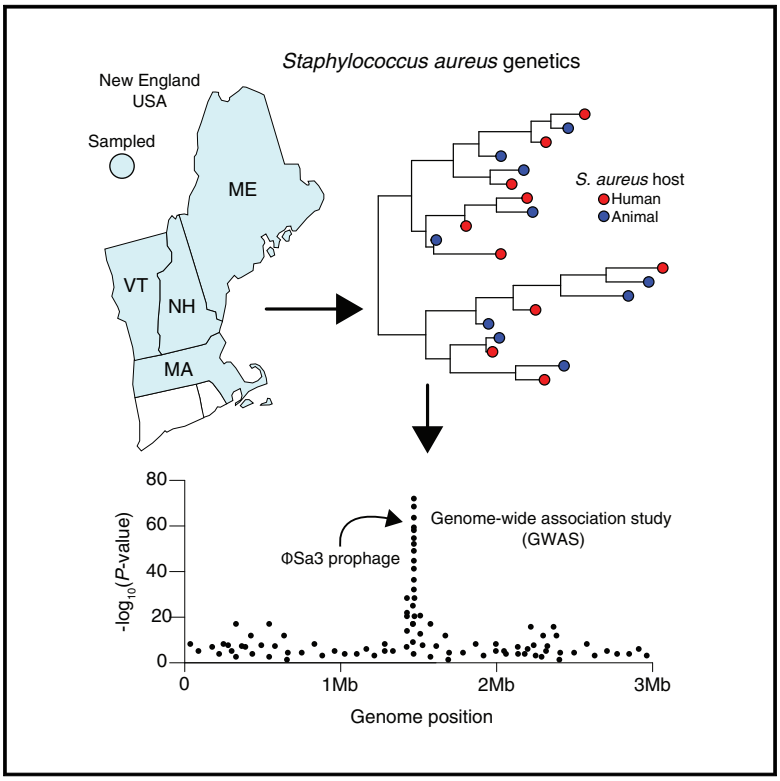
Cell Genomics

Article

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.

Chaguza et al., 2022, Cell Genomics *2*, 100194
November 9, 2022
https://doi.org/10.1016/j.xgen.2022.100194

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Article

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. Andam[3,6,7,*]
[1]Department of Epidemiology of Microbial Diseases, Yale School of Public Health, Yale University, New Haven, CT, USA
[2]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[3]Department of Biological Sciences, University at Albany, State University of New York, New York, USA
[4]New Hampshire Veterinary Diagnostic Laboratory, Durham, NH, USA
[5]Dartmouth-Hitchcock Medical Center and Dartmouth College Geisel School of Medicine, Lebanon, NH, USA
[6]Senior author
[7]Lead 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 animals[1] 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, especially livestock and companion animals living in close proximity to humans, and the extent to which known and novel factors drive cross-species switching. Recent data have revealed the impact of human activities, including animal domestication and antibiotic use in humans and animals, in driving the multi-species evolution and ecology of *S. aureus*.[13] Other genomic studies have also shown variable abundance of virulence-associated genes across different hosts, linked with host-switching and adaptation events in *S. aureus*.[4]

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*,[14] 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.[3,4,12,17–20] 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,[26] healthcare adaptation,[26,27] immune evasion,[28] colonization duration,[29] pathogenicity,[27,30,31] non-communicable disease risk,[32] antimicrobial resistance,[33–35] host adaptation, and transmission[36] 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] (Figure 1D). In contrast, the animal-associated isolates, 114 in total, were from clinical samples of diseased animals sent from four New England states (New Hampshire, Maine, Massachusetts, and Vermont) to the New Hampshire Veterinary Diagnostic Laboratory, United States, from 2017 to 2020.[38] 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,[4–8] 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%, $p = 9.427 \times 10^{-5}$) 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[39–41] (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%, $p = 1.066 \times 10^{-7}$).[42] However, CC1, a known livestock-associated clone,[6] 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.[43,44] 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

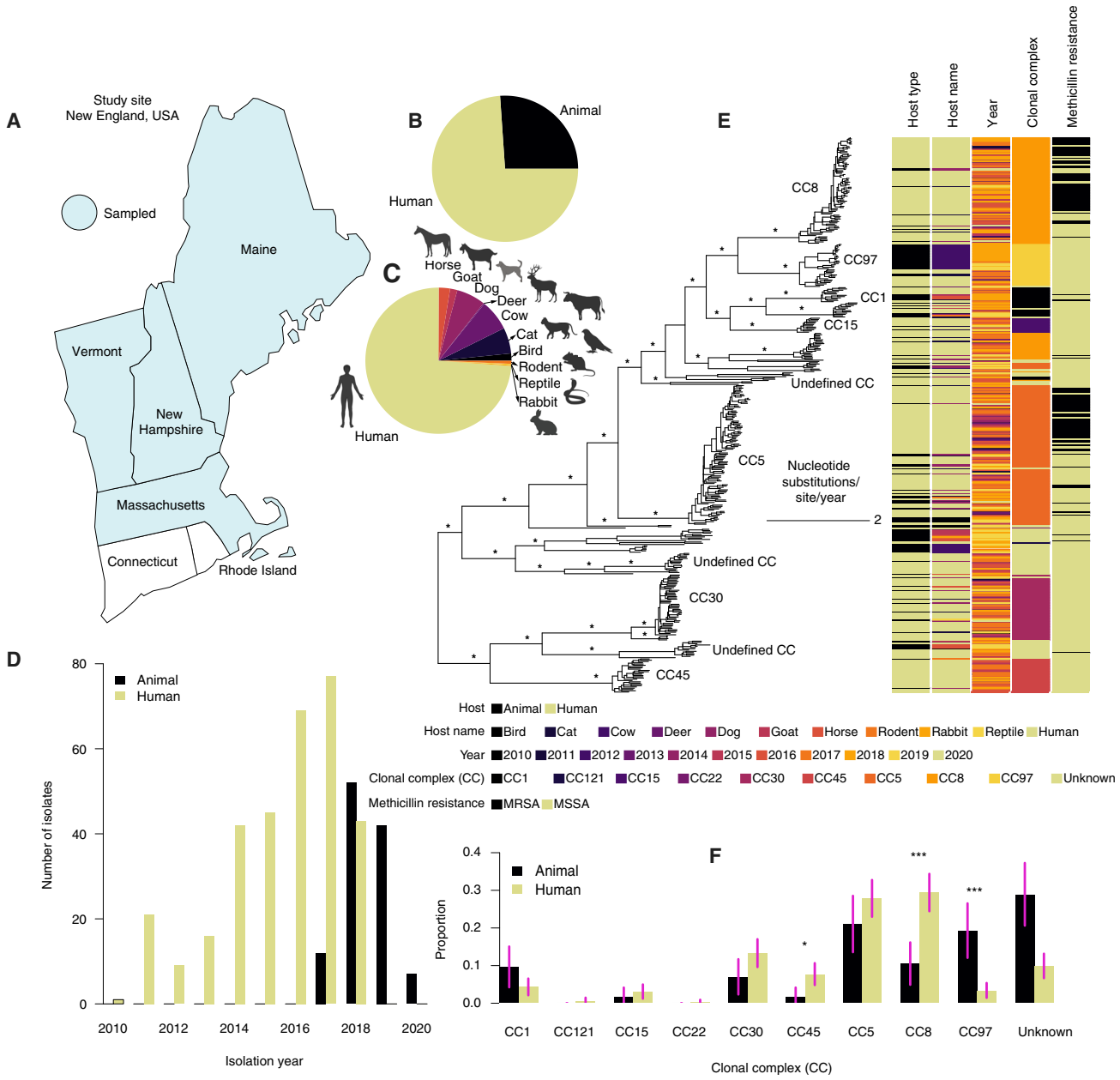


Figure 1. *S. aureus* isolates sampled from human and animal hosts are genetically diverse, intermixed in the phylogeny, and reveal host-adapted clones

(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.

