ARTICLES
https://doi.org/10.1038/s41559-018-0617-0
nature ecology & evolution

Gene exchange drives the ecological success of a multi-host bacterial pathogen

Emily J. Richardson[1,18,19], Rodrigo Bacigalupe[1,19], Ewan M. Harrison[2,19], Lucy A. Weinert[3,19], Samantha Lycett[1], Manouk Vrieling[1], Kirsty Robb[4], Paul A. Hoskisson[4], Matthew T. G. Holden[5], Edward J. Feil[6], Gavin K. Paterson[7], Steven Y. C. Tong[8,9], Adebayo Shittu[10], Willem van Wamel[11], David M. Aanensen[12,13], Julian Parkhill[14], Sharon J. Peacock[15], Jukka Corander[14,16,17], Mark Holmes[3] and J. Ross Fitzgerald[1*]

The capacity for some pathogens to jump into different host-species populations is a major threat to public health and food security. *Staphylococcus aureus* is a multi-host bacterial pathogen responsible for important human and livestock diseases. Here, using a population-genomic approach, we identify humans as a major hub for ancient and recent *S. aureus* host-switching events linked to the emergence of endemic livestock strains, and cows as the main animal reservoir for the emergence of human epidemic clones. Such host-species transitions are associated with horizontal acquisition of genetic elements from host-specific gene pools conferring traits required for survival in the new host-niche. Importantly, genes associated with antimicrobial resistance are unevenly distributed among human and animal hosts, reflecting distinct antibiotic usage practices in medicine and agriculture. In addition to gene acquisition, genetic diversification has occurred in pathways associated with nutrient acquisition, implying metabolic remodelling after a host switch in response to distinct nutrient availability. For example, *S. aureus* from dairy cattle exhibit enhanced utilization of lactose—a major source of carbohydrate in bovine milk. Overall, our findings highlight the influence of human activities on the multi-host ecology of a major bacterial pathogen, underpinned by horizontal gene transfer and core genome diversification.

Many bacterial pathogens are host specialists that co-evolve with a single host species. However, the capacity to switch host species can provide opportunities for expansion into new host populations. The domestication of animals in the Neolithic period (approximately 10,000–2000 BC) and the more recent intensification of livestock farming provided increased opportunities for the transfer of bacterial pathogens between humans and animals[1]. Of note, the majority of emerging human infectious diseases have been traced to an animal origin[2]. *Staphylococcus aureus* is associated with a wide spectrum of diseases in humans, and strains of both methicillin-sensitive and methicillin-resistant *S. aureus* are common causes of nosocomial and community-acquired infection[3,4]. In addition, *S. aureus* causes an array of infections of livestock that are a major burden on the agricultural industry, including mastitis in cows, sheep and goats[5,6], septicaemia and skeletal infections in commercial broiler chickens[7], exudative epidermitis in pigs[8], and skin abscesses and mastitis in rabbits[9].

S. aureus has a clonal population structure defined by a relatively low level of recombination, comprising lineages that have single or multiple host tropisms[10–12]. Inter-host-species transmission can be of critical public health importance, as exemplified by the livestock-associated methicillin-resistant clonal complex CC398, which is associated with pigs and other livestock, but can cause zoonotic infections of pig farmers and their contacts[13,14]. Previous work employed multi-locus sequence typing (MLST) to provide evidence for the occurrence of host-jump events from humans leading to the emergence of *S. aureus* clones in livestock populations[11,12]. More recently, whole-genome sequencing has been employed to investigate the evolution of individual clones, providing insights into the emergence, transmission and acquisition of antibiotic resistance in hospital, community and agricultural settings[13,15–17]. In addition, a role for specific mobile genetic elements (MGEs) and core gene mutations in the host adaptation of *S. aureus* has been identified[9,18,19]. For example, the major porcine and avian clones of *S. aureus* probably originated in humans, and the host jumps were associated with the acquisition of MGEs not found among human isolates[13,18]. Similarly, the major *S. aureus* clone associated with sheep and goats evolved through a combination of gene acquisition and allelic diversification, including loss of gene function[20]. Furthermore, several studies have reported the

[1]The Roslin Institute, Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, UK. [2]Department of Medicine, University of Cambridge, Cambridge, UK. [3]Department of Veterinary Medicine, University of Cambridge, Cambridge, UK. [4]University of Strathclyde, Glasgow, UK. [5]School of Medicine, University of St Andrews, St Andrews, UK. [6]Milner Centre for Evolution, University of Bath, Bath, UK. [7]Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, UK. [8]Victorian Infectious Disease Service, The Royal Melbourne Hospital and The University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia. [9]Menzies School of Health Research, Darwin, Australia. [10]Department of Microbiology, Obafemi Awolowo University, Ile-Ife, Nigeria. [11]Department of Medical Microbiology and Infectious Diseases, Erasmus MC, Rotterdam, The Netherlands. [12]Centre for Genomic Pathogen Surveillance, Hinxton, UK. [13]Department of Infectious Disease Epidemiology, Imperial College London, London, UK. [14]Wellcome Trust Sanger Institute, Hinxton, UK. [15]London School of Hygiene and Tropical Medicine, London, UK. [16]Helsinki Institute for Information Technology, Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland. [17]Department of Biostatistics, University of Oslo, Oslo, Norway. [18]Present address: Institute of Microbiology and Infection, University of Birmingham, Birmingham, UK. [19]These authors contributed equally: Emily J. Richardson, Rodrigo Bacigalupe, Ewan M. Harrison, Lucy A. Weinert. *e-mail: Ross.Fitzgerald@ed.ac.uk

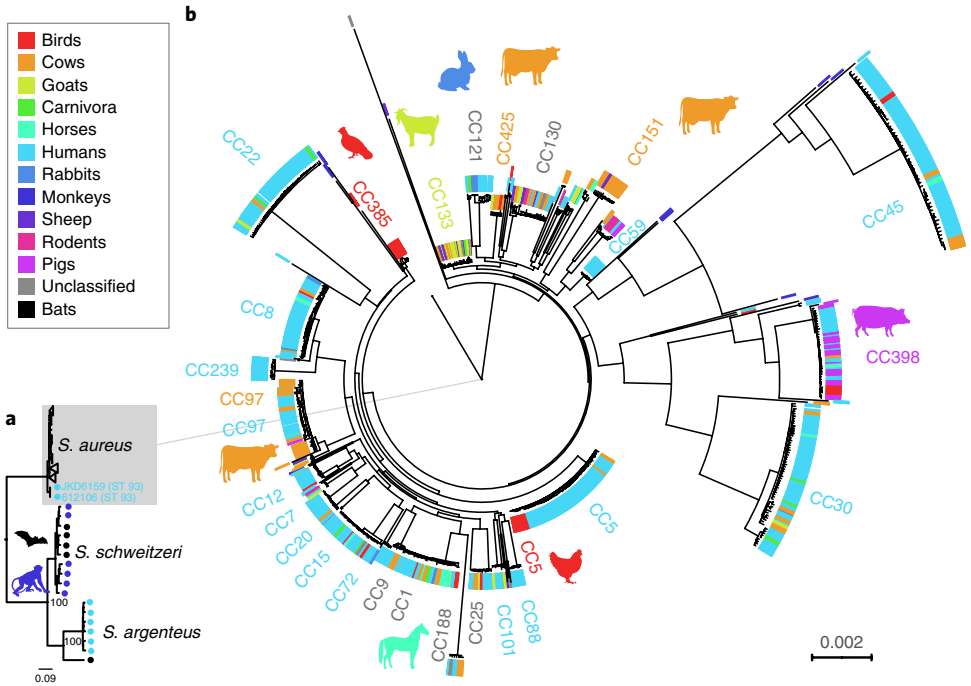

Fig. 1 | *S. aureus* phylogeny according to host-species origin. a, Phylogenetic tree of 800 isolates, constructed using the maximum-likelihood method, demonstrating the relationship between *S. aureus* and other members of the *S. aureus* complex—*S. schweitzeri* and *S. argenteus*. **b**, Phylogenetic (maximum-likelihood) tree of 783 *S. aureus* isolates, with colours indicating the host species. Animal symbols indicate major domesticated animal clones that are largely host specific. The evolutionary history of *S. aureus* was calibrated using well-established substitution rates from published datasets (see Methods).

host-specific functional activity of *S. aureus* effectors, such as leucocidins, superantigens and the von Willebrand factor-binding protein[21–26]. In addition, it was demonstrated that for *S. aureus* strains associated with natural infections of rabbits, a single mutation was responsible for conferring infectivity to the progenitor strain found in human populations[9]. Taken together, these studies highlight the capacity for bacteria to undergo host-switching events and adapt to different species by multiple evolutionary genetic and functional mechanisms. However, a large-scale, genome-based analysis of the evolutionary history of *S. aureus* in the context of its host ecology is lacking, and the scale and molecular basis of host-switching events remains poorly understood.

Here, we carry out a population-genomic analysis of over 800 *S. aureus* isolates selected to represent the known breadth of host-species diversity in order to provide a high-resolution picture of the dynamics of *S. aureus* in the context of its host. The data reveal the impact of human activities such as domestication and the use of antibiotics in medicine and agriculture on the recent evolution of *S. aureus*, and identify the key evolutionary processes underpinning its multi-host-species ecology.

