

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

3
4 Emily J. Richardson^{1,15,16}, Rodrigo Bacigalupe^{1,15}, Ewan M. Harrison^{2,15}, Lucy A.
5 Weinert^{3,15}, Samantha Lycett¹, Manouk Vrieling¹, Kirsty Robb⁴, Paul A. Hoskisson⁴,
6 Matthew TG Holden⁵, Edward J. Feil⁶, Gavin K. Paterson⁷, Steven YC Tong⁸, Adebayo
7 Shittu⁹, Willem van Wamel¹⁰, David M. Aanensen¹¹, Julian Parkhill¹², Sharon J.
8 Peacock¹³, Jukka Corander^{12,14}, Mark Holmes³, and J. Ross Fitzgerald^{1*}
9

10 ¹The Roslin Institute, University of Edinburgh, Easter Bush Campus, Edinburgh EH25 9RG, UK;
11 ²Dept. of Medicine, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge
12 CB20QQ, UK; ³Dept. of Veterinary Medicine, University of Cambridge, Madingley Rd,
13 Cambridge CB30ES, UK; ⁴University of Strathclyde, Glasgow; ⁵School of Medicine, University of
14 St. Andrews, St Andrews KY16 9TF, UK; ⁶Milner Centre for Evolution, 4 South, University of
15 Bath, Claverton Down, Bath BA2 7AY, UK; ⁷Royal (Dick) School of Veterinary Studies,
16 University of Edinburgh, Easter Bush Campus, Edinburgh EH25 9RG; ⁸Victorian Infectious
17 Disease Service, The Royal Melbourne Hospital, and The University of Melbourne, at the Peter
18 Doherty Institute for Infection and Immunity, Victoria, Australia and Menzies School of Health
19 Research, Darwin, Australia; ⁹Department of Microbiology, Obafemi Awolowo University, Ile-Ife,
20 Nigeria; ¹⁰Department of Medical Microbiology and Infectious Diseases, Erasmus MC,
21 Rotterdam, Netherlands; ¹¹Centre for Genomic Pathogen Surveillance, Wellcome Genome
22 Campus, Cambridgeshire, CB10 1QY, UK and Dept. Infectious Disease Epidemiology, Imperial
23 College London, W2 1PG, UK; ¹²Wellcome Trust Sanger Institute, Hinxton; ¹³London School of
24 Hygiene and Tropical Medicine, London; ¹⁴Helsinki Institute for Information Technology,
25 Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland and
26 Department of Biostatistics, University of Oslo, Norway.
27

28 ¹⁵These authors contributed equally

29 ¹⁶Current address: Institute of Microbiology and Infection, University of Birmingham, Edgbaston,
30 Birmingham, B15 2TT, UK
31

32 **Running title:** *Staphylococcus aureus* host adaptation genome evolution

33 **Keywords:** Evolution; *Staphylococcus aureus*; host-species; adaptation; genomics
34

35 *Corresponding author: J. Ross Fitzgerald, The Roslin Institute, University of Edinburgh
36 e-mail: Ross.Fitzgerald@ed.ac.uk
37 Phone: +44 (0)131 6519235

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

Abstract

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-switch 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 remodeling 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.

1 Introduction

2

3 Many bacterial pathogens are host specialists that co-evolve with a single host-species.
4 However, the capacity to switch host species can provide opportunities for expansion
5 into new host populations. The domestication of animals in the Neolithic period (10,000-
6 2,000 BC approximately) and the more recent intensification of livestock farming
7 provided increased opportunities for the movement of bacterial pathogens between
8 humans and animals¹. Of note, the majority of emerging human infectious diseases
9 have been traced to an animal origin². *Staphylococcus aureus* is associated with a wide
10 spectrum of diseases in humans and strains of both methicillin-sensitive (MSSA) and
11 methicillin-resistant *S. aureus* (MRSA) are common causes of nosocomial and
12 community-acquired infection^{3,4}. In addition, *S. aureus* causes an array of infections of
13 livestock that are a major burden on the agricultural industry, including mastitis in cows,
14 sheep and goats^{5,6}, septicemia and skeletal infections in commercial broiler chickens⁷,
15 exudative epidermitis in pigs⁸ and skin abscesses and mastitis in rabbits⁹.

16

17 *S. aureus* has a clonal population structure defined by a relatively low level of
18 recombination, comprised of lineages that have single or multiple host-tropisms¹⁰⁻¹².
19 Inter-host species transmission can be of critical public health importance, as
20 exemplified by the livestock-associated methicillin-resistant clonal complex (CC) 398,
21 which is associated with pigs and other livestock, but can cause zoonotic infections of
22 pig-farmers and their contacts^{13,14}. Previous work employed multi-locus sequence typing
23 (MLST) to provide evidence for the occurrence of host-jump events from humans

1 leading to the emergence of *S. aureus* clones in livestock populations^{11,12}. More
2 recently, whole genome sequencing has been employed to investigate the evolution of
3 individual clones, providing insights into the emergence, transmission and acquisition of
4 antibiotic resistance in hospital, community, and agricultural settings^{13,15-17}. In addition, a
5 role for specific mobile genetic elements (MGEs) and core gene mutations in the host-
6 adaptation of *S. aureus* has been identified^{9,18,19}. For example, the major porcine and
7 avian clones of *S. aureus* likely originated in humans and the host-jumps were
8 associated with acquisition of MGE not found among human isolates^{13,18}. Similarly, the
9 major *S. aureus* clone associated with sheep and goats evolved through a combination
10 of gene acquisition, and allelic diversification including loss of gene function²⁰.
11 Furthermore, several studies have reported the host-specific functional activity of *S.*
12 *aureus* effectors such as leucocidins, superantigens, and the von Willebrand factor-
13 binding protein²¹⁻²⁶. In addition, it was demonstrated that for *S. aureus* strains
14 associated with natural infections of rabbits, a single mutation was responsible for
15 conferring infectivity to the progenitor strain found in human populations⁹. Taken
16 together, these studies highlight the capacity for bacteria to undergo host-switching
17 events and adapt to different species by multiple evolutionary genetic and functional
18 mechanisms. However, a large-scale, genome-based analysis of the evolutionary
19 history of *S. aureus* in the context of its host ecology is lacking, and the scale and
20 molecular basis of host-switching events remains poorly understood.

21
22 Here, we carry out a population genomic analysis of over 800 *S. aureus* isolates
23 selected to represent the known breadth of host-species diversity in order to provide a

1 high-resolution picture of the dynamics of *S. aureus* in the context of its host. The data
2 reveal the impact of human activities such as domestication and the use of antibiotics in
3 medicine and agriculture on the recent evolution of *S. aureus*, and identify the key
4 evolutionary processes underpinning its multi-host species ecology.

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

1 **Results**

2 **Extensive host-switching events define the evolutionary history of *S. aureus*.**

3 We selected *S. aureus* strains to represent the breadth of the known clonal, geographic,
4 and host-species diversity (Isolate selection details in Methods section). Overall, we
5 included 800 isolates representative of 43 different host species, 77 clonal complexes
6 (CCs), isolated in 50 different countries across 5 continents (Supplementary Figure 1-3;
7 Supplementary Table 1). Among the 800 isolates, a total of 115,149 SNPs were
8 identified in a core genome of 711,562 bp and used for reconstruction of the maximum-
9 likelihood (ML) phylogeny for the *S. aureus* species (Fig. 1). The *S. aureus* species tree
10 indicates the existence of highly divergent clades representative of the recently-
11 described *Staphylococcus argenteus* and *Staphylococcus schweitzeri* species which
12 belong to the extended *S. aureus*-related complex (Fig. 1a)²⁷. *S. argenteus*, an
13 emerging cause of human clinical infection²⁸, is more closely related to bat and monkey
14 isolates than to other human *S. aureus* sequence types (STs), consistent with a
15 possible non-human evolutionary origin for *S. argenteus*. Removal of isolates from the
16 divergent clades resulted in a phylogeny of 783 isolates that segregated according to
17 clonal complexes defined by MLST (Fig. 1b). The phylogeny indicates the broad
18 diversity of isolates of human origin with expansion of several successful epidemic
19 hospital and community-associated clones including CC22, CC30, and ST45, as
20 previously described²⁹ (Fig. 1). Animal isolates are typically found in discrete host-
21 specific clades interspersed among human lineages, consistent with ancient and recent
22 host-switching events across the phylogenetic tree (Fig. 1). In order to examine the
23 frequency and timing of host-switching events during the evolution of *S. aureus*, we