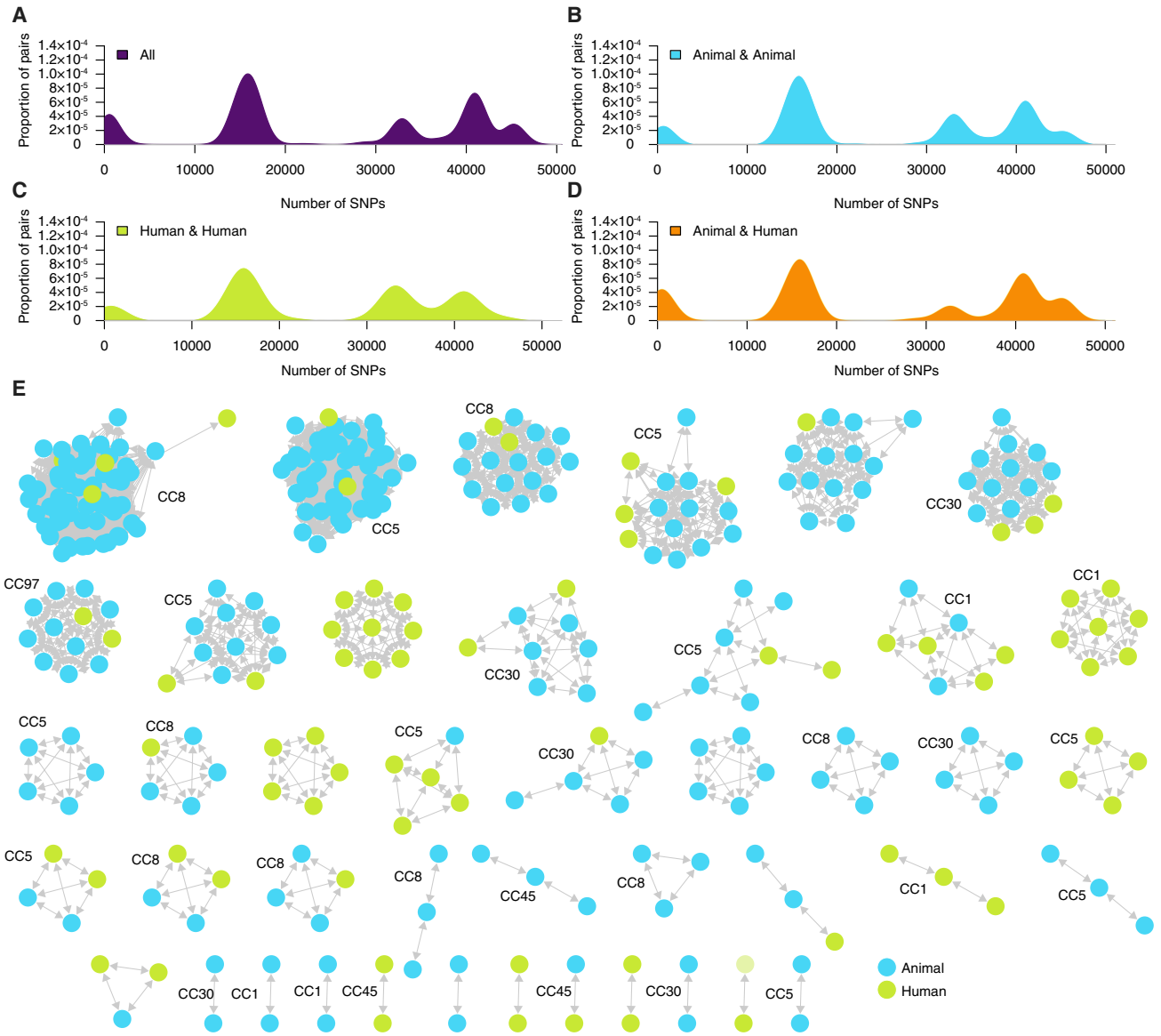


Figure 2. Comparative genomics of *S. aureus* reveals high genetic diversity and transmission between human and animal hosts

(A) Histogram showing the multi-modal distribution of the SNP distance between pairs of all the *S. aureus* isolates in the study.

(B) Histogram showing the multi-modal distribution of the SNP distance between pairs of animal *S. aureus* isolates.

(C) Histogram showing the multi-modal distribution of the SNP distance between pairs of human *S. aureus* isolates.

(D) Histogram showing the multi-modal distribution of the SNP distance between pairs of animal and human *S. aureus* isolates.

(E) Cytoscape network showing connected components or clusters containing both human and animal isolates. The edges in the network represent pairs of *S. aureus* isolates differing by <150 SNPs, which captures recent and non-recent transmissions between animal and human hosts occurring within ~13 years.[45]

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) ($p = 5.435 \times 10^{-14}$), respectively. This suggested that *S. aureus* host-switching occurred more frequently from animal to humans than from

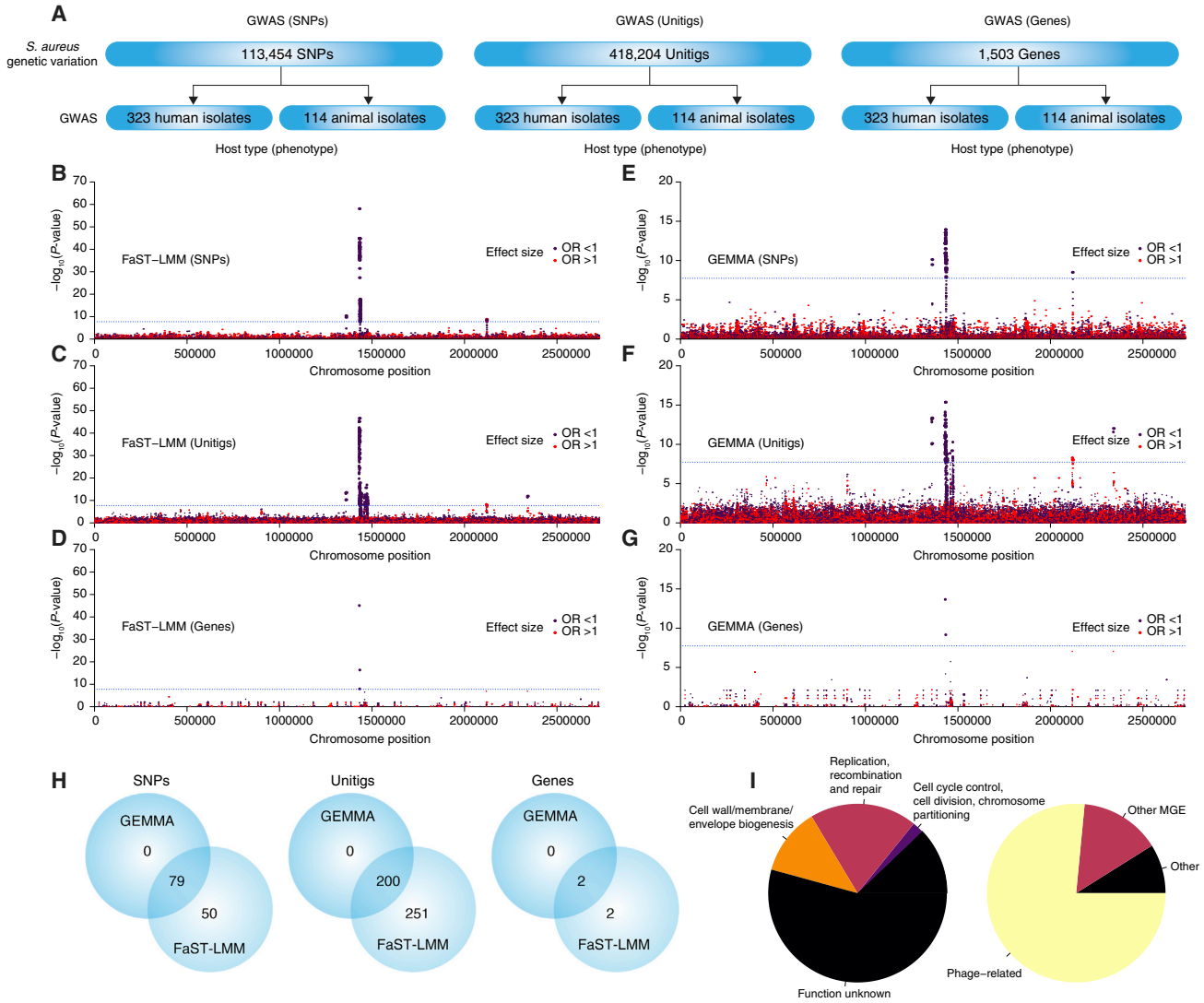


Figure 3. GWAS of human and animal *S. aureus* reveals key independently evolving genetic loci for host transmissibility, infection, switching, and adaptation

(A) Schematic diagram showing the study design, phenotype tested, genetic variation, and tools used for the GWAS.

(B) Manhattan plots for the GWAS of SNPs using FaST-LMM show the log-transformed statistical significance and genomic coordinates based on the human *S. aureus* strain JP080 (GenBank: AP017922.1). The odds ratio in the Manhattan plots is for the minor allele.

