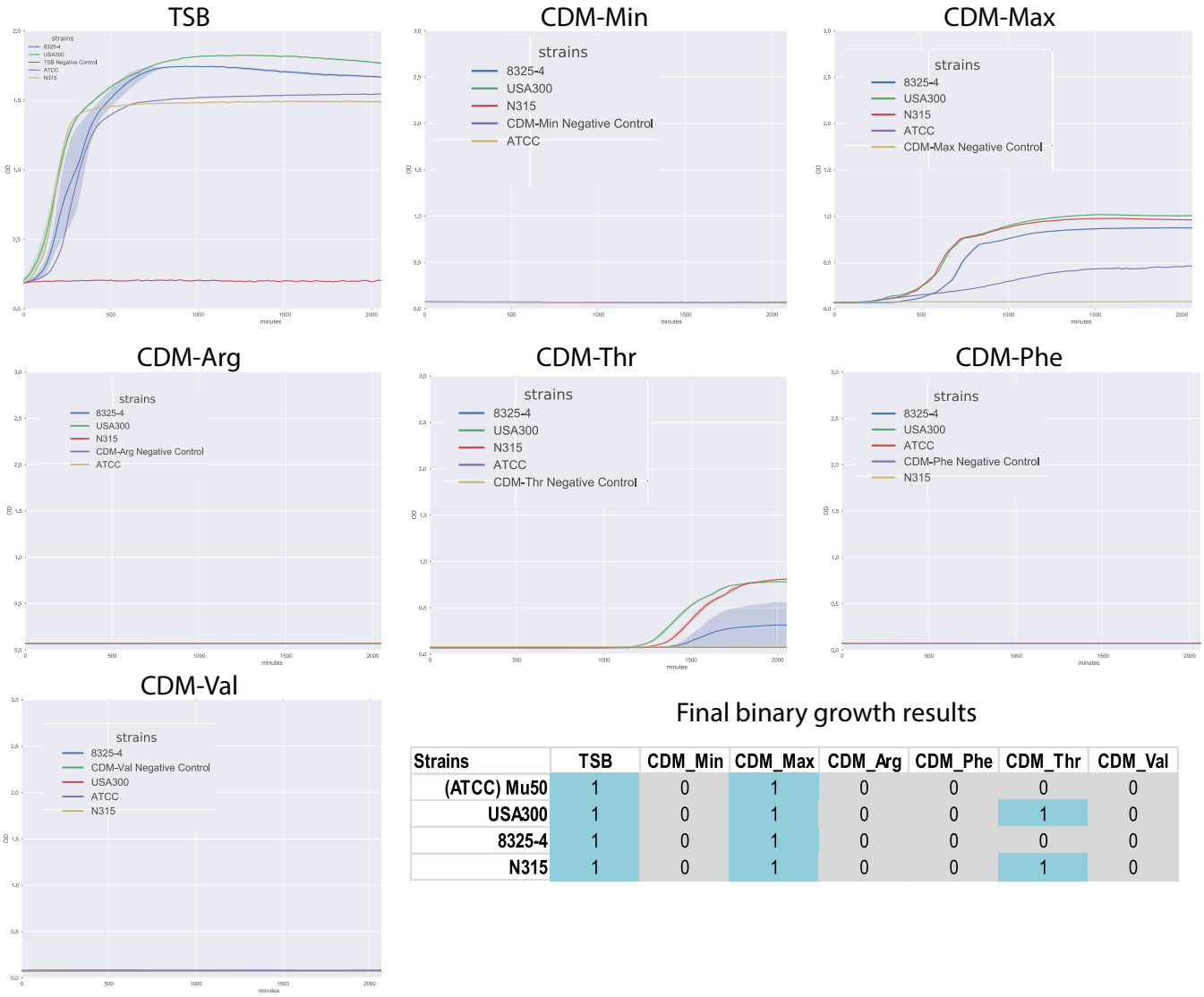


Fig. S2. Reconstruction of gene gain and loss for the selected 64 strains of *S. aureus*. (*A*) A phylogeny for the *S. aureus* species was constructed based on the 1,441 identified core genes. All core genes were aligned to reconstruct the phylogeny of the *S. aureus* species based on its core genome. Gene gain/loss was calculated based on extrapolation of a last common ancestor for (*B*) all genes and (*C*) the virulence factors specifically. Full-size images are available in Dataset S2.