1 employed Bayesian evolutionary analysis by sampling trees (BEAST) using substitution
2 rates from published datasets (Fig. 2; Supplementary Table 2). We estimated the
3 number of cross-species transmissions for 10 major host categories (Supplementary
4 Table 3, Supplementary Figures 2-5) using BEAST with Markov Jumps³⁰. In order to
5 reduce bias caused by the larger numbers of sequences from human and cow hosts
6 compared to the other host types we used 10 stratified subsamples containing 252
7 sequences each, designed to maintain geographic, host-type and temporal diversity
8 while reducing over-representation. To assess the robustness of the main analysis, we
9 performed additional analyses as outlined in Supplementary Material (Supplementary
10 Notes; Supplementary Figures 4-11; Supplementary Tables 4-5), that included 'severe
11 balanced' subsamples of 97 taxa each containing 18-20 taxa of 5 host-types, and
12 ancestral state and host-jumps using the BASTA approximation to the structured
13 coalescent³¹. However, we had difficulty in getting BASTA to run and converge possibly
14 due to its assumptions about the structure of the data and numerical instability. Each
15 subsampled sequence set was analyzed separately within BEAST and resulted in a
16 collection of posterior trees per dataset (Supplementary Figures 6-10). In each case, the
17 analysis revealed extensive host-switching events that occurred over a time-frame
18 spanning several thousand years up to the present decade (Fig. 2a).

19
20 Our analysis identifies humans as a major donor with host-jumps identified from
21 humans into all other host-species groups examined (Fig. 2b, Supplementary Figure 4).
22 The most common recipient for *S. aureus* jumps from humans was cows with a median
23 of 14 jumps (HPD 3-22) between the years -2000 and 2012. Cows also represented a
24 major donor for host-switching events back into humans (n=10; HPD 2-26). In addition,

1 there were numerous *S. aureus* host-switches among ruminants, particularly between
2 cattle and goats in both directions and into sheep. However, host jumps from sheep into
3 other species are rare and not strongly supported by our analyses suggesting that
4 although a common host for *S. aureus*⁵, sheep do not represent a major reservoir for
5 the spread of *S. aureus* to other animals.

6

7 **Host-specific accessory gene pools promote adaptive evolution after host-**

8 **switching events.** In order to investigate the distribution of MGEs on a population level
9 across human and animal isolates, we employed a pangenome-wide association
10 analysis approach to identify genes that were enriched among isolates from specific
11 host-species. First, to account for phylogeny we removed genes identified among all
12 strains within clonal complexes associated with multiple host-species (lineage-
13 dependent genes). Network analysis indicated a remarkable correlation between
14 accessory genome and host-species revealing that diverse clonal complexes can share
15 highly similar accessory genomes that are specific for birds, pigs or horses,
16 respectively. This strongly points to the existence of a host-specific gene pool required
17 for *S. aureus* host-adaptation. Although accessory genomes of *S. aureus* obtained from
18 humans, and from cows, sheep and goats also tended to cluster together in a host-
19 specific manner, there was greater diversity in gene content (Fig. 3). This may reflect
20 the existence of multiple cryptic niches that exist within a single host-species such as
21 those proposed previously for *Campylobacter jejuni*³². We note the existence of a small
22 number of clusters made of isolates from multiple host-species. The existence of these
23 clusters suggests that some accessory gene combinations may confer a more

1 generalist host tropism with the capacity to infect multiple host-species. Alternatively,
2 insufficient time may have passed since the host-transition event for loss of dispensable
3 MGE to occur. Of note, antibiotic resistance gene determinants influenced the clustering
4 of equine and pig isolates suggesting a role for acquisition of resistance in host-
5 adaptation (Supplementary Figure 12).

6
7 Further examination of the impact of the accessory genome on successful host-
8 switching events was carried out by identifying gene acquisition or loss events that
9 correlated with host-switching events identified on the phylogeny of *S. aureus*. A total of
10 36 distinct MGEs including predicted plasmids, transposons, *S. aureus* Pathogenicity
11 Islands (SaPIs) and prophages were identified to be associated with host-switch events
12 ($p < 0.0001$) (Fig. 4a, Supplementary Table 6). Several of the MGEs have previously
13 been identified and demonstrated to encode proteins with host-specific activity. For
14 example, the β -converting phage ϕ Sa3 encodes modulators of the human innate
15 immune response, and pathogenicity islands encode superantigens or von Willebrand
16 factor-binding proteins with ruminant-specific activity^{19,33}. In addition, equine isolates
17 contain a phage encoding a novel equine allele of the staphylococcal inhibitor of
18 complement (*scn*) which also encodes the LukP/Q toxin, recently characterized to have
19 equine-specific activity^{22,25}. However, numerous uncharacterized MGEs have been
20 identified in the current study to be linked to successful host-switch events providing
21 many novel avenues for characterizing the molecular basis of *S. aureus* host-adaptation
22 (Fig 4b). For example, in isolates from pigs, a putative novel plasmid linked to *SCCmec*
23 encoding resistance to heavy metal ions, a common supplement in pig-feed, was linked

1 to host-switching events from humans into pigs (Fig. 4b). Finally, several gene clusters
2 encoding bacteriocins were enriched in isolates from specific host-species ($p < 0.0001$)
3 or were linked to host-switching events ($p < 0.0001$), consistent with the need to compete
4 with resident bacteria for survival (Supplementary Table 6). Taken together, these data
5 suggest that successful host-switch events are associated with acquisition of MGEs
6 from an accessory gene pool that exists in the recipient host-species, and/or loss of
7 MGEs linked to the source species.

8
9 In order to investigate the potential origin of MGEs horizontally acquired after a host-
10 switch event, we examined the codon usage bias of host-specific MGE, and found that
11 MGEs enriched in pig isolates had significantly elevated %GC content and reduced
12 codon adaptive index (CAI) indicative of a distinct genealogical origin (Supplementary
13 Figures 13-15). Of note, an MGE found in pig isolates had highest BLASTn similarity to
14 a putative pathogenicity island previously identified in the pig-associated zoonotic
15 pathogen *Streptococcus suis* (GC content of ~41%) (Supplementary Table 6).

16
17 **Both gain and loss of gene function are associated with *S. aureus* host-**
18 **adaptation.** Determination of the number of predicted functional genes in each *S.*
19 *aureus* genome identified a significantly higher number of genes in bird strains
20 compared to *S. aureus* from any other host-species (Supplementary Figures 16-17).
21 In contrast, the number of pseudogenes per genome is significantly higher ($p < 0.0001$ -
22 0.02) in ruminant strains compared to those from other host-species suggesting that the
23 niche occupied by *S. aureus* in cows may provide stronger selection for loss of gene

1 function compared to the niches for *S. aureus* in birds and pigs. Numerous
2 pseudogenes associated with transport of nutrients in *S. aureus* including
3 carbohydrates, are over-represented in ruminant isolates implying metabolic remodeling
4 in response to distinct nutrient availabilities in the bovine niche (Supplementary Table
5 8).

6

7 **Refinement of host adaptation involves modification of biological pathways in**
8 **response to nutrient availability.** In addition to accessory genes, adaptive mutations
9 in the core genome may be selected for in response to environmental changes such as
10 antibiotic exposure or a switch in host-species^{9,34}. In order to examine the impact of
11 host-species on diversification of the *S. aureus* core genome, we identified groups of
12 related isolates (e.g. within CCs or STs) associated with a specific host-species for
13 genome-wide analysis of positive selection (Supplementary Table 9; Supplementary
14 Figure 18). Positive selection was identified across all host-associated groups
15 examined, with an average of 68 genes (33 to 129) representing approximately 2.7%
16 (1.3% to 5.1%) of a clade-specific core genome (Supplementary Table 10). A limited
17 number of genes were under diversifying selection across multiple host species,
18 including several that encode membrane proteins, lipoproteins and a protein involved in
19 biofilm formation. Some genes were identified as undergoing positive selection in
20 distinct lineages that were associated with the same host-species (mostly human),
21 suggesting strong selective pressure leading to convergent evolution. However, for the
22 most part, our analysis detected distinct sets of genes under positive selection in
23 different lineages, suggesting that signatures of host-adaptation are dependent on the

1 genetic background of the strain, and that host-adaptation can occur via multiple
2 trajectories involving modification of distinct pathways.