(C) Manhattan plots for the GWAS of unitigs using FaST-LMM showing the log-transformed statistical significance and *S. aureus* genome coordinates.

(D) Manhattan plots for the GWAS of genes using FaST-LMM showing the log-transformed statistical significance and *S. aureus* genome coordinates.

(E) Manhattan plots for the GWAS of SNPs using a different GWAS method, GEMMA, showing the log-transformed statistical significance and *S. aureus* genome coordinates.

(F) Manhattan plots for the GWAS of SNPs using GEMMA showing the log-transformed statistical significance and *S. aureus* genome coordinates.

(G) Manhattan plots for the GWAS of SNPs using GEMMA showing the log-transformed statistical significance and *S. aureus* genome coordinates.

(H) Venn diagrams for the SNPs, unitigs, and accessory genes identified by GEMMA and FaST-LMM.

(I) Functional analysis of the genes containing genetic variation statistically associated with host type. Additional information is provided in Figures S2–S6.

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.[13] 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)[47](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 (*hlb*) gene[48,49] (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, amidase, 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 IS*1272* 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 genic 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; $p = 2.25 \times 10^{-39}$), *sak* (odds ratio, 0.750; $p = 1.19 \times 10^{-10}$), and two hypothetical genes: an amidase (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] and

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

Locus tag	Gene name	Reference genome	Total unitigs	*Q* value range[a]	Odds ratio range	Gene product
JP02758_1300	*scn*	AP017922.1	23	2.25×10^{-39} to 2.34×10^{-21}	0.707–0.746	complement inhibitor SCIN
JP02758_1303	*sak*	AP017922.1	30	3.84×10^{-12} to 0.015	0.745–0.834	staphylokinase
JP02758_1973	*tnp*	AP017922.1	5	0.0191–0.0359	1.352	IS*1272* transposase
JP02758_1367	*int*	AP017922.1	57	3.26×10^{-11} to 0.04864	0.742–0.809	integrase
Intergenic		AP017922.1	110	6.56×10^{-41} to 0.0244	0.705–0.840	hypothetical protein
No match		No match	129	1.24×10^{-35} to 0.0484	0.714–0.813	no match
C2G36_RS11395		NZ_CP030138.1	1	4.27×10^{-05} to 4.27×10^{-5}	0.835–0.835	AAA family ATPase
C7M54_RS09660		NZ_CP029685.1	1	0.0051–0.0051	1.448–1.448	IS*1182* family transposase
CGP86_RS04080		NZ_CP022720.1	1	0.0002	1.386–1.386	IS*1182* family transposase
CPC18_RS10665		NZ_CP023561.1	5	8.64×10^{-06} to 0.0439	1.462–1.348	IS*1182* family transposase
CPC18_RS13120		NZ_CP023561.1	4	0.0002–0.0003	1.399–1.390	IS*1182* family transposase
CU118_RS06940		NZ_CP024998.1	1	3.33×10^{-9}	0.799–0.799	AAA family ATPase
E3S65_RS00020		NZ_CP047859.1	2	0.0007–0.0014	0.798–0.803	Clp protease ClpP
E3S65_RS00025		NZ_CP047859.1	2	0.0002–0.0010	0.780–0.791	phage major capsid protein
E3S65_RS14240		NZ_CP047859.1	4	2.62×10^{-08} to 3.81×10^{-6}	0.809–0.823	AAA family ATPase
F6Y18_RS01700		NZ_AP020316.1	1	0.0457–0.0457	1.292–1.292	IS*1182* family transposase
FP479_RS08875		NZ_CP042008.1	1	0.0041–0.0042	1.351–1.351	IS*1182* family transposase
I3K83_RS01375		NZ_CP065199.1	2	2.69×10^{-06} to 0.0489	1.546–1.342	IS*1182* family transposase
I3K83_RS04555		NZ_CP065199.1	1	8.27×10^{-07} to 8.27×10^{-7}	0.667–0.667	transposase
I6J76_RS02915		NZ_CP069470.1	2	0.0002–0.0002	1.399–1.399	IS*1182* family transposase
I6J76_RS05230		NZ_CP069470.1	1	0.043886364–0.043886364	1.348–1.348	IS*1182* family transposase
ILP77_RS09555		NZ_CP062467.1	1	0.017527775–0.017527775	1.288–1.288	IS*1182* family transposase
JP02758_1299		AP017922.1	12	1.24×10^{-35} to 1.15×10^{-16}	0.714–0.761	phage protein
JP02758_1302		AP017922.1	8	3.41×10^{-11} to 0.0022	0.763–0.804	amidase
JP02758_1304		AP017922.1	3	2.01×10^{-06} to 0.032490283	0.794–0.794	lytic enzyme
JP02758_1310		AP017922.1	29	6.02×10^{-08} to 0.0237	0.748–0.805	putative protein SA1764
JP02758_1330		AP017922.1	2	2.7×10^{-05} to 0.010	0.769–0.806	phage protein
JP02758_1340		AP017922.1	6	9.64×10^{-07} to 0.0471	0.817–0.849	hypothetical protein
JP02758_1355		AP017922.1	1	5.91×10^{-10} to 5.91×10^{-10}	0.758–0.758	phage protein
JP02758_2192		AP017922.1	2	3.32×10^{-06} to 9.86×10^{-6}	0.649–0.675	transposase
K9B11_RS08960		NZ_CP083728.1	2	0.0005–0.0005	1.424–1.424	IS*1182* family transposase
LO764_RS13470		NZ_CP087593.1	1	0.0301–0.0301	0.899–0.899	E domain-containing protein
RK77_RS01455		NZ_CP026064.1	1	1.79×10^{-06} to 1.79×10^{-6}	0.673–0.673	transposase

[a]*Q* value represents the adjusted p value based on the Bonferroni correction (see STAR Methods).

pathogenicity in *Staphylococcus epidermidis*.[30] 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 *S. aureus* 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