Results

Extensive host-switching events define the evolutionary history of *S. aureus*. We selected *S. aureus* strains to represent the breadth of the known clonal, geographic and host-species diversity (see Methods for details of isolate selection). Overall, we included 800 isolates representative of 43 different host species and 77 clonal complexes, isolated in 50 different countries across 5 continents (Supplementary Figs. 1–3 and Supplementary Table 1). Among the 800 isolates, a total of 115,149 single nucleotide polymorphisms (SNPs) were identified in a core genome of 711,562 base pairs and used for reconstruction of the maximum-likelihood phylogeny for the *S. aureus* species (Fig. 1). The *S. aureus* species tree indicates the existence of highly divergent clades representative of the recently described *Staphylococcus argenteus* and *Staphylococcus schweitzeri* species, which belong to the extended *S. aureus*-related complex (Fig. 1a)[27]. *S. argenteus*—an emerging cause of human clinical infection[28]—is more closely related to bat and monkey isolates than other human *S. aureus* sequence types, consistent with a possible non-human evolutionary origin for *S. argenteus*. Removal of isolates from the divergent clades resulted in a phylogeny of 783 isolates that segregated according to clonal complexes defined by MLST (Fig. 1b). The phylogeny indicates the broad diversity of isolates of human origin, with expansion of several successful epidemic hospital- and community-associated clones including CC22, CC30 and ST45, as previously described[29] (Fig. 1). Animal isolates are typically found in discrete host-specific clades interspersed among human lineages, consistent with ancient and recent host-switching events across the phylogenetic tree (Fig. 1). To examine the frequency and timing of host-switching events during the evolution of *S. aureus*, we employed Bayesian evolutionary analysis by sampling trees (BEAST) using substitution rates from published datasets (Fig. 2 and Supplementary Table 2). We estimated the number of cross-species transmissions for ten major host categories (Supplementary Table 3 and Supplementary Figs. 2–5) using BEAST with Markov Jumps[30]. To reduce bias caused by the larger numbers of sequences from human and cow hosts compared with the other host types, we used 10 stratified subsamples containing 252 sequences each, designed to maintain geographic, host-type and temporal diversity while reducing over-representation. To assess the robustness of the main analysis, we performed additional analyses as outlined in the Supplementary Materials (Supplementary Notes, Supplementary Figs. 4–11 and Supplementary Tables 4–5), which included 'severe balanced' subsamples of 97 taxa each containing 18–20 taxa of 5 host types, and ancestral state and host jumps using the BASTA approximation to the structured coalescent[31]. However, we had difficulty getting BASTA to run and converge, possibly due to its assumptions about the structure of the data and numerical instability. Each subsampled sequence set was analysed separately within BEAST and resulted

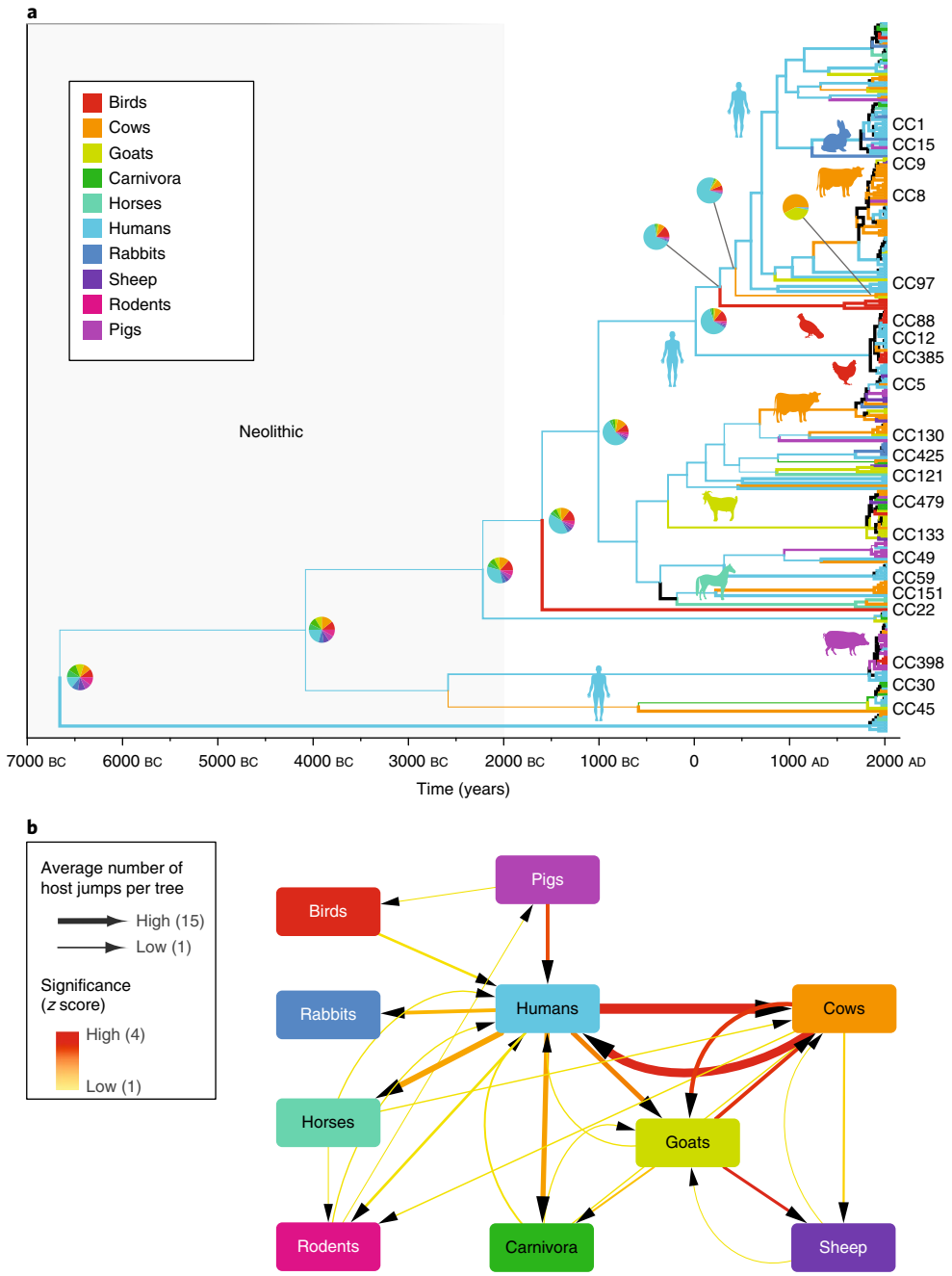

Fig. 2 | *S. aureus* has undergone extensive ancient and recent host-switching events, with humans acting as a major hub. a, Time-scaled phylogeny of a subsample of the *S. aureus* sequences. Clonal complexes are labelled, and branches are coloured according to the host-species group. Pie charts indicate the relative probability of host origin at the ancestral nodes, and line thickness corresponds to the probability of the majority host (see Supplementary Figs. 6–10 for all subsamples). **b**, Quantification of the number of host-switching events, illustrated as a host-transition count network based on BEAST Markov Jumps models averaged over all subsamples of the data. The line width represents the Markov Jump count per tree averaged over all subsamples (Supplementary Figs. 4 and 5), and line colour represents the significance compared with permuted label analysis (*z* score). Only transitions with higher counts compared with models with permuted host labels are shown (*z* score ≥ 0.5).

in a collection of posterior trees per dataset (Supplementary Figs. 6–10). In each case, the analysis revealed extensive host-switching events that occurred over a time frame spanning several thousand years up to the present decade (Fig. 2a).

Our analysis identifies humans as a major donor, with host jumps identified from humans into all other host-species groups examined (Fig. 2b and Supplementary Fig. 4). The most common recipient for *S. aureus* jumps from humans was cows, with a median of 14 jumps (highest posterior density (HPD) interval, 3–22) between the years 2000 BC and 2012 AD. Cows also represented a major donor for host-switching events back into humans ($n = 10$; HPD 2–26). In addition, there were numerous *S. aureus* host switches among ruminants, particularly between cattle and goats in both directions and into sheep. However, host jumps from sheep into other species were rare and not strongly supported by our analyses, suggesting that although sheep are a common host for *S. aureus*[5], they do not represent a major reservoir for the spread of *S. aureus* to other animals.

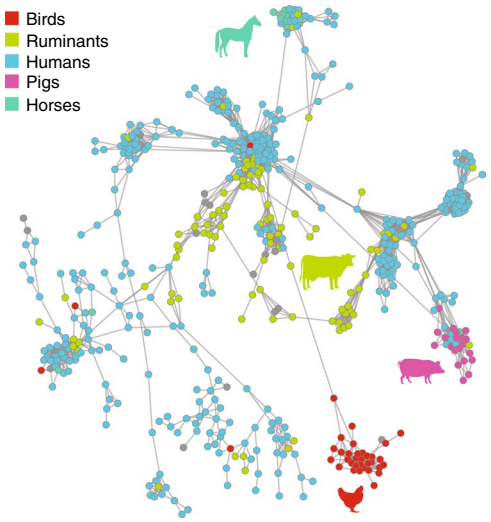

Fig. 3 | Network analysis of the *S. aureus* accessory genome indicates clustering according to host-species group. Network graph of pairwise distances of accessory genome gene content between isolates. Each node represents an isolate, colour-coded to indicate the host-species origin. Each edge indicates >50% shared accessory genome content, with the length of the edges weighted by distance (the proportion of shared accessory genes; shorter edges have more genes in common). All edges with <50% shared accessory genome content were removed.

Host-specific accessory gene pools promote adaptive evolution after host-switching events. To investigate the distribution of MGEs on a population level across human and animal isolates, we employed a pan-genome-wide association analysis approach to identify genes that were enriched among isolates from specific host species. First, to account for phylogeny, we removed genes identified among all strains within clonal complexes associated with multiple host species (lineage-dependent genes). Network analysis indicated a remarkable correlation between accessory genome and host species, revealing that diverse clonal complexes can share highly similar accessory genomes that are specific for birds, pigs or horses, respectively. This strongly points to the existence of a host-specific gene pool required for *S. aureus* host adaptation. Although accessory genomes of *S. aureus* obtained from humans—and from cows, sheep and goats—also tended to cluster together in a host-specific manner, there was greater diversity in gene content (Fig. 3). This may reflect the existence of multiple cryptic niches that exist within a single host species, such as those proposed previously for *Campylobacter jejuni*[32]. We note the existence of a small number of clusters made of isolates from multiple host species. The existence of these clusters suggests that some accessory gene combinations may confer a more generalist host tropism with the capacity to infect multiple host species. Alternatively, insufficient time may have passed since the host-transition event for loss of dispensable MGE to occur. Of note, antibiotic resistance gene determinants influenced the clustering of equine and pig isolates, suggesting a role for the acquisition of resistance in host adaptation (Supplementary Fig. 12).

