# Genome co-adaptation and the evolution of methicillin resistant *Staphylococcus aureus* (MRSA)

Seungwon Ko[1*], Elizabeth A. Cummins[1*], William Monteith[1,2], Samuel K. Sheppard[1]

[1]**Ineos Oxford Institute for Antimicrobial Research, Department of Biology, University of Oxford, Oxford, United Kingdom**

[2]**Department of Biology, University of Bath, Bath, United Kingdom**

***These authors contributed equally.**

**Corresponding author:** samuel.sheppard@biology.ox.ac.uk

## Abstract

**Background:** Antimicrobial resistance (AMR) in bacterial pathogens is a major threat to global health, rendering standard treatments ineffective and increasing the risk of severe infection or death. Resistance is often conferred by genes that are transferred horizontally among species and strains. However, for many bacteria, little is known about the genetic variation that potentiates resistance gene acquisition and accommodates acquired genes in the coadapted recipient genome.

**Results:** Here we introduce a new bioinformatics genome-wide association study approach with Guided Omission of Linkage Disequilibrium (GOLD-GWAS) that masks covarying alleles explained by coinheritance and genome proximity to reveal genes where covarying sequence likely represents functional linkage between loci, consistent with epistasis. Analysing 806 *Staphylococcus aureus* isolate genomes, including methicillin-resistant (MRSA) and methicillin-susceptible (MSSA) strains, we identify genes that covary with the presence of the acquired staphylococcal cassette chromosome *mec* (SCC*mec*) that houses the *mecA* resistance gene.

**Conclusions:** Uncovering known and new gene-gene associations, we demonstrate how resistance can involve genetic coalitions beyond the well-known AMR genes. Understanding how genome change, here extrinsic resistance cassettes, are integrated within coadapted bacterial genomes is an important step towards mitigating AMR evolution by identifying novel genetic targets for risk prediction, diagnosis and therapy.

Keywords: Epistasis; Methicillin resistant *Staphylococcus aureus* (MRSA); staphylococcal cassette chromosome *mec* (SCC*mec*); genome-wide association study (GWAS); co-variation.

## Background

Bacterial populations can exhibit considerable trait variation. In some cases, closely related strains of the same species can vary from harmless commensals to important pathogens (1–3). Understanding how mutation and horizontal gene transfer (HGT) give rise to divergent phenotypes is a major focus in microbiology, with modern genomics linking gene function to trait variation in natural populations (4). Among the most problematic bacterial phenotypes is antimicrobial resistance (AMR), with projections that treatment failure will be associated with an estimated 50 million deaths worldwide by 2050 (5). Many of the genes, alleles, and polymorphisms underlying resistance are known. However, the function of genes depends on genomic context, and there is increasing evidence that the evolution of AMR may involve multiple genes, even for well characterized mechanisms (6–9).

In some well-characterized instances, a single gene or nucleotide polymorphism can increase resistance (10,11). Given this capacity for genomic plasticity, it can be tempting to view genes as modular elements that can be 'plugged in' or 'switched on' to confer resistance. However, this view oversimplifies the reality, as genes interact within genomes to confer phenotypes. This can be a basic additive effect where genes independently contribute to a phenotype, or non-additive effects where the effect of one gene or allele depends on another. Phenotypes conferred by non-additive gene effects can be either synergistic, where the sum of gene effects exceeds their individual contributions, or epistatic, where there is functionally interdependence and the action of one gene depends upon the other(s). Both can be associated with AMR (12–17).

Among the best-known AMR pathogens is methicillin-resistant *Staphylococcus aureus* (MRSA), a major cause of hospital acquired infections (18–20). The principal genetic driver of methicillin resistance is the *mecA* gene, which codes for a modified penicillin-binding protein (PBP2a) (21). This gene is commonly transported among *Staphylococcus* species in the staphylococcal cassette chromosome *mec* (SCC*mec*) (22–24), integrating into the recipient chromosome via sequence-specific recombination with *rlmH* (*orfX*) (25–27) (Supplementary Figure 1). The distribution of SCC*mec* in staphylococcal populations is a balance of forces favouring acquisition (28,29) and the fitness cost to the recipient strains (30–33). Understanding these opposing forces and the genetic consequences of SCC*mec* acquisition in natural populations requires large-scale comparative genomics.

Genome-wide association studies (GWAS) have been used to understand the genetics underlying phenotype variation in bacteria for over a decade (34). Bacterial GWAS has been applied to identify genetic determinants of AMR in pathogens, improving understanding of the emergence and spread of well-known resistance determinants, including SCC*mec* in staphylococci (35–41). While these studies may also highlight genes and alleles that covary with known AMR genes, this has seldom been an explicit aim. With increasing appreciation of the importance of gene-gene interactions in integrated bacterial genomes, there has been more emphasis upon understanding functional gene networks and identifying putative epistasis in population genomic datasets (42,43).

Genome-wide covariation analysis is conceptually simple. Essentially it involves, comparing genomes and identifying alleles that are found together, i.e. when 'A' is present at one locus 'B' is typically present at another. To make these findings biologically relevant it is important to compare the covariation signal to that which is expected by chance. Most co-variation is the result of co-inheritance or physical proximity on in the genome, not epistasis, and this can dominate signals in basic GWAS models. Here, we address this with a novel method that enhances traditional bacterial GWAS for co-variation analyses. Specifically, our approach (GOLD-GWAS) incorporates quantification of genome-wide linkage disequilibrium (LD) decay. That is to say, as the physical distance increases between two alleles they are less likely to have been coinherited because recombination will be more likely to have shuffled segments of DNA, which occurs by HGT in most bacteria. Having masked covariation resulting from LD, our approach identifies genes that covary with *SCCmec* in *S.*

*aureus,* potentially indicative of potentiating or compensatory genome change. Using this approach, we are able to characterize the genomic landscape of AMR coadaptation, infer functional significant genes and identify putative epistatic loci.

## Results

### Testing genome masking with simulated data

To evaluate the performance of the Guided-Omission of LD GWAS (GOLD-GWAS) approach, we tested whether the method could detect an artificially generated covarying site within simulated bacterial genomes. We simulated a dataset comprising 1,000 genomes of 1 Mbp each, generated with a clonal genealogy under a coalescent model with recombination rate of R = 0.01 and a site-specific mutation rate of θ = 0.001 (44). An artificial site of covariation was constructed by identifying a polymorphism with minor allele frequence (MAF) > 0.2 within a 10 kbp range of the 100 kbp position. The polymorphism at 93,696 bp fulfilled these criteria and eleven artificial covarying sites between 600,000 bp and 600,100 bp at 10 bp intervals were created to covary with the 93,696 bp site. These artificial covarying sites were generated independently for each site with a 95% probability, to avoid numerical errors arising from perfect covariation. GOLD-GWAS was then applied to these simulated data with artificial covarying sites. From the GOLD-GWAS output, mapping k-mers to a reference genome demonstrated that our method effectively masked the target region with no k-mers mapping between 83,696 and 103,696 bp. Masking GWAS also identified accurately the artificial covarying site at 600,000 – 600,100 bp with 10 significantly associated k-mers present with -log(p-values) > 5.13 (Supplementary Figure 2).

### LD in *S. aureus* genomes declines to and equilibrates after 8790 bp

A total of 806 *S. aureus* whole genomes were chosen from the NCBI reference sequence database to represent both SCC*mec*-positive (n = 426) and SCC*mec*-negative isolates (n = 380). All assemblies were aligned to the NCTC 8325 reference genome before variant calling and 10% of all identified polymorphisms were randomly sampled for LD analysis. The average $R^2$ value (a measure of LD) was determined for the data set, which fell to approximately 0.149 after 100,000 bp (Figure 1A). The gradient of the log-transformed graph of LD decay was calculated to be -0.0726 (Figure 1B). From these values, the overall LD range of *S. aureus* in our dataset was estimated to be 8790 bp, calculated as the intercept of the average $R^2$ value and the fitted $R^2$ logarithmic bp decay. Our estimate corresponds with a previous study of LD decay in *S. aureus* where no LD was reported between SNPs with distance greater than 10 kbp (45).

### GOLD-GWAS identifies covariation associated with known epistatic sites in *S. aureus*

The GOLD-GWAS method was tested using biological data to identify genomic covariation among known epistatic sites. The sites selected for this test were: (i) *divIVA,* which play critical roles in cell division and polarity (46); (ii) *secA2,* which encodes an accessory secretion ATPase (SecA2) involved in the secretion of major autolysins such as p60 (CwhA) and MurA (NamA) both of which are key factors in bacterial cell wall remodelling (47) (48). The *divIVA* loci were chosen as the target region to evaluate whether the GOLD-GWAS pipeline could detect genomic co-variation within *secA2,* linked to *divIVA*.

Observed p-values from GOLD-GWAS closely followed expected values from a theoretical $\chi^2$-distribution up to approximately -log(p-value) = 2, beyond which points deviated from the null (Figure 2A). The sigmoidal distribution lacked clear ‘shelves’, confirming that there was no poorly controlled confounding population structure. Mapping k-mers to a reference genome (Figure 2B) revealed that k-mers significantly associated with *divIVA* carriage were found in the previously characterised epistatic region of *secA2* and notably *murC*, another

gene involved in peptidoglycan synthesis (47). Removal of the *divIVA* locus and it's LD regions ensured demonstrated that these genes emerged as independent associations - separate of the effects of LD. This is consistent with coadaptation with *divIVA* as previously described (46–48) and the utility of GOLD-GWAS to uncover biologically significant covariation in bacterial genomes. Furthermore, matching k-mers to the genes they are found in (Figure 2C) revealed that many of the top GOLD-GWAS hits had related functions, broadly linked to peptidoglycan biosynthesis (Table 1). Moreover, the genomic location of the top ranked genes (-log(p-value) > 200, beta > 0.5) indicated clearly that the significant associations with *divIVA* have not arisen solely from genomic proximity to the target region (Figure 2D).

**Masking SCC*mec* and the conserved LD region detects the SCC*mec* insertion site**

In this study, the GOLD-GWAS pipeline begins with the computational masking of SCC*mec* and associated LD regions. LD regions are parameterised from the SCC*mec*-positive isolate genomes and observed p-values from GWAS were well controlled at low p-values (p < 1.5) consistent with the expected distribution under the null hypothesis. The even distribution of p-values suggested good control for potential confounding effects of population structure (Figure 3A). Mapping k-mers to the *S. aureus* SCC*mec*-positive reference genome showed that the masking effectively removed significant k-mers within the SCC*mec* and associated LD region (Figure 3Ci). A similar result was observed when k-mers were mapped to the SCC*mec*-negative reference genome apart from a single significant k-mer covering the attB site within *rlmH* (Figure 3Di). As the neighbouring sequences of *rlmH* are highly conserved, the computational masking covers the entire LD region surrounding the SCC*mec* region in both SCC*mec*-positive and SCC*mec*-negative isolates (49) (Figures 3Ci, Di). However, the attB site within *rlmH* is not conserved as it serves as the recombination site for SCC*mec* integration and therefore is present exclusively in SCC*mec*-negative isolates. Hence, the significant -log(p-value) of 9.5 observed for the k-mer that covers attB due to the strong negative correlation of this site with SCC*mec* carriage (Figure 3Di). This association effected the ranking of genes significantly associated with SCC*mec* carriage due to the extremely high (74.2) likelihood ratio test (LRT) -log(p-value) of *rlmH* that outranked covariation signals in other genes (Figure 3B). The genomic locations of the genes covered by significantly associated k-mers in both reference genomes demonstrated that these genes are distributed throughout the genome. The mean distance between genes was 187 kbp, with maximum and minimum distances of 782 kbp and 1.7 kbp respectively (standard deviation = 235 kbp). None of these genes were positioned within 10 kbp of *SCCmec* except for *rlmH*. Excluding the attB site in *rlmH,* this spatial separation indicated that significant hits are highly unlikely to result solely by physical proximity (Figures 3Cii, Dii).

**Masking SCC*mec* and *rlmH* detects genes that covary with SCC*mec* carriage**

To improve the detection of sites that covary with SCC*mec,* beyond the attB site in *rlmH*, we masked SCC*mec* and *rlmH* alongside their respective LD regions using GOLD-GWAS. Again, observed p-values followed the expected distribution under the null hypothesis at low p-values (p < 1.5) with the even distribution confirming adequate control for population structure (Figure 4A). Mapping k-mers to the SCC*mec*-positive reference genome demonstrated effective removal of significant k-mers within SCC*mec* and associated LD regions (Figure 4Ci). Only 2 of the 172,224 total k-mers fell within the masked region with extremely low non-significant -log p-values of 0.04 and 0.33. Mapping k-mers to the SCC*mec*-negative reference genome also showed that no significantly covarying k-mers mapped to the *rlmH* gene and associated LD regions and only 3 of 172,224 k-mers were present with non-significant -log p-values of 0.24, 0.026, and 0.065 (Figure 4Di). The significantly covarying k-mers mapped to 16 genes (Table 2). The *sen* gene, which produces Staphylococcal enterotoxin type N, displayed the highest LRT value (-log(p-value) = 14.2, Figure 4B), while *splE*, which produces serine protease SplE, exhibited the largest effect size (beta = 0.87, Figure 4B). The chromosomal distribution of the 16 top-ranked

genes in both SCC*mec*-positive and -negative reference genomes showed that the covarying genes were widely distributed across the genome. The mean distance between genes was 200 kbp, with maximum and minimum distances of 782 kbp and 1.7 kbp respectively (standard deviation = 239 kbp). No k-mers were located within 10 kbp of the insertion site by *rlmH* (Figures 4Cii, Dii).

K-mers that were significantly associated with SCC*mec* carriage, either more commonly present or more commonly absent, were mapped to a phylogeny of the *S. aureus* isolates (Figure 5). Fisher's exact test was performed for each of the 38 k-mers to determine the nature of the correlation with SCC*mec* carriage (Supplementary Table 1). The majority of k-mers (87.5%, n = 35/40) exhibited positive correlations (OR >1), with 16 of the 35 showing highly significant covariation ($p < 0.001$). The only k-mer with a highly significant negative correlation (OR < 1, $p = 0.0005$) corresponded to *icd.* The higher prevalence of k-mers that are positively associated with SCCmec provides little evidence that there are elements that block SCC*mec* acquisition. Conversely, there is strong evidence that certain genes promote the integration or retention of SCC*mec.* This is consistent with possible potentiation of acquisition or compensatory change to accommodate SCC*mec* in the recipient genome.

