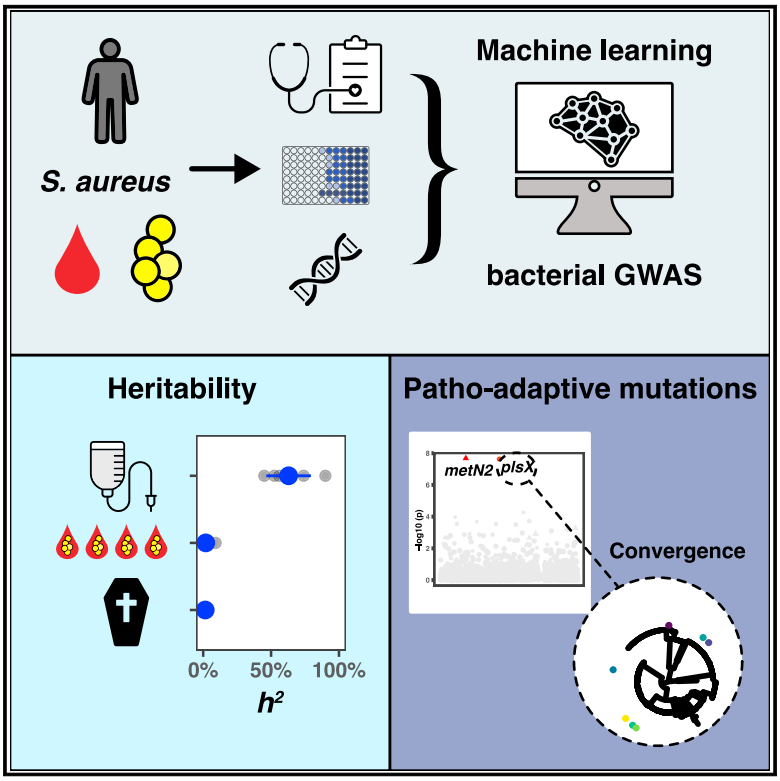
Cell Reports

Article

A statistical genomics framework to trace bacterial genomic predictors of clinical outcomes in *Staphylococcus aureus* bacteremia

Graphical abstract



Authors

Stefano G. Giulieri, Romain Guérillot, Natasha E. Holmes, ..., Steven Y.C. Tong, Timothy P. Stinear, Benjamin P. Howden

Correspondence

stefano.giulieri@unimelb.edu.au

In brief

Giulieri et al. performed a large-scale bacterial GWAS of clinical outcomes of *S. aureus* bacteremia to identify associations between bacterial adaptive mutations and disease severity. While outcomes such as infection mortality had low heritability, their GWAS framework identified new pathoadaptive mutations associated with vancomycin resistance and duration of bacteremia.

Highlights

- Low heritability of clinical outcomes in a large-scale bacterial GWAS of *S. aureus* sepsis
- Vancomycin resistance had high heritability and contributed to infection severity
- Novel pathoadaptive mutations linked to vancomycin resistance and duration of bacteremia

Giulieri et al., 2023, Cell Reports *42*, 113069
September 26, 2023
https://doi.org/10.1016/j.celrep.2023.113069

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

A statistical genomics framework to trace bacterial genomic predictors of clinical outcomes in *Staphylococcus aureus* bacteremia

Stefano G. Giulieri,[1,2,3,4,11,*] Romain Guérillot,[1] Natasha E. Holmes,[3,4] Sarah L. Baines,[1,5] Abderrahman Hachani,[1] Ashleigh S. Hayes,[1] Diane S. Daniel,[1] Torsten Seemann,[5] Joshua S. Davis,[6,7] Sebastiaan Van Hal,[8,9] Steven Y.C. Tong,[2,3] Timothy P. Stinear,[1,10] and Benjamin P. Howden[1,4,5,10]

[1]Department of Microbiology and Immunology, The University of Melbourne at the Doherty Institute for Infection and Immunity, Melbourne, VIC 3000, Australia
[2]Victorian Infectious Disease Service, The Royal Melbourne Hospital at the Peter Doherty Institute for Infection and Immunity, Melbourne, VIC 3000, Australia
[3]Department of Infectious Diseases, The University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, VIC 3000, Australia
[4]Department of Infectious Diseases, Austin Health, Heidelberg, VIC 3084, Australia
[5]Centre for Pathogen Genomics, The University of Melbourne, Melbourne, VIC 3000, Australia
[6]Department of Infectious Diseases, John Hunter Hospital, New Lambton Heights, NSW 2305, Australia
[7]Menzies School of Health Research, Charles Darwin University, Casuarina, NT 0810, Australia
[8]Department of Infectious Diseases and Microbiology, Royal Prince Alfred Hospital, Camperdown, NSW 2050, Australia
[9]Central Clinical School, University of Sydney, Camperdown, NSW 2050, Australia
[10]Senior author
[11]Lead contact
*Correspondence: stefano.giulieri@unimelb.edu.au
https://doi.org/10.1016/j.celrep.2023.113069

SUMMARY

Outcomes of severe bacterial infections are determined by the interplay between host, pathogen, and treatments. While human genomics has provided insights into host factors impacting *Staphylococcus aureus* infections, comparatively little is known about *S. aureus* genotypes and disease severity. Building on the hypothesis that bacterial pathoadaptation is a key outcome driver, we developed a genome-wide association study (GWAS) framework to identify adaptive mutations associated with treatment failure and mortality in *S. aureus* bacteremia (1,358 episodes). Our research highlights the potential of vancomycin-selected mutations and vancomycin minimum inhibitory concentration (MIC) as key explanatory variables to predict infection severity. The contribution of bacterial variation was much lower for clinical outcomes (heritability <5%); however, GWASs allowed us to identify additional, MIC-independent candidate pathogenesis loci. Using supervised machine learning, we were able to quantify the predictive potential of these adaptive signatures. Our statistical genomics framework provides a powerful means to capture adaptive mutations impacting severe bacterial infections.

INTRODUCTION

The course and outcome of severe bacterial infections such as bacteremia and sepsis are determined by the interaction of factors related to the host (e.g., genetics, age, chronic conditions), the pathogen, and the treatment.[1] A deeper understanding of these outcome determinants would help design more effective treatments and inform precision medicine strategies. However, in contrast to progress in the immunogenetics of infections,[2] it has been challenging to identify bacterial genetic factors influencing infection outcomes.[3,4] This is due to the considerable genetic diversity of bacterial pathogens and the strong clonal structure of bacterial populations that complicate genome-wide association studies (GWASs).[5] In addition, many bacterial GWASs addressing clinical outcomes face specific issues like small sample size, retrospective data collection, and unbalanced datasets.[5]

With up to 20% mortality[6,7] and increasing incidence,[8] *Staphylococcus aureus* bloodstream infections are a key challenge to human health.[9] The clinical determinants of *Staphylococcus aureus* bacteremia (SAB) outcomes have been described extensively; in particular, mortality has been the subject of numerous studies (reviewed in van Hal et al.[10]). Small studies have shown the impact of host genetic factors on SAB risk and outcomes; for example, it was shown that a polymorphism in the DNA methyltransferase *DNMT3A* protects from persistent SAB.[11] Studies linking specific clones or bacterial characteristics (vancomycin minimum inhibitory concentration [MIC]) with SAB mortality[12]

provide indirect evidence for the role of bacterial genetic factors in SAB mortality; however, it has been difficult to identify compelling associations between distinctive genes or mutations and clinical outcomes.[13] In a landmark study Recker et al. used a combination of GWASs and machine learning to explore factors associated with SAB mortality. Their work showed that both bacterial genetic and phenotypic information improved the performance of the predictive model. Feature selection from the model indicated several potential pathogenic mutations that could represent the basis for functional validation.[14]

A key concept underlying bacterial phenotype-genotype studies is pathoadaptation, where bacteria acquire mutations that enhance their ability to survive in the host. Upon infection, colonizing *S. aureus* strains are repeatedly subjected to similar selective pressures; thus, mutations promoting survival in the host during infection are expected to be preferentially maintained and become evident when examining *S. aureus* microevolution. The most compelling evidence for the role of pathoadaptation in SAB comes from within-host evolution studies, which have shown enrichment of mutations in a small number of genes including *agr* and antibiotic resistance genes (indicating convergent evolution), both upon transition from colonization to invasion and during persistent infection.[15–17] Bacterial GWASs leverage the concept of adaptation either directly by focusing on homoplastic mutations (phylogeny-based methods)[18–22] or indirectly by correcting the analysis for population structure.[23–26] For example, Farhat et al. identified multiple resistance-associated mutations and genes by applying a phylogenetic convergence testing approach to a collection of *M. tuberculosis* isolates.[18] Convergent evolution was also used in GWASs of *M. tuberculosis* tolerance[27] and vancomycin-intermediate *S. aureus* (VISA) phenotype.[28] Approaches based on population structure (allele-counting methods) do not account for convergent evolution explicitly, but the most robust associations are expected to occur with homoplastic mutations or genes under positive selection. For example, in a GWAS of drug resistance in *M. tuberculosis*, 50% of gene targets identified using regression and correction for population structure were shown to be homoplastic using a convergence-based approach.[29]

The examples described above involve antibiotic resistance, an adaptive phenotype with high heritability and strong selective pressure, where convergent evolution is expected to be apparent. However, it is unclear whether clinical outcomes of bacterial infections are dependent on bacterial genotypes, as shown by bacterial GWASs of clinical outcomes, which have been either negative or have not been reproduced across independent cohorts. Bacterial adaptation can be frequently detected in clinical infections,[30] leading to antibiotic resistance, antibiotic tolerance, or immune evasion and thus potentially driving treatment failure.[31] Here, we propose a statistical genomics framework that incorporates convergent evolution explicitly and tracks molecular signatures of adaptation across three independent SAB patient cohorts to build a model of bacterial genotype-phenotype associations for treatment outcomes. Given the complexity of identifying genetic correlates of clinical outcomes, we progress our analysis from clinically relevant bacterial phenotypes with high expected role of adaptation (vancomycin MIC) to intermediate outcomes (duration of bacteremia) to final endpoints with strong multi-factorial pathogenesis (infection mortality).

RESULTS

S. aureus bacteremia cohorts

To identify bacterial genetic correlates of clinical outcomes of bacteremia, we analyzed 1,358 *S. aureus* sequences from 3 cohorts of SAB with extensive clinical metadata (Table S5). The cohorts included an observational study of vancomycin efficacy in SAB (cohort A, 843 isolates, 738 individual episodes),[12,16,32] an observational study of the VISA/heterogeneous VISA phenotype in sequence type (ST) 239 methicillin-resistant *S. aureus* (MRSA) (cohort B, 296 isolates),[33] and a randomized clinical trial of vancomycin-β-lactam combination therapy in MRSA bacteremia (cohort C, 324 isolates).[34] The microbiological and clinical characteristics of the isolates are described in Table 1 and Figure 1. The cohorts were different in terms of proportion of MRSA, distribution of STs, mortality outcomes, and treatment choice. Therefore, in addition to combining all sequences in a single dataset, we analyzed the three collections separately, with the aim of finding overlapping hits between the independent samples that would provide robust evidence for clinically relevant bacterial adaptation (Figure S1). When multiple isolates per episode were available (cohort A), we used all isolates available in the vancomycin GWAS but only baseline isolates in the GWAS of clinical outcomes.

Lineage effects and bacterial adaptation may impact clinical outcomes of *S. aureus* bacteremia

We first assessed the relative impact of clinical factors on SAB outcomes using cohort A because it included all three outcomes of interest for this study (vancomycin MIC, duration of bacteremia, mortality) (543 episodes) and then fitted a classification model of mortality. In our initial model. we included predictors of SAB mortality that are generally available in the clinic including host factors (age, gender, Charlson comorbidity index, site of acquisition), antistaphylococcal treatment, phenotypic bacterial characteristics (methicillin resistance, vancomycin MIC), and duration of bacteremia.

We used a random forest machine-learning algorithm that was trained using stratified cross-validation on 80% of cohort A episodes. To increase the accuracy of the predictive parameters' estimates, the model was trained on 100 replicate training-testing data splits and produced a mean area under the receiver operating curve (AUROC) of 0.803 and a mean area under the precision recall curve (AUPRC) of 0.395 (Figures 2A and 2B). Thus, traditional predictors of mortality in SAB had an acceptable predictive performance[35]; however, 20% of mortality predictions were incorrect, raising the possibility that the addition of non-clinical variables (including bacterial genomic features) could increase prediction accuracy.

