https://doi.org/10.1038/s44259-025-00172-6

Gene copy-number features generalize better than SNPs for antimicrobial resistance prediction in *Staphylococcus aureus*

Bruna F. Fistarol[1] ✉, Joao D. Gervasio[1] & Gergely J. Szöllősi[1,2] ✉

Rapid prediction of antimicrobial resistance (AMR) from genome sequences is essential for timely therapy, yet models based on curated marker panels or core-genome Single Nucleotide Polymorphisms (SNPs) often fail to generalize to novel bacterial lineages. We evaluate AMR prediction in *Staphylococcus aureus* using pan-genome features that encode homologous gene copy number (including absence) and compare them to SNP-based models across six antibiotics and 4255 isolates. Gradient-boosted decision tree ensembles (XGBoost) trained on gene copy number achieve macro-averaged F1-scores of 0.925–0.988, surpassing SNP-based models (0.838–0.935). Under lineage-held-out evaluation, which withholds entire clades to mimic previously unseen lineages, gene-content models retain markedly higher performance (F1 = 0.875 and 0.904 across two split schemes), whereas SNP-based models degrade substantially (F1 = 0.557 and 0.638). Feature ablation indicates that predictive signal is distributed across many homologous gene families rather than dominated by a few markers, a structure consistent with stronger cross-lineage generalization. Because gene-content features can be robustly obtained even from low-coverage sequencing, this approach extends genome-based AMR prediction to real-world clinical and epidemiological datasets. Together, these results show that copy-number-based pan-genome representations provide a robust alternative to SNP-only approaches, particularly when models must generalize to lineages not represented in training data.

Staphylococcus aureus is a Gram-positive bacterium that can inhabit the human body under physiological conditions, typically as a commensal organism[1]. However, if this balance shifts, *S. aureus* can become pathogenic, leading to disease. Clinical presentations range from skin syndrome[2] to pneumonia[3]; notably, *S. aureus* is a major cause of sepsis[4] and a frequent pathogen in hospital settings [5].

Antimicrobial-resistant (AMR) *S. aureus* was observed not long after the first antimicrobial drug was clinically available[5]. During the last half of the twentieth century and the first decades of the twenty-first, the burden of drug-resistant *S. aureus* has become a global public-health concern: in 2019, methicillin-resistant *S. aureus* (MRSA) alone accounted for more than 100,000 deaths attributable to antimicrobial resistance, and *S. aureus* overall was associated with approximately 1.1 million deaths worldwide[6,7]. Consistent with this global burden, regional surveys from America [8], Africa[9,10], Europe[11] and Asia[12] show that up to 58% of all isolates are methicillin-resistant.

The traditional way for identifying the AMR profile of *S. aureus* involves culturing bacteria in the presence of the antibiotics of interest[13]. Although reliable, culture is time-consuming, delaying diagnosis and potentially leading to inappropriate therapy[8]. Faster molecular diagnostics detect known resistance determinants - specific SNPs or resistance genes - and are already in use for other pathogens such as *Mycobacterium tuberculosis*[14]. To our knowledge, however, there is no such test for *S. aureus*. Moreover, molecular assays are typically restricted to a predefined set of markers[14], which limits their predictive power and motivates the development of models that exploit broader genomic information[15]. Because, as with any other phenotype, machine-learning predictors are at risk of learning to predict resistance solely based on population (sample) structure

[1]Model-Based Evolutionary Genomics Unit, Okinawa Institute of Science and Technology, Okinawa, Japan. [2]Institute of Evolution, HUN-REN Center for Ecological Research, Budapest, Hungary. ✉e-mail: brunaffistarol@gmail.com; gergely.szollosi@oist.jp

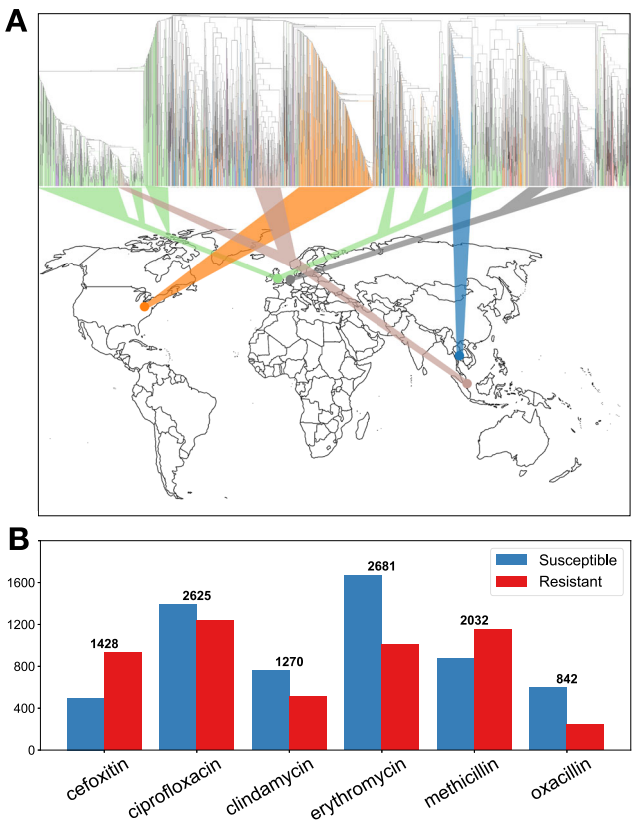

Fig. 1 | Global diversity and phenotype balance of the *S. aureus* dataset. **A** Maximum-likelihood core-genome phylogeny of 4255 *S. aureus* isolates; tip colours denote geographic region. **B** Counts of susceptible (S) and resistant (R) isolates per antibiotic. Bars give the number of genomes per phenotype for each drug; totals per antibiotic are indicated above bars.

and not actual phenotypic drivers, although evaluation should quantify performance on previously unseen lineages.

