BMC Genomics

RESEARCH

Open Access

Inferring *Staphylococcus aureus* host species and cross-species transmission from a genome-based model

Wenyin Du[1], Sitong Chen[1], Rong Jiang[1], Huiliu Zhou[1], Yuehe Li[1], Dejia Ouyang[1], Yajie Gong[1], Zhenjiang Yao[1] and Xiaohua Ye[1*]

Abstract

Background *Staphylococcus aureus* is an important pathogen that can colonize humans and various animals. However, the host-associated determinants of *S. aureus* remain uncertain, which leads to difficulties in inferring its host species and cross-species transmission. We performed a 3-stage genome-wide association study (discovery, confirming, and validation) to compare genetic variation between pig and human *S. aureus*, aiming to elucidate the host-specific genetic elements (k-mers).

Results After 3-stage association analyses, we found a subset of 20 consensus-significant host-associated k-mers, which are significantly overrepresented in a specific host. Surprisingly for host prediction, both the final model with the top 5 k-mers and the simplest model with only the most important k-mer achieved a high classification accuracy of 98%, giving a simple target for predicting host species and cross-species transmission of *S. aureus*. The final classifier with the top 5 k-mers revealed that 97.5% of *S. aureus* isolates from livestock-exposed workers were predicted as pig origin, suggesting a high cross-species transmission risk. The time-based phylogeny inferred the cross-species transmission directions, indicating that ST9 can cross-species spread from animals to humans while ST59 can cross-species spread in the opposite direction.

Conclusion Our findings provide novel insights into host-associated determinants and an accurate model for inferring *S. aureus* host species and cross-species transmission.

Keywords *Staphylococcus aureus*, Bacterial genomes, Genome-wide association study, Cross-species transmission

*Correspondence:
Xiaohua Ye
smalltomato@163.com
[1]Laboratory of Molecular Epidemiology, School of Public Health, Guangdong Pharmaceutical University, Guangzhou, Guangdong, China

Background

Staphylococcus aureus (*S. aureus*) has been regarded as an important human pathogen associated with asymptomatic carriage as well as various diseases ranging from mild skin infections to life-threatening bacteremia, which can result in significant morbidity and mortality [1]. Importantly, *S. aureus* can readily cross species barriers to colonize humans and various animals, with high colonization rates (30–80% for humans, 77% for cows, and 43% for pigs) [2–5]. Previous molecular epidemiological studies have revealed that there are different origins of *S. aureus* isolates, including hospital-, community-, and livestock-associated *S. aureus* (LA-SA) [6, 7]. In addition, there is increasing evidence that zoonotic *S. aureus* can directly and indirectly transmit to humans via occupational animal exposure and multiple environmental pathways, suggesting that the distinction between LA-SA and non-LA-SA has become more and more blurred [8–10]. To identify and prevent cross-species transmission of *S. aureus*, it is urgently needed to elucidate host-specific genetic variants and genes so as to distinguish animal from human isolates.

With the decreasing costs as well as increasing accessibility of high-throughput sequencing, whole-genome sequencing (WGS) has unprecedented power for revealing subtle genomic variants including single nucleotide polymorphisms (SNPs) and pathogenicity genes, which has become a useful tool for bacterial evolution and traceability analyses [11]. In parallel, genome-wide association studies (GWAS) have been increasingly used to explore genotype-phenotype associations so as to identify genetic variation associated with bacterial phenotypes, which may provide insights into preventive and therapeutic interventions [12–15]. Currently, SNPs are the most universal markers for GWAS analyses [16]. However, traditional SNP-based GWAS methods rely on a single reference genome, which can only identify partial genomic variation [17]. The GWAS method based on k-mers (DNA words of length k) has several advantages. First, k-mers do not rely on the reference genomes and can capture various genetic variants and genes associated with bacterial phenotypes [18]. Second, the k-mer-based GWAS offers a useful way to evaluate variation due to SNPs, insertions/deletions, and structural variants [19–21]. Finally, the k-mer-based GWAS infers phenotype-associated variation from raw genome data and therefore reduces potential error-prone variants [18]. Considering the above advantages of k-mer-based GWAS, it has been used to identify specific biomarkers in microbes and humans [22].

Identifying and preventing cross-species transmission of *S. aureus* requires an understanding of the host-specific genetic elements that can infer *S. aureus* host species and distinguish animal from human isolates. Therefore, we employed a 3-stage k-mer-based GWAS analysis (discovery, confirming, and external validation) to compare genomic differences between pig and human *S. aureus* isolates, aiming to identify host-associated genetic elements and predict the host origin of *S. aureus*. In addition, we elucidated the cross-species transmission risk using the k-mer prediction model and inferred the cross-species transmission direction using the Bayesian evolution analysis.

Methods

Sample collection and quality control

Genome assemblies of *S. aureus* isolates collected between 2002 and 2021 in China were downloaded from the NCBI GenBank database (Supplementary Table S1), including 309 pig isolates and 343 human isolates for identifying host-associated variants. The following basic information for each isolate was collected: specimen source, health state of the host, location, and time of sample collection. The eligibility criteria for *S. aureus* isolates included: (1) they provided basic information; (2) they are collected between 2002 and 2021 in China; (3) there are ≥95% of the total sequence belonged to *S. aureus*; and (4) there are ≥90% genome completeness and <10% contamination. *S. aureus* isolates from pig hosts (defined as pig isolates) were sampled from the nasopharynx. *S. aureus* isolates from humans without occupational livestock exposure (defined as human isolates) were sampled from body sites such as the skin, nasopharynx, and anal swabs in asymptomatic individuals. The assembled genomes were analyzed for species annotations to identify the taxonomy of strains using Kraken v.2.1.1 (https://github.com/DerrickWood/kraken2). To further ensure the genome quality, the genome completeness and contamination were assessed using CheckM v.1.0.13 (https://github.com/Ecogenomics/CheckM).

Sequence typing

For *S. aureus* multi-locus sequence typing, genomic sequences were uploaded to the PubMLST database (http://pubmlst.org/saureus/) to determine sequence types (STs) based on the allelic profile of seven housekeeping genes (*arcC*, *aroE*, *glpF*, *gmk*, *pta*, *tpi*, and *yqiL*). The STs were clustered into specific clonal complexes (CCs) using PHYLOViZ v.2.0.0 [23]. Staphylococcal protein A (*spa*) types were inferred using SpaTyper v.1.0 (https://cge.food.dtu.dk/services/spaTyper-1.0/).

Phylogenetic analyses and Bayesian evolution analyses

The whole-genome alignment for the variant sites with SNPs was generated by mapping reads against the *S. aureus* NCTC 8325 reference genome (GenBank accession: NC_007795) using Snippy v.4.6.0 (https://github.com/tseemann/snippy). After screening and removing

recombination regions for the generated whole-genome alignment using Gubbins v.2.3.5 (https://github.com/nickjcroucher/gubbins), the recombination-free maximum likelihood phylogenetic tree was constructed using RaxML v.7.0.4, with a GTR + Γ(Gamma) model and 100 bootstrap replicates for each run [24]. The phylogeny was then visualized and annotated using the ChiPlot online tool (https://www.chiplot.online/).

Counting and annotating k-mers

In order to reveal the host-associated variants of *S. aureus*, we used the alignment-free method based on k-mers, which are computationally efficient and can identify whole-genome variants without the reference genome. Unique k-mers of length 9–100 bp were identified in each assembly using fsm-lite (https://github.com/nvalimak/fsm-lite), and then k-mers were annotated by mapping reads to the reference genomes (RF122 [GenBank accession: NC_007622], MRSA252 [NC_002952], M013 [NC_016928], NCTC 8325 [NC_007795], MSSA476 [NC_002953], Mu3 [NC_009782], Newman [NC_009641], and N315 [NC_002745]) using bwa v.0.7.17 (https://github.com/lh3/bwa). In addition, the gene ontology (GO) annotations were identified using the UniProt (https://beta.uniprot.org/), in which k-mers were divided into three functional GO categories (biological process, molecular function, and cellular component).

