Prophage-encoded immune evasion factors are critical for *Staphylococcus aureus* host infection, switching, and adaptation

**Graphical abstract**

**Highlights**

- *S. aureus* φSa3 prophage immune evasion genes are associated with human hosts

- GWAS of paired isolates identified more variants associated with human hosts

- Overall, *S. aureus* genetics show ~88% heritability for human host association

- φSa3 genes explain ~99.9% heritability for *S. aureus* human host association

**Authors**

Chrispin Chaguza, Joshua T. Smith, Spencer A. Bruce, Robert Gibson, Isabella W. Martin, Cheryl P. Andam

**Correspondence**

chrispin.chaguza@yale.edu (C.C.), candam@albany.edu (C.P.A.)

**In brief**

Chaguza et al. present a population genomic study of *Staphylococcus aureus* from animals and humans to identify genetic signatures for host-switching, transmission, and adaptation. Using multiple genome-wide association study (GWAS) approaches adjusting for the population structure, they found a strong genetic basis in the immune evasion genes carried on a single prophage element. The attribution of the heritability to these prophage-encoded genes suggests that these loci are critical determinants for *S. aureus* host-switching, transmissibility, infection, and adaptation.
Prophage-encoded immune evasion factors are critical for Staphylococcus aureus host infection, switching, and adaptation

Chrispin Chaguza,1,* Joshua T. Smith,2 Spencer A. Bruce,3 Robert Gibson,4 Isabella W. Martin,5 and Cheryl P. Andam3,6,7,*

1Department of Epidemiology of Microbial Diseases, Yale School of Public Health, Yale University, New Haven, CT, USA
2Broad Institute of MIT and Harvard, Cambridge, MA, USA
3Department of Biological Sciences, University at Albany, State University of New York, New York, USA
4New Hampshire Veterinary Diagnostic Laboratory, Durham, NH, USA
5Dartmouth-Hitchcock Medical Center and Dartmouth College Geisel School of Medicine, Lebanon, NH, USA
6Senior author
7Lead contact
*Correspondence: chrispin.chaguza@yale.edu (C.C.), candam@albany.edu (C.P.A.)
https://doi.org/10.1016/j.xgen.2022.100194

SUMMARY

Staphylococcus aureus is a multi-host pathogen that causes infections in animals and humans globally. The specific genetic loci—and the extent to which they drive cross-species switching, transmissibility, and adaptation—are not well understood. Here, we conducted a population genomic study of 437 S. aureus isolates to identify bacterial genetic variation that determines infection of human and animal hosts through a genome-wide association study (GWAS) using linear mixed models. We found genetic variants tagging φSa3 prophage-encoded immune evasion genes associated with human hosts, which contributed ~99.9% of the overall heritability (~88%), highlighting their key role in S. aureus human infection. Furthermore, GWAS of pairs of phylogenetically matched human and animal isolates confirmed and uncovered additional loci not implicated in GWAS of unmatched isolates. Our findings reveal the loci that are critical for S. aureus host transmissibility, infection, switching, and adaptation and how their spread alters the specificity of host-adapted clones.

INTRODUCTION

Staphylococcus aureus is a multi-host pathogen and commonly causes infections in animals1 and community and nosocomial infections in humans.2 Although S. aureus is a generalist species, molecular studies have identified specialist clones predominantly abundant in either human or animal hosts.3 An example of a broadly animal-adapted clone is ST398 (CC398), while ST8 (CC8), ST22 (CC22), and ST36 (CC30) are human adapted.4–8 Previous studies have demonstrated frequent host-switching over different timescales.3,9–12 However, a key question not yet fully addressed is whether there are any additional bacterial-specific factors that influence the transmissibility and pathogenicity of S. aureus strains to different human and animal hosts.4–8 Single-nucleotide mutations in bacteria have been associated with significant phenotypic changes, including host tropism in S. aureus,11 serum resistance and virulence in Salmonella enterica,12 host specificity in Listeria monocytogenes,15 and tissue tropism in Streptococcus pneumoniae.16 Additionally, the acquisition and loss of mobile genetic elements (MGEs), including staphylococcal prophages and pathogenicity genomic islands, have been linked with host adaptation in humans and animals.3,4,9,12,17–20 For example, the emergence of the livestock-associated S. aureus clones ST398 and ST9 from humans and subsequent adaptation in livestock was linked with the loss of phage-associated virulence genes.4,21,22 Similarly, the acquisition of prophages by livestock-associated clones has been linked with increased transmission and adaptation to humans.23 However, it remains unknown whether such differential abundance of these genes is critical for host transmissibility and infection and truly reflects adaptation to different hosts or merely potential interspecies barriers to gene flow between different hosts, such as through restriction-modification systems,24,25 which may result in the unique distribution of genes between hosts. Furthermore, the degree to which they contribute to the phenotypic variability remains unknown. Therefore, uncovering genetic variation critical for host
transmissibility, infection, switching, and adaptation of *S. aureus* and other multi-species pathogens is critical. Such investigations can unravel novel pathogenicity loci for targeting effective prophylactic and therapeutic measures to prevent and control the emergence of virulent strains of serious threat to human and livestock health.

The application of genome-wide association studies (GWASs) has revealed insights regarding the genetic basis of virulence, healthcare adaptation, immune evasion, colonization duration, pathogenicity, non-communicable disease risk, antimicrobial resistance, host adaptation, and transmission of *S. aureus* and related species. Here, we conducted a large-scale GWAS of a genetically diverse collection of 437 *S. aureus* isolates sampled from animals and humans in New England, United States, to explore the genetic basis for transmissibility and infection of animals and human hosts. We applied GWAS based on linear mixed models and a phylogeny-sampled matching of phenotypically distinct isolates to robustly control confounding effects due to the clonal bacterial population structure. This method allowed us to precisely identify and quantify the overall impact of bacterial genetics on transmissibility, host infection, switching, and adaptation of *S. aureus*. Our findings highlight the critical role of horizontal gene transfer in disseminating the prophage-encoded immune evasion factors, which modulate staphylococcal host transmissibility, infection, switching, and adaptation.

**RESULTS**

**Co-circulation of human- and animal-associated *S. aureus* clones**

We constructed a whole-genome phylogeny of 437 *S. aureus* isolates from pure cultures of single colonies, each representing the genetic content of a single cell, sampled from infected humans and animals in New England, United States, from 2010 to 2020 to understand the *S. aureus* population structure (Figures 1A–1C and S1; Data S1). The human isolates, 323 in total, were sampled from the blood of unique pediatric and adult patients with bacteremia at Dartmouth-Hitchcock Medical Center, United States, from 2010 to 2018.37 The dominant clones of the complete dataset were clonal complex (CC) 5 (26.09%; n = 108), CC8 (24.49%; n = 103), CC30 (11.67%; n = 51), CC97 (7.55%; n = 13), CC45 (6.18%; n = 23), and CC1 (5.72%; n = 21), which constituted 81.69% of the *S. aureus* isolates in this study (Figure 1E). Since the distribution of *S. aureus* clones varies by host type, we compared the prevalence of these clones among the animal and human isolates. We found two clones more commonly present in humans than in animals, CC8 (29.41% versus 10.53%, p = 9.427 × 10^{-7}) and CC45 (7.74% versus 1.75%, p = 0.04), consistent with well-established evidence that they are predominantly human-adapted clones associated with the community- and hospital-acquired infections (Figure 1F). In contrast, the only typically animal-associated clone to show a statistically significant higher prevalence in animals was CC97 (19.30% versus 3.41%, p = 1.066 × 10^{-3}). However, CC1, a known livestock-associated clone, also appeared more common in animals than in humans, although the difference was not statistically significant (9.65% versus 4.33%, p = 0.062). These findings suggest potential clonal or lineage effects on host-switching and adaptation of different *S. aureus* strains.