To address these limitations, machine-learning methods have been applied to predict AMR directly from genomic data, with the goal of models that generalize and remain accurate on previously unseen lineages. Core-genome k-mer representations can predict resistance even in the absence of known resistance genes[16], but may miss signals carried by accessory genes. Ren et al. combined single-nucleotide variants (SNVs), k-mer profiles and known resistance genes to train whole-genome models[17], achieving high accuracy relying on complete, well-annotated assemblies. More recently, large cross-species analyses have shown that sampling bias and population structure can confound machine-learning prediction of AMR, especially for SNP-based models[18]. These findings highlight the need for approaches that can generalize and remain accurate on previously unseen lineages.

Antibiotic resistance, however, can arise not only from specific resistance genes or SNPs but also from differences in gene content across lineages driven by horizontal gene transfer, gene gain/loss, and copy-number change. Because the bacterial pan-genome encompasses both shared (core) and variable (accessory) genes, it provides a more complete representation of genetic determinants of resistance. Analysing homologous gene copy number (including absence) therefore, captures duplication - and deletion-driven copy-number variation alongside horizontally acquired genes.

Building on this rationale, we model the pan-genome using gene content - specifically, homologous gene copy number variation (including absence) - as machine-learning features. This approach captures both core and accessory genome elements, allowing us to assess genomic variation that may be overlooked by models relying solely on known resistance genes or conserved *loci*. Previous studies have shown that gene content can predict phenotypes such as aerobicity in bacteria[19], supporting its broader predictive potential. Here, we test whether gene content can be used to predict antibiotic resistance in *S. aureus*, without depending entirely on predefined marker genes. We compare gene-content models to SNP-based models and assess generalization under lineage-held-out evaluation. We show that gene-content features yield higher accuracy and markedly better generalization to novel lineages than SNP-based features, indicating that variation in gene content at the level of homologous clusters likely driven by horizontal gene transfer play key roles in resistance.

Results

Gene-content features enable prediction of antibiotic resistance

We asked whether antimicrobial resistance (AMR) in *S. aureus* can be predicted from genome-wide features rather than curated marker panels of known genes and mutations used by tools such as Mykrobe, ARIBA, ResFinder, and CARD[20–23]. We assembled a dataset of 4255 isolates with laboratory susceptibility calls to six antibiotics (Fig. 1; Methods) and encoded three representations of genomic variation: (i) homologous gene copy-number per homologous gene cluster (including absence; "gene-content"), (ii) binary presence/absence (PA), and (iii) SNPs from core-gene alignments. We trained gradient-boosted decision tree ensembles (XGBoost) to classify resistant versus susceptible isolates using phylogenetically stratified splits (see Materials and Methods for details). Performance was summarized by macro-averaged precision, recall and F1-score on held-out data, with additional lineage-held-out evaluations to quantify generalization to previously unseen lineages (Figs. 2, 3; Supplementary Tables 1, 3, 5).

For homologous gene cluster copy number variation, macro-averaged F1-scores ranged from 0.925 for oxacillin to 0.988 for methicillin (Fig. 2A, Supplementary Table 1). PA representations performed similarly, although HGC yields better scores for four antibiotics (Supplementary Table 1). F1-scores dropped to ~0.5 when susceptibility labels were randomized, suggesting that distinct genomic signatures are associated with resistance phenotypes (Supplementary Fig. 1).

We next evaluated SNP-based models trained on core-gene alignments. Across the six antibiotics, macro-averaged F1-scores ranged from 0.838 for erythromycin to 0.935 for methicillin (Fig. 2B; Supplementary Table 1). Under a label-permutation control, F1-scores collapsed to ~0.5 (Supplementary Fig. 1), indicating that the classifier learned non-trivial genotype-phenotype signal from core-genome variation. Although SNP-based models underperformed gene-content models overall, they remained clearly above chance. The variation in performance across drugs suggests that the concentration of predictive signal in core versus accessory genomic regions differs by phenotype; we did not map SNP features to specific *loci* and therefore refrain from mechanistic interpretation.

Antibiotic resistance is predicted by a broad set of genomic features

To assess how concentrated the predictive signal is, we performed an importance-ranked, cumulative feature ablation. At each step we removed the current top-ranked features (by XGBoost feature importance) and re-evaluated performance, continuing until the macro-averaged F1-score fell below 0.5. We ran this procedure for both gene-content (copy number, including absence) and core-genome SNPs. For gene content, accuracy declined gradually: F1-scores remained above 0.5 even after removing more than 30% of the most influential clusters in both representations (Fig. 2C). This indicates a distributed - and partly redundant - predictive architecture, rather than reliance on a few dominant markers; such architectures are naturally more portable across lineages and are consistent with the stronger lineage-held-out performance of gene-content models. In contrast, for SNP-based predictions F1-scores declined rapidly with feature ablation, reaching 0.5 after removing 8–17% of SNPs.