Three-stage GWAS analyses of identifying host-associated k-mers

Due to high-dimensional and high-correlated genomic data, a 3-stage GWAS analysis process (discovery, confirming, and external validation; Fig. 1) was performed to identify host-associated k-mers by multiple GWAS methods including the linear mixed model (LMM), Scoary, least absolute shrinkage and selection operator regression (LASSO), and extreme gradient boosting (XGBoost) [25–28], which can avoid over-fitting and reduce the model complexity. In the GWAS models, we used the k-mer matrix (presence or absence) as the independent variable and the *S. aureus* host species as the outcome variable (pig or human). In the discovery stage, we employed the univariate LMM (Pyseer v.1.2.0) to test for genetic associations between k-mers and host species so as to initially screen significant host-associated k-mers, which can control for the important covariate (study time) as a fixed effect and the clone population structure in terms of a genetic-relatedness matrix as a random effect to improve the power in bacterial association analyses [19]. The resulting QQ plot confirmed that the

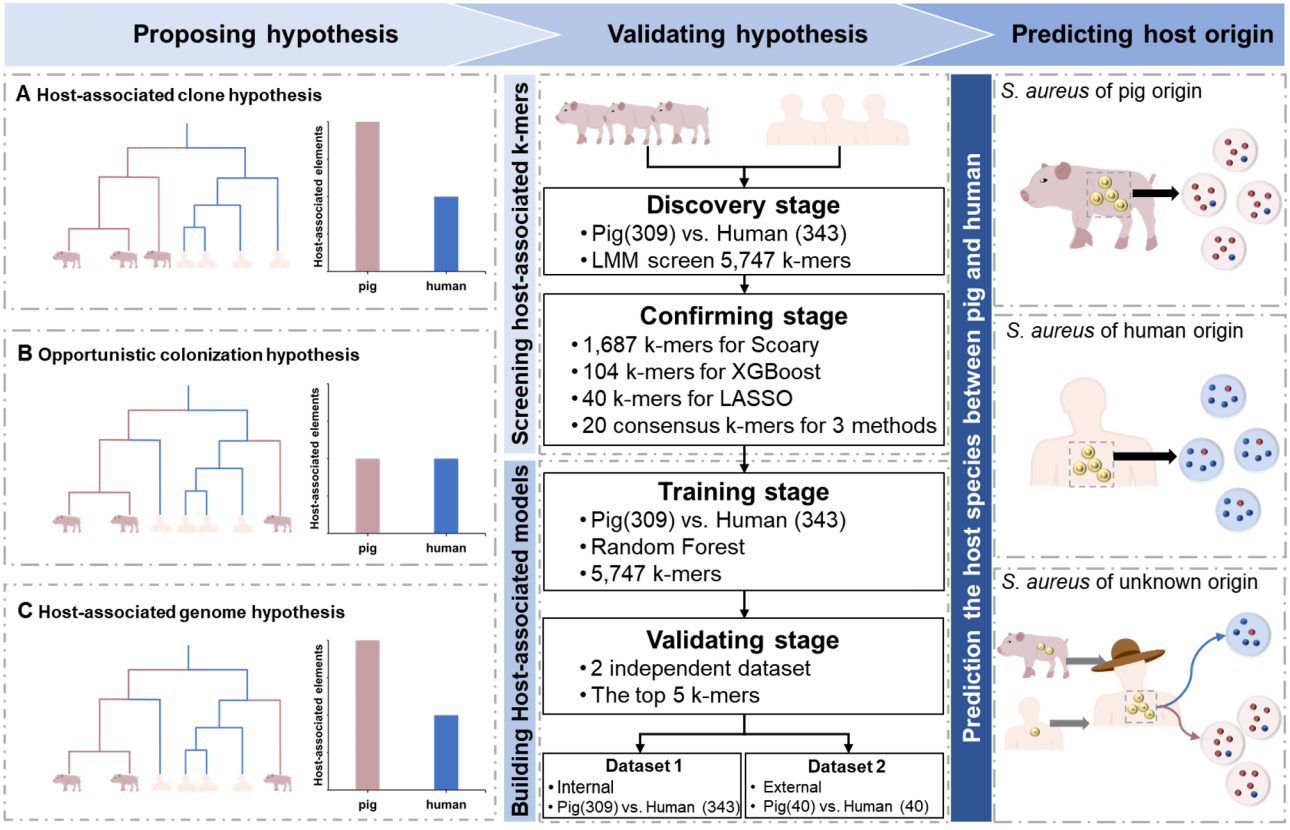

Fig. 1 Flow chart of host-associated colonization hypotheses and statistical analyses. LMM, linear mixed model; LASSO, least absolute shrinkage and selection operator; XGBoost, extreme gradient boosting

population structure was well controlled. In the confirming stage, we used multiple methods (Scoary, LASSO, and XGBoost) to further identify consensus host-associated k-mers, aiming to obtain a simple and accurate prediction model. Scoary is a widely applicable and ultrafast software for bacterial GWAS analyses, which can use the phylogenetic structure to adjust bacterial population structure [26]. LASSO regression can effectively reduce the model complexity and prevent over-fitting by shrinking the coefficients toward zero, which is suitable for the high-correlated and high-dimensional genome data [29]. XGBoost is a highly effective machine learning method based on a scalable end-to-end tree boosting system, which can take advantage of out-of-core computing and smoothly scale to handle massive and high-dimensional genome data [28]. Bonferroni correction (α/N) was used to reduce the probability of false positive rates due to multiple testing of 375,673 k-mers (adjusted *P* threshold being 1.33×10^{-7}). In the external validation stage, we conducted a validation analysis using Random Forest (RF) on an independent dataset available from the NCBI GenBank database (40 pig vs. 40 human isolates; Supplementary Table S2). We built RF classifiers to evaluate the power of prediction models with different k-mer combinations using the "randomForest" R package (https://cran.r-project.org/doc/Rnews/) [30]. We used resubstitution and ten-fold cross-validation estimations to evaluate the power of prediction models with different k-mer combinations. The importance of the k-mers was sorted by the Mean Decrease Gini (MDG).

Table 1 Association analysis between predominant genotypes and host species

Genotypes(*n*)	Pig isolates, no.(%), *n* = 309	Human isolates, no.(%), *n* = 343	x^2	*P* value
Clonal complexes(CCs)/sequence types(STs)				
CC9(277)	277(89.6)	0(0.0)	534.60	< 0.001
ST9(229)	229(74.1)	0(0.0)	391.81	< 0.001
ST3597(41)	41(13.3)	0(0.0)	48.57	< 0.001
CC59(84)	1(0.3)	83(24.2)	82.56	< 0.001
ST59(72)	1(0.3)	71(20.7)	68.70	< 0.001
CC398(66)	16(5.2)	50(19.8)	15.79	< 0.001
ST398(55)	16(5.2)	39(11.4)	8.07	0.005
CC239(44)	2(0.6)	42(12.2)	34.74	< 0.001
ST239(44)	2(0.6)	42(12.2)	34.74	< 0.001
CC5(35)	3(1.0)	32(9.3)	22.36	< 0.001
ST5(24)	3(1.0)	21(6.1)	12.17	< 0.001
CC508(22)	0(0.0)	22(6.4)	20.51	< 0.001
ST45(18)	0(0.0)	18(5.2)	16.68	< 0.001
Spa types				
t899(266)	266(86.1)	0(0.0)	498.74	< 0.001
t437(61)	1(0.3)	60(17.5)	56.50	< 0.001
t30(31)	2(0.7)	29(7.5)	21.88	< 0.001

Notes: Data are presented as no. (%) or as otherwise indicated

Predicting the cross-species transmission risk and direction

Considering the potential risk of cross-species transmission of LA-SA by occupational livestock exposure, we used the RF classifier to predict the host species of unknown-origin *S. aureus* isolated from farm workers with occupational pig exposure (Supplementary Table S3). The cross-species evolution directions were inferred by BactDating v1.1.1 (https://xavierdidelot.github.io/BactDating/), which can perform Bayesian inference of ances tral dates and the root location on a time-based phylogenetic tree.