## Discussion

There has been extensive work identifying genetic determinants of antimicrobial resistance in several bacterial species. While robust genotype-phenotype prediction is possible in some cases, accuracy is often less than 100% (50–53). Furthermore, where AMR is predicted from genome data, this is usually inferred for resistance above a certain threshold based on clinically relevant minimum inhibitory concentrations of antimicrobials used in laboratory bacterial growth assays. Of course, breaking phenotypes down into 'yes / no' metadata is an oversimplification. Even for relatively simple AMR phenotypes, where a single gene or allele may be the principal agent, there are complex genomic interactions that govern if expression occurs and to what extent (17,54–57).

Differentiating the nature of multiple gene interactions, such as synergistic additive effects and epistasis, is challenging from genome data alone. One of the main reasons for this is that the genomic signature of covariation is the same for coinheritance and LD, and functional co-adaption - here epistasis. Bench-marking the GOLD-GWAS approach with simulated data, we showed that it was possible to identify covarying alleles that were not explained by LD (Supp. Fig 1). Furthermore, extending the method to real data of *S. aureus* isolate genomes identified multiple significant associations among loci known to be functionally linked. In particular, *divIVA* and other genes involved in the peptidoglycan synthesis (*murC, yfhO, asp2, clfB, atl, bshB2*), including variants of genes encoding enzymes for uridine diphosphate N-acetylglucosamine (UDP-*GlcNAc*) biosynthesis (Table 1) (23,58–62).

Having demonstrated utility for detecting known adaptive covariation signatures, GOLD-GWAS was applied to identify genes that covary with SCC*mec* (Figure 3). Unsurprisingly, sequence variation at the known insertion site (*rlmH*) was strongly associated with SCC*mec* carriage, consistent with an established role in mobile genetic element integration and excision (25,49,63). Extending analyses with or without masking *rlmH,* detected k-mers that mapped to other genes, including some annotated as encoding hypothetical proteins (Figure 3B and Figure 4B). The most significant covariation with SCC*mec* was observed with the staphylococcal enterotoxin N gene (*sen*) (Figure 4B). This is associated with toxic shock-like syndromes and food poisoning (64) caused by staphylococci and epidemiological studies show that enterotoxin genes can be enriched in MRSA populations (65). This suggests a possible correlation between toxin production and antibiotic resistance, supported by the GOLD-GWAS analyses.

Among the other genes with sequence that co-varied with SCC*mec* were those linked directly linked to antibiotic resistance. The response regulator component of the glycopeptide resistance associated two-component system, *graR*, regulates susceptibility to vancomycin and daptomycin and is linked to cell wall stress responses (66–70). The relatively large average effect size reported here (beta = 0.8385) is likely attributable to the significant down-regulation of *mecA* in the absence of *graRS* (70). Furthermore, the *norC* gene encodes a multidrug efflux pump (71) and the von Willebrand factor binding protein (vWbp) have been linked to biofilm formation which improves survival rate under various stress conditions (72,73). These functions highlight the complex network of antibiotic resistance mechanisms that may interact with methicillin resistance conferred by *mecA*.

Quantifying and differentiating potential synergistic effects among virulence and resistance-associated loci is extremely important for understanding pathogen evolution. While direct evidence of epistasis requires phenotypic validation, we propose three adaptive scenarios that may explain putative SCC*mec* gene interactions identified in this study. First, specific loci could have a direct role in SCC*mec* insertion or excision by facilitating or hindering integration at the attB site. In this scenario, the presence or absence of specific regulatory or structural elements near the integration site may determine how efficiently SCC*mec* is inserted or excised from the bacterial chromosome. Second, some genes may undergo compensatory mutations following SCC*mec* insertion. In this case, mutations may balance the disruptive effects of integrating the large SCC*mec* element, restoring cell function and preserving bacterial fitness while enabling resistance (74–76). This scenario is consistent with the prevalence of positive correlations among k-mers that covary with SCC*mec*. Finally, there may be potentiation or functional synergy with methicillin resistance. Here, regulatory and virulence pathways, such as those govern by the GraRS two-component system, may interact with beta-lactam resistance, enhancing SCC*mec* effects by boosting resistance or altering the cell envelope to further reduce antibiotic susceptibility (77).

It is important to recognise the methodological limitations of GOLD-GWAS. Most importantly, computational genome analyses only provide statistical inference based upon sequence covariation. However, epistasis is measured phenotypically. Therefore, while our approach can identify candidates for further study, mechanistic understanding requires functional microbiological validation. There are also limitations of the analysis methodology. First, there is an emphasis on covariation with a target locus, rather than an all against all comparison. While this gives enhanced computational efficiency over some existing methods (42,78–81) and targeted gene identification, it inevitably overlooks multi-locus interactions that are not related to the gene(s) under investigation. Second, while masking LD regions helps to minimise confounding signals of coinheritance it may also mask potential functional interactions among genes in physical proximity.

## Conclusions

Integrated genome covariation analyses, such as that presented here, are an important step towards improved understanding of pathogen evolution. Context-free gene-centric approaches often fall short of accurate functional inference – such as AMR. With ever larger genome datasets and deeper understanding of gene function it is possible to move closer to accurate genotype-phenotype maps that account for gene network interactions.

## Methods

### Isolate genomes

All available *S. aureus* whole genome sequences were retrieved from the National Center for Biotechnology Institution (NCBI) reference sequence database with higher than 95% completeness or x30 coverage (n = 1,001, accessed 3rd October 2023). SCC*mec* regions

were identified and typed using a custom database (github.com/Sheppard-Lab/sccmec_classifier) with minimap2 (v2.28) (82). Genomes with untypable SCC*mec* regions were removed from further analysis (n = 195). The final data set of 806 isolates consisted of 426 SCC*mec*-positive and 380 SCC*mec*-negative isolates. Isolate details including accession numbers are included in Supplementary Table 2. Core genome alignments were generated using PIRATE (v1.0.5) (83). A core genome maximum-likelihood phylogeny was created using RaxML (v8.2.12) (84) with GTRGAMMA as a substitution model. ClonalFrameML (v1.12) was used to account for recombination (85). All phylogenies were visualised using MicroReact (86)(87).

**Simulation data**

Simulation data was generated using SimBac (44). Specifically, we simulated bacterial genomes with a recombination rate of R = 0.01 and a site-specific mutation rate of θ = 0.001 to generate 1,000 genomes, each spanning 1 Mbp. To introduce artificial genomic covariation, we identified a polymorphism located within 100 ±10 kbp with a minor allele frequency exceeding 0.2. Then, we generated eleven artificial covarying sites between 600,000 bp and 600,100 bp at 10 bp intervals. The nucleotide at each covarying site was determined based on the nucleotide present at the identified polymorphic site (at ~100 kbp). The covariation was simulated with 95% accuracy to avoid perfect correlation, which could result in p-values of 0 and lead to numerical instability. The target region for masking was specified with start and end coordinates of 85896 bp and 101896 bp respectively.

**Computational masking**

GOLD-GWAS is conceptually simple, involving: (i) estimating LD around a locus, here SCC*mec*; (ii) building a directory of unitigs and masking those within the LD region and the target region; (iii) conducting GWAS (Supplementary Figure 3). Linkage disequilibrium (LD) values were calculated using a custom pipeline (github.com/Sheppard-Lab/GOLD-GWAS/blob/main/run_ld_calculation.sh) Briefly, assemblies were aligned to the *S. aureus* NCTC 8325 reference genome using BWA (v0.7.17) (88)**.** Alignments were then converted to BAM format, merged, sorted, and indexed with SAMtools (v1.16.1) (89). Variant-calling was performed by BCFtools (v1.14) (89). All single nucleotide polymophisms (SNPs) were counted, and 10% were randomly sampled to generate the variant call format (VCF) input file for LD analysis with PLINK (v1.9) (90). Only SNPs no more than 200,000 bps apart, using an LD window of 2,900 Mb for all SNPs, were analysed. LD decay plots were created with a custom R (v4.2.2) script (github.com/Sheppard-Lab/GOLD-GWAS/blob/main/plotting_lddecay. r).

The LD threshold was calculated as the intersection of the fitted $R^2$ logarithmic bp decay and the average LD value, for distances exceeding 100 kbp, which are considered sufficiently large to represent a non-LD region (Figure 6A) (91). Due to potential regional variations in LD across the genome, a conservative cut-off of 10,000 bp was chosen for the masking procedure. The SCC*mec* target regions for masking were specified using the direct repeat and reverse complement inverse repeat ($DR_{SCC}$-R and $IR_{SCC}$-L, Supplementary Figure 1B) sequences as the start and end coordinates respectively, to capture all SCC*mec* types. From here, SCC*mec* and associated LD regions in SCC*mec*-positive isolates and LD regions in SCC*mec*-negative isolates were detected using the custom database option in ABRicate (v1.0.0)(92) with 80% coverage and 80% sequence identity thresholds (Figure 6B). Unitigs were generated from genome assemblies using the call mode of unitig-caller (v1.3.0) (93) with the pyseer flag and a k-mer size of 31. Unitigs were then mapped to the SCC*mec* (or *divIVA)* and associated LD region sequences using BWA (v0.7.17). Unitigs with over 80% sequence identity and 80% coverage were removed from further analysis (Figure 6C).

**Genome-wide association studies**

GWAS was performed on the filtered set of unitigs using pyseer (v1.3.11) (94) with the linear mixed model flag. Manhattan plots were generated using Phandango (87) with JKD6159 (NCBI Reference Sequence: NC_017338.2) and HOU1444-VR (NCBI Reference Sequence: NZ_CP012593.1) as reference genomes for SCC*mec*-positive and SCC*mec*-negative strains, respectively. Gene hits were annotated using BWA (v0.7.17) with a minimum match length of 8 bp (88,94). The threshold for significance was calculated as 0.05 divided by the number of unique unitigs. From the hits that exceeded this threshold, the results were classified into two groups: those with -log(p-values) greater or less than the 3rd quantile (Q3) + 1.5x the interquartile range (IQR). To mitigate lineage effects (identified by poor chi-square values), we applied a minimum minor allele frequency threshold of 0.05%. To enhance biological relevance and reduce background noise, only genes with known names located in the vicinity of the identified k-mers were retained, while those with uncharacterised protein structures or functions were excluded from further analysis due to lack of information. To determine the nature of k-mer associations, Fisher's exact test was performed using Python scipy.stats module fishers_exact with multiple testing correction implemented by statsmodel.stats.multitest. All scripts used for this analysis are available at github.com/Sheppard-Lab/GOLD-GWAS.

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Availability of data and materials

All data analysed during the current study are included in this published article and its supplementary information files**.** Custom Python scripts used to perform analyses are available at github.com/Sheppard-Lab/GOLD-GWAS and zenodo (DOI: 10.5281/zenodo.15451691) unless otherwise stated in the text.

### Competing interests

The authors declare that they have no competing interests.

### Funding

SK is a self-funded PhD student and EAC is supported by a BBSRC grant (BB/W020602/1), awarded to SKS.

**Consent to publish**

Not applicable.

**Authors' contributions**

SK: methodology, software, validation, formal analysis, writing. EAC: methodology, writing, visualisation. WM: conceptualisation, methodology. SKS: conceptualisation, writing.

**References**

1. Espadinha D, Sobral RG, Mendes CI, Méric G, Sheppard SK, Carriço JA, et al. Distinct phenotypic and genomic signatures underlie contrasting pathogenic potential of staphylococcus epidermidis clonal lineages. Front Microbiol. 2019 Aug 27;10(AUG):463823.

2. Méric G, Mageiros L, Pensar J, Laabei M, Yahara K, Pascoe B, et al. Disease-associated genotypes of the commensal skin bacterium Staphylococcus epidermidis. Nature Communications 2018 9:1. 2018 Nov 28;9(1):1–11.

3. Sheppard SK. Strain wars and the evolution of opportunistic pathogens. Curr Opin Microbiol. 2022 Jun 1;67:102138.

4. Kobras CM, Fenton AK, Sheppard SK. Next-generation microbiology: from comparative genomics to gene function. Genome Biology 2021 22:1. 2021 Apr 29;22(1):1–16.

5. Naghavi M, Vollset SE, Ikuta KS, Swetschinski LR, Gray AP, Wool EE, et al. Global burden of bacterial antimicrobial resistance 1990–2021: a systematic analysis with forecasts to 2050. The Lancet. 2024 Sep 28;404(10459):1199–226.

6. Munita JM, Arias CA. Mechanisms of Antibiotic Resistance. Kudva IT, Zhang Q, editors. Microbiol Spectr. 2016 Mar 25;4(2):10.1128/microbiolspec.VMBF-0016–2015.

7. Darby EM, Trampari E, Siasat P, Gaya MS, Alav I, Webber MA, et al. Molecular mechanisms of antibiotic resistance revisited. Nature Reviews Microbiology 2022 21:5. 2022 Nov 21;21(5):280–95.

8. Kobras CM, Monteith W, Somerville S, Delaney JM, Khan I, Brimble C, et al. Loss of Pde1 function acts as an evolutionary gateway to penicillin resistance in Streptococcus pneumoniae. Proc Natl Acad Sci U S A. 2023 Oct 10;120(41):e2308029120.

9. Papkou A, Hedge J, Kapel N, Young B, MacLean RC. Efflux pump activity potentiates the evolution of antibiotic resistance across S. aureus isolates. Nature Communications 2020 11:1. 2020 Aug 7;11(1):1–15.

10. Sköld O. Sulfonamide resistance: Mechanisms and trends. Drug Resistance Updates. 2000;3(3):155–60.

11. Sproston EL, Wimalarathna HML, Sheppard SK. Trends in fluoroquinolone resistance in campylobacter. Microb Genom. 2018 Aug 1;4(8):e000198.