Further examination of the impact of the accessory genome on host adaptation was carried out by identifying gene acquisition or loss events that correlated with host-switching events identified on the phylogeny of *S. aureus*. A total of 36 distinct MGEs, including predicted plasmids, transposons, *S. aureus* pathogenicity islands and prophages, were identified to be associated with host-switching events ($P < 0.0001$) (Fig. 4a and Supplementary Table 6). Several of the MGEs have previously been identified and shown to encode proteins with host-specific activity. For example, the β-converting phage ϕSa3 encodes modulators of the human innate immune response, and pathogenicity islands encode superantigens or von Willebrand factor-binding proteins with ruminant-specific activity[19,33]. In addition, equine isolates contain a phage encoding a novel equine allele of the staphylococcal inhibitor of complement (*scn*), which also encodes the LukPQ toxin recently characterized to have equine-specific activity[22,25]. However, numerous uncharacterized MGEs have been identified in the current study to be linked to successful host-switching events, providing many novel avenues for characterizing the molecular basis of *S. aureus* host adaptation (Fig. 4b). For example, in isolates from pigs, a putative novel plasmid linked to SCC*mec*, encoding resistance to heavy metal ions (a common supplement in pig-feed), was linked to host-switching events from humans into pigs (Fig. 4b). Finally, several gene clusters encoding bacterocins were enriched in isolates from specific host species ($P < 0.0001$) or were linked to host-switching events ($P < 0.0001$), consistent with the need to compete with resident bacteria for survival (Supplementary Table 6). Taken together, these data suggest that successful host-switching events are associated with the acquisition of MGEs from an accessory gene pool that exists in the recipient host species, and/or loss of MGEs linked to the source species.

To investigate the potential origin of MGEs horizontally acquired after a host-switching event, we examined the codon usage bias of host-specific MGEs. We found that MGEs enriched in pig isolates had a significantly elevated percentage of guanine–cytosine content and a reduced codon adaptation index (CAI), indicative of a distinct genealogical origin (Supplementary Figs. 13–15). Of note, an MGE found in pig isolates had the highest BLASTn similarity to a putative pathogenicity island previously identified in the pig-associated zoonotic pathogen *Streptococcus suis* (guanine–cytosine content of ~41%) (Supplementary Table 6).

Both gain and loss of gene function are associated with *S. aureus* host adaptation. Determination of the number of predicted functional genes in each *S. aureus* genome identified a significantly higher number of genes in bird strains compared with any other host species (Supplementary Figs. 16 and 17).

In contrast, the number of pseudogenes per genome was significantly higher ($P < 0.0001$–0.02) in ruminant strains compared with those from other host species, suggesting that the niche occupied by *S. aureus* in cows may provide stronger selection for loss of gene function compared with the niches for *S. aureus* in birds and pigs. Numerous pseudogenes associated with the transport of nutrients in *S. aureus*, including carbohydrates, are over-represented in ruminant isolates, implying metabolic remodelling in response to distinct nutrient availabilities in the bovine niche (Supplementary Table 8).

Refinement of host adaptation involves the modification of biological pathways in response to nutrient availability. In addition to accessory genes, adaptive mutations in the core genome may be selected for in response to environmental changes, such as antibiotic exposure or a switch in host species[9,34]. To examine the impact of host species on diversification of the *S. aureus* core genome, we identified groups of related isolates (for example, within clonal complexes or sequence types) associated with a specific host species for genome-wide analysis of positive selection (Supplementary Table 9 and Supplementary Fig. 18). Positive selection was identified across all host-associated groups examined, with an average of 68 genes (33–129) representing approximately 2.7% (1.3–5.1%) of a clade-specific core genome (Supplementary Table 10). A limited number of genes were under diversifying selection across multiple host species, including several that encode membrane proteins, lipoproteins and a protein involved in biofilm formation. Some genes were identified as undergoing positive selection in distinct lineages that were associated with the same host species (mostly human), suggesting

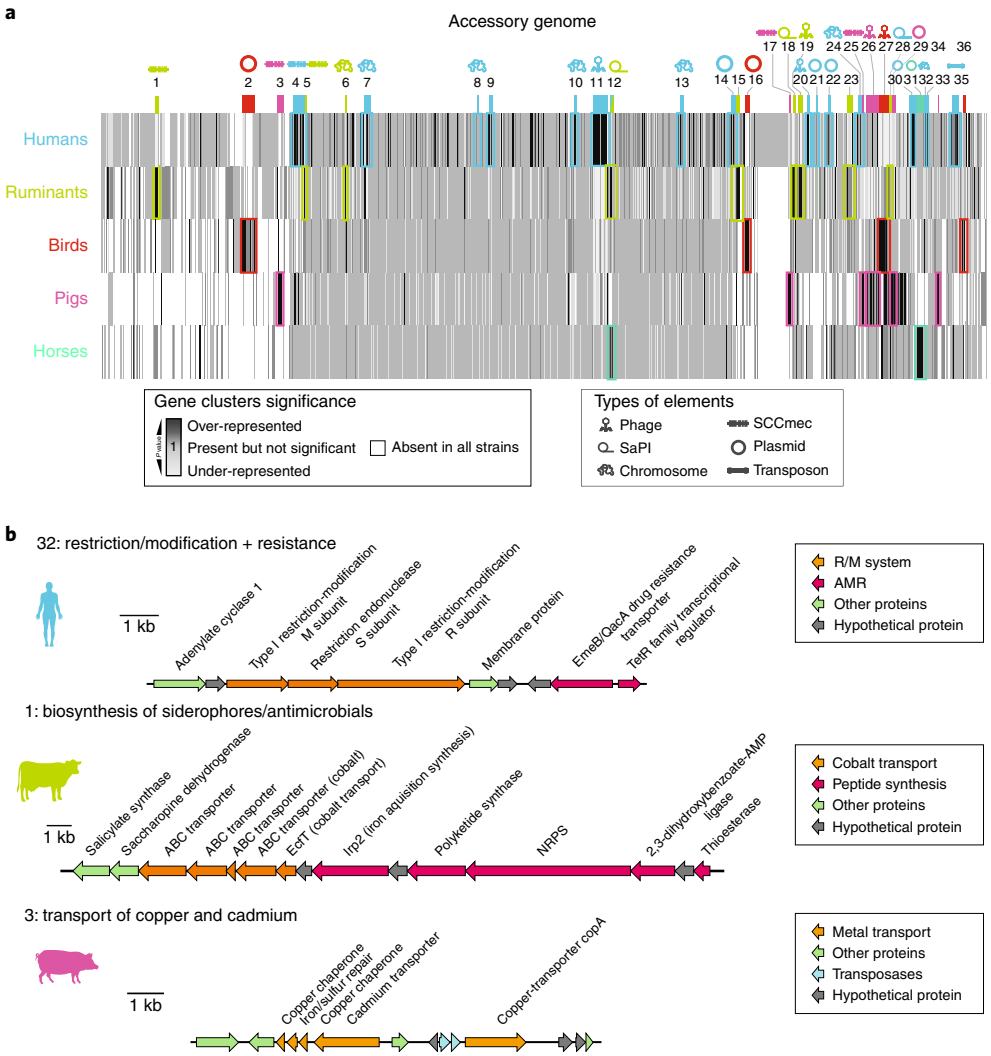

Fig. 4 | Identification of horizontally acquired genetic elements correlated with host adaptation. a, Schematic of the *S. aureus* pan-genome. Gene clusters linked to host species are indicated by shading. Coloured symbols indicate the nature of the MGE and the associated host species. SaPI, *S. aureus* pathogenicity island; SCCmec, staphylococcal cassette chromosome mec. **b**, Annotated gene maps of selected novel genetic elements linked to specific host species and associated with the acquisition of MGEs from an accessory gene pool that exists in the recipient host species and/or the loss of MGEs linked to the source species. kb, kilobase; R/M, restriction/modification; AMR, anti-microbial resistance-associated; ABC, ATP-binding cassette transporters; NRPS, non-ribosomal peptide synthase.

strong selective pressure leading to convergent evolution. However, for the most part, our analysis detected distinct sets of genes under positive selection in different lineages, suggesting that signatures of host adaptation are dependent on the genetic background of the strain, and that host adaptation can occur via multiple trajectories involving the modification of distinct pathways.

We predicted functional categories of genes under positive selection and the biological pathways affected, revealing several functional groups that were enriched for positively selected genes independent of the host species, including genes linked to pathogenesis, immune evasion and the maintenance of MGEs (Supplementary Table 11 and Supplementary Fig. 19). However, the majority of the functional categories were host-species dependent, consistent with distinct mechanisms underpinning adaptation to different host species (Supplementary Table 11; summarized in Fig. 5). In particular, biological pathways associated with amino acid metabolism and iron acquisition were under positive selection in several host species, suggesting diversification in response to distinct nutrient availability in different host niches. In addition, genes associated with the transport and metabolism of carbohydrates demonstrated signatures of positive selection in *S. aureus* clones from humans and cows (Fig. 5).