**Transmission of *S. aureus* frequently occurs between humans and animals**

We next compared the pairwise genetic distances based on the single-nucleotide polymorphisms (SNP) to identify potential zoonotic and reverse zoonotic transmission events of *S. aureus* between human and animal hosts. The pairwise SNP distances distinguishing pairs of *S. aureus* isolates showed a multi-modal distribution for the isolates sampled from the same host and different host types (Figures 2A–2D). Previous studies of bacterial transmission have primarily used a lower SNP threshold, typically ~50 SNPs, which efficiently detects recent transmission events, especially within households and healthcare settings.14,44 However, since we were interested in capturing both recent and non-recent direct or indirect transmissions, we used a threshold of 150 SNPs, which would allow us to capture potential transmissions within ~13 years, assuming a cutoff of ~24 core genome SNPs to resolve transmissions occurring within one year.45 We identified 19 potential transmission clusters when we applied this SNP threshold (Figure 2E). These transmission clusters were associated with several major clonal complexes, including CC1 (one cluster), CC5 (10), CC8 (eight), CC30 (six), CC45 (three), and CC97 (one), and other undefined clonal complexes (10). The mean number of SNPs between human and animal *S. aureus* isolates associated with each transmission event was 107.04 (range, 35–149), which suggested that the transmissions were predominantly indirect and not recent occurrences. The sharing of *S. aureus* clones between the human and animal hosts in New England, United States, suggested the occurrence of potential zoonotic and reverse zoonotic transmission events (Figure 1E). Although the *S. aureus* isolates were not collected simultaneously and had no available epidemiological information linking them, the isolates were sampled from the same region. Therefore, the high genetic similarity between the human and animal strains likely represented direct or indirect transmission events. Together, these findings showed a history of non-recent transmissions of *S. aureus* clones between human and animal hosts.

**GWAS reveals the critical role of prophage-encoded immune evasion and novel genes in *S. aureus* host transmissibility, infection, switching, and adaptation**

We next investigated whether the genetics of *S. aureus* influenced the transmission of the strains between human and animal hosts. We also measured the association between the host type phenotype and phylogeny. We estimated the phylogenetic signal by mapping the phenotype onto the phylogeny to estimate Pagel’s λ statistic.46 To minimize bias due to the unequal number of human and animal isolates in the phylogenetic tree, we randomly subsampled the phylogeny of 437 isolates to select an equal
Figure 1. *S. aureus* isolates sampled from human and animal hosts are genetically diverse, intermixed in the phylogeny, and reveal host-adapted clones

(A) The geographical location of the human and animal *S. aureus* isolates in New England, United States. The animal isolates were collected from 2017 to 2020, while the human isolates were collected between 2010 and 2018.

(B) Pie chart showing the number of isolates from humans and animals.

(C) Pie chart showing the proportion of human and animal *S. aureus* isolates.

(D) Bar plot showing the number of *S. aureus* isolates by year of isolation and host species.

(E) Maximum-likelihood phylogeny generated using 141,232 SNPs showing genetic similarity of *S. aureus* isolates sampled from humans and animals in New England, United States. The phylogeny is annotated by sampled host type, host, and sequence type based on multilocus sequence typing (MLST). For visual clarity, we performed a square root transformation of the phylogenetic branches as the terminal taxon tips were obscured by the long deep branches. Phylogenetic branches with bootstrap values equal to 100% are marked with an asterisk.

(F) Bar plot showing the relative frequency of *S. aureus* clonal complexes among human and animal hosts. *p < 0.05, ***p < 0.001 (testing for equality of proportions). The error bars in the graph represent 95% confidence intervals (CIs). Additional information for the isolates is provided in Figure S1.
number of 80 isolates from humans and animal hosts 50 times and inferred Pagel’s λ values based on this dataset using the all-rates-different (ARD) discrete character models, assuming unequal transition rates between states. We estimated Pagel’s λ of 0.87 (95% confidence interval [CI], 0.46–1.00), which indicated a strong correlation between the phylogeny and the host type (i.e., a strong phylogenetic signal). This strong phylogenetic signal implied potential strain or lineage effects whereby certain clusters of S. aureus strains were associated with the same host type.

These observations were consistent with the findings in Figure 1E, which showed that some clonal complexes were more commonly found in humans (for example, CC97). In contrast, other clones, including CC8 and CC45, were more commonly associated with animals. Overall, the transition rates of S. aureus from animal to human hosts and vice versa were approximately 32.56 (95% CI, 24.00–44.55) and 21.2 (95% CI, 14.00–29.78) (p = 5.435 × 10⁻⁵), respectively. This suggested that S. aureus host-switching occurred more frequently from animal to humans than from...
humans to animals. These findings provided further evidence for the effect of strains or lineages on the infection of different host types with *S. aureus*.

Frequent genetic exchange of antimicrobial resistance (AMR) and virulence-associated determinants promote the ecological adaptation of *S. aureus* adaptation. We next investigated whether certain genetic variations influenced infection, switching, and adaptation of the human and animal hosts. We performed a GWAS of the 323 human and 114 animal *S. aureus* isolates to identify genetic signals for infection of human and animal hosts independent of the genetic background using linear mixed models implemented in Factored Spectrally Transformed Linear Mixed Models (FaST-LMM) (Figures 3A and S2). Such genetic variation is likely homoplastic due to independent and...
convergent evolution. Since *S. aureus* evolves primarily through mutation and MGEs, we first investigated the association of SNPs and host type. We found 129 SNPs out of the 113,454 SNPs present in 5%–95% of the isolates that were statistically associated with host type after correcting for multiple testing (Figure 3B). The Q-Q plots for SNP-based GWAS showed no issues due to the population structure (Figure S3). We found SNPs with the lowest statistical significance in the genomic region containing the MGE φSa3 prophage, a ~43-kb prophage harboring the immune evasion cluster genes, which inserts into the β-hemolysin (hbl) gene (Table S1; Figure S4; Data S2). These genes included *scn, chp, sak*, and *sea*, which encode the innate immune modulators, including staphylococcal complement inhibitor (SCIN), chemotaxis inhibitory protein of *Staphylococcus* (CHIPS), staphylokinase (SAK), and enterotoxin A (SEA) proteins, respectively.\(^{48,50,51}\) The effect sizes associated with these genes ranged from 0.707 to 0.729 (adjusted *p* value, 3.59 × 10\(^{-39}\) to 2.05 × 10\(^{-29}\)) for the SAK gene (*scn*), 0.739 to 0.753 (adjusted *p* value, 4.29 × 10\(^{-12}\) to 3.16 × 10\(^{-10}\)) for the SAK gene (*sak*), and 1.63 × 10\(^{-9}\) to 1.32 × 10\(^{-8}\) (*p* value, 0.0044–0.0360) for the transposase gene (*tnp*) (Table S1). Additional genes containing SNPs associated with host type were hypothetical and included those encoding a lytic enzyme, amide, putative phage protein, and transposase (Data S2). We also identified statistically significant SNPs in the intergenic regions within the prophage region, which were associated with the infection of human and animal host types. These prophage-associated SNPs possibly reflected strong linkage disequilibrium due to the translocation of the bacteriophage from strain to strain as a single intact unit. We also found six SNPs outside the φSa3 prophage region that were associated with host type. Two SNPs were associated with a transposase sequence upstream of the integration site of the φSa3 prophage. In comparison, four SNPs associated with the IS1272 transposase downstream of the prophage were likely acquired from other staphylococcal species.\(^{52}\) Notably, the number and statistical significance of the SNPs identified outside the prophage sequence were lower than those located within the prophage. Therefore, these findings indicate that genetic variation in *S. aureus* has a major effect on infection of different host types. We propose that the host transmissibility and infection are primarily driven by genetic variation within the φSa3 prophage.

We then conducted a complementary GWAS using unitigs, defined as variable-length contiguous k-mer sequences, as markers of *S. aureus* genetic variation (Figure 3A and Table 1, and Figure S5). Unlike SNPs, which are determined based on a reference genome, unitigs capture additional genetic and intergenic variation not available in the reference genome, including accessory genes, insertion and deletions of any size, and genomic rearrangements.\(^{53}\) Crucially, the unitigs capture genomic variation arising through horizontal gene transfer (HGT), such as prophages and other MGEs, which play a critical role in bacterial evolution.\(^{54}\) Based on the GWAS of 418,204 unitigs detected in 5%–95% of the isolates, we found 451 unitigs statistically associated with *S. aureus* host type (Figure 3C and Data S2). The number of statistically significant unitigs was about three times higher than the associations detected by the SNPs, which demonstrated an increased ability to detect the host-associated genetic variation. Interestingly, mapping the unitig sequences to a reference genome to annotate them showed that they tagged all the genes containing the variants identified in the SNP-based GWAS, particularly the immune evasion genes in the φSa3 prophage. However, the identified unitigs also mapped to additional genes not found in the SNP-based GWAS, demonstrating the increased power and resolution of the unitig-based GWAS. Similarly, the Q-Q plots for unitig-based GWAS revealed no issues due to the population structure (Figure S3). Altogether, these findings provided further evidence that *S. aureus* genetic variation contributes to the differential infection of human and animal hosts.