12. Trubenová B, Roizman D, Rolff J, Regoes RR. Modeling Polygenic Antibiotic Resistance Evolution in Biofilms. Front Microbiol. 2022 Jul 7;13.

13. Martinez JL, Baquero F. Mutation frequencies and antibiotic resistance. Antimicrob Agents Chemother. 2000 Jul;44(7):1771–7.

14. Schubert B, Maddamsetti R, Nyman J, Farhat MR, Marks DS. Genome-wide discovery of epistatic loci affecting antibiotic resistance in Neisseria gonorrhoeae using evolutionary couplings. Nat Microbiol. 2018 Feb 1;4(2):328.

15. Igler C, Rolff J, Regoes R. Multi-step vs. single-step resistance evolution under different drugs, pharmacokinetics, and treatment regimens. Elife. 2021 May 1;10:e64116.

16. Wong A. Epistasis and the evolution of antimicrobial resistance. Front Microbiol. 2017 Feb 17;8(FEB):235611.

17. DelaFuente J, Diaz-Colunga J, Sanchez A, San Millan A. Global epistasis in plasmid-mediated antimicrobial resistance. Mol Syst Biol. 2024 Apr 2;20(4):311–20.

18. Klein E, Smith DL, Laxminarayan R. Hospitalizations and deaths caused by methicillin-resistant Staphylococcus aureus, United States, 1999-2005. Emerg Infect Dis. 2007;13(12):1840–6.

19. Kennedy AD, Otto M, Braughton KR, Whitney AR, Chen L, Mathema B, et al. Epidemic community-associated methicillin-resistant Staphylococcus aureus: Recent clonal expansion and diversification. Proc Natl Acad Sci U S A. 2008 Jan 29;105(4):1327–32.

20. van Hal SJ, Jensen SO, Vaska VL, Espedido BA, Paterson DL, Gosbell IB. Predictors of mortality in staphylococcus aureus bacteremia. Clin Microbiol Rev. 2012 Apr;25(2):362–86.

21. Peacock SJ, Paterson GK. Mechanisms of methicillin resistance in Staphylococcus aureus. Annu Rev Biochem. 2015 Jun 2;84(Volume 84, 2015):577–601.

22. Katayama Y, Ito T, Hiramatsu K. A new class of genetic element, staphylococcus cassette chromosome mec, encodes methicillin resistance in Staphylococcus aureus. Antimicrob Agents Chemother. 2000;44(6):1549–55.

23. Büttner FM, Zoll S, Nega M, Götz F, Stehle T. Structure-Function Analysis of Staphylococcus aureus Amidase Reveals the Determinants of Peptidoglycan Recognition and Cleavage. Journal of Biological Chemistry. 2014 Apr 18;289(16):11083–94.

24. Méric G, Miragaia M, De Been M, Yahara K, Pascoe B, Mageiros L, et al. Ecological Overlap and Horizontal Gene Transfer in Staphylococcus aureus and Staphylococcus epidermidis. Genome Biol Evol. 2015 May 1;7(5):1313–28.

25. Uehara Y. Current Status of Staphylococcal Cassette Chromosome mec (SCCmec). Antibiotics. 2022 Jan 1;11(1).

26. McClure JA, Conly JM, Zhang K. Characterizing a Novel Staphylococcal Cassette Chromosome mec with a Composite Structure from a Clinical Strain of Staphylococcus hominis, C34847. Antimicrob Agents Chemother. 2021 Nov 1;65(11):e00777-21.

27. Jansen WTM, Beitsma MM, Koeman CJ, Van Wamel WJB, Verhoef J, Fluit AC. Novel Mobile Variants of Staphylococcal Cassette

Chromosome mec in Staphylococcus aureus. Antimicrob Agents Chemother. 2006 Jun;50(6):2072.

28. Hanssen AM, Ericson Sollid JU. SCCmec in staphylococci: genes on the move. FEMS Immunol Med Microbiol. 2006 Feb 1;46(1):8–20.

29. Chambers HF, DeLeo FR. Waves of resistance: Staphylococcus aureus in the antibiotic era. Nature Reviews Microbiology 2009 7:9. 2009;7(9):629–41.

30. Durhan E, Korcan SE, Altindis M, Konuk M. Fitness and competitive growth comparison of methicillin resistant and methicillin susceptible Staphylococcus aureus colonies. Microb Pathog. 2017 May 1;106:69–75.

31. Jia K, Zhu H, Wang J, Qin X, Wang X, Dong Q. Fitness cost and compensatory evolution of penicillin-induced resistant Staphylococcus aureus. Food Research International. 2025 Feb 1;203:115841.

32. Noto MJ, Fox PM, Archer GL. Spontaneous deletion of the methicillin resistance determinant, mecA, partially compensates for the fitness cost associated with high-level vancomycin resistance in Staphylococcus aureus. Antimicrob Agents Chemother. 2008 Apr;52(4):1221–9.

33. Lee SM, Ender M, Adhikari R, Smith JMB, Berger-Bächi B, Cook GM. Fitness cost of staphylococcal cassette chromosome mec in methicillin-resistant Staphylococcus aureus by way of continuous culture. Antimicrob Agents Chemother. 2007 Apr;51(4):1497–9.

34. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, et al. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. Proc Natl Acad Sci U S A. 2013 Jul 16;110(29):11923–7.

35. Mosquera-Rendón J, Moreno-Herrera CX, Robledo J, Hurtado-Páez U. Genome-Wide Association Studies (GWAS) Approaches for the Detection of Genetic Variants Associated with Antibiotic Resistance: A Systematic Review. Microorganisms. 2023 Dec 1;11(12).

36. Farhat MR, Freschi L, Calderon R, Ioerger T, Snyder M, Meehan CJ, et al. GWAS for quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions. Nature Communications 2019 10:1. 2019 May 13;10(1):1–11.

37. Chen PE, Shapiro BJ. The advent of genome-wide association studies for bacteria. Curr Opin Microbiol. 2015 Jun 1;25:17–24.

38. Power RA, Parkhill J, De Oliveira T. Microbial genome-wide association studies: lessons from human GWAS. Nat Rev Genet. 2016 Dec 13;18(1):41–50.

39. Jaillard M, Lima L, Tournoud M, Mahé P, van Belkum A, Lacroix V, et al. A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. PLoS Genet. 2018 Nov 1;14(11).

40. Post V, Harris LG, Morgenstern M, Mageiros L, Hitchings MD, Méric G, et al. Comparative genomics study of Staphylococcus epidermidis isolates from orthopedic-device-related infections correlated with patient outcome. J Clin Microbiol. 2017 Oct 1;55(10):3089–103.

41. Post V, Pascoe B, Mourkas E, Calland JK, Hitchings MD, Erichsen C, et al. Multiple hypervirulent methicillin-sensitive Staphylococcus aureus lineages contribute towards poor patient outcomes in orthopedic device-related infections. medRxiv. 2022 Oct 22;2022.10.21.22280349.

42. Pensar J, Puranen S, Arnold B, MacAlasdair N, Kuronen J, Tonkin-Hill G, et al. Genome-wide epistasis and co-selection study using mutual information. Nucleic Acids Res. 2019 Oct 10;47(18):e112–e112.

43. Taylor AJ, Yahara K, Pascoe B, Ko S, Mageiros L, Mourkas E, et al. Epistasis, core-genome disharmony, and adaptation in recombining bacteria. mBio. 2024 Jun 1;15(6).

44. Brown T, Didelot X, Wilson DJ, De Maio N. SimBac: simulation of whole bacterial genomes with homologous recombination. Microb Genom. 2016 Jan 19;2(1):e000044.

45. Takuno S, Kado T, Sugino RP, Nakhleh L, Innan H. Population Genomics in Bacteria: A Case Study of Staphylococcus aureus. Mol Biol Evol. 2012 Feb 1;29(2):797–809.

46. Chaudhary R, Mishra S, Kota S, Misra H. Molecular interactions and their predictive roles in cell pole determination in bacteria. Crit Rev Microbiol. 2021;47(2):141–61.

47. Lenz LL, Mohammadi S, Geissler A, Portnoy DA. SecA2-dependent secretion of autolytic enzymes promotes Listeria monocytogenes pathogenesis. Proc Natl Acad Sci U S A. 2003 Oct 14;100(21):12432–7.

48. Kaval KG, Rismondo J, Halbedel S. A function of DivIVA in Listeria monocytogenes division site selection. Mol Microbiol. 2014 Nov 1;94(3):637–54.

49. Wang L, Safo M, Archer GL. Characterization of DNA sequences required for the CcrAB-mediated integration of staphylococcal cassette

chromosome mec, a Staphylococcus aureus genomic island. J Bacteriol. 2012 Jan;194(2):486–98.

50. Yee R, Bard JD, Simner PJ. The Genotype-to-Phenotype Dilemma: How Should Laboratories Approach Discordant Susceptibility Results? J Clin Microbiol. 2021 Jun 1;59(6):e00138-20.

51. Neuert S, Nair S, Day MR, Doumith M, Ashton PM, Mellor KC, et al. Prediction of phenotypic antimicrobial resistance profiles from whole genome sequences of non-typhoidal Salmonella enterica. Front Microbiol. 2018 Mar 27;9(MAR):592.

52. Terrance Walker G, Quan J, Higgins SG, Toraskar N, Chang W, Saeed A, et al. Predicting Antibiotic Resistance in Gram-Negative Bacilli from Resistance Genes. Antimicrob Agents Chemother. 2019 Apr 1;63(4).

53. Brochier C, Philippe H, Moreira D. The evolutionary history of ribosomal protein RpS14: Horizontal gene transfer at the heart of the ribosome. Trends in Genetics. 2000 Dec 1;16(12):529–33.

54. Ogunlana L, Kaur D, Shaw LP, Jangir P, Walsh T, Uphoff S, et al. Regulatory fine-tuning of mcr-1 increases bacterial fitness and stabilises antibiotic resistance in agricultural settings. The ISME Journal 2023 17:11. 2023 Sep 18;17(11):2058–69.

55. Jangir PK, Yang Q, Shaw LP, Caballero JD, Ogunlana L, Wheatley R, et al. Pre-existing chromosomal polymorphisms in pathogenic E. coli potentiate the evolution of resistance to a last-resort antibiotic. Elife. 2022;11.

56. Rubin DHF, Ma KC, Westervelt KA, Hullahalli K, Waldor MK, Grad YH. CanB is a metabolic mediator of antibiotic resistance in Neisseria gonorrhoeae. Nat Microbiol. 2023 Jan 1;8(1):28–39.

57. Yokoyama M, Stevens E, Laabei M, Bacon L, Heesom K, Bayliss S, et al. Epistasis analysis uncovers hidden antibiotic resistance-associated fitness costs hampering the evolution of MRSA. Genome Biol. 2018 Jul 18;19(1):1–12.

58. Gaballa A, Newton GL, Antelmann H, Parsonage D, Upton H, Rawat M, et al. Biosynthesis and functions of bacillithiol, a major low-molecular-weight thiol in Bacilli. Proc Natl Acad Sci U S A. 2010 Apr 6;107(14):6482.

59. Thomer L, Becker S, Emolo C, Quach A, Kim HK, Rauch S, et al. N-Acetylglucosaminylation of Serine-Aspartate Repeat Proteins Promotes Staphylococcus aureus Bloodstream Infection. J Biol Chem. 2013 Feb 7;289(6):3478.

60. Seepersaud R, Bensing BA, Yen YT, Sullam PM. The Accessory Sec Protein Asp2 Modulates GlcNAc Deposition onto the Serine-Rich Repeat Glycoprotein GspB. J Bacteriol. 2012 Oct;194(20):5564.

61. Rismondo J, Percy MG, Gründling A. Discovery of genes required for lipoteichoic acid glycosylation predicts two distinct mechanisms for wall teichoic acid glycosylation. J Biol Chem. 2018 Mar 2;293(9):3293.

62. Weber F, Motzkus NA, Brandl L, Möhler M, Alempijevic A, Jäschke A. Identification and in vitro characterization of UDP-GlcNAc-RNA cap-modifying and decapping enzymes. Nucleic Acids Res. 2024 Jun 10;52(10):5438–50.

63. Chongtrakool P, Ito T, Ma XX, Kondo Y, Trakulsomboon S, Tiensasitorn C, et al. Staphylococcal cassette chromosome mec (SCCmec) typing of methicillin-resistant Staphylococcus aureus strains isolated in 11 Asian countries: A proposal for a new nomenclature for SCCmec elements. Antimicrob Agents Chemother. 2006 Mar;50(3):1001–12.

64. Balaban N, Rasooly A. Staphylococcal enterotoxins. Int J Food Microbiol. 2000 Oct 1;61(1):1–10.

65. Ortega E, Abriouel H, Lucas R, Gálvez A. Multiple Roles of Staphylococcus aureus Enterotoxins: Pathogenicity, Superantigenic Activity, and Correlation to Antibiotic Resistance. Toxins (Basel). 2010 Aug;2(8):2117.

66. Doddangoudar VC, Boost M V., Tsang DNC, O'Donoghue MM. Tracking changes in the vraSR and graSR two component regulatory systems during the development and loss of vancomycin non-susceptibility in a clinical isolate. Clin Microbiol Infect. 2011;17(8):1268–72.

67. Cafiso V, Bertuccio T, Spina D, Purrello S, Campanile F, Di Pietro C, et al. Modulating activity of vancomycin and daptomycin on the expression of autolysis cell-wall turnover and membrane charge genes in hVISA and VISA strains. PLoS One. 2012 Jan 9;7(1).

68. Mensa B, Howell GL, Scott R, DeGrado WF. Comparative mechanistic studies of brilacidin, daptomycin, and the antimicrobial peptide LL16. Antimicrob Agents Chemother. 2014;58(9):5136–45.

69. Müller A, Grein F, Otto A, Gries K, Orlov D, Zarubaev V, et al. Differential daptomycin resistance development in Staphylococcus aureus strains with active and mutated gra regulatory systems. Int J Med Microbiol. 2018 Apr 1;308(3):335–48.

70. Chen L, Wang Z, Xu T, Ge H, Zhou F, Zhu X, et al. The Role of graRS in Regulating Virulence and Antimicrobial Resistance in Methicillin-Resistant Staphylococcus aureus. Front Microbiol. 2021 Aug 16;12:727104.

