

Phylogenomics and gene association analysis support probiotic potential in a subset of *Staphylococcus xylosus* isolates

Kaylee Dawn Rich¹, Zhangbin Cai², Diego B. Nobrega¹, Jennifer Ronholm² and Jeroen De Buck^{1,*}

Abstract

Mastitis is a major disease in dairy cattle, often caused by *Staphylococcus aureus* infection and increasingly complicated by antimicrobial resistance. As a result, bacterial strains able to effectively provide colonization resistance in the bovine mammary gland are being explored as potential probiotics to reduce reliance on antimicrobial treatments. Among these, some non-*S. aureus* staphylococci species have shown promise; however, their effects appear to be highly strain-dependent. In this study, a combination of machine learning approaches alongside standard genomic and phylogenetic analyses was used to compare and select putative probiotic *Staphylococcus xylosus* strains. The genomes of 82 *S. xylosus* and 16 closely related *Staphylococcus* isolates from milk samples obtained from the Canadian Bovine Mastitis Research Network were assembled and annotated. In addition to identifying orthologous gene families and reconstructing phylogenetic relationships, each genome was screened for virulence factors, antimicrobial resistance (AMR) genes and bacteriocins. Random forest modelling and association rule learning were then applied to identify combinations of genes associated with isolates from milk samples collected from quarters exhibiting low inflammation, assessed using somatic cell counts (SCC). This approach identified 63 genes that frequently co-occurred in isolates from low SCC samples (low SCC <25,000 cells ml⁻¹) but were largely absent in those from high SCC samples (≥200,000 cells ml⁻¹). These gene sets were used as biomarkers in conjunction with phylogenetic and clustering analyses to guide the selection of a subset of *S. xylosus* isolates with potential probiotic properties.

Impact Statement

Mastitis is a common and costly disease in dairy cows that affects milk production and animal health. *Staphylococcus aureus* is a major cause of bovine mastitis, but treatment success is often low, in part due to rising antimicrobial resistance, and new ways to prevent and manage infection are needed. One approach is the use of protective bacteria as probiotics. Non-*S. aureus* staphylococci (NAS) are common in the udder, but their effects vary widely between strains. While some NAS appear beneficial, others can still cause or exacerbate disease. Understanding the genetic basis of these differences is key to identifying strains suitable for probiotic development. We examined genomic variation among *Staphylococcus xylosus* and the closely related *Staphylococcus shinnii* and *Staphylococcus pseudoxylosus* isolates to identify gene combinations associated with low inflammation in the udder, which may indicate probiotic potential. These findings could help detect protective NAS strains to use as probiotics that improve udder health without relying on antibiotics.

DATA SUMMARY

Draft genome assemblies and raw reads can be found under National Center for Biotechnology Information (NCBI) BioProject PRJNA1336871, available at <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1336871>. Code and documentation for bioinformatic

Received 01 January 2026; Accepted 13 April 2026; Published 30 April 2026

Author affiliations: ¹Faculty of Veterinary Medicine, Foothills Campus, University of Calgary, Calgary, AB, Canada; ²Faculty of Agricultural and Environmental Sciences, Macdonald Campus, McGill University, Montréal, QC, Canada.

***Correspondence:** Jeroen De Buck, jdebuck@ucalgary.ca

Keywords: genomics; machine learning; mastitis; phylogenetics; probiotics; *Staphylococcus*.

Abbreviations: AIPs, autoinducing peptides; AMR, antimicrobial resistance; ANI, average nucleotide identity; ARL, association rules machine learning; FP, frequent pattern growth; IMI, intramammary infection; MDS, multidimensional scaling; NAS, non-*Staphylococcus aureus* staphylococci; RF, random forest; SCC, somatic cell counts; SCC, somatic cell count; T7SS, type VII secretion system; VF, virulence factor.

All supporting data, code and protocols have been provided within the article or through supplementary data files. Three supplementary tables are available with the online version of this article.

001708 © 2026 The Authors



This is an open-access article distributed under the terms of the Creative Commons Attribution License. This article was made open access via a Publish and Read agreement between the Microbiology Society and the corresponding author's institution.

methods are available at <https://github.com/kayleerich/biomARkers>.

INTRODUCTION

Mastitis is a costly disease in dairy cattle that is often caused by intramammary infection (IMI). It can present as either clinical mastitis, which shows visible signs such as swelling, redness and abnormal milk, or subclinical mastitis, which occurs without obvious symptoms but is typically detected through elevated somatic cell counts (SCC) [1–3]. This disease has significant impacts on animal health, milk production and farm economics. *Staphylococcus aureus* is a major cause of bovine mastitis, and though treating these infections with antimicrobials is common, bacteriological cure rates of IMI caused by *S. aureus* are often poor. Additionally, increasing antimicrobial resistance (AMR) has further challenged the success of therapies, and new treatment strategies are required [4–6]. Various approaches are being investigated, including the use of protective bacteria as a potential probiotic treatment [7–9].

Severity of IMI, including mastitis, is associated with microbiome composition in the udder, where some bacterial species appear to be protective against mastitis pathogens [10–13]. Among these are the frequently isolated non-*S. aureus* staphylococci (NAS). However, the relationships between NAS, other bacterial species and IMI severity are difficult to evaluate [2, 10, 14]. The role of NAS in bovine udder health is complex and nuanced. While traditionally considered minor IMI pathogens and frequently implicated as IMI agents [15–17], some NAS strains may confer protective effects by inhibiting other mastitis-causing pathogens [18, 19]. Recent studies have illustrated a strain-specific nature of NAS interactions with mastitis pathogens. For instance, Persson Waller et al. [20] found that certain *Staphylococcus chromogenes* genotypes were associated with more severe IMI than other *S. chromogenes* genotypes, whereas Jung et al. [12] found a negative correlation between NAS occurrence with *Escherichia coli* clinical mastitis, supporting the notion that some NAS may contribute positively to udder health. These accounts emphasize the need to distinguish between potentially virulent and protective NAS. They also support the increasingly accepted view that the effect of NAS on udder health is highly strain- or genotype-dependent [10].

Identifying beneficial NAS strains requires a deeper understanding of the mechanisms by which they interact with pathogenic species. Genes from protective strains could serve as biomarkers to predict a strain's probiotic potential or its risk of contributing to severe IMI. Among the mechanisms of antagonism, several have been described, including quorum quenching, antimicrobial production and biofilm inhibition. Quorum quenching is the disruption of bacterial communication systems that regulate gene expression based on population density. By interfering with these signalling pathways, antagonistic bacteria have been shown to attenuate the virulence of other bacteria [21]. NAS are capable of inhibiting the *S. aureus* quorum-sensing system, which is responsible for the expression of a range of virulence factors by interfering with its accessory gene regulator (*agr*) [22, 23]. By producing analogues of the *S. aureus* autoinducing peptides (AIPs) that mediate *agr* signalling, NAS suppress *S. aureus* quorum sensing and pathogenicity. The presence of AIP analogues in NAS genomes could be a promising lead in the search for probiotics. However, since these peptides target specific strains, reliance on AIP-producing NAS as a prophylactic may lead to the selection of AIP-resistant *S. aureus* strains [23].

Another antagonistic mechanism employed by NAS is the production of bacteriocins. These are antimicrobial peptides produced by a wide range of bacteria and have garnered attention for their potential in novel therapeutic development. NAS-derived bacteriocins have been shown to inhibit *S. aureus* [24], and bacteriocin-producing NAS have been proposed as potential probiotics [24–26]. However, the relationship between bacteriocin gene content and inhibitory activity is not straightforward. For instance, two strains of *Staphylococcus simulans* with differing bacteriocin gene profiles were found to have similar inhibitory effects on colonization of the mammary gland by *S. aureus* [24, 27], indicating that other factors may contribute to their antagonistic activity *in vivo*. Recent studies also suggest that biofilm inhibition may play a more critical role in the prophylactic potential of certain NAS strains [28]. Though the underlying genes involved are unknown, this effect appears to be independent of bacteriocin production [29]. These findings indicate that while bacteriocins are a promising avenue for drug development, they are unlikely to be the sole determinant of probiotic efficacy.

The mechanisms underlying the effect that NAS have on IMI are complex, and a complete understanding of the genes involved remains elusive. While some NAS strains inhibit mastitis-causing pathogens, they may still possess pathogenic potential themselves. Additionally, interactions with the host immune response must also be considered when evaluating probiotic suitability [30]. For this reason, focusing solely on genes with known functions may overlook subtle but important genetic patterns that contribute to protective effects. In this study, we focused on the NAS species *Staphylococcus xylosus* as a prospective probiotic because it is one of the most common NAS species isolated from bovine milk samples, it has been associated with lower milk SCC than the other commonly-isolated NAS species *S. chromogenes* and *S. simulans* and it is not associated with persistent IMI [31, 32]. Furthermore, though the reported effect that *S. xylosus* has on udder health varies, some strains have been shown to interfere with *S. aureus* quorum sensing and biofilm formation [13, 33–35]. To identify strains with probiotic potential, we characterized genomic and phylogenetic diversity within *S. xylosus* and two species formerly considered conspecific with *S. xylosus*: *Staphylococcus shinii* and *Staphylococcus pseudoxylosus*. Because these taxa are highly similar (ANI >90%) but differ in SCC phenotypes, we included them to improve the stratification of isolates and enhance our ability to distinguish genes associated with high versus low SCC.

We used SCC to categorize isolates according to host inflammatory response. Using random forest modelling and association rules learning to identify groups of genes associated with low inflammation, we aimed to identify genetic markers that could differentiate protective from potentially harmful *S. xylosus* strains.

METHODS

Genome extraction, isolation and sequencing

The 98 *Staphylococcus* isolates used in this study were sourced from the Mastitis Pathogen Culture Collection located at the Faculté de médecine vétérinaire à Saint-Hyacinthe in Québec, and data and samples were originally collected during a study conducted by the National Cohort of Dairy Farms of the Canadian Bovine Mastitis Research Network [36]. Isolates were divided into two groups based on sample SCC of the corresponding milk samples: $\text{SCC}_{\text{low}} < 25,000 \text{ cells ml}^{-1}$ or $\text{SCC}_{\text{high}} \geq 200,000 \text{ cells ml}^{-1}$. Genomic DNA was extracted from isolates using the DNeasy UltraClean Microbial Kit (QIAGEN, Germany) according to the manufacturer's instructions. Libraries were prepared with the Nextera DNA Flex kit (Illumina, USA) and sequenced on an Illumina MiSeq to generate 2×300 bp paired-end reads.