To estimate the relative role of these predictors of mortality, we calculate variable importance (impurity index) based on 100 replicate analyses. While age and comorbidities were the most important features in the final model (as expected), we were surprised that vancomycin MIC was as important as comorbidities

Table 1. Characteristics of *S. aureus* isolates included in the bacteremia cohorts

	Cohort A (n = 738)	Cohort B (n = 296)	Cohort C (n = 324)
Sample period			
1990–1999	0 (0)	32 (11)	0 (0)
2000–2009	517 (70)	264 (89)	0 (0)
2010–2019	221 (30)	0 (0)	324 (100)
Country			
Australia/NZ	738 (100)	296 (100)	239 (74)
Singapore	0 (0)	0 (0)	53 (16)
Israel	0 (0)	0 (0)	32 (10)
Patient age (years), median (IQR)	61 (46–74)	64 (50–73)	64 (49–76)
MRSA	252 (34)	296 (100)	324 (100)
Charlson comorbidity index, median (IQR)	2 (0–3)	2 (1–3)	5 (2–7)
Vancomycin MIC (mg/L), median (IQR)	1.5 (1–2)	1 (0.75–1.5)	N/A
Main antibiotic			
Flucloxacillin	384 (52)	0 (0)	0 (0)
Vancomycin	301 (41)	252 (85)	158 (49)
Daptomycin	0 (0)	0 (0)	3 (1)
Vancomycin + β-lactam	0 (0)	0 (0)	159 (49)
Daptomycin + β-lactam	0 (0)	0 (0)	4 (1)
Other	53 (7)	11 (4)	0 (0)
No treatment	0 (0)	28 (9)	0 (0)
Duration of bacteremia (days), median (IQR)	1 (1–2)	N/A	1 (1–3)
Mortality at 30 days	109 (15)	88 (30)	35 (11)

Values are counts (percentage) unless otherwise specified. MRSA, methicillin-resistant *S. aureus*; N/A, not available.

and more important than methicillin resistance (Figure 2C). This ranking of variables was also confirmed when fitting a logistic regression model (Table S1). This highlights the role of bacterial factors in SAB mortality and suggests that partially adaptive antibiotic resistance (such as increased vancomycin MIC)[36] is at least as important as acquired antibiotic resistance genes (Figure S2). This finding further supported our choice of including vancomycin MIC as an intermediate bacterial phenotype in the bacterial GWAS.

To explore the relative role of bacterial genetic factors in SAB outcomes, we first considered whether there were "high-risk" clones or lineages that were associated with either of the three GWAS outcomes. To assess lineage effects in cohort A, we first performed a multi-dimensional scaling analysis (MDS) of a distance matrix generated by Mash, an approach that compares genetic similarity based on a subsample of 1,000 k-mers.[37] This allowed us to include the whole genome content rather than just the core genome SNP matrix. We then extracted the top 10 components (85% of genetic variance explained) and assessed their correlation both with mortality and bacterial covariates. This correlation analysis showed robust lineage effects for vancomycin MIC, in particular MDS2 and MDS3 ($p < 10^{-15}$; Figure 2D; Table S3), while a weak lineage effect was found for mortality (MDS2 and MDS4, $p < 0.01$). As expected, there was a strong correlation between MDS axes and clonal complexes (CCs) (Figures 1B, S3, and S4). For example, the most prevalent CC (CC239) was positively correlated with MDS2 and negatively correlated with MDS1, while CC93 was strongly correlated with MDS3.

We then calculated the narrow-sense heritability (h^2) of clinical outcomes and bacterial phenotypes to determine what proportion of these parameters variation was a result of genetic variation. With both methods used (linear mixed model and whole-genome elastic net model; Table S10), vancomycin MIC had a high heritability, as expected with a microbiological phenotype that is genetically encoded,[38] while duration of bacteremia and mortality had substantially lower values (Figure 2E; Table S2). These values are consistent with bacterial GWASs of *S. aureus* infection vs. colonization (2%)[39] and *E. coli* bacteremia mortality (<1%).[4]

Identification of loci associated with increased vancomycin MIC

We first analyzed loci associated with vancomycin MIC (as a continuous variable) in two cohorts. Vancomycin MIC was chosen due to its clinical relevance both in our model and in the literature[40] and its higher bacterial heritability. In addition, we used E-test data that have a larger dynamic range and allowed us to capture subtle adaptive changes,[41] further enhancing the power to detect significant associations using a linear regression model. The MIC distribution was slightly shifted between the two cohorts, with a higher median MIC in cohort A than in cohort B (1.5 vs. 1 mg/L, $p = 2.7 \times 10^{-13}$, Wilcoxon test) (Table 1; Figures 1C and S6), possibly due to the different clonal structure

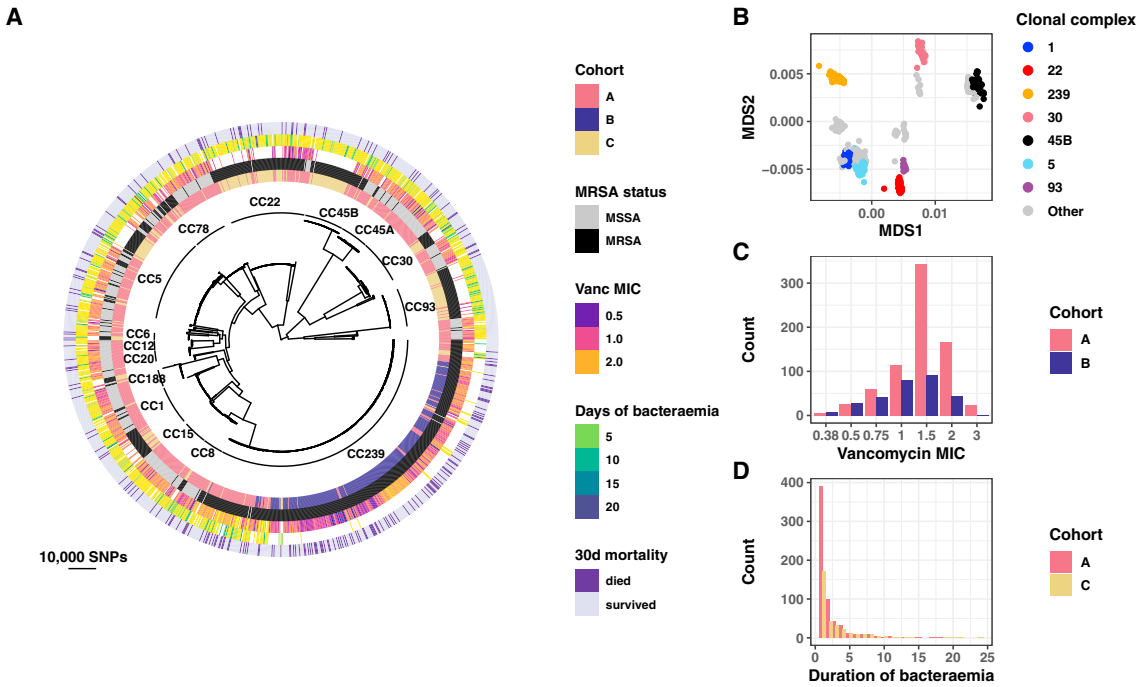


Figure 1. Genomic and phenotypic characterization of 3 *S. aureus* bacteremia cohorts

(A) Maximum likelihood phylogenetic trees of 1,358 bloodstream *S. aureus* strains included in the analysis. The tree is annotated with (from the inside to the outside): clonal complex (CC), cohort, MRSA status, vancomycin MIC, duration of bacteraemia, mortality at 30 days.

(B) First two principal components of multi-dimensional scaling (MDS) of the Mash distance matrix. Dots represent sequences and are colored according to the 7 most prevalent clonal complexes.

(C) Distribution of vancomycin MIC (in mg/L).

(D) Distribution of duration of bacteremia (in days).

of the two populations (Figure 1A); however, the shape of the MIC distribution was similar, suggesting that it would be possible to compare the results from the analyses performed in both cohorts separately.

Consistent with the heritability estimate, we identified several mutations that were significantly associated with vancomycin MIC; however, the analysis was strongly biased by the population structure, as shown by the quantile-quantile plot (Figure S5). While this indicates inflation (despite correction for population structure in the linear mixed model) and confirms the critical role of lineage effects in antibiotic resistance in *S. aureus*, close inspection of variants reaching genome-wide significance showed mutations in *mprF* (known to be associated with daptomycin and vancomycin resistance in *S. aureus*[42–44]) and *pbp2* (thought to be associated with β-lactam resistance and vancomycin resistance[42,45]), confirming the ability of this approach to identify resistance-associated loci (Table 2). In addition, we found an intergenic mutation 345 bp upstream of *mprF*, further supporting the importance of *mprF* in our datasets.

We also considered significant mutations outside known vancomycin resistance loci to explore whether our GWASs had the ability to identify new targets of vancomycin resistance. The most significant association in cohort A was found with the mutation V141I in the iron-sulfur cluster synthesis gene *sufU*[47] and with an intergenic mutation at position 1,293,856 of the reference genome *S. aureus* BPH2947 that was located within a promoter upstream of the predicted operon that contains the serine-threonine phosphatase *stp1*, which has been previously associated with vancomycin resistance.[46,48] Only one locus reached genome-wide significance in cohort B: an intergenic mutation at position 2,959,009 within a predicted promoter upstream of the aureolysin gene *aur*.

Inflation of p values and incomplete correction for population structure indicated that the GWAS at variant level was not sufficiently accurate to identify new targets. Moreover, it is known from within-host adaptive studies that convergence at the mutational level is weak.[15] Therefore, we aggregated mutations at the gene level, an approach that has been used by others[22,24,49] and that may implicitly consider convergent evolution since different mutations are expected to arise independently across the phylogeny. Because the effect of non-synonymous substitutions could be either a gain or loss of function, we analyzed protein-modifying (all mutations excluding synonymous substitutions) and truncations separately and aggregated only rare mutations (allele fraction below 1%), which is consistent with gene-burden test practices.[50,51]

This analysis revealed 9 genes that reached genome-wide significance (Figure 3). Some of these genes (such as *walK* and *vraG* in cohort B) are known to be associated with vancomycin resistance[52,53] but were not identified in the mutations analysis, highlighting the potential of the gene-burden test to increase the power of GWASs. A strong association was found with the

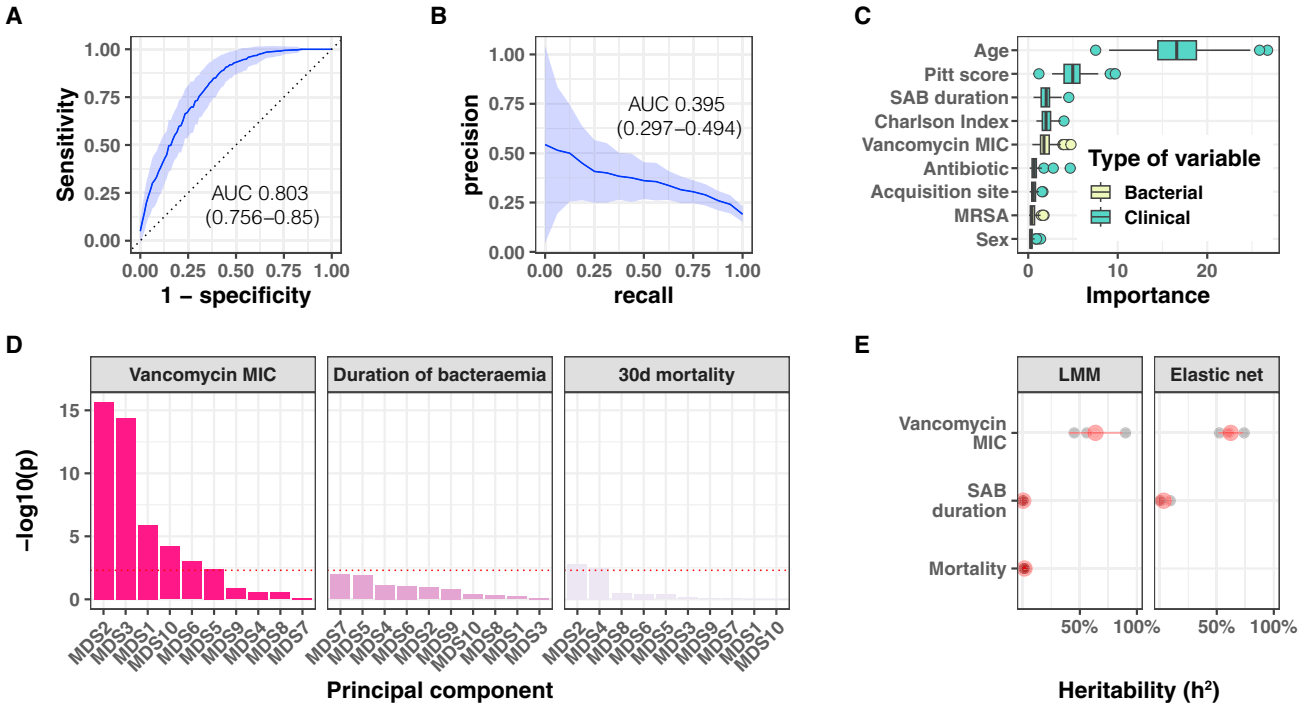


Figure 2. Impact of clinical characteristics, bacterial phenotypes, and bacterial genomes on outcomes of *S. aureus* bacteremia

(A and B) Performance of the random forest model of 30 day mortality when using clinical and microbiological phenotypic data only (values are mean ± standard deviation). The model was trained and tested using cohort A data. The performance metrics (mean area under the curve [AUC] ± standard deviation) in the test dataset were calculated from 100 replicate analyses.