Having uncovered the SNPs and unitigs associated with the infection of different host types, we undertook further analyses to investigate whether it was the presence and absence of the entire genes tagged by the SNPs and unitigs associated with the host type rather than allelic variation at nucleotide substitution level (Figures 3A, S3, and S6; Table S2). We achieved this through a pan-genome analysis of the *S. aureus* isolates to determine the presence and absence patterns of genes. A GWAS based on the presence and absence patterns of 1,503 intermediate frequency genes found in 5%–95% of the isolates revealed four genes, namely, *scn* (odds ratio, 0.707; *p* = 2.25 × 10\(^{-39}\)), *sak* (odds ratio, 0.750; *p* = 1.19 × 10\(^{-15}\)), and two hypothetical genes: an amida (odds ratio, 0.814; *p* = 0.0308) and ATPases associated with diverse cellular activities (AAA) family protein (odds ratio, 0.835; *p* = 0.0025) (Figure 3D; Table S2; Data S2). The Q-Q plots for gene-based GWAS showed robust control of the population structure (Figure S3).

We also repeated the GWAS analysis using a different tool, genome-wide efficient mixed-model analysis (GEMMA),\(^{55}\) to check for consistency in the identified genetic variation and confirm that our approach worked correctly (Figures 3E–3G). Similarly, the population structure was efficiently controlled (Figure S3). We found similar results between different approaches. All the SNPs, unitigs, and genes statistically associated with *S. aureus* infection of different host types identified by GEMMA were also identified by FaST-LMM. However, FaST-LMM had greater power, as shown by the discovery of additional variants associated with host type (Figure 3H; Data S3). Functional analysis of the combined genetic variants identified by both genes revealed that most genes were associated with MGEs, particularly prophages (Figure 3I). These demonstrate that mobilization of the φSa3 prophage-associated genes through HGT, particularly the human innate immune evasion genes, rather than allelic variation is the primary determinant of infection, switching, and adaptation of *S. aureus* to human and animal hosts.

**A phylogeny-based sampling of matched pairs of human and animal isolates confirms the association of GWAS loci with host transmissibility and infection**

To confirm whether the genetic loci implicated are associated with *S. aureus* infection of different host types, we next undertook a GWAS of phylogenetically matched pairs of human and animal isolates (Figures 4A and 4B). We based our phylogeny-based matched sampling on a similar approach used to identify loci for drug resistance and host adaptation in *Mycobacterium tuberculosis* and *Campylobacter* species, respectively.\(^{56}\)
pathogenicity in *Staphylococcus epidermidis.* This approach improves statistical power to uncover hidden genotype-phenotype associations even for small datasets due to efficient controlling of the residual confounding effect caused by cryptic bacterial population structure. We hypothesized that phylogenetic matching of the human and animal isolates would result in a distribution of the genetic variation among the human and animal isolates if no genetic variation affects the infection of different host types.

We identified 36 pairs of phylogenetically matched human and animal isolates from the phylogeny of 437 *S. aureus* isolates (Figures 4A and 4B). The median number of SNPs between the paired human and animal isolates was ~320. Considering the small number of identified isolate pairs, genetic variation with unadjusted p value <0.05 was denoted statistically significant based on the exact McNemar’s test to avoid over-penalizing the p values for statistical significance when correcting for multiple testing using Bonferroni correction. Our findings rejected the hypothesis that no genetic variation influenced *S. aureus* infection in animal and human hosts. We found 20 genes encoding proteins with diverse functions, which were overrepresented in human and animal isolates (Figure 4C; Table S3; Data S4). Three of the four genes found in the GWAS of unmatched isolates were also identified in the GWAS of the matched isolates, which provides further evidence for the role of the effect of these genes in *S. aureus* of humans and animal hosts. These findings demonstrate the improved power of the efficient study design based on the

### Table 1. Summary of unitigs associated with human and animal *S. aureus* isolates based on GWAS using FaST-LMM

<table>
<thead>
<tr>
<th>Locus tag</th>
<th>Gene name</th>
<th>Reference genome</th>
<th>Total unitigs</th>
<th>Q value range$^a$</th>
<th>Odds ratio range</th>
<th>Gene product</th>
</tr>
</thead>
<tbody>
<tr>
<td>JP02758_1300</td>
<td><em>scn</em></td>
<td>AP017922.1</td>
<td>23</td>
<td>$2.25 \times 10^{-39}$ to $2.34 \times 10^{-21}$</td>
<td>0.707–0.746</td>
<td>complement inhibitor SCIN</td>
</tr>
<tr>
<td>JP02758_1303</td>
<td><em>sak</em></td>
<td>AP017922.1</td>
<td>30</td>
<td>$3.84 \times 10^{-12}$ to 0.015</td>
<td>0.745–0.834</td>
<td>staphylocinase</td>
</tr>
<tr>
<td>JP02758_1973</td>
<td><em>tnp</em></td>
<td>AP017922.1</td>
<td>5</td>
<td>0.0191–0.0359</td>
<td>1.352</td>
<td>IS1272 transposase</td>
</tr>
<tr>
<td>JP02758_1367</td>
<td><em>int</em></td>
<td>AP017922.1</td>
<td>57</td>
<td>$3.26 \times 10^{-1}$ to 0.04864</td>
<td>0.742–0.809</td>
<td>integrase</td>
</tr>
<tr>
<td>Intergenic</td>
<td></td>
<td>AP017922.1</td>
<td>110</td>
<td>$6.56 \times 10^{-1}$ to 0.0244</td>
<td>0.705–0.840</td>
<td>hypothetical protein</td>
</tr>
<tr>
<td>No match</td>
<td></td>
<td>No match</td>
<td>129</td>
<td>$1.24 \times 10^{-35}$ to 0.0484</td>
<td>0.714–0.813</td>
<td>no match</td>
</tr>
<tr>
<td>C2G36_RS11395</td>
<td></td>
<td>NZ_CP030138.1</td>
<td>1</td>
<td>$4.27 \times 10^{-06}$ to $4.27 \times 10^{-6}$</td>
<td>0.835–0.835</td>
<td>AAA family ATPase</td>
</tr>
<tr>
<td>C7M54_RS09660</td>
<td></td>
<td>NZ_CP029668.1</td>
<td>1</td>
<td>0.0051–0.0051</td>
<td>1.448–1.448</td>
<td>IS1182 family transposase</td>
</tr>
<tr>
<td>CGP86_RS04080</td>
<td></td>
<td>NZ_CP022720.1</td>
<td>1</td>
<td>0.0002</td>
<td>1.386–1.386</td>
<td>IS1182 family transposase</td>
</tr>
<tr>
<td>CPC18_RS10665</td>
<td></td>
<td>NZ_CP023561.1</td>
<td>5</td>
<td>$8.64 \times 10^{-06}$ to 0.0439</td>
<td>1.462–1.348</td>
<td>IS1182 family transposase</td>
</tr>
<tr>
<td>CPC18_RS13120</td>
<td></td>
<td>NZ_CP023561.1</td>
<td>4</td>
<td>0.0002–0.0003</td>
<td>1.399–1.390</td>
<td>IS1182 family transposase</td>
</tr>
<tr>
<td>CU118_RS06940</td>
<td></td>
<td>NZ_CP024998.1</td>
<td>1</td>
<td>$3.33 \times 10^{-9}$</td>
<td>0.799–0.799</td>
<td>AAA family ATPase</td>
</tr>
<tr>
<td>E3S65_RS00020</td>
<td></td>
<td>NZ_CP047859.1</td>
<td>2</td>
<td>0.0007–0.0014</td>
<td>0.798–0.803</td>
<td>Clp protease ClpP</td>
</tr>
<tr>
<td>E3S65_RS00025</td>
<td></td>
<td>NZ_CP047859.1</td>
<td>2</td>
<td>0.0002–0.0010</td>
<td>0.780–0.791</td>
<td>phage major capsid protein</td>
</tr>
<tr>
<td>E3S65_RS14240</td>
<td></td>
<td>NZ_CP047859.1</td>
<td>4</td>
<td>$2.62 \times 10^{-08}$ to $3.81 \times 10^{-6}$</td>
<td>0.809–0.823</td>
<td>AAA family ATPase</td>
</tr>
<tr>
<td>F6Y18_RS001700</td>
<td></td>
<td>NZ_AP020316.1</td>
<td>1</td>
<td>0.0457–0.0457</td>
<td>1.292–1.292</td>
<td>IS1182 family transposase</td>
</tr>
<tr>
<td>FP479_RS08875</td>
<td></td>
<td>NZ_CP042008.1</td>
<td>1</td>
<td>0.0041–0.0042</td>
<td>1.351–1.351</td>
<td>IS1182 family transposase</td>
</tr>
<tr>
<td>I3K83_RS01375</td>
<td></td>
<td>NZ_CP065199.1</td>
<td>2</td>
<td>$2.69 \times 10^{-06}$ to 0.0489</td>
<td>1.546–1.342</td>
<td>IS1182 family transposase</td>
</tr>
<tr>
<td>I3K83_RS04555</td>
<td></td>
<td>NZ_CP065199.1</td>
<td>1</td>
<td>$8.27 \times 10^{-07}$ to $8.27 \times 10^{-7}$</td>
<td>0.667–0.667</td>
<td>transposase</td>
</tr>
<tr>
<td>I6J76_RS02915</td>
<td></td>
<td>NZ_CP069470.1</td>
<td>2</td>
<td>0.0002–0.0002</td>
<td>1.399–1.399</td>
<td>IS1182 family transposase</td>
</tr>
<tr>
<td>I6J76_RS05230</td>
<td></td>
<td>NZ_CP069470.1</td>
<td>1</td>
<td>0.043886364–0.43886364</td>
<td>1.348–1.348</td>
<td>IS1182 family transposase</td>
</tr>
<tr>
<td>ILP77_RS09555</td>
<td></td>
<td>NZ_CP062467.1</td>
<td>1</td>
<td>0.017527775–0.01752777</td>
<td>1.288–1.288</td>
<td>IS1182 family transposase</td>
</tr>
<tr>
<td>JP02758_1299</td>
<td></td>
<td>AP017922.1</td>
<td>12</td>
<td>$1.24 \times 10^{-06}$ to $1.15 \times 10^{-16}$</td>
<td>0.714–0.761</td>
<td>phage protein</td>
</tr>
<tr>
<td>JP02758_1302</td>
<td></td>
<td>AP017922.1</td>
<td>8</td>
<td>$3.41 \times 10^{-11}$ to 0.0022</td>
<td>0.763–0.804</td>
<td>amidase</td>
</tr>
<tr>
<td>JP02758_1304</td>
<td></td>
<td>AP017922.1</td>
<td>3</td>
<td>$2.01 \times 10^{-06}$ to 0.032490283</td>
<td>0.794–0.794</td>
<td>lytic enzyme</td>
</tr>
<tr>
<td>JP02758_1310</td>
<td></td>
<td>AP017922.1</td>
<td>29</td>
<td>$6.02 \times 10^{-08}$ to 0.0237</td>
<td>0.748–0.805</td>
<td>putative protein SA1764</td>
</tr>
<tr>
<td>JP02758_1330</td>
<td></td>
<td>AP017922.1</td>
<td>2</td>
<td>$2.7 \times 10^{-05}$ to 0.010</td>
<td>0.769–0.806</td>
<td>phage protein</td>
</tr>
<tr>
<td>JP02758_1340</td>
<td></td>
<td>AP017922.1</td>
<td>6</td>
<td>$9.64 \times 10^{-07}$ to 0.0471</td>
<td>0.817–0.849</td>
<td>hypothetical protein</td>
</tr>
<tr>
<td>JP02758_1355</td>
<td></td>
<td>AP017922.1</td>
<td>1</td>
<td>$5.91 \times 10^{-10}$ to $5.91 \times 10^{-10}$</td>
<td>0.758–0.758</td>
<td>phage protein</td>
</tr>
<tr>
<td>JP02758_2192</td>
<td></td>
<td>AP017922.1</td>
<td>2</td>
<td>$3.32 \times 10^{-08}$ to $9.86 \times 10^{-8}$</td>
<td>0.649–0.675</td>
<td>transposase</td>
</tr>
<tr>
<td>K9B11_RS08960</td>
<td></td>
<td>NZ_CP083728.1</td>
<td>2</td>
<td>0.0005–0.0005</td>
<td>1.424–1.424</td>
<td>IS1182 family transposase</td>
</tr>
<tr>
<td>LO764_RS13470</td>
<td></td>
<td>NZ_CP087593.1</td>
<td>1</td>
<td>0.0301–0.0301</td>
<td>0.899–0.899</td>
<td>E domain-containing protein</td>
</tr>
<tr>
<td>RK77_RS01455</td>
<td></td>
<td>NZ_CP026064.1</td>
<td>1</td>
<td>$1.79 \times 10^{-06}$ to $1.79 \times 10^{-6}$</td>
<td>0.673–0.673</td>
<td>transposase</td>
</tr>
</tbody>
</table>