71. Truong-Bolduc QC, Strahilevitz J, Hooper DC. NorC, a New Efflux Pump Regulated by MgrA of Staphylococcus aureus. Antimicrob Agents Chemother. 2006 Mar;50(3):1104.

72. Wang D, Wang L, Liu Q, Zhao Y. Virulence factors in biofilm formation and therapeutic strategies for Staphylococcus aureus: A review. Animals and Zoonoses. 2024 Nov 28;

73. Evans DCS, Khamas AB, Payne-Dwyer A, Wollman AJM, Rasmussen KS, Klitgaard JK, et al. Cooperation between coagulase and von willebrand factor binding protein in Staphylococcus aureus fibrin pseudocapsule formation. Biofilm. 2024 Dec 1;8:100233.

74. Andersson DI, Hughes D. Antibiotic resistance and its cost: is it possible to reverse resistance? Nature Reviews Microbiology 2010 8:4. 2010 Mar 8;8(4):260–71.

75. Mwangi MM, Shang WW, Zhou Y, Sieradzki K, De Lencastre H, Richardson P, et al. Tracking the in vivo evolution of multidrug resistance in Staphylococcus aureus by whole-genome sequencing. Proc Natl Acad Sci U S A. 2007 May 29;104(22):9451–6.

76. Reynolds MG. Compensatory evolution in rifampin-resistant Escherichia coli. Genetics. 2000;156(4):1471–81.

77. Jiang JH, Cameron DR, Nethercott C, Aires-De-Sousa M, Peleg AY. Virulence attributes of successful methicillin-resistant Staphylococcus aureus lineages. Clin Microbiol Rev. 2023 Dec 20;36(4).

78. Puranen S, Pesonen M, Pensar J, Xu YY, Lees JA, Bentley SD, et al. SuperDCA for genome-wide epistasis analysis. Microb Genom. 2018 Jun 1;4(6):e000184.

79. Skwark MJ, Croucher NJ, Puranen S, Chewapreecha C, Pesonen M, Xu YY, et al. Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis. PLoS Genet. 2017 Feb 1;13(2).

80. Kuronen J, Horsfield ST, Pöntinen AK, Mallawaarachchi S, Arredondo-Alonso S, Thorpe H, et al. Pangenome-spanning epistasis and coselection analysis via de Bruijn graphs. Genome Res. 2024 Jul 1;34(7):1081.

81. Mallawaarachchi S, Tonkin-Hill G, Pöntinen AK, Calland JK, Gladstone RA, Arredondo-Alonso S, et al. Detecting co-selection through excess linkage disequilibrium in bacterial genomes. NAR Genom Bioinform. 2024 Apr 4;6(2):61.

82. Li H. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics. 2018 Sep 15;34(18):3094–100.

83. Bayliss SC, Thorpe HA, Coyle NM, Sheppard SK, Feil EJ. PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. Gigascience. 2019 Oct 1;8(10):1–9.

84. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014 May 1;30(9):1312–3.

85. Didelot X, Wilson DJ. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. PLoS Comput Biol. 2015;11(2):e1004041.

86. Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C, et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. Microb Genom. 2016 Nov 1;2(11):e000093.

87. Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, Harris SR. Phandango: an interactive viewer for bacterial population genomics. Bioinformatics. 2018 Jan 1;34(2):292–3.

88. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754–60.

89. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021 Feb 1;10(2).

90. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007;81(3):559.

91. van Heel DA, Hunt K, Greco L, Wijmenga C. Genetics in coeliac disease. Best Pract Res Clin Gastroenterol. 2005;19(3):323–39.

92. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, et al. Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. Antimicrob Agents Chemother. 2019;63(11):e00483-19.

93. Holley G, Melsted P. Bifrost: Highly parallel construction and indexing of colored and compacted de Bruijn graphs. Genome Biol. 2020 Sep 17;21(1):1–20.

94. Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. Bioinformatics. 2018 Dec 15;34(24):4310–2.

## Figures and Tables

**Table 1: Genes containing k-mers significantly associated with the presence of *divIVA* in *S. aureus* genomes (-log(p-value) > 200), ordered by descending maximum p-value.**

| Gene | Description |
|---|---|
| *secA2* | Accessory Sec translocase SecA2 |
| *murC* | UDP-N-acetylmuramate-L-alanine ligase |
| *yfhO* | Lipoteichoic acid-specific glycosyltransferase YfhO |
| *asp2* | Accessory Sec system protein Asp2 |
| *clfB* | MSCRAMM adhesin clumping factor ClfB |
| *atl* | Bifunctional autolysin; cleaves peptidoglycan during cell division |
| *argJ* | Bifunctional glutamate N-acetyltransferase/amino acid acetyl-transferase |
| *coa* | Staphylocoagulase; cleaves fibrinogen |
| *nsaS* | Nisin susceptibility-associated sensor histidine kinase NsaS |
| *bshB2* | Bacillithiol biosynthesis deacetylase BshB2 |
| *dapA* | 4-hydroxy-tetrahydrodipicolinate synthase |
| *gap* | Type I glyceraldehyde-3-phosphate dehydrogenase |
| *vraD* | Peptide resistance ABC transporter ATPase subunit VraD |
| *rnr* | Ribonuclease R |

MSCRAMM: Microbial Surface Components Recognizing Adhesive Matrix Molecule.

**Table 2: Genes containing k-mers significantly associated with the presence of SCC*mec* in *S. aureus* genomes (-log(p-value) > 200), ordered by descending maximum p-value.**

| Gene | Description |
|---|---|
| *sen* | Staphylococcal enterotoxin type N |
| *graR* | Response regulator transcription factor GraR/ApsR |
| *vwb* | von Willebrand factor binding protein Vwb |
| *fmhC* | FemA/FemB family glycyltransferase FmhC |
| *seo* | Staphylococcal enterotoxin type O |
| *scpA* | Cysteine protease staphopain A |
| *purB* | Adenylosuccinate lyase |
| *scb* | Staphylococcal complement inhibitor SCIN-B |
| *nrdI* | Class Ib ribonucleoside-diphosphate reductase assembly flavoprotein NrdI |
| *rpsP* | 30S ribosomal protein S16 |
| *ffh* | Signal recognition particle protein |
| *ald* | Alanine dehydrogenase |
| *lip1* | YSIRK domain-containing triacylglycerol lipase Lip1 |
| *norC* | Multidrug efflux MFS transporter NorC |
| *splE* | Serine protease SplE |
| *icd* | Isocitrate dehydrogenase NADP-dependent isocitrate dehydrogenase |

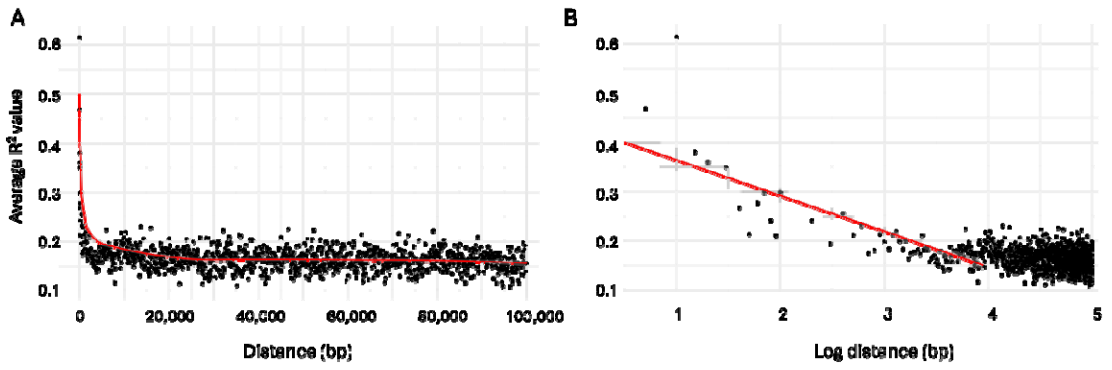



**Figure 1: Distribution of linkage disequilibrium values ($R^2$) across the genome as a function of the distance between single nucleotide polymorphism pairs in *S. aureus*.** (A) Relationship between the linear distance between two SNPs and the average $R^2$ value where each point is the $R^2$ value for an individual distance category. Red line represents a polynomial trend line. (B) Relationship between the logarithm of the distance between two SNPs and the average $R^2$ value. Red line represents the linear trend line fitted after excluding data points with $R^2 < 0.2$.

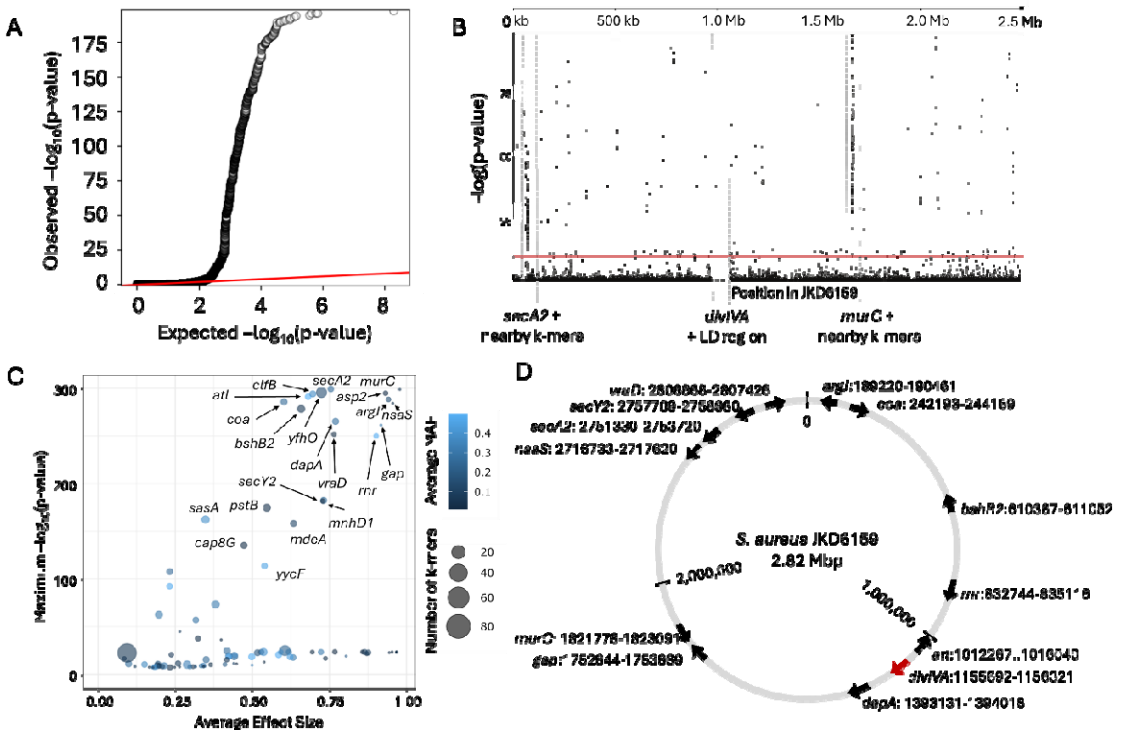



**Figure 2: Summary of GOLD-GWAS after masking *divIVA* and associated LD regions**. (A) Quantile-quantile plot comparing the expected -log(p-values) with observed -log(p-values) from GOLD-GWAS. The red diagonal line indicates where the expected and observed values are equal. (B) Manhattan plot demonstrating the statistical significance association for selected variants arranged in order on the reference genome JKD6159. Each dot represents a k-mer and the red line represents the threshold for significance. The positions of *murC*, *secA2*, *divIVA* and its LD region are indicated. (C) Plot of genes

associated with *divIVA*. Minor allele frequencies (MAF) are shown by colour gradient and dot size represents the number of k-mers mapped to the gene. (D) Genomic location and orientation of genes covered by associated k-mers with a likelihood ratio test -log(p-value) > 200 and average effect size (beta) > 0.5 that exist within the *S. aureus* JKD6159 genome. Position of *divIVA* is indicated by the red arrow.

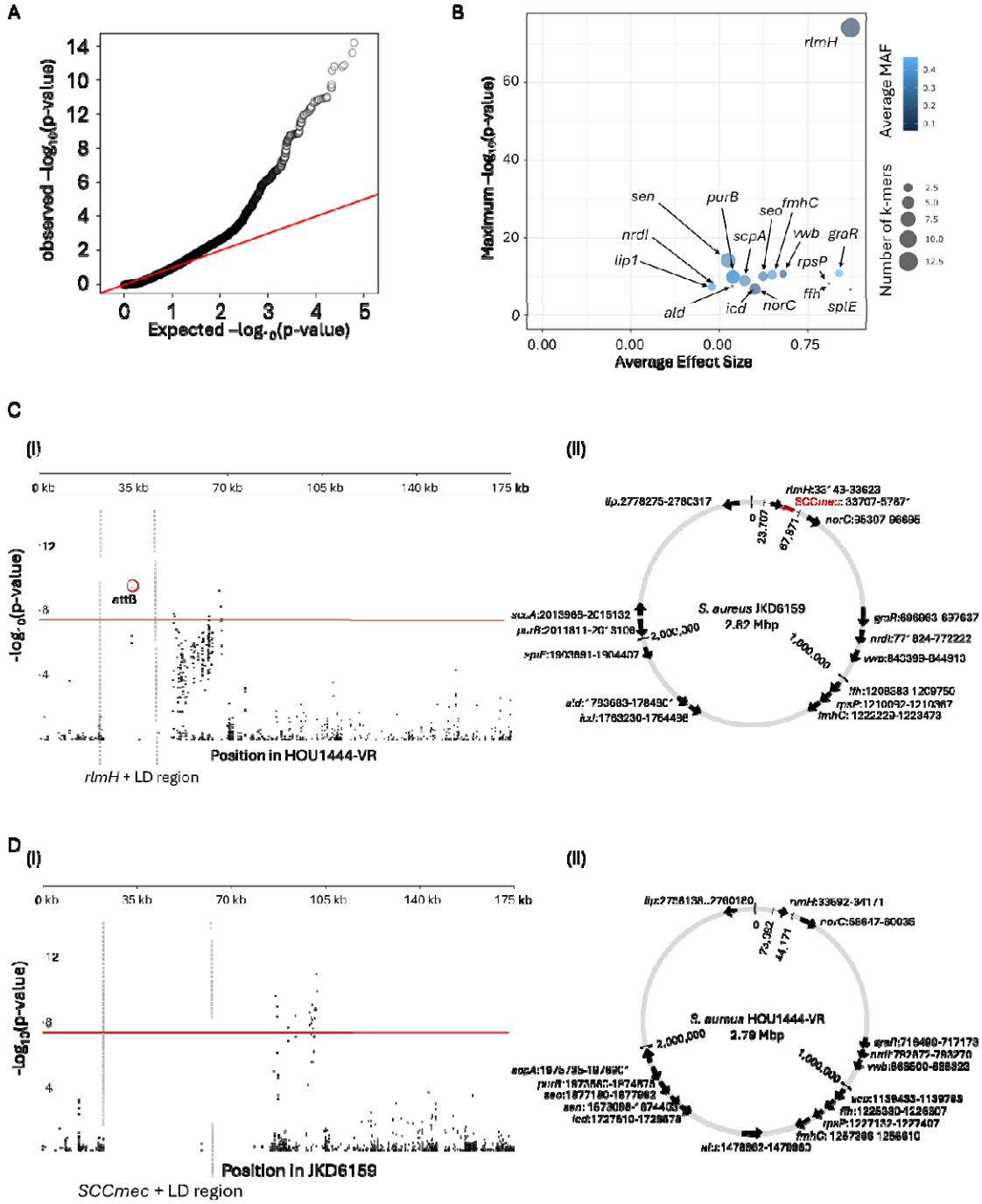



**Figure 3: Summary of GOLD-GWAS results after masking SCC*mec* and conserved LD regions.** (A) Quantile-quantile plot comparing the expected -log(p-values) with observed -log(p-values) from GOLD-GWAS. The red diagonal line indicates where the expected and observed values are equal. (B) Plot of genes associated with SCC*mec* carriage. Minor