(A) Receiver operator characteristic curve (ROC).

(B) Precision recall curve.

(C) Variable importance plot (Gini impurity index) based on 100 replicate analyses (boxes represent median, first, and third quartiles, whiskers are based on ×1.5 interquartile range).

(D) Lineage effects for the vancomycin MIC, duration of bacteremia, and the SAB mortality outcomes for cohort A. The bars show the significance of the association between the ten first principal components (based on multi-dimensional scaling) and the outcome.

(E) Narrow-sense heritability (SNP heritability) of the GWAS outcomes of interest: vancomycin MIC, duration of bacteremia (persistence), and SAB mortality, according to the linear mixed model (LMM) and the whole-genome elastic net model implemented in Pyseer. Red dots and lines represent mean and standard deviation across the 3 cohorts, and gray dots represent individual cohort measures.

SAUSA300_0956 locus, a GNAT-family acetyltransferase.[54,55] The function of the protein in *S. aureus* is not well understood; however, some indirect evidence may point to a role in cell-wall synthesis or turnover. First, it is located between the autolysin *atl* and the methicillin resistance factor *fmtA*.[56] Second, an *E. coli* homolog of SAUSA300_0956 was linked to membrane biosynthesis.[55] Another interesting association involved the SAUSA300_1220 locus. This gene (also known as *desR*) was recently found to be potentially associated with regulation of β-lactam resistance in MRSA.[57] The SAUSA300_2445 locus is part of the gluconate operon, which has been shown to be activated in response to salicylate challenge,[58] and in the VISA strain Mu50.[59] Two further genes were significantly associated with changes in vancomycin MIC: *thrD*, an aspartate kinase with no further characterization in *S. aureus*, and the ribosome silencing factor *rsfS*, which has been shown to be important for the stress response.[60]

We used the Nebraska Transposon Mutant Library[61] to validate 5 out of 9 genes (*vraG*, SAUSA300_1220 = *desR*, SAUSA300_2248, *rsfS*, *thrD*) that reached genome-wide significance in the gene-burden GWAS for vancomycin MIC. We confirmed that the *vraG* transposon mutant had a lower vancomycin MIC than the wild-type JE2, a finding that is consistent with the association of *vraG* gain-of-function substitutions with vancomycin resistance. Beside this known association, we found that the *thrD* mutant had a vancomycin MIC decrease comparable to the impact of the *vraG* inactivation (and consistent with the direction of the effect size in the gene-burden GWAS) and that inactivation of *rsfS* led to a vancomycin MIC increase (Table S4).

Gene-burden GWAS identifies novel potential genetic signatures of persistence in *S. aureus* bacteremia

We next assessed genetic correlates of duration of bacteremia in cohorts A and C. As expected, the distribution of this phenotype was skewed, with 59% of strains being associated with a duration of bacteremia of 1 day or less (Figure 1D). Using automated normalization optimization (based on Pearson P statistics), we

Table 2. Selection of loci significantly associated with vancomycin MIC

Reference position	Gene	Product	Mutation	Antibiotic resistance	p value (effect size)		
					Cohort A	Cohort B	Combined analysis
976,503	*sufU*	NifU-family SUF system FeS assembly protein	V141I	no	3.2×10^{-12} (0.886)	–	2.3×10^{-9} (0.870)
1,293,856	intergenic	upstream of *def2-fmt-sun-rlmN-stp1-pknB* operon	−31G>T	vancomycin[46]	–	–	4.76×10^{-14} (0.815)
2,959,009	intergenic	upstream of *aur* gene	−60C>T	no	–	1.12×10^{-4} (0.737)	–
1,458,535	intergenic	upstream of mprF	−322G>A	vancomycin,[42] daptomycin[44]	1.3×10^{-11} (0.837)	–	1.87×10^{-8} (0.818)
1,459,663	*mprF*	oxacillin resistance-related FmtC protein	E269D	vancomycin,[42] daptomycin[44]	1.04×10^{-7} (0.74)	–	–
1,460,765	*mprF*	oxacillin resistance-related FmtC protein	G637S	vancomycin,[42] daptomycin[44]	–	–	4.28×10^{-7} (−0.463)
1,583,057	*pbp2*	penicillin-binding protein 2	A246T	β-lactams, vancomycin[42,45]	–	–	4.28×10^{-7} (−0.463)

Loci are displayed if they have the lowest p value within the analysis or if they are located near or in known vancomycin resistance loci.

found that none of the most common available strategies could improve the distribution of the variable (Figure S7).

Just one mutation (non-synonymous substitution in the isomerase *galE*) barely achieved genome-wide significance, in accordance with a lower heritability and an expected polygenic signal associated with this complex trait (Table S7). However, we identified loci that were associated with duration of bacteremia on the truncation burden test. The strongest association in cohort A was found with protein truncations in *murQ*, an esterase that is part of the peptidoglycan recycling machinery.[62] Mapping of *murQ* truncations showed that they were independently acquired multiples times across the phylogeny, suggesting convergent evolution (Figure 4B).

We found two significant associations in cohort C. These involved truncations in *metN2*, a methionine transporter, and protein-modifying mutations in *plsX*. The latter gene is a phosphate acyltransferase and is part of the phospholipid synthesis pathway.[55] It interacts with the fatty-acid kinase *fakA*, which has been linked to virulence factor expression[63] and resistance to antimicrobial peptides.[64] Underscoring the relevance of these two cohort C signatures of bacteremia duration are the low p values ($\sim 2 \times 10^{-8}$) and the findings of the whole-genome elastic net model, where the heritability of duration of bacteremia was higher than expected (20%) and mutations in *metN2* and *plsX* were selected in the final model.

Bacterial genomic predictors do not improve mortality prediction of *S. aureus* bacteremia

We assessed whether there were any mutations or loci associated with SAB mortality in the three cohorts and the combined dataset (Table S8). Only in cohort C could we detect a genome-wide significant association with rare mutations in four putative metabolic genes (SAUSA300_1088, *hxlA*, *glcU*, *ctaB*) (Figure S8).

This suggested that single locus effects do not have a strong impact on mortality or that this impact could not be detected with the size of our datasets. To investigate whether combination of mutations in multiple genetic loci could influence SAB mortality, we assessed the predictive power of bacterial genetic variation in our classification model of mortality. Since dimensionality can affect the performance of machine-learning algorithms, we did not use single mutations as features. Instead, we aggregated mutations in core genes and calculated a score based on the predicted impact of the mutation (lowest score for no mutation, highest score for predicted protein truncations). The performance of this unfiltered gene-based approach was only slightly higher than chance alone (mean AUROC: 0.56, mean AUPRC: 0.17) (Figure 5) and did not improve when restricting to rare mutations, using the same approach as with the gene-burden GWAS (mean AUROC: 0.52, mean AUPRC: 0.17). Similarly, feature selection based on GWAS hits did not improve the performance (Figure S9). We also considered a model that used only the first ten principal components with similar predictive performance (mean AUROC: 0.58, mean AUPRC: 0.21). Furthermore, the performance of the model based on clinical predictors was not improved by adding genomic predictors (mean AUROC: 0.77–0.78, mean AUPRC: 0.31–0.35).

DISCUSSION

We used a multi-faceted strategy including bacterial GWAS and supervised machine learning to assess the role of bacterial genetic factors in clinically relevant phenotypes and outcomes of *S. aureus* bacteremia. We found that the bacterial genotype has limited impact on these outcomes, in particular mortality, as demonstrated by low heritability, poor predictive performance in our classification model of SAB, and a limited number of significant findings in the bacterial GWAS. Despite the weak genotype-outcomes association, we were able to identify a small number of candidate pathoadaptive loci associated with duration of bacteremia, an important intermediate outcome of SAB

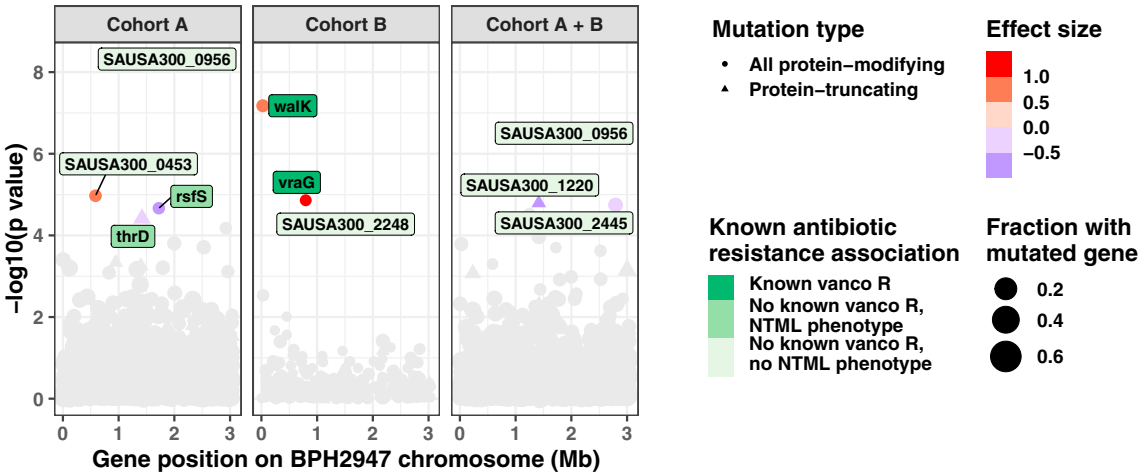


Figure 3. Gene-burden GWAS of vancomycin MIC in cohorts A, B, and A and B combined

The association was computed using a linear mixed model (LMM) with vancomycin MIC as continuous outcome variable. All mutations with predicted impact on the protein sequence were aggregated by coding regions (circles: all protein modifying, triangles: protein truncating only). The size of the points represents the fractions of strains with a mutated genes and the color indicated the effect size (impact on vancomycin MIC) in the linear regression. Gene labels are colored according to following criteria: known association with vancomycin resistance and presence of a vancomycin phenotype in cognate Nebraska Transposon Library mutants (NTLMs). The complete GWAS output is shown in Table S6 and Table S12. No vancomycin E-test data were available for cohort C.

that can reflect treatment response. In addition, our GWAS approach expanded the current picture of the polygenic background of vancomycin resistance in clinical *S. aureus* with several new genetic loci implicated.

Identifying causative genomic correlates of bacterial phenotypes has been difficult even when dealing with phenotypes with expected strong heritability (like antibiotic resistance),[21] but it is particularly challenging for clinical phenotypes such as invasiveness, treatment response, or infection-related mortality.[39,65] These outcomes are expected to be multifactorial, and the relative contribution of bacterial genetics can be extremely variable. Moreover, because the size of the datasets are usually small, negative findings might reflect an insufficient sample size, not an absence of contributing bacterial effects.