Final binary growth results

Strains	TSB	CDM_Min	CDM_Max	CDM_Arg	CDM_Phe	CDM_Thr	CDM_Val
(ATCC) Mu50	1	0	1	0	0	0	0
USA300	1	0	1	0	0	1	0
8325-4	1	0	1	0	0	0	0
N315	1	0	1	0	0	1	0

Fig. S3. Experimental growth screens on chemically defined media. Four *S. aureus* strains were grown for 24 h in different media compositions. TSB media is a standard chemically undefined media for *S. aureus* growth. A minimal media (CDM-Min) was defined based on M9-glucose media with addition of vitamins and proline, serine, and leucine. This media did not support growth of any *S. aureus* strains. Additional amino acids were added to the CDM-Min media to test growth capabilities including arginine, threonine, phenylalanine, and valine. A CDM-max media consisted of all amino acids except tyrosine and cysteine.

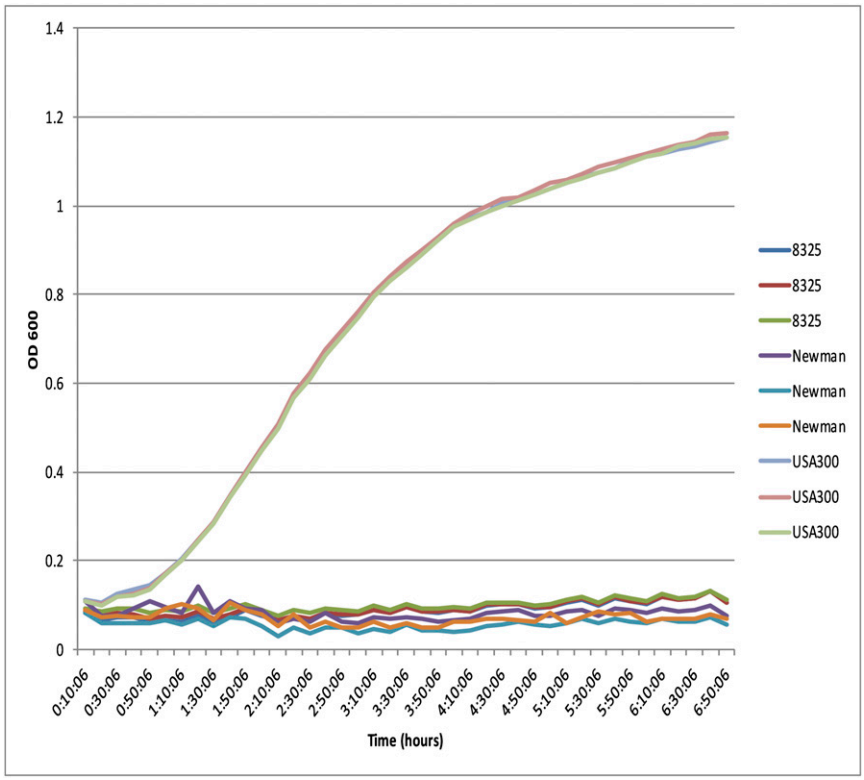


Fig. S4. Experimental confirmation of USA300's ability to grow in the presence of spermidine. *S. aureus strains* USA300, Newman, and ATCC8325 were grown in TSB media supplemented with 6 mM spermidine. Only USA300 strains were able to survive. Our metabolic models indicate that this is due to the presence of the *speG* gene encoding spermidine acetyltransferase (see main text).