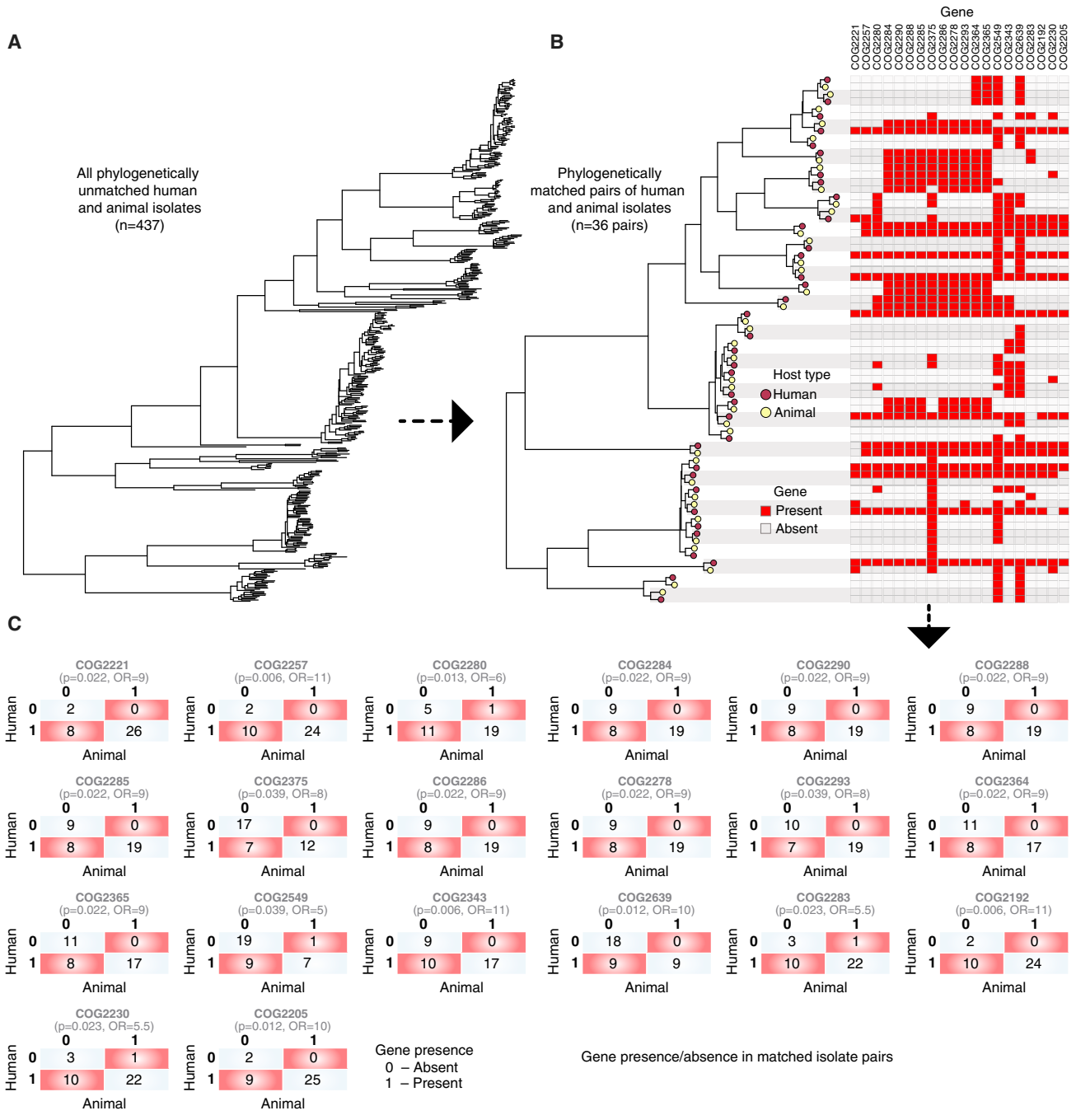


Figure 4. A phylogeny-based sampling of matched human and animal isolates confirms genes implicated in the GWAS of unmatched isolates and reveals improved power to discover hidden genes associated with host transmissibility, infection, switching, and adaptation

(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.

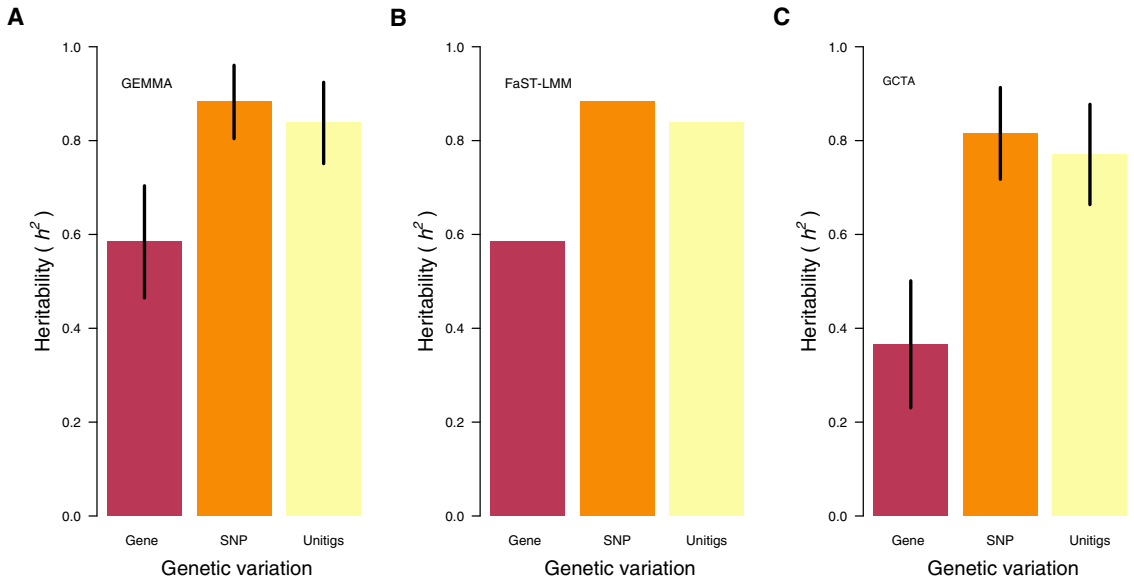


Figure 5. High narrow-sense heritability highlights the remarkable contribution of *S. aureus* genetics to host transmissibility, infection, switching, and adaptation. Estimates of the narrow-sense heritability (h^2) using different genetic variants and methods

(A) Heritability was estimated for accessory genes, SNPs, and unitigs using GEMMA.

(B) Heritability was estimated for accessory genes, SNPs, and unitigs using FaST-LMM.

(C) Heritability was estimated for accessory genes, SNPs, and unitigs using GCTA. The error bars in the graph represent 95% CIs. FaST-LMM reported no standard errors for the heritability estimates; therefore, the plots show no CIs. Heritability is expressed as a proportion with values ranging from 0 to 1.

matched sampling scheme in identifying cryptic genetic variation associated with *S. aureus* 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),[57] FaST-LMM,[57] and GEMMA.[55] 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 based on GEMMA (h^2 = 0.88; 95% CI, 0.80–0.96), FaST-LMM (h^2 = 0.88), CGTA (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.[58] Such high heritability is similar to the loci associated with AMR phenotypes in several bacterial pathogens, including *M. tuberculosis*[34] and *Neisseria gonorrhoeae*.[59] 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 φSa3 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 *scn* (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 φSa3 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.[27] 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

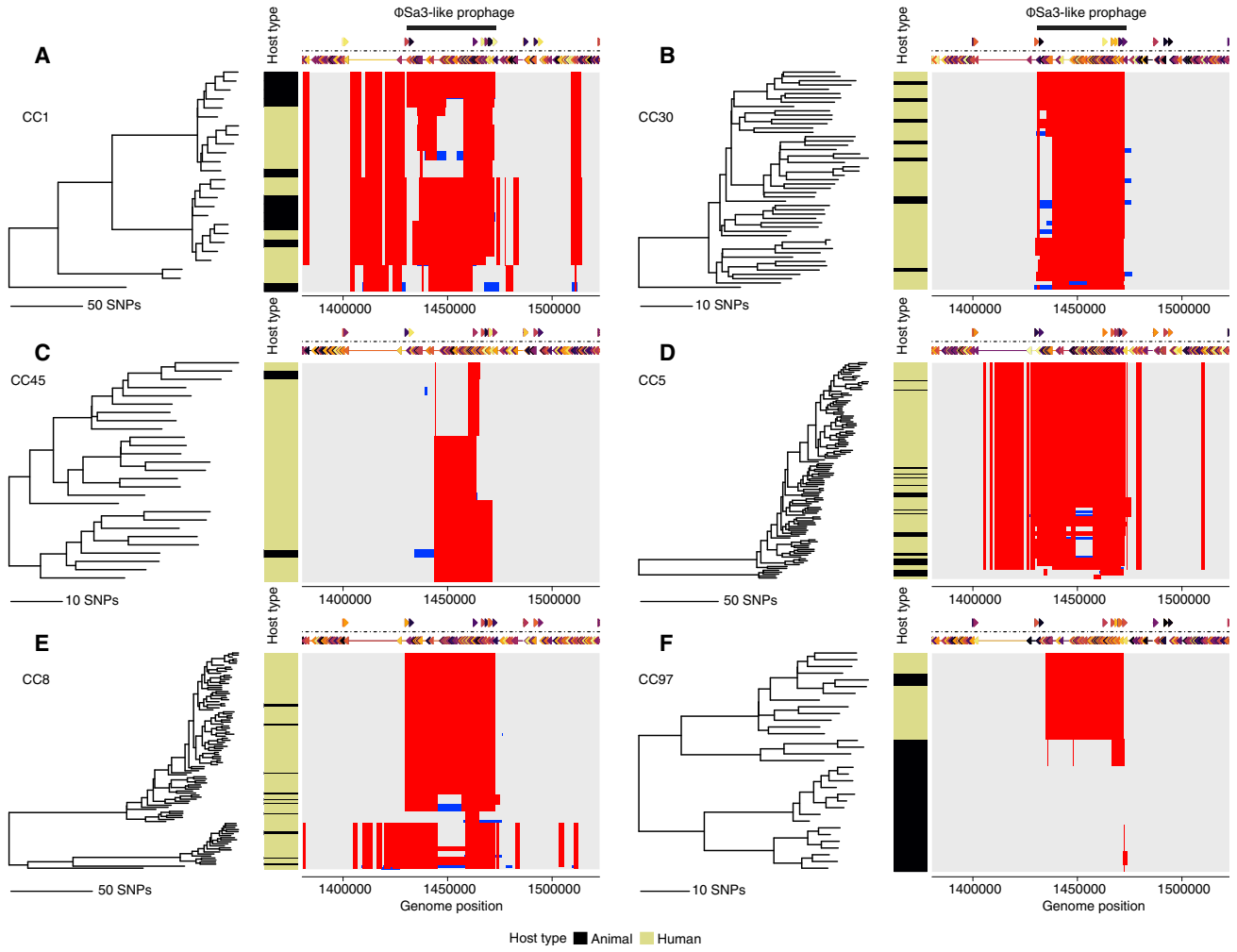


Figure 6. The φSa3 prophage encoding genetic variation associated with *S. aureus* host transmissibility, infection, switching, and adaptation is the major hotspot for genetic exchange via recombination

(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 φSa3 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.