Among these multi-factorial clinical outcomes, invasiveness is thought to have a detectable bacterial molecular signature.[51] Indeed, even if bacterial invasion was exclusively determined by host and environmental factors (or by chance), some form of adaptive evolution would be expected as the result of the transition from the colonizing niche to blood or human tissues,[30] as confirmed by within-host evolution studies where a proportion of invasive isolates carried molecular signatures of adaptation that were absent in colonizing isolates.[15,17,66] In accordance with this premise, two adequately sized bacterial GWASs of *S. pneumoniae* meningitis have identified point mutations in both pathogenesis and antibiotic resistance genes that were significantly associated with bacterial meningitis.[51,67] Strikingly, in one study, the association was replicated in an independent cohort.[67] In their study, Lees et al. estimated the heritability of bacterial meningitis at a staggering 70%.[51] By contrast, heritability of *S. aureus* invasiveness was much lower (4%) in the GWAS study by Young et al.,[68] who identified a highly significant association between invasiveness and mutations in the dihydrofolate reductase gene *dfrB* (associated with antibiotic resistance). This result is consistent with the association of penicillin-binding protein mutations with invasive *S. pneumoniae* infections and suggests an intriguing association between antibiotic resistance and invasiveness in bacteria, possibly due to the pleomorphic effects of antibiotic resistance and its impact on immune evasion and persistence.[69,70]

The attempt of applying bacterial GWASs to infection outcomes (such as mortality) has been less successful than GWASs on antibiotic resistance mechanisms. While Recker et al. were able to identify some molecular correlates of mortality from SAB,[14] a Danish study found no association between bacterial genetic determinants and occurrence of infectious endocarditis during SAB.[71] Similarly, Lees et al. found a stark contrast between the high heritability of the invasiveness phenotype and the null heritability of the meningitis outcome phenotype.[51] In a bacterial GWAS of a cohort of 912 patients with *E. coli* bacteremia, no bacterial genetic variant was associated with mortality, septic shock, or intensive care unit admission. This correlated with very low heritability estimates for these outcomes.[4]

Our heritability estimates and machine-learning model confirmed the weak association between bacterial genotype and clinical outcomes. However, we hypothesized that it was possible to detect a role for bacterial variation based on the assumption that niche adaptation of invasive *S. aureus* promotes antibiotic resistance or tolerance. Building on this framework, we performed a multi-cohort GWASs with a series of phenotypes with expected decreasing importance of bacterial factors (vancomycin MIC, duration of infection, infection mortality). Specifically, we focused on the vancomycin MIC because this phenotype was an important feature in our random forest model of mortality from SAB, in agreement with previous analyses of the cohort A.[12] In addition, vancomycin MIC is an example of

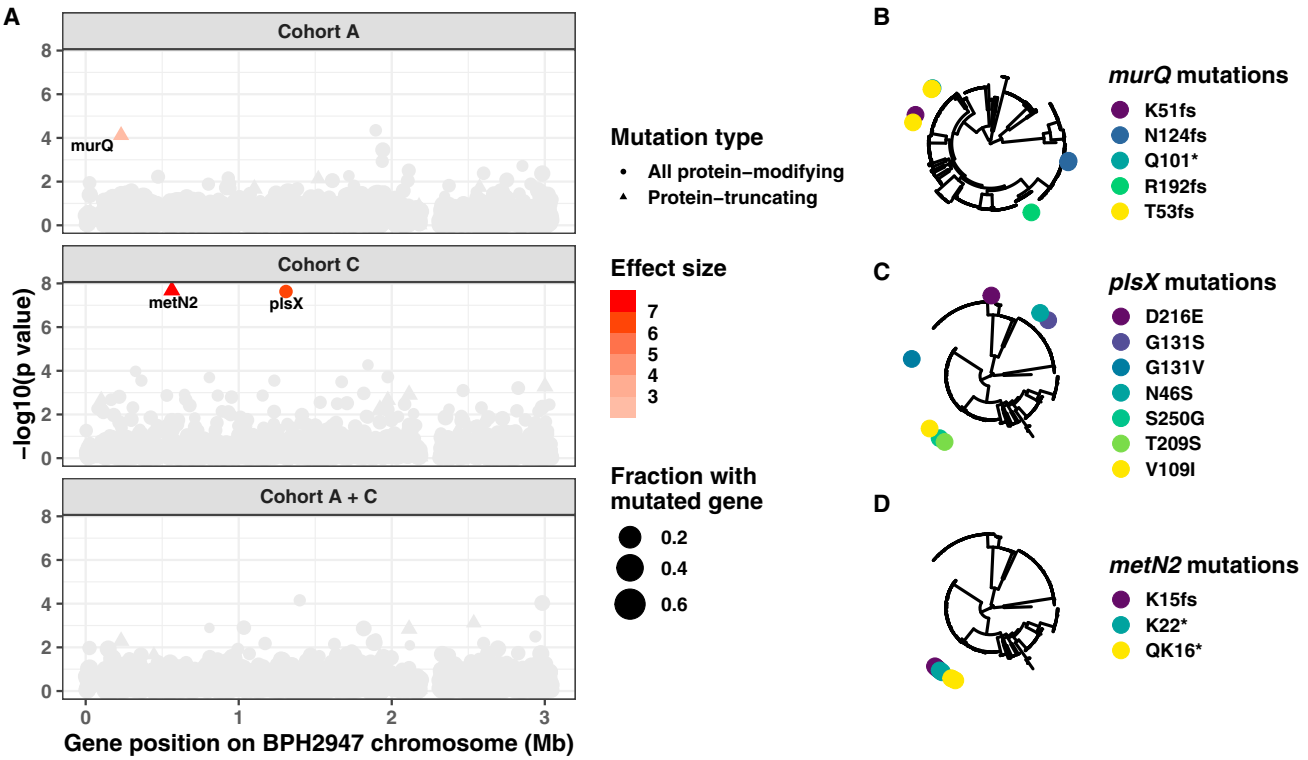


Figure 4. Gene-burden GWAS of duration of bacteremia in cohorts A, C, and A and C combined

(A) Manhattan plot showing the strength of the significance of mutated genes. The association was computed using a linear mixed model (LMM) with duration of bacteremia as continuous outcome variable. All mutations with predicted impact on the protein sequence were aggregated by coding regions (circles: all protein modifying, triangles: protein truncating only). The size of the points represents the fractions of strains with a mutated gene, and the color indicates the effect size (impact on duration of bacteremia) in the linear regression.

(B–D) Positions of *murQ* (B), *plsX* (C), and *metN2* (D) mutations on the phylogenetic tree. The complete GWAS output is shown in Table S7.

Duration of bacteremia was not recorded for cohort B.

adaptative phenotype that is relatively easy to measure and that is available for hundreds of clinical isolates. We then progressed to duration of bacteremia, a key intermediary outcome of SAB[72] that is related to antibiotic efficacy and is often integrated into clinical management.

We first identified a key role of lineage effects, in particular the vancomycin MIC phenotype. Identifying lineage effects is important because they provide an estimate of the risk of false positive GWAS findings related to the confounding from the population structure.[23] On the other hand, the presence of lineage effects might signify that there are high-risk clones for resistance or virulence phenotype, a feature that is important for pathogen surveillance and possibly for outcome prediction.[73]

Previous GWASs of vancomycin MIC have shown the role of known determinants of vancomycin resistance including the *walKR*[49] and *rpoB*[28] genes. Here, in a large GWAS of vancomycin MIC, we confirm the role of some of these established determinants of vancomycin resistance (*walK*, *mprF*) and provide evidence for new loci, the most promising being *thrD*, an aspartate kinase, and *rsfS*, a ribosome silencing factor involved in the stress response. Our results confirm the potential for bacterial GWASs to assist the discovery of non-canonical antibiotic resistance loci.[29]

Overall, only a few molecular markers achieve genome-wide significance in our GWASs of SAB mortality, indicating that larger sample sizes including thousands of sequences will be necessary to draw firm conclusions on the impact the bacterial genotype on mortality. Therefore, we considered duration of bacteremia, an intermediate outcome that strongly associates with mortality.[72] We found that truncations in *murQ*, a gene linked to the degradation of MurNAc-6p, an intermediate metabolite in peptidoglycan recycling,[62] were significantly associated with persistent SAB, potentially suggesting a role in antibiotic resistance or tolerance. Two candidate loci of persistent bacteremia were the methionine biosynthesis locus *metN2* and the phospholipid biosynthesis gene *plsX*. MetN2 is an ABC transporter; a potential clue to its role in bacteremia comes from a study where expression of *metN2* was strongly reduced in *S. aureus* mutants that were resistant to antimicrobial peptides.[74] Interestingly, *plsX* has also a potential association with antimicrobial peptide resistance: it is closely connected with the fatty-acid kinase *fakA*,[63] which was linked to resistance to the antimicrobial peptide dermcidin in a transposon mutant screen.[64]

There is ongoing uncertainty around the role of bacterial genetic factors in human infections. Here, we show that despite a

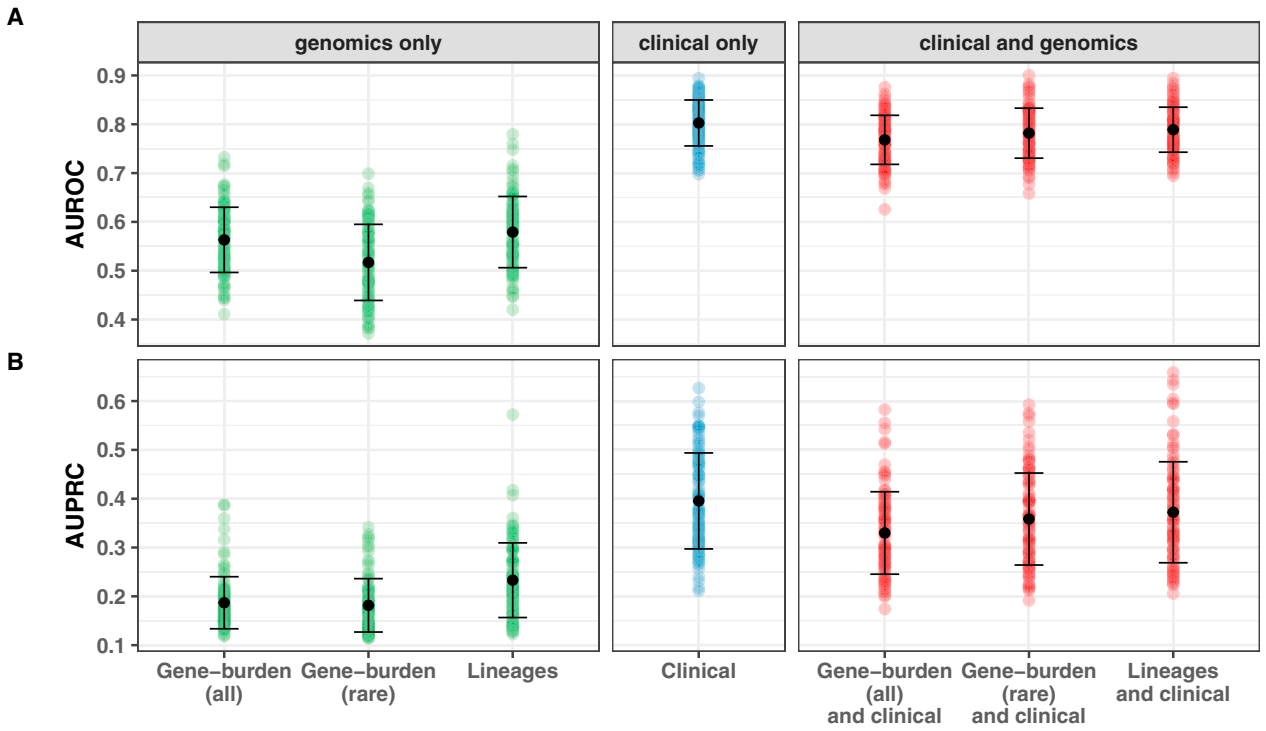


Figure 5. Predictive performance of seven random forest models of SAB mortality in cohort A

The data shown are the summary of 100 replicated random data splits. Error bars represent mean and standard deviation.

(A) Area under the receiver operating characteristic curve (AUROC).