3

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

16

17 **Bovine *S. aureus* strains utilize lactose with higher efficiency compared to human**
18 **or avian strains.** Considering the signatures of positive selection identified among
19 pathways associated with carbohydrate and amino acid metabolism, we investigated
20 differences in growth phenotype of selected host-specific *S. aureus* strains using a
21 metabolic phenotype microarray (Biolog), and observed preliminary strain-dependent
22 differences in growth that were influenced by the availability of specific amino acids or
23 carbohydrates. For example, *S. aureus* strains from cows had higher relative growth in

1 the presence of lactose, the primary disaccharide available in bovine milk. The genome-
2 wide positive selection analysis indicated that in bovine strains, genes associated with
3 the functional category of transport of disaccharides and oligosaccharides were
4 impacted by positive selection. To further investigate this, we carried out phenotypic
5 analysis of *S. aureus* strains from bovine, human and avian host- species of different
6 clonal complexes when grown in the presence of lactose (Fig. 5e). As lactic acid is
7 produced by *S. aureus* as a by-product of fermentation, we measured pH levels in
8 culture media containing lactose and identified a decrease in pH levels for bovine *S.*
9 *aureus* clones in comparison to human or avian clones, consistent with increased
10 efficiency of fermentation of lactose (Fig. 5f). These data support the concept that *S.*
11 *aureus* undergoes genetic diversification in response to the nutrients that differ in
12 availability in different niches.

13

14 **Resistance to antimicrobials differs among human and pig *S. aureus*.** Our
15 understanding of the relative contribution of the use of antibiotics in human medicine
16 and agriculture to the emergence of antibiotic resistance is very limited. To address this
17 question for the model human and animal pathogen *S. aureus*, we examined the
18 distribution of antibiotic, antiseptic and heavy metal-ion resistance determinants among
19 human and livestock isolates, and then accounted for phylogenetic relatedness for
20 resistance to different classes of antibiotic (Supplementary Table 12). An array of
21 resistance determinants were significantly enriched in human, ruminant, and pig
22 isolates, respectively, but not among avian isolates, consistent with a limited role for the
23 poultry industry in the emergence of antibiotic resistance in *S. aureus*.(Fig. 6; Table

1 S12). When testing for phylogenetic independence, we aimed to maximize statistical
2 power by including all gene determinants into groups specific for each class of
3 antimicrobial, and also examined selected individual determinants *str* and *sdrM*. The
4 analysis indicated that resistance to streptomycin, antiseptics, and tetracyclines were all
5 significantly associated with pig isolates, whereas *sdrM* was
6 enriched in human isolates. However, fluoroquinolone and heavy metal ion resistance
7 did not correlate with hosts after correction for phylogeny implying that expansion of
8 specific clones has contributed to the high frequency of those resistance determinants
9 among human and pig hosts, respectively. Taken together, these data demonstrate that
10 resistance to specific classes of antimicrobial in *S. aureus* is host species-dependent
11 providing evidence for distinct antibiotic selective pressures in humans and livestock. Of
12 note tetracyclines, and aminoglycosides (such as streptomycin) are used in much
13 higher amounts in farmed animals compared to human medicine³⁵. Zoonotic
14 transmission of *S. aureus* is a relatively common occurrence for some clones,
15 particularly between pigs and humans in the case of CC398, providing a route for the
16 transmission of resistant strains and associated resistance determinants to humans³⁶.

17

18 **Discussion**

19

20 Many new pathogens emerge following zoonotic or anthroponotic events providing the
21 opportunity for spread within a new host population². *S. aureus* is considered a
22 generalist bacterial species, capable of colonizing a wide range of hosts⁵. However, the
23 species is composed of distinct sub-lineages that are commonly associated with

1 particular hosts or host groups^{10,14}. Accordingly, *S. aureus* represents an excellent
2 model to explore the dynamics of a bacterial pathogen at the human-animal interface.
3 Here, we demonstrate that the segregated host-specialism of *S. aureus* arose via
4 multiple cross-species transmission events that occurred over the last 5,000-6,000
5 years, leading to the emergence of successful endemic and epidemic clones circulating
6 in distinct host-species populations. We identify humans as a major reservoir for the
7 spread of *S. aureus* to livestock, reflecting the role of humans in domestication of
8 animals, and subsequent opportunities for cross-species transmission events consistent
9 with analysis using MLST¹². Importantly, we also identify cows as the main animal
10 source for the emergence of *S. aureus* clones that are epidemic in human populations
11 consistent with a previous study that identified a bovine origin for emergent CC97
12 clones causing human infections across multiple continents¹⁷.

13

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

23

1 In addition to gene acquisition associated with host-switch events, we identified
2 evidence of adaptive evolution in the core genome consistent with host-specific
3 selective pressure driving the diversification of biological pathways that are involved in
4 survival or transmission. Furthermore, in some cases, distinct pathways were under
5 positive selective pressure in different clones associated with the same host-species,
6 implying that multiple distinct pathways may mediate host-adaptation depending on the
7 genetic background of the strain. In particular, pathways linked to carbohydrate
8 transport exhibited signatures of host-adaptation and phenotypic analysis revealed
9 enhanced utilization by bovine *S. aureus* clones of the disaccharide lactose, the major
10 carbohydrate available in bovine milk.

11

12 These findings inform a model of *S. aureus* host-adaptation in which acquisition of a
13 specific set of MGEs occur rapidly after a host-switch event (although we can't rule out
14 this could occur prior to the jump in some cases), conferring the capacity for survival in
15 the new host, largely through targeting of the innate immune response via bacterial
16 effectors such as leukocidins, superantigens and other immune-modulators. Other
17 MGEs confer resistance to antibiotics and heavy metal ions allowing survival under
18 strong anti-microbial selective pressures. Subsequently, positive selection acts on the
19 core genome via point mutation and/or recombination³⁷ causing allelic variation and
20 loss of gene function that results in modification of metabolism in response to distinct
21 nutrient availability.

22

1 Our findings suggest that since human-driven domestication, interactions with livestock
2 have provided opportunities for numerous successful host-switch events between
3 humans and livestock hosts. Further, industrialization of agriculture including use of
4 antibiotics and feed supplements in intensive farming have directly influenced the
5 evolution of *S. aureus* clones resulting in the emergence of resistance in response to
6 distinct antibiotic selective pressures in human medicine and agriculture^{18,38}. These data
7 support the idea that surveillance could play a critical role in the early identification of
8 emerging clones that have jumped host.

9
10 Taken together, our data provide a high-resolution view of the capacity for a model
11 multi-host pathogen to undergo radical changes in host ecology by genetic adaptation.
12 Investigation into the functional basis of these genetic changes will reveal key host-
13 pathogen interactions that could be targeted for novel therapies. Further, the
14 identification of the common routes for *S. aureus* livestock-human host-species
15 switches and distinct types of antimicrobial resistance in humans and livestock species
16 could inform the design of more effective farm security and antibiotic treatment practices
17 to limit the emergence of new resistant clones. These findings will be relevant to other
18 major bacterial pathogens with the capacity to spread between livestock and humans.