We next examined the identity of high-importance homologous gene cluster copy number features. Among the top ten gene-content features per antibiotic, we found 38 unique Homologous Gene Clusters (HGCs), 14

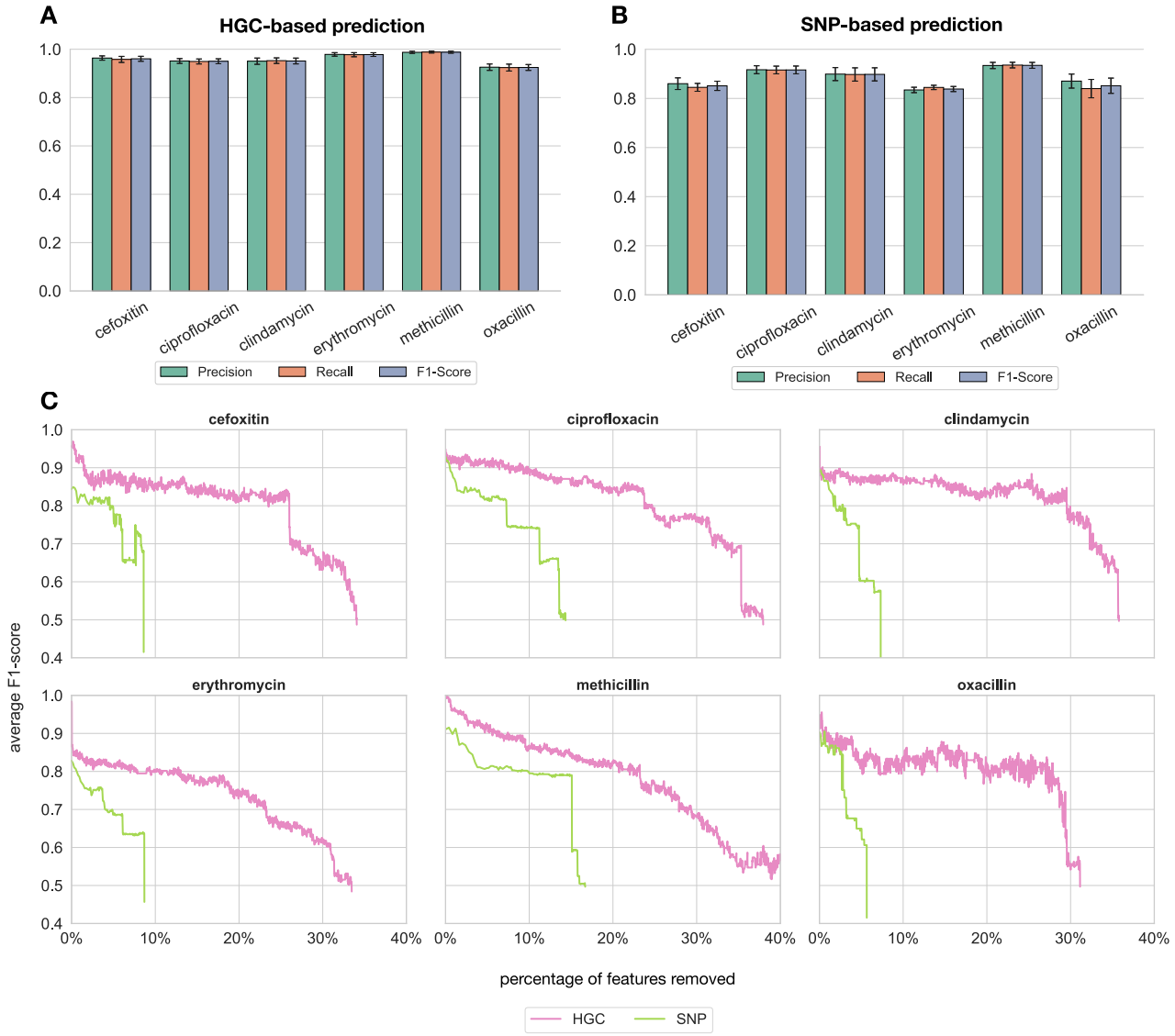

Fig. 2 | Model performance and feature ablation. A Mean and standard deviation bars of macro-averaged precision, recall, and F1-score across 10 train/test replicates of gradient-boosted decision trees (XGBoost) trained on homologous gene copy numbers (gene content, including absence). **B** Mean and standard deviation bars of macro-averaged precision, recall, and F1-score across 10 train/test replicates of SNP-based models using core-gene SNPs. Gene-content-based models consistently outperformed SNP-based models across antibiotics, with statistical significance evaluated using the Wilcoxon rank-sum test (Supplementary Table 2). **C** Feature ablation (importance-ranked, cumulative): macro-averaged F1-score as top-ranked features are sequentially removed until reaching 0.5. Pink, copy-number (gene-content) model; green, SNP-based model.

recurrent across at least two antibiotics. Using BLASTX against NCBI protein databases[24], ten HGCs matched gene families previously implicated in antimicrobial resistance (Table 1). The direction of association between copy number (or presence) and phenotype varied by antibiotic (Supplementary Fig. 2). These associations are descriptive rather than causal; together with the ablation results they indicate that numerous accessory *loci* contribute predictive signal, which plausibly supports better performance on previously unseen lineages.