(B) Area under the precision recall curve (AUPRC). A summary of the random forest models performance is shown in Table S9.

likely preponderant role for host and treatment factors in shaping infection outcomes, it is also possible to disentangle the multifactorial networks underlying these infections and detect specific bacterial genetic loci that help drive particular clinical outcomes. While these single loci likely have a low global predictive value, they might be more relevant in specific, high-risk settings (e.g., persistent or recurrent infection) and might act in concert with other bacterial variants, host genetic variants, and treatment factors. In addition, the genetic loci that were enriched in our analysis might represent new clinically relevant molecular targets of antibiotic resistance and tolerance and important candidates for functional validation.

Limitations of the study

Several units of bacterial genetic variation can be used in bacterial GWASs, including accessory genes, point mutations, large structural variants, or k-mers. Here, we focused on point mutations in core genes because we have previously shown that this type of variation correlates well with adaptive phenotypes including vancomycin and oxacillin resistance and loss of cytotoxicity. While combining three cohorts increased the statistical power of our analysis, it also introduced heterogeneities in the data and limited the number of co-variates we could use in the combined dataset. We addressed this shortcoming by performing both combined and per-cohort analyses. The distribution of duration of bacteremia was heavily skewed, and the variable could not be normalized using multiple approaches. In the future, this could be addressed by obtaining more precise duration data (e.g., systematic blood culture collection or cell-free DNA studies[75]). Finally, we could validate vancomycin resistance GWAS hits but not genes associated with duration of bacteremia. Validation of signatures of clinical outcomes is challenging due to the lack of reliable experimental phenotypes. We acknowledge that these findings should be validated in a second collection of clinical strains.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Study cohorts
  - Ethics approval
- METHOD DETAILS
  - Phenotypic testing
  - Whole-genome sequencing
- QUANTIFICATION AND STATISTICAL ANALYSIS

- Machine learning model to select predictors of mortality/treatment response
- Variant calling
- Preparation of the bacterial genotype files
- Variant annotation
- Bacterial GWAS design
- Bacterial GWAS pipeline
- Lineage effects and heritability

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2023.113069.

ACKNOWLEDGMENTS

We acknowledge all the patients and healthcare providers who were involved in the ANZCOSS and VANESSA cohorts. Patients were included in the cohorts by the following investigators and centers in Australia and New Zealand: Natasha E. Holmes, Paul D. R. Johnson, and Benjamin P. Howden (Austin Health, Heidelberg); Wendy J. Munckhof (Princess Alexandra Hospital, Woolloongabba); J. Owen Robinson (Royal Perth Hospital, Perth); Tony M. Korman (Southern Health, Clayton); Matthew V.N. Sullivan (Westmead Hospital, Westmead); Tara L. Anderson and Sanchia Warren (Royal Hobart Hospital, Hobart); Sally A. Roberts (Auckland District Health Board, Auckland); S.V.H. (Liverpool Hospital, Liverpool and Royal Prince Alfred Hospital, Sydney); Allen C. Cheng (Alfred Health, Melbourne); Eugene Athan (Barwon Health, Geelong); and John D. Turnidge (Australian Commission on Safety and Quality in Healthcare, Sydney). We thank Takehiro Tomita and Susan Ballard (Microbiological Diagnostic Unit Public Health Laboratory, Melbourne) for performing whole-genome sequencing of the ANZCOSS isolates. We acknowledge the CAMERA2 Study Group for sharing sequences and clinical metadata of trial participants with multiple sequential bacteremia strains: Nick Anagostou, David Andresen, Sophia Archuleta, Narin Bak, Alan Cass, Mark Chatfield, Alan Cheng, Jane Davies, Joshua Davis, Yael Dishon, Ravindra Dotel, Patricia Ferguson, Hong Foo, Vance Fowler, Niladri Ghosh, Timothy Gray, Stephen Guy, N.E.H., Benjamin Howden, Sandra Johnson, Shirin Kalimuddin, David Lye, Stephen McBride, Genevieve McKew, Niamh Meagher, Jane Nelson, Matthew O'Sullivan, David Paterson, Mical Paul, David Price, Anna Ralph, Matthew Roberts, Owen Robinson, Ben Rogers, Naomi Runnegar, Simon Smith, Archana Sud, S.Y.C.T., Adrian Tramontana, S.V.H., Genevieve Walls, Morgyn Warner, Dafna Yahav, and Barnaby Young. This work was supported by a Research Fellowship from the National Health and Medical Research Council, Australia, to B.P.H. (GNT1196103) and T.P.S. (GNT1105525). S.G.G. was supported by a PhD scholarship of the University of Melbourne. A.H. was supported by the National Health and Medical Research Council, Australia (GNT2018880).

AUTHOR CONTRIBUTIONS

Conceptualization, S.G.G., R.G., T.P.S., and B.P.H.; methodology, S.G.G. and R.G.; software, S.G.G. and T.S.; validation, S.G.G., R.G., and A.S.H.; formal analysis, S.G.G.; investigation, S.G.G., N.E.M., S.L.B., A.S.H., D.S.D., J.S.D., S.V.H., and S.Y.C.T.; resources, N.E.M., J.S.D., S.V.H., S.Y.C.T., and B.P.H.; data curation, S.G.G., N.E.M., S.L.B., D.S.D., J.S.D., S.V.H., and S.Y.C.T.; writing – original draft, S.G.G.; writing – review & editing, S.G.G., R.G., N.E.M., S.L.B., A.H., J.S.D., S.V.H., S.Y.C.T., T.P.S., and B.P.H.; visualization, S.G.G.; supervision, R.G., T.P.S., and B.P.H.; project administration, S.G.G.; funding acquisition, T.P.S. and B.P.H.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in their field of research or within their geographical location. One or more of the authors of this paper self-identifies as a member of the LGBTQIA+ community. One or more of the authors of this paper self-identifies as living with a disability.

Received: December 1, 2022
Revised: June 29, 2023
Accepted: August 18, 2023
Published: September 12, 2023

REFERENCES

1. Cohen, J., Vincent, J.L., Adhikari, N.K.J., Machado, F.R., Angus, D.C., Calandra, T., Jaton, K., Giulieri, S., Delaloye, J., Opal, S., et al. (2015). Sepsis: a roadmap for future research. Lancet Infect. Dis. *15*, 581–614. https://doi.org/10.1016/S1473-3099(15)70112-X.
2. Kwok, A.J., Mentzer, A., and Knight, J.C. (2021). Host genetics and infectious disease: new tools, insights and translational opportunities. Nat. Rev. Genet. *22*, 137–153. https://doi.org/10.1038/s41576-020-00297-6.
3. Giulieri, S.G., Tong, S.Y.C., and Williamson, D.A. (2020). Using genomics to understand meticillin- and vancomycin-resistant *Staphylococcus aureus* infections. Microb. Genom. *6*, e000324. https://doi.org/10.1099/mgen.0.000324.
4. Denamur, E., Condamine, B., Esposito-Farèse, M., Royer, G., Clermont, O., Laouenan, C., Lefort, A., de Lastours, V., and Galardini, M.; the, C., and SEPTICOLI groups (2022). Genome wide association study of Escherichia coli bloodstream infection isolates identifies genetic determinants for the portal of entry but not fatal outcome. PLoS Genet. *18*, e1010112. https://doi.org/10.1371/journal.pgen.1010112.
5. Power, R.A., Parkhill, J., and de Oliveira, T. (2017). Microbial genome-wide association studies: lessons from human GWAS. Nat. Rev. Genet. *18*, 41–50. https://doi.org/10.1038/nrg.2016.132.
6. Battle, S.E., Shuping, M., Withers, S., Justo, J.A., Bookstaver, P.B., and Al-Hasan, M.N. (2022). Prediction of Mortality in *Staphylococcus aureus* Bloodstream Infection using Quick Pitt Bacteremia Score. J. Infect. *84*, 131–135. https://doi.org/10.1016/j.jinf.2021.12.002.
7. Bai, A.D., Lo, C.K.L., Komorowski, A.S., Suresh, M., Guo, K., Garg, A., Tandon, P., Senecal, J., Del Corpo, O., Stefanova, I., et al. (2022). *Staphylococcus aureus* bacteremia mortality: A systematic review and meta-analysis. Clin. Microbiol. Infect. *28*, 1076–1084. https://doi.org/10.1016/j.cmi.2022.03.015.
8. Imam, N., Tempone, S., Armstrong, P.K., McCann, R., Johnson, S., Worth, L.J., and Richards, M.J. (2019). Increased incidence of community-associated *Staphylococcus aureus* bloodstream infections in Victoria and Western Australia, 2011–2016. Med. J. Aust. *210*, 87–88. https://doi.org/10.5694/mja2.12057.
9. Tong, S.Y.C., Davis, J.S., Eichenberger, E., Holland, T.L., and Fowler, V.G., Jr. (2015). *Staphylococcus aureus* infections: epidemiology, pathophysiology, clinical manifestations, and management. Clin. Microbiol. Rev. *28*, 603–661. https://doi.org/10.1128/CMR.00134-14.
10. van Hal, S.J., Jensen, S.O., Vaska, V.L., Espedido, B.A., Paterson, D.L., and Gosbell, I.B. (2012). Predictors of mortality in *Staphylococcus aureus* Bacteremia. Clin. Microbiol. Rev. *25*, 362–386. https://doi.org/10.1128/cmr.05022-11.
11. Mba Medie, F., Sharma-Kuinkel, B.K., Ruffin, F., Chan, L.C., Rossetti, M., Chang, Y.L., Park, L.P., Bayer, A.S., Filler, S.G., Ahn, R., et al. (2019). Genetic variation of DNA methyltransferase-3A contributes to protection against persistent MRSA bacteremia in patients. Proc. Natl. Acad. Sci. USA. *116*, 20087–20096. https://doi.org/10.1073/pnas.1909849116.
12. Holmes, N.E., Turnidge, J.D., Munckhof, W.J., Robinson, J.O., Korman, T.M., O'Sullivan, M.V.N., Anderson, T.L., Roberts, S.A., Gao, W.,

Christiansen, K.J., et al. (2011). Antibiotic choice may not explain poorer outcomes in patients with *Staphylococcus aureus* bacteremia and high vancomycin minimum inhibitory concentrations. J. Infect. Dis. *204*, 340–347. https://doi.org/10.1093/infdis/jir270.

13. Giulieri, S.G., Holmes, N.E., Stinear, T.P., and Howden, B.P. (2016). Use of bacterial whole-genome sequencing to understand and improve the management of invasive *Staphylococcus aureus* infections. Expert Rev. Anti Infect. Ther. *14*, 1023–1036. https://doi.org/10.1080/14787210.2016.1233815.

14. Recker, M., Laabei, M., Toleman, M.S., Reuter, S., Saunderson, R.B., Blane, B., Török, M.E., Ouadi, K., Stevens, E., Yokoyama, M., et al. (2017). Clonal differences in *Staphylococcus aureus* bacteraemia-associated mortality. Nat. Microbiol. *2*, 1381–1388. https://doi.org/10.1038/s41564-017-0001-x.

15. Young, B.C., Wu, C.-H., Gordon, N.C., Cole, K., Price, J.R., Liu, E., Sheppard, A.E., Perera, S., Charlesworth, J., Golubchik, T., et al. (2017). Severe infections emerge from commensal bacteria by adaptive evolution. Elife *6*, e30637. https://doi.org/10.7554/eLife.30637.

16. Giulieri, S.G., Baines, S.L., Guerillot, R., Seemann, T., Gonçalves da Silva, A., Schultz, M., Massey, R.C., Holmes, N.E., Stinear, T.P., and Howden, B.P. (2018). Genomic exploration of sequential clinical isolates reveals a distinctive molecular signature of persistent *Staphylococcus aureus* bacteraemia. Genome Med. *10*, 65. https://doi.org/10.1186/s13073-018-0574-x.

17. Giulieri, S.G., Guérillot, R., Duchene, S., Hachani, A., Daniel, D., Seemann, T., Davis, J.S., Tong, S.Y.C., Young, B.C., Wilson, D.J., et al. (2022). Niche-specific genome degradation and convergent evolution shaping *Staphylococcus aureus* adaptation during severe infections. Elife *11*, e77195. https://doi.org/10.7554/eLife.77195.

18. Farhat, M.R., Shapiro, B.J., Kieser, K.J., Sultana, R., Jacobson, K.R., Victor, T.C., Warren, R.M., Streicher, E.M., Calver, A., Sloutsky, A., et al. (2013). Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. Nat. Genet. *45*, 1183–1189. https://doi.org/10.1038/ng.2747.