19

20 **Methods**

21

22 **Isolate selection.** For selection of isolates, the literature was reviewed (date:
23 November 2013) and all available *S. aureus* strains associated with animals and

1 humans for which genomes had been determined were identified. We aimed to include
2 isolates to represent the breadth of clonal complexes, host-species diversity,
3 geographical locations and as wide a temporal scale as possible (Supplementary
4 Tables 1-3). Publicly available sequences were selected as follows; 74 reference
5 genomes, 302 from the EARSS project²⁹, and 252 from other published studies of the
6 authors (Supplementary Table 1). Furthermore, to be as representative of the known *S.*
7 *aureus* host, clonal, and geographic diversity as possible we selected an additional 172
8 isolates for whole genome sequencing (Supplementary Table 1). Our dataset is biased
9 towards human isolates which represent approximately 60% of the total with 40%
10 approximately from animal sources. This reflects that fact that much of the known
11 diversity of the *S. aureus* species is of human origin¹², and also that fewer number of
12 isolates that have been obtained from animals. Given the predominant European origin
13 of the animal isolates (due to the contemporary interest in animal *S. aureus* in Europe),
14 we chose to enrich the number of human isolates with the EARSS collection
15 representative of the diversity of invasive *S. aureus* circulating among humans in
16 Europe in 2006²⁹. Accordingly, there is a European bias to the sample dataset and we
17 can't rule out that we have under-sampled the *S. aureus* diversity that exists in other
18 parts of the world. Nonetheless, our dataset contained isolates from 50 different
19 countries across 5 continents and many sequence types are widely distributed on an
20 intercontinental scale. In addition, our dataset includes isolates from the years 1930 to
21 2014, although the majority have been isolated since 2005 reflecting the greater
22 availability of recent clinical isolates (particularly from animals). It should therefore also
23 be considered that the dataset is biased towards contemporary *S. aureus* and that older

1 lineages that are now less abundant or extinct may not be represented in our dataset. In
2 order to partially address the uneven distributions of isolates by host, space and time,
3 where appropriate, we have carried out experimental replicates based on severe
4 subsampling of the dataset that provide more evenly distributed groups. In addition, we
5 have drawn conclusions that are consistent across subsampled data and, when
6 appropriate, multiple different analytic approaches. Overall, we included 800 isolates
7 representative of 43 different host species and 77 clonal complexes (CCs), isolated in
8 50 different countries across 5 continents (Supplementary Table 1). All sequences and
9 associated metadata have been uploaded to Microreact a publicly accessible database
10 that allows visualization and analysis of the data <https://microreact.org/project/shacdata>
11 ³⁹.

13 **Sequencing, genome assemblies, variant calling and phylogenetic**

14 **reconstruction.** For the current study, bacterial DNA was extracted and sequenced
15 using Illumina HiSeq2000 with 100-cycle paired-end runs at the Wellcome Trust Sanger
16 Institute or Illumina HiSeq2000 at Edinburgh Genomics. The nucleotide sequence data
17 were submitted to the European Nucleotide Archive (ENA) (www.ebi.ac.uk/ena) with the
18 accession numbers listed in Supplementary Table 1. Completed genomes downloaded
19 from the NCBI database were converted into pseudo-fastq files using Wgsim
20 (<https://github.com/lh3/wgsim>). For each isolate the sequence reads were used to
21 create multiple assemblies using VelvetOptimiser v2.2.5⁴⁰ and Velvet v1.2⁴¹. The
22 assemblies were improved by scaffolding the best N50 and contigs using SSPACE⁴²
23 and sequence gaps filled using GapFiller⁴³. Isolates were excluded from the analysis for

1 the following reasons that are indicative of contamination or poor quality sequence data;
2 a large number of contigs and a large number of 'N's in the assemblies or large genome
3 size (>2.9 Mb). Sequence types were determined from the assemblies using MLST
4 check (https://github.com/sanger-pathogens/mlst_check), which was used to compare
5 the assembled genomes against the MLST database for *S. aureus*
6 (<http://pubmlst.org/saureus/>). Sequence reads were mapped to a relevant reference
7 genome (European Nucleotide Archive (ENA) ST425 (strain LGA251, accession
8 number FR821779), using SMALT (<http://www.sanger.ac.uk/science/tools/smalt-0>)
9 following the default settings to identify single nucleotide polymorphisms (SNPs).
10 Consensus sequences were obtained using samtools and concatenated into core
11 genome alignments⁴⁴. SNPs located in mobile genetic elements were removed from the
12 alignments and a maximum likelihood tree was constructed using RAxML following
13 default settings and 1000 bootstrap replicates⁴⁵.

14

15 **Time scaled trees and estimation of the number of host jumps.** Time scaled trees
16 were generated using BEAST 1.8.2⁴⁶. All isolates with unknown date, unknown host
17 species or unknown geographical location were removed in addition to the diverse
18 BAPs groups 12 and 14 leaving a total of 696 isolates. Sites determined to be affected
19 by recombination from the BNG analysis of the individual BAPs groups, were coded as
20 missing data. Since missing data can effect phylogenetic inference and contribute to
21 heavy likelihood calculations, sites that had more than one missing state in the
22 alignment (either from missing mapped reads or recombination) were excluded from
23 further analyses, leaving a total of 55,778 sites (4306 segregating sites).

1 To account for different evolutionary processes acting at synonymous and non-
2 synonymous sites, RNA and noncoding sites, the evolutionary model was partitioned in
3 to 1+2nd sites, 3rd sites, non-coding sites, and RNAs according to the reference strain
4 LGA251. Pseudogenes were partitioned in to 1+2nd and 3rd sites with the rationale that
5 they may be functional in other isolates. For overlapping reading frames, sites were
6 assigned to the region of highest constraint (e.g. when coding and RNA, sites assigned
7 as RNA, when 1st and 3rd, sites assigned as 1+2nd etc.). For all partitions, we used an
8 HKY + Γ substitution model.

9
10 **Dating.** We treated all sequences as contemporaneous but assigned a median prior of
11 1.61 [0.604, 2.9] substitutions per site per million years on to 3rd positions, which are
12 less likely to be subject to strong purifying selection (known to affect rates over different
13 timescales). The prior comes from previous studies of *S. aureus* using tip dates on
14 different strain types (Supplementary Table 3). An uncorrelated lognormal model of
15 changes in substitution rate across different branches was employed. An initial MCMC
16 run with this model was performed in BEAST v1.8.2⁴⁶ using Beagle⁴⁷ with two
17 independent chains, removing the appropriate burnin and run for approximately
18 100,000,000 generations.

19 In addition, a 10 further subsampled datasets were produced that included only
20 sequences from the 10 major host types. These were stratified subsamples containing
21 252 samples each, designed to maintain the host species, geographic and temporal
22 diversity. The major hosts types are birds, cows, goats, carnivores, horses, humans,

1 rabbits, sheep, rodents and pigs. From these analyses, we subsampled an empirical
2 distribution of 1000 trees post burnin which were used for all further BEAST analyses.
3

4 **Markov Jump analysis.** In order to reconstruct host transition events, we used an
5 asymmetric discrete state phylogeographic analysis with Markov Jumps (REF: 51)
6 applied to the 10 major host types with default priors. We employed the Markov jump
7 analysis to estimate the posterior expectation of the number of host change events
8 across the branches of the phylogeny³⁰, using posterior sets of 1000 time scaled trees
9 from the initial BEAST analyses on the subsampled data sets. The trait models were
10 used in an MCMC chain of 1,100,000 steps, sampling every 100 steps and discarding
11 the first 10% as burn-in, leaving 10,000 trees annotated with the host information (i.e.
12 approximately 10 model instances per tree of the original posterior set). Since biased
13 sampling can lead to biased results when using these trait models, in addition to using
14 the 10 stratified subsamples of 252 sequences each we performed additional analyses
15 with host state randomization and using 100 bootstrapped maximum likelihood trees
16 (RAxML) in place of the 1000 original BEAST trees (for the 10 stratified subsamples) To
17 balance the numbers of isolates per host category further, we also created 10 'severe'
18 stratified subsamples containing 97 sequences each with 20 from Humans, Cows and
19 Sheep+Goats combined and 19 and 18 from Birds and Pigs respectively. Necessarily in
20 these severe subsamples it was not possible to maintain the full human and cattle
21 diversity, although sequences from different geographic locations and years were
22 chosen. We applied BEAST with Markov Jumps on these 5 host categories using: full
23 joint inference of trees using sequences and traits together; trees using the sequences

1 only followed by the trait mapping as before; and the BASTA structured coalescent
2 approximation ³¹.

3

4 **Pseudogene analysis.** Pseudogenes were predicted during the PROKKA annotation
5 process⁴⁸. Specifically, each protein in a genome was searched against UniProtKB
6 (Swiss-Prot) using BLASTp⁴⁹ or UniProtKB (TrEMBL). If no significant hits were
7 identified, proteins were examined for conserved motifs. Any proteins exhibited less
8 than 95% coverage of their top hit were listed as potential pseudogenes. The region of
9 the top hit that was not present in the protein sequence was then interrogated against
10 all contigs using BLASTn⁴⁹. Hits that were in the correct orientation and on the same
11 contigs were accepted as pseudogenes and labelled according to their type (frameshift,
12 stop codon, insertion). Proteins that were less 95% coverage of their top hit and on the
13 edge of a contig with their counterpart on another contigs were not labelled as
14 pseudogenes, rather CDS that have split due to the assembly breaking at this point.