Table S1. Biomass composition

Metabolite	Coefficient
1,2-Daicylglycerol (Saureus)	0.0042
10-Formyltetrahydrofolate	0.000223
1_2-Dianteisopentadecanoyl-sn-glycerol_3-phosphate	0.01
5,10-Methylenetetrahydrofolate	0.000223
5,6,7,8-Tetrahydrofolate	0.000223
AMP	3.11
ATP	54.124831
Acetyl-CoA	0.0000462
Ammonium	0.013013
CMP	2.2
CTP	0.133508
Calcium	0.005205
Cardiolipin (Saureus)	0.00251
Chloride	0.005205
Co2+	0.000025
CoA	0.000576
Cu2+	0.000709
D-Glucosamine 1-phosphate	0.23
D-Glucose	0.06685
Diglucosyl-diacylglycerol (SA) 2	0.0012068
Diphosphate	0.139258
Fe2+	0.006715
Fe3+	0.007808
Flavin adenine dinucleotide oxidized	0.000223
GTP	0.05952
Glucosyl Phosphoglycerol (SA)	0.053
Glycine	0.612638
H_2O	48.601527
Heme O	0.01
L-Alanine	0.513689
L-Arginine	0.295792
L-Asparagine	0.241055
L-Aspartate	0.241055
L-Cysteine	0.09158
L-Glutamate	0.26316
L-Glutamine	0.26316
L-Histidine	0.094738
L-Isoleucine	0.290529
L-Leucine	0.450531
L-Lysine	0.343161
L-Methionine	0.153686
L-Phenylalanine	0.185265
L-Proline	0.221055
L-Serine	0.215792
L-Threonine	0.253687
L-Tryptophan	0.056843
L-Tyrosine	0.137896
L-Valine	0.423162
LL-2,6-Diaminoheptanedioate	1.27
Mn2+	0.000691
Molybdate	0.000007
N-Acetyl-D-glucosamine	0.38
N-Acetyl-D-glucosamine 1-phosphate	1.22
Nicotinamide adenine dinucleotide	0.001831
Nicotinamide adenine dinucleotide - reduced	0.0000462
Nicotinamide adenine dinucleotide phosphate	0.000447
Nicotinamide adenine dinucleotide phosphate - reduced	0.0003738
Phosphatidylalanine_SA2	0.01
Phosphatidylethanolamine (Saureus)	0.0089
Phosphatidylglycerol (Saureus)	0.0231
Phosphatidylglycerophosphate (Saureus)	0.01
Phosphatidylglycine_SA2	0.01

Table S1. Cont.

Metabolite	Coefficient
Phosphatidylleucine_SA2	0.01
Phosphatidyllysine_SA2	0.015
Putrescine	0.01
SA free fatty acids	0.01
Siroheme	0.000223
Sodium	0.27853
Succinyl-CoA	0.0000462
Sulfate	0.004338
Thiamin	0.000223
Thiamine diphosphate	0.000223
UDP-*N*-acetylmuramate	1.78
UMP	2.39
UTP	0.06072
Undecaprenyl diphosphate	0.000055
Undecaprenyl-diphospho-*N*-acetylmuramoyl-(*N*-acetylglucosamine)-L-ala-D-glu-meso-2, 6-diaminopimeloyl-D-ala-D-ala	0.000055
Zinc	0.000341
dAMP	0.76
dATP	0.02028
dCMP	0.53
dCTP	0.0099
dGMP	0.6
dGTP	0.0099
dTMP	0.73
dTTP	0.02028
Magnesium	0.008675
Minor teichoic acid (acetylgalactosamine glucose phosphate, $n = 30$)	0.0286
Nickel	0.000323
Potassium	0.195193