$^a$Q value represents the adjusted p value based on the Bonferroni correction (see STAR Methods).
Figure 4. A phylogeny-based sampling of matched human and animal isolates confirms genes implicated in the GWAS of unmatched isolates and reveals improved power to discover hidden genes associated with host transmissibility, infection, switching, and adaptation

(A) Maximum-likelihood phylogeny showing genetic similarity of S. aureus isolates sampled from humans and animals in New England, United States (see Figure 1E). The phylogenetic branches are square root transformed for clarity.

(B) Phylogeny of the selected pairs of genetically matched human and animal isolates for the matched GWAS using the exact McNemar’s test.

(C) Contingency tables show the presence (1) and absence (0) of genes statistically overrepresented among S. aureus isolates collected from humans and animals based on the GWAS of matched human and animal isolates.
from 0 to 1, to quantify the contribution of specific key genetic determinants critical for host transmissibility, infection, switching, and adaptation. We next calculated the narrow-sense heritability, which ranges between animal and human hosts, we hypothesized that genomic variation associated with the infection of human and animal hosts.

High heritability pinpoints the remarkable contribution of specific key genetic determinants critical for host transmissibility, infection, switching, and adaptation. We next calculated the narrow-sense heritability, which ranges from 0 to 1, to quantify the contribution of S. aureus genetic variation to the infection of different host types using three approaches: genome-wide complex trait analysis (GCTA), FaST-LMM, and GEMMA. Since S. aureus clones typically switch or jump between animal and human hosts, we hypothesized that genomic variation associated with these events is under strong natural selection and, therefore, would exhibit high heritability. Consistent with our hypothesis, we found a remarkably high heritability for infection of different host types based on the SNP variation measured by GEMMA ($h^2 = 0.88$; 95% CI, 0.75–0.92), FaST-LMM ($h^2 = 0.88$), and GCTA ($h^2 = 0.72$; 95% CI, 0.82–0.91) (Figure 5A). Similarly, heritability estimates based on the genetic variation captured by the unitig sequences were high, with the highest estimate calculated by GEMMA ($h^2 = 0.84$; 95% CI, 0.75–0.92) followed by FaST-LMM ($h^2 = 0.84$) and GCTA ($h^2 = 0.77$; 95% CI, 0.66–0.88) (Figure 5B). However, the heritability estimates based on the kinship matrix generated based on the presence and absence of the accessory gene were lower than for the SNPs and unitigs (Figure 5C). Therefore, the heritability estimates based on the presence and absence of genes underestimated the heritability compared with the estimates based on the SNP and unitig genetic variation, which is not unexpected because genes do not capture variability in the intergenic variation. Despite this, for context, the heritability estimates based on gene content were much higher than estimates reported for other bacterial phenotypes, such as disease severity. Such high heritability is similar to the loci associated with AMR phenotypes in several bacterial pathogens, including M. tuberculosis and Neisseria gonorrhoeae. Such AMR-associated loci are typically subject to strong natural selection pressure, implying that the implicated loci in S. aureus are similarly subjected to substantial selective forces. These findings demonstrate that genetic variation is the primary driver of S. aureus infection of different host types, highlighting its impact on host transmissibility, infection, switching, and adaptation.

We then determined the amount of heritability attributable to the S. aureus prophage-encoded immune evasion genes. First, we repeated the GWAS using GEMMA by including a covariate for the presence and absence of the four genes statistically significantly associated with the host type in the GWAS using FaST-LMM, namely the SCIN (locus tag: JP02758_0997), staphylokinase sak (locus tag: JP02758_1303), and a hypothetical amidase gene (locus tag: JP02758_1302) (Figures 3B–3G). This analysis allowed us to calculate the point estimate for the narrow-sense heritability not explained by the immune evasion genes in the Sa3 prophage. This approach was previously used to quantify the heritability attributable to the Panton-Valentine leukocidin (PVL) locus encoding a cytotoxin crucial for staphylococcal pyomyositis infection. The point estimates for the heritability were 0.0003 and 0.0007 based on the SNP and unitig genetic variation measured by GEMMA. These estimates correspond to a 99.97% (0.882–0.0003) and 99.92% (0.838–0.0007) decrease in the heritability explained by factors...
other than the immune evasion genes. These findings attributed nearly all the heritability for host transmissibility and infection to the immune evasion genes in the \( \Phi \)Sa3 prophage. These results highlight the critical role of variability in the presence and absence patterns of these genes in \( S. \) aureus host transmissibility, infection, switching, and adaptation.