15

16 The UniProt ID Mapping tool was used to assign Gene ontology (GO) terms to all
17 pseudogenes by transferring the GO terms assigned to the closest reference (identified
18 during the annotation process described above). GO was assigned to all non-
19 pseudogenes (CDS features) using the same method and InterProScan⁵⁰. The R
20 package topGO⁵¹ with Fisher's exact test was used to identify enriched GO terms whilst
21 taking into account the GO hierarchy (the p-value was adjusted using Bonferroni
22 correction).

23

1 **Pan-genome association analysis.** All genomes in this study were organised into a list
2 of reference genomes followed by assembled contigs. The second genome in the list
3 was aligned to the first genome using Nucmer⁵² and any regions larger than 100 bp that
4 did not map to the first genome were appended to the end of it to produce a pangenome
5 representing the unique regions in the first two genomes. Each subsequent genome
6 aligned to the combination of all unique regions from the previously aligned genomes in
7 the list producing a pangenome that represents all of the nucleotide sequences of all
8 genomes. All genes were organised into groups of orthologues using the bi-directional
9 best hits algorithm in Get_homologues with a minimum coverage setting of 50% and
10 minimum sequence identity setting of 80%⁵³. The pangenome was used as the
11 reference and the coding sequences (CDS) predicted in the annotations described
12 previously were compared to all CDS within the pangenome. Features annotated as
13 pseudogenes were excluded from this analysis. The Get_homologues
14 compare_clusters perl script was used to create a pangenome matrix of all identified
15 gene clusters against all genomes. All core gene clusters (clusters that contain genes
16 from every genome) were removed from the pangenome matrix. Further to this all
17 clusters that only contained genes from one genome or all genomes except one were
18 removed. Furthermore, gene clusters that were found in all members of any STs
19 associated with multiple host species were removed on the basis that they are not
20 specific to a single host species. This has the effect of removing lineage-associated
21 genes resulting in a set of gene clusters that are strain-dependent and largely
22 independent of phylogeny. Hypergeometric testing was used to find over- and under-
23 represented gene clusters for each host (the p-value was adjusted using Bonferroni

1 correction). All gene clusters were searched against the NCBI non-redundant nucleotide
2 database using blastn to provide the most up to date annotation and to examine the
3 likely bacterial species origin of each MGE.

4
5 A pairwise distance matrix was calculated from the pangenome matrix using the distmat
6 function in EMBOSS⁵⁴. The matrix was converted into a bidirectional graph with
7 distance as the edge weight parameter. The graph was processed in BioLayout with an
8 edge weight threshold of 0.5⁵⁵.

9 10 **Identification of gene acquisitions or losses associated with host-switching**

11 **events.** The R package APE (Analysis of Phylogenetics and Evolution)⁵⁶ was used to fit
12 a single discrete trait model and get the ancestral state of each node for each gene
13 cluster against the phylogenetic tree. From this, a vector for every gene cluster was
14 created with gene acquisition/loss events by comparing every child node in the tree to
15 its parent node to determine if there was no change, a gene acquisition or a gene loss
16 event. This was performed separately for each host type (i.e. human, ruminant, bird,
17 horse and pig) to identify which nodes are associated with a host-switching event. All
18 gene state vectors were compared to all host state vectors using a Fisher's exact test to
19 show whether a gene loss/acquisition even is related to a host switch event. The p-
20 values were adjusted using Bonferroni correction.

21
22 **Codon usage bias analysis.** The codon adaptation index is used to calculate codon
23 usage bias by comparing the CAI of a gene against the codon usage table of a

1 reference set of genes. The codon usage table was calculated using the EMBOSS tool
2 cusp⁵⁴. For this study, the codon usage table was comprised of all genes that were not
3 significantly over represented in a host or significantly associated with a host switch.
4 The codon adaptation index (CAI) for all genes significantly over represented in a host
5 and significantly associated with a host jump was calculated using the EMBOSS tool
6 cai⁵⁴. The codon adaptation index was also calculated using five random subsets of 50
7 genes as controls. A one-way ANOVA test was used to test whether there was a
8 significant effect of host upon CAI. A Tukey HSD test was then applied to compare the
9 CAIs between host species.

10

11 **Distribution of antibiotic resistance genes analysis.** Antimicrobial resistance genes
12 were identified as described by Holden et al¹⁵. Resistance genes were identified by a
13 combination of BlastN and mapping against assemblies and as previously described ⁵⁷
14 and resistance SNPs identified by mapping against a pseudomolecule of genes with
15 previously reported resistance-conferring mutations. Isolates were grouped into human
16 isolates and all animals and them human, rabbits, companion animals (horses, dogs,
17 cats), marine, pigs, primates, ruminants (goats, sheep, cows) and small mammals (rats,
18 mice, other small mammals). The proportions of isolates with each resistance gene and
19 ≥ 1 resistance conferring SNP for each antibiotic was compared to identify enrichment
20 using a two-tailed Fishers Exact test with a Bonferroni correction for multiple testing.
21 Determinates with a P value $< 9.9 \times 10^{-5}$ were considered statistically significantly
22 enriched. To examine whether the Fisher Exact tests of independence were robust
23 when accounting for population structure, we tested whether resistance phenotypes and

1 host were correlated across the *S. aureus* phylogeny. We conducted these for
2 pig/human and ruminant/human since these were the only comparisons where
3 significant differences were observed according to the Fisher's exact test. In order to
4 maximise statistical power, we grouped all gene determinants into specific classes of
5 antimicrobial (e.g. Tetracycline-resistant if encoding any *tet* allele) and tested for
6 correlation with host species using the program BayesTraits⁵⁸ (using the posterior
7 sample of trees from our earlier BEAST analysis). We note that the correlated
8 evolutionary analysis may be overly conservative in cases where horizontal gene
9 transfer is rampant or homoplasies are high. BayesTraits uses a continuous-time
10 Markov model to estimate transition rates between the presence and absence of a gene
11 or SNP and between human and non-human hosts. We allowed the transition rates to
12 evolve in either a correlated fashion (where the rate of change in one trait depends on
13 the state found in the other trait) or independently. Posterior distributions of parameters
14 were estimated from up to 4 million iterations of the MCMC with default priors. After
15 discarding burn-in, the marginal likelihoods of the dependent and independent models
16 were obtained using the Akaike Information Criterion (AICM) estimated using the
17 methods-of-moment estimator in Tracer 1.6⁵⁹

18
19 **Genome-wide positive selection analysis.** To identify genes under positive selection
20 in different host groups, we first identified lineages (STs or CCs) correlated with
21 particular hosts. As the power of the selection analysis is determined by the number of
22 isolates included, only clades with more than 10 isolates associated with a host were
23 considered. Based on these criteria, 15 CCs from four groups of hosts were analyzed: 9

1 for humans (CC30, CC5, CC59, CC15, CC12, ST239, ST8, CC22 and CC45), 3 for
2 ruminants (CC133, primarily associated with sheep and goats and the cows related
3 CC151 and CC97), 2 for birds (CC5 and CC385) and one for pigs (CC398)
4 (Supplementary Table 7). Although the CC398 clade also contained several human
5 isolates, these mostly represent spill-over events rather than an established association
6 so the CC398-human group was not included for the analysis. Given the variable
7 number of isolates of each CC-host group, in order to standardize the analysis while
8 preventing the underestimation of genes under positive selection, 10 isolates linked with
9 a host were analyzed at a time. Replicates or triplicates of different subsets of genomes
10 using sampling with replacement was carried out if the number of isolates for that
11 lineage was large enough. Next, we identified orthologous genes in each of these
12 groups using the algorithm OrthoMCL integrated in get_homologues (identity >70%,
13 similarity >75%, f50, e-value = $1e-5$)⁵³. Genes were considered orthologous if they were
14 present in at least 70% of the genomes. Since alignment of coding DNA sequences
15 may insert gaps in codons and produce frame-shifts, we aligned genes at the protein
16 level using MUSCLE 3.8.31⁶⁰ and translated these sequences back to DNA using
17 pal2nal v14⁶¹. Genes identified as inparalogous that turned out to be duplications were
18 kept for further analyses, otherwise discarded. For every alignment, recombination was
19 detected using the NSS, Max Chi and Phi tests included in PhiPack⁶² and recombinant
20 genes removed from further analyses. For the gene clusters containing 10 isolates,
21 phylogenetic trees were extracted from the 783 isolates ML tree. For clusters with less
22 than 10 genomes, subtrees were produced from the general tree using the tree prune
23 function in ete2⁶³. The DNA alignments and trees were used for PAML analysis⁶⁴. We