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 φSa3 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 through recombination and HGT. 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

events.[63] 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[64–66] (Figures 6 and S7). Altogether, these findings indicate that recombination and HGT drive the rapid acquisition and loss of the genes, thereby modulating rapid host transmissibility, infection, switching, and adaptation of *S. aureus* over short and long timescales.

S. aureus φSa3 prophages are ubiquitous and genetically diverse

We next investigated the genetic diversity of the φSa3 prophage containing the implicated immune evasion genes. First, we mapped genomic data of each isolate against a reference φSa3 sequence 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 φSa3 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 φSa3-like and other prophage sequences, including φSa1, φSa3, φSa5, φSa6, φSa7, and φSa9, and especially those lacking the immune evasion genes.[51] 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 φSa3 prophage sequence, next we performed a complementary analysis using *de novo* assemblies. We extracted full φSa3 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 φSa3-like prophages are more common in human than animal *S. aureus* isolates and exhibit substantial genetic diversity, consistent with findings elsewhere.[51,64,66]

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 φSa3 β-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 lysin or amidase and an AAA ATPases family protein).[48,50] While other genetic loci located outside the φSa3 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 ∼88% of the heritability in the infection of human and animal host types, of which ∼99.9% was attributed to the immune evasion genes within the φSa3 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.[24]

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,[48] specifically the SCIN (*scn*)[67] and SAK (*sak*) genes[50] encoded by the φSa3 prophage in *S. aureus* host-switching and adaptation. These results are consistent with previous pan-genome studies.[4,17–19,22,68] The *scn* gene encodes an efficient complement inhibitor for the classical, alternative, and lectin complement pathways, which prevents opsonophagocytosis, neutrophil-mediated killing and neutrophil chemotaxis, and generation of complement component 5a (C5a).[49,69] The *sak* gene interferes with human innate immune defenses by exhibiting anti-opsonic activity.[70] 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.[70] The ability of these MGEs to rapidly spread from strain to strain provides an opportunity for rapid adaptation of *S. aureus* over short timescales to survive in a host. Surprisingly, we did not identify any statistically significant signals for host transmissibility, infection, and adaptation in the AMR-associated elements and genes, such as the SCC*mec*,[71] which have been widely linked with the successful clonal expansion of several *S. aureus* clones, including CC398,[19] although they likely play a major role in transmission in the human population.[72]

A key aspect of our study is the ability to control for the population structure when correlating the genetic variation and host type, which typically leads to spurious associations, especially for bacterial data. Our analysis of phylogenetically matched pairs of human and animal isolates, which likely represented the recent or non-recent zoonotic transmission, supported the findings from the GWAS of unmatched isolates that the prophage-encoded innate immune evasion genes are overrepresented in the isolates collected from humans and animals. This sampling approach removes residual confounding effects due to the cryptic clonal population structure of the isolates, which ultimately improves the power to detect real genetic signals. This approach has been successfully demonstrated to discover genetic factors for AMR of *M. tuberculosis* and adaptation of *Campylobacter* species to different livestock hosts,[56] and

pathogenicity of *S. epidermidis*.[30] 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.[73] Although the animal-associated *S. aureus* isolates are not entirely devoid of the φSa3-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, φSa3-containing isolates in animals and φSa3-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.[36,74] Thus, our findings indicate that the acquisition of the φSa3 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.[13] 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*,[4,13,17–19,23,69,74] the overall contribution of the pathogen genetics remains less understood. Our findings suggest remarkably high narrow-sense heritability, reaching ∼88% based on SNP and unitig 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 ∼84% and ∼97% in *M. tuberculosis* and *N. gonorrhoeae* respectively.[34,59] Similarly, the establishment of *S. aureus* pyomyositis infection has also shown a high heritability of ∼64%.[27] 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 φSa3 prophage contributed ∼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 φSa3 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[75] 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)[30] and *S. aureus* (101 cases and 417 controls; unmatched),[27] 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)[56] 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.

Received: May 2, 2022
Revised: June 13, 2022
Accepted: September 14, 2022
Published: October 4, 2022

REFERENCES

1. Smith, T.C. (2015). Livestock-associated Staphylococcus aureus: the United States experience. PLoS Pathog. *11*, e1004564.
2. Wertheim, H.F.L., Vos, M.C., Ott, A., van Belkum, A., Voss, A., Kluytmans, J.A.J.W., van Keulen, P.H.J., Vandenbroucke-Grauls, C.M.J.E., Meester, M.H.M., and Verbrugh, H.A. (2004). Risk and outcome of nosocomial Staphylococcus aureus bacteraemia in nasal carriers versus non-carriers. Lancet *364*, 703–705.
3. Matuszewska, M., Murray, G.G.R., Harrison, E.M., Holmes, M.A., and Weinert, L.A. (2020). The evolutionary genomics of host specificity in Staphylococcus aureus. Trends Microbiol. *28*, 465–477.
4. Price, L.B., Stegger, M., Hasman, H., Aziz, M., Larsen, J., Andersen, P.S., Pearson, T., Waters, A.E., Foster, J.T., Schupp, J., et al. (2012). Staphylococcus aureus CC398: host adaptation and emergence of methicillin resistance in livestock. mBio *3*, e00305–e00311.
5. Feltrin, F., Alba, P., Kraushaar, B., Ianzano, A., Argudín, M.A., Di Matteo, P., Porrero, M.C., Aarestrup, F.M., Butaye, P., Franco, A., and Battisti, A. (2016). A livestock-associated, multidrug-resistant, methicillin-resistant Staphylococcus aureus clonal complex 97 lineage spreading in dairy cattle and pigs in Italy. Appl. Environ. Microbiol. *82*, 816–821.
6. Alba, P., Feltrin, F., Cordaro, G., Porrero, M.C., Kraushaar, B., Argudín, M.A., Nykäsenoja, S., Monaco, M., Stegger, M., Aarestrup, F.M., et al. (2015). Livestock-associated methicillin resistant and methicillin susceptible Staphylococcus aureus sequence type (CC)1 in European farmed animals: high genetic relatedness of isolates from Italian cattle herds and humans. PLoS One *10*, e0137143.
7. Moodley, A., Espinosa-Gongora, C., Nielsen, S.S., McCarthy, A.J., Lindsay, J.A., and Guardabassi, L. (2012). Comparative host specificity of human- and pig- associated Staphylococcus aureus clonal lineages. PLoS One *7*, e49344.
8. Chen, H., Yin, Y., van Dorp, L., Shaw, L.P., Gao, H., Acman, M., Yuan, J., Chen, F., Sun, S., Wang, X., et al. (2021). Drivers of methicillin-resistant Staphylococcus aureus (MRSA) lineage replacement in China. Genome Med. *13*, 171.
9. Koop, G., Vrieling, M., Storisteanu, D.M.L., Lok, L.S.C., Monie, T., van Wigcheren, G., Raisen, C., Ba, X., Gleadall, N., Hadjirin, N., et al. (2017). Identification of LukPQ, a novel, equid-adapted leukocidin of Staphylococcus aureus. Sci. Rep. *7*, 40660.
10. Viana, D., Blanco, J., Tormo-Más, M.A., Selva, L., Guinane, C.M., Baselga, R., Corpa, J.M., Lasa, I., Novick, R.P., Fitzgerald, J.R., and Penadés, J.R. (2010). Adaptation of Staphylococcus aureus to ruminant and equine hosts involves SaPI-carried variants of von Willebrand factor-binding protein. Mol. Microbiol. *77*, 1583–1594.
11. Viana, D., Comos, M., McAdam, P.R., Ward, M.J., Selva, L., Guinane, C.M., González-Muñoz, B.M., Tristan, A., Foster, S.J., Fitzgerald, J.R., and Penadés, J.R. (2015). A single natural nucleotide mutation alters bacterial pathogen host tropism. Nat. Genet. *47*, 361–366.
12. Lowder, B.V., Guinane, C.M., Ben Zakour, N.L., Weinert, L.A., Conway-Morris, A., Cartwright, R.A., Simpson, A.J., Rambaut, A., Nübel, U., and Fitzgerald, J.R. (2009). Recent human-to-poultry host jump, adaptation, and pandemic spread of Staphylococcus aureus. Proc. Natl. Acad. Sci. USA *106*, 19545–19550.
13. Richardson, E.J., Bacigalupe, R., Harrison, E.M., Weinert, L.A., Lycett, S., Vrieling, M., Robb, K., Hoskisson, P.A., Holden, M.T.G., Feil, E.J., et al. (2018). Gene exchange drives the ecological success of a multi-host bacterial pathogen. Nat. Ecol. Evol. *2*, 1468–1478.
14. Hammarlöf, D.L., et al. (2018). Role of a single noncoding nucleotide in the evolution of an epidemic African clade of. Proc. Natl. Acad. Sci. USA *115*, E2614–E2623.
15. Lecuit, M., Dramsi, S., Gottardi, C., Fedor-Chaiken, M., Gumbiner, B., and Cossart, P. (1999). A single amino acid in E-cadherin responsible for host specificity towards the human pathogen Listeria monocytogenes. EMBO J. *18*, 3956–3963.
16. Tikhomirova, A., Trappetti, C., Paton, J.C., Watson-Haigh, N., Wabnitz, D., Jervis-Bardy, J., Jardeleza, C., and Kidd, S.P. (2021). A single nucleotide polymorphism in an IgA1 protease gene determines Streptococcus pneumoniae adaptation to the middle ear during otitis media. Pathog. Dis. *79*, ftaa077.
17. Guinane, C.M., Ben Zakour, N.L., Tormo-Mas, M.A., Weinert, L.A., Lowder, B.V., Cartwright, R.A., Smyth, D.S., Smyth, C.J., Lindsay, J.A., Gould, K.A., et al. (2010). Evolutionary genomics of Staphylococcus aureus reveals insights into the origin and molecular basis of ruminant host adaptation. Genome Biol. Evol. *2*, 454–466.
18. Resch, G., François, P., Morisset, D., Stojanov, M., Bonetti, E.J., Schrenzel, J., Sakwinska, O., and Moreillon, P. (2013). Human-to-bovine jump of Staphylococcus aureus CC8 is associated with the loss of a β-hemolysin converting prophage and the acquisition of a new staphylococcal cassette chromosome. PLoS One *8*, e58187.
19. Spoor, L.E., McAdam, P.R., Weinert, L.A., Rambaut, A., Hasman, H., Aarestrup, F.M., Kearns, A.M., Larsen, A.R., Skov, R.L., and Fitzgerald, J.R.