Bovine *S. aureus* strains utilize lactose with higher efficiency compared with human or avian strains. Considering the signatures of positive selection identified among pathways associated with carbohydrate and amino acid metabolism, we investigated differences in the growth phenotype of selected host-specific *S. aureus* strains using a metabolic phenotype microarray (Biolog), and observed preliminary strain-dependent differences in growth that were influenced by the availability of specific amino acids or carbohydrates. For example, *S. aureus* strains from cows had higher relative growth in the presence of lactose—the primary disaccharide available in bovine milk. The genome-wide positive selection analysis indicated that in bovine strains, genes associated with the functional category of transport of disaccharides and oligosaccharides were impacted by positive selection. To further investigate this, we carried out phenotypic analysis of *S. aureus*

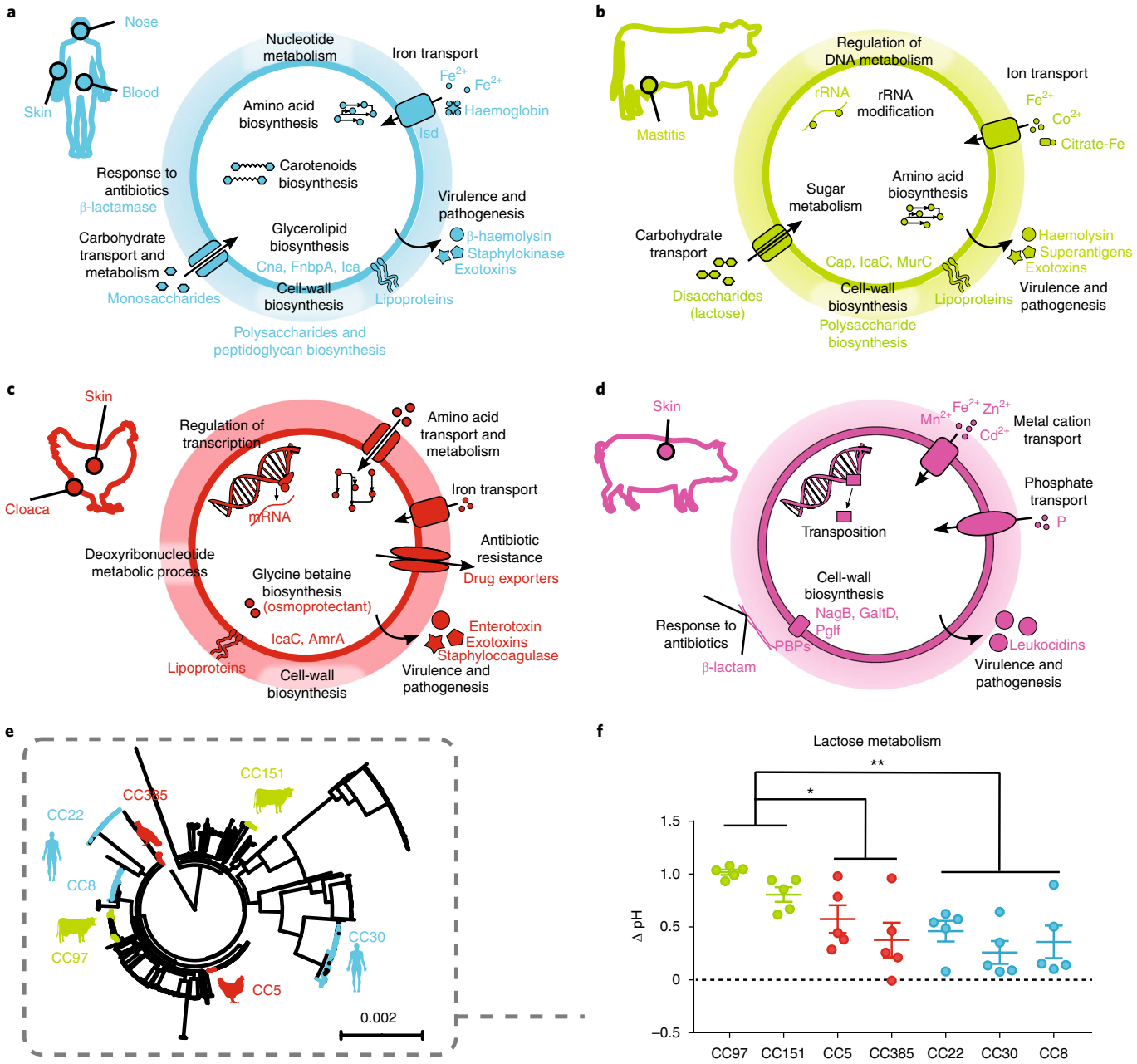

Fig. 5 | Summary of biological pathways under positive selection in different host species, and evidence for phenotypic adaptation. a–d, The main anatomical isolation sites on each host group are indicated by filled circles. The functional categories virulence and pathogenesis, resistance to antibiotics, transport of ions and cell-wall biosynthesis were under positive selection in all four host-species groups. In humans (**a**) and ruminants (**b**), the categories amino acid biosynthesis and transport/metabolism of carbohydrates were positively selected. The categories amino acid transport/metabolism and biosynthesis of osmoprotectants were under positive selection in birds (**c**), and the category transposable elements was under positive selection in pigs (**d**). mRNA, messenger RNA; rRNA, ribosomal RNA; Cna, collagen adhesin; FnbpA, fibronectin-binding protein A; Ica, intercellular adhesin; Cap, capsule biosynthesis protein; IcaC, intercellular adhesin protein C; MurC, UDP-N-acetylmuramoyl-L-alanine synthetase; AmrA, activation of Mga regulon expression locus A; PG1F, UDP-N-acetyl-alpha-D-glucosamine C6 dehydratase; GalD, galactose dehydratase; NagB, glucosamine-6-phosphate deaminase; PBPs, penicillin-binding proteins. **e**, Phylogenetic tree indicating the distinct lineages selected for comparative analysis of lactose fermentation. **f**, Acidification of the *S. aureus* culture supernatant in the presence of 100 mM lactose, as indicated by ΔpH. Fermentation of the disaccharide lactose is enhanced in bovine lineages. Experiments were performed in triplicate with five strains per clonal lineage. Each dot represents the average ΔpH per strain and bars indicate the s.e.m. per clonal lineage ($n=5$). Asterisks indicate significant differences between bovine (CC97 and CC151; $n=10$), avian (CC5 and CC385; $n=10$) and human lineages (CC22, CC30 and CC8; $n=15$): $^{*}P<0.005$, $^{**}P<0.0001$, one-way ANOVA followed by Tukey's multiple comparison test.

strains from bovine, human and avian host species of different clonal complexes when grown in the presence of lactose (Fig. 5e). As lactic acid is produced by *S. aureus* as a by-product of fermentation, we measured pH levels in culture media containing lactose and identified a decrease in pH levels for bovine *S. aureus* clones compared with human or avian clones, consistent with an increased efficiency of lactose fermentation (Fig. 5f). These data support the concept that *S. aureus* undergoes genetic diversification in response to the nutrients that differ in availability in different niches.

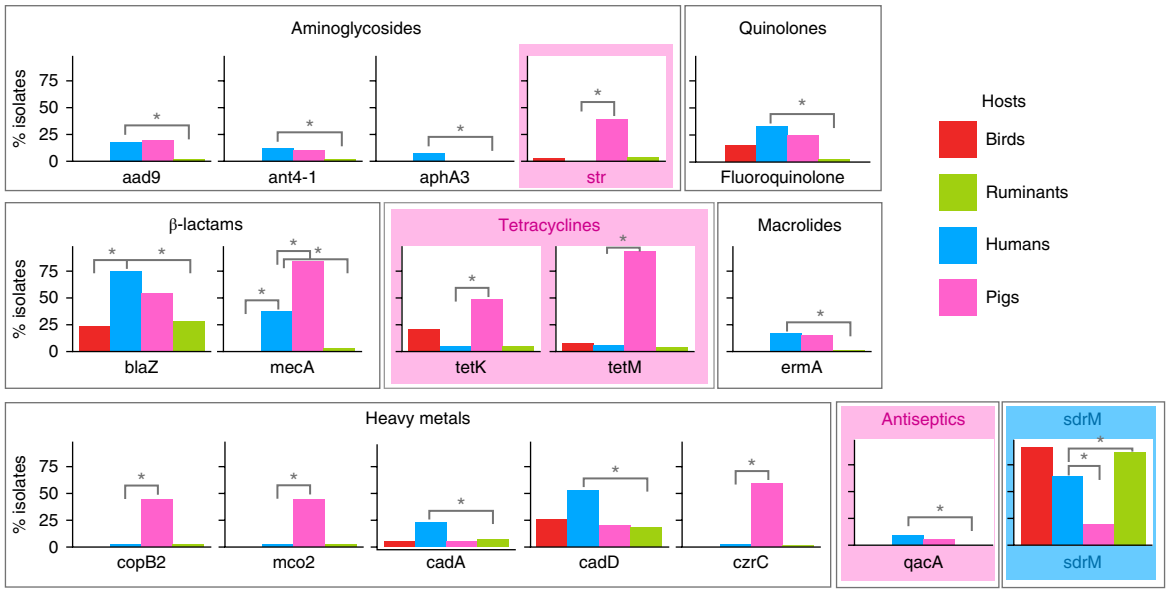

Fig. 6 | Resistance to antimicrobials is non-randomly associated with host species. Proportion (%) of isolates examined that contained the specified resistance determinant (Supplementary Table 12). Asterisks indicate a significant association ($P < 0.05$) between resistance determinants and the host species (Fisher's exact test). Coloured borders indicate the antibiotic class or single determinants (*sdrM* and *str*) that are associated with the host-species group after testing for phylogenetic independence.