Phylogenetic analysis reveals multiple independent gains of resistance

Trees constructed from core-gene SNPs reflect sequence-level divergence among strains, whereas trees based on gene content capture gene gain/loss dynamics (Supplementary Fig. 3). The two trees are broadly concordant, with differences mostly among closely related isolates in the gene-content tree (Fig. 3A, B; Supplementary Fig. 4). Most genomes were collected in the UK, followed by the USA, the Netherlands, Thailand and Singapore; isolates from the same country often cluster, indicating geographic structure (Fig. 1A).

We propagated resistant/susceptible states over each tree using unordered (Wagner) parsimony[25] and counted inferred transitions. Both the gene-content and core-genome trees showed partial clustering by phenotype (Supplementary Fig. 3). Visually inspecting the reconstructed states from gene-content, multiple and independent gains of resistance can be identified more clearly. (Supplementary Fig. 3).

Gene-content models outperform SNP-based models under lineage-held-out evaluation

We evaluated generalization under two lineage-held-out schemes that withhold clusters of related strains to mimic deployment to previously unseen lineages. In practice, we derived partitions from pairwise similarity: (i) gene-content similarity (presence/absence of homologous clusters) and

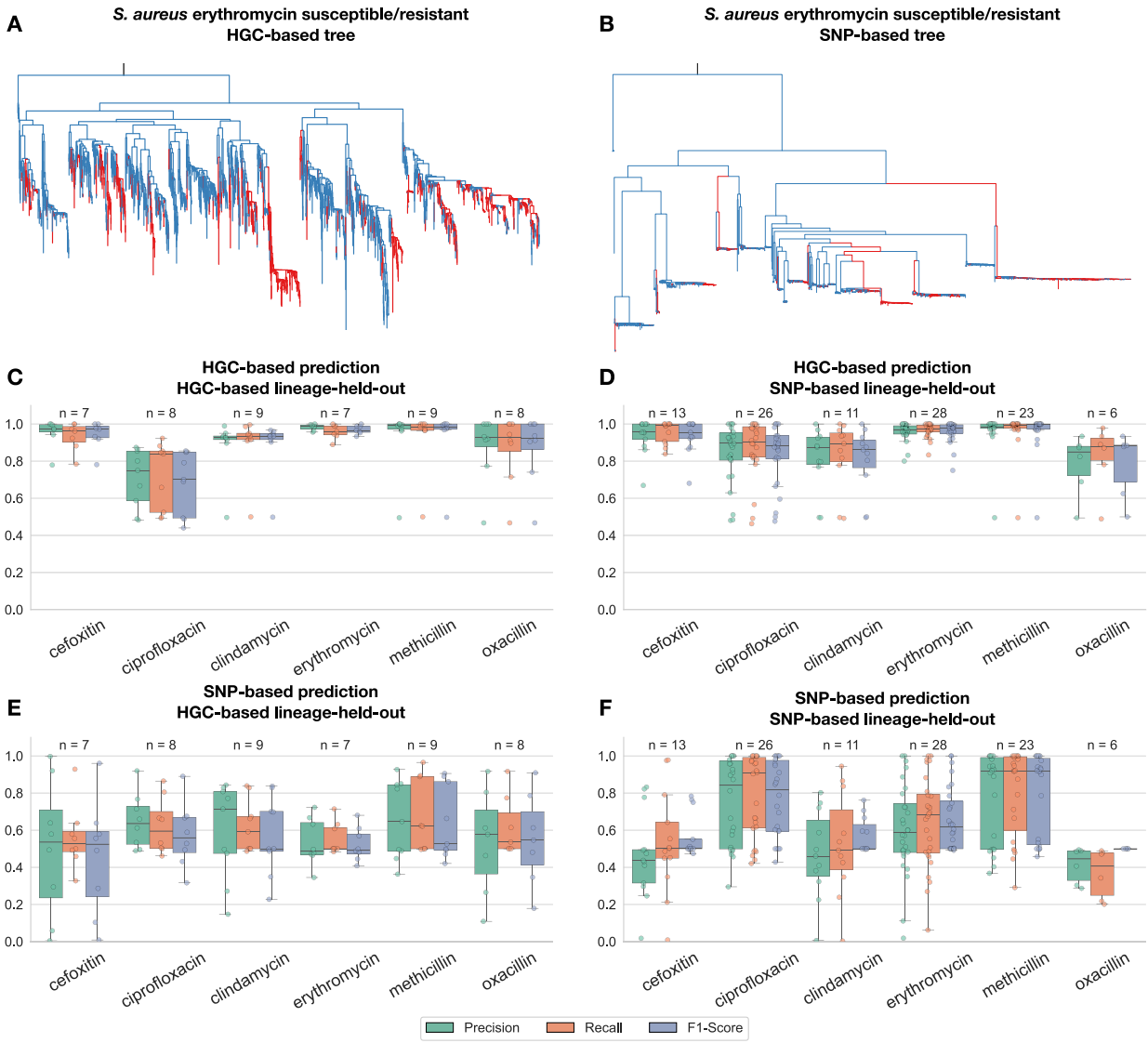

Fig. 3 | Lineage-held-out evaluation and cross-lineage performance. A Gene-content (copy number) tree coloured by erythromycin phenotype (resistant, red; susceptible, blue). **B** Core-gene SNP tree coloured by the same phenotype. **C** Distribution of precision, recall and F1-scores for XGBoost trained on gene-content features under lineage-held-out splits derived from homologous gene cluster PA pairwise similarity. **D** As in (**C**), but lineage-held-out splits derived from core sequence similarity. **E** Distribution of precision, recall, and F1-scores for SNP-based models under the splits used in (**C**). Except for ciprofloxacin, gene-content-based models consistently outperformed SNP-based models across antibiotics, with statistical significance evaluated using the Wilcoxon signed-rank test (Supplementary Table 4). **F** As in (**E**), but under the splits used in (**D**). Gene-content-based models consistently outperformed SNP-based models across antibiotics, with statistical significance evaluated using the Wilcoxon signed-rank test (Supplementary Table 6).

(ii) core-gene sequence similarity. Because the trees were constructed from the same similarity information, this approach is analogous to using clade-based splits while allowing more partitions and stable evaluation. Under both partition schemes, SNP-based models degraded substantially (mean macro F1 = 0.556 and 0.638, respectively; Fig. 3E, F; Supplementary Tables 3, 5). In contrast, gene-content models retained high performance (0.875 and 0.904; Fig. 3C, D; Supplementary Tables 3, 5). These results indicate that signals encoded by gene copy-number variation generalize more readily across lineages than core SNP patterns.

Taken together, the distributed nature of the predictive features, the recurrence of high-importance accessory *loci* and the stability of performance under lineage-held-out evaluation support the conclusion that copy-number-based pan-genome representations capture portable signals that generalize across *S. aureus* lineages, whereas SNP-only models are more sensitive to population structure and strain turnover.