Whole-genome assembly and annotation

Snakemake v7.32.4 [37] was used to automate the assembly and annotation process for all isolates. The raw sequence reads were screened for low quality and adapter sequences and then trimmed using Fastp v0.24.0 [38, 39]. Quality of the trimmed sequences was assessed with FastQC v0.12.1 [40] then *de novo* genomes were assembled using SPAdes v4.0.0 [41] implemented by Unicycler v0.5.1 [42] with default parameters. Genome assembly quality and completeness were assessed using Quast v5.3.0 [43, 44] and BUSCO v5.8.0 [45] (parameters: -m geno -l bacillales_odb10), and contigs shorter than 500 bp were removed using seqkit v2.5.1. Annotation of protein-coding genes, rRNAs and tRNAs in the assembled genomes was done using Pyrodigal v3.6.3 [46], Infernal v1.1.5 [47] and tRNAscan v2.0.12 [48] implemented via Bakta v1.10.1 [49] (additional parameters: --genus *Staphylococcus* --species *xylosus* --gram +, full database). The species of each isolate was assessed by calculating average nucleotide identity (ANI) between each genome and the reference genomes for *S. xylosus*, *S. shinii* and *S. pseudoxylosus* (NCBI accessions: GCF_000709415.1, GCA_017583065.1, GCF_044794325.1), where two genomes were considered from the same species if ANI >95%.

Plasmid sequences were identified and extracted from the genome assembly of each isolate using MOB-suite v3.1.9 [50]. The identified plasmid sequences were manually examined, and those that were annotated solely with non-plasmid genes (e.g. rRNA genes, MOB closest-neighbour match to NCBI accession CP042023, CP042061 or CP042065) were removed. Differences in the distributions of the number of plasmids per isolate between high and low SCC isolates were assessed using the exact two-sample Kolmogorov–Smirnov test.

Virulence factors, bacteriocins and AMR genes were identified using a manually curated database that included sequences from VRprofile, BAGEL4 [51] and CARD [52] databases, based on predicted orthology (see the next section for orthology prediction methods).

Pan-genome and phylogenetic analysis

Orthologous gene families (orthogroups) were determined via Orthofinder v3.0.1 [53, 54] using the annotated protein-coding sequences obtained from Bakta. Orthogroups were then used to characterize the pan-genome of *S. xylosus*, represented by the core genome (hard core: shared by ≥99% of isolates, soft core: 99%>isolates≥95%), accessory genome (95%>isolates≥15%) and cloud genome (shared by <15% of isolates). Pan-genome analyses were assessed in R using the pagoo v0.3.18 [55] and micropan v2.1 [56] packages.

To assess diversity and evolutionary relationships among the *S. xylosus*, *S. shinii* and *S. pseudoxylosus* isolates, phylogenetic trees were constructed from single-copy protein-coding genes present in ≥95% of isolates. The *S. xylosus* reference genome (accession GCF_000709415.1) was included for the phylogenetic context of the isolates' placement within the tree. The coding sequences of annotated protein-coding genes were clustered using CD-HIT v4.8.1 [57, 58] at ≥90% sequence identity over ≥80% sequence length (parameters: -c 0.90 s 0.80 n 5). Sequences in clusters with only one representative from ≥95% of isolates were used to generate a database of marker proteins for phylogenetic tree construction with PhyloPhlAn v3.1.1 [59] (non-default parameters: --db_aa diamond --min_num_entries 85 --diversity low --trim greedy --remove_fragmentary_entries --fragmentary_threshold 0.67). Visual comparisons to the gene-tree-derived isolate tree generated by Orthofinder were done to assess if there were notable discrepancies between the two. Trees were visualized in R using the ggtree package v3.16.0 [60, 61].

Isolate similarity and orthogroup correlations

Multidimensional scaling (MDS) for all isolates was done using accessory genome orthogroups (in <95% of isolates), as well as all virulence factor, AMR and bacteriocin orthogroups. MDS distances and *k*-means clusters from those distances were calculated

using the `cmdscale` and `k-means` functions from the `stats` package v4.3.2 in R. The `k-means` clusters for the virulence factor, AMR and bacteriocin MDS distances were coloured based on similarity to the accessory genome `k-means` clusters. Similarity was determined by considering the number of shared isolates between clusters, guided by calculating the Rand Index for cluster similarity using the `rand_score` function from the `scikit-learn` module v1.4.2 in Python, and plots were visualized using `ggplot2` v3.5.2 in R.

Pairwise correlation for all virulence factors, AMR and bacteriocin orthogroups was determined by calculating the phi coefficient with a confidence level of 0.95 for orthogroup presence in SCC_{high} and SCC_{low} isolates using the `cor` function of the `stats` package v4.3.2 in R. The plot was visualized in R using `corrplot` v0.95.

Biomarker identification using machine learning algorithms

A random forest (RF) classifier model was trained using the presence or absence of all orthogroups to identify important predictors for SCC level. The model was implemented in Python 3 using the `RandomForestClassifier` from the `scikit-learn` module v1.4.2 [62]. To improve computational efficiency, two filtering steps were applied: (i) orthogroups shared by more than 95% and fewer than 5% of isolates were removed, and (ii) one representative was chosen for groups of orthogroups that were present/absent in all of the same isolates. To mitigate the risk of overfitting the model to this dataset, the data were split, and 80% of the isolates were used to train the model, while the remaining 20% of isolates were used to test the resulting model for precision and accuracy.

To differentiate orthogroup presence that was predictive for either low or high SCC (SCC_{low} or SCC_{high}), a random forest clustering method was implemented using the `FGClustering` module v1.1.1 [63] in Python. The resulting RF groups were analysed to determine isolates grouped based on SCC level. These were labelled as RF_{high} and RF_{low} for the group with the highest number of isolates from high and low SCC samples, respectively. Each isolate was designated as SCC_{low}/RF_{high} , SCC_{low}/RF_{low} , SCC_{high}/RF_{low} or SCC_{high}/RF_{high} to indicate its SCC status and RF group. Isolates where SCC and RF group did not match (SCC_{low}/RF_{high} and SCC_{high}/RF_{low}) were considered more likely to have had their SCC status influenced by host physiological state and environmental exposures (such as immune competence, stress responses, management conditions, co-infection or microbiome composition) rather than isolate genomic features and were not used in the next steps.

Associations between important predictive orthogroups (P -value < 0.001) for each group were identified using the frequent pattern growth and association rules learning algorithms (`fpgrowth` and `association_rules`) from the `mlxtend` Python module v0.23.4 [64]. Rules were filtered by *lift* (a measure of how often orthogroups in the set are present in the same isolate than would be expected if they were statistically independent) and *confidence* (the probability of seeing that orthogroup in an isolate given that the other orthogroup is also present) scores, where *lift* ≥ 1.3 and *confidence* ≥ 0.8 . Rules for the RF_{low} group were examined, and those containing an orthogroup also appearing in a rule from the RF_{high} group were removed. The remaining rules with *lift* > 1.5 were used to identify unique combinations of five orthogroups.

Data management was primarily done in Python using `pandas` v2.3.1 and `numpy` v1.24.3. Bar graphs and pie charts were generated in Python using `matplotlib` v3.7.1, and the networks were visualized using the `gravis` v0.1.0 module [65].

Data Availability

Draft genome assemblies and raw reads were deposited under BioProject PRJNA1336871 [<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1336871>]. BioSample accessions for each isolate are provided in Table S1 (available in the online Supplementary Material). Code and documentation are available at <https://github.com/kayleerich/biomARKers>.

RESULTS

Pangenome analysis

After genome assembly and annotation, we assessed the completeness of each assembly using BUSCO and calculated an average completeness score of 98.7% (99.3–100%) (Table S1). Across the 98 isolates, we found an average of 2679 protein-coding genes, 58 tRNA genes and 20 rRNA genes per isolate. We confirmed the species of each isolate via ANI and found that the *S. xylosus* isolates had an average ANI of 98.58%, 93.46% and 91.22% to the *S. xylosus*, *S. pseudoxylosus* and *S. shinii* reference genomes, respectively. ANI between the *S. xylosus* reference genome and the *S. pseudoxylosus* or *S. shinii* isolates was 93.47% and 91.20%, respectively.

We identified 3,672 protein-coding orthogroups in the pangenome of 82 *S. xylosus* isolates (Table S2). From these, we identified genes belonging to the core (present in $\geq 95\%$ of isolates, $n \geq 78$) and accessory genomes (present in $< 95\%$ of isolates, $1 \leq n < 78$), where 2,253 (61.4%) orthogroups were part of the core genome and 1,419 (38.6%) were part of the accessory genome. Of the core genome, we found that only 43 (1.2%) orthogroups were present in 95–99%, designated the soft-core genome ($78 \leq n < 81$). We further divided the accessory genome into the shell (present in 15–95% of isolates, $12 \leq n < 78$) and cloud (present in 0–15% of isolates, $1 \leq n < 12$) genomes, which contained 410 (11.2%) and 1,009 (27.5%) orthogroups, respectively (Fig. 1a). Of the latter, 58 were unique, i.e. found in only one isolate. When genomes were added one at a time, we found that the number of core orthogroups