Table S2. Minimal media and strain-specific auxotrophies

Strain	trp	phe	cys	his	met	pro	4abz	Guanine	Riboflavin	Spermidine	tyr	arg	ile
6850	X	X	X	X	X	X	X	X	X	X			
Bmb9393	X	X	X	X	X	X	X	X	X	X			
Btn1260	X	X	X	X	X	X	X	X	X	X			
CN1	X	X	X	X	X	X	X	X	X	X			
Newbould-305	X	X	X	X	X	X	X	X	X	X			
SA40	X	X	X	X	X	X	X	X	X	X			
SA957		X	X	X	X	X		X	X	X			
SA-118	X	X	X	X	X	X	X				X		
SA-120	X	X	X	X	X	X	X				X		
SA-333		X	X	X	X	X					X		
55–2053	X	X	X	X	X	X	X				X		
71193	X	X	X	X	X	X	X					X	
C160	X	X	X	X	X	X	X				X		
GR1	X	X	X	X	X	X	X				X	X	
LCT	X	X	X	X	X	X	X				X	X	
M1	X	X	X	X	X	X	X						
Mu3	X	X	X	X	X	X	X						
Newman	X	X					X						
TCH60	X	X	X	X	X	X	X						
VH60	X	X	X	X	X	X	X				X		
ST228-16125	X							X	X				X
ST228-18341	X							X	X				X
ST228-18583	X							X	X				X
ST398	X	X	X	X	X	X	X	X	X	X			

X indicates that this strain has a model-predicted auxotrophy for the given compound.

Dataset S1. Model information in XLSX format

Dataset S1

Worksheet 1 shows strain-specific reconstruction information. A total of 64 strain-specific reconstructions were created for *S. aureus* strains. This sheet contains their names, taxonomic identifications, clonal complexes, NCBI genome identifications, and other factors such as antibiotic sensitivity, lifestyle, and whether they are animal or human associated. Worksheet 2 shows pan-*S. aureus* reactions. All reaction information associated with the 1,519 reactions in the panreactome plus the 234 exchange reactions required for substrates to enter or leave the cell is shown, including reaction abbreviation, name, formula, and EC numbers. Reactions are assigned to metabolic systems and subsystems according to Orth et al. (96). Reactions have also been cross referenced to major biochemical databases such as ModelSEED (63), KEGG (65, 66), and BIOCYC (67). Worksheet 3 shows pan-*E. coli* metabolites. All metabolite information associated with the 1,521 metabolites in the panreactome is shown, including metabolite abbreviation, name, compartment, charge, formula, inchi and smile string, molecular weight, and CAS number. Metabolites have also been cross referenced to major biochemical databases such as ModelSEED (63), KEGG (65, 66), and BIOCYC (67). Worksheet 4 shows each strain's model-specific GPR association for reactions present in the panreactome. Blank cells indicate that the model of the strain does not encode an enzyme catalyzing the reaction. Worksheet 5 shows in silico formulation of the biomass composition of *S. aureus*. These compounds represent those that must be produced in order for *S. aureus* to grow and proliferate. Units are in mmol/gDW. Worksheet 6 shows in silico M9 minimal media formulation and strain-specific auxotrophies. In silico formulation of M9 minimal media used to perform the growth screens in each different condition is shown. This dataset also includes the specific compound that was removed to test each different carbon, nitrogen, phosphorous, and sulfur source. It also includes the specific exchange reactions that were opened in models of auxotrophic strains to allow the models to support growth on M9 minimal media. Worksheet 7 shows in silico growth rates on LB and M9 supplemented media. Results of growth simulations are shown for each of the 64 strains in LB and M9 minimal media. The M9 media is supplemented with all compounds the model is auxotrophic for (see worksheet 5). In silico predicted growth rate is presented in units of hr^{-1}. Worksheet 8 shows in silico growth screens. Results of growth simulations are shown for each of the 64 strains on 302 different growth-supporting carbon, nitrogen, sulfur, and phosphorous sources. All simulations were performed in both aerobic and anaerobic conditions. Each condition is named with the element it is testing as well as whether the simulation was conducted in aerobic or anaerobic conditions. In silico predicted growth rate is presented in units of hr^{-1}. Worksheet 9 shows singly essential reactions for each model growth on LB media. Each model was grown in simulated LB media, and the predicted growth rate for systematic knockout of all reactions was performed. The results are presented in this worksheet.

Dataset S2. Zip file of all 64 models in SBML format and PDFs of Fig. S2

Dataset S2

Models are also available publicly at the BIGG database (70): bigg.ucsd.edu.