Resistance to antimicrobials differs among human and pig *S. aureus*. Our understanding of the relative contribution of the use of antibiotics in human medicine and agriculture to the emergence of antibiotic resistance is very limited. To address this question for the model human and animal pathogen *S. aureus*, we examined the distribution of antibiotic, antiseptic and heavy metal ion resistance determinants among human and livestock isolates, and then accounted for phylogenetic relatedness for resistance to different classes of antibiotic (Supplementary Table 12). An array of resistance determinants were significantly enriched in human, ruminant and pig isolates, respectively, but not among avian isolates, consistent with a limited role for the poultry industry in the emergence of antibiotic resistance in *S. aureus* (Fig. 6 and Supplementary Table 12). When testing for phylogenetic independence, we aimed to maximize statistical power by including all gene determinants in groups specific for each class of antimicrobial, and also examined selected individual determinants *str* and *sdrM*. The analysis indicated that resistances to streptomycin, antiseptics and tetracyclines were all significantly associated with pig isolates, whereas *sdrM* was enriched in human isolates. However, fluoroquinolone and heavy metal ion resistance did not correlate with hosts after correction for phylogeny, implying that the expansion of specific clones has contributed to the high frequency of those resistance determinants among human and pig hosts, respectively. Taken together, these data demonstrate that resistance to specific classes of antimicrobial in *S. aureus* is host-species dependent, providing evidence for distinct antibiotic selective pressures in humans and livestock. Of note, tetracyclines and aminoglycosides (such as streptomycin) are used in much higher amounts in farmed animals compared with human medicine[35]. Zoonotic transmission of *S. aureus* is a relatively common occurrence for some clones, particularly between pigs and humans in the case of CC398, providing a route for the transmission of resistant strains and associated resistance determinants to humans[36].

Discussion

Many new pathogens emerge following zoonotic or anthroponotic events, providing the opportunity for spread within a new host population[2]. *S. aureus* is considered a generalist bacterial species capable of colonizing a wide range of hosts[5]. However, the species is composed of distinct sublineages that are commonly associated with particular hosts or host groups[10,14]. Accordingly, *S. aureus* represents an excellent model for exploring the dynamics of a bacterial pathogen at the human–animal interface. Here, we demonstrate that the segregated host specialism of *S. aureus* arose via multiple cross-species transmission events that occurred over the past 5,000–6,000 years, leading to the emergence of successful endemic and epidemic clones circulating in distinct host-species populations. We identify humans as a major reservoir for the spread of *S. aureus* to livestock, reflecting the role of humans in the domestication of animals, and subsequent opportunities for cross-species transmission events consistent with analyses using MLST[12]. Importantly, we also identify cows as the main animal source for the emergence of *S. aureus* clones that are epidemic in human populations, consistent with a previous study that identified a bovine origin for emergent CC97 clones causing human infections across multiple continents[17].

The identification of combinations of MGEs that are associated with specific host species and linked to host-switching events provides compelling evidence for the key role of horizontal gene acquisition in the adaptation of *S. aureus* to its hosts. While several MGEs have been identified to be associated with host-specific clones[18,19,22,24], our species-wide analysis reveals combinations of MGEs linked to specific host species, providing many new avenues for investigating the mechanisms of bacterial host adaptation. Overall, the data suggest that host-specific accessory gene pools that are presumably present in the microbiota of the new host species promote the host-adaptive evolution of *S. aureus*.

In addition to gene acquisition associated with host-switching events, we identified evidence of adaptive evolution in the core genome consistent with host-specific selective pressure driving the diversification of biological pathways that are involved in survival or transmission. Furthermore, in some cases, distinct pathways were under positive selective pressure in different clones associated with the same host species, implying that multiple distinct pathways may mediate host adaptation depending on the genetic background of the strain. In particular, pathways linked to carbohydrate transport exhibited signatures of host adaptation, and phenotypic

analysis revealed enhanced utilization by bovine *S. aureus* clones of the disaccharide lactose—the major carbohydrate available in bovine milk.

These findings inform a model of *S. aureus* host adaptation in which the acquisition of a specific set of MGEs occurs rapidly after a host-switching event (although we cannot rule out that this could occur before the jump in some cases), conferring the capacity for survival in the new host, largely through targeting of the innate immune response via bacterial effectors such as leukocidins, superantigens and other immune modulators. Other MGEs confer resistance to antibiotics and heavy metal ions allowing survival under strong antimicrobial selective pressures. Subsequently, positive selection acts on the core genome via point mutation and/or recombination[37] causing allelic variation and loss of gene function that results in modification of metabolism in response to distinct nutrient availability.

Our findings suggest that since human-driven domestication, interactions with livestock have provided opportunities for numerous successful host-switching events between humans and livestock hosts. Furthermore, the industrialization of agriculture, including the use of antibiotics and feed supplements in intensive farming, has directly influenced the evolution of *S. aureus* clones, resulting in the emergence of resistance in response to distinct antibiotic selective pressures in human medicine and agriculture[18,38]. These data support the idea that surveillance could play a critical role in the early identification of emerging clones that have jumped host.

Taken together, our data provide a high-resolution view of the capacity for a model multi-host pathogen to undergo radical changes in host ecology by genetic adaptation. Investigation into the functional basis of these genetic changes will reveal key host–pathogen interactions that could be targeted for novel therapies. Furthermore, the identification of the common routes for *S. aureus* livestock–human host-species switches and distinct types of antimicrobial resistance in humans and livestock species could inform the design of more effective farm security and antibiotic treatment practices to limit the emergence of new resistant clones. These findings will be relevant to other major bacterial pathogens with the capacity to spread between livestock and humans.

Methods

Isolate selection. To select the isolates the literature was reviewed in November 2013, and all available *S. aureus* strains associated with animals and humans for which genomes had been determined were identified. We aimed to include isolates to represent the breadth of clonal complexes, host-species diversity, geographical locations and as wide a temporal scale as possible (Supplementary Tables 1–3). Publicly available sequences were selected as follows: 74 reference genomes, 302 from the EARSS project[29] and 252 from other published studies of the authors (Supplementary Table 1). Furthermore, to best represent the known *S. aureus* host, as well as clonal and geographic diversity, we selected an additional 172 isolates for whole-genome sequencing (Supplementary Table 1). Our dataset was biased towards human isolates, which represented approximately 60% of the total (approximately 40% were from animal sources). This reflects the fact that much of the known diversity of the *S. aureus* species is of human origin[12], and also that fewer numbers of isolates have been obtained from animals. Given the predominant European origin of the animal isolates (due to the contemporary interest in animal *S. aureus* in Europe), we chose to enrich the number of human isolates with the EARSS collection, representative of the diversity of invasive *S. aureus* circulating among humans in Europe in 2006[29]. Accordingly, there is a European bias to the sample dataset and we cannot rule out that we have undersampled the *S. aureus* diversity that exists in other parts of the world. Nonetheless, our dataset contained isolates from 50 different countries across 5 continents, and many sequence types are widely distributed on an intercontinental scale. In addition, our dataset includes isolates from the years 1930 to 2014, although the majority have been isolated since 2005, reflecting the greater availability of recent clinical isolates (particularly from animals). It should therefore also be considered that the dataset is biased towards contemporary *S. aureus* and that older lineages that are now less abundant or extinct may not be represented in our dataset. To partially address the uneven distributions of isolates by host, space and time, where appropriate, we carried out experimental replicates based on severe subsampling of the dataset to provide more evenly distributed groups. In addition, we drew conclusions that were consistent across subsampled data and, when appropriate, multiple different analytic approaches. Overall, we included 800 isolates representative of 43 different host species and 77 clonal complexes, isolated in 50 countries across 5 continents (Supplementary Table 1). All sequences and associated metadata have been uploaded to Microreact—a publicly accessible database that allows visualization and analysis of the data https://microreact.org/project/shacdata[39].

Sequencing, genome assemblies, variant calling and phylogenetic reconstruction. For the current study, bacterial DNA was extracted and sequenced using an Illumina HiSeq 2000 with 100-cycle paired-end runs at the Wellcome Trust Sanger Institute or an Illumina HiSeq 2000 at Edinburgh Genomics. The nucleotide sequence data were submitted to the European Nucleotide Archive (ENA; www.ebi.ac.uk/ena) with the accession numbers listed in Supplementary Table 1. Completed genomes downloaded from the National Center for Biotechnology Information database were converted into pseudo-FASTQ files using Wgsim (https://github.com/lh3/wgsim). For each isolate, the sequence reads were used to create multiple assemblies using VelvetOptimiser version 2.2.5 (ref. [40]) and Velvet version 1.2 (ref. [41]). The assemblies were improved by scaffolding the best N50 and contigs using SSPACE[42], and sequence gaps were filled using GapFiller[43]. Isolates were excluded from the analysis for the following reasons that are indicative of contamination or poor-quality sequence data: a large number of contigs and a large number of 'N's in the assemblies; or large genome size (>2.9 megabases). Sequence types were determined from the assemblies using MLST check (https://github.com/sanger-pathogens/mlst_check), which was used to compare the assembled genomes against the MLST database for *S. aureus* (http://pubmlst.org/saureus/). Sequence reads were mapped to a relevant reference genome (ENA ST425 (strain LGA251; accession number FR821779) using SMALT (http://www.sanger.ac.uk/science/tools/smalt-0) following the default settings to identify SNPs. Consensus sequences were obtained using SAMtools and concatenated into core genome alignments[44]. SNPs located in MGEs were removed from the alignments and a maximum-likelihood tree was constructed using RAxML following default settings and 1,000 bootstrap replicates[45].

Time-scaled trees and estimation of the number of host jumps. Time-scaled trees were generated using BEAST 1.8.2 (ref. [46]). All isolates with unknown date, unknown host species or unknown geographical location were removed in addition to the diverse Bayesian analysis of population structure (BAPs)[47] groups 12 and 14, leaving a total of 696 isolates. Sites determined to be affected by recombination from the BratNextGen (BNG)[48] analysis of the individual BAPs groups were coded as missing data. Since missing data can effect phylogenetic inference and contribute to heavy likelihood calculations, sites that had more than one missing state in the alignment (either from missing mapped reads or recombination) were excluded from further analyses, leaving a total of 55,778 sites (4,306 segregating sites).

To account for different evolutionary processes acting at synonymous and non-synonymous sites, RNA, and non-coding sites, the evolutionary model was partitioned into 1 + second sites, third sites, non-coding sites and RNAs according to the reference strain LGA251. Pseudogenes were partitioned into 1 + second and third sites with the rationale that they may be functional in other isolates. For overlapping reading frames, sites were assigned to the region of highest constraint (for example, for coding and RNA, sites were assigned as RNA; for first and third, sites were assigned as 1 + second, and so on). For all partitions, we used an HKY + Γ substitution model.