19. Collins, C., and Didelot, X. (2018). A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. PLoS Comput. Biol. *14*, e1005958. https://doi.org/10.1371/journal.pcbi.1005958.

20. Brynildsrud, O., Bohlin, J., Scheffer, L., and Eldholm, V. (2016). Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol. *17*, 238. https://doi.org/10.1186/s13059-016-1108-8.

21. Chen, P.E., and Shapiro, B.J. (2021). Classic genome-wide association methods are unlikely to identify causal variants in strongly clonal microbial populations. Preprint at bioRxiv, 2021.2006.2030.450606. https://doi.org/10.1101/2021.06.30.450606.

22. Saund, K., and Snitkin, E.S. (2020). Hogwash: three methods for genome-wide association studies in bacteria. Microb. Genom. *6*, mgen000469. https://doi.org/10.1099/mgen.0.000469.

23. Earle, S.G., Wu, C.-H., Charlesworth, J., Stoesser, N., Gordon, N.C., Walker, T.M., Spencer, C.C.A., Iqbal, Z., Clifton, D.A., Hopkins, K.L., et al. (2016). Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nat. Microbiol. *1*, 16041. https://doi.org/10.1038/nmicrobiol.2016.41.

24. Lees, J.A., Galardini, M., Bentley, S.D., Weiser, J.N., and Corander, J. (2018). pyseer: a comprehensive tool for microbial pangenome-wide association studies. Bioinformatics *34*, 4310–4312. https://doi.org/10.1093/bioinformatics/bty539.

25. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575. https://doi.org/10.1086/519795.

26. Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. *44*, 821–824. https://doi.org/10.1038/ng.2310.

27. Hicks, N.D., Yang, J., Zhang, X., Zhao, B., Grad, Y.H., Liu, L., Ou, X., Chang, Z., Xia, H., Zhou, Y., et al. (2018). Clinically prevalent mutations in Mycobacterium tuberculosis alter propionate metabolism and mediate multidrug tolerance. Nat. Microbiol. *3*, 1032–1042. https://doi.org/10.1038/s41564-018-0218-3.

28. Alam, M.T., Petit, R.A., 3rd, Crispell, E.K., Thornton, T.A., Conneely, K.N., Jiang, Y., Satola, S.W., and Read, T.D. (2014). Dissecting vancomycin-intermediate resistance in *Staphylococcus aureus* using genome-wide association. Genome Biol. Evol. *6*, 1174–1185. https://doi.org/10.1093/gbe/evu092.

29. Farhat, M.R., Freschi, L., Calderon, R., Ioerger, T., Snyder, M., Meehan, C.J., de Jong, B., Rigouts, L., Sloutsky, A., Kaur, D., et al. (2019). GWAS for quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions. Nat. Commun. *10*, 2128. https://doi.org/10.1038/s41467-019-10110-6.

30. Culyba, M.J., and Van Tyne, D. (2021). Bacterial evolution during human infection: Adapt and live or adapt and die. PLoS Pathog. *17*, e1009872. https://doi.org/10.1371/journal.ppat.1009872.

31. Balaban, N.Q., Helaine, S., Lewis, K., Ackermann, M., Aldridge, B., Andersson, D.I., Brynildsen, M.P., Bumann, D., Camilli, A., Collins, J.J., et al. (2019). Definitions and guidelines for research on antibiotic persistence. Nat. Rev. Microbiol. *17*, 441–448. https://doi.org/10.1038/s41579-019-0196-3.

32. Holmes, N.E., Robinson, J.O., van Hal, S.J., Munckhof, W.J., Athan, E., Korman, T.M., Cheng, A.C., Turnidge, J.D., Johnson, P.D.R., and Howden, B.P.; VANESSA study group on behalf of the Australasian Society for Infectious Diseases ASID Clinical Research Network CRN (2018). Morbidity from in-hospital complications is greater than treatment failure in patients with *Staphylococcus aureus* bacteraemia. BMC Infect. Dis. *18*, 107. https://doi.org/10.1186/s12879-018-3011-2.

33. van Hal, S.J., Jones, M., Gosbell, I.B., and Paterson, D.L. (2011). Vancomycin Heteroresistance Is Associated with Reduced Mortality in ST239 Methicillin-Resistant *Staphylococcus aureus* Blood Stream Infections. PLoS One *6*, e21217. https://doi.org/10.1371/journal.pone.0021217.

34. Tong, S.Y.C., Lye, D.C., Yahav, D., Sud, A., Robinson, J.O., Nelson, J., Archuleta, S., Roberts, M.A., Cass, A., Paterson, D.L., et al. (2020). Effect of Vancomycin or Daptomycin With vs Without an Antistaphylococcal β-Lactam on Mortality, Bacteremia, Relapse, or Treatment Failure in Patients With MRSA Bacteremia: A Randomized Clinical Trial. JAMA *323*, 527–537. https://doi.org/10.1001/jama.2020.0103.

35. Lemeshow, S., Sturdivant, R.X., Hosmer, D.W., Jr., and Hosmer, D.W., Jr. (2013). Applied Logistic Regression (Incorporated: John Wiley & Sons).

36. Liu, P., Hao, Z., Liu, M., Niu, M., Sun, P., Yan, S., Zhao, L., and Zhao, X. (2021). Genetic mutations in adaptive evolution of growth-independent vancomycin-tolerant *Staphylococcus aureus*. J. Antimicrob. Chemother. *76*, 2765–2773. https://doi.org/10.1093/jac/dkab260.

37. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. *17*, 132. https://doi.org/10.1186/s13059-016-0997-x.

38. Mai, T.T., Lees, J.A., Gladstone, R.A., and Corander, J. (2023). Inferring the heritability of bacterial traits in the era of machine learning. Bioinform. Adv. *3*, vbad027. https://doi.org/10.1093/bioadv/vbad027.

39. Young, B.C., Wu, C.-H., Charlesworth, J., Earle, S., Price, J.R., Gordon, N.C., Cole, K., Dunn, L., Liu, E., Oakley, S., et al. (2021). Antimicrobial resistance determinants are associated with *Staphylococcus aureus* bacteraemia and adaptation to the healthcare environment: a bacterial genome-wide association study. Microb. Genom. *7*, 000700. https://doi.org/10.1099/mgen.0.000700.

40. van Hal, S.J., Lodise, T.P., and Paterson, D.L. (2012). The clinical significance of vancomycin minimum inhibitory concentration in

Staphylococcus aureus infections: a systematic review and meta-analysis. Clin. Infect. Dis. *54*, 755–771. https://doi.org/10.1093/cid/cir935.

41. Holmes, N.E., Johnson, P.D.R., and Howden, B.P. (2012). Relationship between Vancomycin-Resistant *Staphylococcus aureus*, Vancomycin-Intermediate *S. aureus*, High Vancomycin MIC, and Outcome in Serious *S. aureus* Infections. J. Clin. Microbiol. *50*, 2548–2552. https://doi.org/10.1128/JCM.00775-12.

42. Machado, H., Seif, Y., Sakoulas, G., Olson, C.A., Hefner, Y., Anand, A., Jones, Y.Z., Szubin, R., Palsson, B.O., Nizet, V., and Feist, A.M. (2021). Environmental conditions dictate differential evolution of vancomycin resistance in *Staphylococcus aureus*. Commun. Biol. *4*, 793. https://doi.org/10.1038/s42003-021-02339-z.

43. Thitiananpakorn, K., Aiba, Y., Tan, X.E., Watanabe, S., Kiga, K., Sato'o, Y., Boonsiri, T., Li, F.Y., Sasahara, T., Taki, Y., et al. (2020). Association of mprF mutations with cross-resistance to daptomycin and vancomycin in methicillin-resistant *Staphylococcus aureus* (MRSA). Sci. Rep. *10*, 16107. https://doi.org/10.1038/s41598-020-73108-x.

44. Bayer, A.S., Schneider, T., and Sahl, H.G. (2013). Mechanisms of daptomycin resistance in *Staphylococcus aureus*: role of the cell membrane and cell wall. Ann. N. Y. Acad. Sci. *1277*, 139–158. https://doi.org/10.1111/j.1749-6632.2012.06819.x.

45. Sieradzki, K., and Tomasz, A. (1999). Gradual Alterations in Cell Wall Structure and Metabolism in Vancomycin-Resistant Mutants of *Staphylococcus aureus*. J. Bacteriol. *181*, 7566–7570. https://doi.org/10.1128/jb.181.24.7566-7570.1999.

46. Cameron, D.R., Ward, D.V., Kostoulias, X., Howden, B.P., Moellering, R.C., Jr., Eliopoulos, G.M., and Peleg, A.Y. (2012). Serine/threonine phosphatase Stp1 contributes to reduced susceptibility to vancomycin and virulence in *Staphylococcus aureus*. J. Infect. Dis. *205*, 1677–1687. https://doi.org/10.1093/infdis/jis252.

47. Roberts, C.A., Al-Tameemi, H.M., Mashruwala, A.A., Rosario-Cruz, Z., Chauhan, U., Sause, W.E., Torres, V.J., Belden, W.J., and Boyd, J.M. (2017). The Suf Iron-Sulfur Cluster Biosynthetic System Is Essential in *Staphylococcus aureus*, and Decreased Suf Function Results in Global Metabolic Defects and Reduced Survival in Human Neutrophils. Infect. Immun. *85*, e00100-17. https://doi.org/10.1128/iai.00100-17.

48. Passalacqua, K.D., Satola, S.W., Crispell, E.K., and Read, T.D. (2012). A mutation in the PP2C phosphatase gene in a *Staphylococcus aureus* USA300 clinical isolate with reduced susceptibility to vancomycin and daptomycin. Antimicrob. Agents Chemother. *56*, 5212–5223. https://doi.org/10.1128/AAC.05770-11.

49. Baines, S.L., Holt, K.E., Schultz, M.B., Seemann, T., Howden, B.O., Jensen, S.O., van Hal, S.J., Coombs, G.W., Firth, N., Powell, D.R., et al. (2015). Convergent adaptation in the dominant global hospital clone ST239 of methicillin-resistant *Staphylococcus aureus*. mBio *6*, e00080. https://doi.org/10.1128/mBio.00080-15.

50. Weiner, D.J., Nadig, A., Jagadeesh, K.A., Dey, K.K., Neale, B.M., Robinson, E.B., Karczewski, K.J., and O'Connor, L.J. (2023). Polygenic architecture of rare coding variation across 394,783 exomes. Nature *614*, 492–499. https://doi.org/10.1038/s41586-022-05684-z.

51. Lees, J.A., Ferwerda, B., Kremer, P.H.C., Wheeler, N.E., Serón, M.V., Croucher, N.J., Gladstone, R.A., Bootsma, H.J., Rots, N.Y., Wijmega-Monsuur, A.J., et al. (2019). Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. Nat. Commun. *10*, 2176. https://doi.org/10.1038/s41467-019-09976-3.

52. Howden, B.P., McEvoy, C.R.E., Allen, D.L., Chua, K., Gao, W., Harrison, P.F., Bell, J., Coombs, G., Bennett-Wood, V., Porter, J.L., et al. (2011). Evolution of multidrug resistance during *Staphylococcus aureus* infection involves mutation of the essential two component regulator WalKR. PLoS Pathog. *7*, e1002359. https://doi.org/10.1371/journal.ppat.1002359.

53. Falord, M., Karimova, G., Hiron, A., and Msadek, T. (2012). GraXSR Proteins Interact with the VraFG ABC Transporter To Form a Five-Component System Required for Cationic Antimicrobial Peptide Sensing and Resistance in *Staphylococcus aureus*. Antimicrob. Agents Chemother. *56*, 1047–1058. https://doi.org/10.1128/AAC.05054-11.

54. Majorek, K.A., Osinski, T., Tran, D.T., Revilla, A., Anderson, W.F., Minor, W., and Kuhn, M.L. (2017). Insight into the 3D structure and substrate specificity of previously uncharacterized GNAT superfamily acetyltransferases from pathogenic bacteria. Biochim. Biophys. Acta. Proteins Proteom. *1865*, 55–64. https://doi.org/10.1016/j.bbapap.2016.10.011.