(2013). Livestock origin for a human pandemic clone of community-associated methicillin-resistant Staphylococcus aureus. mBio *4*, 003566.

20. Moon, B.Y., Park, J.Y., Hwang, S.Y., Robinson, D.A., Thomas, J.C., Fitzgerald, J.R., Park, Y.H., and Seo, K.S. (2015). Phage-mediated horizontal transfer of a Staphylococcus aureus virulence-associated genomic island. Sci. Rep. *5*, 9784.
21. Larsen, J., Stegger, M., Andersen, P.S., Petersen, A., Larsen, A.R., Westh, H., Agersø, Y., Fetsch, A., Kraushaar, B., Käsbohrer, A., et al. (2016). Evidence for human adaptation and foodborne transmission of livestock-associated methicillin-resistant Staphylococcus aureus. Clin. Infect. Dis. *63*, 1349–1352.
22. Yu, F., Cienfuegos-Gallet, A.V., Cunningham, M.H., Jin, Y., Wang, B., Kreiswirth, B.N., and Chen, L. (2021). Molecular evolution and adaptation of livestock-associated methicillin-resistant Staphylococcus aureus (LA-MRSA) sequence type 9. mSystems *6*, e0049221.
23. Sieber, R.N., Urth, T.R., Petersen, A., Møller, C.H., Price, L.B., Skov, R.L., Larsen, A.R., Stegger, M., and Larsen, J. (2020). Phage-mediated immune evasion and transmission of livestock-associated methicillin-resistant Staphylococcus aureus in humans. Emerg. Infect. Dis. *26*.
24. Park, S., and Ronholm, J. (2021). Staphylococcus aureus in agriculture: lessons in evolution from a multispecies pathogen. Clin. Microbiol. Rev. *34*, 001822.
25. Dempsey, R.M., Carroll, D., Kong, H., Higgins, L., Keane, C.T., and Coleman, D.C. (2005). Sau42I, a BcgI-like restriction-modification system encoded by the Staphylococcus aureus quadruple-converting phage Phi42. Microbiology *151*, 1301–1311.
26. Laabei, M., Recker, M., Rudkin, J.K., Aldeljawi, M., Gulay, Z., Sloan, T.J., Williams, P., Endres, J.L., Bayles, K.W., Fey, P.D., et al. (2014). Predicting the virulence of MRSA from its genome sequence. Genome Res. *24*, 839–849.
27. Young, B.C., Earle, S.G., Soeng, S., Sar, P., Kumar, V., Hor, S., Sar, V., Bousfield, R., Sanderson, N.D., Barker, L., et al. (2019). Panton-Valentine leucocidin is the key determinant of pyomyositis in a bacterial GWAS. Elife *8*, e42486.
28. Wee, B.A., Alves, J., Lindsay, D.S.J., Klatt, A.B., Sargison, F.A., Cameron, R.L., Pickering, A., Gorzynski, J., Corander, J., Marttinen, P., et al. (2021). Population analysis of Legionella pneumophila reveals a basis for resistance to complement-mediated killing. Nat. Commun. *12*, 7165.
29. Lees, J.A., Croucher, N.J., Goldblatt, D., Nosten, F., Parkhill, J., Turner, C., Turner, P., and Bentley, S.D. (2017). Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. Elife *6*, e26255.
30. Méric, G., Mageiros, L., Pensar, J., Laabei, M., Yahara, K., Pascoe, B., Kittiwan, N., Tadee, P., Post, V., Lamble, S., et al. (2018). Disease-associated genotypes of the commensal skin bacterium Staphylococcus epidermidis. Nat. Commun. *9*, 5034.
31. Earle, S.G., Lobanovska, M., Lavender, H., Tang, C., Exley, R.M., Ramos-Sevillano, E., Browning, D.F., Kostiou, V., Harrison, O.B., Bratcher, H.B., et al. (2021). Genome-wide association studies reveal the role of polymorphisms affecting factor H binding protein expression in host invasion by Neisseria meningitidis. PLoS Pathog. *17*, e1009992.
32. Berthenet, E., Yahara, K., Thorell, K., Pascoe, B., Meric, G., Mikhail, J.M., Engstrand, L., Enroth, H., Burette, A., Megraud, F., et al. (2018). A GWAS on Helicobacter pylori strains points to genetic variants associated with gastric cancer risk. BMC Biol. *16*, 84.
33. Chewapreecha, C., Marttinen, P., Croucher, N.J., Salter, S.J., Harris, S.R., Mather, A.E., Hanage, W.P., Goldblatt, D., Nosten, F.H., Turner, C., et al. (2014). Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. PLoS Genet. *10*, e1004547.
34. Farhat, M.R., Freschi, L., Calderon, R., Ioerger, T., Snyder, M., Meehan, C.J., de Jong, B., Rigouts, L., Sloutsky, A., Kaur, D., et al. (2019). GWAS for quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions. Nat. Commun. *10*, 2128.
35. Coll, F., Phelan, J., Hill-Cawthorne, G.A., Nair, M.B., Mallard, K., Ali, S., Abdallah, A.M., Alghamdi, S., Alsomali, M., Ahmed, A.O., et al. (2018). Author Correction: genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. Nat. Genet. *50*, 764.
36. Sieber, R.N., Larsen, A.R., Urth, T.R., Iversen, S., Møller, C.H., Skov, R.L., Larsen, J., and Stegger, M. (2019). Genome investigations show host adaptation and transmission of LA-MRSA CC398 from pigs into Danish healthcare institutions. Sci. Rep. *9*, 18655.
37. Smith, J.T., Eckhardt, E.M., Hansel, N.B., Eliato, T.R., Martin, I.W., and Andam, C.P. (2021). Genomic epidemiology of methicillin-resistant and -susceptible Staphylococcus aureus from bloodstream infections. BMC Infect. Dis. *21*, 589.
38. Bruce, S.A., Smith, J.T., Mydosh, J.L., Ball, J., Needle, D.B., Gibson, R., and Andam, C.P. (2022). Shared antibiotic resistance and virulence genes in Staphylococcus aureus from diverse animal hosts. Sci. Rep. *12*, 4413.
39. Blomfeldt, A., Aamot, H.V., Eskesen, A.N., Müller, F., and Monecke, S. (2013). Molecular characterization of methicillin-sensitive Staphylococcus aureus isolates from bacteremic patients in a Norwegian university hospital. J. Clin. Microbiol. *51*, 345–347.
40. Carrel, M., Perencevich, E.N., and David, M.Z. (2015). USA300 methicillin-resistant Staphylococcus aureus, United States, 2000-2013. Emerg. Infect. Dis. *21*, 1973–1980.
41. Manara, S., Pasolli, E., Dolce, D., Ravenni, N., Campana, S., Armanini, F., Asnicar, F., Mengoni, A., Galli, L., Montagnani, C., et al. (2018). Whole-genome epidemiology, characterisation, and phylogenetic reconstruction of Staphylococcus aureus strains in a paediatric hospital. Genome Med. *10*, 82.
42. Smyth, D.S., Feil, E.J., Meaney, W.J., Hartigan, P.J., Tollersrud, T., Fitzgerald, J.R., Enright, M.C., and Smyth, C.J. (2009). Molecular genetic typing reveals further insights into the diversity of animal-associated Staphylococcus aureus. J. Med. Microbiol. *58*, 1343–1353.
43. Gouliouris, T., Coll, F., Ludden, C., Blane, B., Raven, K.E., Naydenova, P., Crawley, C., Török, M.E., Enoch, D.A., Brown, N.M., et al. (2021). Quantifying acquisition and transmission of Enterococcus faecium using genomic surveillance. Nat. Microbiol. *6*, 103–111.
44. De Silva, D., Peters, J., Cole, K., Cole, M.J., Cresswell, F., Dean, G., Dave, J., Thomas, D.R., Foster, K., Waldram, A., et al. (2016). Whole-genome sequencing to determine transmission of Neisseria gonorrhoeae: an observational study. Lancet Infect. Dis. *16*, 1295–1303.
45. Coll, F., Raven, K.E., Knight, G.M., Blane, B., Harrison, E.M., Leek, D., Enoch, D.A., Brown, N.M., Parkhill, J., and Peacock, S.J. (2020). Definition of a genetic relatedness cutoff to exclude recent transmission of meticillin-resistant Staphylococcus aureus: a genomic epidemiology analysis. Lancet. Microbe *1*, e328–e335.
46. Pagel, M. (1999). Inferring the historical patterns of biological evolution. Nature *401*, 877–884.
47. Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. Nat. Methods *8*, 833–835.
48. van Wamel, W.J.B., Rooijakkers, S.H.M., Ruyken, M., van Kessel, K.P.M., and van Strijp, J.A.G. (2006). The innate immune modulators staphylococcal complement inhibitor and chemotaxis inhibitory protein of Staphylococcus aureus are located on β-hemolysin-converting bacteriophages. J. Bacteriol. *188*, 1310–1315.
49. Rooijakkers, S.H.M., Ruyken, M., Roos, A., Daha, M.R., Presanis, J.S., Sim, R.B., van Wamel, W.J.B., van Kessel, K.P.M., and van Strijp, J.A.G. (2005). Immune evasion by a staphylococcal complement inhibitor that acts on C3 convertases. Nat. Immunol. *6*, 920–927.
50. Jin, T., Bokarewa, M., Foster, T., Mitchell, J., Higgins, J., and Tarkowski, A. (2004). Staphylococcus aureus resists human defensins by production