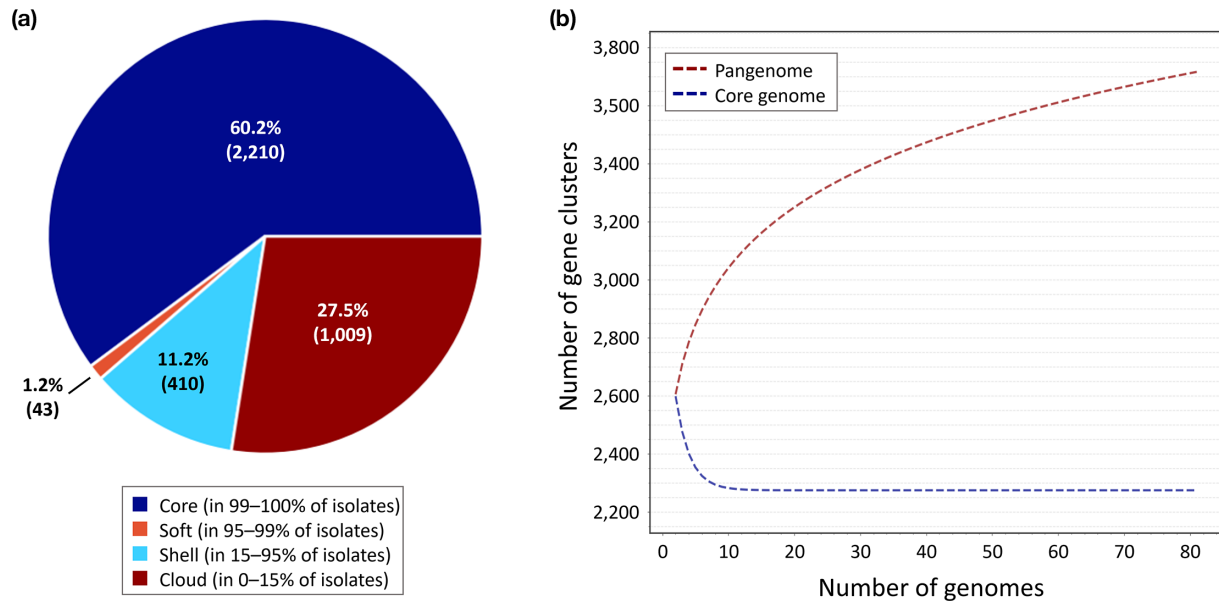


Fig. 1. Pangenome of 82 *S. xylosoy* isolates. (a) Pangenome composition of core (dark blue), soft-core (orange), shell (light blue) and cloud (red) orthogroups. (b) Rarefaction curve for 77 permutations of 82 isolates depicting the change in the total number of orthogroups in the pangenome (red) and the core genome (blue) with the addition of each *S. xylosoy* genome.

decreased until the addition of new genomes had no further effect. The total number of orthogroups in the pangenome continued to increase with each inclusion (Fig. 1b), indicating an open pangenome (Heap's law $\alpha=0.95$).

Phylogenetic analysis and clustering

Using the orthogroups present in more than 95% of the isolates ($n=2,279$), we constructed a phylogenetic tree with PhyloPhlAn (Fig. 2a). From this phylogeny, we identified seven distinct clades, which we labelled A through G. These formed two separate lineages; one comprised of Clades A through E and the other of Clades F and G. We also found that the *S. shini* and *S. pseudoxylosoy* isolates each formed their own outgroup clades, which we collectively refer to as NX (non-*xylosoy*) for the sake of clarity. We then calculated MDS distances for all isolates using orthogroups present in <95% of the isolates and found that, except for Clade E, isolates in a phylogenetic clade clustered together (Fig. 2b).

We observed some clear distinctions between clades. All but one of the NX isolates were SCC_{high} , and these isolates formed a clear group (k -means cluster 1) using genes present in <95% of all isolates (Fig. 2b). Clade F contained both SCC_{low} and SCC_{high} isolates, which were not segregated by subclade. They also did not cluster with Clades A–D after MDS analysis. Instead, we found that Clade F clustered more closely with Clade G and with some of the isolates from Clade E (k -means cluster 2). The remaining Clade E isolates clustered with Clades A–D (k -means cluster 3), which were primarily SCC_{low} .

When we used only virulence factor (VF), AMR or bacteriocin orthogroups to calculate the MDS distances between isolates (Fig. 2c–e), the k -means clusters did not correspond to the phylogeny as strictly. We also noticed that the k -means clusters were not consistent with the number of orthogroups present in each isolate. Though the NX isolates and Clade E had the most AMR genes (10–13 and 9–11, respectively), the isolates from these clades clustered separately, indicating different AMR gene profiles for these groups. Clade A isolates also had more AMR genes than the remaining clades (8–13), and the majority of these isolates clustered with Clade E. Isolates in Clades A, B, C and D were primarily SCC_{low} and had between 38 and 49 virulence factor orthogroups each. All of the isolates in these clades clustered together, indicating similar virulence factor profiles. This was different from our observations of NX and Clade E isolates, which clustered together and had fewer virulence factors – between 32–39 and 32–34, respectively. However, the NX clades are characterized by SCC_{high} isolates, and all but one of the isolates from Clade E were SCC_{low} .

For all previously discussed MDS distance clustering, the NX isolates largely clustered together. Interestingly, we did not observe this trend when we calculated the MDS distances using only the presence of bacteriocin genes (Fig. 2c). The NX isolates had the greatest variation in the number of encoded bacteriocin genes, and those in the same k -means cluster were not the most closely related isolates. Lantibiotic presence in particular varied greatly in NX isolates, ranging from 15 to 19 lantibiotic genes encoded. In comparison, isolates from the other clades encoded a maximum of 17 lantibiotics, and within-clade variation was only 3 for Clade B, 2 for Clade A, 0 for Clade C and 1 for all other clades.

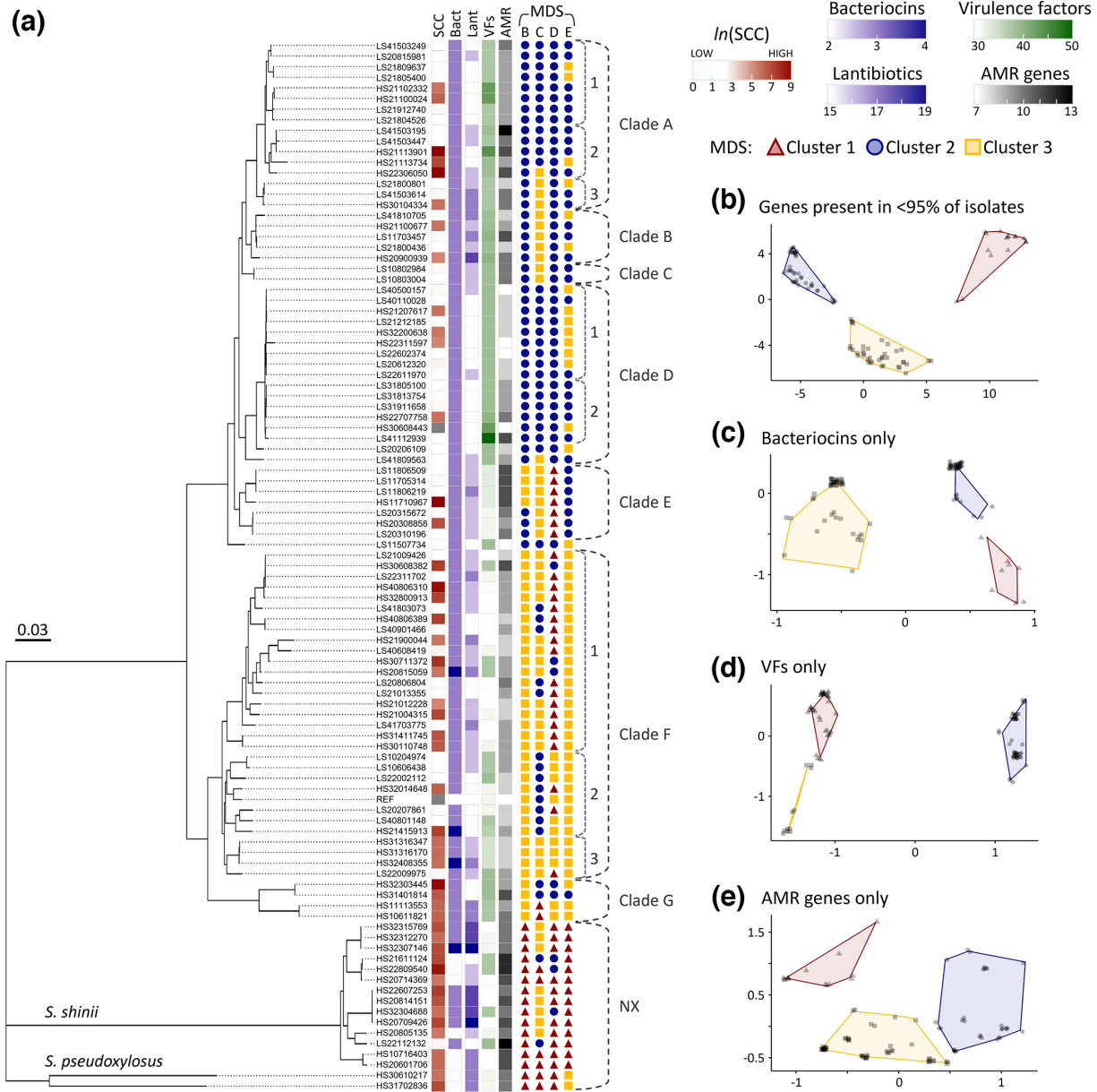


Fig. 2. Phylogeny and MDS clustering of 98 *Staphylococcus* isolates. (a) Phylogenetic tree split into seven distinct *S. xylosus* clades, labelled Clades A through G. *S. shini* and *S. pseudoxylus* clade branches are labelled with species name and collectively referred to as NX. Phylogeny was generated with PhyloPhlan using 1,816 markers comprised of 43,839 aligned positions from proteins present in $\geq 95\%$ of isolates. The heatmap displays the SCC (where darker red indicates higher SCC) for each isolate, and the number of bacteriocins and lantibiotics (purple), virulence factors (green) and AMR (black) orthogroups present in its genome. Symbols correspond to *k*-means clusters of MDS distances in (b–d) (red/triangles, cluster 1; blue/circles, cluster 2; yellow/squares, cluster 3) to visualize whether the different gene categories (accessory genes, bacteriocin genes, virulence genes or AMR genes) follow similar patterns between and within phylogenetic clades and clustering in the MDS plots. MDS distances were calculated as a measure of similarity between isolates using (b) orthogroups present in at least 5% of isolates and less than 95% of isolates, (c) bacteriocin orthogroups only, (d) virulence factor orthogroups only and (e) AMR gene orthogroups only.

Associations between virulence factors, AMR and bacteriocin orthogroups

We did not find a significant correlation between SCC level and number of virulence factors or bacteriocins. We did, however, find a positive correlation between SCC and number of AMR genes ($R^2=0.28$, $P=0.0059$) and number of AMR genes and lantibiotics ($R^2=0.31$, $P=0.0019$) as well as a negative correlation between number of VFs and lantibiotics ($R^2=-0.41$, $P=4.44e-05$).