Discussion

Several studies have shown that pan-genome features paired with tree-ensemble models can predict antimicrobial resistance (AMR). Moradigaravand et al.[26] reported good performance in *Escherichia coli* using allele-level features on a relatively modest dataset, and Do et al.[27] combined gene content with SNP to obtain competitive accuracy in *E. coli* and *Klebsiella pneumoniae*. However, these approaches adopt different operational definitions of the pan-genome: Moradigaravand et al. treat distinct alleles as separate features, whereas Do et al. restrict to homologous gene clusters present in > 65% of strains. Here we follow a broader definition similar to Tettelin[28], including all homologous groups (HGs) - core and accessory - and we encode *homologous gene copy number* (including absence). Direct numerical comparison across studies is not straightforward given species, feature, and split differences; nevertheless, for the overlapping antibiotic ciprofloxacin our accuracy is comparable despite focusing on *S. aureus*, a Gram-positive pathogen.

Table 1 | Functional annotations of the top gene clusters contributing to antibiotic resistance prediction

HG Cluster	Description based on Blast results	Comment	References
Cluster2814*	Transposes for iteration sequence-like element IS431mec	Associated with bleomycin resistance	https://doi.org/10.1007/s002530050783
Cluster3468*	GdpD - Glycerophosphodiester phosphodiesterase domain containing	Associated with vancomycin resistance in enterococci	https://doi.org/10.1056/NEJMoa1011138
Cluster687*	IS1182 family transposase	Associated with immune evasion and its deletion might be related to antimicrobial resistance	https://doi.org/10.1371/journal.pone.0187288
Cluster696*	Beta-lactam sensor/signal transducer MecR1	Associated with methicillin susceptibility	https://doi.org/10.1111/j.1440-1681.2007.04705.x
Cluster502*	Penicillin-binding protein PBP2a	Associated with penicillin resistance, presence is not enough to resist methicillin	https://doi.org/10.1002/iub.1289
Cluster5269	MaoC family dehydratase	Proposed as a target for drugs, to disrupt the cell membrane	https://doi.org/10.1016/j.bbapap.2025.141082
Cluster5022*	Replication/maintenance of protein RepL	Mobile Genetic Element associated with AMR	https://doi.org/10.3389/fmicb.2021.656306
Cluster5773	Hydroxymethylglutaryl-CoA synthase	Proposed as a target for drugs	https://doi.org/10.1073/pnas.84.21.7488, https://doi.org/10.1038/s42003-023-04639-y
Cluster3813	Class I SAM-dependent methyltransferase	Proposed as target, ribosomal protein	https://doi.org/10.1126/science.1200877
Cluster3529*	23S rRNA (adenine(2058)-N(6))-methyltransferase Erm(C)	Associated with AMR	https://doi.org/10.1128/jb.169.8.3857-3860.1987
Cluster3520*	23S rRNA (adenine(2058)-N(6))-methyltransferase Erm(A)	Associated with AMR	https://doi.org/10.1128/jb.169.8.3857-3860.1987
Cluster3154*	Aminoglycoside nucleotidyltransferase ANT(9)	Confer resistance to spectinomycin	https://doi.org/10.1128/spectrum.00620-23
Cluster583*	Tyrosine-type recombinase/integrase(9)	Mobile Genetic Element associated with AMR	https://doi.org/10.1371/journal.pone.0001315
Cluster125	SdrD adhesin	Not related to antimicrobial resistance, but as it is an adhesin there are some hypothesis that could be a target, to avoid the formation of biofilm	

Annotations are based on BlastX results against NCBI's nr database. Where available, links to known resistance functions or gene families are included. Genes associated with AMR are marked with "*".