of staphylokinase, a novel bacterial evasion mechanism. J. Immunol. *172*, 1169–1176.

51. Goerke, C., Pantucek, R., Holtfreter, S., Schulte, B., Zink, M., Grumann, D., Bröker, B.M., Doskar, J., and Wolz, C. (2009). Diversity of prophages in dominant Staphylococcus aureus clonal lineages. J. Bacteriol. *191*, 3462–3468.
52. Archer, G.L., Thanassi, J.A., Niemeyer, D.M., and Pucci, M.J. (1996). Characterization of IS1272, an insertion sequence-like element from Staphylococcus haemolyticus. Antimicrob. Agents Chemother. *40*, 924–929.
53. Jaillard, M., Lima, L., Tournoud, M., Mahé, P., van Belkum, A., Lacroix, V., and Jacob, L. (2018). A fast and agnostic method for bacterial genome-wide association studies: bridging the gap between k-mers and genetic events. PLoS Genet. *14*, e1007758.
54. Xia, G., and Wolz, C. (2014). Phages of Staphylococcus aureus and their impact on host evolution. Infect. Genet. Evol. *21*, 593–601.
55. Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. *44*, 821–824.
56. Farhat, M.R., Shapiro, B.J., Sheppard, S.K., Colijn, C., and Murray, M. (2014). A phylogeny-based sampling strategy and power calculator informs genome-wide associations study design for microbial pathogens. Genome Med. *6*, 101.
57. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
58. Kremer, P.H.C., Lees, J.A., Ferwerda, B., van de Ende, A., Brouwer, M.C., Bentley, S.D., and van de Beek, D. (2020). Genetic variation in Neisseria meningitidis does not influence disease severity in meningococcal meningitis. Front. Med. *7*.
59. Ma, K.C., Mortimer, T.D., Duckett, M.A., Hicks, A.L., Wheeler, N.E., Sánchez-Busó, L., and Grad, Y.H. (2020). Adaptation to the cervical environment is associated with increased antibiotic susceptibility in Neisseria gonorrhoeae. Nat. Commun. *11*, 5374.
60. Feil, E.J., Cooper, J.E., Grundmann, H., Robinson, D.A., Enright, M.C., Berendt, T., Peacock, S.J., Smith, J.M., Murphy, M., Spratt, B.G., et al. (2003). How clonal is Staphylococcus aureus. J. Bacteriol. *185*, 3307–3316.
61. Spoor, L.E., Richardson, E., Richards, A.C., Wilson, G.J., Mendonca, C., Gupta, R.K., McAdam, P.R., Nutbeam-Tuffs, S., Black, N.S., O'Gara, J.P., et al. (2015). Recombination-mediated remodelling of host–pathogen interactions during Staphylococcus aureus niche adaptation. Microb. Genom. *1*, e000036.
62. Young, B.C., Earle, S.G., Soeng, S., Sar, P., Kumar, V., Hor, S., Sar, V., Bousfield, R., Sanderson, N.D., Barker, L., and Stoesser, N. (2019). Author response: Panton–Valentine leucocidin is the key determinant of Staphylococcus aureus pyomyositis in a bacterial GWAS. eLife. https://doi.org/10.7554/elife.42486.019.
63. Croucher, N.J., Page, A.J., Connor, T.R., Delaney, A.J., Keane, J.A., Bentley, S.D., Parkhill, J., and Harris, S.R. (2015). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. *43*, e15.
64. van Alen, S., Ballhausen, B., Kaspar, U., Köck, R., and Becker, K. (2018). Prevalence and genomic structure of bacteriophage phi3 in human-derived livestock-associated methicillin-resistant Staphylococcus aureus isolates from 2000 to 2015. J. Clin. Microbiol. *56*, 001400.
65. Holden, M.T.G., Hsu, L.Y., Kurt, K., Weinert, L.A., Mather, A.E., Harris, S.R., Strommenger, B., Layer, F., Witte, W., de Lencastre, H., et al. (2013). A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant Staphylococcus aureus pandemic. Genome Res. *23*, 653–664.
66. Kashif, A., McClure, J.A., Lakhundi, S., Pham, M., Chen, S., Conly, J.M., and Zhang, K. (2019). Staphylococcus aureus ST398 virulence is associated with factors carried on prophage φSa3. Front. Microbiol. *10*, 2219.
67. Jongerius, I., Puister, M., Wu, J., Ruyken, M., van Strijp, J.A.G., and Rooijakkers, S.H.M. (2010). Staphylococcal complement inhibitor modulates phagocyte responses by dimerization of convertases. J. Immunol. *184*, 420–425.
68. Nepal, R., Houtak, G., Shaghayegh, G., Bouras, G., Shearwin, K., Psaltis, A.J., Wormald, P.J., and Vreugde, S. (2021). Prophages encoding human immune evasion cluster genes are enriched in Staphylococcus aureus isolated from chronic rhinosinusitis patients with nasal polyps. Microb. Genom. *7*.
69. Rooijakkers, S.H., van Kessel, K.P., and van Strijp, J.A. (2005). Staphylococcal innate immune evasion. Trends Microbiol. *13*, 596–601.
70. Rooijakkers, S.H.M., van Wamel, W.J.B., Ruyken, M., van Kessel, K.P.M., and van Strijp, J.A.G. (2005). Anti-opsonic properties of staphylokinase. Microb. Infect. *7*, 476–484.
71. Ito, T., Katayama, Y., Asada, K., Mori, N., Tsutsumimoto, K., Tiensasitorn, C., and Hiramatsu, K. (2001). Structural comparison of three types of staphylococcal cassette chromosome mec integrated in the chromosome in methicillin-resistant Staphylococcus aureus. Antimicrob. Agents Chemother. *45*, 1323–1336.
72. Loftus, R.W., Dexter, F., and Robinson, A.D.M. (2018). Methicillin-resistant Staphylococcus aureus has greater risk of transmission in the operating room than methicillin-sensitive S aureus. Am. J. Infect. Control *46*, 520–525.
73. Rooijakkers, S.H.M., Ruyken, M., van Roon, J., van Kessel, K.P.M., van Strijp, J.A.G., and van Wamel, W.J.B. (2006). Early expression of SCIN and CHIPS drives instant immune evasion by Staphylococcus aureus. Cell Microbiol. *8*, 1282–1293.
74. Matuszewska, M., Murray, G.G.R., Ba, X., Wood, R., Holmes, M.A., and Weinert, L.A. (2022). Stable antibiotic resistance and rapid human adaptation in livestock-associated MRSA. Elife *11*, e74819.
75. Zinsstag, J., Schelling, E., Waltner-Toews, D., and Tanner, M. (2011). From "one medicine" to "one health" and systemic approaches to health and well-being. Prev. Vet. Med. *101*, 148–156.
76. Enright, M.C., Day, N.P., Davies, C.E., Peacock, S.J., and Spratt, B.G. (2000). Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of Staphylococcus aureus. J. Clin. Microbiol. *38*, 1008–1015.
77. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. *19*, 455–477.
78. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics *27*, 578–579.
79. Nadalin, F., Vezzi, F., and Policriti, A. (2012). GapFiller: a de novo assembly approach to fill the gap within paired reads. BMC Bioinf. *13*, S8.
80. Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics *29*, 1072–1075.
81. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055.
82. Yu, L., Hisatsune, J., Hirakawa, H., Mizumachi, E., Toyoda, A., Yahara, K., and Sugai, M. (2017). Complete genome sequence of super biofilm-elaborating Staphylococcus aureus isolated in Japan. Genome Announc. *5*, 010433.
83. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069.
84. Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J.A., Gladstone, R.A., Lo, S., Beaudoin, C., Floto, R.A., et al. (2020). Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol. *21*, 180.