allele frequency (MAF) values are depicted by the colour gradient and dot size represent the number of k-mers mapped to the gene. (C) Manhattan plot and corresponding genomic map for the SCC*mec*-positive JKD6159 genome. (i) Manhattan plot demonstrating the statistical significance association for selected variants arranged in order on the 0–175 kbp region of the *S. aureus* JKD6159 genome. Th position of *SCCmec* and its associated LD regions are indicated. Each dot represents a k-mer and the red line represents the threshold for significance. (ii) Genomic location and orientation of significantly associated genes in SCC*mec*-positive JDK6159. Position of SCC*mec* is indicated by the red block and masked regions by dotted lines. (D) Manhattan plot and corresponding genomic map for the SCC*mec*-negative HOU1444-VR genome. (i) Manhattan plot demonstrating the statistical significance association for selected variants arranged in order on the 0–175 kbp region of the *S. aureus* HOU1444-VR genome. The position of *rlmH* and its associated LD region are indicated. The red circle highlights the k-mer identified as the attB site. (ii) Genomic location and orientation of significantly associated genes in SCC*mec*-negative HOU1444-VR. Masked regions are indicated by dotted lines.

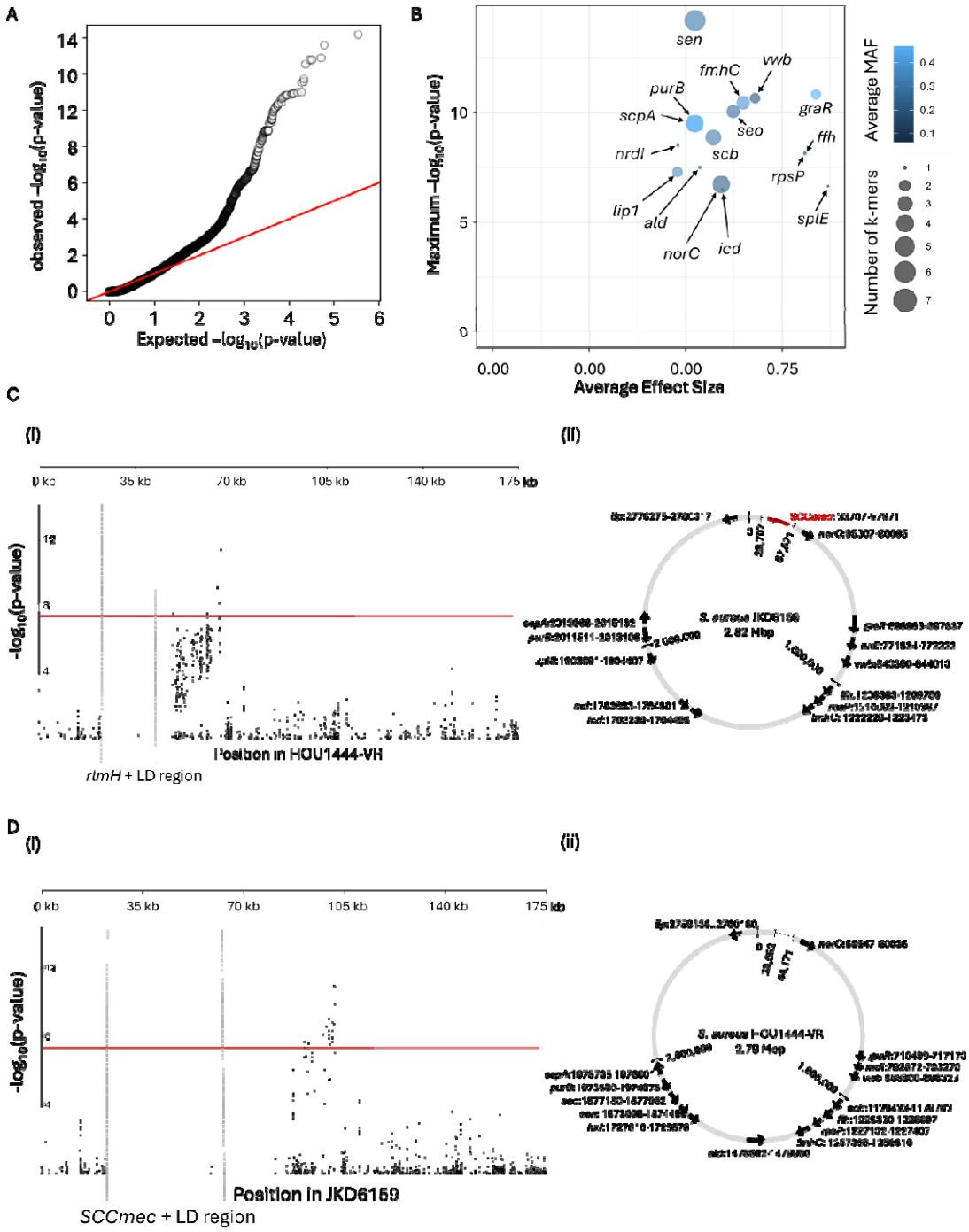



**Figure 4: Summary of GOLD-GWAS results after masking SCC*mec*, *rlmH* and associated LD regions.** (A) Quantile-quantile plot comparing the expected -log(p-values) with observed -log(p-values) from GOLD-GWAS. The red diagonal line indicates where the expected and observed values are equal. (B) Plot of genes associated with SCC*mec* carriage. Minor allele frequency (MAF) values are depicted by the colour gradient and dot size represent the number of k-mers mapped to the gene. (C) Manhattan plot and corresponding genomic map for the SCC*mec*-positive JKD6159 genome. (i) Manhattan plot demonstrating the statistical significance association for selected variants arranged in order on the 0–175 kbp region of the *S. aureus* JKD6159 genome. Th position of *SCCmec* and its

associated LD regions are indicated. Each dot represents a k-mer and the red line represents the threshold for significance. (ii) Genomic location and orientation of significantly associated genes in SCC*mec*-positive JDK6159. Position of SCC*mec* is indicated by the red block and masked regions by dotted lines. (D) Manhattan plot and corresponding genomic map for the SCC*mec*-negative HOU1444-VR genome. (i) Manhattan plot demonstrating the statistical significance association for selected variants arranged in order on the 0–175 kbp region of the *S. aureus* HOU1444-VR genome. The position of *rlmH* and its associated LD region are indicated. (ii) Genomic location and orientation of significantly associated genes in SCC*mec*-negative HOU1444-VR. Masked regions are indicated by dotted lines.

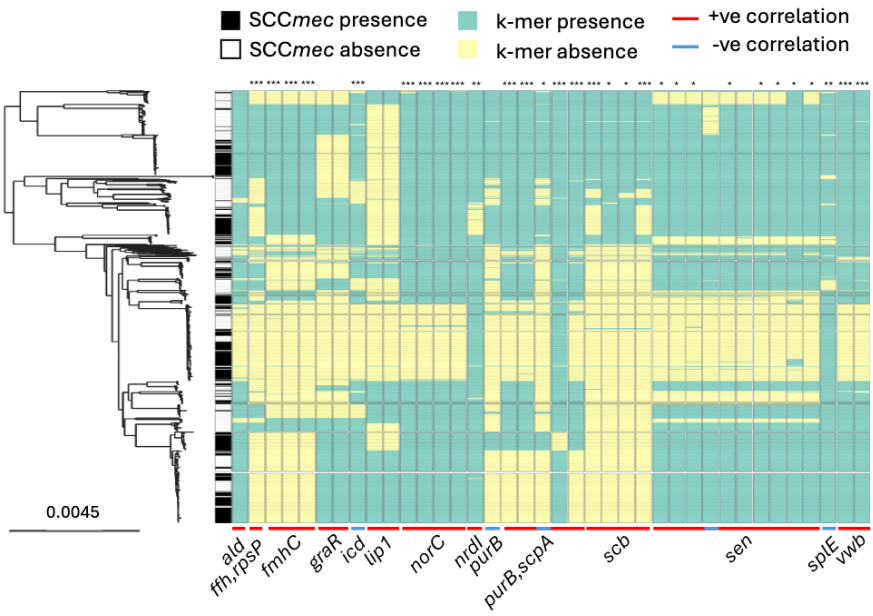



**Figure 5: Presence of 38 k-mers significantly associated with SCC*mec* carriage**. Maximum-likelihood phylogenetic heatmap showing the presence/absence of SCC*mec* (black/white) and 38 k-mers (green/yellow) that lie in coding regions and are significantly associated with SCC*mec* carriage. The corresponding genes are annotated below the heatmap. Created using Microreact (86). *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