For *S. aureus*, k-mer-based models[29,30] and core-genome feature sets[16] have achieved strong performance. A recent preprint extended allele-based features to *S. aureus* and compared deep learners with tree ensembles, finding that gradient boosting was typically superior[31]. Our results are consistent with that pattern: gradient-boosted decision trees applied to gene-content features perform well across six antibiotics.

An important caveat, highlighted by Yu et al.[18], is that AMR predictors can overfit population structure and perform poorly on novel lineages. We therefore evaluated models under lineage-held-out schemes that withhold groups of related strains, which approximate clade-level partitions. Under these conditions, SNP-based models degraded substantially, whereas gene-content models retained high F1. Two observations help explain this difference. First, feature ablation showed that compared to SNPs a larger fraction of high importance HGC copy number features needed to be removed before predictive accuracy declined, indicating a distributed (and partly redundant) predictive architecture rather than dependence on a few markers. Second, among high-importance HGCs we recovered a mixture of families previously implicated in AMR and additional accessory loci. Together, these patterns are consistent with transferable signals carried by the accessory genome - including copy-number changes and horizontally acquired elements - being less tied to specific lineages than core-SNP patterns[32,33].

When the analysis is restricted to variation among core homologous gene clusters, predictive accuracy declines substantially, still these core-based models generalize better than SNP models under lineage-held-out evaluation (Supplementary Fig. 5). This pattern supports the view that the accessory genome captures the determinants of resistance that are not entirely present in the conserved genomic background. It has been previously demonstrated, agreeing with our findings, that the use of the core genome alone incurs a lower recall metric[16]. From a healthcare perspective, misclassifying a resistant bacteria as susceptible is a much bigger issue than the other way around, therefore low recall rates should be kept to a minimum, which is the case for the models presented here.

A common concern is whether models "learn phylogeny" rather than resistance. In practice, these goals intersect. Gene content is correlated with clade structure, but lineage-held-out evaluation reduces direct leakage of lineage identifiers, and the stability of gene-content performance under such splits suggests that a substantial fraction of the predictive signal transfers beyond the training phylogeny.

This work has limitations. Copy-number estimates can be affected by assembly fragmentation and coverage variation; although we treat absence explicitly and use tree ensembles that down-weight noisy features, prospective assessment on read-depth-normalized copy-number variation calls would be informative. Our lineage-held-out schemes rely on clustering genomes by pairwise gene-content or sequence similarity; alternative partitions or external cohorts would provide additional validation. We did not benchmark other learners (e.g., LightGBM/CatBoost) or perform calibrated probability assessment, and we did not pursue causal attribution for specific HGCs. These are natural extensions.

Beyond its predictive accuracy, an important practical implication of this framework is that homologous gene copy number can be robustly inferred even from draft assemblies or low-coverage sequencing data. Unlike SNP-based approaches, which require high-quality alignments and high coverage for core regions, gene-content features tolerate assembly incompleteness and heterogeneous sequencing depth. This makes copy-number-based pan-genome modelling potentially highly suitable for clinical datasets, where sequencing conditions and quality can vary widely.

In summary, modelling the pan-genome with homologous gene copy number yields accurate AMR prediction in *S. aureus* and, critically, markedly better generalization to previously unseen lineages than SNP-based models. The distributed contribution of many HGCs, together with

recurrence of AMR-linked families, supports the view that accessory-genome variation provides portable predictive signal. Given its robustness under lineage-held-out evaluation, this representation is a plausible alternative for genomic susceptibility prediction and a practical starting point for discovering candidate markers for downstream validation.

Methods

Genome data acquisition

Antimicrobial resistance (AMR) phenotypes for *S. aureus* were downloaded from the Bacterial and Viral Bioinformatics Resource Center (BV-BRC) (accessed on January 21, 2025)[34]. We filtered for bacterial strains containing the phenotypes *susceptible* or *resistant* to six antibiotics with evidence from laboratory-derived experiments. We chose to work with cefoxitin, ciprofloxacin, clindamycin, erythromycin, methicillin and oxacillin, since they have more than 840 samples each and a balanced distribution of resistant and susceptible strains (Fig. 1B). Assembled genomes for each bacterial strain were retrieved from the BV-BRC FTP server.

Genome data processing and feature extraction

Coding DNA sequences (CDS) for the genome assemblies were predicted using Prodigal (v2.6.3)[35]. The predicted CDSs were clustered using CD-Hit (v4.8.1 - 2019)[36] with 80% of sequence similarity and word-length of 5, resulting in non-redundant clusters of similar genes across genomes. The gene content was represented as a two-dimensional matrix, where each column is a cluster and each row is a strain. Each entry in the matrix contains the number of copies of a given cluster in a genome, with a value of 0 assigned when the cluster is absent. The matrix of cluster presence/absence is constructed substituting each count different than 0 by 1.

Core genes were defined as those that were present in at least 99% of the strains with exactly one copy. These core gene clusters were individually aligned across strains, then concatenated to form a single multiple sequence alignment. This alignment was encoded into a feature matrix by converting each nucleotide or gap character into an integer: 'A'=0, 'C'=1, 'T'=2, 'G'=3, and '-'=4.