85. Page, A.J., Taylor, B., Delaney, A.J., Soares, J., Seemann, T., Keane, J.A., and Harris, S.R. (2016). SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb. Genom. *2*, e000056.

86. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. *37*, 1530–1534.

87. Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. Bioinformatics *20*, 289–290.

88. Chaguza, C., Tonkin-Hill, G., Lo, S.W., Hadfield, J., Croucher, N.J., Harris, S.R., and Bentley, S.D. (2021). RCandy: an R package for visualising homologous recombinations in bacterial genomes. Bioinformatics *38*, 1450–1451. https://doi.org/10.1093/bioinformatics/btab814.

89. Keck, F., Rimet, F., Bouchez, A., and Franc, A. (2016). phylosignal: an R package to measure, test, and explore the phylogenetic signal. Ecol. Evol. *6*, 2774–2780.

90. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.

91. Pennell, M.W., Eastman, J.M., Slater, G.J., Brown, J.W., Uyeda, J.C., FitzJohn, R.G., Alfaro, M.E., and Harmon, L.J. (2014). Geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. Bioinformatics *30*, 2216–2218.

92. Revell, L.J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. *3*, 217–223.

93. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

94. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

95. Holley, G., and Melsted, P. (2020). Bifrost: highly parallel construction and indexing of colored and compacted de Bruijn graphs. Genome Biol. *21*, 249.

96. D Turner, S. (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. J. Open Source Softw. *3*, 731. https://doi.org/10.21105/joss.00731.

97. Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M.J.L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423.

98. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

99. Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., von Mering, C., and Bork, P. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol. Biol. Evol. *34*, 2115–2122.

100. Arndt, D., Marcu, A., Liang, Y., and Wishart, D.S. (2019). PHAST, PHASTER and PHASTEST: tools for finding prophage in bacterial genomes. Briefings Bioinf. *20*, 1560–1567.

101. Overbeek, R., Olson, R., Pusch, G.D., Olsen, G.J., Davis, J.J., Disz, T., Edwards, R.A., Gerdes, S., Parrello, B., Shukla, M., et al. (2014). The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). Nucleic Acids Res. *42*, D206–D214.

102. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

103. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

104. Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.A., Barrell, B.G., and Parkhill, J. (2005). ACT: the artemis comparison tool. Bioinformatics *21*, 3422–3423.

105. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. *17*, 132.

STAR★METHODS

KEY RESOURCES TABLE

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

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
IQ-TREE	GitHub	https://github.com/Cibiv/IQ-TREE
Snp-dists	GitHub	https://github.com/tseemann/snp-dists
SPAdes	GitHub	https://github.com/ablab/spades
SSPACE	GitHub	https://github.com/nsoranzo/sspace_basic
GapFiller	Sourceforge	https://sourceforge.net/projects/gapfiller/
QUAST	GitHub	https://github.com/ablab/quast
CheckM	GitHub	https://github.com/Ecogenomics/CheckM
exact2x2	CRAN	https://cran.r-project.org/web/packages/exact2x2/
eggNOG-mapper	EMBL	http://eggnog-mapper.embl.de/
in_silico_PCR.pl	GitHub	https://github.com/egonozer/in_silico_pcr
MASH	GitHub	https://github.com/marbl/Mash
pheatmap	CRAN	https://CRAN.R-project.org/package=pheatmap
Other		
Scripts and GWAS datasets	GitHub, Zenodo	Zenodo (https://doi.org/10.5281/zenodo.6944550), GitHub (https://github.com/ChrispinChaguza/SAUREUS_NE_USA)

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,[37] while 114 isolates were sampled from animals.[38] 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*.[76] 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 ×150 per base coverage, trimming adapters, correcting sequencing errors, and assembly using SPAdes (version 3.13.0).[77] We then used SSPACE (version 3.0)[78] and GapFiller (version 1.10)[79] to scaffold and close gaps to improve the generated assemblies. Genome quality was assessed using the programs QUAST (version 5.0.2)[80] and CheckM (version 1.1.3)[81] 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).[82] 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).[83] 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).[84]

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)[82] 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).[85] 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)[86] with the general time-reversible (GTR) and Gamma models and 100 bootstraps. We visualised the phylogenetic trees using the APE package (version 4.3)[87] and RCandy (version 1.0.0).[88] We annotated the trees with isolate metadata using the "gridplot" and "phylo4d" functions available in the phylosignal (version 1.3)[89] 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)[84] 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).[90] We quantified the correlation between the phylogenetic tree and the host type phenotype, i.e., phylogenetic signal, using Pagel's λ statistic[46] using the

"fitDiscrete" functions in the Geiger (version 2.0.6.4).[91] 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).[92] We used Gubbins to detect recombination events in whole-genome alignments of different isolates from different clonal complexes with at least 25 sequenced genomes.[63]

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 0.1.16).[93] We excluded rare SNPs with minor allele frequency <5% and missingness >5% using PLINK (version 1.90b4).[94] 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).[84] 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).[95] 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)[47] for the GWAS followed by validation using GEMMA (version 0.98.1).[55] 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,[34] 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 $\alpha = 0.05$ 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 qqman (version 0.1.7).[96] 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),[47] GEMMA (version 0.98.1),[55] and GCTA (version 1.93.2).[57] 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 φSa3 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.[27] 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).[97] In addition, we used BLASTN (version 2.5.0+)[98] to identify genomic regions containing the gene and unitig sequences. Functional annotation of the genes identified in the GWAS was done using EggNOG-mapper.[99] 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) (https://www.R-project.org/) 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/).[100] We annotated the identified reference prophage sequence using the RAST server.[101] 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).[102] We extracted the genomic coordinates for the mapped sequences using the ''bamtobed'' option in BEDTools (version 2.30.0)[103] 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+)[98] to compare the prophage sequence against the genome sequence of the *S. aureus* isolates and visualised the results using ACT (version 18.1.0).[104] 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).[105] 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).