**HGT drives the dissemination of the genes for host transmissibility, infection, switching, and adaptation**

HGT and homologous recombination play a crucial role in disseminating pathogenicity loci of bacteria,\(^{60,61}\) including \( S. \) aureus.\(^{62}\) We hypothesized that the rapid acquisition and loss of the prophage through HGT contributes to the rapid dissemination of the immune evasion genes between \( S. \) aureus strains. Therefore, we next investigated whether the prophage region harboring the immune evasion genes implicated in the GWAS was located within a hotspot for genetic exchange via recombination.

We generated whole-genome sequence alignments of the \( S. \) aureus isolates belonging to the major clonal complexes, namely, CC1, CC5, CC8, CC30, CC45, and CC97. We selected these clones because they contained at least 25 isolates, sufficient for a robust phylogenomic analysis to detect recombination.

**Figure 6. The \( \Phi \)Sa3 prophage encoding genetic variation associated with \( S. \) aureus host transmissibility, infection, switching, and adaptation is the major hotspot for genetic exchange via recombination**

(A) Maximum-likelihood phylogeny of 25 \( S. \) aureus CC1 isolates annotated by host type. The rectangular matrix adjacent to the color strips at the tips of the phylogeny represents a zoomed plot showing recombination events detected by Gubbins in the genomic region 1,430,443 to 1,472,709 containing the \( \Phi \)Sa3 prophage in the reference sequence for the human \( S. \) aureus strain JP080 (GenBank: AP017922.1). The presence of the prophage was inferred using PHASTER. Recombination blocks colored in red were found in more than one isolate in the phylogeny, while those colored in blue were unique to a single isolate. The genes in the forward and reverse genome strands are shown in different colors for clarity.

(B) Recombination events in the whole-genome alignment of 30 CC30 isolates.

(C) Recombination events in the whole-genome alignment of 27 CC45 isolates.

(D) Recombination events in the whole-genome alignment of 114 CC5 isolates.

(E) Recombination events in the whole-genome alignment of 107 CC8 isolates.

(F) Recombination events in the whole-genome alignment of 33 CC97 isolates. Recombination plots showing the whole genome are shown in Figure S7.
We found that the genomic region harboring the prophage region was the primary hotspot for homologous recombination consistent across different S. aureus clonal complexes, consistent with findings elsewhere. We isolated prophage sequences from S. aureus strain JP080 (GenBank: AP017922.1) isolated from a human infection to determine the presence and absence of genes and intergenic sequences. We found the presence of highly variable prophage sequences in different isolates, with marked differences in the presence and absence of the prophage-encoded genes (Figure S8). There was some noticeable homology at the 5' and 3' ends of the prophage, suggesting the potential presence of Sa3-like and other prophage sequences, including Sa1, Sa3, Sa5, Sa6, Sa7, and Sa9, and especially those lacking the immune evasion genes.

The observed correlation between the presence and absence of the immune evasion genes and host type was apparent, supporting the association reported in the GWAS. Since the mapping-based approach may not capture genetic variation outside the reference Sa3 prophage sequence, next we performed a complementary analysis using de novo assemblies. We extracted full prophage sequences found on a single contig from de novo assemblies of the isolates and compared their genetic diversity using the Jaccard index (i.e., proportion of shared k-mers). We found 173 (53.56%) human isolates containing a complete prophage sequence compared with 20 (17.54%) of the animal isolates. The mean length of the extracted prophage sequences was 42,541.35 bp (range: 40,467–44,418 bp), and clustering analysis showed a high genetic diversity between the sequences (Figure S9). These findings confirmed that the Sa3-like prophages are more common in human than animal S. aureus isolates and exhibit substantial genetic diversity, consistent with findings elsewhere.

**DISCUSSION**

Understanding the factors critical for host transmissibility, infection, switching, and adaptation of pathogens with multi-host ecologies, such as S. aureus, remains vital to inform measures to improve human and animal health. Here, using a comprehensive population genomics approach, we identified non-recent transmissions of S. aureus and uncovered key Sa3 \( \beta \)-hemolysin-converting bacteriophage-encoded genetic loci critical for host-switching and infection in humans and animals. Such host-switching and infection are driven by the acquisition and loss of the human innate immune evasion factors, specifically the SCIN (scn) and staphylokinase (sak) genes, and two hypothetical genes (one encoded a phage lysis or amidase and an AAA ATPases family protein). While other genetic loci located outside the Sa3 prophage were associated with S. aureus infection of the human and animal host types, we found the strongest genetic signal in the prophage region. Importantly, S. aureus genetics explained \( \sim 88 \% \) of the heritability in the infection of human and animal host types, of which \( \sim 99.9 \% \) was attributed to the immune evasion genes within the Sa3 prophage, highlighting the remarkable contribution of these genes to S. aureus host transmissibility, infection, switching, and adaptation. Considering the rapid dissemination of the prophages between S. aureus strains in the population, our findings suggest that any S. aureus clone, including those typically considered specialist lineages, can rapidly evolve and become endemic in livestock and humans upon acquisition or loss of the genes implicated in this study, thereby promoting the spread of antibiotic-resistant and virulent genes to strains in animals and humans.

Our GWAS approach based on different types of genetic variation (SNPs, genes, and unitigs) has revealed the role of the innate immune evasion genes, which display high specificity to the human immune system, specifically the SCIN (scn) and SAK (sak) genes encoded by the Sa3 prophage in S. aureus host-switching and adaptation. These results are consistent with previous pan-genome studies. Specifically, it encodes a plasminogen activator converting plasminogen to plasmin, which cleaves and removes opsonic molecules, including human immunoglobulin (Ig) G and complement component 3b (C3b), thereby preventing neutrophil-mediated phagocytosis, and generation of complement component 5a (C5a). The sak gene interferes with human innate immune defenses by exhibiting anti-opsonic activity. It encodes a plasminogen activator converting plasminogen to plasmin, which cleaves and removes opsonic molecules, including human immunoglobulin (Ig) G and complement component 3b (C3b), thereby preventing neutrophil-mediated phagocytosis, and generation of complement component 5a (C5a). The sak gene interferes with human innate immune defenses by exhibiting anti-opsonic activity. It encodes a plasminogen activator converting plasminogen to plasmin, which cleaves and removes opsonic molecules, including human immunoglobulin (Ig) G and complement component 3b (C3b), thereby preventing neutrophil-mediated phagocytosis.
pathogenicity of *S. epidermidis*. Several experimental studies, which showed the effect of these prophages in the infection of different host types, support our findings indicating the key role of the immune evasion factors in *S. aureus* host transmissibility, infection, switching, and adaptation. Although the animal-associated *S. aureus* isolates are not entirely devoid of the \( \phi S a3 \)-like prophages, these are found at low frequency, enough to facilitate spread to other animal isolates, and this may promote infection to human hosts. However, such a low prevalence of these prophages, \( \phi S a3 \)-containing isolates in animals and \( \phi S a3 \)-devoid isolates in humans, implies that carriage of the prophage incurs a substantial fitness cost to the bacteria in the animal host environment, likely under strong negative and positive selection in animals and humans, respectively. Thus, our findings indicate that the acquisition of the \( \phi S a3 \) prophage-encoded human innate immune evasion genes drives rapid host-switching and adaptation of *S. aureus* over short and long timescales, facilitated by the repeated transmissions between animals and humans. However, further studies are required to understand specific factors modulating the spread of these prophages between *S. aureus* strains in different hosts.

Although several studies have postulated the importance of bacterial loci in host-switching and adaptation of *S. aureus*, the overall contribution of the pathogen genetics remains less understood. Our findings suggest remarkably high narrow-sense heritability, reaching \( \sim 88\% \) based on SNP and unigene sequence variation, highlighting the major contribution of *S. aureus* to infection of different host types. Such high heritability is consistent with estimates reported for highly penetrant phenotypes, typically under strong natural selection pressure. An example of such a phenotype is AMR, whose heritability reached \( \sim 84\% \) and \( \sim 97\% \) in *M. tuberculosis* and *N. gonorrhoeae* respectively. Similarly, the establishment of *S. aureus* pyomyositis infection has also shown a high heritability of \( \sim 64\% \). Consistent with our findings, *S. aureus* pyomyositis infection is also driven by a single bacteriophage-associated locus, the staphylococcal PVL locus, which can rapidly spread between strains over short timescales to promote rapid phenotypic changes and adaptation. An in-depth investigation of the heritability revealed that the immune evasion genes in the \( \phi S a3 \) prophage contributed \( \sim 99.9\% \) of the phenotypic variability, clearly pinpointing their role in *S. aureus* infection in animals and human hosts. Overall, these findings demonstrate the critical role of the human innate immune evasion genes in the \( \phi S a3 \) prophage for *S. aureus* host transmissibility, infection, switching, and adaptation.