Results

Characteristics of *S. aureus* isolates

In this study, we analyzed the genomes of 652 *S. aureus* isolates including 309 pig isolates and 343 human isolates (Fig. 1). All *S. aureus* isolates were collected from China between 2002 and 2021, mainly from Hubei ($n = 447$), Shanghai ($n = 65$), and Shandong ($n = 54$). All pig isolates were from nasal swabs; and human isolates were from nasal swabs (323 isolates), anal swabs (16 isolates), and skin swabs (4 isolates). Among all *S. aureus* isolates, we identified 56 unique STs belonging to 31 CCs inferred from whole-genome sequences. The most common genotype for pig isolates was CC9 (ST9), while the predominant CCs for human isolates were non-CC9 clones including CC59 (ST9), CC398 (ST398), and CC239(ST239). With regards to spa typing, the predominant spa type for pig isolates was t899, but the major spa type for human isolates was t437 (Table 1).

Identification of host-associated genotypes by association analysis

In terms of CCs and STs, we observed that livestock-associated CC9 (89.6% vs. 0.0%), ST9 (74.1% vs. 0.0%), and ST3597 (a single-locus variant of ST9, 13.3% vs. 0.0%) were more prone to colonize pigs successfully; while other clones were more prone to colonize humans, including CC59 (0.3% vs. 24.2%), CC398 (5.2% vs. 19.8%), CC239 (0.6% vs. 12.2%), CC5 (1.0% vs. 9.3%), and CC508 (0.0% vs. 6.4%). Moreover, there were significant differences in the proportion of specific spa types (t899, t437, and t30) between pig and human isolates (all $P < 0.05$), with t899 only found in pig isolates. The RF classifiers for predicting host species (pig or human) reached classification accuracy of 94.48% for the model based on all 56 ST genotypes and 87.73% for a single ST9 predictor (ST9 or non-ST9), suggesting that ST9 is associated with livestock specificity. However, the phylogenetic tree based on core SNPs revealed that pig isolates clustered in the same clones with human isolates (Fig. 2), indicating that

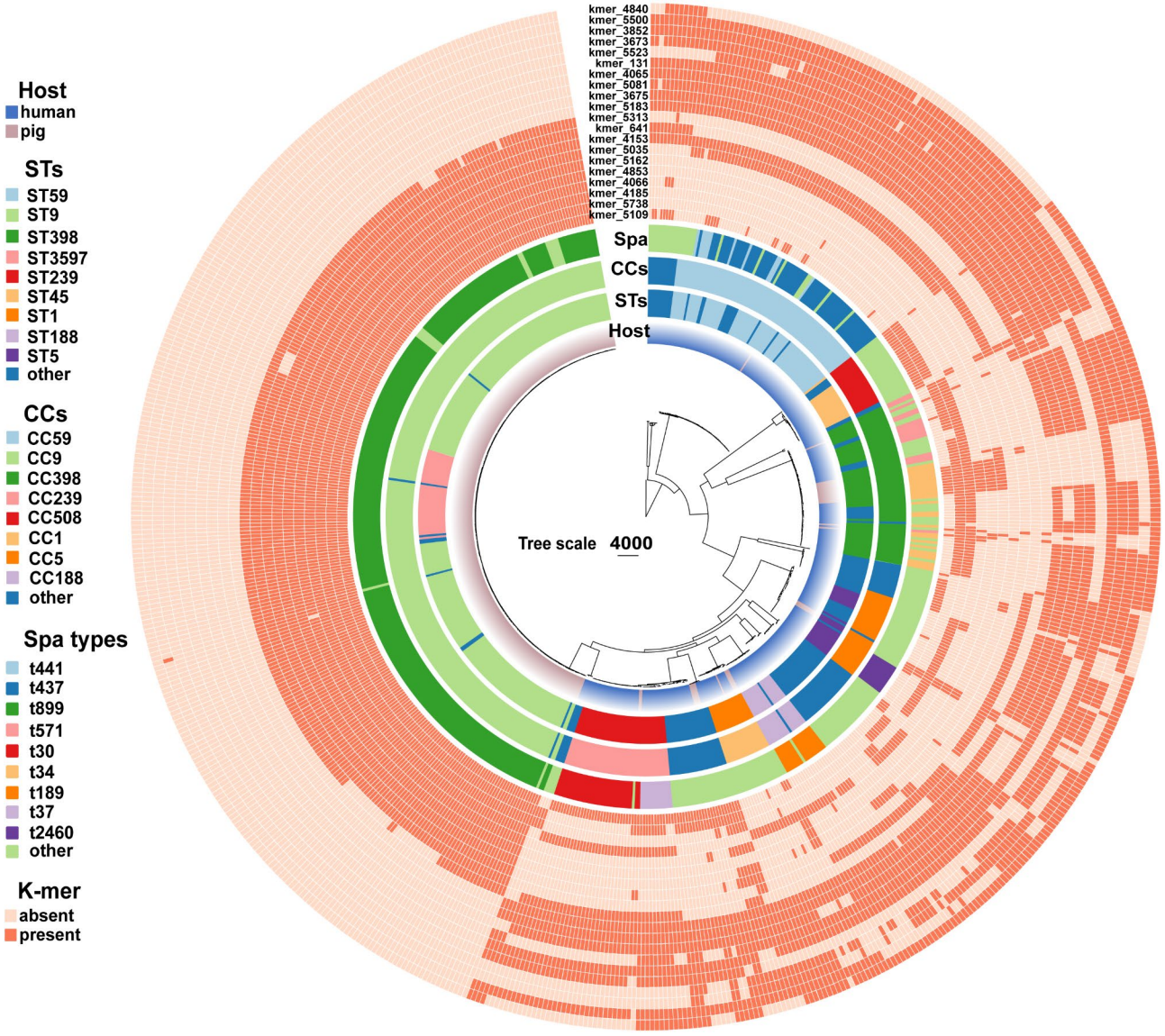

Fig. 2 Whole-genome phylogenetic tree showing genetic relatedness of 652 *S. aureus* isolates. The colored strips at the tips of the tree (from inner to outer) represent isolate metadata (host, STs, CCs, and spa types) and the presence/absence of host-associated k-mers

traditional genotyping techniques provide little power to reveal subtle genetic differences between pig and human isolates.

Discovery of host-associated k-mers by LMM

Based on the assemblies of 652 *S. aureus* isolates, we identified 24,670,041 k-mers. After filtering out low frequency k-mers, 375,673 k-mers were subjected to the k-mers-based GWAS analysis. In the discovery stage, we use the univariate LMM to preliminarily screen 34,992 significant k-mers after controlling for the clone population structure (Fig. 3A), with 5,747 k-mers successfully mapped to 479 unique genes with known functions. Due to the considerable redundancy among the genetic elements in risk prediction, we used a simple model with only 5,747 k-mers, with the classification accuracy being 99.08% and the AUC value being 1.00. The top 100 host-associated k-mers were sorted according to the estimated importance (Fig. 3B), which were mainly associated with immune modulation (53%), effector delivery system (20%), antibiotic resistance (7%), and exoenzyme (4%). In addition, the GO annotations of the top 100 k-mers showed that the cellular component was significantly enriched in the plasma membrane and extracellular region, the biological process mainly enriched in response to arsenic-containing substance, and the molecular function mainly enriched in DNA binding (Fig. 3C).