1 employed the site evolution models of Codeml (M1a, M2a, M7, M8 and M8a) to perform
2 codon-by-codon analysis of dN/dS ratios (nonsynonymous to synonymous substitution,
3 ω) of genes and a likelihood ratio test (LRT) was used to determine significant
4 differences between nested models M1a-M2a, M7-M8, M8a-M8, where one accounts
5 for positive selection (alternative hypothesis) and the other specifies a neutral model
6 (null hypothesis). Statistic tests were assessed to a chi-square distribution with 2 and 1
7 degrees of freedom⁶⁴. Bayes Empirical Bayes⁶⁵ was used to calculate the posterior
8 probabilities of amino acid sites under positive selection of proteins that had significant
9 LRTs. As independent replicates from similar CC/Host groups resulted in slightly
10 different genes positively selected, we used get_homologues to merge the core
11 genomes and genes selected for each group using same parameters as above. Genes
12 under positive selection were considered when they were in common for different
13 replicates with a p-value of 0.05 or were identified in different replicates with a stringent
14 p-value (0.05/number of genes per core genome).

15 To explore functional categories under positive selection we performed classification of
16 Clusters of Orthologous Groups (COGs), annotated Gene Ontology terms (GO) and
17 analysed metabolic pathways (KEGG). To assign COG terms, we performed BLASTp of
18 single representatives of the orthologous clusters against the prot2003-2014 database,
19 retrieving the top 5 hits to include alternative annotations. We mapped the gene IDs
20 obtained to the cog2003-2014.csv database from which the COGs were inferred.

21 Frequencies of COGs for positively selected genes in each CC-host were compared
22 with the average COG frequencies in the respective core genomes. GO annotations
23 were obtained by mapping the genes to the go_20151121-seqdb, uniprot_sprot and

1 uniprot_trembl databases using BLASTp. From these, the UniProtKB were mapped to
2 the gene_association_goa database and filtered by bacteria domain to obtain the GO
3 categories. To visualize and identify overrepresented GO categories of positively
4 selected genes in different hosts, we used BiNGO⁶⁶. We identified overrepresented
5 categories using the hypergeometric test with the Benjamini & Hochberg False
6 Discovery Rate (FDR) multiple testing correction at a significance level of 5%. We
7 chose the 'Biological Process' category and the prokaryotic ontology file
8 (gosubset_prok.obo). However, as most groups did not show significant
9 overrepresentation, we visualized all the GO categories of genes under positive
10 selection and used REVIGO⁶⁷ with the p-values from BiNGO in order to obtain
11 summaries of non-redundant GO terms classified into functional categories.

12
13 **Analysis of lactose fermentation.** *S. aureus* was cultured in Tryptic Soy Broth (TSB) in
14 presence or absence of 100 mM lactose at 37°C for 17 h with shaking at 200 rpm. OD₆₀₀
15 was measured and culture supernatants were collected by centrifugation. Subsequently,
16 the pH of the supernatants was measured using a pH meter (Sartorius, UK). Delta pH
17 values were calculated by subtracting the pH values of TSB cultures supplemented with
18 100 mM lactose from the pH values of normal TSB cultures. Statistical analysis was
19 performed in Graphpad Prism 7 using One-Way ANOVA followed by Tukey's multiple
20 comparison test.

21
22 **Data availability.** The sequence datasets generated during the current study are
23 available in the European Nucleotide Archive (ENA) (www.ebi.ac.uk/ena) with the

1 accession number PRJEB20741. Accession numbers of previously published
2 sequences analyzed in the current study are listed in Supplementary Table 1. All data
3 analysed during this study are included in this published article (and its supplementary
4 information files).

5 **References**

- 6
7
- 8 1. Morand, S., McIntyre, K.M. & Baylis, M. Domesticated animals and human infectious
9 diseases of zoonotic origins: domestication time matters. *Infect Genet Evol* **24**, 76-81
10 (2014).
 - 11 2. Woolhouse, M.E., Haydon, D.T. & Antia, R. Emerging pathogens: the epidemiology and
12 evolution of species jumps. *Trends Ecol Evol* **20**, 238-44 (2005).
 - 13 3. Lowy, F.D. Staphylococcus aureus Infections. *New England Journal of Medicine* **339**,
14 520-532 (1998).
 - 15 4. Chambers, H.F. & Deleo, F.R. Waves of resistance: Staphylococcus aureus in the
16 antibiotic era. *Nat Rev Microbiol* **7**, 629-41 (2009).
 - 17 5. Peton, V. & Le Loir, Y. Staphylococcus aureus in veterinary medicine. *Infect Genet Evol*
18 **21**, 602-15 (2014).
 - 19 6. Bradley, A.J., Leach, K.A., Breen, J.E., Green, L.E. & Green, M.J. Survey of the
20 incidence and aetiology of mastitis on dairy farms in England and Wales. *Vet Rec* **160**,
21 253-7 (2007).
 - 22 7. McNamee, P.T. & Smyth, J.A. Bacterial chondronecrosis with osteomyelitis ('femoral
23 head necrosis') of broiler chickens: a review. *Avian Pathol* **29**, 477-95 (2000).
 - 24 8. van Duijkeren, E. *et al.* Methicillin-resistant Staphylococcus aureus in pigs with exudative
25 epidermitis. *Emerg Infect Dis* **13**, 1408-10 (2007).
 - 26 9. Viana, D. *et al.* A single natural nucleotide mutation alters bacterial pathogen host
27 tropism. *Nat Genet* **47**, 361-6 (2015).
 - 28 10. Feil, E.J. *et al.* How clonal is Staphylococcus aureus? *J Bacteriol* **185**, 3307-16 (2003).
 - 29 11. Shephard, M.A. *et al.* Historical zoonoses and other changes in host tropism of
30 Staphylococcus aureus, identified by phylogenetic analysis of a population dataset.
31 *PLoS One* **8**, e62369 (2013).
 - 32 12. Weinert, L.A. *et al.* Molecular dating of human-to-bovid host jumps by Staphylococcus
33 aureus reveals an association with the spread of domestication. *Biol Lett* **8**, 829-32
34 (2012).
 - 35 13. Price, L.B. *et al.* Staphylococcus aureus CC398: host adaptation and emergence of
36 methicillin resistance in livestock. *mBio* **3**(2012).
 - 37 14. Fitzgerald, J.R. Livestock-associated Staphylococcus aureus: origin, evolution and
38 public health threat. *Trends Microbiol* **20**, 192-8 (2012).
 - 39 15. Holden, M.T. *et al.* A genomic portrait of the emergence, evolution, and global spread of
40 a methicillin-resistant Staphylococcus aureus pandemic. *Genome Res* **23**, 653-64
41 (2013).
 - 42 16. McAdam, P.R. *et al.* Molecular tracing of the emergence, adaptation, and transmission
43 of hospital-associated methicillin-resistant Staphylococcus aureus. *Proc Natl Acad Sci U*
44 *S A* **109**, 9107-12 (2012).