When we did pairwise comparisons of the presence and associations between bacteriocin-encoding genes, we found that there were some bacteriocins not encoded by any SCC_{low} isolates, and no significant associations between the bacteriocins were present

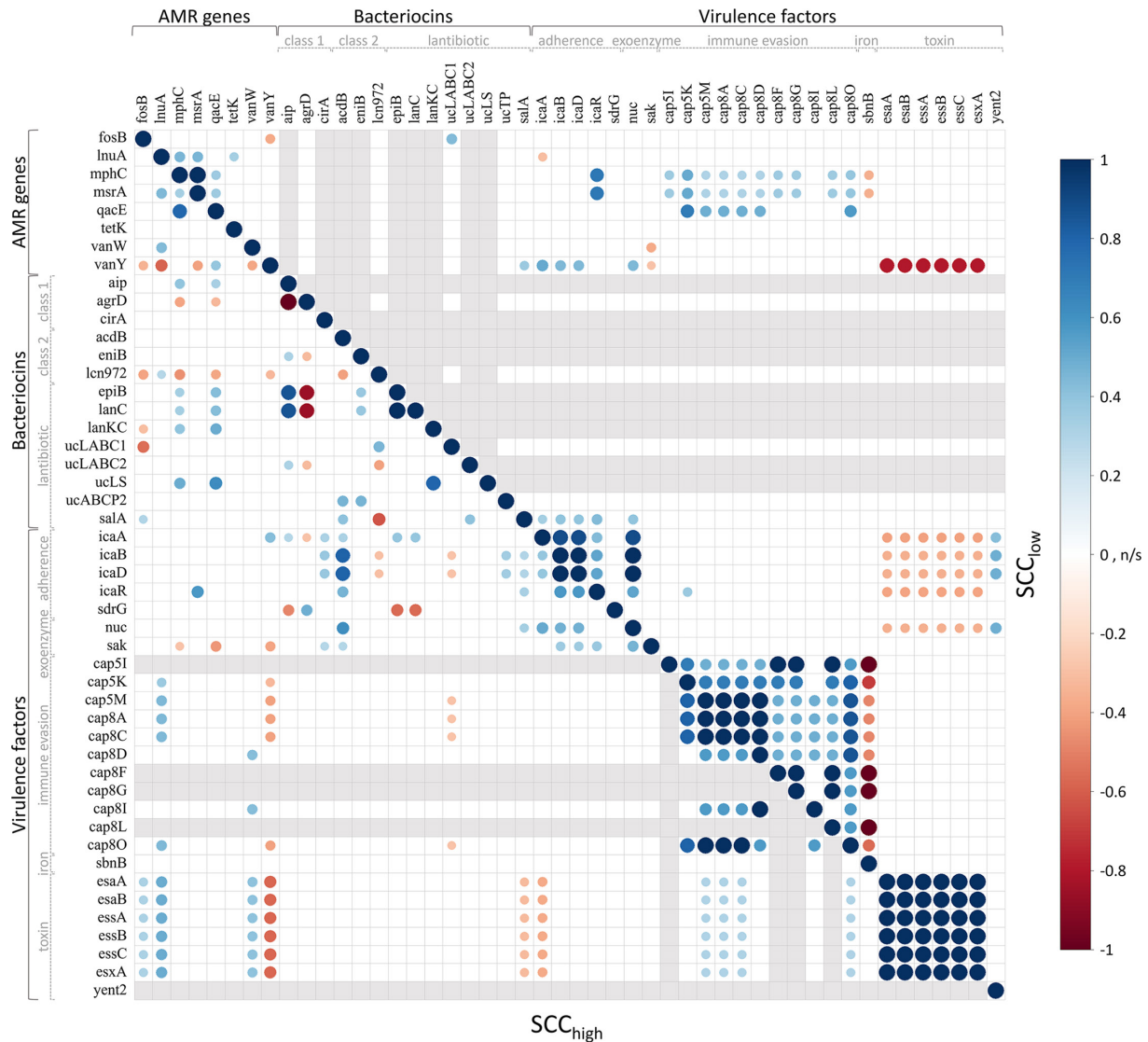


Fig. 3. Significant pairwise associations between AMR genes, bacteriocins and virulence factors for SCC_{high} isolates (bottom left matrix) vs. SCC_{low} isolates (top right matrix). Associations ($P > 0.05$) are represented by phi correlations for the presence of orthogroups, and positive association is represented in blue and negative in red. The column or row is greyed out if the orthogroup was not found in that group of isolates (i.e. SCC_{high} or SCC_{low}). Orthogroup IDs associated with gene symbols can be found in Table S3.

(Fig. 3). In contrast, there were multiple associations between bacteriocins in SCC_{high} isolates. These included positive associations between autoinducing peptide IV (*aip*), enterocin X- β (*eniB*), lantibiotic synthetase component C (*lanC*) and lantibiotic dehydratase/lantibiotic epidermin biosynthesis protein (*epiB*), which were all negatively associated with cyclic lactone autoinducer peptide (*agrD*).

We found that associations between capsule orthogroups such as *cap8A*, *cap8C*, *cap8D* and *cap5M* were prevalent in both SCC_{low} and SCC_{high} isolates. However, other capsule genes (*cap8F*, *cap8G*, *cap8L* and *cap5I*) were not found in isolates from SCC_{high} samples. We also found that another capsule gene, *cap5K*, was positively associated with the quaternary ammonium compound resistance gene *qacE* in SCC_{low} isolates but had a weak negative association in SCC_{high} isolates.

In both SCC_{low} and SCC_{high} isolates, we found that type VII secretion system (T7SS) genes (*esaA*, *esxA*, *essA*, *esaB* and *essC*) were negatively associated with the vancomycin resistance gene *vanY*. These T7SS genes were positively associated with other AMR genes, including another vancomycin resistance gene (*vanW*), a thiol transferase conferring fosfomycin resistance (*fosB*) and lincosamide nucleotidyltransferase (*lnuA*) in SCC_{high} isolates with no significant association in SCC_{low} isolates. *lnuA* was also the only gene significantly positively associated with the tetracycline resistance gene *tetK* and only in SCC_{low} isolates.

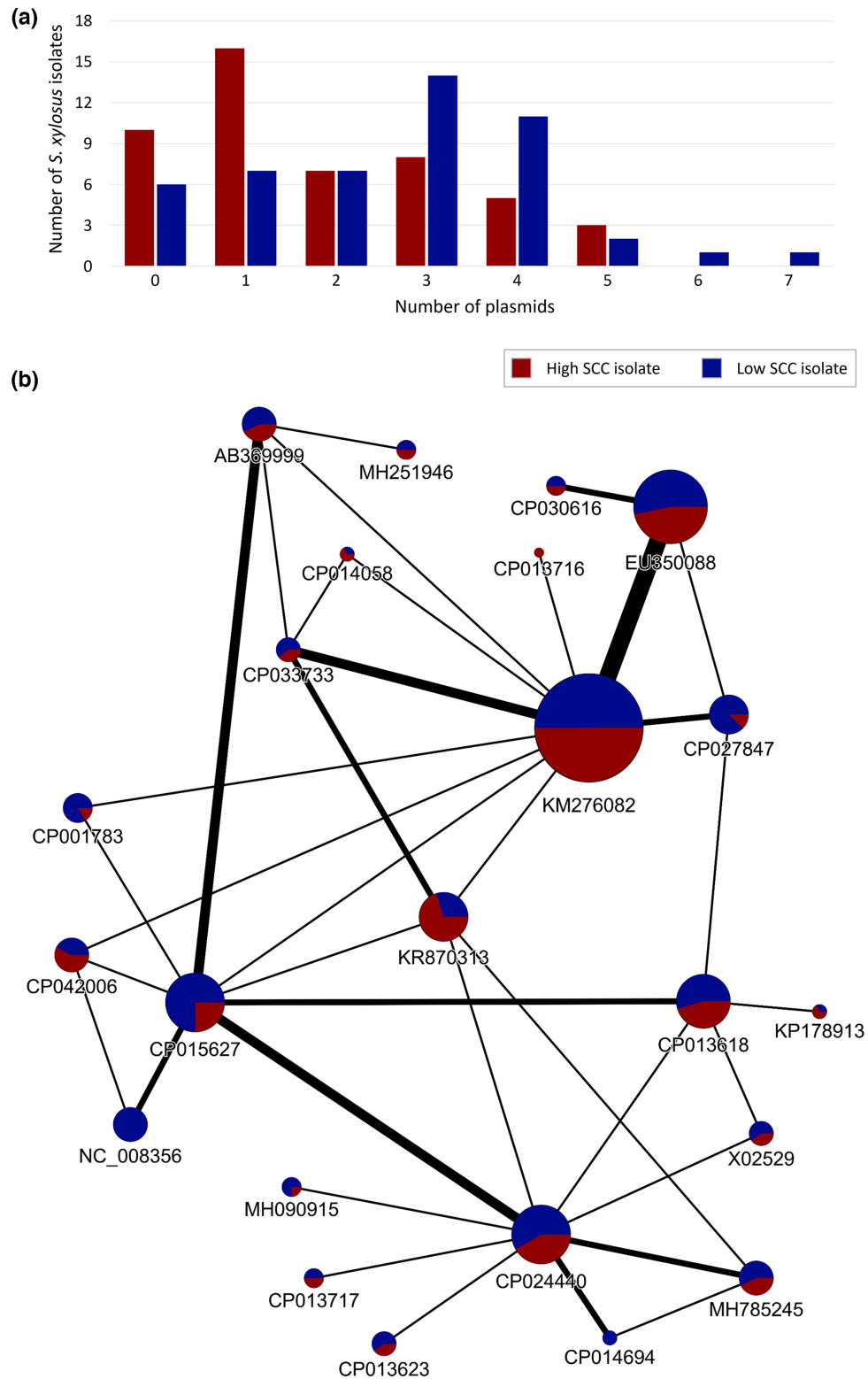


Fig. 4. (a) Number of plasmids per isolate coloured by either SCC_{high} (red) or SCC_{low} (blue). (b) Network of plasmids found in the same isolate. Each node represents a plasmid, with size indicating the number of isolates carrying that plasmid and colours representing the proportion of SCC_{high} (red) or SCC_{low} (blue) isolates. Edges link co-occurring plasmids, with thickness representing the number of isolates in which that pair of plasmids was found. The network has been filtered to remove edges representing a single link and plasmids only found in a single isolate.