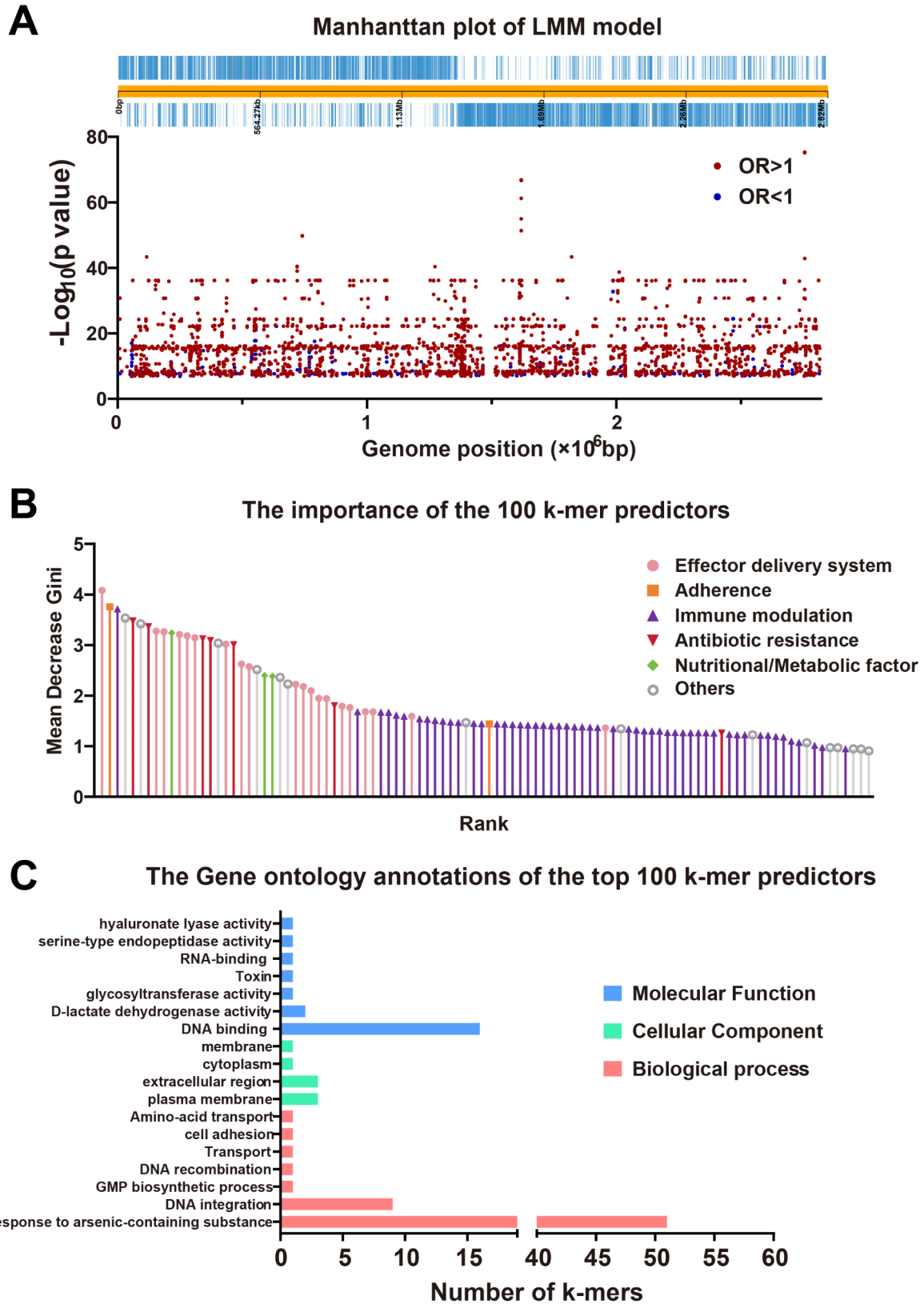

Fig. 3 Screening for host-associated k-mers by linear mixed model. **A** Manhattan plot showing the significant host-associated k-mers (mapping to a reference genome NCTC 8325); **B** Importance of the top 100 k-mer predictors; **C** Gene ontology annotations of the top 100 k-mer predictors

Confirmation of host-associated k-mers by three GWAS methods

In order to reduce the complexity of the initial model with 5,747 k-mers identified by LMM in the discovery stage, we used three GWAS methods (Scoary, LASSO, and XGBoost) to further identify consensus host-associated k-mers in the confirming stage. According to the Venn diagram (Fig. 4A), we observed 1,687 k-mers identified by Scoary, 40 k-mers identified by LASSO and 104 k-mers identified by XGBoost, with 20 consensus host-associated k-mers identified by all three methods. This model based on 20 k-mer predictors reached a classification accuracy of 98.78% (Table 2) and an AUC value of 0.99. As shown in Fig. 4B, the importance of the top 5 k-mers is significantly higher than that of other predictors. So we used these k-mers to obtain a simpler model, with a classification accuracy of 98.17% and an AUC value of 0.99 (Fig. 4C), indicating that a small set of 5 k-mer predictors is sufficient to distinguish different host species. Interestingly, the classification accuracy was 98.15% for the most important predictor (kmer_5162 in *pre*; Table 2) and 98.15% for the model with two important predictors (kmer_5162 and ST9), indicating that a single k-mer classifier is very powerful, irrespective of ST9. There were statistically significant differences in the proportion of the top 5 k-mer predictors between pig and human isolates (all $P < 0.05$; Fig. 4D), suggesting that certain k-mers were over-expressed in a specific host. Figure 4E shows changes in the estimated risk score for the top 5 k-mer predictors, in which a point above the diagonal line indicates that the risk score is increased when the k-mer is present. The presence of pig-specific k-mers mapped to antibiotic resistance (kmer_4853 in *ccrA*) and effector delivery system (kmer_5162 in *pre*; Table 3) were associated with increased risk of invading pigs.

External validation of host-associated k-mers by an independent dataset

To further validate the above results, an independent dataset (40 pig vs. 40 human isolates) was used to perform an additional external validation analysis using the RF classifier. The classification accuracy was 97.5% for the final model with the top 5 k-mers and 83.8% for a single highest-ranked k-mer predictor (kmer_5162 in *pre*), which is similar to that in the larger original dataset (98.17% and 98.15%, respectively).

Predicting the cross-species transmission risk of *S. aureus*

The RF classifier with 20 k-mer predictors was used to predict the host species of *S. aureus* of unknown origin (40 isolates from farm workers with occupational pig exposure), with 38 isolates (95.0%) predicted as pig origin, suggesting a high cross-species transmission risk of LA-SA among occupational livestock-exposed workers. Similar prediction results were observed for the RF classifier with the top 5 k-mer predictors (39 isolates predicted as pig origin; 97.5%), indicating that the prediction results are robust. As shown in Fig. 4F, the prevalence of the pig-specific k-mers (kmer_5162 and kmer_4853) was significantly higher in pig isolates than in human isolates, indicating that the enrichment of pig-specific elements may increase the risk of cross-species transmission.

Inferring the cross-species transmission direction of *S. aureus*

Most noteworthy, both livestock-associated ST9 and human-associated ST59 were observed in occupational livestock-exposed workers (Fig. 4F), but the potential evolutionary direction between humans and livestock is still unclear. To infer the evolutionary directions of ST9 and ST59 isolates, we used the Bayesian inference to estimate the evolutionary history. The time-measured phylogeny of ST9 isolates revealed that ST9 human isolates may originate from pig isolates (Fig. 5A), which occurred around 1991 (95% CI: 1977–1997); and the main host switching events from animals to humans possibly appeared in 2007 (95% CI: 2002–2009), 2008 (95% CI: 2003–2010), 2012 (95% CI: 2010–2013), 2013 (95% CI: 2011–2014), and 2014 (95% CI: 2013–2015). The time-measured phylogeny of ST59 isolates showed that the ST59 pig isolates may originate from human isolates (Fig. 5B), with the host switching occurred around 1987 (95% CI: 1981–1994). These findings suggest that ST9 isolates can cross-species spread from animals to humans but ST59 isolates can cross-species spread from humans to animals.

Discussion

Cross-species transmission of *S. aureus* is a multi-factorial process that involves various hosts and bacterial factors. To address the potential genetic backgrounds for cross-species transmission, it is necessary to consider different colonization models based on species-specific clones and genomics (Fig. 1). First, a host-associated clone model (Fig. 1A), in which only certain host-associated clones can colonize pigs or humans. Second, an opportunistic colonization model (Fig. 1B), in which all clones have equal capability of colonizing various hosts and host-associated determinants evenly enriched in the two group isolates. Third, a host-associated genome model (Fig. 1C), in which enrichment of host-associated elements may increase the risk of colonizing a specific host.