In conclusion, we have identified and quantified the contribution of the primary genetic determinants for *S. aureus* host transmissibility, infection, switching, and adaptation. These findings have direct implication for designing therapeutic measures to limit the emergence and spread of pandemic *S. aureus* clones, especially those associated with antimicrobial resistance, and to improve human and animal health and sustainable food security. Our population genomic study of human and animal-associated isolates underscores the need for a broader “one-health” perspective to understand and manage *S. aureus* and other multi-host pathogens with complex ecologies.

**Limitations of the study**

Our study benefits from a targeted focus on *S. aureus* isolates sampled from diseased humans and animals in the same region. Our results agree with previous studies suggesting the impact of the immune evasion genes, but we present newer and more robust genomic approaches to study host adaptation, which have revealed genetic variation important for adaptation, quantified their heritability, and shown unrecognized diversity in the species. However, our study has several limitations. First, the isolates were sampled from individuals with disease; therefore, they may not have captured the full extent of *S. aureus* genetic diversity colonizing human and animal hosts. We also acknowledge the partly overlapping sampling times and the lesser representation of the animal isolates, which may reduce the likelihood of detecting more recent *S. aureus* transmission between human and animal hosts. However, our analysis showed sufficient power to detect genetic variation significantly associated with human and animal hosts because our phenotype had high penetrance, large effect sizes, and high heritability. Other GWAS studies elsewhere that have used smaller or similarly sized datasets, for example, *S. epidermidis* (76 controls and 76 controls; genetically matched) and *S. aureus* (101 cases and 417 controls; un-matched), showed sufficient power to detect genetic signals associated with phenotypes. We also utilized a phylogeny-based sampling of matched closely related pairs of human and animal isolates to eliminate residual confounding effects due to the cryptic population structure and uneven dataset sizes for the human and animal isolates. This approach has previously been applied to datasets of *M. tuberculosis* (eight cases and eight controls; phylogenetically matched) and *Campylobacter* species (eight cases and eight controls; phylogenetically matched) and showed improved power to detect genetic signals. We also recognize that different animal species have a unique physiology that requires corresponding unique adaptive traits of bacterial pathogens. Nonetheless, by grouping all the isolates from animals, our dataset is biased toward human samples, thus the genetic signatures identified in this study could be specific to human hosts. Therefore, future work should include systematic sampling from different animal species to better understand host adaptation and switching between non-human hosts. Our work provides an essential framework for future investigations on host adaptation in *S. aureus* and other bacterial pathogens.

**STAR METHODS**

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

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
  - Lead contact
  - Materials availability
  - Data and code availability
- **EXPERIMENTAL MODEL AND SUBJECT DETAILS**
  - Ethics statement
- **METHOD DETAILS**
  - Sample characteristics and microbiological processing
  - Whole-genome sequencing of *S. aureus* isolates
QUANTIFICATION AND STATISTICAL ANALYSIS

- Molecular typing, de novo genome assembly, annotation, and pan-genome construction
- Population structure, phylogenetic construction, and recombination
- Generating variant data for bacterial GWAS
- Genome-wide association analysis
- Comparative genomics of prophage sequences

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2022.100194.

ACKNOWLEDGMENTS

The authors would like to thank the study participants and guardians, and the clinical and laboratory staff who collected and processed the samples at various laboratories. We would also like to thank Dr. Bernadette Young at the University of Oxford for providing advice regarding estimating heritability attributable to a single locus and two anonymous reviewers for their careful reviewing of our manuscript and their insightful comments and suggestions. The study was supported by the National Institutes of Health (NIH) award no. 1R35GM142924 to C.P.A. The funders had no role in study design, data collection, analysis, decision to publish, or manuscript preparation, and the findings do not necessarily reflect the official views and policies of the authors’ institutions and funders. For the purpose of open access, the authors have applied a CC BY public copyright license to any author-accepted manuscript version arising from this submission.

AUTHOR CONTRIBUTIONS

C.C. and C.P.A. conceived and designed the study. C.C., J.T.S., and S.A.B. performed statistical and bioinformatics analyses. J.T.S., R.G., and I.W.M. performed sampling, culturing, and DNA extractions. C.C. and C.P.A. wrote the manuscript. C.P.A. supervised the project and acquired the funding. All authors reviewed and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Livestock origin for a human pandemic clone of community-associated meticillin-resistant Staphylococcus aureus. mBio 4, 003566.


### KEY RESOURCES TABLE

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bacterial and virus strains</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S. aureus isolates from animal and human hosts</td>
<td>PubMed</td>
<td>PubMed: 35292708 and 34154550</td>
</tr>
<tr>
<td><strong>Biological samples</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S. aureus isolates</td>
<td>PubMed</td>
<td>PubMed: 35292708 and 34154550</td>
</tr>
<tr>
<td><strong>Critical commercial assays</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QuickDNA Fungal/Bacterial Miniprep Kit</td>
<td>Zymo Research, Irvine, CA</td>
<td>Cat. No. D6005</td>
</tr>
<tr>
<td>Tryptic soy agar</td>
<td>Remel, Lenexa, KS</td>
<td>Cat. No. R111007</td>
</tr>
<tr>
<td>Brain heart infusion broth</td>
<td>BD Difco, Franklin Lakes, NJ</td>
<td>Cat. No. 211059</td>
</tr>
<tr>
<td>Qubit fluorometer</td>
<td>Invitrogen, Grand Island, NY</td>
<td>Cat. No. Q33266</td>
</tr>
<tr>
<td>RipTide High Throughput Rapid DNA Library Prep kit</td>
<td>iGenomX, Carlsbad, CA</td>
<td>Cat. No. 104950</td>
</tr>
<tr>
<td><strong>Deposited data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S. aureus isolates</td>
<td>This manuscript</td>
<td>NCBI: PRJNA673382 and PRJNA741582</td>
</tr>
<tr>
<td><strong>Software and algorithms</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>CRAN</td>
<td><a href="https://www.R-project.org/">https://www.R-project.org/</a></td>
</tr>
<tr>
<td>APE</td>
<td>CRAN</td>
<td><a href="https://cran.r-project.org/web/packages/ape/">https://cran.r-project.org/web/packages/ape/</a></td>
</tr>
<tr>
<td>Phylosignal</td>
<td>CRAN</td>
<td><a href="https://cran.r-project.org/web/packages/phylosignal/">https://cran.r-project.org/web/packages/phylosignal/</a></td>
</tr>
<tr>
<td>Phytools</td>
<td>CRAN</td>
<td><a href="https://cran.r-project.org/web/packagesphytools/">https://cran.r-project.org/web/packagesphytools/</a></td>
</tr>
<tr>
<td>Geiger</td>
<td>CRAN</td>
<td><a href="http://cran.r-project.org/web/packages/geiger/">http://cran.r-project.org/web/packages/geiger/</a></td>
</tr>
<tr>
<td>FaST-LMM</td>
<td>GitHub</td>
<td><a href="https://fastlmm.github.io/">https://fastlmm.github.io/</a></td>
</tr>
<tr>
<td>GEMMA</td>
<td>GitHub</td>
<td><a href="https://github.com/genes-statistics/GEMMA">https://github.com/genes-statistics/GEMMA</a></td>
</tr>
<tr>
<td>PLINK</td>
<td>Harvard University</td>
<td><a href="https://zzz.bwh.harvard.edu/plink/">https://zzz.bwh.harvard.edu/plink/</a></td>
</tr>
<tr>
<td>RCandy</td>
<td>GitHub</td>
<td><a href="https://github.com/ChrispinChaguza/RCandy">https://github.com/ChrispinChaguza/RCandy</a></td>
</tr>
<tr>
<td>VCFTools</td>
<td>Sourceforge</td>
<td><a href="http://vcftools.sourceforge.net/">http://vcftools.sourceforge.net/</a></td>
</tr>
<tr>
<td>Samtools</td>
<td>GitHub</td>
<td><a href="https://samtools.github.io/">https://samtools.github.io/</a></td>
</tr>
<tr>
<td>Prokka</td>
<td>GitHub</td>
<td><a href="https://github.com/tseemann/prokka">https://github.com/tseemann/prokka</a></td>
</tr>
<tr>
<td>Shovill</td>
<td>GitHub</td>
<td><a href="https://github.com/tseemann/shovill">https://github.com/tseemann/shovill</a></td>
</tr>
<tr>
<td>BWA MEM</td>
<td>Anaconda</td>
<td><a href="https://anaconda.org/tseemann/bwa">https://anaconda.org/tseemann/bwa</a></td>
</tr>
<tr>
<td>Snippy</td>
<td>GitHub</td>
<td><a href="https://github.com/tseemann/snippy">https://github.com/tseemann/snippy</a></td>
</tr>
<tr>
<td>Cytoscape</td>
<td>Cytoscape Team</td>
<td><a href="https://cytoscape.org/">https://cytoscape.org/</a></td>
</tr>
<tr>
<td>RAST server</td>
<td>Argonne National Laboratory</td>
<td><a href="https://rast.nmpdr.org/">https://rast.nmpdr.org/</a></td>
</tr>
<tr>
<td>PHASTER</td>
<td>University of Alberta</td>
<td><a href="https://phaster.ca/">https://phaster.ca/</a></td>
</tr>
<tr>
<td>ACT</td>
<td>Wellcome Sanger Institute</td>
<td><a href="https://www.sanger.ac.uk/tool/artemis-comparison-tool-act/">https://www.sanger.ac.uk/tool/artemis-comparison-tool-act/</a></td>
</tr>
<tr>
<td>BEDTools</td>
<td>University of Utah</td>
<td><a href="https://bedtools.readthedocs.io/en/latest/">https://bedtools.readthedocs.io/en/latest/</a></td>
</tr>
<tr>
<td>GCTA</td>
<td>Westlake University</td>
<td><a href="https://yanglab.westlake.edu.cn/software/gcta/">https://yanglab.westlake.edu.cn/software/gcta/</a></td>
</tr>
<tr>
<td>qqman</td>
<td>CRAN</td>
<td><a href="https://cran.r-project.org/web/packages/qqman/">https://cran.r-project.org/web/packages/qqman/</a></td>
</tr>
<tr>
<td>MLST</td>
<td>PubMLST</td>
<td><a href="https://pubmlst.org/multilocus-sequence-typing">https://pubmlst.org/multilocus-sequence-typing</a></td>
</tr>
<tr>
<td>Bifrost</td>
<td>GitHub</td>
<td><a href="https://github.com/pmelsted/bifrost">https://github.com/pmelsted/bifrost</a></td>
</tr>
<tr>
<td>Panaroo</td>
<td>GitHub</td>
<td><a href="https://github.com/gtonkinhill/panaroo">https://github.com/gtonkinhill/panaroo</a></td>
</tr>
<tr>
<td>Snp-sites</td>
<td>GitHub</td>
<td><a href="https://github.com/sanger-pathogens/snp-sites">https://github.com/sanger-pathogens/snp-sites</a></td>
</tr>
</tbody>
</table>