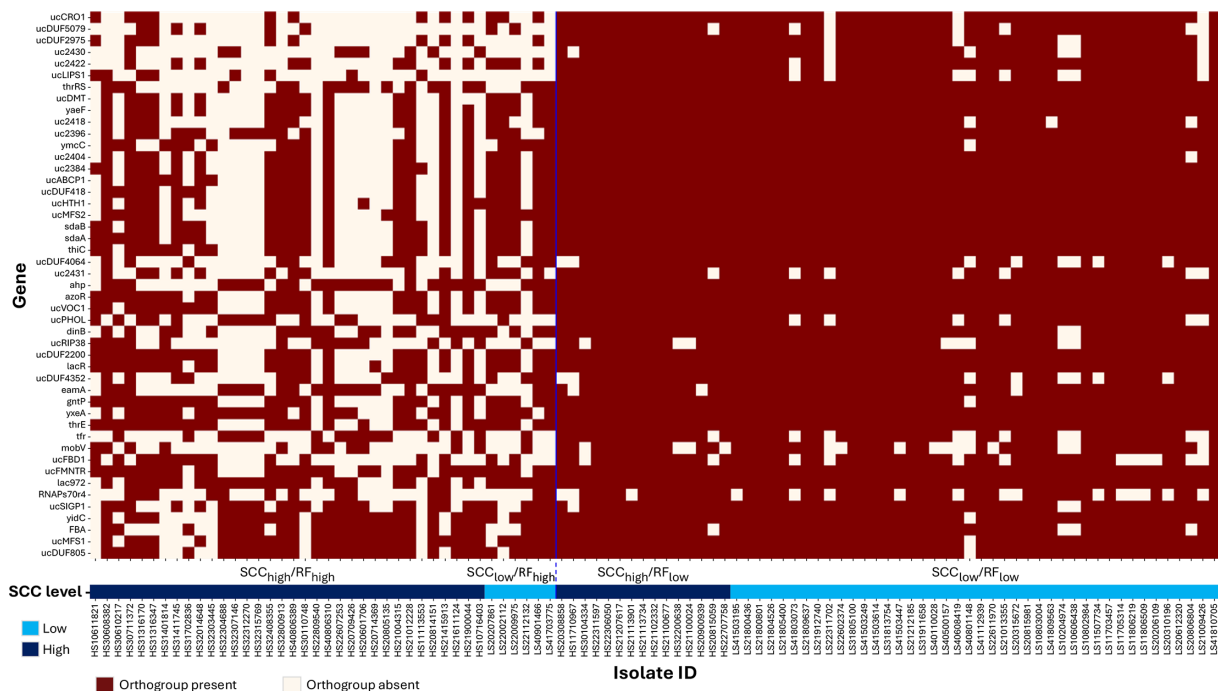
Plasmids

When we examined the association between *lnuA* and *tetK* further, we found that both genes were commonly encoded on plasmids. Though they were never encoded on the same plasmid, we found that 15 isolates (12 SCC_{low}, 3 SCC_{high}) carried a plasmid for each gene. Furthermore, we found that there were no orthogroups solely encoded on plasmids and that SCC_{low} isolates carried more plasmids than SCC_{high} isolates (exact two-sample Kolmogorov–Smirnov test, $P=0.02597$) (Fig. 4a). We also found that the median number of plasmids per isolate was 2 and 16 isolates had no identified plasmids. Though there were plasmids that we found only in SCC_{low} or SCC_{high} isolates, such as NC_008356, which we found in 7 SCC_{low} isolates, there was no single plasmid found in every SCC_{low} or SCC_{high} isolate. We also observed that some plasmids were never carried together by the same isolate, and some plasmids were only carried by *S. xyloso* or NX isolates. For example, we found the plasmids EU350088 and CP024440 in 15 and 12 isolates, respectively, but never in the same isolate (Fig. 4b). Of the two, EU350088 was only found in *S. xyloso* isolates, but CP024440 was found in both *S. xyloso* and NX isolates.

Random forest modelling and clustering

There were no orthogroups that were found in all SCC_{low} isolates and not in any SCC_{high} isolates. Due to the limitations of using SCC as a proxy for NAS-induced IMI, we sought to reduce single-sample noise from external factors by enriching for isolates carrying genes consistently associated with low SCC. To this end, we trained an RF classifier followed by RF-based clustering to identify genes whose presence was an important predictor for SCC_{low} and grouped isolates based on similarity in decision paths within the model. The overall accuracy of our RF model was 0.85, with 0.89 for target (i.e. SCC_{low}) precision and 0.90 for target recall. We identified two groups within the RF model: the first consisted of 34 SCC_{high} isolates and 6 SCC_{low} isolates and the second consisted of 15 SCC_{high} isolates and 43 SCC_{low} isolates. Since the first group contained the majority of SCC_{high} isolates and the second group most of the SCC_{low} isolates, we refer to these as RF_{high} and RF_{low}, respectively. All the NX isolates and ~30% of the *S. xyloso* isolates were included in RF_{high}, indicating that the primary influence on RF clustering was not phylogenetic distance.

We selected orthogroups that were important predictors for RF_{low} ($P<0.001$) as candidate biomarkers for SCC_{low} (Fig. 5). Some notable orthogroups included a TetR/AcrR family transcriptional regulator (*tfr*, OG0002443), lactococcin 972 (*lcn972*, OG0002332), alkylhydroperoxidase (*ahp*, OG0002375) and an FMN-dependent NADH:quinone oxidoreductase (*azoR*, OG0002367).



Frequent patterns and association rules

To find orthogroups that are commonly found in the same isolates (co-occurring biomarkers), we implemented two algorithms: frequent pattern growth (FP) to calculate the frequency of all combinations of orthogroups for each group and association rules machine learning (ARL) to find statistically significant combinations. These two algorithms use a significant amount of computing power; the amount of time and memory required for FP and ARL calculation increases exponentially with the number of items in the dataset. This has posed a substantial obstacle in applying these algorithms to large genomic datasets. Therefore, to reduce the complexity of the input dataset, we chose to only use orthogroups that were identified as important predictors ($P < 0.001$) by the clustered RF model: 80 for RF_{high} and 149 for RF_{low}.

After calculating frequencies and association rules (maximum number of genes per rule=5) for each RF group, we filtered by *lift*, where *lift* > 1 indicates a positive correlation for orthogroup presence in the same isolate, and removed any rules for RF_{low} containing an orthogroup that also appeared in a rule for RF_{high}, since these are not informative for differentiating between SCC_{low} and SCC_{high} isolates. This reduced the number of association rules from 170,369,670 to 161,690, and we identified 6,455 potential biomarker sets created from 63 orthogroups when we filtered for rules with unique combinations of 5 genes and *lift* > 1.5. From the final set of rules, we identified the orthogroups most frequently found in association rules together (Fig. 6, Table S3). Among these, we found multiple notable associated orthogroups, including the T7SS proteins (*esaA*, *esxA*, *essA*, *esaB*, *essB* and *essC*; collectively referred to as T7SS), *tfr*, lysophospholipase (*lpl*) and T7SS effector LXG polymorphic toxin (*t7lxg*).

We further investigated which isolates encoded different combinations of the 63 orthogroups, as isolates encoding these gene combinations may be suitable candidates for use as probiotics. As expected, we found that combinations of these orthogroups were

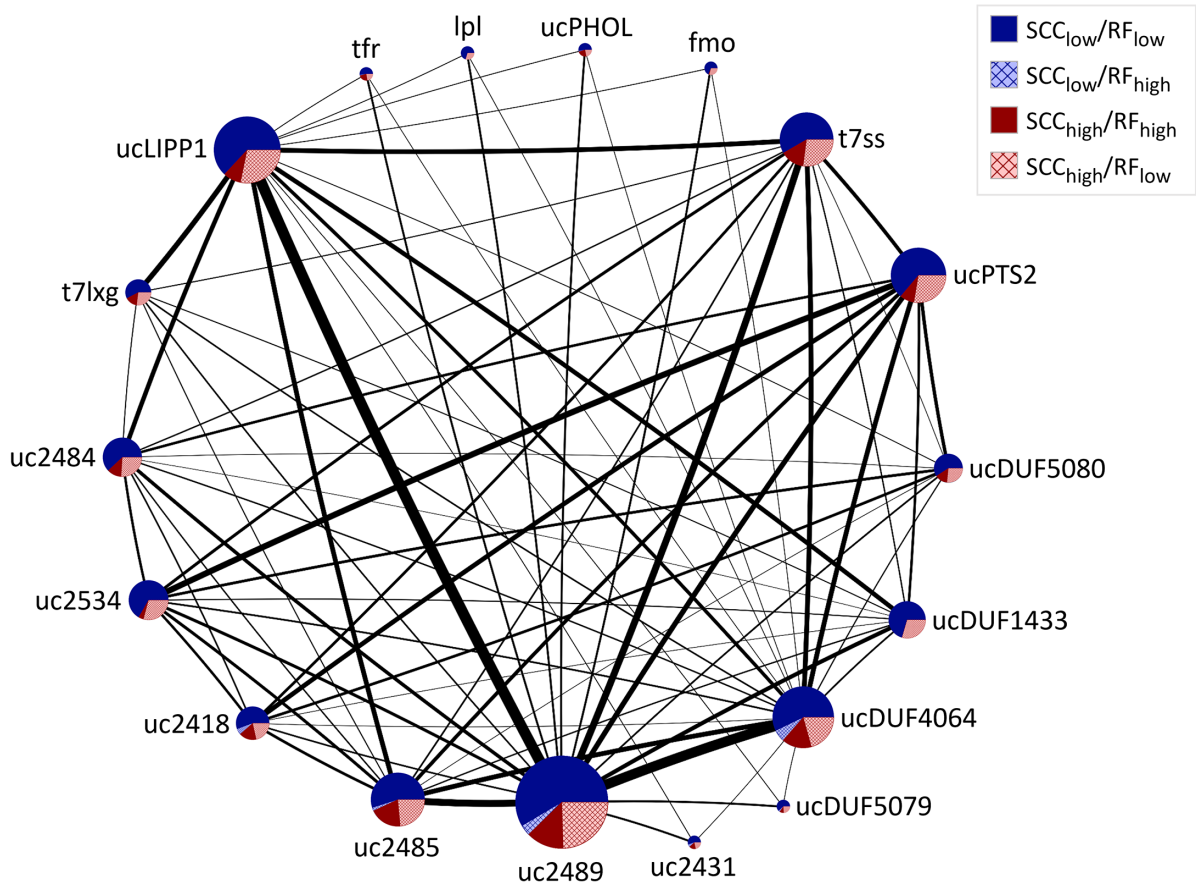


Fig. 6. Network of orthogroups from the set of potential biomarker groups. Nodes represent single orthogroups with colours representing SCC_{low}/RF_{low} (dark blue), SCC_{high}/RF_{high} (dark red), SCC_{low}/RF_{high} (hatched blue) or SCC_{high}/RF_{low} (hatched red) isolates; edges connect orthogroups that appear together in the final set of association rules. Node size is indicative of the relative number of rules containing that orthogroup, and edge thickness indicates the relative number of times those two orthogroups appear in the same rule. Network has been filtered for readability to only include edges and nodes for orthogroups that appeared with more than one of the other 62 orthogroups in ≥ 200 rules. Orthogroup IDs associated with gene symbols can be found in Table S3.