The prevalent genotypes of *S. aureus* vary among different geographical areas and hosts. Previous studies showed that the most prevalent clone of LA-SA in Asia was CC9 (ST9), while CC398 (ST398) predominated in Europe as well as the United States [31, 32], indicating

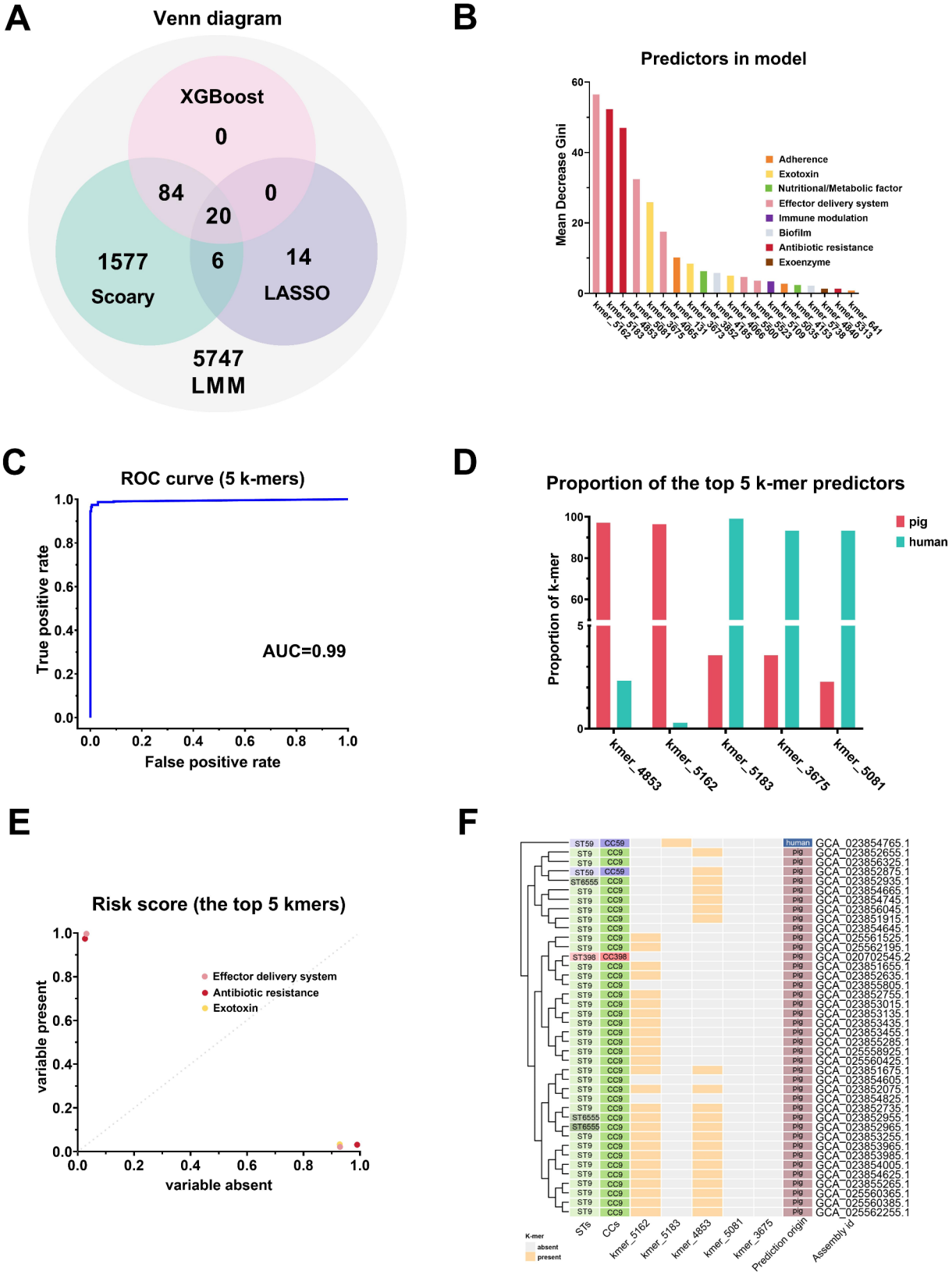

Fig. 4 Confirming host-associated k-mers by multiple GWAS analyses. **A** Venn diagram visualization of the k-mers identified by GWAS methods. **B** Importance of the 20 k-mers in the final model. **C** ROC curve of the final model based on the top 5 k-mers. **D** Proportion of the top 5 k-mer predictors between pig and human isolates. **E** Change in risk score for the top 5 predictors when the k-mer is present (y-axis) compared to absent (x-axis). **F** Cluster heatmap showing the host prediction of 40 *S. aureus* isolates of unknown origin and the presence/absence of host-associated k-mers

Table 2 Resubstitution and cross-validation results on Random Forest models

Evaluation indicators	20 predictors		5 predictors		Single predictor (kmer_5162)	
	Resubstitution estimate	Ten-fold cross-validation estimate	Resubstitution estimate	Ten-fold cross-validation estimate	Resubstitution estimate	Ten-fold cross-validation estimate
Accuracy	99.69	98.78	98.47	98.17	98.16	98.15
Sensitivity	99.68	99.69	99.34	99.36	99.67	99.68
Specificity	99.71	98.25	97.71	97.40	96.88	96.97
PPV	99.68	97.74	97.41	96.77	96.44	96.43
NPV	99.71	99.71	99.42	99.42	99.71	99.71
Kappa	0.99	0.98	0.97	0.96	0.96	0.96

Notes: Values are percentages except for kappa, which is reported as a value ranging from – 1 to 1

PPV, positive predictive value; NPV, negative predictive value

Table 3 Association analyses between the top 5 k-mers and host species

K-mers	Genes	Pig isolates, no.(%),*n* = 309	Human isolates no. (%), *n* = 343	*P* value	OR (95% CI)[a]
kmer_4853	*ccrA*	300(97.1)	8(2.3)	4.72×10^{-12}	1.34(1.23–1.45)
kmer_5162	*pre*	298(96.4)	1(0.3)	6.72×10^{-19}	2.54(2.08–3.10)
kmer_5183	*ccrB*	11(3.6)	340(99.1)	3.55×10^{-34}	0.63(0.59–0.68)
kmer_3675	*set3*	11(3.6)	320(93.3)	1.11×10^{-19}	0.76(0.72–0.80)
kmer_5081	*int*	7(2.3)	320(93.3)	1.21×10^{-17}	0.79(0.75–0.83)

[a]OR, odds ratio

that specific clones may be useful for identifying livestock association. Similarly in our study, livestock-associated CC9 (ST9) was the predominant genotype of pig isolates, while non-CC9 clones (such as CC59, CC398, and CC239) were prevalent in human isolates, giving evidence for identifying host-specific clones [32]. Consistent with previous studies [32, 33], we observed significant associations between *S. aureus* genotypes and host species, suggesting that specific clones are associated with increasing its ability to colonize specific hosts. In the simple clone model, all pig clones would appear as specific clades of genetically related isolates (Fig. 1A), which have been used for identifying disease-associated genotypes of *Streptococcus pneumoniae* and *Staphylococcus epidermidis* [34, 35]. However, this simple host-specific clone model is not suitable for all *S. aureus* isolates because many pig isolates clustered in the same clones of the phylogenetic tree with human isolates. One explanation of this genetic clustering is that some *S. aureus* clones can colonize multiple host species rather than a specific host, which is consistent with previous studies in Australia and Danish [36, 37]. For the opportunistic colonization model (Fig. 1B), all clones have equal capability of colonizing in both pigs and humans, with host-associated determinants evenly enriched in the two groups. According to the phylogenetic tree, we found that human isolates unevenly distributed and clustered in highly uniform clades with pig isolates, indicating that most clones have equal ability to cause colonization in both pigs and humans. If this opportunistic colonization model is suitable for *S. aureus*, host-associated genetic determinants are equally enriched in pig and human isolates (no significant difference). However, our LMM-based GWAS analysis identified numerous host-associated k-mers, suggesting that the enrichment of host-associated determinants can increase the risk of invading specific hosts (the host-associated genome model; Fig. 1C), which is similar to the divided genome model for the pathogenicity of *Streptococcus pneumoniae* and *Escherichia coli* [34, 38]. These findings indicate that the host-associated genome model (that is, the divided genome model) is suitable for *S. aureus.* In the divided genome model, horizontal gene transfer could spread genetic determinants in several bacteria [39–42], leading diverse clones to successfully colonize various hosts.