(Continued on next page)
RESOURCE AVAILABILITY

Lead contact
Further information and requests for data, resources, and reagents should be directed to and will be fulfilled by the Lead Contact, Cheryl P. Andam (candam@albany.edu).

Materials availability
This study did not generate new unique reagents. However, the raw data and code for this study can be found in the Supplemental Materials and repository specified in Data and code availability.

Data and code availability
- The sequence data used in this study were deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive under the BioProject accession numbers NCBI: PRJNA673382 (human isolates) and NCBI: PRJNA741582 (animal isolates). The accession numbers for individual isolates are provided in the supplementary material.
- A summary of the isolates used in this study and GWAS hits are available in Data S1, S2, S3, and S4. Additional code and data are publicly available from GitHub (https://github.com/ChrispinChaguza/SAUREUS_NE_USA) and Zenodo (https://doi.org/10.5281/zenodo.6944550).
- All other data supporting the findings of this study are available within the paper and its supplementary information files. Any additional information required is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Ethics statement

Human isolates
Ethical approval was granted by the Committee for the Protection of Human Subjects of Dartmouth-Hitchcock Medical Center and Dartmouth College. This study protocol was deemed not to be human subjects research. Samples used in the study were subcultured bacterial isolates that had been archived in the routine course of clinical laboratory operations. No patient specimens were used, and patient protected health information was not collected.

Animal isolates
The clinical specimens were received from multiple veterinary practices in the states of Connecticut, New Hampshire, Maine, Massachusetts, and Vermont. All isolates were from animals with confirmed clinical infections. No live vertebrates were used by the New Hampshire Veterinary Diagnostic Laboratory (NHVDL) in this study; hence, the NHVDL was exempt from the Institutional Animal Care and Use Committee (IACUC) approval process at the University of New Hampshire.
METHOD DETAILS

Sample characteristics and microbiological processing
Four hundred and thirty-seven isolates collected from humans and animals in New England, USA were available for this study. The isolates were collected between 2010 and 2020. Of these, 323 were collected from humans with bacteraemia infection, while 114 isolates were sampled from animals. We collected a single isolate from each patient and animal. For the human isolates, the first significant blood culture isolate from each patient is routinely archived at the Dartmouth-Hitchcock Medical Center, New Hampshire, USA, in case of future need for patient care, epidemiologic, public health, or laboratory quality studies. Upon subculture, isolates were assigned a study number and all patient identifiers were removed with only the date of collection and results of clinical antimicrobial susceptibility testing linked to the study number. We selected a convenience sample of approximately half of the unique patient isolates distributed throughout the study period for this study. For the animal isolates, isolates were obtained as culture swabs from routine clinical specimen submissions to the New Hampshire Veterinary Diagnostic Laboratory, New Hampshire, USA. Clinical specimens were received from multiple veterinary practices in four states (New Hampshire, Maine, Massachusetts, Vermont).

Whole-genome sequencing of S. aureus isolates
The S. aureus isolates were subcultured from glycerol stocks onto commercially prepared tryptic soy agar with 10% sheep red blood cells (Remel, Lenexa, KS) and in brain heart infusion broth (BD Difco, Franklin Lakes, NJ) at 37°C for 24 h. DNA was extracted and purified from the liquid culture using the Zymo Research QuickDNA Fungal/Bacterial Miniprep Kit (Irvine, CA) following the manufacturer’s protocol. We used a Qubit fluorometer (Invitrogen, Grand Island, NY) to measure DNA concentration. DNA libraries were prepared using the RipTide High Throughput Rapid DNA Library Prep kit (iGenomX, Carlsbad, CA). DNA samples were sequenced as multiplexed libraries on the Illumina HiSeq platform operated per the manufacturer’s instructions. Sequencing resulted in 250 nucleotides long paired-end reads. Sequencing was carried out at the UNH Hubbard Center for Genome Studies at the University of New Hampshire, Durham, NH, USA.

QUANTIFICATION AND STATISTICAL ANALYSIS

Molecular typing, de novo genome assembly, annotation, and pan-genome construction
The isolates were assigned into clones or sequence types and clonal complexes using the multilocus sequence typing (MLST) scheme for S. aureus. We assembled the reads into contigs using Shovill (version 1.1.0) with the no “matchtrim” option to generate high-quality draft genomes (https://github.com/tseemann/shovill). The Shovill pipeline implements subsampling reads to depth x 150 per base coverage, trimming adapters, correcting sequencing errors, and assembly using SPAdes (version 3.13.0). We then used SSPACE (version 3.0) and GapFiller (version 1.10) to scaffold and close gaps to improve the generated assemblies. Genome quality was assessed using the programs QUAST (version 5.0.2) and CheckM (version 1.1.3) to assess the quality of our sequences and exclude genomes with <90% completeness and >5% contamination. Overall, genomes were at least 97.7% complete with no more than 2.86% contamination. We also excluded assemblies with >200 contigs and an N50 < 40,000 bp and percent of heterozygous sites over total number of SNPs >15% when mapped against a JP080 S. aureus reference genome (GenBank accession: AP017922.1). After filtering out the genomes with low coverage and of poor quality, a total of 437 S. aureus genomes, 114 from animals and 323 from humans were used for downstream analyses. The resulting contigs were annotated using Prokka (version 1.14.6). The median number of genes per isolate was 2,596 (range: 2,429 to 2,814), the median number of contigs was 35 (range: 13 to 130), the median of the largest contig was 488,890 bp (range: 136,232 to 1,320,700 bp), the median total assembly length was 2,782,737 bp (range: 2,666,799 to 2,974,399 bp), while the median assembly N50 value was 2,096,692 bp (range: 40,538 to 1,029,233 bp) (Data S1). We generated the core, accessory, and pan-genomes with the moderate stringency mode using Panaroo (version 1.2.2).