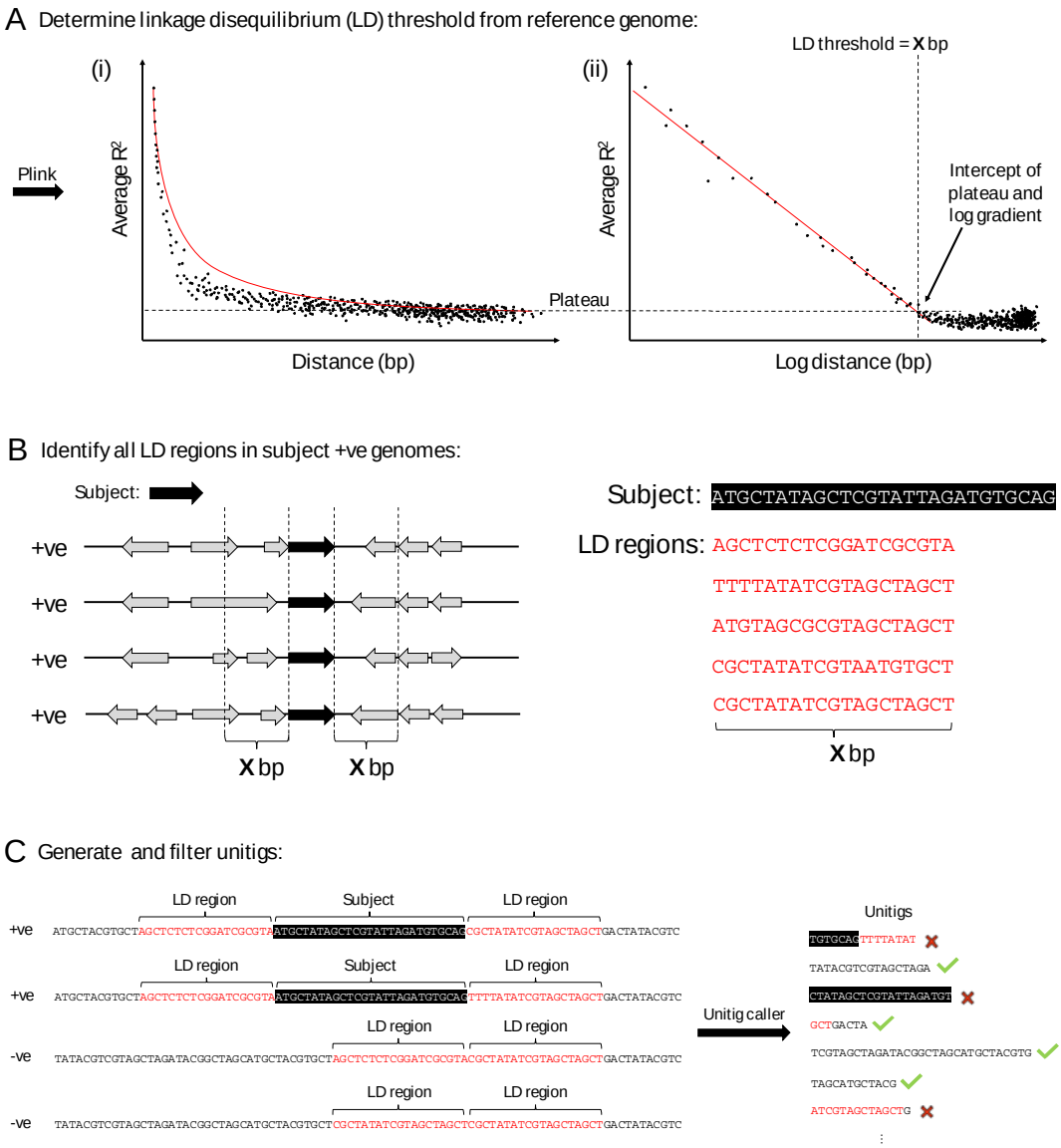
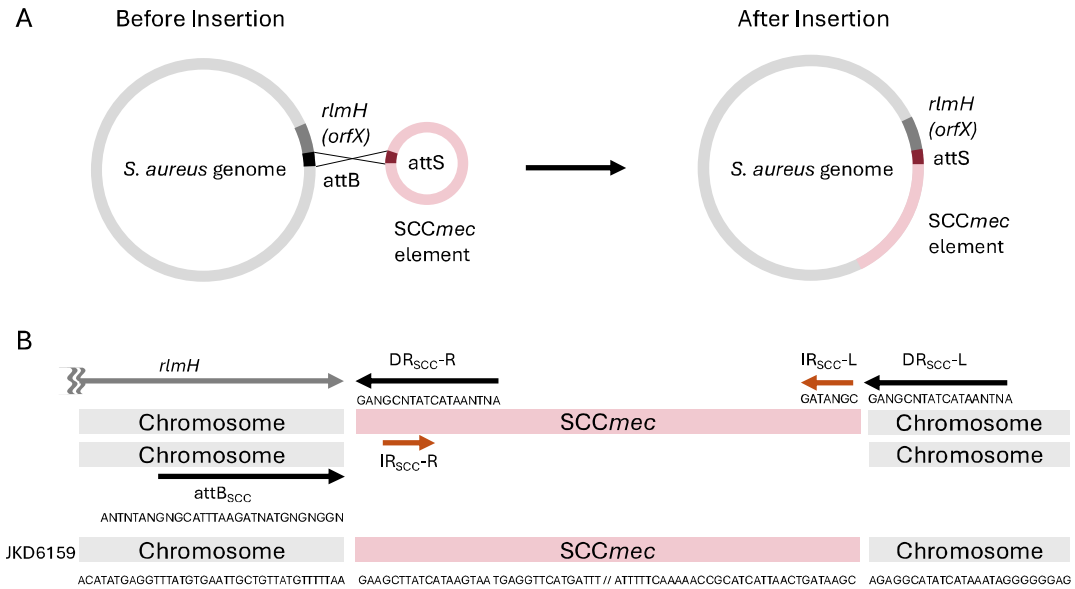
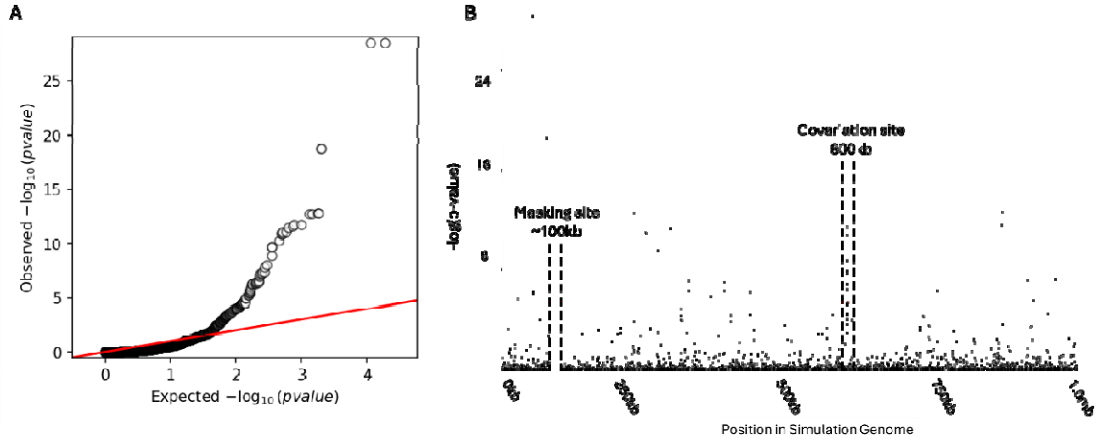



**Figure 6: Computational Masking Method**. (A) (i) Relationship between the linear distance between two SNPs and the average $R^2$ value. Red line represents a polynomial trend line. (ii) Relationship between the logarithm of the distance between two SNPs and the average $R^2$ value. Red line indicates the linear trend computed from the subset of samples excluding $R^2 < 0.2$. The horizontal dotted line denotes the average $R^2$ value observed beyond 100,000 bp, and the vertical dotted line marks the intersection of this with the trend line in the logarithmic distance plot. (B) Schematic demonstrating how sequences of LD regions are collected in association with the target gene based on the LD threshold from (A). (C) Unitigs with ≥80% identity with any target gene or LD region sequence are filtered out. Black boxes indicate target sequences, red text indicates the LD regions.

**Supplementary**

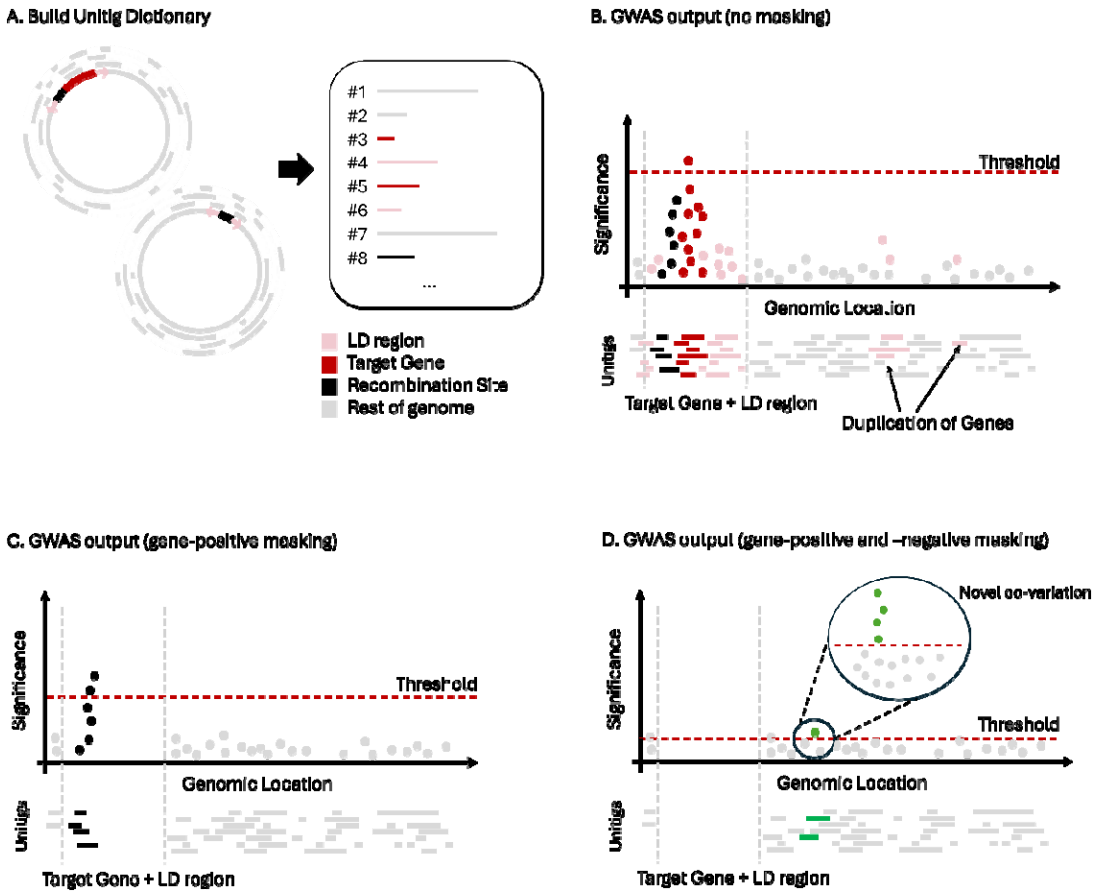



**Supplementary Figure 1: Site-specific integration of SCC*mec***. (A) The *S. aureus* genome before and after SCC*mec* insertion. The attB (blue) and attS (red) sites are responsible for SCC*mec* recombination. Before insertion, the attB site is located downstream of the *rlmH* sequence, which is replaced by attS once integration has occurred. (B) Structure of SCC*mec*. Within SCC*mec* there is one direct repeat (DR) sequence, and two reverse complement inverted repeats (IRs). An additional DR is present at the boundary of the downstream chromosomal region. Arrows indicate the orientation and positions of these repeats. Sequences based on the well-characterised clinical strain of *S. aureus* JKD6159 (NCBI Reference Sequence: NC_017338.2).




**Supplementary Figure 2: Summary of the genome-wide association study after masking an artificially created site of covariation.** (A) Quantile-quantile plot comparing the expected -log(p-values) with observed -log(p-values). The red diagonal line indicates where the expected and observed values are equal. (B) Manhattan plot demonstrating the statistical significance association for selected variants arranged in order on a simulated genome. Each dot represents a k-mer. Red dotted line indicates the threshold for

significance. Masked region near 100kb and artificial covariation region at 600kb are labelled.




**Supplementary Figure 3: Summary of the GOLD-GWAS process.** An overview conceptualising the methodology and advantages of GOLD-GWAS over standard GWAS. (A) A dictionary of unitigs is constructed that captures all genomic variation present in a pangenome. (B) Performing standard GWAS on this unitig dictionary using the presence of a target gene as the binary classifier will produce strong correlations for any unitigs that map to the target gene itself. Consequently, the extremely high p-values associated with these unitigs will result in a high threshold for statistical significance. Strongly associated unitigs may also include a recombination site, the target gene's linkage disequilibrium regions, and any target gene unitig duplicated within the genome. (C) After identifying all target gene and LD region unitigs within gene-positive isolates, GOLD-GWAS removes these unitigs from the dictionary before performing GWAS. Unitig filtering now allows detection of recombination site or insertion sequence due to strong correlation with gene-negative isolates only. (D) By identifying all target gene and LD region unitigs in both gene-positive and -negative isolates and removing them prior to GWAS, GOLD-GWAS reveals previously hidden unitigs that significantly covary with the target gene.

Figure 1

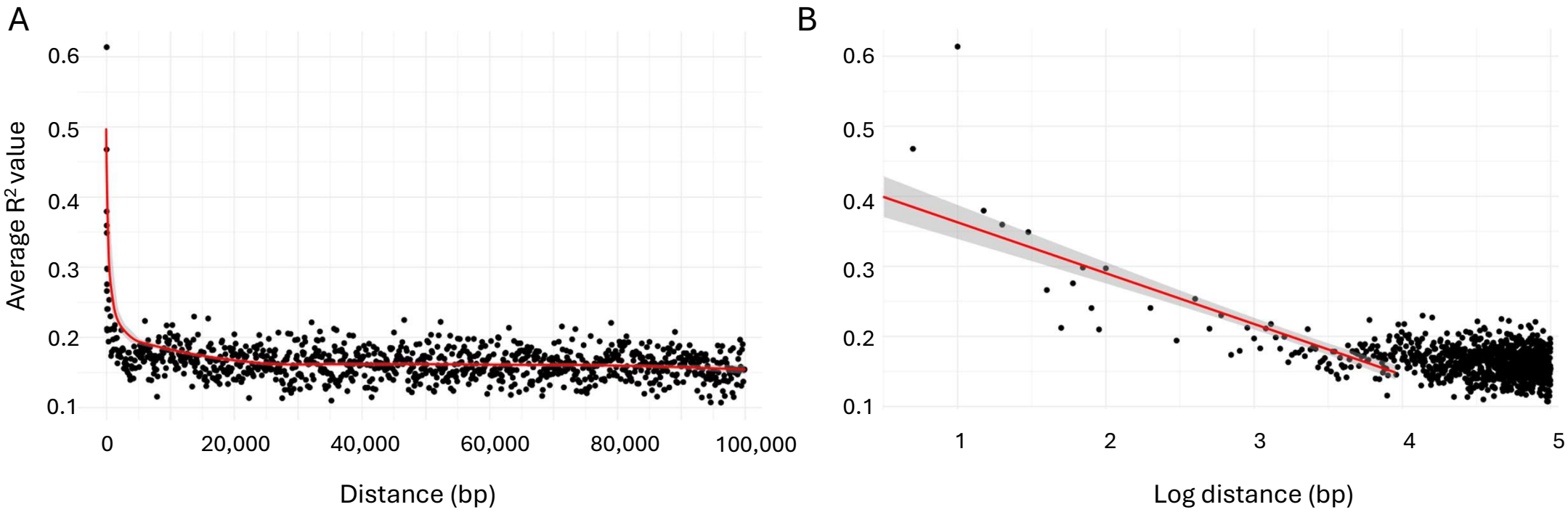

A
B
Average R2 value
0.6
0.5
0.4
0.3
0.2
0.1
0
20,000
40,000
60,000
80,000
100,000
Distance (bp)
1
2
3
4
5
Log distance (bp)