To obtain a simple and accurate host-associated genome model, a 3-stage GWAS analysis process (initial discovery, further confirming, and external validation) was performed to uncover 20 consensus host-associated k-mers. Our findings suggest that the enrichment of host-associated genetic elements may increase the risk of invading a specific host. The final model based on a k-mer profile of the top 5 k-mers achieved a high classification accuracy (98.17%), which is significantly higher than that of pathogenicity prediction for *Escherichia coli* (76.9%) and *Staphylococcus epidermidis* (79.8%) [35, 38], offering an accurate model for identifying livestock-associated isolates. Surprisingly, the best k-mer classifier

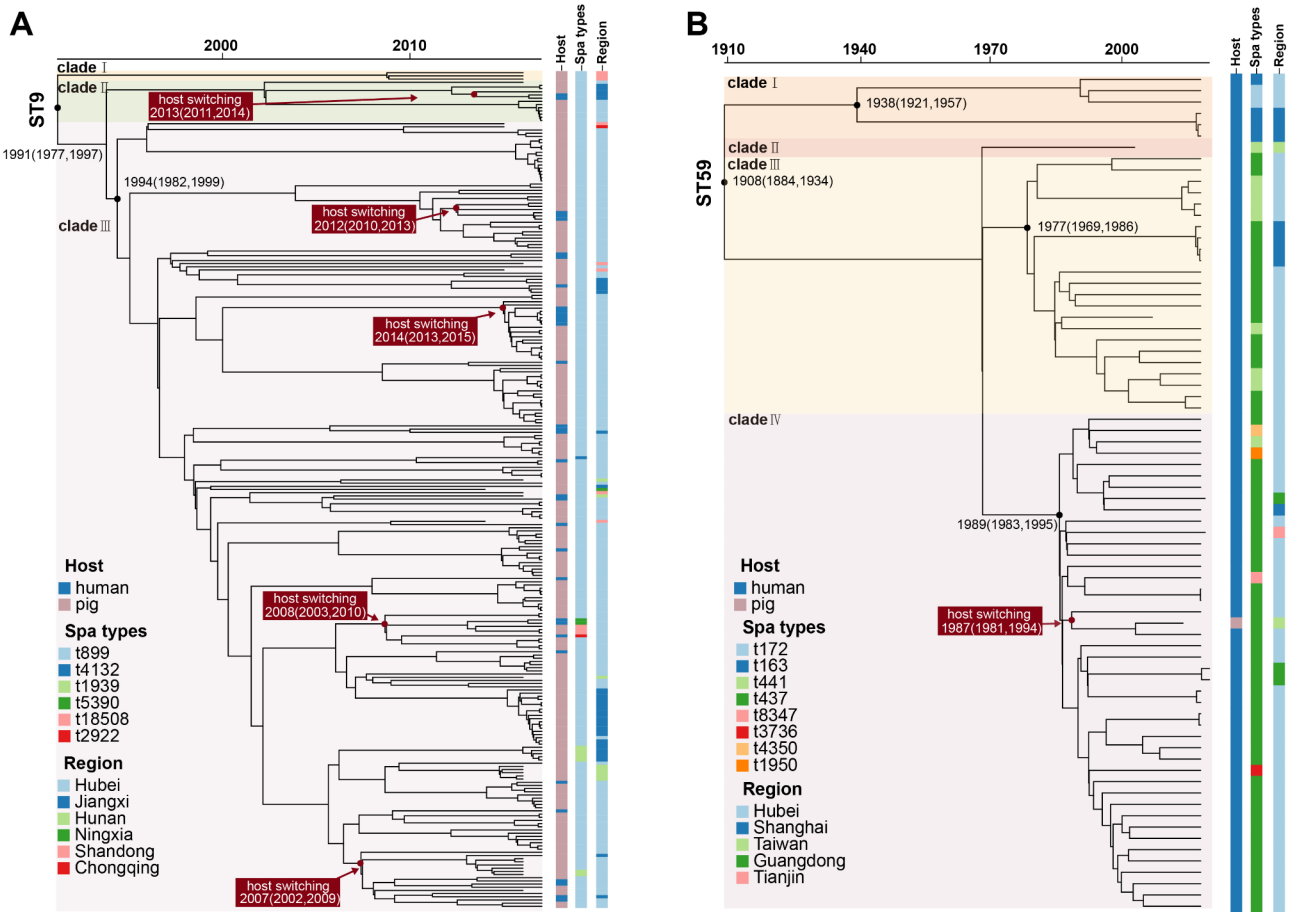

Fig. 5 Bayesian phylogenetic analyses of ST9 (**A**) and ST59 (**B**) isolates using BactDating. The characteristics of each strain are shown on the right (host, spa types, and region)

(kmer_5162 in *pre*) achieved a significantly high classification accuracy of 98.15% on its own and a similar accuracy was observed in the external validation analysis, which is significantly higher than that of the prediction model based on the best clone (87.73% for ST9) predicted by seven housekeeping genes in this study, giving a simple target for predicting host species.

Previous studies have reported that occupational livestock exposure is associated with the high carriage risk of *S. aureus* [6, 9, 43], but the risk of cross-species transmission is still unknown. In this study, we used an RF classifier to predict the host species of *S. aureus* of unknown origin (40 isolates from farm workers with occupational pig exposure). Similar prediction results were observed for the RF classifiers based on 20 k-mer predictors (95.0%) and 5 k-mer predictors (97.5%), which may provide more evidence for obtaining a robust risk prediction as well as elucidating the risk of cross-species transmission of LA-SA among occupational livestock-exposed workers. Importantly, ST9 and ST59 isolates were found in both humans and livestock, but their cross-species transmission directions are still unclear. Hence, we used the Bayesian inference to estimate the evolutionary history, suggesting that ST9 can cross-species spread from animals to humans but ST59 can cross-species spread from humans to animals. These findings give new ideas for identifying and predicting the cross-species transmission risk and direction of *S. aureus*.

S. aureus multi-host colonization is a complex multifactorial process. Previous studies suggest that multi-host colonization is associated with the mobile genetic elements, particularly for elements that encode virulence, resistance, and immune evasion pathways [44, 45]. For example, the human-specific host immune evasion proteins (Sak, Chp, and Scn) have been proven to be strongly associated with human isolates [46–48]. In our previous study [49], we also found that the animal-related isolates were significantly less likely to possess these proteins, which was consistent with the LMM-based results in this study. Additionally, the analysis of SNP-, gene-, and pathway-based approaches identified epithelial cell responses to mechanical and non-mechanical stress, indicating that certain genes overlap with pathways [50]. Considering that there are also some overlaps

between gene and k-mer analyses, we used 8 reference genomes to annotate identified k-mers, suggesting that these k-mers were associated with the effector delivery system, antibiotic resistance, exotoxin, adherence and so on. It is known that *S. aureus* cross-species transmission is associated with horizontal gene transfer, which provides the required traits for *S. aureus* colonization in specific hosts [2]. The plasmid recombination enzyme (*pre*) has been shown to mediate plasmid recombination and transfer [51, 52]. It is one of the elements of rolling circle plasmid that plays an important role in horizontal gene transfer [53]. The site-specific integrase (*int*) promotes site-specific integration and excision of genetic elements and genes into the human chromosome, which is necessary for involving λ integrase site-specific recombination pathway [54, 55]. These functions indicate potential roles for *pre* and *int* contribute to *S. aureus* adherence and colonization in special hosts. In terms of SCCmec-related genes, the *ccrA* and *ccrB* genes belong to cassette chromosome recombinases (Ccr), which comprise an unusual site-specific DNA recombination system [56]. Remarkably, the CcrA allows for the modular selection of new target sites in the bacterial chromosome and helps target recombination to the appropriate host site, which may play an important role in facilitating successful colonization in various hosts [57]. The cell wall associated fibronectin binding protein (*ebh*) belongs to the microbial surface component recognizing adhesive matrix molecules family, which is associated with adhesion, avoiding host defenses and contributing to the characteristic cell growth and envelope assembly pathways, thereby promoting *S. aureus* adaptation to different host species [58–61]. Exotoxin 3 (*set3*) as a Staphylococcal superantigen-like protein and a key regulator in the classical pathway of complement activation, has played an important role in inducing perturbation of the host immune system and facilitating *S. aureus* colonization [62, 63]. These findings may provide genetic evidence for understanding *S. aureus* cross-species transmission.