Dating. We treated all sequences as contemporaneous, but assigned a median prior of 1.61 (HPD interval 0.604 to 2.9) substitutions per site per million years on to third positions, which are less likely to be subject to strong purifying selection (known to affect rates over different timescales). The prior comes from previous studies of *S. aureus* using tip dates on different strain types (Supplementary Table 3). An uncorrelated log-normal model of changes in the substitution rate across different branches was employed. An initial Markov chain Monte Carlo (MCMC) run with this model was performed in BEAST version 1.8.2 (ref. [46]) using Beagle[49] with 2 independent chains, removing the appropriate burn-in and run for approximately 100,000,000 generations.

In addition, ten further subsampled datasets were produced that included only sequences from the ten major host types. These were stratified subsamples containing 252 samples each, designed to maintain the host species, and geographic and temporal diversity. The major host types were birds, cows, goats, carnivores, horses, humans, rabbits, sheep, rodents and pigs. From these analyses, we subsampled an empirical distribution of 1,000 trees post burn-in, which were used for all further BEAST analyses.

Markov Jump analysis. To reconstruct host-transition events, we used an asymmetric discrete state phylogeographic analysis with Markov Jumps[50] applied to the ten major host types with default priors. We employed the Markov Jump analysis to estimate the posterior expectation of the number of host change events across the branches of the phylogeny[30], using posterior sets of 1,000 time-scaled trees from the initial BEAST analyses on the subsampled datasets. The trait models were used in an MCMC chain of 1,100,000 steps, sampling every 100 steps and discarding the first 10% as burn-in, leaving 10,000 trees annotated

with the host information (that is, approximately 10 model instances per tree of the original posterior set). Since biased sampling can lead to biased results when using these trait models, in addition to using the 10 stratified subsamples of 252 sequences each, we performed analyses with host state randomization and using 100 bootstrapped maximum-likelihood trees (RAxML) in place of the 1,000 original BEAST trees (for the 10 stratified subsamples). To balance the numbers of isolates per host category further, we also created 10 'severe' stratified subsamples containing 97 sequences each with 20 from humans, cows and sheep + goats combined, and 19 and 18 from birds and pigs, respectively. Necessarily, in these severe subsamples, it was not possible to maintain the full human and cattle diversity, although sequences from different geographic locations and years were chosen. We applied BEAST with Markov Jumps on these five host categories using: full joint inference of trees using sequences and traits together; trees using the sequences only followed by the trait mapping as before; and the BASTA structured coalescent approximation[51].

Pseudogene analysis. Pseudogenes were predicted during the PROKKA annotation process[51]. Specifically, each protein in a genome was searched against UniProtKB (Swiss-Prot) using BLASTp[52] or UniProtKB (TrEMBL). If no significant hits were identified, proteins were examined for conserved motifs. Any proteins that exhibited less than 95% coverage of their top hit were listed as potential pseudogenes. The region of the top hit that was not present in the protein sequence was then interrogated against all contigs using BLASTn[52]. Hits that were in the correct orientation and on the same contigs were accepted as pseudogenes and labelled according to their type (frameshift, stop codon or insertion). Proteins that exhibited less than 95% coverage of their top hit and were on the edge of a contig with their counterpart on another contig were not labelled as pseudogenes, rather coding sequences that had split due to the assembly breaking at this point.

The UniProt ID Mapping tool was used to assign Gene Ontology terms to all pseudogenes by transferring the Gene Ontology terms assigned to the closest reference (identified during the annotation process described above). Gene Ontology was assigned to all non-pseudogenes (coding sequence features) using the same method and InterProScan[53]. The R package topGO[50] with Fisher's exact test was used to identify enriched Gene Ontology terms while taking into account the Gene Ontology hierarchy (the *P* value was adjusted using Bonferroni correction).

Pan-genome association analysis. All genomes in this study were organized into a list of reference genomes followed by assembled contigs. The second genome in the list was aligned to the first genome using Nucmer[54], and any regions larger than 100 base pairs that did not map to the first genome were appended to the end of it to produce a pan-genome representing the unique regions in the first two genomes. Each subsequent genome aligned to the combination of all unique regions from the previously aligned genomes in the list, producing a pan-genome that represents all of the nucleotide sequences of all genomes. All genes were organized into groups of orthologues using the bi-directional best-hits algorithm in get_homologues with a minimum coverage setting of 50% and a minimum sequence identity setting of 80%[55]. The pan-genome was used as the reference, and the coding sequences predicted in the annotations described previously were compared with all coding sequences within the pan-genome. Features annotated as pseudogenes were excluded from this analysis. The get_homologues compare_clusters perl script was used to create a pan-genome matrix of all identified gene clusters against all genomes. All core gene clusters (clusters that contained genes from every genome) were removed from the pan-genome matrix. Further to this, all clusters that only contained genes from one genome or all genomes except one were removed. Furthermore, gene clusters that were found in all members of any sequence types associated with multiple host species were removed on the basis that they were not specific to a single host species. This had the effect of removing lineage-associated genes, resulting in a set of gene clusters that was strain dependent and largely independent of phylogeny. Hypergeometric testing was used to find over- and under-represented gene clusters for each host (the *P* value was adjusted using Bonferroni correction). All gene clusters were searched against the National Center for Biotechnology Information non-redundant nucleotide database using BLASTn to provide the most up-to-date annotation and to examine the probable bacterial species origin of each MGE.

A pairwise distance matrix was calculated from the pan-genome matrix using the distmat function in EMBOSS[56]. The matrix was converted into a bi-directional graph with distance as the edge weight parameter. The graph was processed in BioLayout with an edge weight threshold of 0.5 (ref. [57]).

Identification of gene acquisitions or losses associated with host-switching events. The R package APE (Analysis of Phylogenetics and Evolution)[58] was used to fit a single discrete trait model and get the ancestral state of each node for each gene cluster against the phylogenetic tree. From this, a vector for every gene cluster was created with gene acquisition/loss events by comparing every child node in the tree with its parent node to determine whether there was no change, a gene acquisition or a gene loss event. This was performed separately for each host type (that is, human, ruminant, bird, horse and pig) to identify which nodes were associated with a host-switching event. All gene state vectors were compared with all host state vectors using a Fisher's exact test to show whether a gene loss/acquisition was related to a host-switching event. The *P* values were adjusted using Bonferroni correction.

Codon usage bias analysis. The codon adaptation index was used to calculate codon usage bias by comparing the CAI of a gene against the codon usage table of a reference set of genes. The codon usage table was calculated using the EMBOSS tool cusp[56]. For this study, the codon usage table comprised all genes that were not significantly over-represented in a host or significantly associated with a host switch. The CAI for all genes significantly over-represented in a host and significantly associated with a host jump was calculated using the EMBOSS tool cai[56]. The codon adaptation index was also calculated using 5 random subsets of 50 genes as controls. A one-way analysis of variance (ANOVA) test was used to test whether there was a significant effect of host upon CAI. A Tukey's honest significant difference test was then applied to compare the CAIs between host species.

Distribution of antibiotic resistance genes analysis. Antimicrobial resistance genes were identified as described by Holden et al.[15]. Resistance genes were identified by a combination of BLASTn and mapping against assemblies, and as previously described[59]. Resistance SNPs were identified by mapping against a pseudomolecule of genes with previously reported resistance-conferring mutations. Isolates were grouped into human isolates and all animals, and then human, rabbits, companion animals (horses, dogs and cats), marine, pigs, primates, ruminants (goats, sheep and cows) and small mammals (rats, mice and other small mammals). The proportions of isolates with each resistance gene and ≥1 resistance-conferring SNP for each antibiotic was compared to identify enrichment using a two-tailed Fisher's exact test with Bonferroni correction for multiple testing. Determinates with a P value $< 9.9 \times 10^{-5}$ were considered statistically significantly enriched. To examine whether the Fisher's exact tests of independence were robust when accounting for population structure, we tested whether resistance phenotypes and host were correlated across the *S. aureus* phylogeny. We conducted these for pig/human and ruminant/human, since these were the only comparisons where significant differences were observed according to the Fisher's exact test. To maximize statistical power, we grouped all gene determinants into specific classes of antimicrobial (that is, tetracycline-resistant if encoding any *tet* allele) and tested for correlation with host species using the programme BayesTraits[60] (using the posterior sample of trees from our earlier BEAST analysis). We note that the correlated evolutionary analysis may be overly conservative in cases where horizontal gene transfer is rampant or homoplasies are high. BayesTraits uses a continuous-time Markov model to estimate transition rates between the presence and absence of a gene or SNP, and between human and non-human hosts. We allowed the transition rates to evolve in either a correlated fashion (where the rate of change in one trait depends on the state found in the other trait) or independently. Posterior distributions of parameters were estimated from up to 4 million iterations of the MCMC with default priors. After discarding burn-in, the marginal likelihoods of the dependent and independent models were obtained using the Akaike information criterion estimated using the methods-of-moment estimator in Tracer 1.6 (ref. [61]).