Figure 2

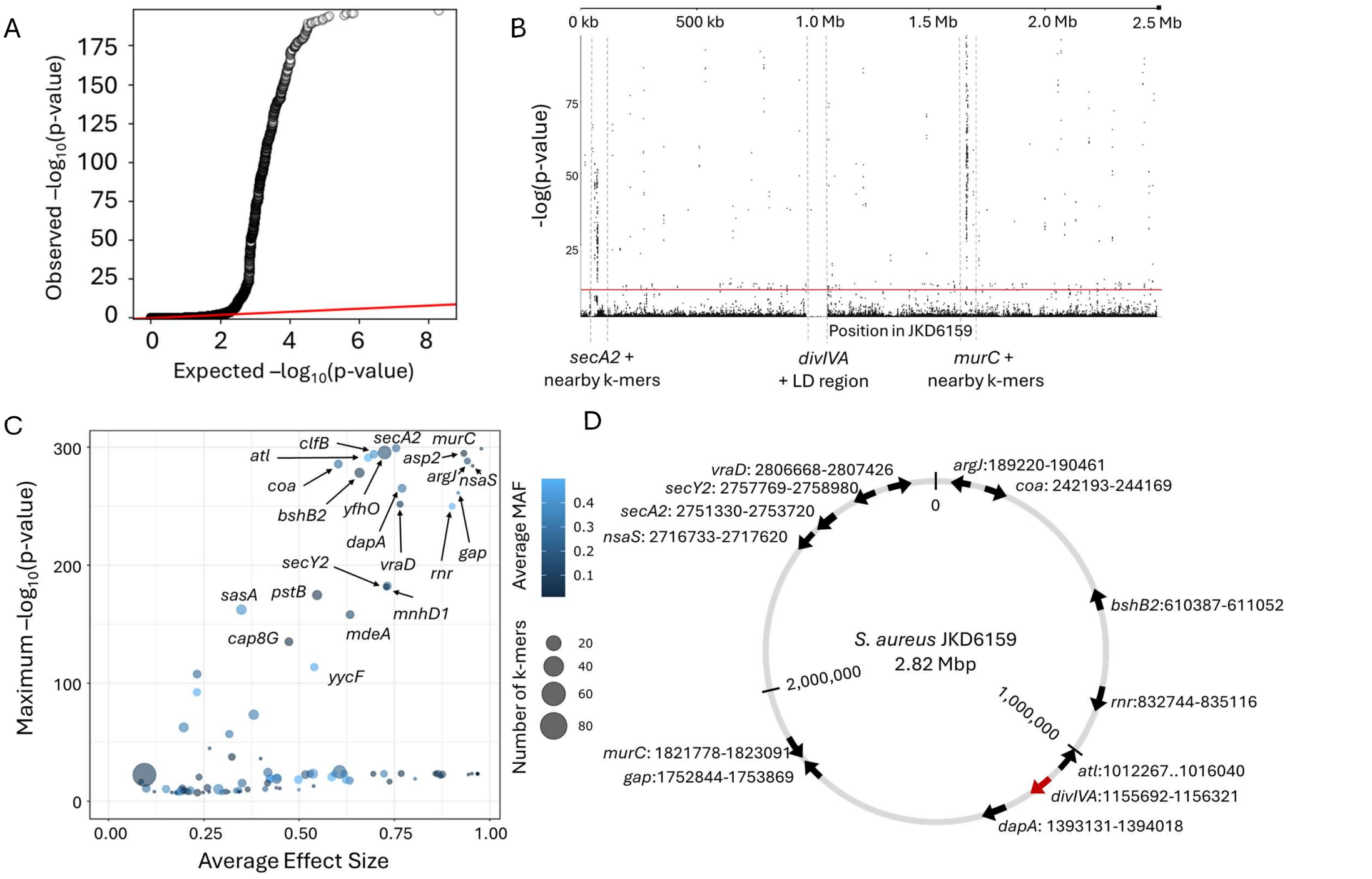
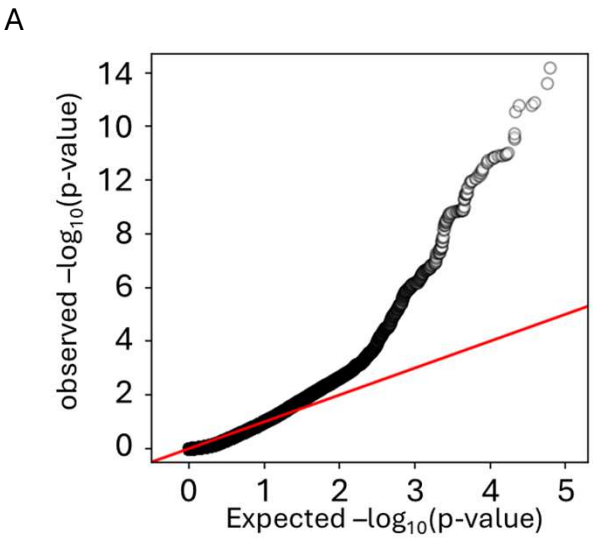
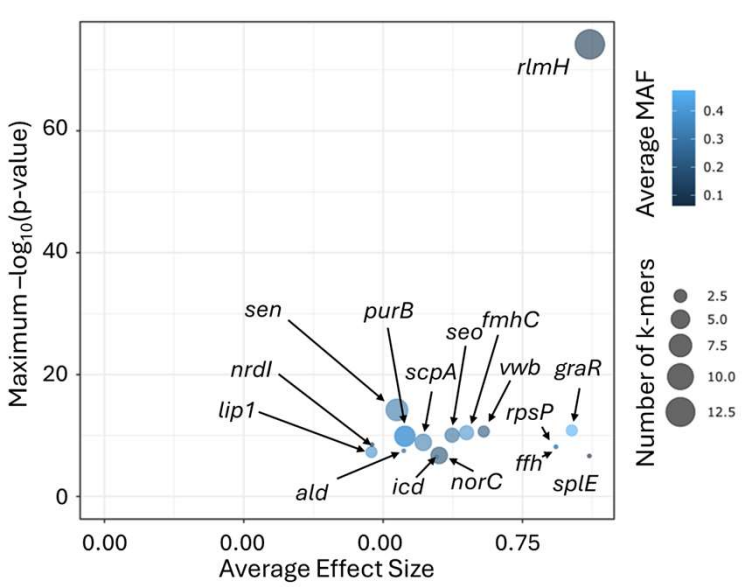
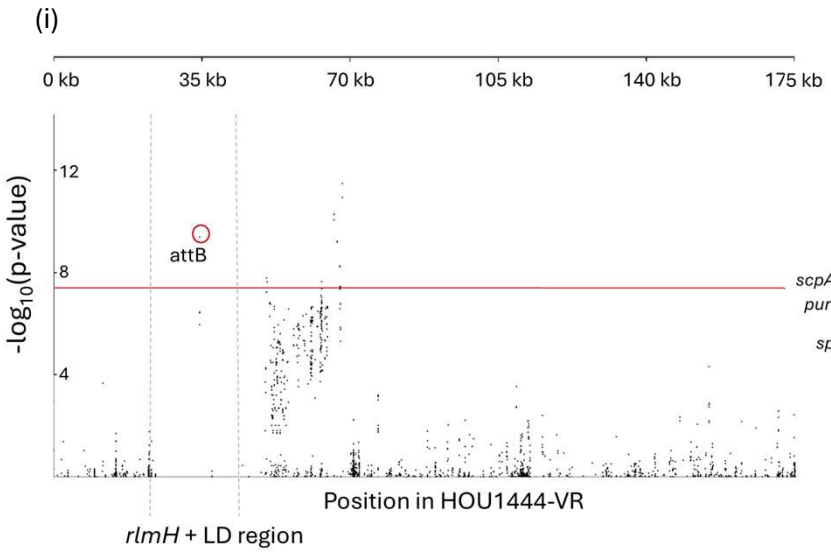

A
Observed –log10(p-value)
Expected –log10(p-value)
B
0 kb
500 kb
1.0 Mb
1.5 Mb
2.0 Mb
2.5 Mb
-log(p-value)
Position in JKD6159
secA2 + nearby k-mers
divIVA + LD region
murC + nearby k-mers
C
Maximum –log10(p-value)
Average Effect Size
clfB
secA2
murC
atl
asp2
argJ
nsaS
coa
bshB2
yfhO
dapA
vraD
rnr
gap
secY2
sasA
pstB
mnhD1
mdeA
cap8G
yycF
Average MAF
Number of k-mers
D
S. aureus JKD6159
2.82 Mbp
vraD: 2806668-2807426
secY2: 2757769-2758980
secA2: 2751330-2753720
nsaS: 2716733-2717620
argJ:189220-190461
coa: 242193-244169
bshB2:610387-611052
rnr:832744-835116
atl:1012267..1016040
divIVA:1155692-1156321
dapA: 1393131-1394018
murC: 1821778-1823091
gap:1752844-1753869

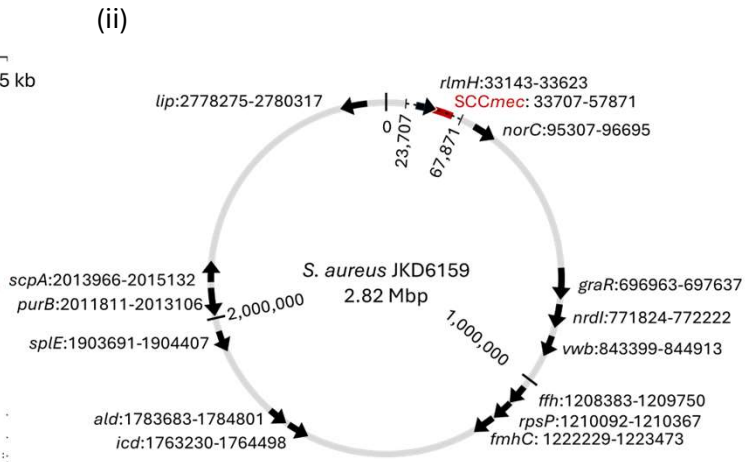
Figure 3


A
observed −log10(p-value)
Expected −log10(p-value)
0
2
4
6
8
10
12
14
0
1
2
3
4
5

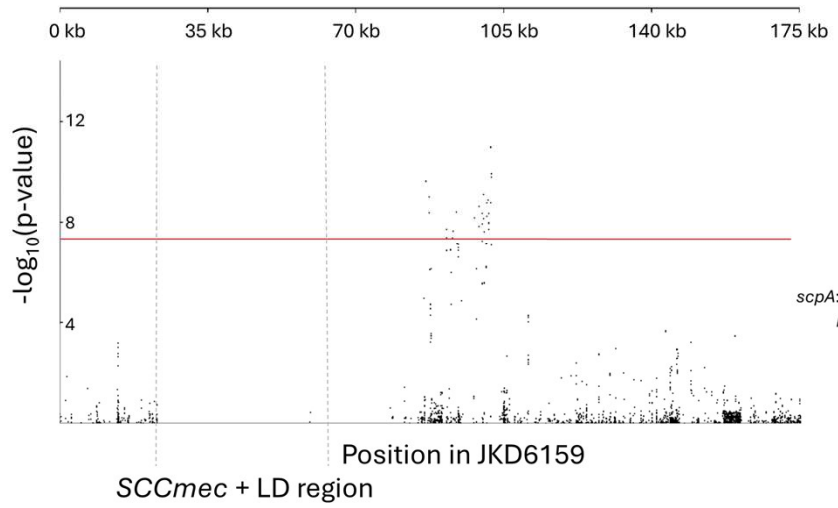

B
Maximum −log10(p-value)
Average Effect Size
0
20
40
60
0.00
0.00
0.00
0.75
rlmH
sen
purB
seo
fmhC
nrdI
scpA
vwb
graR
lip1
rpsP
ffh
ald
icd
norC
splE
Average MAF
0.4
0.3
0.2
0.1
Number of k-mers
2.5
5.0
7.5
10.0
12.5

C


(i)
0 kb
35 kb
70 kb
105 kb
140 kb
175 kb
-log10(p-value)
12
8
4
attB
Position in HOU1444-VR
rlmH + LD region


(ii)
rlmH:33143-33623
SCCmec: 33707-57871
lip:2778275-2780317
norC:95307-96695
0
23,707
67,871
S. aureus JKD6159
2.82 Mbp
scpA:2013966-2015132
purB:2011811-2013106
2,000,000
1,000,000
splE:1903691-1904407
graR:696963-697637
nrdI:771824-772222
vwb:843399-844913
ffh:1208383-1209750
rpsP:1210092-1210367
fmhC: 1222229-1223473
ald:1783683-1784801
icd:1763230-1764498

D


(i)
0 kb
35 kb
70 kb
105 kb
140 kb
175 kb
-log10(p-value)
12
8
4
Position in JKD6159
SCCmec + LD region

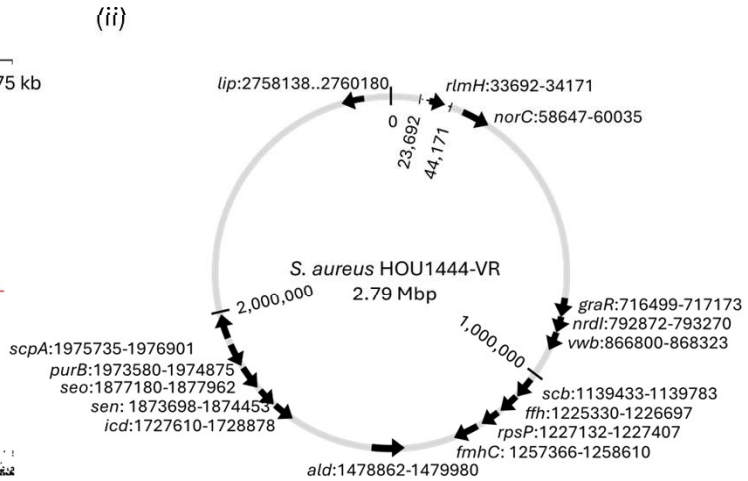

(ii)
lip:2758138..2760180
rlmH:33692-34171
norC:58647-60035
0
23,692
44,171
S. aureus HOU1444-VR
2.79 Mbp
2,000,000
1,000,000
graR:716499-717173
nrdI:792872-793270
vwb:866800-868323
scpA:1975735-1976901
purB:1973580-1974875
seo:1877180-1877962
sen: 1873698-1874453
icd:1727610-1728878
scb:1139433-1139783
ffh:1225330-1226697
rpsP:1227132-1227407
fmhC: 1257366-1258610
ald:1478862-1479980