AMR prediction

The XGBoost classifier (v2.1.3)[37] was used to classify strains as either susceptible or resistant to each antibiotic. We tested three feature representations. The first two used gene content: one as a matrix of gene cluster counts, and the second as a binary presence/absence matrix, where nonzero counts were converted to one. The third approach used SNP data, represented by the integer-encoded alignment of concatenated core genes. For each antibiotic, strains were divided into non-overlapping training (80%) and testing (20%) sets to evaluate model performance in ten replicates.

To assess the robustness of the models, resistance phenotypes were randomly shuffled among genomes, and the classification procedure was repeated. A drop in performance to near-random levels was used to confirm the biological relevance of the original features. This process were repeated 10 times for each one of the three approaches described above.

To evaluate model generalization to unseen lineages, we simulated prediction on novel strains. We used clustalO (v1.2.4)[38] to cluster the aligned and concatenated set of core genes to create clusters with at most 200 sequences, which we can use as test data in both the gene content and the point-mutation approach. Given that every sequence in a cluster is distinguished enough from every other cluster, we treat them as novel strains. We filtered only novel strain sets with both resistant and susceptible strains. Regarding gene content, we calculated the pairwise distance of the feature matrix and used KMeans (implemented in scikit-learn[39]), to create 10 clusters of similar strains. Similarly to the sequence, each cluster is viewed as a novel strain set. Clusters representing more than 30% of the dataset or having only susceptible or resistant strains were excluded from evaluation.

Feature importance analysis

To assess the impact of individual gene clusters and SNPs, an iterative feature removal strategy was applied on gene content datasets at each iteration, the most important features (according to XGBoost's feature importance by weight) were removed, and the model was retrained. This process continued until macro F1-score dropped below 0.5. Analyses were repeated using both raw gene count and SNPs.

Phylogenetic tree construction

Two phylogenetic frameworks were constructed for each antibiotic: one based on concatenated core gene alignments, and another using binary presence/absence matrices of all gene cluster alignments were constructed using concatenated sequences from 313 core homologous gene cluster that were present in ≥99% of strains with exactly one copy, while gene cluster matrices were binarized for each genome. Trees were inferred using IQ-TREE (v2.4.0)[40], employing the GTR+F+G4 model for nucleotide alignments and the GTR2 model for binary data. These trees were used to explore evolutionary relationships and the distribution of resistance phenotypes. To decide where the root of the tree lies, we retrieved 10 genomes of *Staphylococcus pasteur* from BV-BRC, and processed them as it is describe on the Genome Data Processing. Cophylogenetic plots were generated using the `cophylo` function in the phytools R package (v2.4-4)[41], which aligns the tips of paired trees and provides a visual framework to assess topological congruence between gene-content and sequence-based phylogenies.

Data availability

Processed datasets and analysis scripts are available at https://doi.org/10.6084/m9.figshare.30230515.

Received: 6 October 2025; Accepted: 21 November 2025;
Published online: 16 December 2025

References

1. Wertheim, H. F. et al. The role of nasal carriage in staphylococcus aureus infections. *Lancet Infect. Dis.* **5**, 751–762 (2005).
2. Leung, A. K. C., Barankin, B. & Leong, K. F. Staphylococcal-scalded skin syndrome: evaluation, diagnosis, and management. *World J. Pediatrics* **14**, 116–120 (2018).
3. Paling, F. P. et al. Association of staphylococcus aureus colonization and pneumonia in the intensive care unit. *JAMA Netw. Open* **3**, e2012741 (2020).
4. Powers, M. E. & Wardenburg, J. B. Igniting the fire: Staphylococcus aureus virulence factors in the pathogenesis of sepsis. *PLoS Pathog.* **10**, e1003871 (2014).
5. Guo, Y., Song, G., Sun, M., Wang, J. & Wang, Y. Prevalence and therapies of antibiotic-resistance in staphylococcus aureus. *Front. Cell. Infect. Microbiol.* **10**, 107 (2020).
6. Murray, C. J. et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* **399**, 629–655 (2022).
7. Ikuta, K. S. et al. Global mortality associated with 33 bacterial pathogens in 2019: a systematic analysis. *Lancet* **400**, 1517–1535 (2022).
8. Hassoun, A., Linden, P. K. & Friedman, B. Incidence, prevalence, and management of mrsa bacteremia across patient populations-a review of recent developments in mrsa management and treatment. *Crit. Care* **21**, https://doi.org/10.1186/s13054-017-1801-3 (2017).
9. Ambachew, A., Gebrecherkos, T. & Ayalew, G. Prevalence and clindamycin resistance profile of staphylococcus aureus and associated factors among patients attending the university of gondar comprehensive specialized hospital, gondar, northwest ethiopia. *Interdiscip. Perspect. Infect. Dis.* **2022**, 1–10 (2022).
10. Ezeh, C. K., Eze, C. N., Dibua, M. E. U. & Emencheta, S. C. A meta-analysis on the prevalence of resistance of staphylococcus aureus to different antibiotics in Nigeria. *Antimicrob. Resist. Infect. Control* **12**, https://doi.org/10.1186/s13756-023-01243-x (2023).
11. Köck, R. et al. Methicillin-resistant staphylococcus aureus (mrsa): burden of disease and control challenges in europe. *Eurosurveillance* **15**, https://doi.org/10.2807/ese.15.41.19688-en (2010).