This is a new attempt to use a 3-stage k-mer-based GWAS analysis strategy to identify host-associated k-mers, so as to infer *S. aureus* host species and cross-species transmission. This study has potential limitations. First, the sample size of pig and human isolates in this study was limited by the current public genomes in the NCBI database. However, this is a novel, large-scale genomic analysis of *S. aureus* isolates to identify host-associated determinants. Second, strong population structure is the most important confounding factor in bacterial GWAS, leading to potential false associations. To minimize these false associations, we used a robust LMM to adjust clone population structure in the discovery stage and multiple GWAS methods to identify consensus host-associated variants in the confirming stage. Third, although it is a large-scale genomic GWAS on this topic in China, it only represents data from one country. Future multinational multi-host studies are required to confirm the results of this study and infer the host-associated determinants of other animals. Finally, national data from heterogeneous settings may lead to potential heterogeneity. So we adjusted for potential bias by adding study time as a fixed covariate and population structure as a random effect. Although we did not correct for geographical origin of isolates, the effect of geographical origin would be minimal as all individuals came from the same country.

Conclusions

In summary, our 3-stage GWAS analysis strategy (discovery, confirming, and external validation) identified a subset of consensus-significant host-associated k-mers, suggesting that the enrichment of genetic determinants may increase the risk of invading a specific host. Notably, both the final model with the top 5 k-mers and the simplest model with only the highest-ranked k-mer achieved a very high classification accuracy of 98%, giving a simple target for predicting the host origin of *S. aureus*. The time-measured phylogeny inferred the cross-species transmission directions, with animal-to-human transmission for ST9 and human-to-animal transmission for ST59. Our findings provide novel insights into host-associated genetic traits and give an accurate model for inferring host species and cross-species transmission of *S. aureus.*

Abbreviations

CC	Clonal complex
GO	Gene ontology
GWAS	Genome-wide association study
LA-SA	Livestock-associated *S. aureus*
LASSO	Least absolute shrinkage and selection operator regression
LMM	Linear mixed model
MDG	Mean Decrease Gini
RF	Random Forest
S. aureus	*Staphylococcus aureus*
SNPs	Single nucleotide polymorphisms
spa	Staphylococcal protein A
ST	Sequence type
XGBoost	Extreme gradient boosting

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11331-4.

Supplementary Material 1

Acknowledgements

Not applicable.

Author contributions

WYD and XHY designed the study and wrote the manuscript. WYD, STC, and RJ performed all bioinformatic analyses. HLZ, DJO, and YHL took charge of

data curation. XHY, ZJY and YJG took charge of supervision and reviewed the data. All authors have read and approved the final manuscript.

Funding
This work was supported by the Key Scientific Research Foundation of Guangdong Educational Committee (No. 2022ZDZX2033), the Guangdong Basic and Applied Basic Research Foundation (No. 2023A1515011583), and the National Natural Science Foundation of China (Nos. 81973069 and 81602901). The funders had no role in the study design, data collection and analysis, and interpretation of the data.

Data availability
All data generated or analyzed in this study are available as Supplementary Information (Supplementary Tables S1–S6).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 5 October 2024 / Accepted: 5 February 2025
Published online: 17 February 2025

References

1. Cheung GYC, Bae JS, Otto M. Pathogenicity and virulence of Staphylococcus aureus. Virulence. 2021;12:547–69.
2. Richardson EJ, Bacigalupe R, Harrison EM, Weinert LA, Lycett S, Vrieling M, et al. Gene exchange drives the ecological success of a multi-host bacterial pathogen. Nat Ecol Evol. 2018;2:1468–78.
3. Wang Y, Zhang P, Wu J, Chen S, Jin Y, Long J, et al. Transmission of livestock-associated methicillin-resistant Staphylococcus aureus between animals, environment, and humans in the farm. Environ Sci Pollut Res. 2023;30:86521–39.
4. Ren Q, Liao G, Wu Z, Lv J, Chen W. Prevalence and characterization of Staphylococcus aureus isolates from subclinical bovine mastitis in southern Xinjiang, China. J Dairy Sci. 2020;103:3368–80.
5. Kobusch I, Schröter I, Linnemann S, Schollenbruch H, Hofmann F, Boelhauve M. Prevalence of LA-MRSA in pigsties: analysis of factors influencing the (De) colonization process. Sci Rep. 2022;12:18000.
6. Dong Q, Liu Y, Li W, Liu Y, Ye X. Cross-species transmission risk of livestock-associated MRSA: a systematic review and Bayesian meta-analysis of global data. Prev Vet Med. 2021;194:105429.
7. Liu Y, Han C, Chen Z, Guo D, Ye X. Relationship between livestock exposure and methicillin-resistant Staphylococcus aureus carriage in humans: a systematic review and dose–response meta-analysis. Int J Antimicrob Agents. 2020;55:105810.
8. Howden BP, Giulieri SG, Wong Fok Lung T, Baines SL, Sharkey LK, Lee JYH, et al. Staphylococcus aureus host interactions and adaptation. Nat Rev Microbiol. 2023;21:380–95.
9. Liu Y, Li W, Dong Q, Liu Y, Ye X. Livestock-associated and non-livestock-associated Staphylococcus aureus carriage in humans is associated with pig exposure in a dose-response manner. Infect Drug Resist. 2021;14:173–84.
10. García-Álvarez L, Holden MTG, Lindsay H, Webb CR, Brown DFJ, Curran MD, et al. Meticillin-resistant Staphylococcus aureus with a novel mecA homologue in human and bovine populations in the UK and Denmark: a descriptive study. Lancet Infect Dis. 2011;11:595–603.
11. Guo Q, Yang J, Forsythe SJ, Jiang Y, Han W, He Y, et al. DNA sequence-based re-assessment of archived Cronobacter sakazakii strains isolated from dairy products imported into China between 2005 and 2006. BMC Genomics. 2018;19:506.
12. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet. 2019;20:467–84.
13. Burgaya J, Marin J, Royer G, Condamine B, Gachet B, Clermont O, et al. The bacterial genetic determinants of Escherichia coli capacity to cause bloodstream infections in humans. PLoS Genet. 2023;19:e1010842.
14. Dyzenhaus S, Sullivan MJ, Alburquerque B, Boff D, van de Guchte A, Chung M, et al. MRSA lineage USA300 isolated from bloodstream infections exhibit altered virulence regulation. Cell Host Microbe. 2023;31:228–e2428.
15. Månsson E, Bech Johannesen T, Nilsdotter-Augustinsson Å, Söderquist B, Stegger M. Comparative genomics of Staphylococcus epidermidis from prosthetic-joint infections and nares highlights genetic traits associated with antimicrobial resistance, not virulence. Microb Genom. 2021;7:000504.
16. Gupta PK, Kulwal PL, Jaiswal V. Association mapping in crop plants: opportunities and challenges. Adv Genet. 2014;85:109–47.
17. Karikari B, Lemay M-A, Belzile F. k-mer-based genome-wide association studies in plants: advances, challenges, and perspectives. Genes. 2023;14:1439.
18. Lemay M-A, de Ronne M, Bélanger R, Belzile F. k-mer-based GWAS enhances the discovery of causal variants and candidate genes in soybean. Plant Genome. 2023;16:e20374.
19. Gupta PK. GWAS for genetics of complex quantitative traits: genome to pangenome and SNPs to SVs and k-mers. BioEssays. 2021;43:e2100109.
20. Wang Y, Chen Q, Deng C, Zheng Y, Sun F. KmerGO: a tool to identify group-specific sequences with k-mers. Front Microbiol. 2020;11:2067.
21. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, et al. De novo assembly of haplotype-resolved genomes with trio binning. Nat Biotechnol. 2018;22:101038.
22. Gupta PK. Quantitative genetics: pan-genomes, SVs, and k-mers for GWAS. Trends Genet. 2021;37:868–71.
23. Nascimento M, Sousa A, Ramirez M, Francisco AP, Carriço JA, Vaz C. PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. Bioinformatics. 2017;33:128–9.
24. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.
25. Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J. Pyseer: a comprehensive tool for microbial pangenome-wide association studies. Bioinformatics. 2018;34:4310–2.
26. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol. 2016;17:238.
27. Gui J, Li H. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics. 2005;21:3001–8.
28. Chen T, Guestrin C, XGBoost:. A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA: ACM; 2016. pp. 785–94.
29. Hu J-Y, Wang Y, Tong X-M, Yang T. When to consider logistic LASSO regression in multivariate analysis? Eur J Surg Oncol. 2021;47:2206.
30. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
31. Yu F, Cienfuegos-Gallet AV, Cunningham MH, Jin Y, Wang B, Kreiswirth BN, et al. Molecular evolution and adaptation of livestock-associated methicillin-resistant Staphylococcus aureus (LA-MRSA) sequence type 9. mSystems. 2021;6:e0049221.
32. Aires-de-Sousa M. Methicillin-resistant Staphylococcus aureus among animals: current overview. Clin Microbiol Infect. 2017;23:373–80.
33. Peton V, Le Loir Y. Staphylococcus aureus in veterinary medicine. Infect Genet Evol. 2014;21:602–15.
34. Jonczyk MS, Simon M, Kumar S, Fernandes VE, Sylvius N, Mallon A-M, et al. Genetic factors regulating lung vasculature and immune cell functions associate with resistance to pneumococcal infection. PLoS ONE. 2014;9:e89831.
35. Méric G, Mageiros L, Pensar J, Laabei M, Yahara K, Pascoe B, et al. Disease-associated genotypes of the commensal skin bacterium Staphylococcus epidermidis. Nat Commun. 2018;9:5034.
36. Sahibzada S, Abraham S, Coombs GW, Pang S, Hernández-Jover M, Jordan D, et al. Transmission of highly virulent community-associated MRSA ST93 and livestock-associated MRSA ST398 between humans and pigs in Australia. Sci Rep. 2017;7:5273.
37. Larsen J, Stegger M, Andersen PS, Petersen A, Larsen AR, Westh H, et al. Evidence for human adaptation and foodborne transmission of livestock-associated methicillin-resistant *Staphylococcus aureus*: Table 1. Clin Infect Dis. 2016;63:1349–52.
38. Mageiros L, Méric G, Bayliss SC, Pensar J, Pascoe B, Mourkas E, et al. Genome evolution and the emergence of pathogenicity in avian Escherichia coli. Nat Commun. 2021;12:765.