55. Whaley, S.G., Radka, C.D., Subramanian, C., Frank, M.W., and Rock, C.O. (2021). Malonyl-acyl carrier protein decarboxylase activity promotes fatty acid and cell envelope biosynthesis in Proteobacteria. J. Biol. Chem. *297*, 101434. https://doi.org/10.1016/j.jbc.2021.101434.

56. Qamar, A., and Golemi-Kotra, D. (2012). Dual roles of FmtA in *Staphylococcus aureus* cell wall biosynthesis and autolysis. Antimicrob. Agents Chemother. *56*, 3797–3805. https://doi.org/10.1128/AAC.00187-12.

57. El-Halfawy, O.M., Czarny, T.L., Flannagan, R.S., Day, J., Bozelli, J.C., Kuiack, R.C., Salim, A., Eckert, P., Epand, R.M., McGavin, M.J., et al. (2020). Discovery of an antivirulence compound that reverses β-lactam resistance in MRSA. Nat. Chem. Biol. *16*, 143–149. https://doi.org/10.1038/s41589-019-0401-8.

58. Riordan, J.T., Muthaiyan, A., Van Voorhies, W., Price, C.T., Graham, J.E., Wilkinson, B.J., and Gustafson, J.E. (2007). Response of *Staphylococcus aureus* to Salicylate Challenge. J. Bacteriol. *189*, 220–227. https://doi.org/10.1128/JB.01149-06.

59. Katayama, Y., Sekine, M., Hishinuma, T., Aiba, Y., and Hiramatsu, K. (2016). Complete Reconstitution of the Vancomycin-Intermediate *Staphylococcus aureus* Phenotype of Strain Mu50 in Vancomycin-Susceptible *S. aureus*. Antimicrob. Agents Chemother. *60*, 3730–3742. https://doi.org/10.1128/AAC.00420-16.

60. Khusainov, I., Fatkhullin, B., Pellegrino, S., Bikmullin, A., Liu, W.-t., Gabdulkhakov, A., Shebel, A.A., Golubev, A., Zeyer, D., Trachtmann, N., et al. (2020). Mechanism of ribosome shutdown by RsfS in *Staphylococcus aureus* revealed by integrative structural biology approach. Nat. Commun. *11*, 1656. https://doi.org/10.1038/s41467-020-15517-0.

61. Fey, P.D., Endres, J.L., Yajjala, V.K., Widhelm, T.J., Boissy, R.J., Bose, J.L., Bayles, K.W., and Bush, K. (2013). A Genetic Resource for Rapid and Comprehensive Phenotype Screening of Nonessential *Staphylococcus aureus* Genes. mBio *4*, e00537–e00512. https://doi.org/10.1128/mBio.00537-12.

62. Borisova, M., Gaupp, R., Duckworth, A., Schneider, A., Dalügge, D., Mühleck, M., Deubel, D., Unsleber, S., Yu, W., Muth, G., et al. (2016). Peptidoglycan Recycling in Gram-Positive Bacteria Is Crucial for Survival in Stationary Phase. mBio *7*, e00923-16. https://doi.org/10.1128/mBio.00923-16.

63. Parsons, J.B., Broussard, T.C., Bose, J.L., Rosch, J.W., Jackson, P., Subramanian, C., and Rock, C.O. (2014). Identification of a two-component fatty acid kinase responsible for host fatty acid incorporation by *Staphylococcus aureus*. Proc. Natl. Acad. Sci. USA. *111*, 10532–10537. https://doi.org/10.1073/pnas.1408797111.

64. Li, M., Rigby, K., Lai, Y., Nair, V., Peschel, A., Schittek, B., and Otto, M. (2009). *Staphylococcus aureus* Mutant Screen Reveals Interaction of the Human Antimicrobial Peptide Dermcidin with Membrane Phospholipids. Antimicrob. Agents Chemother. *53*, 4200–4210. https://doi.org/10.1128/AAC.00428-09.

65. Lapp, Z., Han, J.H., Wiens, J., Goldstein, E.J.C., Lautenbach, E., and Snitkin, E.S. (2021). Patient and Microbial Genomic Factors Associated with Carbapenem-Resistant Klebsiella pneumoniae Extraintestinal Colonization and Infection. mSystems *6*, e00177-21–e00121. https://doi.org/10.1128/mSystems.00177-21.

66. Lees, J.A., Kremer, P.H.C., Manso, A.S., Croucher, N.J., Ferwerda, B., Serón, M.V., Oggioni, M.R., Parkhill, J., Brouwer, M.C., van der Ende, A., et al. (2017). Large scale genomic analysis shows no evidence for pathogen adaptation between the blood and cerebrospinal fluid niches during bacterial meningitis. Microb. Genom. *3*, e000103. https://doi.org/10.1099/mgen.0.000103.

67. Li, Y., Metcalf, B.J., Chochua, S., Li, Z., Walker, H., Tran, T., Hawkins, P.A., Gierke, R., Pilishvili, T., McGee, L., and Beall, B.W. (2019). Genome-wide association analyses of invasive pneumococcal isolates identify a missense bacterial mutation associated with meningitis. Nat. Commun. *10*, 178. https://doi.org/10.1038/s41467-018-07997-y.

68. Young, B.C., Wu, C.-H., Charlesworth, J., Earle, S., Price, J.R., Gordon, N.C., Cole, K., Dunn, L., Liu, E., Oakley, S., et al. (2021). Antimicrobial resistance determinants are associated with *Staphylococcus aureus* bacteraemia and adaptation to the hospital environment: a bacterial genome-wide association study. medRxiv *2001*, 2013.21249734. https://doi.org/10.1101/2021.01.13.21249734.

69. Gao, W., Cameron, D.R., Davies, J.K., Kostoulias, X., Stepnell, J., Tuck, K.L., Yeaman, M.R., Peleg, A.Y., Stinear, T.P., and Howden, B.P. (2013). The RpoB H(4)(8)(1)Y rifampicin resistance mutation and an active stringent response reduce virulence and increase resistance to innate immune responses in *Staphylococcus aureus*. J. Infect. Dis. *207*, 929–939. https://doi.org/10.1093/infdis/jis772.

70. Guérillot, R., Kostoulias, X., Donovan, L., Li, L., Carter, G.P., Hachani, A., Vandelannoote, K., Giulieri, S., Monk, I.R., Kunimoto, M., et al. (2019). Unstable chromosome rearrangements in *Staphylococcus aureus* cause phenotype switching associated with persistent infections. Proc. Natl. Acad. Sci. USA. *116*, 20135–20140. https://doi.org/10.1073/pnas.1904861116.

71. Lilje, B., Rasmussen, R.V., Dahl, A., Stegger, M., Skov, R.L., Fowler, V.G., Ng, K.L., Kiil, K., Larsen, A.R., Petersen, A., et al. (2017). Whole-genome sequencing of bloodstream *Staphylococcus aureus* isolates does not distinguish bacteraemia from endocarditis. Microb. Genom. *3*, e000138. https://doi.org/10.1099/mgen.0.000138.

72. Kuehl, R., Morata, L., Boeing, C., Subirana, I., Seifert, H., Rieg, S., Kern, W.V., Kim, H.B., Kim, E.S., Liao, C.H., et al. (2020). Defining persistent *Staphylococcus aureus* bacteraemia: secondary analysis of a prospective cohort study. Lancet Infect. Dis. *20*, 1409–1417. https://doi.org/10.1016/s1473-3099(20)30447-3.

73. Lam, M.M.C., Wick, R.R., Watts, S.C., Cerdeira, L.T., Wyres, K.L., and Holt, K.E. (2021). A genomic surveillance framework and genotyping tool for Klebsiella pneumoniae and its related species complex. Nat. Commun. *12*, 4188. https://doi.org/10.1038/s41467-021-24448-3.

74. Kubicek-Sutherland, J.Z., Lofton, H., Vestergaard, M., Hjort, K., Ingmer, H., and Andersson, D.I. (2017). Antimicrobial peptide exposure selects for *Staphylococcus aureus* resistance to human defence peptides. J. Antimicrob. Chemother. *72*, 115–127. https://doi.org/10.1093/jac/dkw381.

75. Mourad, A., Fowler, V.G., and Holland, T.L. (2023). Which trial do we need? Next Generation Sequencing to Individualize Therapy in *Staphylococcus aureus* Bacteremia. Clin. Microbiol. Infect. *29*, 955–958. https://doi.org/10.1016/j.cmi.2023.04.004.

76. Turnidge, J.D., Kotsanas, D., Munckhof, W., Roberts, S., Bennett, C.M., Nimmo, G.R., Coombs, G.W., Murray, R.J., Howden, B., Johnson, P.D.R., et al. (2009). *Staphylococcus aureus* bacteraemia: a major cause of mortality in Australia and New Zealand. Med. J. Aust. *191*, 368–373.

77. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., and Korobeynikov, A. (2020). Using SPAdes de novo assembler. Current Protocols in Bioinformatics *70*, e102. https://doi.org/10.1002/cpbi.102.

78. Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biol *20*, 257. https://doi.org/10.1186/s13059-019-1891-0.

79. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology *17*, 132. https://doi.org/10.1186/s13059-016-0997-x.

80. Whalen, S., Schreiber, J., Noble, W.S., and Pollard, K.S. (2022). Navigating the pitfalls of applying machine learning in genomics. Nat. Rev. Genet. *23*, 169–181. https://doi.org/10.1038/s41576-021-00434-9.

81. Lemoine, F., and Gascuel, O. (2021). Gotree/Goalign : Toolkit and Go API to facilitate the development of phylogenetic workflows. bioRxiv *3*, 2009.447704. https://doi.org/10.1101/2021.06.09.447704.

82. Page, A.J., Taylor, B., Delaney, A.J., Soares, J., Seemann, T., Keane, J.A., and Harris, S.R. (2016). SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb. Genom. *2*, e000056. https://doi.org/10.1099/mgen.0.000056.

83. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. *37*, 1530–1534. https://doi.org/10.1093/molbev/msaa015.

84. Crispell, J., Balaz, D., and Gordon, S.V. (2019). HomoplasyFinder: a simple tool to identify homoplasies on a phylogeny. Microb. Genom. *5*, e000245. https://doi.org/10.1099/mgen.0.000245.

85. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

86. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly *6*, 80–92.

87. Fuchs, S., Mehlan, H., Bernhardt, J., Hennig, A., Michalik, S., Surmann, K., Pané-Farré, J., Giese, A., Weiss, S., Backert, L., et al. (2018). AureoWiki The repository of the *Staphylococcus aureus* research and annotation community. Int. J. Med. Microbiol. *308*, 558–568. https://doi.org/10.1016/j.ijmm.2017.11.011.

88. Dehal, P.S., Joachimiak, M.P., Price, M.N., Bates, J.T., Baumohl, J.K., Chivian, D., Friedland, G.D., Huang, K.H., Keller, K., Novichkov, P.S., et al. (2010). MicrobesOnline: an integrated portal for comparative and functional genomics. Nucleic Acids Res. *38*, D396–D400. https://doi.org/10.1093/nar/gkp919.

89. Peterson, R.A., and Cavanaugh, J.E. (2020). Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. J. Appl. Stat. *47*, 2312–2327. https://doi.org/10.1080/02664763.2019.1630372.

90. McCaw, Z.R., Lane, J.M., Saxena, R., Redline, S., and Lin, X. (2020). Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. Biometrics *76*, 1262–1272. https://doi.org/10.1111/biom.13214.