encoded by more SCC_{low} than SCC_{high} isolates. For example, while individual orthogroups containing the genes *t7lxg*, *lac972*, *tfr*, *azoR* and T7SS were common in both SCC groups (in 41–80% of SCC_{high} isolates and 59–100% of SCC_{low}), the co-occurrence of all five was found in just 24% of SCC_{high} isolates but 56% of SCC_{low} isolates. We found that all of the SCC_{high} isolates encoding this combination of orthogroups were from the RF_{low} group, as depicted in Fig. 5. Another gene combination of *tfr*, signal transduction protein TRAP and 2-hydroxy-3-oxopropionate reductase *garR* was encoded by 17 SCC_{high} isolates and 33 SCC_{low} isolates. Three of the SCC_{high} isolates were SCC_{high}/RF_{high}, though the remaining 14 SCC_{high} isolates and all the SCC_{low} isolates belonged to RF_{low}. When we considered isolates that encoded T7SS genes as well, the number of SCC_{high} isolates dropped to 12, all of which belonged to RF_{low}. Similarly, 33 SCC_{low} isolates and 14 SCC_{high} isolates (69% of SCC_{low} isolates compared to only 29% of SCC_{high} isolates) encoded the combination of *fmo*, *lac972* and damage-inducible protein DinB (*dinB*). When we examined these isolates further, we again found that all of them belonged to RF_{low}. Interestingly, 97% of all isolates carry the gene for cyclic lactone autoinducer peptide (*agrD*). However, when we looked at the presence of *agrD* with *t7lxg*, we found that both were present in 77% of the RF_{low} group, but only 20% of RF_{high} isolates carried both.

Though the number of isolates, particularly NX isolates, included in this study is insufficient to determine whether a gene is *S. xyloso*-specific, we investigated whether any of the 63 genes were found in all *S. xyloso* isolates but not in any NX isolates. We found that only one (*thiC*) was carried by every *S. xyloso* isolate and no NX isolates, and two uncharacterized genes (*uc2384* and *ucABCP1*) were found in isolates from every *S. xyloso* phylogenetic clade but not found in any NX isolates.

Probiotic evaluation

To identify the specific isolates with probiotic potential, we examined the phylogenetic relationships, results from our MDS analyses, random forest isolate clustering and the presence of gene combinations identified through association rules learning. We also considered isolates with fewer AMR genes to be preferential due to the positive correlation we found between the number of AMR genes and SCC, but also because a probiotic carrying AMR genes has the potential to transfer those genes to

Table 1. Number of SCC_{high} and SCC_{low} isolates encoding the specified genes and gene combinations

Gene	Gene product	RF _{low} group			RF _{high} group		
		SCC _{high}	SCC _{low}	Total RF _{low}	SCC _{high}	SCC _{low}	Total RF _{high}
<i>tfr</i>	TetR/AcrR family transcriptional regulator	14	33	47	15	1	16
<i>TRAP</i>	Signal transduction protein TRAP	15	42	57	28	5	33
<i>garR</i>	2-hydroxy-3-oxopropionate reductase	15	42	57	28	5	33
<i>T7SS</i>	Type VII secretion proteins	13	28	41	7	0	7
<i>lac972</i>	Lactococcin 972 family bacteriocin	15	42	57	24	6	30
<i>lpl</i>	Lysophospholipase	14	33	47	0	0	0
<i>azoR</i>	FMN-dependent NADH:quinone oxidoreductase	15	42	57	20	5	25
<i>fmo</i>	FAD-dependent monooxygenase	14	33	47	1	0	1
<i>dinB</i>	Damage-inducible protein DinB	15	40	55	18	4	22
<i>t7lxg</i>	T7SS effector LXG polymorphic toxin	14	30	44	8	0	8
<i>ahp</i>	Alkylhydroperoxidase	15	42	57	22	15	37
<i>agrD*</i>	Cyclic lactone autoinducer peptide	15	42	57	30	6	36
Gene combinations							
<i>tfr</i> , <i>TRAP</i> , <i>garR</i> , <i>T7SS</i>		12	28	0	0	0	0
<i>t7lxg</i> , <i>lac972</i> , <i>tfr</i> , <i>azoR</i> , <i>T7SS</i>		12	27	39	0	0	0
<i>ahp</i> , <i>lac972</i> , <i>dinB</i> , <i>t7lxg</i>		14	30	44	2	0	2
<i>t7lxg</i> , <i>lpl</i>		13	29	42	0	0	0
<i>fmo</i> , <i>lac972</i> , <i>dinB</i>		14	33	47	0	0	0
<i>agrD*</i> , <i>t7lxg</i>		14	30	44	8	0	8

*Mentioned in text, not identified as an important predictor of SCC or RF group.

more pathogenic bacteria. Taking these factors into account, we excluded all NX from consideration since there were vastly more SCC_{high} NX isolates than SCC_{low} isolates and all NX isolates belonged to RF_{high}. Similarly, Clades F and G contained more SCC_{high} isolates than SCC_{low} isolates, and no Clade F subclades contained isolates only belonging to RF_{low}, which gave us less confidence in selecting any of these isolates as a potential probiotic. All of the isolates in the remaining clades belonged to RF_{low}. From the remaining clades, we considered similarity in the presence of VFs, bacteriocins or AMR genes to the NX clades, Clade F or Clade G as a possible exclusion factor (i.e. isolates that clustered with Clades F, G or NX in the MDS analyses). Consequently, we excluded Clade E because these isolates clustered with the presence of the NX virulence factor in the MDS analyses.

After these exclusions, we looked at the presence of the aforementioned genes and gene combinations in the isolates of each clade. We found that the T7SS protein-encoding genes were present in all Clade A, B, C and D isolates but missing from all isolates in Clades F2 and E, as well as all NX isolates. Similarly, both *fmo* and *lpl* were missing from isolates in Clades F, G and NX but present in all others. Furthermore, none of the NX, Clade F or G isolates and less than half of the isolates in Clade E contained any of the gene combinations detailed in Table 1. Though it is impractical to compare and document all combinations – we have provided a parsing tool to investigate gene combinations in the supplemental materials – isolates from Clades A1, B and D1 consistently encoded all genes in a given combination, were SCC_{low} with few exceptions and carried fewer AMR genes on average than isolates from other clades.

DISCUSSION

Somatic cell count in milk samples is widely used as an indicator for IMI in dairy cows and serves as a general measure of udder health. Factors such as days in milk, antibiotic treatment history, housing conditions and previous infections can elevate SCC, though infection with mastitis pathogens such as *S. aureus* and *Streptococcus* species is a primary reason for increased inflammation [66–68]. SCC has been used as a metric to distinguish minor pathogenic bacteria causing IMI, including NAS species that are more likely to be mastitis pathogens [13]. Elevated SCC does not always reflect inflammation caused by an isolated bacterial strain, however. In some cases, the primary pathogen may not be detected if it is no longer shed in milk, was eliminated before or during sampling procedures or is non-culturable [69]. Despite these limitations, SCC can provide a useful, albeit imperfect, starting point for exploring bacterial strains associated with reduced inflammatory responses or protective effects [70]. To identify potentially probiotic isolates, we compared the genetic profiles of *S. xylosum*, *S. pseudoxylum* and *S. shinnii* isolates from samples with varying SCC in the corresponding milk samples and aimed to uncover bacterial traits linked to lower inflammation.

We explored the potential isolate-level genetic contributors to SCC variation in *S. xylosum* by dividing our isolates by SCC level (SCC_{high} and SCC_{low}), using the standard threshold for subclinical mastitis [1]. We then employed multiple analyses in our approach; we built a phylogenetic tree to identify isolate groupings based on lineage, while MDS helped visualize similarities in virulence factor, bacteriocin and AMR gene content. To better distinguish isolates that were more likely to influence the level of inflammation from those where SCC levels may have been influenced by external, host or unrelated factors, we used RF classification to identify the genetic features most predictive of SCC category and then grouped isolates based on these predictors using RF clustering. This approach enabled us to focus on isolates with genomic profiles consistently associated with either high or low SCC rather than relying on SCC level as a strict indicator of NAS-triggered inflammation. Although SCC levels associated with our isolates may still partly reflect host or environmental influences, the resulting RF_{high} and RF_{low} groups offered a more nuanced interpretation for the selection of important genes and subsequent probiotic candidate identification. Finally, we used association rule learning to detect combinations of genes that were overrepresented in isolates from SCC_{low} samples with genomic profiles also associated with RF_{low}.