- 1 17. Spoor, L.E. *et al.* Livestock origin for a human pandemic clone of community-associated
2 methicillin-resistant *Staphylococcus aureus*. *MBio* **4**(2013).
- 3 18. Lowder, B.V. *et al.* Recent human-to-poultry host jump, adaptation, and pandemic
4 spread of *Staphylococcus aureus*. *Proc Natl Acad Sci U S A* **106**, 19545-50 (2009).
- 5 19. Viana, D. *et al.* Adaptation of *Staphylococcus aureus* to ruminant and equine hosts
6 involves SaPI-carried variants of von Willebrand factor-binding protein. *Mol Microbiol* **77**,
7 1583-94 (2010).
- 8 20. Guinane, C.M. *et al.* Evolutionary genomics of *Staphylococcus aureus* reveals insights
9 into the origin and molecular basis of ruminant host adaptation. *Genome Biol Evol* **2**,
10 454-66 (2010).
- 11 21. Koymans, K.J., Vrieling, M., Gorham, R.D., Jr. & van Strijp, J.A. Staphylococcal Immune
12 Evasion Proteins: Structure, Function, and Host Adaptation. *Curr Top Microbiol Immunol*
13 (2016).
- 14 22. Koop, G. *et al.* Identification of LukPQ, a novel, equid-adapted leukocidin of
15 *Staphylococcus aureus*. *Sci Rep* **7**, 40660 (2017).
- 16 23. Loffler, B. *et al.* *Staphylococcus aureus* panton-valentine leukocidin is a very potent
17 cytotoxic factor for human neutrophils. *PLoS Pathog* **6**, e1000715 (2010).
- 18 24. Vrieling, M. *et al.* LukMF¹ is the major secreted leukocidin of bovine *Staphylococcus*
19 *aureus* and is produced in vivo during bovine mastitis. *Sci Rep* **6**, 37759 (2016).
- 20 25. de Jong, N.W.M. *et al.* Identification of a staphylococcal complement inhibitor with broad
21 host specificity in equid *Staphylococcus aureus* strains. *J Biol Chem* **293**, 4468-4477
22 (2018).
- 23 26. Wilson, G.J. *et al.* A novel core genome-encoded superantigen contributes to lethality of
24 community-associated MRSA necrotizing pneumonia. *PLoS Pathog* **7**, e1002271 (2011).
- 25 27. Tong, S.Y. *et al.* Novel staphylococcal species that form part of a *Staphylococcus*
26 *aureus*-related complex: the non-pigmented *Staphylococcus argenteus* sp. nov. and the
27 non-human primate-associated *Staphylococcus schweitzeri* sp. nov. *Int J Syst Evol*
28 *Microbiol* **65**, 15-22 (2015).
- 29 28. Thaipadungpanit, J. *et al.* Clinical and molecular epidemiology of *Staphylococcus*
30 *argenteus* infections in Thailand. *J Clin Microbiol* **53**, 1005-8 (2015).
- 31 29. Aanensen, D.M. *et al.* Whole-Genome Sequencing for Routine Pathogen Surveillance in
32 Public Health: a Population Snapshot of Invasive *Staphylococcus aureus* in Europe.
33 *MBio* **7**(2016).
- 34 30. Minin, V.N. & Suchard, M.A. Counting labeled transitions in continuous-time Markov
35 models of evolution. *J Math Biol* **56**, 391-412 (2008).
- 36 31. De Maio, N., Wu, C.H., O'Reilly, K.M. & Wilson, D. New Routes to Phylogeography: A
37 Bayesian Structured Coalescent Approximation. *PLoS Genet* **11**, e1005421 (2015).
- 38 32. Sheppard, S.K. *et al.* Cryptic ecology among host generalist *Campylobacter jejuni* in
39 domestic animals. *Mol Ecol* **23**, 2442-51 (2014).
- 40 33. Deringer, J.R., Ely, R.J., Monday, S.R., Stauffacher, C.V. & Bohach, G.A. Vbeta-
41 dependent stimulation of bovine and human T cells by host-specific staphylococcal
42 enterotoxins. *Infect Immun* **65**, 4048-54 (1997).
- 43 34. Howden, B.P., Peleg, A.Y. & Stinear, T.P. The evolution of vancomycin intermediate
44 *Staphylococcus aureus* (VISA) and heterogenous-VISA. *Infect Genet Evol* **21**, 575-82
45 (2014).
- 46 35. Directorate, P.H.E.a.V.M. UK One Health report: antibiotics use in humans and animals.
47 (2015).
- 48 36. Ward, M.J. *et al.* Time-scaled evolutionary analysis of the transmission and antibiotic
49 resistance dynamics of *Staphylococcus aureus* CC398. *Appl Environ Microbiol* (2014).
- 50 37. Murray, S. *et al.* Recombination-Mediated Host Adaptation by Avian *Staphylococcus*
51 *aureus*. *Genome Biol Evol* **9**, 830-842 (2017).

- 1 38. Ward, M.J. *et al.* Identification of source and sink populations for the emergence and
2 global spread of the East-Asia clone of community-associated MRSA. *Genome Biol* **17**,
3 160 (2016).
- 4 39. Argimon, S. *et al.* Microreact: visualizing and sharing data for genomic epidemiology and
5 phylogeography. *Microb Genom* **2**, e000093 (2016).
- 6 40. Zerbino, D.R. Using the Velvet de novo assembler for short-read sequencing
7 technologies. *Curr Protoc Bioinformatics* **Chapter 11**, Unit 11 5 (2010).
- 8 41. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de
9 Bruijn graphs. *Genome Res* **18**, 821-9 (2008).
- 10 42. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. & Pirovano, W. Scaffolding pre-
11 assembled contigs using SSPACE. *Bioinformatics* **27**, 578-9 (2011).
- 12 43. Nadalin, F., Vezzi, F. & Policriti, A. GapFiller: a de novo assembly approach to fill the
13 gap within paired reads. *BMC Bioinformatics* **13 Suppl 14**, S8 (2012).
- 14 44. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
15 2078-9 (2009).
- 16 45. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
17 large phylogenies. *Bioinformatics* **30**, 1312-3 (2014).
- 18 46. Drummond, A.J., Suchard, M.A., Xie, D. & Rambaut, A. Bayesian phylogenetics with
19 BEAUti and the BEAST 1.7. *Mol Biol Evol* **29**, 1969-73 (2012).
- 20 47. Ayres, D.L. *et al.* BEAGLE: an application programming interface and high-performance
21 computing library for statistical phylogenetics. *Syst Biol* **61**, 170-3 (2012).
- 22 48. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-9
23 (2014).
- 24 49. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421
25 (2009).
- 26 50. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification.
27 *Bioinformatics* **30**, 1236-40 (2014).
- 28 51. Aibar, S., Fontanillo, C., Droste, C. & De Las Rivas, J. Functional Gene Networks:
29 R/Bioc package to generate and analyse gene networks derived from functional
30 enrichment and clustering. *Bioinformatics* **31**, 1686-8 (2015).
- 31 52. Delcher, A.L., Phillippy, A., Carlton, J. & Salzberg, S.L. Fast algorithms for large-scale
32 genome alignment and comparison. *Nucleic Acids Res* **30**, 2478-83 (2002).
- 33 53. Contreras-Moreira, B. & Vinuesa, P. GET_HOMOLOGUES, a versatile software
34 package for scalable and robust microbial pangenome analysis. *Appl Environ Microbiol*
35 **79**, 7696-701 (2013).
- 36 54. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open
37 Software Suite. *Trends Genet* **16**, 276-7 (2000).
- 38 55. Wright, D.W., Angus, T., Enright, A.J. & Freeman, T.C. Visualisation of BioPAX
39 Networks using BioLayout Express (3D). *F1000Res* **3**, 246 (2014).
- 40 56. Paradis, E. Analysis of diversification: combining phylogenetic and taxonomic data. *Proc*
41 *Biol Sci* **270**, 2499-505 (2003).
- 42 57. David, S. *et al.* Evaluation of an Optimal Epidemiological Typing Scheme for Legionella
43 pneumophila with Whole-Genome Sequence Data Using Validation Guidelines. *J Clin*
44 *Microbiol* **54**, 2135-48 (2016).
- 45 58. Barker, D., Meade, A. & Pagel, M. Constrained models of evolution lead to improved
46 prediction of functional linkage from correlated gain and loss of genes. *Bioinformatics* **23**,
47 14-20 (2007).
- 48 59. Rambaut A, S.M., Xie D, and Drummond AJ. Tracer v 1.6. (2014).
- 49 60. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high
50 throughput. *Nucleic Acids Res* **32**, 1792-7 (2004).