Figure 4

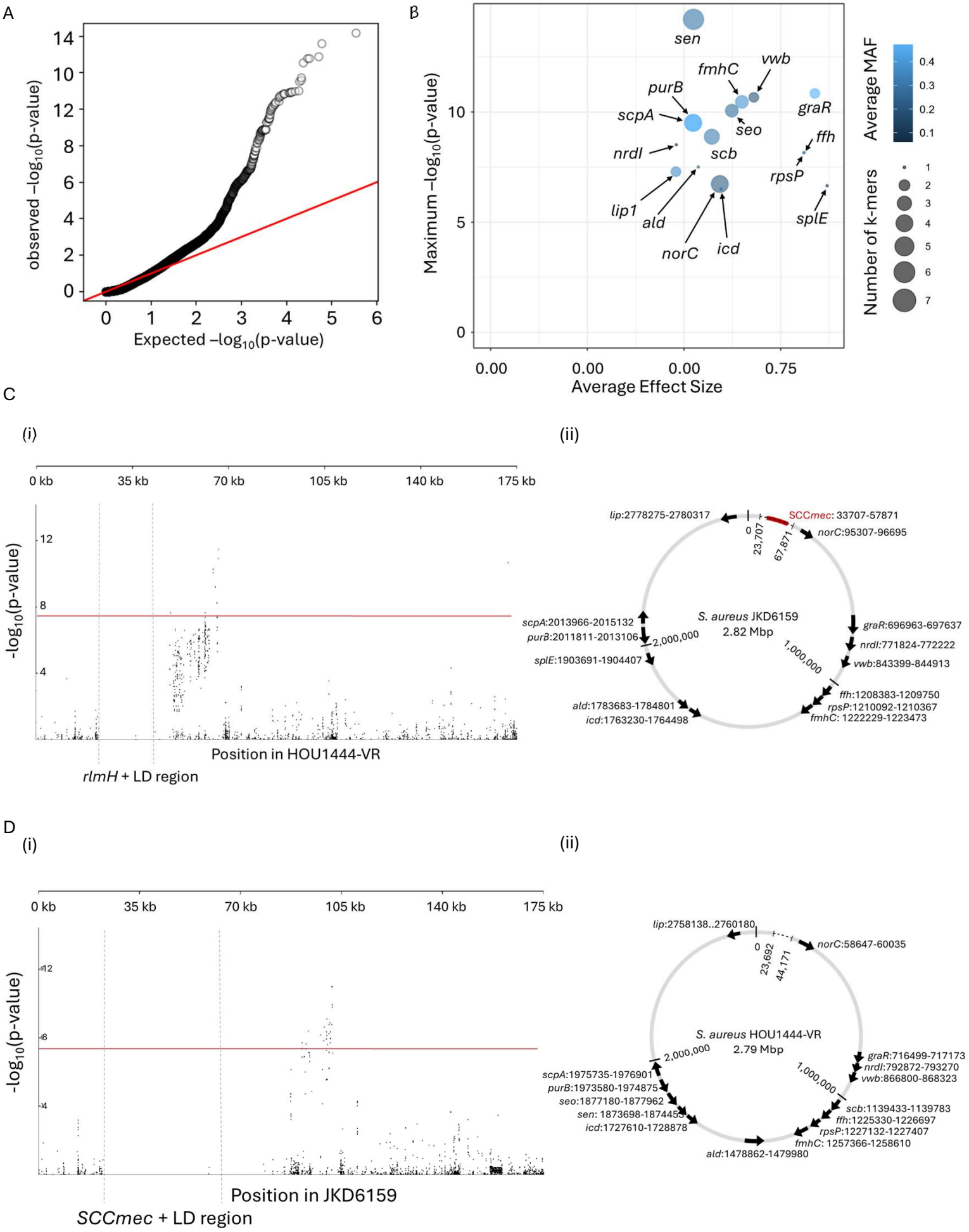

A
observed −log10(p-value)
Expected −log10(p-value)
B
Maximum −log10(p-value)
Average Effect Size
Average MAF
Number of k-mers
sen
fmhC
vwb
purB
scpA
seo
graR
ffh
nrdI
scb
rpsP
lip1
ald
norC
icd
splE
C
(i)
(ii)
-log10(p-value)
Position in HOU1444-VR
rlmH + LD region
S. aureus JKD6159
2.82 Mbp
SCCmec: 33707-57871
lip:2778275-2780317
norC:95307-96695
scpA:2013966-2015132
purB:2011811-2013106
splE:1903691-1904407
ald:1783683-1784801
icd:1763230-1764498
graR:696963-697637
nrdI:771824-772222
vwb:843399-844913
ffh:1208383-1209750
rpsP:1210092-1210367
fmhC: 1222229-1223473
D
(i)
(ii)
-log10(p-value)
Position in JKD6159
SCCmec + LD region
S. aureus HOU1444-VR
2.79 Mbp
lip:2758138..2760180
norC:58647-60035
scpA:1975735-1976901
purB:1973580-1974875
seo:1877180-1877962
sen: 1873698-1874455
icd:1727610-1728878
graR:716499-717173
nrdI:792872-793270
vwb:866800-868323
scb:1139433-1139783
ffh:1225330-1226697
rpsP:1227132-1227407
fmhC: 1257366-1258610
ald:1478862-1479980

Figure 5

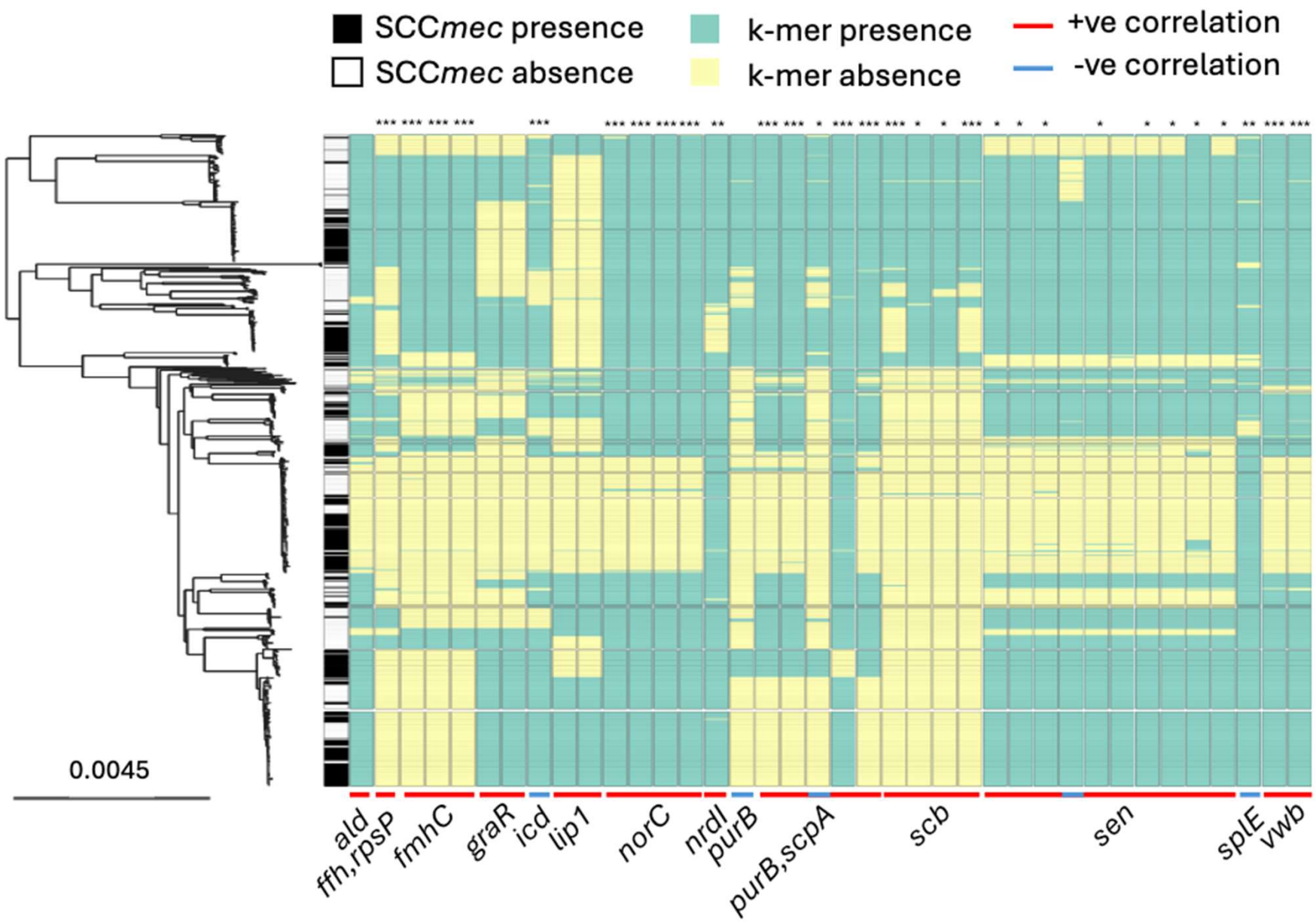
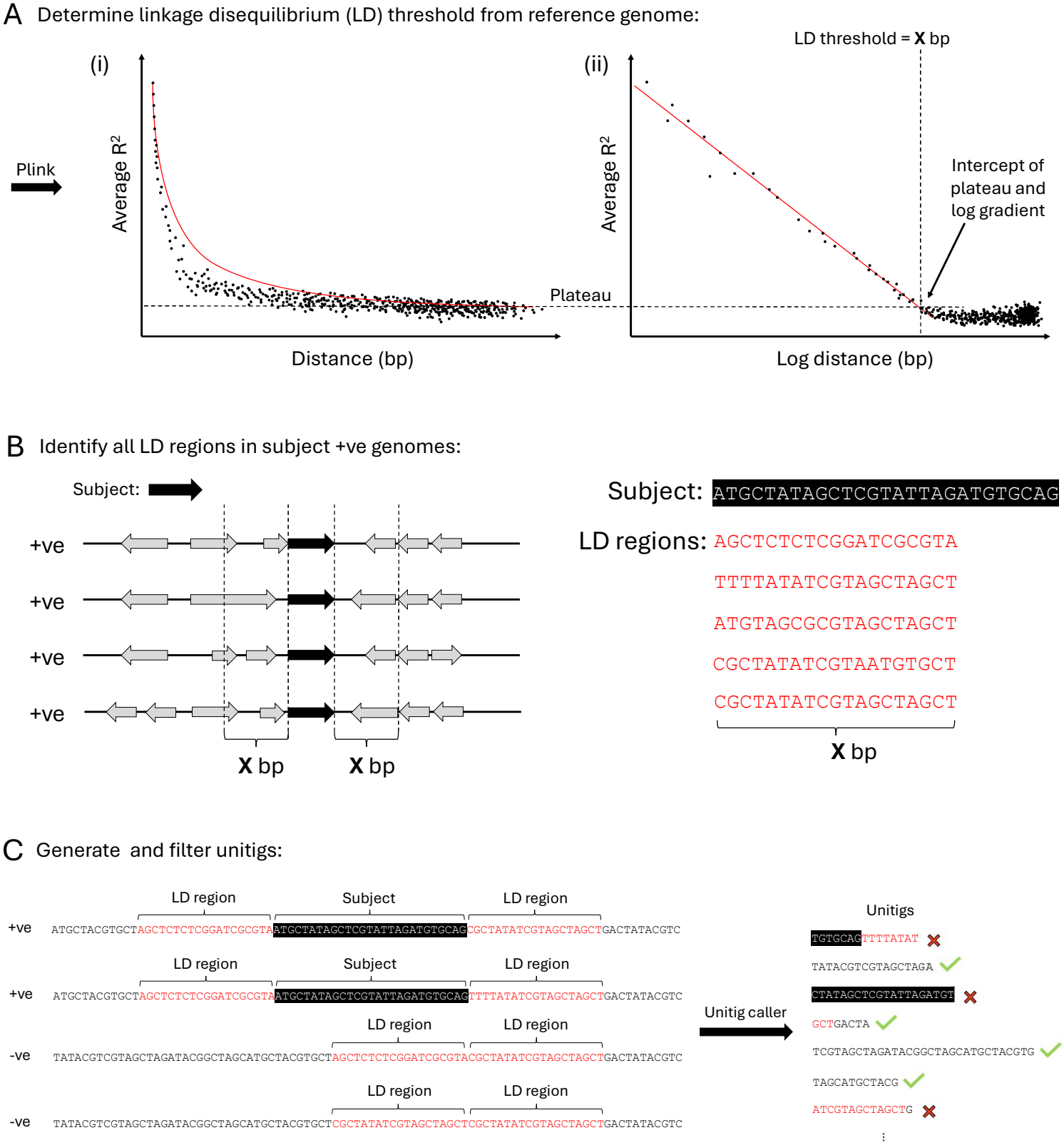

SCCmec presence
SCCmec absence
k-mer presence
k-mer absence
+ve correlation
-ve correlation
0.0045
ald
ffh,rpsP
fmhC
graR
icd
lip1
norC
nrdI
purB
purB,scpA
scb
sen
splE
vwb

Figure 6


A Determine linkage disequilibrium (LD) threshold from reference genome:
Plink
(i)
Average R2
Distance (bp)
Plateau
(ii)
LD threshold = X bp
Average R2
Intercept of plateau and log gradient
Log distance (bp)
B Identify all LD regions in subject +ve genomes:
Subject:
+ve
+ve
+ve
+ve
X bp
X bp
Subject: ATGCTATAGCTCGTATTAGATGTGCAG
LD regions: AGCTCTCTCGGATCGCGTA
TTTTATATCGTAGCTAGCT
ATGTAGCGCGTAGCTAGCT
CGCTATATCGTAATGTGCT
CGCTATATCGTAGCTAGCT
X bp
C Generate and filter unitigs:
LD region
Subject
LD region
+ve ATGCTACGTGCTAGCTCTCTCGGATCGCGTAATGCTATAGCTCGTATTAGATGTGCAGCGCTATATCGTAGCTAGCTGACTATACGTC
LD region
Subject
LD region
+ve ATGCTACGTGCTAGCTCTCTCGGATCGCGTAATGCTATAGCTCGTATTAGATGTGCAGTTTTATATCGTAGCTAGCTGACTATACGTC
LD region
LD region
-ve TATACGTCGTAGCTAGATACGGCTAGCATGCTACGTGCTAGCTCTCTCGGATCGCGTACGCTATATCGTAGCTAGCTGACTATACGTC
LD region
LD region
-ve TATACGTCGTAGCTAGATACGGCTAGCATGCTACGTGCTCGCTATATCGTAGCTAGCTCGCTATATCGTAGCTAGCTGACTATACGTC
Unitig caller
Unitigs
TGTGCAGTTTTATAT
TATACGTCGTAGCTAGA
CTATAGCTCGTATTAGATGT
GCTGACTA
TCGTAGCTAGATACGGCTAGCATGCTACGTG
TAGCATGCTACG
ATCGTAGCTAGCTG