Across all isolates, we observed a positive correlation between SCC and the number of AMR genes. We also found that the number of AMR genes was positively correlated with lantibiotic gene presence. We did not find a significant correlation between SCC level and the number of virulence factors or bacteriocins, however. This suggests that the mere presence of virulence factors or bacteriocins as a group does not directly predict the pathogenicity or protective capacity of an isolate. This is consistent with previous studies, which have not found an association between the abundance of VFs and SCC [71–74]. Rather, specific groups of virulence factors produced by pathogenic bacteria are linked to increased SCC [75], though our sample size limited our ability to examine any association between specific categories of virulence factors and SCC in greater detail. Surprisingly, we observed a negative correlation between the number of virulence factors and lantibiotic genes, and we found that the most common virulence factors were type VII secretion system genes, which have been shown to mediate interbacterial competition in *Firmicutes* and have high inter- and intra-species variation [76]. This inverse relationship could suggest an evolutionary trade-off between broad-spectrum antimicrobial production (e.g. lantibiotics) for more specific niche competition (e.g. T7SS LXG toxins, *t7lxl*). Further research into these correlations, including investigation of historical antimicrobial exposure and the possible co-selection of resistance and competitive traits, is needed.

Regardless, antimicrobial-resistant strains are not an appropriate choice for probiotic development due to the potential transfer of AMR genes to pathogenic bacteria. The correlation between lantibiotics and AMR genes also emphasizes that

finding an effective probiotic is more nuanced than simply identifying bacteriocin-encoding strains. To account for this, we identified gene combinations that were disproportionately present in RF_{low} isolates. Although many of these genes were individually found in both SCC_{high} and SCC_{low} isolates, combinations of two or more were carried primarily by isolates in the RF_{low} group. These findings underscore the importance of examining multiple gene associations over individual gene presence and provide a basis for identifying probiotic potential using multigene (genomic) signatures. Further investigation of these gene combinations may yield valuable insights into traits important for probiotic activity, such as the co-occurrence of genes for niche competition and surviving host immune responses. For example, we found that the niche competition genes *lac972* and *t7lxg* often co-occurred in RF_{low} isolates alongside the oxidative stress resistance gene, *ahp*, and genes coding for lipoproteins, which sometimes modulate immune responses [77, 78]. This combination of genes in a probiotic could contribute either to beneficial, localized immune stimulation or to immune suppression and evasion. Further work examining how these genes interact (e.g. through synergistic effects and changes in expression) may also help explain the observed variation in activity between strains encoding similar antimicrobial gene clusters [24, 27]. Moreover, because some of the identified combinations include genes of unknown function, they offer a promising starting point for characterizing previously unannotated genes.

While our findings are based on statistical associations that have yet to be experimentally validated, bioinformatic methods have played an increasingly important role in characterizing probiotic bacteria [8, 79, 80]. Though there are few examples of large-scale screening for novel candidates, genomic-scale metabolic models have been used to predict candidate probiotic strains for aquaculture applications [81] and for use in human therapeutics [82]. Additionally, in a metagenomic analysis of clinical mastitis infections, Park *et al.* [11] suggest that bacterial species negatively correlated with *S. aureus* may have probiotic potential, though employing metagenomics as a screening strategy requires refinement. We build on the growing collection of screening techniques by using machine learning algorithms alongside traditional genomic analyses to differentiate potentially beneficial *S. xylosus* isolates from those that may exacerbate IMI. To the best of our knowledge, this is the first study to use association rule learning and random forest clustering to identify biomarkers associated with inflammation or to use the absence of inflammation as a selection criterion for probiotic bacteria. It is also the first to apply machine learning and bioinformatic methods specifically for selecting novel probiotic NAS in mastitis research. Our methodology also offers broader potential applications for distinguishing more virulent strains among pathogenic species and identifying novel inflammation-related genes.

In conclusion, we found that isolates classified as SCC_{low}/RF_{low} that carry combinations of the genes we have identified, particularly those from phylogenetic Clades A1, B and D1, warrant further investigation as prospective probiotic candidates. While our findings highlight promising targets, these associations remain exploratory, and both the probiotic potential of the isolates and the functional relevance of the identified gene combinations require experimental validation. Although validation was beyond the scope of this study, the antagonistic activity of these isolates against *S. aureus* is currently being assessed using assays previously developed in the Ronholm lab [25]. These gene combinations may also prove useful for detecting protective strains among other NAS species, and additional gene combinations relevant to udder health may be discovered through similar analytical approaches. This study was limited by the available dataset; to improve the resolution of associations between bacterial genes, host response and protective potential, future studies would benefit from larger sample sizes with more isolates from non-*S. xylosus* species, samples collected across multiple timepoints and the inclusion of metadata such as antimicrobial treatment history and prior infection status.

Funding information

This work was funded through the Natural Sciences and Engineering Research Council of Canada (NSERC) and Fonds de Recherche du Québec – Nature and Technologies sector (FRQNT) NOVA program for junior researchers awarded to J.R. (grant no. 324681). We would also like to acknowledge the support from an NSERC-CREATE Genome Editing for Food Security and Environmental Sustainability (GEFSSES) scholarship awarded to Z.C. and a Postdoctoral Innovation Fellowship from the University of Calgary awarded to K.D.R.

Acknowledgements

We would like to thank all of the dairy producers, animal health technicians and everyone with the Canadian Bovine Mastitis and Milk Quality Research Network (CBMQRN) who participated in the original sample and data collection. We also thank Dr François Malouin for his contributions and role in developing the overarching collaborative project that served as the foundation for this study.

Conflicts of interest

The authors declare that there are no conflicts of interest.

References

- Sharma N, Singh NK, Bhadwal MS. Relationship of somatic cell count and mastitis: an overview. *Asian Australas J Anim Sci* 2011;24:429–438.
- Kerro Dego O, Vidlund J. Staphylococcal mastitis in dairy cows. *Front Vet Sci* 2024;11:1356259.
- Keane OM. Symposium review: Intramammary infections-Major pathogens and strain-associated complexity. *J Dairy Sci* 2019;102:4713–4726.
- Cheng WN, Han SG. Bovine mastitis: risk factors, therapeutic strategies, and alternative treatments - A review. *Asian-Australas J Anim Sci* 2020;33:1699–1713.

5. Rainard P, Foucras G, Fitzgerald JR, Watts JL, Koop G, et al. Knowledge gaps and research priorities in *Staphylococcus aureus* mastitis control. *Transbound Emerg Dis* 2018;65 Suppl 1:149–165.
6. Moradi A, Bidarian B, Mohammadian F, Akbarian F, Kalateh Rahmani H, et al. Trends in antimicrobial resistance of major mastitis-causing pathogens: a nine-year study. *Vet Med Sci* 2025;11:e70417.
7. Pellegrino MS, Froila ID, Natanael B, Gobelli D, Nader-Macias MEF, et al. In Vitro characterization of lactic acid bacteria isolated from bovine milk as potential probiotic strains to prevent bovine mastitis. *Probiotics & Antimicro Prot* 2019;11:74–84.
8. Rahman MM, Siddique N, Gilman MAA, Hasnat S, Haider MG, et al. Genomic and in vitro analysis of *Pediococcus pentosaceus* MBBL4 implicated its therapeutic use against mastitis pathogens and as a potential probiotic. *Probiotics Antimicrob Proteins* 2025;17:3595–3613.
9. Almeida JM de, Maffei JT, Gebara C, Minafra C, Toledo-Silva B, et al. Exploring probiotic potential and antimicrobial properties of lactic acid bacteria from cow's milk. *Appl Food Res* 2024;4:100461.
10. De Buck J, Ha V, Naushad S, Nobrega DB, Luby C, et al. Non-aureus staphylococci and bovine udder health: current understanding and knowledge gaps. *Front Vet Sci* 2021;8:658031.
11. Park S, Jung D, Altshuler I, Kurban D, Dufour S, et al. A longitudinal census of the bacterial community in raw milk correlated with *Staphylococcus aureus* clinical mastitis infections in dairy cattle. *Anim Microbiome* 2022;4:59.
12. Jung D, Park S, Kurban D, Dufour S, Ronholm J. The occurrence of *Aerococcus urinaeequi* and non-aureus staphylococci in raw milk negatively correlates with *Escherichia coli* clinical mastitis. *mSystems* 2024;9:e0036224.
13. Condas LAZ, De Buck J, Nobrega DB, Carson DA, Roy J-P, et al. Distribution of non-aureus staphylococci species in udder quarters with low and high somatic cell count, and clinical mastitis. *J Dairy Sci* 2017;100:5613–5627.
14. Vanderhaeghen W, Piepers S, Leroy F, Van Coillie E, Haesebrouck F, et al. Identification, typing, ecology and epidemiology of coagulase negative staphylococci associated with ruminants. *Vet J* 2015;203:44–51.
15. Pyörälä S, Taponen S. Coagulase-negative staphylococci-emerging mastitis pathogens. *Vet Microbiol* 2009;134:3–8.
16. De Visscher A, Piepers S, Haesebrouck F, De Vlieghe S. Intramammary infection with coagulase-negative staphylococci at parturition: species-specific prevalence, risk factors, and effect on udder health. *J Dairy Sci* 2016;99:6457–6469.
17. Piepers S, De Vlieghe S, de Kruif A, Opsomer G, Barkema HW. Impact of intramammary infections in dairy heifers on future udder health, milk production, and culling. *Vet Microbiol* 2009;134:113–120.
18. Matthews KR, Harmon RJ, Smith BA. Protective effect of *Staphylococcus chromogenes* infection against *Staphylococcus aureus* infection in the lactating bovine mammary gland. *J Dairy Sci* 1990;73:3457–3462.
19. De Vlieghe S, Opsomer G, Vanrolleghem A, Devriese LA, Sampimon OC, et al. In vitro growth inhibition of major mastitis pathogens by *Staphylococcus chromogenes* originating from teat apices of dairy heifers. *Vet Microbiol* 2004;101:215–221.
20. Persson Waller K, Myrenäs M, Börjesson S, Kim H, Widerström M, et al. Genotypic characterization of *Staphylococcus chromogenes* and *Staphylococcus simulans* from Swedish cases of bovine subclinical mastitis. *J Dairy Sci* 2023;106:7991–8004.
21. Qazi S, Middleton B, Muharram SH, Cockayne A, Hill P, et al. N-acylhomoserine lactones antagonize virulence gene expression and quorum sensing in *Staphylococcus aureus*. *Infect Immun* 2006;74:910–919.
22. Paharik AE, Parlet CP, Chung N, Todd DA, Rodriguez EI, et al. Coagulase-negative staphylococcal strain prevents *Staphylococcus aureus* colonization and skin infection by blocking quorum sensing. *Cell Host & Microbe* 2017;22:746–756.
23. Otto M, Echner H, Voelter W, Götz F. Pheromone cross-inhibition between *Staphylococcus aureus* and *Staphylococcus epidermidis*. *Infect Immun* 2001;69:1957–1960.
24. Carson DA, Barkema HW, Naushad S, De Buck J. Bacteriocins of non-aureus staphylococci isolated from bovine milk. *Appl Environ Microbiol* 2017;83:e01015–17.
25. Park S, Classen A, Gohou HM, Maldonado R, Kretschmann E, et al. A new, reliable, and high-throughput strategy to screen bacteria for antagonistic activity against *Staphylococcus aureus*. *BMC Microbiol* 2021;21:189.
26. Brito MAVP, Somkuti GA, Renye JA Jr. Production of antilisterial bacteriocins by staphylococci isolated from bovine milk. *J Dairy Sci* 2011;94:1194–1200.
27. Caddey B, Li M, De Buck B, Han J, Gao HW, et al. Evaluation of *Staphylococcus simulans* inhibition of *Staphylococcus aureus* infection by an in vivo murine mastitis model. *Appl Environ Microbiol* 2025:e0067825.
28. Brouillette E, Goetz C, Droppa-Almeida D, Chamberland S, Jacques M, et al. Secondary *Staphylococcus aureus* intramammary colonization is reduced by non-aureus staphylococci exoproducts. *Microbes Infect* 2022;24:104879.
29. Goetz C, Tremblay YDN, Lamarche D, Blondeau A, Gaudreau AM, et al. Coagulase-negative staphylococci species affect biofilm formation of other coagulase-negative and coagulase-positive staphylococci. *J Dairy Sci* 2017;100:6454–6464.
30. Srithanasuwan A, Pangprasit N, Mektrirat R, Suriyasathaporn W, Chuammitri P. Divergent immune responses to minor bovine mastitis-causing pathogens. *Vet Sci* 2024;11:262.
31. Fry PR, Middleton JR, Dufour S, Perry J, Scholl D, et al. Association of coagulase-negative staphylococcal species, mammary quarter milk somatic cell count, and persistence of intramammary infection in dairy cattle. *J Dairy Sci* 2014;97:4876–4885.
32. Thorberg B-M, Danielsson-Tham M-L, Emanuelson U, Persson Waller K. Bovine subclinical mastitis caused by different types of coagulase-negative staphylococci. *J Dairy Sci* 2009;92:4962–4970.
33. Wuytack A, De Visscher A, Piepers S, Boyen F, Haesebrouck F, et al. Distribution of non-aureus staphylococci from quarter milk, teat apices, and rectal feces of dairy cows, and their virulence potential. *J Dairy Sci* 2020;103:10658–10675.
34. Leroy S, Lebert I, Andant C, Talon R. Interaction in dual species biofilms between *Staphylococcus xylosum* and *Staphylococcus aureus*. *Int J Food Microbiol* 2020;326:108653.
35. Mahmmod YS, Klaas IC, Svennensen L, Pedersen K, Ingmer H. Communications of *Staphylococcus aureus* and non-aureus *Staphylococcus* species from bovine intramammary infections and teat apex colonization. *J Dairy Sci* 2018;101:7322–7333.
36. Reyher KK, Dufour S, Barkema HW, Des Côteaux L, Devries TJ, et al. The national cohort of dairy farms--a data collection platform for mastitis research in Canada. *J Dairy Sci* 2011;94:1616–1626.
37. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, et al. Sustainable data analysis with snakemake. *F1000Res* 2021;10:33.
38. Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *Imeta* 2023;2:e107.
39. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;34:i884–i890.
40. Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data; (n.d.). <https://www.bioinformatics.babraham.ac.uk/projects/fastqc>
41. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.
42. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017;13:e1005595.

43. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* 2018;34:i142–i150.
44. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;29:1072–1075.
45. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol* 2021;38:4647–4654.
46. Larralde M. Pyrodigal: python bindings and interface to Prodigal, an efficient method for gene prediction in prokaryotes. *JOSS* 2022;7:4296.
47. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 2013;29:2933–2935.
48. Chan PP, Lin BY, Mak AJ, Lowe TM. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res* 2021;49:9077–9096.
49. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, et al. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb Genom* 2021;7:000685.
50. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom* 2018;4.
51. van Heel AJ, de Jong A, Song C, Viel JH, Kok J, et al. BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. *Nucleic Acids Res* 2018;46:W278–W281.
52. Alcock BP, Huynh W, Chalil R, Smith KW, Raphenya AR, et al. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the comprehensive antibiotic resistance database. *Nucleic Acids Res* 2023;51:D690–D699.
53. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 2019;20:238.
54. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 2015;16:157.
55. Ferrés I, Iraola G. An object-oriented framework for evolutionary pangenome analysis. *Cell Rep Methods* 2021;1:100085.
56. Snipen L, Liland KH. micropan: an R-package for microbial pangenomics. *BMC Bioinformatics* 2015;16:79.
57. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;28:3150–3152.
58. Wei Z-G, Chen X, Zhang X-D, Zhang H, Fan X-G, et al. Comparison of methods for biological sequence clustering. *IEEE/ACM Trans Comput Biol Bioinform* 2023;20:2874–2888.
59. Segata N, Börnigen D, Morgan XC, Huttenhower C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat Commun* 2013;4:2304.
60. Yu G, Smith DK, Zhu H, Guan Y, Lam TT. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 2017;8:28–36.
61. Yu G. Using ggtree to visualize data on tree-like structures. *CP in Bioinformatics* 2020;69:e96.
62. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, et al. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;12:2825–2830.
63. Barros de Andrade e Sousa L, Pelin H, Thalmeier D, Piraud M. Forest-guided clustering - explainability for random forest models. 2023. <https://doi.org/10.5281/ZENODO.7823042>
64. Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. *SIGMOD Rec* 2000;29:1–12.
65. Haas R. gravis; (n.d.). <https://github.com/robert-haas/gravis> [accessed 20 June 2025].
66. Zadoks RN, Allore HG, Barkema HW, Sampimon OC, Wellenberg GJ, et al. Cow- and quarter-level risk factors for *Streptococcus uberis* and *Staphylococcus aureus* mastitis. *J Dairy Sci* 2001;84:2649–2663.
67. Breen JE, Bradley AJ, Green MJ. Quarter and cow risk factors associated with a somatic cell count greater than 199,000 cells per milliliter in United Kingdom dairy cows. *J Dairy Sci* 2009;92:3106–3115.
68. Fadillah A, van den Borne BHP, Poetri ON, Hogeveen H, Slijper T, et al. Evaluation of factors associated with bulk milk somatic cell count and total plate count in Indonesian smallholder dairy farms. *Front Vet Sci* 2023;10.
69. Zorah KT, Daniel RC, Frost AJ. Detection of bacterial antigens in milk samples from clinical cases of bovine mastitis in which culture is negative. *Vet Rec* 1993;132:208–210.
70. Paduch J-H, Lücking J, Mansion-de Vries E, Zinke C, Wente N, et al. Prevention of intramammary infections by prepartum external application of a teat dip containing lactic acid bacteria with antimicrobial properties in dairy heifers. *Pathogens* 2020;9:288.
71. Haveri M, Roslöf A, Rantala L, Pyörälä S. Virulence genes of bovine *Staphylococcus aureus* from persistent and nonpersistent intramammary infections with different clinical characteristics. *J Appl Microbiol* 2007;103:993–1000.
72. Naushad S, Nobrega DB, Naqvi SA, Barkema HW, De Buck J. Genomic analysis of bovine *Staphylococcus aureus* isolates from milk to elucidate diversity and determine the distributions of antimicrobial and virulence genes and their association with mastitis. *mSystems* 2020;5:e00063-20.
73. Naushad S, Naqvi SA, Nobrega DB, Luby C, Kastelic JP, et al. Comprehensive virulence gene profiling of bovine non-*aureus* staphylococci based on whole-genome sequencing data. *mSystems* 2019;4:e00098-18.
74. Ävall-Jääskeläinen S, Taponen S, Kant R, Paulin L, Blom J, et al. Comparative genome analysis of 24 bovine-associated *Staphylococcus* isolates with special focus on the putative virulence genes. *PeerJ* 2018;6:e4560.
75. Tong X, Barkema HW, Nobrega DB, Xu C, Han B, et al. Virulence of bacteria causing mastitis in dairy cows: a literature review. *Microorganisms* 2025;13:167.
76. Spencer BL, Doran KS. Evolving understanding of the type VII secretion system in Gram-positive bacteria. *PLoS Pathog* 2022;18:e1010680.
77. Skabytska Y, Wölbing F, Günther C, Köberle M, Kaesler S, et al. Cutaneous innate immune sensing of Toll-like receptor 2-6 ligands suppresses T cell immunity by inducing myeloid-derived suppressor cells. *Immunity* 2014;41:762–775.
78. Gong Z, Yu Z, Ren P, Zhang S, Gao R, et al. An underlying mechanism of bovine mastitis: PGE₂ regulates *Staphylococcus aureus*-induced inflammatory response through TLR2, TLR4, and NLRP3 in macrophages. *Vet Q* 2026;46:2615759.
79. Kapse NG, Engineer AS, Gowdaman V, Wagh S, Dhakephalkar PK. Functional annotation of the genome unravels probiotic potential of *Bacillus coagulans* HS243. *Genomics* 2019;111:921–929.
80. Fakruddin Md, Karim MdAB, Ul Islam B, Safa A, Sultana J, et al. Genomic face-off: an in silico comparison of the probiotic potential of *Lactobacillus* spp. and *Akkermansia muciniphila*. *CG* 2025;26:129–143.
81. Devika NT, Jangam AK, Katneni VK, Patil PK, Nathamuni S, et al. In Silico prediction of novel probiotic species limiting pathogenic *Vibrio* growth using constraint-based genome scale metabolic modeling. *Front Cell Infect Microbiol* 2021;11:752477.
82. Valderrama B, Daly I, Gunnigle E, O'Riordan KJ, Chichlowski M, et al. From in silico screening to in vivo validation in zebrafish – a framework for reeling in the right psychobiotics. *Food & Fun* 2025;16:2018–2030.