- 1 61. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence
2 alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**, W609-12
3 (2006).
- 4 62. Bruen, T., Bruen, T. PhiPack: PHI test and other tests of recombination. (McGill
5 University, Montreal, Quebec, 2005).
- 6 63. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and
7 Visualization of Phylogenomic Data. *Mol Biol Evol* **33**, 1635-8 (2016).
- 8 64. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-
9 91 (2007).
- 10 65. Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an improved branch-site likelihood
11 method for detecting positive selection at the molecular level. *Mol Biol Evol* **22**, 2472-9
12 (2005).
- 13 66. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess
14 overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**,
15 3448-9 (2005).
- 16 67. Supek, F., Bosnjak, M., Skunca, N. & Smuc, T. REVIGO summarizes and visualizes
17 long lists of gene ontology terms. *PLoS One* **6**, e21800 (2011).
- 18 68. Reuter, S. *et al.* Building a genomic framework for prospective MRSA surveillance in the
19 United Kingdom and the Republic of Ireland. *Genome Res* (2015).
- 20

21 **Acknowledgments**

22 The study was supported by a project grant (BB/K00638X/1) and institute strategic grant
23 funding ISP2: BB/P013740/1 from the Biotechnology and Biological Sciences Research
24 Council (UK) to J.R.F, Medical Research Council (UK) grant MRNO2995X/1 to J.R.F.
25 and Wellcome Trust collaborative award 201531/Z/16/Z to J.R.F. S.Y.C.T. is an
26 Australian National Health and Medical Research Council Career Development Fellow
27 (#1065736). L.A.W is supported by a Dorothy Hodgkin Fellowship funded by the Royal
28 Society (Grant Number DH140195) and a Sir Henry Dale Fellowship jointly funded by
29 the Wellcome Trust and the Royal Society (Grant Number 109385/Z/15/Z). S.L. is
30 supported by a Chancellor's Fellowship from the University of Edinburgh. M.T.G.H was
31 supported by the Scottish Infection Research Network and Chief Scientist Office
32 through the Scottish Healthcare Associated Infection Prevention Institute consortium
33 funding (CSO Reference: SIRN10). E.M.H. and S.J.P were funded by The Health
34 Innovation Challenge Fund (WT098600, HICF-T5-342), a parallel funding partnership
35 between the Department of Health and Wellcome Trust, and the UKCRC Translational
36 Infection Research Initiative, and the Medical Research Council (Grant Number
37 G1000803). S.J.P. is a National Institute for Health Research Senior Investigator.

1 M.T.G.H was supported by the Scottish Infection Research Network and Chief Scientist
2 Office through the Scottish Healthcare Associated Infection Prevention Institute
3 consortium funding (CSO Reference: SIRN10) M.T.G.H was supported by the
4 Scottish Infection Research Network and Chief Scientist Office through
5 the Scottish Healthcare Associated Infection Prevention Institute consortium
6 funding (CSO Reference: SIRN10). P.A.H. is supported by Natural Environment Research
7 Council for Grant NE/M001415/1. We thank Beth Blane, Nick Brown and Estee Torok for
8 their role in the original study that isolated and sequenced *S. aureus* from patients at the
9 Cambridge University Hospitals NHS Foundation Trust ⁶⁸, from which 76 genomes were
10 downloaded from the ENA and used in this study. We also thank Edinburgh Genomics
11 for sequencing, and all those who made isolates available for the study including
12 Zoological Society London, G. Foster, H. Hasman, S. Monecke, E. Smith, D. Smyth,
13 and H. Jorgensen.

14

15 **Author contributions**

16 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
17 study. E.J.R., R.B., E.M.H., L.A.W., S.L , M.V. and K.R. carried out experiments. E.J.R.,
18 R.B., E.M.H., L.A.W., S.L., G.K.P, D.M.A., M.T.H., E.J.F., J.C., M.V., P.A.H., K.R. and
19 J.R.F analysed data. S.Y.C.T., A.S., and W.vW. provided isolates. E.J.R., R.B., E.M.H.,
20 S.L. and J.R.F. wrote the manuscript. All authors contributed to manuscript editing.

21

22 **Competing Interests**

23

24 The authors declare no competing interests.

25

26

27 **Figure legends**

28

29 **Figure 1. *Staphylococcus aureus* phylogeny according to host-species origin.** (a)
30 Phylogenetic tree of 800 isolates constructed using the maximum likelihood (ML) method
31 demonstrating the relationship between *S. aureus* and other members of the *S. aureus*
32 complex; *S. schweitzeri* and *S. argenteus*; (b) Phylogenetic (ML) tree of 783 *S. aureus*
33 isolates, with host species indicated in colour or animal symbols indicating major

1 domesticated animal clones that are largely host-specific. The evolutionary history of *S.*
2 *aureus* was calibrated using well-established substitution rates from published datasets
3 (see methods).

4
5 **Figure 2. *S. aureus* has undergone extensive ancient and recent host-switching**
6 **events with humans acting as a major hub.**

7 (a) Time-scaled phylogeny of a
8 subsample of the *S. aureus* sequences with clonal complexes (CCs) labeled, branches
9 colored according to host-species group. Pie charts indicate relative probability of host
10 origin at the ancestral nodes, and line thickness corresponds to probability of the
11 majority host (see Supplementary Figures 6-10 for all subsamples). Major clonal
12 complexes (CC)s are indicated. (b) Quantification of the number of host-switch events:
13 Host transition count network from BEAST Markov Jumps models averaged over all
14 subsamples of the data. Line-width represents the average Markov Jump count per
15 tree, averaged over all subsamples (Supplementary Figures 4-5) and line color
16 represents the significance compared to permuted label analysis (Z-score). Only
17 transitions with higher counts compared to models with permuted host-labels are shown
18 (Z-score ≥ 0.5).

19 **Figure 3. Network analysis of *S. aureus* accessory genome indicates clustering**
20 **according to host species group.**

21 Network graph of pairwise distances of accessory
22 genome gene content between isolates. Each node represents an isolate, colour-coded
23 to indicate host species origin, and each edge indicates greater than 50% of shared
24 accessory genome content with the length of the edges weighted by distance
25 (proportion of shared accessory genes; shorter edges have more genes in common). All
26 edges with $<50\%$ shared accessory genome content were removed.

27 **Figure 4. Identification of horizontally-acquired genetic elements correlated with**
28 **host-adaptation.**

29 (a) Schematic representation of the *S. aureus* pan-genome with gene
30 clusters linked to host-species indicated by shading. Coloured symbols indicate the
31 nature of the mobile genetic element and the associated host-species (b) Annotated
gene maps of selected novel genetic elements linked to specific host species.

1 associated with acquisition of MGEs from an accessory gene pool that exists in the
2 recipient host-species, and/or loss of MGEs linked to the source species.

3

4 **Figure 5. Summary of biological pathways under positive selection in different**
5 **host-species and evidence for phenotypic adaptation.** The main anatomical
6 isolation sites on each host group are indicated by filled circles. Functional categories
7 *virulence and pathogenesis, resistance to antibiotics, transport of ions and cell wall*
8 *biosynthesis* were under positive selection in all 4 host-species groups. In humans (a)
9 and ruminants (b) the categories *amino acids biosynthesis* and *transport/metabolism of*
10 *carbohydrates* were positively selected. The categories *amino-acid*
11 *transport/metabolism* and *biosynthesis of osmoprotectants* were under positive
12 selection in birds (c) and *transposable elements* in pigs (d). (e) Phylogenetic tree
13 indicating the distinct lineages selected for comparative analysis of lactose
14 fermentation, (f) Fermentation of the disaccharide lactose is enhanced in bovine
15 lineages. Acidification of *S. aureus* culture supernatant in presence of 100 mM lactose
16 as indicated by the delta pH. Experiments were performed in triplicate with 5 strains per
17 clonal lineage. Each dot represents the average delta pH per strain and bars indicate
18 the SEM per clonal lineage (n = 5). Asterisks indicate significant differences between
19 bovine (CC97 and CC151, n = 10), avian (CC5 and CC385, n = 10) and human
20 lineages (CC22, CC30, and CC8, n = 15) with **P<0.005, ***P<0.001 and ****P<0.0001
21 using One-Way ANOVA followed by Tukey's multiple comparison test.

22

23 **Figure 6. Resistance to antimicrobials is non-randomly associated with host**
24 **species.** Proportion (%) of isolates examined which contain the specified resistance
25 determinant (Supplementary Table 12). Asterisks indicate significant association of
26 resistance determinants with host-species (Fisher exact test), and colored borders
27 indicate antibiotic class or single determinants (*sdrM* and *str*) that are associated with
28 host-species group after testing for phylogenetic independence.

29