91. Törönen, P., Medlar, A., and Holm, L. (2018). PANNZER2: a rapid functional annotation web server. Nucleic Acids Res. *46*, W84-w88. https://doi.org/10.1093/nar/gky350.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
Staphylococcus aureus BPH2947	NCBI	GenBank: GCF_900620245.1
Staphylococcus aureus FPR3757	NCBI	GenBank: GCF_000013465.1
Nebraska Transposon Mutants Library (NTML)	University of Nebraska Medical Center	https://app1.unmc.edu/fgx/tools.html
Deposited data		
Staphylococcus aureus raw reads: cohort A	This manuscript	GenBank: PRJEB22792, PRJEB27932, PRJEB63385
Staphylococcus aureus raw reads: cohort B	This manuscript	GenBank: PRJEB2561
Staphylococcus aureus raw reads: cohort C	This manuscript	GenBank: PRJEB50796, PRJEB63381
Software and algorithms		
R	CRAN	https://www.R-project.org/
tidymodels	Github	https://www.tidymodels.org/
Kraken	Github	https://github.com/DerrickWood/kraken
Shovill	Github	https://github.com/tseemann/shovill
Abricate	Github	https://github.com/tseemann/abricate
Snippy	Github	https://github.com/tseemann/snippy
Goalign	Github	https://github.com/evolbioinfo/goalign
Snp-sites	Github	https://github.com/sanger-pathogens/snp-sites
IQ-tree	Github	https://github.com/Cibiv/IQ-TREE
HomoplasyFinder	Github	https://github.com/JosephCrispell/homoplasyFinder
BCFTools	Github	https://github.com/samtools/bcftools
Blast	NCBI	https://blast.ncbi.nlm.nih.gov/doc/blast-help/downloadblastdata.html
BestNormalise	CRAN	https://cran.r-project.org/web/packages/bestNormalize/
RNOmni	Github	https://github.com/zrmacc/RNOmni
Pyseer	Github	https://github.com/mgalardini/pyseer/
GEMMA	Github	https://github.com/genetics-statistics/GEMMA
Mash	Github	https://github.com/marbl/Mash/
Bugwas	Github	https://github.com/sgearle/bugwas/
Other		
Aureowiki	Website	https://aureowiki.med.uni-greifswald.de
MicrobesOnline	Website	http://www.microbesonline.org/
National Database of Antibiotic Resistant Organisms	NCBI	https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Stefano G. Giulieri (stefano.giulieri@unimelb.edu.au).

Materials availability

Strains used in this study will be shared by the lead contact upon request.

Data and code availability

- The raw reads of the strains included in this analysis have been deposited at the European Nucleotide Archive and are available as of date of publication. Bioproject accession numbers are listed in the Key resources table.
- All original code for the preparation of bacterial genotype files and for running Pyseer has been deposited at Github and is publicly available as of the date of publication. DOIs are listed in the Key resources table.
- Any additional information required to reanalyse the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Study cohorts

Cohort A was assembled from two multicentre observational studies of SAB that were performed to assess the impact of vancomycin MIC on clinical outcomes.[16] The vancomycin substudy of the Australian and New Zealand Cooperative on Outcome in Staphylococcal Sepsis (ANZCOSS) study included 532 cases of SAB from a larger retrospective study of SAB between 2007-2008.[76] The selection of episodes for the substudy has been described previously.[12] Briefly, to ensure a balanced proportion of cases, each SAB case treated with vancomycin was matched with the next available case treated with flucloxacillin at each participating center. The Vancomycin Efficacy in Staphylococcal Sepsis in Australasia (VANESSA) study included prospectively 226 patients with SAB between 2012-2013.[32] Data collected within cohort A included patient demographics, site of acquisition, comorbidities, clinical syndrome, antibiotic treatment, 30-day mortality and duration of bacteraemia. Baseline isolates were collected for all patients, while supplementary isolates (colonising isolates from nasal swabs and serial blood isolates) were obtained when available.

Cohort B included 296 episodes of MRSA bacteraemia due to sequence type 239 that were collected at a single center in Australia between 1996 and 2008 to assess the impact of vancomycin MIC and vancomycin heteroresistance on 30 day mortality.[33]

Cohort C included data and sequences from the CAMERA2 trial (Combination Antibiotics for Methicillin Resistant *Staphylococcus aureus*) that was performed between 2015 and 2018 in Australia, New Zealand, Singapore and Israel, and randomised participants with MRSA bacteraemia to either monotherapy with vancomycin or daptomycin or combination therapy with vancomycin or daptomycin plus an antistaphylococcal beta-lactam (flucloxacillin, cloxacillin or cefazolin).[34] The primary outcome was a composite endpoint ascertained 90 days after randomisation and included mortality, persistent bacteraemia after 5 days, microbiological relapse (positive blood culture after a preceding negative culture), and microbiological treatment failure (positive sterile site culture 14 or more days after randomisation). In addition, the trial collected information on the day of death, duration of bacteraemia and day of relapse. Relevant clinical data assessed were comorbidities, severity scores, primary source and secondary foci, source control. Details of the antibiotic treatment were collected up to day 90.

Ethics approval

Human research committee approval was obtained at each site for the three cohorts. Written informed consent was collected for the prospective component of cohort A (VANESSA) and for cohort C (CAMERA2).

METHOD DETAILS

Phenotypic testing

Bacterial isolates were stored in glycerol broth at −80C, after confirmation of *S. aureus* using coagulase and DNase tests or Matrix-assisted laser desorption-ionisation time-of-flight (MALDI-TOF). Vancomycin minimum inhibitory concentration (MIC) was determined by E-test (bioMerieux) following manufacturer's instructions. An inoculum of 0.5 McFarland was used. Vancomycin E-test was performed for cohort A and B strains and for Nebraska Transposon Library mutants[61] with transposon insertions in genes that were genome-wide significant in the vancomycin MIC GWAS.

Whole-genome sequencing

Bacterial strains were subcultured twice from −80C glycerol stock. The Janus automated workstation (PerkinElmer) or manual extraction kits (Invitrogen PureLink genomic DNA kit or the Sigma GenElute kit) were used to extract genomic DNA from single colonies. DNA concentrations were normalised to 0.2 ng/ul for library preparation using Nextera XT DNA (Illumina). Whole-genome sequencing was performed on the Illumina MiSeq and NextSeq platforms. Reads were assembled using Shovill v1.1.0 (https://github.com/tseemann/shovill), a pipeline that streamlines and optimises Spades v3.14.1.[77] To assess the quality of the sequences, mean read depth, assembly metrics and the percentage of *S. aureus* reads using Kraken 2 v2.1.1,[78] were calculated. Genomes were discarded if read depth was <20, the fraction of *S. aureus* reads was <80% or the size of the assembly was >3.5 Mb or <2.5 Mb. We used Mash v2.3[79] to generate a distance matrix from the Shovill assemblies using 1,000 sketches. We used Abricate v1.0.1 (https://github.com/tseemann/abricate) to screen assemblies for resistance genes using the NCBI database (https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/).

QUANTIFICATION AND STATISTICAL ANALYSIS

Machine learning model to select predictors of mortality/treatment response

Using data from cohort A, we fitted a classification model of 30-day mortality of SAB to assess the predictive performance of non-genomic features including demographic and clinical data (age, Pitt bacteraemia score, Charlson comorbidity index, site of acquisition, gender), microbiological phenotypes (methicillin resistance, vancomycin MIC) and treatment response variables (duration of bacteraemia). We also tested the impact of antibiotic resistance more broadly by adding presence/absence of resistance genes based on Abricate (see above). The R package tidymodels was used to build a random forest model of SAB mortality. The model was trained on 80% of the data using stratified 10-fold cross-validation and hyperparameter tuning to optimise the area under the receiver operator characteristics curve (AUROC). Tuning of the random forest hyperparameters "mtry" (number of randomly selected predictors) and "min_n" (minimal node size) was done using a 5 × 5 grid, with range of hyperparameters determined empirically based on initial tuning using 20 automatically selected candidate combinations. Categorical predictors were converted into binary variables and all numeric predictors were centered and scaled. Zero variance predictors were removed. Downsampling was used to compensate for class inbalance (mortality rate 15%). The model was fitted on 100 replicated data-splits to generate estimates of the performance and variable importance as in.[65] To assess whether feature selection based on GWAS hists would improve the performance the mortality model, we performed a gene-burden GWAS as described below on the training dataset. We then train the models as above using following logarithmically transformed p value filters: no filter (0), 1, 1.5, 2, 2.5. The fitted model was then tested on the testing dataframe. This approach was designed to minimise overfitting, a critical issue when using filter methods for feature selection.[80] Because of its computational intensity, we performed this analysis in 10 data-splits replicates.

To confirm the output of the random forest model we fitted a logistic regression using the same dataset and variables using the logistic regression recipe in the R package parsnip, with default parameters.

Variant calling

Reads were mapped to reference genome *S. aureus* BPH2947 (GenBank: GCF_900620245.1) using Snippy v4.6.0 (https://github.com/tseemann/snippy). For each dataset, a core genome alignment including all positions with at least 90% non-gaps was constructed using Snippy-core v4.6.0, Goalign v0.3.4[81] and SNP-sites v2.5.1.[82] The alignment was then used to infer a maximum-likelihood phylogenetic tree with IQ-TREE v2.0.3[83] under a GTR+G4 model. Support was calculated using the ultrafast support feature and SH-aLRT test. For each variant at positions included in the core genome alignment, we computed the number of independent acquisitions across the phylogeny using HomoplasyFinder.[84]

Preparation of the bacterial genotype files

For each dataset, variants at core genome positions (minimum 90% of sequences with a coverage of at least 10 reads) were merged in a single variant call format (VCF) file using BCFtools v 1.10.2,[85] without merging multiallelic records. A catalog of mutations with predicted impact on protein function was generated by masking synonymous variants using the command bcftools view -e 'ANN[0] ~ "[^&]synonymous"'. Mutations with predicted loss of function (LOF) were extracted using the command bcftools view -i 'ANN[0] ~ "frameshift_variant" | ANN[0] ~ "stop_gained" | ANN[0] ~ "stop_lost" | ANN[0] ~ "start_lost"'.

Variant annotation

In addition to the variant annotation based on BPH2947 generated by SnpEff[86] (implemented in Snippy), we used Blast v 2.10.1 to determine homologs in reference genome *S. aureus* USA300 FPR3757 (GenBank: GCF_000013465.1) that were further annotated using the AureoWiki[87] and MicrobesOnline[88] databases. Blast hits were kept if alignment coverage was above 50% and identity was above 90%. When multiple hits were available, the hits with highest identity were selected.

Bacterial GWAS design

We ran bacterial GWAS to test the association between mutations with predicted impact on protein function (filtered as described above) with the phenotypes of interest, as continuous (vancomycin MIC, duration of bacteraemia) or binary traits (30 day mortality). For clinical outcomes, we analyzed the baseline sample (i.e., the first available sample per episode). For vancomycin MIC GWAS, we included more than one sample per episode when they were available. For duration of bacteraemia, potential transformations to achieve a normal distribution were assessed using the bestNormalize package in R.[89] For vancomycin MIC, normalisation was attempted using the RNOmni package in R.[90] The analysis was done on each cohort separately and on the combined dataset.

Bacterial GWAS pipeline

We used the factored spectrally transformed linear mixed models (FaST-LMM) as implemented in Pyseer v 1.1.2,[24] which allows to correct for population structure by adding a kinship matrix as a random effect. The kinship matrix was computed from a VCF file including all core-genome variants using Gemma, version 0.98.1.[26] When analysing the combined dataset, the model was adjusted for cohort membership ('–covariates' option). We assessed phenotype associations both with individual mutations and with protein-modifying and loss-of-function aggregated in BPH2947 genes, operons and gene ontology categories as implemented with the '–burden' option in Pyseer. Genes annotations were retrieved from the annotated BPH2947 assembly

(GenBank: GCF_900620245.1), operons were annotated based on FPR3757 annotation in MicrobesOnline[88] and gene ontology were obtained using PANNZER[91] to retrieve GO annotations for BPH2947. For the mutation GWAS, we excluded synonymous substitutions and mutations that were acquired only once across the phylogeny (based on homoplasy counts obtained using HomoplasyFinder[84]). For the gene-burden GWAS, only rare variants (i.e., those with a minim allele fraction below 1%) were considered, as suggested by the Pyseer authors.[51]

In addition to the preferred pipeline described above (LMM-based), we ran the two other models implemented in Pyseer: the fixed-effect model correcting for population structure using the first 4 components of the multidimensional scaling of the pairwise Mash distance matrix (Table S11) and the whole genome elastic net model (Table S10) with an alpha of 1 (lasso regression), 10-fold cross-validation and without correction for population structure.

Lineage effects and heritability

Heritability (SNP-heritability, h^2) - the proportion of phenotypic variance that is correlated with genotypic variance - was estimated by Pyseer from the linear mixed model using the kinship matrix computed in Gemma (see above). Lineages of cohort A isolates were inferred from the Mash distance matrix using the 'cmdscale' function in R. The association between the first 10 components and the bacterial GWAS outcomes was computed by running Pyseer with the '–lineage' option. Correlation between the 10 first components and internal nodes of the tree or major clonal complexes was calculated as performed in the R package bugwas (https://github.com/sgearle/bugwas/blob/master/bugwas/R/BUGWAS_functions.R#L962).[23]