Genome-wide positive selection analysis. To identify genes under positive selection in different host groups, we first identified lineages (sequence types or clonal complexes) correlated with particular hosts. As the power of the selection analysis was determined by the number of isolates included, only clades with more than ten isolates associated with a host were considered. Based on these criteria, 15 clonal complexes from 4 groups of hosts were analysed: 9 for humans (CC30, CC5, CC59, CC15, CC12, ST239, ST8, CC22 and CC45), 3 for ruminants (CC133 (primarily associated with sheep and goats), and the cow-related CC151 and CC97), 2 for birds (CC5 and CC385) and 1 for pigs (CC398) (Supplementary Table 7). Although the CC398 clade also contained several human isolates, these mostly represented spill-over events rather than an established association, so the CC398-human group was not included in the analysis. Given the variable number of isolates of each clonal complex host group, to standardize the analysis while preventing the underestimation of genes under positive selection, ten isolates linked with a host were analysed at a time. Replicates or triplicates of different subsets of genomes using sampling with replacement were carried out if the number of isolates for that lineage was large enough. Next, we identified orthologous genes in each of these groups using the algorithm OrthoMCL integrated in get_homologues (identity >70%, similarity >75%, f50, e-value $= 1 \times 10^{-5}$)[55]. Genes were considered orthologous if they were present in at least 70% of the genomes. Since alignment of coding DNA sequences may insert gaps in codons and produce frame shifts, we aligned genes at the protein level using MUSCLE 3.8.31 (ref. [62]) and translated these sequences back to DNA using pal2nal version 14 (ref. [63]). Genes identified as inparalogous that turned out to be duplications were kept for further analyses, or else discarded. For every alignment, recombination was detected using the NSS, Max Chi and Phi tests included in PhiPack[64], and recombinant genes were removed from further analyses. For the gene clusters containing 10 isolates, phylogenetic trees were extracted from the 783-isolate maximum-likelihood tree. For clusters with

fewer than ten genomes, subtrees were produced from the general tree using the tree prune function in ete2 (ref. [65]). The DNA alignments and trees were used for PAML analysis[66]. We employed the site evolution models of Codeml (M1a, M2a, M7, M8 and M8a) to perform codon-by-codon analysis of dN/dS ratios (non-synonymous to synonymous substitution, ω) of genes, and a likelihood-ratio test was used to determine significant differences between the nested models M1a–M2a, M7–M8 and M8a–M8, where one accounts for positive selection (the alternative hypothesis) and the other specifies a neutral model (the null hypothesis). Statistic tests were assessed to a chi-squared distribution with 2 and 1 degrees of freedom[66]. Bayes Empirical Bayes[67] was used to calculate the posterior probabilities of amino acid sites under positive selection of proteins that had significant likelihood-ratio tests. As independent replicates from similar clonal complex/host groups resulted in slightly different genes being positively selected, we used get_homologues to merge the core genomes and genes selected for each group using the same parameters as above. Genes under positive selection were considered when they were in common for different replicates with a P value of 0.05 or were identified in different replicates with a stringent P value (0.05/number of genes per core genome).

To explore functional categories under positive selection, we performed classification of Clusters of Orthologous Groups (COGs), annotated Gene Ontology terms and analysed metabolic pathways (KEGG). To assign COG terms, we performed BLASTp of single representatives of the orthologous clusters against the prot2003-2014 database, retrieving the top five hits to include alternative annotations. We mapped the gene identifications (IDs) obtained to the cog2003-2014.csv database from which the COGs were inferred. Frequencies of COGs for positively selected genes in each clonal complex host were compared with the average COG frequencies in the respective core genomes. Gene Ontology annotations were obtained by mapping the genes to the go_20151121-seqdb, uniprot_sprot and uniprot_trembl databases using BLASTp. From these, the UniProtKB were mapped to the gene_association_goa database and filtered by bacteria domain to obtain the Gene Ontology categories. To visualize and identify over-represented Gene Ontology categories of positively selected genes in different hosts, we used BiNGO[68]. We identified over-represented categories using the hypergeometric test with the Benjamini and Hochberg false discovery rate multiple testing correction at a significance level of 5%. We chose the 'biological process' category and the prokaryotic ontology file (gosubset_prok.obo). However, as most groups did not show significant over-representation, we visualized all the Gene Ontology categories of genes under positive selection and used REVIGO[69] with the P values from BiNGO to obtain summaries of non-redundant Gene Ontology terms classified into functional categories.

Analysis of lactose fermentation. *S. aureus* was cultured in tryptic soy broth (TSB) in the presence or absence of 100 mM lactose at 37 °C for 17 h with shaking at 200 r.p.m. The optical density at a wavelength of 600 nm was measured and culture supernatants were collected by centrifugation. Subsequently, the pH of the supernatants was measured using a pH meter (Sartorius). ΔpH values were calculated by subtracting the pH values of TSB cultures supplemented with 100 mM lactose from the pH values of normal TSB cultures. Statistical analysis was performed in Graphpad Prism 7 using one-way ANOVA followed by Tukey's multiple comparison test.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability. The sequence datasets generated during the current study are available in the ENA (www.ebi.ac.uk/ena) with the accession number PRJEB20741. Accession numbers of previously published sequences analysed in the current study are listed in Supplementary Table 1. All data analysed during this study are included in this published article and its Supplementary Information files.

Received: 24 May 2017; Accepted: 25 June 2018;
Published online: 23 July 2018

References

1. Morand, S., McIntyre, K. M. & Baylis, M. Domesticated animals and human infectious diseases of zoonotic origins: domestication time matters. *Infect. Genet. Evol.* **24**, 76–81 (2014).
2. Woolhouse, M. E., Haydon, D. T. & Antia, R. Emerging pathogens: the epidemiology and evolution of species jumps. *Trends Ecol. Evol.* **20**, 238–244 (2005).
3. Lowy, F. D. *Staphylococcus aureus* infections. *N. Engl. J. Med.* **339**, 520–532 (1998).
4. Chambers, H. F. & Deleo, F. R. Waves of resistance: *Staphylococcus aureus* in the antibiotic era. *Nat. Rev. Microbiol.* **7**, 629–641 (2009).
5. Peton, V. & Le Loir, Y. *Staphylococcus aureus* in veterinary medicine. *Infect. Genet. Evol.* **21**, 602–615 (2014).
6. Bradley, A. J., Leach, K. A., Breen, J. E., Green, L. E. & Green, M. J. Survey of the incidence and aetiology of mastitis on dairy farms in England and Wales. *Vet. Rec.* **160**, 253–257 (2007).
7. McNamee, P. T. & Smyth, J. A. Bacterial chondronecrosis with osteomyelitis ('femoral head necrosis') of broiler chickens: a review. *Avian Pathol.* **29**, 477–495 (2000).
8. Van Duijkeren, E. et al. Methicillin-resistant *Staphylococcus aureus* in pigs with exudative epidermitis. *Emerg. Infect. Dis.* **13**, 1408–1410 (2007).
9. Viana, D. et al. A single natural nucleotide mutation alters bacterial pathogen host tropism. *Nat. Genet.* **47**, 361–366 (2015).
10. Feil, E. J. et al. How clonal is *Staphylococcus aureus*? *J. Bacteriol.* **185**, 3307–3316 (2003).
11. Shepheard, M. A. et al. Historical zoonoses and other changes in host tropism of *Staphylococcus aureus*, identified by phylogenetic analysis of a population dataset. *PLoS ONE* **8**, e62369 (2013).
12. Weinert, L. A. et al. Molecular dating of human-to-bovid host jumps by *Staphylococcus aureus* reveals an association with the spread of domestication. *Biol. Lett.* **8**, 829–832 (2012).
13. Price, L. B. et al. *Staphylococcus aureus* CC398: host adaptation and emergence of methicillin resistance in livestock. *mBio* **3**, e00305-11 (2012).
14. Fitzgerald, J. R. Livestock-associated *Staphylococcus aureus*: origin, evolution and public health threat. *Trends Microbiol.* **20**, 192–198 (2012).
15. Holden, M. T. et al. A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic. *Genome Res.* **23**, 653–664 (2013).
16. McAdam, P. R. et al. Molecular tracing of the emergence, adaptation, and transmission of hospital-associated methicillin-resistant *Staphylococcus aureus*. *Proc. Natl Acad. Sci. USA* **109**, 9107–9112 (2012).
17. Spoor, L. E. et al. Livestock origin for a human pandemic clone of community-associated methicillin-resistant *Staphylococcus aureus*. *mBio* **4**, e00356-13 (2013).
18. Lowder, B. V. et al. Recent human-to-poultry host jump, adaptation, and pandemic spread of *Staphylococcus aureus*. *Proc. Natl Acad. Sci. USA* **106**, 19545–19550 (2009).
19. Viana, D. et al. Adaptation of *Staphylococcus aureus* to ruminant and equine hosts involves SaPI-carried variants of von Willebrand factor-binding protein. *Mol. Microbiol.* **77**, 1583–1594 (2010).
20. Guinane, C. M. et al. Evolutionary genomics of *Staphylococcus aureus* reveals insights into the origin and molecular basis of ruminant host adaptation. *Genome Biol. Evol.* **2**, 454–466 (2010).
21. Koymans, K. J., Vrieling, M., Gorham, R. D.Jr & van Strijp, J. A. Staphylococcal immune evasion proteins: structure, function, and host adaptation. *Curr. Top. Microbiol. Immunol.* **409**, 441–489 (2017).
22. Koop, G. et al. Identification of LukPQ, a novel, equid-adapted leukocidin of *Staphylococcus aureus*. *Sci. Rep.* **7**, 40660 (2017).
23. Loffler, B. et al. *Staphylococcus aureus* panton-valentine leukocidin is a very potent cytotoxic factor for human neutrophils. *PLoS Pathog.* **6**, e1000715 (2010).
24. Vrieling, M. et al. LukMF' is the major secreted leukocidin of bovine *Staphylococcus aureus* and is produced in vivo during bovine mastitis. *Sci. Rep.* **6**, 37759 (2016).
25. De Jong, N. W. M. et al. Identification of a staphylococcal complement inhibitor with broad host specificity in equid *Staphylococcus aureus* strains. *J. Biol. Chem.* **293**, 4468–4477 (2018).
26. Wilson, G. J. et al. A novel core genome-encoded superantigen contributes to lethality of community-associated MRSA necrotizing pneumonia. *PLoS Pathog.* **7**, e1002271 (2011).
27. Tong, S. Y. et al. Novel staphylococcal species that form part of a *Staphylococcus aureus*-related complex: the non-pigmented *Staphylococcus argenteus* sp. nov. and the non-human primate-associated *Staphylococcus schweitzeri* sp. nov. *Int. J. Syst. Evol. Microbiol.* **65**, 15–22 (2015).
28. Thaipadungpanit, J. et al. Clinical and molecular epidemiology of *Staphylococcus argenteus* infections in Thailand. *J. Clin. Microbiol.* **53**, 1005–1008 (2015).
29. Aanensen, D. M. et al. Whole-genome sequencing for routine pathogen surveillance in public health: a population snapshot of invasive *Staphylococcus aureus* in Europe. *mBio* **7**, e00444-16 (2016).
30. Minin, V. N. & Suchard, M. A. Counting labeled transitions in continuous-time Markov models of evolution. *J. Math. Biol.* **56**, 391–412 (2008).
31. De Maio, N., Wu, C. H., O'Reilly, K. M. & Wilson, D. New routes to phylogeography: a Bayesian structured coalescent approximation. *PLoS Genet.* **11**, e1005421 (2015).
32. Sheppard, S. K. et al. Cryptic ecology among host generalist *Campylobacter jejuni* in domestic animals. *Mol. Ecol.* **23**, 2442–2451 (2014).
33. Deringer, J. R., Ely, R. J., Monday, S. R., Stauffacher, C. V. & Bohach, G. A. Vβ-dependent stimulation of bovine and human T cells by host-specific staphylococcal enterotoxins. *Infect. Immun.* **65**, 4048–4054 (1997).
34. Howden, B. P., Peleg, A. Y. & Stinear, T. P. The evolution of vancomycin intermediate *Staphylococcus aureus* (VISA) and heterogenous-VISA. *Infect. Genet. Evol.* **21**, 575–582 (2014).
35. *UK One Health Report: Antibiotics Use in Humans and Animals* (Public Health England & Veterinary Medicines Directorate, 2015).