39. Arnold BJ, Huang I-T, Hanage WP. Horizontal gene transfer and adaptive evolution in bacteria. Nat Rev Microbiol. 2022;20:206–18.
40. Hiller NL, Ahmed A, Powell E, Martin DP, Eutsey R, Earl J, et al. Generation of genic diversity among Streptococcus pneumoniae strains via horizontal gene transfer during a chronic polyclonal pediatric infection. PLoS Pathog. 2010;6:e1001108.
41. Obolski U, Swarthout TD, Kalizang'oma A, Mwalukomo TS, Chan JM, Weight CM, et al. The metabolic, virulence and antimicrobial resistance profiles of colonising Streptococcus pneumoniae shift after PCV13 introduction in urban Malawi. Nat Commun. 2023;14:7477.
42. Yang S, Chen J, Fu J, Huang J, Li T, Yao Z, et al. Disease-associated Streptococcus pneumoniae genetic variation. Emerg Infect Dis. 2024;30:39–49.
43. Cortimiglia C, Luini M, Bianchini V, Marzagalli L, Vezzoli F, Avisani D, et al. Prevalence of Staphylococcus aureus and of methicillin-resistant S. aureus clonal complexes in bulk tank milk from dairy cattle herds in Lombardy Region (Northern Italy). Epidemiol Infect. 2016;144:3046–51.
44. McCarthy AJ, Loeffler A, Witney AA, Gould KA, Lloyd DH, Lindsay JA. Extensive horizontal gene transfer during Staphylococcus aureus co-colonization in vivo. Genome Biol Evol. 2014;6:2697–708.
45. Lindsay JA. Genomic variation and evolution of Staphylococcus aureus. Int J Med Microbiol. 2010;300:98–103.
46. van Wamel WJB, Rooijakkers SHM, Ruyken M, van Kessel KPM, van Strijp JAG. The innate immune modulators staphylococcal complement inhibitor and chemotaxis inhibitory protein of Staphylococcus aureus are located on beta-hemolysin-converting bacteriophages. J Bacteriol. 2006;188:1310–5.
47. Sung JM-L, Lloyd DH, Lindsay JA. Staphylococcus aureus host specificity: comparative genomics of human versus animal isolates by multi-strain microarray. Microbiol (Reading). 2008;154:1949–59.
48. McCarthy AJ, Witney AA, Lindsay JA. Staphylococcus aureus temperate bacteriophage: carriage and horizontal gene transfer is lineage associated. Front Cell Infect Microbiol. 2012;2:6.
49. Ye X, Wang X, Fan Y, Peng Y, Li L, Li S, et al. Genotypic and phenotypic markers of livestock-associated methicillin-resistant Staphylococcus aureus CC9 in humans. Appl Environ Microbiol. 2016;82:3892–9.
50. Ye Z, Vasco DA, Carter TC, Brilliant MH, Schrodi SJ, Shukla SK. Genome wide association study of SNP-, gene-, and pathway-based approaches to identify genes influencing susceptibility to Staphylococcus aureus infections. Front Genet. 2014;5:125.
51. Grohmann E, Muth G, Espinosa M. Conjugative plasmid transfer in Gram-positive bacteria. Microbiol Mol Biol Rev. 2003;67:277–301.
52. Gennaro ML, Kornblum J, Novick RP. A site-specific recombination function in Staphylococcus aureus plasmids. J Bacteriol. 1987;169:2601–10.
53. Wawrzyniak P, Płucienniczak G, Bartosik D. The different faces of rolling-circle replication and its multifunctional initiator proteins. Front Microbiol. 2017;8:2353.
54. LANDY A. The λ integrase site-specific recombination pathway. Microbiol Spectr. 2015;3:MDNA3–0051.
55. Frumerie C, Sylwan L, Helleday T, Yu A, Haggård-Ljungquist E. Bacteriophage P2 integrase: another possible tool for site-specific recombination in eukaryotic cells. J Appl Microbiol. 2008;105:290–9.
56. Wu Z, Li F, Liu D, Xue H, Zhao X. Novel type XII staphylococcal cassette chromosome *mec* harboring a new cassette chromosome recombinase, CcrC2. Antimicrob Agents Chemother. 2015;59:7597–601.
57. Wang L, Archer GL. Roles of CcrA and CcrB in excision and integration of Staphylococcal cassette chromosome mec, a Staphylococcus aureus genomic island. J Bacteriol. 2010;192:3204–12.
58. McCarthy AJ, Lindsay JA. Genetic variation in Staphylococcus aureus surface and immune evasion genes is lineage associated: implications for vaccine design and host-pathogen interactions. BMC Microbiol. 2010;10:173.
59. Asante J, Abia ALK, Anokwah D, Hetsa BA, Fatoba DO, Bester LA, et al. Phenotypic and genomic insights into Biofilm formation in antibiotic-resistant clinical coagulase-negative Staphylococcus species from South Africa. Genes. 2023;14:104.
60. Cheng AG, Missiakas D, Schneewind O. The giant protein ebh is a determinant of Staphylococcus aureus cell size and complement resistance. J Bacteriol. 2014;196:971–81.
61. Foster TJ, Geoghegan JA, Ganesh VK, Höök M. Adhesion, invasion and evasion: the many functions of the surface proteins of Staphylococcus aureus. Nat Rev Microbiol. 2014;12:49–62.
62. Oku T, Kurisaka C, Ando Y, Tsuji T. Identification of human plasma C1 inhibitor as a target protein for staphylococcal superantigen-like protein 5 (SSL5). Biochem Biophys Res Commun. 2019;508:1162–7.
63. Davis AE, Mejia P, Lu F. Biological activities of C1 inhibitor. Mol Immunol. 2008;45:4057–63.

Publisher's note