Population structure, phylogenetic construction, and recombination
A multi-sequence whole-genome alignment was generated based on consensus sequences of each isolate inferred after mapping reads against a complete reference genome for a human S. aureus strain JP080 (GenBank accession: AP017922.1) using Snippy (version 4.6.0) (https://github.com/tseemann/snippy). The sites containing SNPs in the whole genome alignment were extracted from the alignment in FASTA and variant call format (VCF) formats using SNP-sites (version 2.3.2). We used the SNPs in the FASTA file to generate a maximum-likelihood phylogenetic tree of the S. aureus isolates using IQ-TREE (version 2.1.2) with the general time-reversible (GTR) and Gamma models and 100 bootstraps. We visualised the phylogenetic trees using the APE package (version 4.3) and RCandy (version 1.0.0). We annotated the trees with isolate metadata using the “gridplot” and “phylo4d” functions available in the phylosignal (version 1.3) and phylobase (version 0.8.6) packages (https://cran.r-project.org/package=phylobase), respectively. We calculated the number of SNPs between pairs of isolates based on the core genome alignment generated with Panaroo (version 1.2.2) using snp-dists (version 0.7.0) (https://github.com/tseemann/snp-dists). We generated networks of genetically similar human and animal isolates distinguished by < 150 SNPs using Cytoscape (version 3.8.2). We quantified the correlation between the phylogenetic tree and the host type phenotype, i.e., phylogenetic signal, using Pagel’s λ statistic using the
“fitDiscrete” functions in the Geiger (version 2.0.6.4). The number of transitions between human and animal host type states were inferred using “make.simmap” and “describe.simmap” functions in phytools (version 0.7.70). We used Gubbins to detect recombination events in whole-genome alignments of different isolates from different clonal complexes with at least 25 sequenced genomes.

Generating variant data for bacterial GWAS

To prepare the dataset for the GWAS analysis, we converted the biallelic SNPs into the pedigree file format using VCFtools (version 1.90b4). We excluded rare SNPs with minor allele frequency <5% and missingness >5% using PLINK (version 1.90b4). To prepare the GWAS dataset based on the gene presence and absence patterns, we used custom Python scripts to generate pedigree files based on the presence-absence patterns of orthologous genes generated using Panaroo (version 1.2.2). To identify the maximal unitig sequences, i.e., non-branching paths in a compacted De Bruijn graph, we first build a graph for the entire dataset based on 31 bp k-mer sequences using Bifrost (version 1.0.1). The unitig sequences generated based on the entire isolate collection were queried against a De Bruijn graph of each genome using Bifrost to determine the presence and absence patterns of each unitig sequence in the genomes. A unitig was considered present when exact matches for all the k-mers in the unitig sequence were found in each genome graph. The presence and absence patterns of the unitigs were merged with the affection status to generate the pedigree data files required for the GWAS. As similarly done with the SNP variant data, the genes and unitigs found in <5% of the isolates were excluded from the final dataset for the GWAS using PLINK.

Genome-wide association analysis

We first compared the relative frequency of the MLST clonal complexes using Fisher’s exact test. To identify genomic variation, i.e., SNPs, genes, and unitigs, associated with infection of humans and animals, we performed a univariate analysis based on linear mixed models efficient at correcting the clonal arterial population structure. We initially used FaST-LMM (FastLmmC, version 2.07.20140723) for the GWAS followed by validation using GEMMA (version 0.98.1). To control the population structure of the isolates, which is a major confounder in bacterial GWAS analyses, we specified a random covariate based on the kinship matrix of the pairwise SNP distances between the isolates. All the GWAS analyses based on SNPs, genes, and unitigs used the same genetic similarity matrix. Since the S. aureus genome is haploid, we coded the variants as mitochondrial DNA in the human genome, i.e. designating it chromosome 26, to allow the use of the aforementioned GWAS tools initially developed to primarily handle diploid human genome data. We adjusted the raw statistical significance (p-values) for each variant, inferred using the likelihood ratio test using the Bonferroni correction method to control the false discovery rate due to multiple testing. Since the frequency of genomic variants tested, i.e., accessory genes, SNPs, and unitigs, varied greatly, we used a fixed value for the S. aureus genome size to represent the possible maximum number of realised genomic variants. This approach is more conservative than adjusting based on the observed variants, minimising false positives but potentially increasing false negatives slightly. However, crucially, our approach ensures the use of a consistent p-value threshold when interpreting the statistical significance of different types of genomic variation.

Genetic variants with p-value < 1.83 × 10⁻⁸, i.e., α/G where the statistical significance threshold α = 0.05 and the genome size G = 2,729,352 bp for the S. aureus reference genome of the strain JP080 (GenBank accession: AP017922), were deemed statistically significant. We compared the observed and expected statistical significance to visually inspect potential issues when controlling the population structure using the Q-Q plots generated with gqman (version 0.1.7). The overall proportion of phenotypic variability explained by variation in the genome (narrow-sense heritability) was estimated using FaST-LMM (FastLmmC, version 2.07.20140723), GEMMA (version 0.98.1), and GCTA (version 1.93.2). Since the heritability uses the genetic similarity matrix of the isolates, we used a separate matrix generated for the different variant types to obtain reliable estimates of heritability for each type of genetic variation. To determine the relative decrease in the heritability due to specific variants in the Sa3 prophage, we included covariate in GEMMA for the presence and absence patterns of the prophage-associated genes, an approach similarly used by Young et al. We selected genetically matching pairs of human and animal S. aureus isolates in the phylogenetic tree of all the isolates. We used the Exact McNemar’s test to test whether certain genes were overrepresented in the human or animal isolates.

The genomic features associated with each SNP, accessory gene, and unitig were identified by comparing them with a panel of S. aureus reference genomes using BioPython (version 1.78). In addition, we used BLASTN (version 2.5.0+) to identify genomic regions containing the gene and unitig sequences. Functional annotation of the genes identified in the GWAS was done using EggNOG-mapper. We summarised the functional annotations and generated Manhattan plots for the GWAS results using R (version 4.0.3) (https://www.R-project.org/).

Statistical analysis was performed using R (version 4.0.3) using the test of equal proportions for the comparison of proportions between two groups. The analysis of genetic variation from the matched human-animal isolates was done using the Exact McNemar’s test in the exact2x2 (version 1.6.6) package (https://cran.r-project.org/web/packages/exact2x2/). The statistical significance in the GWAS was corrected for multiple comparisons using the Bonferroni correction. All error bars show the mean and the 95% confidence interval. We tested the difference in the mean number of S. aureus transitions between human and animal host type states using the Kruskal Wallis test. All replicates are biological. Sample numbers are given in the figure legends. No blinding was used. No statistical methods were used to predetermine sample sizes.
Comparative genomics of prophage sequences

We used the PHASTER web tool to identify prophage sequences in S. aureus genomes (https://phaster.ca/). We annotated the identified reference prophage sequence using the RAST server. We mapped unitig sequences for each isolate to a reference ϕSa3 prophage identified at position 1,430,443 to 1,472,709 in the genome of the human S. aureus strain JP080 (GenBank accession: AP017922.1) using BWA MEM (version 0.7.17-r1188). We extracted the genomic coordinates for the mapped sequences using the “bamtobed” option in BEDTools (version 2.30.0) to identify and extract the prophage sequence in the draft S. aureus assemblies for the isolates used in this study. We used a combination of BLASTN (version 2.9.0+) to compare the prophage sequence against the genome sequence of the S. aureus isolates and visualised the results using ACT (version 18.1.0). We generated a visualisation of the presence and absence of sequences in the ϕSa3 prophage using R (version 4.0.3) (https://www.R-project.org/). In addition, we also compared the genetic diversity of the prophage sequences using de novo assemblies. We extracted the ϕSa3 prophage sequences from the genomes using in_silico_PCR.pl (version 0.5.1) (https://github.com/egonozer/in_silico_pcr) and compared their pairwise genetic similarity using the Jaccard index, i.e., fraction of shared k-mers, using the ‘sketch’ (-s 5000) and ‘dist’ functions in MASH (version 2.0). We created a heatmap with clustered dendrograms generated using ‘hclust’ function to show the similarity of the prophage sequences based on the Jaccard indices using pheatmap (version 1.0.12) (https://CRAN.R-project.org/package=pheatmap).