12. Hosaka, Y. et al. Surveillance of multi-drug resistance phenotypes in staphylococcus aureus in Japan and correlation with whole-genome sequence findings. *J. Hospital Infect.* **123**, 34–42 (2022).
13. Brown, D. F. J. et al. Guidelines for the laboratory diagnosis and susceptibility testing of methicillin-resistant staphylococcus aureus (mrsa). *J. Antimicrobial Chemother.* **56**, 1000–1018 (2005).
14. Nguyen, T. N. A., Anton-Le Berre, V., Bañuls, A.-L. & Nguyen, T. V. A. Molecular diagnosis of drug-resistant tuberculosis; a literature review. *Front. Microbiol.***10**, https://doi.org/10.3389/fmicb.2019.00794 (2019).
15. Xie, F. et al. Large-scale genomic analysis reveals significant role of insertion sequences in antimicrobial resistance of acinetobacter baumannii. *mBio***16**, https://doi.org/10.1128/mbio.02852-24 (2025).
16. Nguyen, M., Olson, R., Shukla, M., VanOeffelen, M. & Davis, J. J. Predicting antimicrobial resistance using conserved genes. *PLOS Computational Biol.* **16**, e1008319, https://doi.org/10.1371/journal.pcbi.1008319 (2020).
17. Ren, Y. et al. Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. *Bioinformatics* **38**, 325–334 (2022).
18. Yu, Y., Wheeler, N. E. & Barquist, L. Biased sampling confounds machine learning prediction of antimicrobial resistance. *bioRXiv* https://doi.org/10.1101/2025.01.07.631773 (2025).
19. Davín, A. A. et al. A geological timescale for bacterial evolution and oxygen adaptation. *Science* **388**, eadp1853 (2025).
20. Bradley, P. et al. Rapid antibiotic-resistance predictions from genome sequence data for extitStaphylococcus aureus and extitMycobacterium tuberculosis. *Nat. Commun.* **6**, 10063 (2015).
21. Hunt, M. et al. Ariba: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb. Genomics* **3**, e000131 (2017).
22. Zankari, E. et al. Identification of acquired antimicrobial resistance genes. *J. Antimicrobial Chemother.* **67**, 2640–2644 (2012).
23. Alcock, B. P. et al. Card 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **48**, D517–D525 (2020).
24. Altschul, S. F. et al. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402, https://doi.org/10.1093/nar/25.17.3389 (1997).
25. Eck, R. V. & Dayhoff, M. O. *Atlas of protein sequence and structure, 1966* (National Biomedical Research Foundation, Silver Spring, MD, 1966).
26. Moradigaravand, D. et al. Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data. *PLOS Computational Biol.* **14**, e1006258 (2018).
27. Do, V. H. et al. Panka: Leveraging population pangenome to predict antibiotic resistance. *iScience* **27**, 110623 (2024).
28. Tettelin, H. et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". *Proc. Natl Acad. Sci.* **102**, 13950–13955 (2005).
29. Davis, J. J. et al. Antimicrobial resistance prediction in patric and rast. *Sci. Rep.* **6**, https://doi.org/10.1038/srep27930 (2016).
30. Wang, S. et al. A practical approach for predicting antimicrobial phenotype resistance in staphylococcus aureus through machine learning analysis of genome data. *Front. Microbiol.* **13**, https://doi.org/10.3389/fmicb.2022.841289 (2022).
31. Ghosh, J., Taneja, J. & Kant, R. Decoding host-pathogen interactions instaphylococcus aureus: Insights into allelic variation and antimicrobial resistance prediction using artificial intelligence and machine learning based approaches. *bioRXiv*.https://doi.org/10.1101/2025.02.18.638850 (2025).
32. Dubois, V., Debreyer, C., Litvak, S., Quentin, C. & Parissi, V. A new in vitro strand transfer assay for monitoring bacterial class 1 integron recombinase inti1 activity. *PLoS ONE* **2**, e1315 (2007).
33. Asante, J. et al. Genomic analysis of antibiotic-resistant staphylococcus epidermidis isolates from clinical sources in the KwaZulu-Natal province, South Africa. *Front. Microbiol.* **12**, https://doi.org/10.3389/fmicb.2021.656306 (2021).
34. Olson, R. D. et al. Introducing the bacterial and viral bioinformatics resource center (bv-brc): a resource combining PATRIC, IRD and VIPR. *Nucl. Acids Res.* **51**, D678–D689 (2022).
35. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformat.* **11**, 119 (2010).
36. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
37. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, 2016).
38. Sievers, Y. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
39. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
40. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2014).
41. Revel, L. J. phytools 2.0: an updated r ecosystem for phylogenetic comparative methods (and other things). *PeerJ.* **12**, e16505 (2024).

Acknowledgements

We are grateful for the help and support provided by the Scientific Computing and Data Analysis section of Core Facilities at OIST.

Author contributions

B.F., J.G. and G.J.Sz conceived the study, B.F. and J.G. carried out bioinformatics analyses, B.F., J.G. and G.J.S.z wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44259-025-00172-6.

Correspondence and requests for materials should be addressed to Bruna F. Fistarol or Gergely J. Szöllősi.