36. Ward, M. J. et al. Time-scaled evolutionary analysis of the transmission and antibiotic resistance dynamics of *Staphylococcus aureus* CC398. *Appl. Environ. Microbiol.* **80**, 7275–7282 (2014).
37. Murray, S. et al. Recombination-mediated host adaptation by avian *Staphylococcus aureus*. *Genome Biol. Evol.* **9**, 830–842 (2017).
38. Ward, M. J. et al. Identification of source and sink populations for the emergence and global spread of the East-Asia clone of community-associated MRSA. *Genome Biol.* **17**, 160 (2016).
39. Argimon, S. et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb. Genom.* **2**, e000093 (2016).
40. Zerbino, D. R. Using the Velvet de novo assembler for short-read sequencing technologies. *Curr. Protoc. Bioinformatics* **11**, 11.5 (2010).
41. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
42. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
43. Nadalin, F., Vezzi, F. & Policriti, A. GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinform.* **13**, S8 (2012).
44. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
45. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
46. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).
47. Corander, J., Marttinen, P., Sirén, J. & Tang, J. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinform.* **9**, 539 (2008).
48. Marttinen, P. et al. Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res.* **40**, e6 (2012).
49. Ayres, D. L. et al. BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Syst. Biol.* **61**, 170–173 (2012).
51. Aibar, S., Fontanillo, C., Droste, C. & De Las Rivas, J. Functional gene networks: R/Bioc package to generate and analyse gene networks derived from functional enrichment and clustering. *Bioinformatics* **31**, 1686–1688 (2015).
52. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
53. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421 (2009).
50. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
54. Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30**, 2478–2483 (2002).
55. Contreras-Moreira, B. & Vinuesa, P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl. Environ. Microbiol.* **79**, 7696–7701 (2013).
56. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
57. Wright, D. W., Angus, T., Enright, A. J. & Freeman, T. C. Visualisation of BioPAX networks using BioLayout Express^{3D}. *F1000Res* **3**, 246 (2014).
58. Paradis, E. Analysis of diversification: combining phylogenetic and taxonomic data. *Proc. Biol. Sci.* **270**, 2499–2505 (2003).
59. David, S. et al. Evaluation of an optimal epidemiological typing scheme for *Legionella pneumophila* with whole-genome sequence data using validation guidelines. *J. Clin. Microbiol.* **54**, 2135–2148 (2016).
60. Barker, D., Meade, A. & Pagel, M. Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. *Bioinformatics* **23**, 14–20 (2007).
61. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. *System. Biol.* https://doi.org/10.1093/sysbio/syy032 (2018).
62. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
63. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612 (2006).
64. Bruen, T. & Bruen, T. *PhiPack: PHI Test and Other Tests of Recombination* (McGill University, Montreal, 2005).
65. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
66. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
67. Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22**, 2472–2479 (2005).
68. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks. *Bioinformatics* **21**, 3448–3449 (2005).
69. Supek, F., Bosnjak, M., Skunca, N. & Smuc, T. REVIGO summarizes and visualizes long lists of Gene Ontology terms. *PLoS ONE* **6**, e21800 (2011).

Acknowledgements

This study was supported by a project grant (BB/I013873/1) and institute strategic grant funding ISP2: BBS/E/D/20002173 from the Biotechnology and Biological Sciences Research Council (UK) to J.R.F., Medical Research Council (UK) grant MRNO2995X/1 to J.R.F. and Wellcome Trust collaborative award 201531/Z/16/Z to J.R.F. S.Y.C.T. is an Australian National Health and Medical Research Council Career Development Fellow (number 1065736). L.A.W. is supported by a Dorothy Hodgkin Fellowship funded by the Royal Society (grant number DH140195) and a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and Royal Society (grant number 109385/Z/15/Z). S.L. is supported by a Chancellor's Fellowship from the University of Edinburgh. M.T.G.H. was supported by the Scottish Infection Research Network and Chief Scientist Office through Scottish Healthcare Associated Infection Prevention Institute consortium funding (CSO reference: SIRN10). E.M.H. and S.J.P. were funded by The Health Innovation Challenge Fund (WT098600, HICF-T5-342), a parallel funding partnership between the Department of Health and Wellcome Trust, the UKCRC Translational Infection Research Initiative, and the Medical Research Council (grant number G1000803). S.J.P. is a National Institute for Health Research senior investigator. P.A.H. is supported by Natural Environment Research Council grant NE/M001415/1. We thank B. Blane, N. Brown and E. Torok for their role in the original study isolating and sequencing *S. aureus* from patients at the Cambridge University Hospitals NHS Foundation Trust, from which 76 genomes were downloaded from the ENA and used in this study. We also thank Edinburgh Genomics for sequencing, and all those who made isolates available for the study, including the Zoological Society London, G. Foster, H. Hasman, S. Monecke, E. Smith, D. Smyth and H. Jorgensen.

Author contributions

J.R.F., S.J.P., J.P., M.H., E.M.H., L.A.W. and M.T.G.H. conceived and designed the study. E.J.R., R.B., E.M.H., L.A.W., S.L., M.V. and K.R. carried out the experiments. E.J.R., R.B., E.M.H., L.A.W., S.L., G.K.P., D.M.A., M.T.G.H., E.J.F., J.C., M.V., P.A.H., K.R. and J.R.F. analysed the data. S.Y.C.T., A.S. and W.v.W. provided isolates. E.J.R., R.B., E.M.H., S.L. and J.R.F. wrote the manuscript. All authors contributed to editing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-018-0617-0.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.R.F.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): J. Ross Fitzgerald

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software used

Data analysis	Completed genomes downloaded from the NCBI database were converted into pseudo-fastq files using Wgsim (https://github.com/lh3/wgsim). Multiple assemblies using VelvetOptimiser v2.2.5 and Velvet v1.2. The assemblies were improved by scaffolding the best N50 and contigs using SSPACE and sequence gaps filled using GapFiller. Sequence reads were mapped to a relevant reference genome using SMALT. Consensus sequences were obtained using samtools. A maximum likelihood tree was constructed using RAxML. Time scaled trees were generated using BEAST 1.8.2. and Beagle. Pseudogenes were predicted during the PROKKA. Proteins in a genome was searched against UniProtKB (Swiss-Prot) using BLASTp 54 or UniProtKB (TrEMBL). The UniProt ID Mapping tool was used to assign Gene ontology (GO) terms to all pseudogenes. GO was assigned to all non-pseudogenes (CDS features) using InterProScan. The R package topGO with Fisher's exact test was used to identify enriched GO terms. Pan-genome association analysis were performed using Nucmer, Get_homologues, the distmat function in EMBOSS and the graph was processed in BioLayout. Identification of gene acquisitions or losses associated with host-switching events were performed using the R package APE. Codon usage bias analysis was performed using the EMBOSS tools cusp and cai. Genome-wide positive selection analysis were performed using get_homologues, MUSCLE 3.8.31, pal2nal v14, PhiPack and PAML. Functional categories were annotated using the COGs and GO databases as reference. Identification of overrepresented GO categories of positively selected genes in different hosts, we used BiNGO and REVIGO.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The sequence datasets generated during the current study are available in the European Nucleotide Archive (ENA) (www.ebi.ac.uk/ena) with the accession number PRJEB20741.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Population genomic study of bacterial evolution in the context of the host-species
Research sample	For selection of isolates, the literature was reviewed (date: November 2013) and all available S. aureus strains associated with animals and humans for which genomes had been determined were identified.
Sampling strategy	We aimed to include isolates to represent the breadth of clonal complexes, host-species diversity, geographical locations and as wide a temporal scale as possible. Publicly available sequences were selected as follows; 74 reference genomes, 302 from the EARSS project29, and 252 from other published studies of the authors. Furthermore, to be as representative of the known S. aureus host, clonal, and geographic diversity as possible we selected an additional 172 isolates for whole genome sequencing (Supplementary Table 1).
Data collection	sequencing of existing bacterial isolates
Timing and spatial scale	n/a
Data exclusions	Isolates were excluded from the analysis for the following reasons that are indicative of contamination or poor quality sequence data; a large number of contigs and a large number of 'N's in the assemblies or genome size larger than expected for S. aureus (>2.9 Mb).
Reproducibility	All findings are reproducible
Randomization	Isolates were allocated into groups according to their respective host-species. Random sub-sampling was carried out to limit the effects of sample bias.
Blinding	n/a

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging