# Global phylogenomics of multidrug-resistant *Staphylococcus aureus* sequence type 772: the Bengal Bay clone

Steinig E.J.[1,2], Duchene S.[3], Robinson D.A.[4], Monecke S.[5,6,7], Yokoyama M.[8], Laabei M.[8], Slickers P.[5,6], Andersson P.[1], Williamson D.[9], Kearns A.[10], Goering R.[11], Dickson E.[12], Ehricht R.[5,6], Ip M.[13], O'Sullivan M.V.N.[14], Coombs G.W.[15], Petersen A.[16], Brennan G.[17], Shore A.C.[18], Coleman D.C.[18], Pantosti A.[19], de Lencastre H.[20,21], Westh H.[22,23], Kobayashi N.[24], Heffernan H.[25], Strommenger B.[26], Layer F.[26], Weber S.[27], Aamot H.[28], Skakni L.[29], Peacock S.J.[30], Sarovich D.[1,31], Giffard P.[1,32], Harris S.[33], Parkhill J.[33], Massey R.C.[34], Holden M.T.G.[33,35], Bentley S.D.[33], and Tong S.Y.C.[1,36,*]

[1]*Menzies School of Health Research, Darwin, Australia,* [2]*Australian Institute of Tropical Health and Medicine, Townsville, Australia,* [3]*Department of Biochemistry & Molecular Biology, University of Melbourne, Melbourne, Australia,* [4]*University of Mississippi Medical Center, Jackson, United States,* [5]*Abbott (Alere Technologies GmbH), Jena, Germany,* [6]*InfectoGnostics Research Campus, Jena, Germany,* [7]*Technical University of Dresden, Dresden, Germany,* [8]*Milner Centre for Evolution, University of Bath, Bath, United Kingdom,* [9]*Doherty Applied Microbial Genomics, Department of Microbiology & Immunology, The University of Melbourne at The Peter Doherty Institute for Infection and Immunity, Melbourne, Australia; Microbiological Diagnostic Unit Public Health Laboratory, Department of Microbiology & Immunology, The University of Melbourne at The Peter Doherty Institute for Infection and Immunity, Melbourne, Australia,* [10]*Public Health England, National Infection Service, London, United Kingdom,* [11]*Creighton University, Omaha, United States,* [12]*Scottish Microbiology Reference Laboratories, Glasgow, United Kingdom,* [13]*The Chinese University of Hong Kong, Hong Kong,* [14]*The University of Sydney, Sydney, Australia,* [15]*School of Veterinary and Laboratory Sciences, Murdoch University, Murdoch, Western Australia,* [16]*Statens Serum Institut, Copenhagen, Denmark,* [17]*National MRSA Reference Laboratory, St. James's Hospital, Dublin, Ireland,* [18]*Microbiology Research Unit, School of Dental Science, University of Dublin, Trinity College Dublin, Ireland,* [19]*Istituto Superiore di Sanità, Rome, Italy,* [20]*Instituto de Tecnologia Química e Biológica, Oeiras, Portugal,* [21]*The Rockefeller University, New York City, United States of America,* [22]*University of Copenhagen, Copenhagen, Denmark,* [23]*Hvidovre University Hospital, Hvidovre, Denmark,* [24]*Sapporo Medical University, Sapporo, Japan,* [25]*Institute of Environmental Science and Research, Wellington, New Zealand,* [26]*Robert Koch Institute, Wernigerode, Germany,* [27]*Sheikh Khalifa Medical City, Abu Dhabi, United Arab Emirates,* [28]*Akershus University Hospital, Lørenskog, Norway,* [29]*King Fahd Medical City, Riyadh, Kingdom of Saudi Arabia,* [30]*London School of Hygiene and Tropical Medicine, United Kingdom,* [31]*Sunshine Coast University, Sippy Downs, Australia,* [32]*The School of Psychological and Clinical Sciences, Charles Darwin University, Darwin, Australia,* [33]*Wellcome Trust Sanger Institute, Cambridge, United Kingdom,* [34]*School of Cellular and Molecular Medicine, University of Bristol, United Kingdom,* [35]*University of St Andrews, St Andrews, United Kingdom,* [36]*Victorian Infectious Disease Service, The Royal Melbourne Hospital, and The University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia*

## Introduction

The global spread of antimicrobial resistance (AMR) has been well documented in Gram-negative bacteria and healthcare-associated epidemic pathogens, with emergence often from regions with high levels of antimicrobial use and lack of effective stewardship[1–5]. However, the degree to which similar processes occur with Gram-positive bacteria in the community setting, such as with community-associated (CA-) MRSA, is less well understood. Given the heavy burden and costs associated with MRSA infections[6,7], there is an urgent need to elucidate the patterns and drivers of the spread of novel virulent and multidrug-resistant MRSA clones. Here, we present whole-genome sequence data of 340 sequence type (ST) 772 *S. aureus* isolates, also known as the Bengal Bay clone. The collection encompasses the global distribution of a recently emerged, multidrug-resistant CA-MRSA lineage first reported from India[8] and Bangladesh[9] in 2004. Genomic and epidemiological data support an origin and spread of ST772 from the Indian subcontinent, often associated with travel and family contacts. We show that there is potential for short-term outbreaks to occur following intercontinental transmission, although ongoing endemic transmission is uncommon. Acquisition of a multidrug resistance integrated plasmid was instrumental in the emergence of a dominant clade (ST772-A) in the early 1990s. Phenotypic data suggest that the integrated plasmid did not incur a fitness cost. The Bengal Bay clone therefore combines the multidrug resistance of traditional healthcare-associated clones with the epidemiological and virulence potential of CA-MRSA.

## Main

In 2004, a novel *S. aureus* clone, sequence type (ST) 772, was isolated from two hospitals in Bangladesh[9] and from a community-setting in India[8]. The clone continued to be reported in community- (CA-) and healthcare-associated (HA-) environments in India, where it has become one of the dominant epidemic lineages of CA-MRSA[10]. Similar to other *S. aureus*, ST772 primarily causes skin and soft tissue infections, but more severe manifestations such as bacteraemia and necrotising pneumonia have been observed. Its potential for infiltration into nosocomial environments[11–14] and resistance to multiple classes of commonly used antibiotics (aminoglycosides, β-lactams, fluoroquinolones, macrolides and trimethoprim)[14–16] has resulted in ST772 becoming a serious public health concern in South Asia and elsewhere. Over the last decade, the clone has been isolated from community- and hospital-environments of countries in Asia, Australasia, Africa, the Middle East and Europe with patient records frequently indicating travel or family background in South Asia (Supplementary Map 1, Supplementary Table 1). As a consequence of its discovery, distribution and epidemiology, the lineage has been informally dubbed the “Bengal

Bay clone"[17]. Despite clinical and epidemiological hints for a recent and widespread dissemination of ST772, a global perspective on the evolutionary history and emergence of the clone is lacking.

We generated whole genome sequence data of 354 *S. aureus* ST772 isolates collected across Australasia, South Asia, Hong Kong, the Middle East and Europe between 2004 and 2013 (Supplementary Map 2, Supplementary Table 2). Fourteen isolates were excluded after initial quality control due to contamination (Supplementary Tables 2, 3). The remainder mapped with 165x average coverage against the PacBio reference genome DAR4145[16] from Mumbai (Supplementary Tables 2, 3). Phylogenetic analysis using core-genome SNPs (n = 7,063) revealed little geographic structure within the lineage (Figure 1a). A minority of ST772-methicillin-susceptible *S. aureus* (MSSA) and -MRSA strains (n = 11) were basal to a single globally distributed clade (ST772-A, n = 329) that harbored an integrated resistance plasmid (IRP) described in the reference genome DAR4145[16] (Figures 1a, 1b). Population network analysis distinguished three distinct subgroups within ST772-A (Figures 1a, 1c): an early-branching subgroup harboring multiple subtypes of the staphylococcal cassette chromosome (SCC*mec*) (A1, n = 81), a dominant subgroup (A2, n = 153) and an emerging subgroup (A3, n = 56), that exclusively harbors a short variant of SCC*mec*-V.

Epidemiological and genomic characteristics of ST772 were consistent with an evolutionary origin from the Indian subcontinent. 60% of isolates in this study were collected from patients with family- or travel-background in Bangladesh, India, Nepal or Pakistan, compared to unknown (19%) or other countries (21%) (Figure 2a, Supplementary Table 2). We found significantly more isolates from India and Bangladesh among the basal strains, compared to clade ST772-A (Fisher's exact test, 5/11 vs. 47/291, p = 0.026). In particular, three isolates from India and Bangladesh were basal in the (outgroup-rooted) maximum-likelihood phylogeny (Figure 1b, Supplementary Figure 1), including two MSSA samples from the original isolations in 2004 (RG28, NKD22). Isolates recovered from South Asia were genetically more diverse than isolates from Australasia and Europe, supporting an origin from the Indian subcontinent (Figure 2b, Supplementary Figure 2).

Consistent with a methicillin-susceptible progenitor, a significantly higher proportion of MSSA was found in the basal isolates (Fisher's exact test, 4/11 vs. 31/291, p = 0.028) and MSSA isolates demonstrated a lower patristic distance to the root of the maximum likelihood phylogeny compared to MRSA (Supplementary Figure 3a). Although it appears that MSSA is proportionately more common in South Asia (Supplementary Figure 3b), it is also possible that the observed distribution may be related to non-structured sampling. Recent studies have detected ST772-MSSA and -MRSA in Nepal[18] and ST772-MRSA in Pakistan[19], but it is unclear whether the lineage has been endemic in

these countries prior to its emergence in India. Deeper genomic surveillance of ST772-MSSA and –MRSA in the region will be necessary to understand the local epidemiology and evolutionary history of the clone on the Indian subcontinent.

Phylogenetic dating suggests an initial divergence of the ancestral ST772 population in 1970 (age of root node: 1970.02, CI: 1955.43 – 1982.60) with a core-genome substitution rate of 1.61 x $10^{-6}$ substitutions/site/year after removing recombination (Figure 2c, 2d, Supplementary Figures 4, 5). This was followed by the emergence of the dominant clade ST772-A and its population subgroups in the early 1990s (ST772-A divergence, 1990.83, 95% CI: 1980.38 – 1995.08). The geographic pattern of dissemination is heterogeneous (Figure 1a). There was no evidence for widespread endemic dissemination of the clone following intercontinental transmission, although localised healthcare-associated outbreak clusters occurred in neonatal intensive care units in Ireland (NICU-1 and NICU-2, Figure 1a, Supplementary Figure 6)[20] and have been reported from other countries in Europe[14] and South Asia[11–13]. While some localised spread in the community was observed among our isolates, patients in local transmission clusters often had traveled to or had family in South Asia (19/27 clusters, Supplementary Figure 6). Small transmission clusters in both hospitals and house-holds were also recently reported from a comprehensive surveillance study of ST772 in Norway[14].

Given the available epidemiological data (Figure 2a), phylogeographic heterogeneity (Figure 1a) and the clone's limited success to establish itself in regions outside its endemic range in South Asia (Figure 1a), there appears to be ongoing exportation of ST772 from the Indian subcontinent. This hypothesis is in line with MRSA importation in travelers, including direct observations of ST772 importation by returnees from India[21]. The pattern of spread mirrors other CA-MRSA lineages such as USA300[22,23], ST80-MRSA[24] and ST59[25] where clones emerge within a particular geographic region, are exported elsewhere, but rarely become established and endemic outside of their place of origin. In contrast, HA-MRSA clones such as CC22-MRSA-IV (EMRSA-15)[4] and ST239-MRSA-III[26,27] demonstrate much stronger patterns of phylogeographic structure, consistent with importation into a country followed by local dissemination through the healthcare system.

We examined the distribution of virulence factors, antibiotic resistance determinants and mutations in coding regions to identify the genomic drivers in the emergence and dissemination of ST772. Nearly all isolates (336/340) carried the Panton-Valentine leucocidin (PVL) genes *lukS/F*, most isolates (326/340) carried the associated enterotoxin A (*sea*) and all isolates carried *scn* (Supplementary Table 5). This indicates a nearly universal carriage, across all clades, of both, a truncated *hlb*-converting prophage (the typically associated staphylokinase gene *sak* was only present

in one isolate) and the PVL/*sea* prophage φ-IND772[28]. Amongst other virulence factors, the enterotoxin genes *sec* and *sel*, the gamma-hemolysin locus, *egc* cluster enterotoxins and the enterotoxin homologue ORF CM14 were ubiquitous in ST772 (Supplementary Table 7). We detected no statistically significant difference between core virulence factors present in the basal group and ST772-A (Supplementary Table 5, Supplementary Figure 7).

We noted a pattern of increasing antimicrobial resistance as successive clades of ST772 emerged. Predicted resistance phenotypes across ST772 were common for ciprofloxacin (97.4%), erythromycin (96.2%), gentamicin (87.7%), methicillin (89.7%), penicillin (100%) and trimethoprim (98.8%), with a corresponding resistome composed of acquired and chromosomally encoded genes and mutations (Figure 3a, Figure 3b, Supplementary Table 6). There was significantly less predicted resistance in the basal strains compared to ST772-A, including overall multidrug-resistance (≥ 3 classes, 8/11 vs. 291/291, Fisher's exact test, $p < 0.001$) (Figure 3d). The key resistance determinants of interest were the SCC*mec* variants, an integrated resistance plasmid, and other smaller mobile elements and point mutations.

MRSA isolates predominantly harbored one of two subtypes of SCC*mec*-V: a short variant (5C2) or a composite cassette (5C2&5), which encodes a type 5 *ccr* complex containing *ccrC1* (allele 8) between the *mec* gene complex and *orfX*[29] (Supplementary Figure 8). Integration of the Tn*4001* transposon encoding aminoglycoside resistance gene *aadA-aphD* occurred across isolates with different SCC*mec* types (260/267), but not in MSSA (0/35). All MRSA isolates (n = 7) within the basal group carried the larger composite cassette SCC*mec*-V (5C2&5), with two of these strains lacking *ccrC* and one isolate carrying a remnant of SCC*mec*-IV (Figure 1a).

The diversity of SCC*mec* types decreased as ST772-A diverged into subgroups (Figure 1a, c, Supplementary Table 6). ST772-A1 included MSSA (n = 30) as well as SCC*mec*-V (5C2) (n = 22) and (5C2&5) (n = 18) strains. Four isolates harbored a putative composite SCC element that included SCCmec-V (5C2), as well as *pls* and the *kdp* operon previously known from SCCmec II. One isolate harbored a composite SCCmec-V (5C2&5) with copper + zinc resistance element, known from the European livestock associated CC398-MRSA[30]. Another six isolates yielded irregular and/or composite SCC elements (Supplementary Table 6).

In contrast, the dominant subgroups ST772-A2 and -A3 exclusively carried the short SCC*mec*-V (5C2) element. In 11 of these isolates (including all isolates in NICU-2) the SCC*mec*-V (5C2) element lacked *ccrC* and two isolates carried additional recombinase genes (*ccrA/B2* and *ccrA2*). In light of

earlier studies demonstrating a fitness advantage in having a smaller SCC*mec* element[31–33], the fixation of the shorter SCC*mec*-V (5C2) subtype in ST772-A2 and –A3 may be a contributing factor to their success.

ST772-A was characterized by the acquisition of an integrated multidrug resistance plasmid (IRP, Figure 3c), encoding the macrolide-resistance locus *msrA*, as well as determinants against β-lactams (*blaZ*) and aminoglycosides (*aadE-sat4-aphA3*). Thus predicted resistance to erythromycin was uniquely found in ST772-Aand not in any of the basal strains (Fisher's exact test, 289/291 vs 0/11, $p < 0.001$, Figure 3d). The IRP element is highly similar to the extrachromosomal plasmid 18809-p03 in USA300[34] and the integrated plasmid in the dominant European lineage ST80[24] (Figure 3c, Supplementary File 1). Unlike in ST80, the ST772 IRP is not integrated in SCC*mec*[16].

Three basal strains were not multi-drug resistant and included two isolates from the original collections in India (RG28) and Bangladesh (NKD122) (Figure 1a, 3a). These two strains lacked the trimethoprim determinant *dfrG* and the fluoroquinolone mutations in *grlA* or *gyrA,* encoding only a penicillin-resistance determinant *blaZ* on a Tn*554*-like transposon. However, seven of the strains more closely related to ST772-A did harbor elements and mutations conferring trimethoprim (*dfrG*) and quinolone resistance (*grlA* and *gyrA* mutations). Interestingly, we observed a shift from the quinolone resistance *grlA* S80F mutation in basal strains and ST772-A1, to the *grlA* S80Y mutation in ST772-A2 and -A3 (Figure 3a).

Thus, the phylogenetic distribution of the key resistance elements suggests acquisition of the IRP by a PVL-positive MSSA strain in the early 1990s (ST772-A1 divergence, 1990.83, 95% CI: 1980.38 – 1995.08), followed by fixation of both the shorter variant of SCC*mec*-V (5C2) and the *grlA* S80Y mutation in a PVL- and IRP-positive MSSA ancestor in the late 1990s (ST772-A2 divergence, 1999.18, 95% CI: 1993.26 – 2001.56) (Figure 1a, Figure 2c).

We found three other mutations of interest that were present exclusively in ST772-A strains (Supplementary Table 7). The first mutation caused a non-synonymous change in *fbpA* (L55P), encoding a fibrinogen-binding protein that mediates surface adhesion in *S. aureus*[35]. The second comprised a non-synonymous change (L67V) in the *plc* gene, encoding a phospholipase associated with survival in human blood cells and abscess environments in USA300[36]. The third encoded a non-synonymous mutation (S273G) in *tet(38),* an efflux pump that promotes resistance to tetracyclines as well as survival in abscess environments and skin colonisation[37]. The functional implication of genes harboring these canonical mutations might suggest a modification of the clone's ability to colonise

and cause SSTIs.

In light of these canonical SNPs, we selected five basal strains and 10 strains from ST772-A to screen for potential phenotypic differences that may contribute to the success of ST772-A. We assessed *in vitro* growth, biofilm formation, cellular toxicity, and lipase activity (Figure 4, Supplementary Table 8). We found no statistically significant differences between the basal strains and ST772-A in these phenotypic assays, apart from significantly lower lipase activity among ST772-A strains (Welch's two-sided t-test, $t = 3.4441$, $df = 6.0004$, $p = 0.0137$, Figure 4e), which may be related to the canonical non-synonymous mutation in *plc*. However, it is increased rather than decreased lipase activity that has been associated with viability of *S. aureus* USA300 in human blood and neutrophils[36].

We found no difference in the median growth rate of ST772-A compared to the basal strains (Figure 4, Mann-Whitney, $W = 27$, $p = 0.8537$, Supplementary Table 8), although there were two ST772-A strains that grew more slowly suggesting the possibility of some strain to strain variability. However, overall, it appears that acquisition of resistance determinants on the IRP has not incurred a significant cost to *in vitro* growth of strains from ST772-A. This raises the possibility that members of this clade will both survive in environments where antibiotics are heavily used, such as hospitals or in the community where antibiotic stewardship is poor, but also be at little disadvantage in environments where there is less antibiotic use, because its growth rate is comparable to that of non-resistant strains. While we only assayed for a limited number of phenotypic differences, our data suggest that acquisition of antibiotic resistance was a key driver in the emergence and persistence of ST772-A.

Considering the widespread use of antibiotics and associated poor antibiotic regulation, poor public health infrastructure, and high population density in parts of South Asia, the emergence and global dissemination of multidrug resistant bacterial clones (both Gram-positive and Gram-negative) is alarming, and perhaps not surprising. Here, we demonstrate that the acquisition of specific antimicrobial resistance determinants has been instrumental in the evolution of a multidrug resistant CA-MRSA clone. Global initiatives and funding to monitor the occurrence of emerging clones and resistance mechanisms, and support for initiatives in antimicrobial stewardship at community, healthcare and agricultural levels are urgently needed.

## Figures

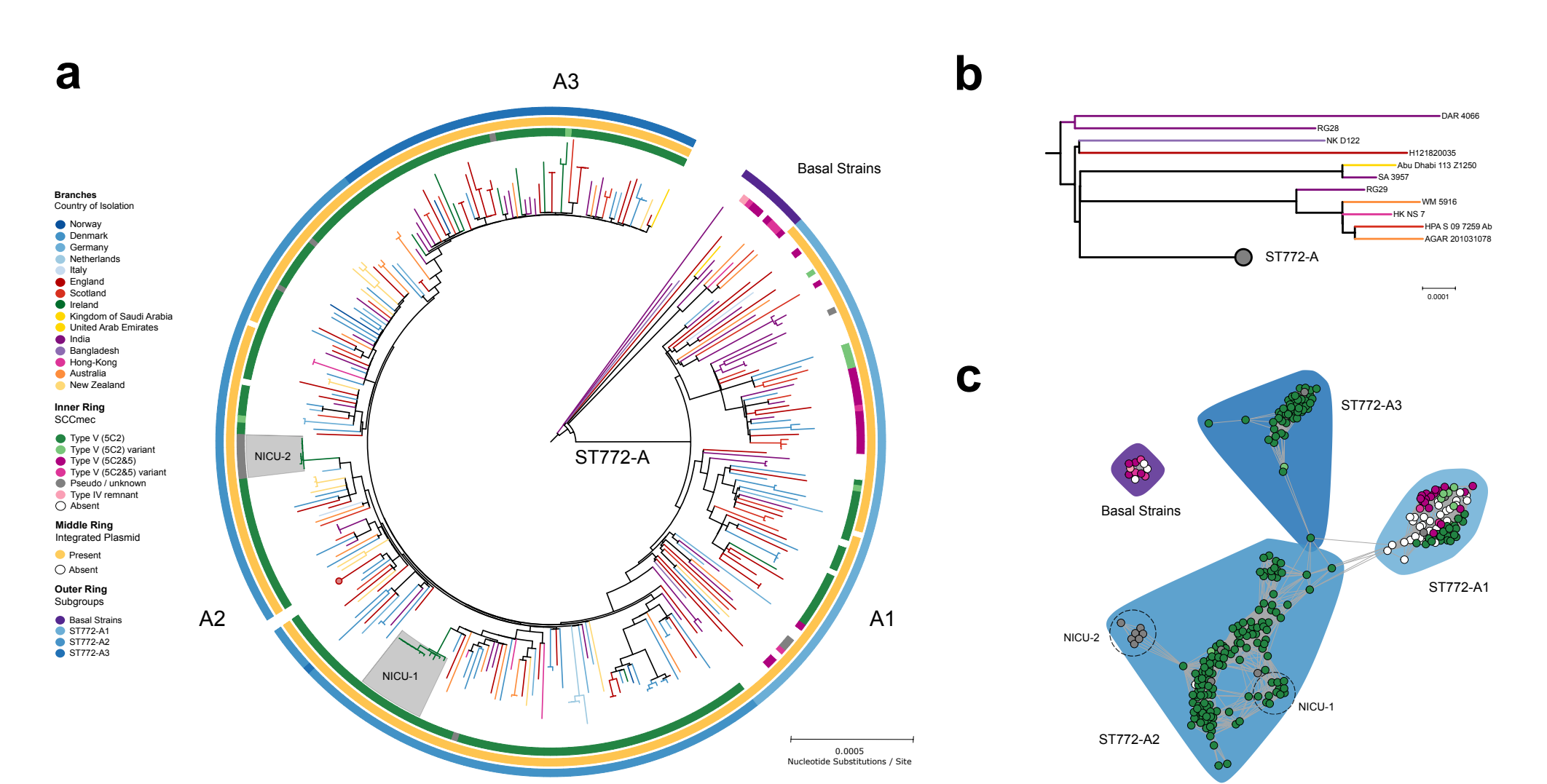
**Figure 1:** Evolutionary history and population structure of ST772**. (a)** Maximum likelihood phylogeny of ST772 (n = 340) based on 7,063 core-genome SNPs. Branch colors denote country of isolation, the inner ring delineates presence and type of SCC*mec*, the middle ring shows presence of the integrated resistance plasmid and the outer ring denotes community-membership of the population graph shown in (c). Communities match the tree topology, with several basal isolates (n = 11) and a single derived clade ST772-A (n = 329) composed of three population subgroups (A1 – A3). Isolates from two outbreaks in neonatal intensive care units in Ireland are indicated in grey (NICU-1 and NICU-2). Only one representative isolate from longitudinal sampling of a single healthcare worker (n = 39) is included (red circle). **(b)** Basal strains of ST772 showing positions of isolates from India and Bangladesh at the root of the phylogeny (RG28, DAR4066, NKD122). **(c)** Population graph based on pairwise SNP distances, showing SCC*mec* type (node color as for Figure 1a legend) and population subgroups (polygons, A1-A3). Dashed circles denote hospital-associated outbreaks in Ireland (NICU-1 and NICU-2).

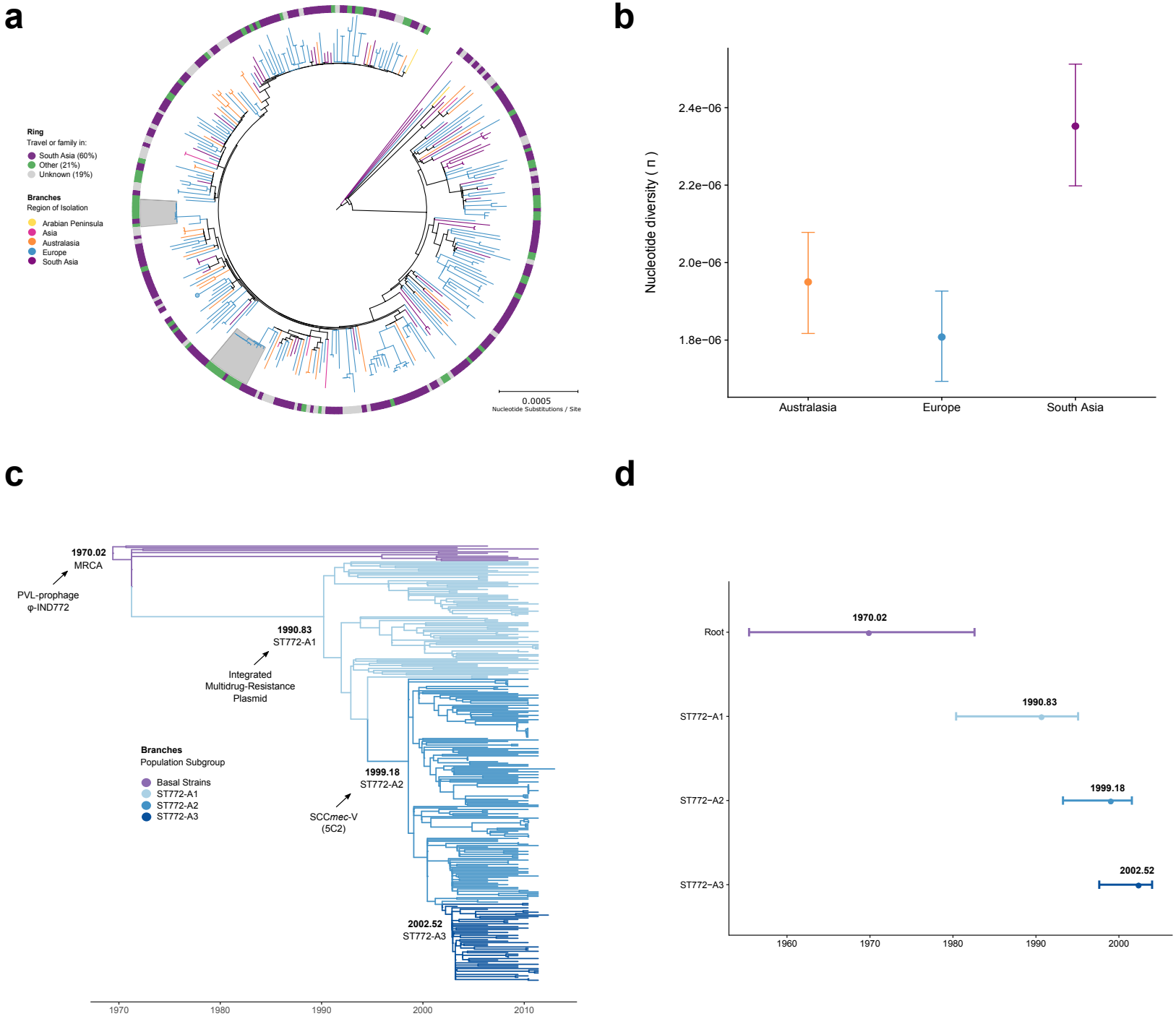
**Figure 2**: Molecular epidemiology of ST772. (**a**) Patient family- or travel-background in South Asia (India, Pakistan, Nepal, Bangladesh) (59.5%, purple), other countries (21.2%, green) or unknown status (19.3%, gray), is widely distributed across the phylogenetic topology of ST772 (n = 340). Only one representative isolate from longitudinal sampling of a single healthcare worker(n = 39) is included (circle). (**b**) Average pairwise nucleotide diversity per site ($\pi$), measured by region (Australasia: orange, n = 36; Europe: blue, n = 244; South Asia: purple, n = 52). Error bars indicate 95% confidence intervals using non-parametric bootstrapping. Isolates from the Arabian Peninsula (n = 2) and Hong Kong (n = 6) were excluded from the diversity analysis due to the small number of samples from these regions. **(c)** Phylogenetic time-tree with the timescale estimated in Least Squares Dating (LSD). The annotations for nodes represent the time of origin (in years) of basal strains and subgroups A1, A2, A3. Times to the most recent common ancestor (TMRCA) for these lineages are shown. Tips are colored according to the subgroup as per Figure 1a. The position of the root was optimised during the analysis. Arrows indicate acquisition of three critical mobile genetic elements: the PVL/*sea*-prophage φ-IND772, an integrated multidrug resistance plasmid and the short staphylococcal cassette chromosome SCC*mec*-V (5C2). **(d)** Times to the most recent common ancestor of sub-groups in ST772 after removing recombination. Horizontal bars indicate 95% confidence intervals for nodes (CI) using parametric bootstrapping in LSD.

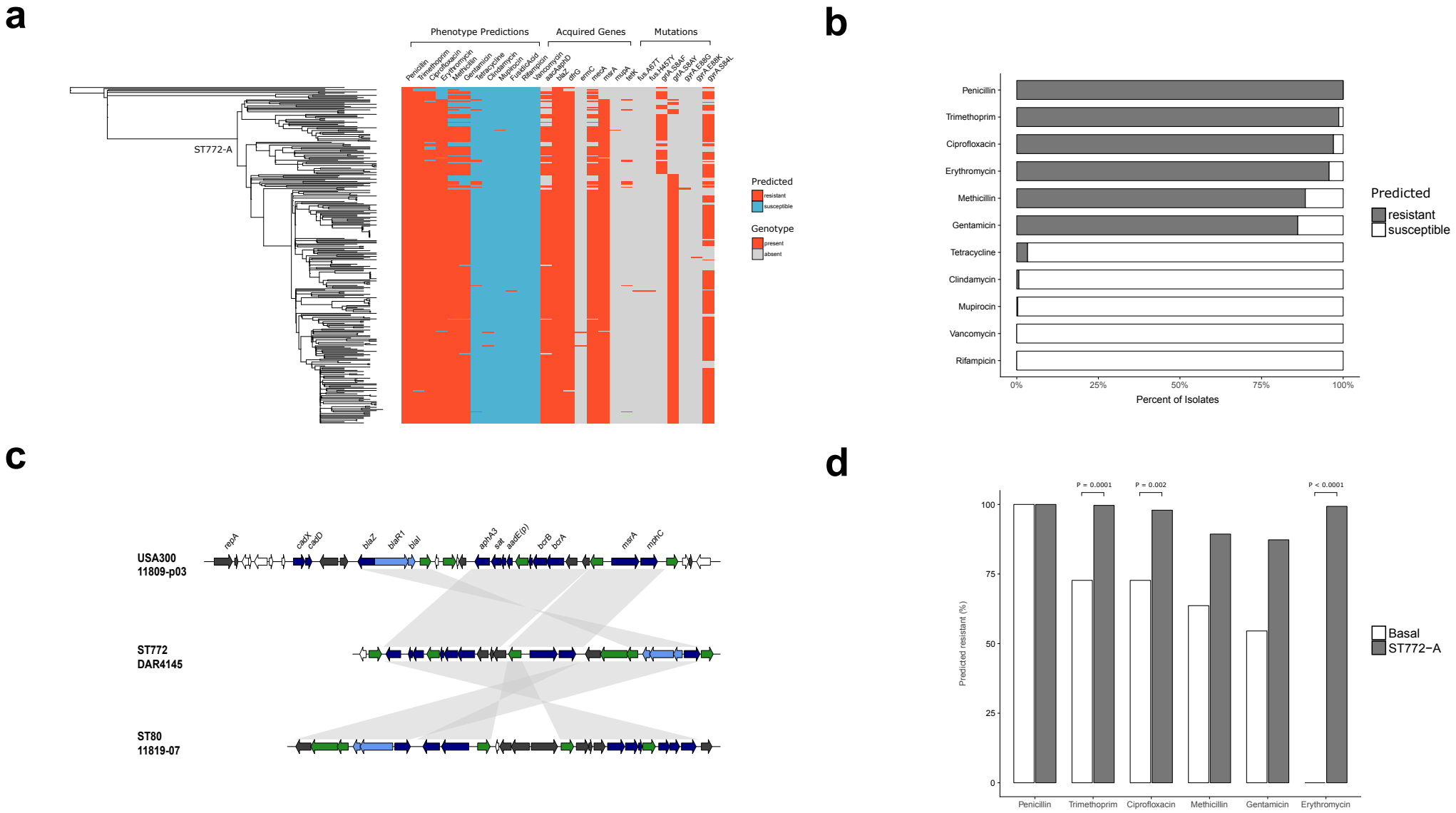
**Figure 3:** Resistome and predicted resistance phenotypes across ST772. (**a**) Resistome mapped to

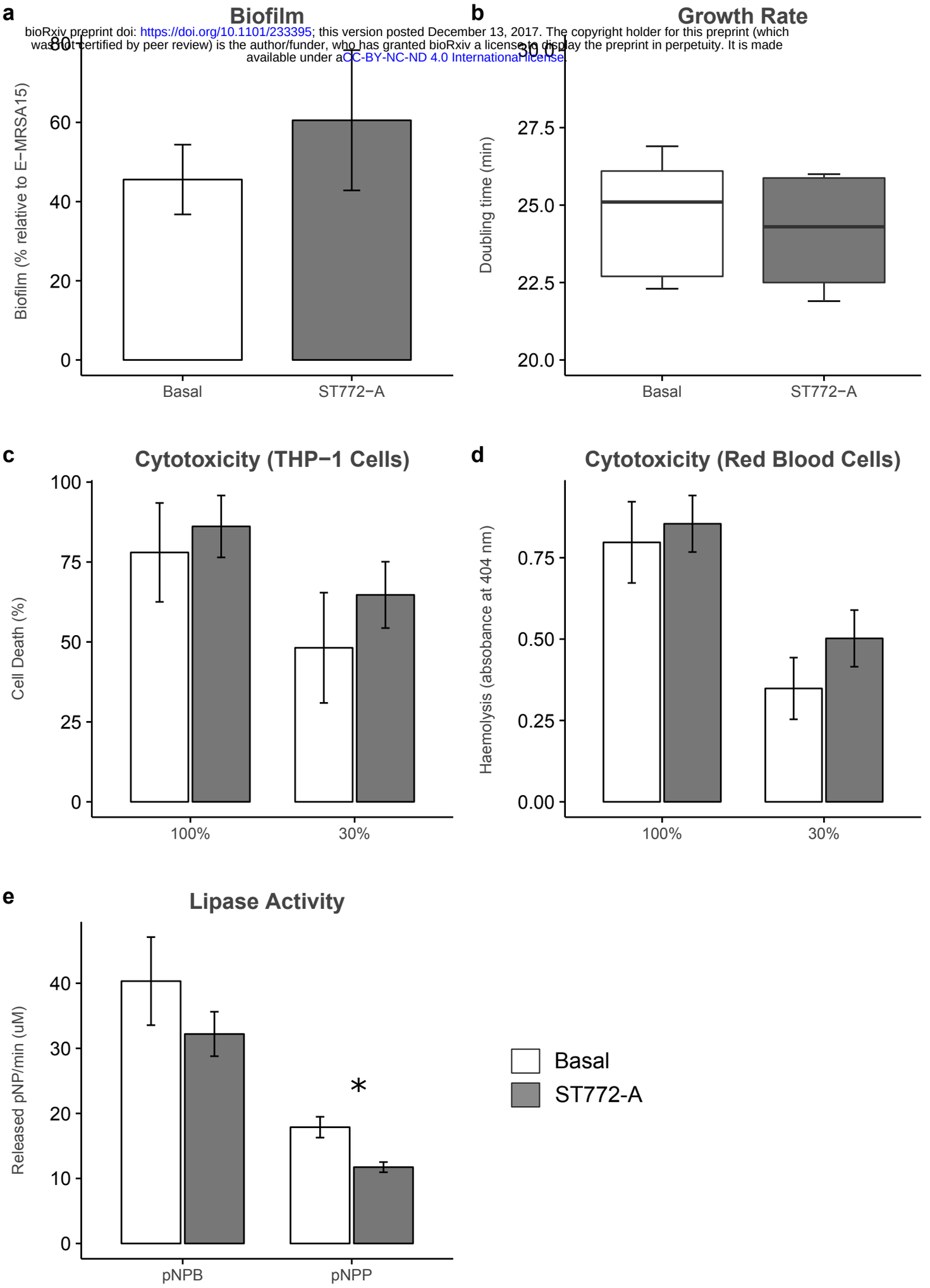
maximum likelihood phylogeny of ST772. Predicted resistant phenotype is depicted in red, while susceptible phenotype is depicted in blue. Presence of acquired resistance genes and mutations responsible for phenotype predictions are shown in red, while absence of these determinants is shown in gray. (**b**) Percent of isolates predicted resistant (gray) or susceptible (white) for all antimicrobials included in Mykrobe predictor (**c**) BLAST comparison of the multidrug-resistance plasmid in DAR4145 (middle) with the extrachromosomal plasmid 11809-p03 (top) and the SCC*mec*-IV integrated plasmid in ST80 (bottom), showing alignments > 1000 bp and > 95% nucleotide identity. The comparison highlights three regions harboring resistance genes (dark blue) and their regulators (light blue), which are flanked by transposition elements (green) and appear to have integrated with reversions and rearrangements into ST80 and ST772. Resistance genes include the β-lactam *blaZ* complex, aminoglycoside cluster *aphA3-sat4-aadE* and bacitracin resistance loci *bcrA/B,* as well as macrolide efflux genes *msrA* and *mphC*. Hypothetical proteins and genes of other annotated function are shown in white and dark gray, respectively. **(d)** Proportion of isolates predicted resistant to common antibiotics for basal isolates (white, n = 11) and isolates from ST772-A (gray, n = 291). Values above bars denote statistically significant differences between groups using Fisher's exact test where $p < 0.01$.

**Figure 4**: Phenotypic assays for representative strains from the basal group (white, n = 5) and ST772-A (gray, n = 10) for **(a)** optical density measurements (595 nm) of biofilm formation, accounting for day to day variability relative to control strain E-MRSA15 (%), **(b)** overnight growth in tryptic soy broth (doubling time per minute) measured by optical density (600 nm), **(c)** cytotoxicity of neat (100%) and diluted (30%) bacterial supernatant to THP-1 cells measured as cell death by flow cytometry, **(d)** absorbance measurements (404 nm) of erythrocyte haemolysis in neat (100%) and diluted (30%) bacterial supernatant, **(e)** lipase activity of *para*-nitrophenyl butyrate (pNPB) or *para*-nitrophenyl palmitate (pNPP) (release of pNP per minute) in neat bacterial supernatant measured by absorbance (410 nm). Slow growing strains H104580604 and HPAS101177P were considered outliers and removed from the boxplot for clarity after calculation of median and interquartile ranges and assessment of significance. Error bars show standard error; the asterisk denotes a significant difference in pNPP release (Welch's two-sided t-test, $t = 3.4441$, $df = 6.0004$, $p = 0.0137$) between basal strains and ST772.


a
A3
Basal Strains
A2
A1
ST772-A
NICU-2
NICU-1
0.0005
Nucleotide Substitutions / Site
Branches
Country of Isolation
Norway
Denmark
Germany
Netherlands
Italy
England
Scotland
Ireland
Kingdom of Saudi Arabia
United Arab Emirates
India
Bangladesh
Hong-Kong
Australia
New Zealand
Inner Ring
SCCmec
Type V (5C2)
Type V (5C2) variant
Type V (5C2&5)
Type V (5C2&5) variant
Pseudo / unknown
Type IV remnant
Absent
Middle Ring
Integrated Plasmid
Present
Absent
Outer Ring
Subgroups
Basal Strains
ST772-A1
ST772-A2
ST772-A3
b
DAR 4066
RG28
NK D122
H121820035
Abu Dhabi 113 Z1250
SA 3957
RG29
WM 5916
HK NS 7
HPA S 09 7259 Ab
AGAR 201031078
ST772-A
0.0001
c
ST772-A3
Basal Strains
ST772-A1
NICU-2
NICU-1
ST772-A2


a
Ring
Travel or family in:
South Asia (60%)
Other (21%)
Unknown (19%)
Branches
Region of Isolation
Arabian Peninsula
Asia
Australasia
Europe
South Asia
0.0005
Nucleotide Substitutions / Site
b
Nucleotide diversity ( π )
2.4e−06
2.2e−06
2.0e−06
1.8e−06
Australasia
Europe
South Asia
c
1970.02
MRCA
PVL-prophage
φ-IND772
1990.83
ST772-A1
Integrated
Multidrug-Resistance
Plasmid
1999.18
ST772-A2
SCCmec-V
(5C2)
2002.52
ST772-A3
Branches
Population Subgroup
Basal Strains
ST772-A1
ST772-A2
ST772-A3
1970
1980
1990
2000
2010
d
Root
ST772−A1
ST772−A2
ST772−A3
1970.02
1990.83
1999.18
2002.52
1960
1970
1980
1990
2000


a
Phenotype Predictions
Acquired Genes
Mutations
Penicillin
Trimethoprim
Ciprofloxacin
Erythromycin
Methicillin
Gentamicin
Tetracycline
Clindamycin
Mupirocin
Fusidic Acid
Rifampicin
Vancomycin
aac4ag4nD
blaZ
dfrG
ermC
mecA
msrA
mupA
tetK
fus.A57T
fus.H457Y
grlA.S84F
grlA.S84V
gyrA.E88G
gyrA.E88K
gyrA.S84L
ST772-A
Predicted
resistant
susceptible
Genotype
present
absent
b
Penicillin
Trimethoprim
Ciprofloxacin
Erythromycin
Methicillin
Gentamicin
Tetracycline
Clindamycin
Mupirocin
Vancomycin
Rifampicin
0%
25%
50%
75%
100%
Percent of Isolates
Predicted
resistant
susceptible
c
USA300
11809-p03
repA
cadX
cadD
blaZ
blaR1
blaI
aphA3
sat
aadE(p)
bcrB
bcrA
msrA
mphC
ST772
DAR4145
ST80
11819-07
d
P = 0.0001
P = 0.002
P < 0.0001
Predicted resistant (%)
100
75
50
25
0
Penicillin
Trimethoprim
Ciprofloxacin
Methicillin
Gentamicin
Erythromycin
Basal
ST772−A



## Methods

### *Isolates*

Isolates were obtained from Australia (21), Bangladesh (3), Denmark (70), England (103), Germany (16), Hong Kong (6), India (44), Ireland (28), Italy (2), Netherlands (4), New Zealand (17), Norway (3), Saudi Arabia (1), Scotland (29) and the United Arab Emirates (1) between 2004 and 2012 (Supplementary Table 2). The collection was supplemented with six previously published genome sequences from India[29,38,39]. Notable samples include the initial isolates from Bangladesh and India[8,9], two hospital-associated (NICU) clusters from Ireland[20] and longitudinal isolates from a single healthcare worker at a veterinary clinic sampled over two consecutive weeks (VET) [40]. Geographic regions were designated as Australasia (Australia, New Zealand), East Asia (Hong Kong), South Asia (India, Bangladesh), Arabian Peninsula (Saudi Arabia, United Arab Emirates) and Europe (Denmark, England, Germany, Ireland, Italy, Netherlands, Norway and Scotland).

### *Clinical data and epidemiology*

Anonymised patient data was obtained for the date of collection, clinical symptoms, geographic location, epidemiological connections based on family or travel-history, and acquisition in nosocomial- or community-environments, where available (Supplementary Table 2). Clinical symptoms were summarized as SSTI (abscesses, boils, ulcers, exudates, pus, ear and eye infections), urogenital- (vaginal swabs, urine), bloodstream- (bacteremia) or respiratory-infections (pneumonia, lungs abscesses) and colonization (swabs from ear, nose, throat, perineum or environmental) (Supplementary Table 2, Supplementary Figure 9). Literature and sample maps (Supplementary Maps 1 and 2) were constructed with *geonet,* a wrapper for geographic projections with Leaflet in R (https://github.com/esteinig/geonet).

Where available, acquisition in community- or healthcare-environments was recorded in accordance with guidelines from the CDC. CA-MRSA is therein classified as an infection in a person who has none of the following established risk factors for MRSA infection: isolation of MRSA more than 48 h after hospital admission; history of hospitalization, surgery, dialysis or residence in a long-term care facility within one year of the MRSA culture date; the presence of an indwelling catheter or a percutaneous device at the time of culture; or previous isolation of MRSA[41,42] (Supplementary Figure 9).

A valid epidemiological link to South Asia was declared if either travel- or family-background could be reliably traced to Bangladesh, India, Nepal or Pakistan. If both categories (travel and family) were unknown or one did not show a link to the region, we conservatively declared the link as unknown or absent, respectively. The longitudinal collection (n = 39) from a staff member at a veterinary hospital in England was treated as a single patient sample.

*Sequencing, quality control and assembly*

Unique index-tagged libraries were created for each isolate, and multiplexed libraries were sequenced on the Illumina HiSeq with 100 bp paired-end reads. Read quality control was conducted with Trimmomatic[43], Kraken[44] and FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc). Quality control identified a large proportion of reads classified as *Enterococcus faecalis* in sample HWM2178 (Supplementary Table 3). *In silico* micro-array typing (see below) identified an additional 13 isolates with possible intra-specific contamination due to simultaneous presence of *agr I* and *II,* as well as capsule types 5 and 8 (Supplementary Table 2). We excluded these isolates from all genomic analyses. Raw Illumina data were sub-sampled to 100x coverage and assembled with the SPAdes[45] pipeline Shovill (https://github.com/tseemann/shovill), which wraps SPAdes, Lighter[46], FLASH[47], BWA MEM[48], SAMtools[49], KMC[50] and Pilon[51]. Final assemblies were annotated with Prokka v1.11[52]. Samples from the veterinary staff member were processed and sequenced as described by Paterson et al.[40].

*MLST and SCC typing*

*In silico* multi-locus sequence typing (MLST) was conducted using mlst (https://github.com/tseemann/mlst) on the assembled genomes with the *S. aureus* database from PubMLST (https://pubmlst.org/saureus/). Three single locus variants (SLVs) of ST772 were detected and retained for the analysis, describing sequence types ST1573, ST3362 and ST3857 (Supplementary Table 2). Sequences of experimentally verified sets of probes for SCC- related and other *S. aureus* specific markers[53,54] were blasted against SPAdes assemblies (*in silico* micro-array typing), allowing prediction of presence or absence of these markers and detailed typing of SCC elements. We assigned MRSA to four isolates that failed precise SCC classification based on presence of *mecA* on the probe array and detection of the gene with Mykrobe predictor[55].

*Variant calling*

Samples passing quality control (n = 340) were aligned to the PacBio reference genome DAR4145 from Mumbai and variants were called with the pipeline Snippy (available at https://github.com/tseemann/snippy) which wraps BWA MEM, SAMtools, SnpEff[56] and Freebayes[57]. Core SNPs (n = 7,063) were extracted with *snippy-core* at default settings. We assigned canonical SNPs for ST772-A, as those present exclusively in all isolates of ST772-A, but not in the basal strains. Annotations of variants were based on the reference genome DAR4145.

*Phylogenetics and recombination*

A maximum-likelihood (ML) tree under the General Time Reversible model of nucleotide substitution with among-site rate heterogeneity across 4 categories (GTR + Γ), ascertainment bias correction (Lewis) and 100 bootstrap (BS) replicates was generated based on 7,063 variant sites (core-genome SNPs) in RaxML-NG 0.5.0 (available at https://github.com/amkozlov/raxml-ng), which implements the core functionality of RAxML[58]. The tree with the highest likelihood out of ten replicates was midpoint-rooted and visualized with interactive Tree of Life (ITOL) (Figure 1a, 2a, Supplementary Figure 6, 12a)[59]. In all phylogenies (Figures 1a, 2a, 3a, Supplementary Figures 6, 10, 12a) samples from the veterinary staff member were collapsed for clarity.

A confirmation alignment (n = 351) was computed as described above for resolving the pattern of divergence in the basal strains of ST772. The alignment included the CC1 strain MW2 as outgroup, as well as another known SLV of CC1, sequence type 573 (n = 10). The resulting subset of core SNPs (n = 25,701) was used to construct a ML phylogeny with RaxML-NG (GTR + Γ) and 100 bootstrap replicates (Supplementary Figure 1). We also confirmed the general topology of our main phylogeny as described above using the whole genome alignment of 2,545,215 nucleotides generated by Snippy, masking sites if they contained missing (-) or uncertain (N) characters across ST772 (not shown).

Gubbins[60] was run on a complete reference alignment with all variant sites defined by Snippy to detect homologous recombination events, using a maximum of five iterations and the GTR + Γ model in RaxML (Supplementary Figure 10). A total of 205 segments were identified as recombinant producing a core alignment of 7,928 SNPs. Phylogenies were visualized using ITOL, *ape*[61], *phytools*[62], *ggtree*[63] or *plotTree* (https://github.com/holtlab/plotTree/). Patristic distances to the root of the phylogeny (Supplementary Figure 2) were computed in the *adephylo*[64] function *distRoot*.

*Dating analysis*

We used LSD v0.3 [65] to obtain a time-scaled phylogenetic tree. This method fits a strict molecular clock to the data using a least-squares approach. Importantly, LSD does not explicitly model rate variation among lineages and it does not directly account for phylogenetic uncertainty. However, its accuracy is similar to that obtained using more sophisticated Bayesian approaches [66], with the advantage of being computationally less demanding.

LSD typically requires a phylogenetic tree with branch lengths in substitutions per site, and calibrating information for internal nodes or for the tips of the tree. We used the phylogenetic tree inferred using Maximum likelihood in PhyML[67] (before and after removing recombination with Gubbins, as described above) using the GTR+Γ substitution model with 4 categories for the Γ distribution. We used a combination of nearest-neighbour interchange and subtree-prune-regraft to search tree space. Because PhyML uses a stochastic algorithm, we repeated the analyses 10 times and selected that with the highest phylogenetic likelihood. To calibrate the molecular clock in LSD, we used the collection dates of the samples (i.e. heterochronous data). The position of the root can be specified *a priori*, using an outgroup or by optimising over all branches. We chose the latter approach. To obtain uncertainty around node ages and evolutionary rates we used the parametric bootstrap approach with 100 replicates implemented in LSD.

An important aspect of analysing heterochronous data is that the reliability of estimates of evolutionary rates and timescales is contingent on whether the data have temporal structure. In particular, a sufficient amount of genetic change should have accumulated over the sampling time. We investigated the temporal structure of the data by conducting a regression of the root-to-tip distances of the Maximum likelihood tree as a function of sampling time[68], and a date-randomisation test[69]. Under the regression method, the slope of the line is a crude estimate of the evolutionary rate, and the extent to which the points deviate from the regression line determines the degree of clocklike behaviour, typically measured using the $R^2$[70]. The date randomisation test consists in randomising the sampling times of the sequences and re-estimating the rate each time. The randomisations correspond to the distribution of rate estimates under no temporal structure. As such, the data have strong temporal structure if the rate estimate using the correct sampling times is not within the range of those obtained from the randomisations[71]. We conducted 100 randomisations, which suggested strong temporal structure for our data (Supplementary Figure 3). We also verified that the data did not display phylogenetic-temporal clustering, a pattern which sometimes misleads the date-randomisation test[72].

Results from this analysis (substitution rates, and node age estimates) using phylogenies before and

after removing recombination were nearly identical (Supplementary Figure 4, 5). We therefore chose to present results from our analysis after removing recombination.

*Nucleotide diversity*

Pairwise nucleotide diversity and SNP distance distributions for each region with n > 10 (Australasia, Europe, South Asia) were calculated as outlined by Stucki et al.[73]. Pairwise SNP distances were computed using the SNP alignment from Snippy (n = 7,063) and the *dist.dna* function from *ape* with raw counts and deletion of missing sites in a pairwise fashion. An estimate of average pairwise nucleotide diversity per site ($\pi$) within each geographic region was calculated from the SNP alignments using raw counts divided by the alignment length. Confidence intervals for each region were estimated using 1000 bootstrap replicates across nucleotide sites in the original alignment via the *sample* function (with replacement) and 2.5% - 97.5% quantile range (Figure 2b).

*Population structure*

We used the network-analysis and -visualization tool NetView[74,75] (available at http://github.com/esteinig/netview) to delineate population subgroups in ST772. Pairwise Hamming distances were computed from the core SNP alignment derived from Snippy. The distance matrix was used to construct mutual k-nearest-neighbour networks from $k$ = 1 to $k$ = 100. We ran three commonly used community detection algorithms as implemented in *igraph* to limit the parameter choice to an appropriate range for detecting fine-scale population structure: fast-greedy modularity optimization[76], Infomap[77] and Walktrap[78]. We thereby accounted for differences in the mode of operation and resolution of algorithms. Plotting the number of detected communities against $k$, we were able to select a parameter value at which the results from the community detection were approximately congruent (Supplementary Figure 11).

Since we were interested in the large-scale population structure of ST772, we selected $k$ = 40 and used the low-resolution fast-greedy modularity optimisation to delineate final population subgroups. Community assignments were mapped back to the ML phylogeny of ST772 (Figure 1a). All subgroups agreed with the phylogenetic tree structure and were supported by ≥ 99% bootstrap values (Supplementary Figure 12). One exception was isolate HW_M2760 located within ST772-A2 by phylogenetic analysis, but assigned to ST772-A3 by network analysis (Supplementary Figures 11, 12). This appeared to be an artefact of the algorithm, as its location and connectivity in the network representation matched its phylogenetic position within ST772-A2. The network and communities

were visualized using the Fruchtermann-Reingold algorithm (Figure 1c), excluding samples from the veterinary staff member in Figure 1c (Supplementary Figure 11).

*Local transmission clusters*

We obtained approximate transmission clusters by employing a network approach supplemented with the ML topology and patient data, including date of collection, location of collection and patient family links and travel or family links to South Asia. We used pairwise SNP distances to define a threshold of 4 SNPs, corresponding to the maximum possible SNP distance obtained within one year under a core genome substitution rate of $1.61 \times 10^{-6}$ nucleotide substitutions/site/year. We then constructed the adjacency matrix for a graph, in which isolates were connected by an undirected edge, if they had a distance of less or equal to 4 SNPs. All other isolates were removed from the graph and we mapped the resulting connected components to the ML phylogeny, showing that in each case the clusters were also reconstructed in the phylogeny, where isolates diverged from a recent common ancestor (gray highlights, Supplementary Figure 6). We then traced the identity of the connected components in the patient meta-data and added this information to each cluster. NICU clusters were reconstructed under these conditions.

*Antimicrobial resistance, virulence factors and pan-genome*

Mykrobe predictor was employed for antibiotic susceptibility prediction and detection of associated resistance determinants and mutations. Mykrobe predictor has demonstrated sensitivity and specificity > 99% for predicting phenotypic resistance and is comparable to gold-standard phenotyping in *S. aureus*[55]. Predicted phenotypes were therefore taken as a strong indication for actual resistance phenotypes in ST772. Genotype predictions also reflect multidrug resistance profiles (aminoglycosides, β-lactams, fluoroquinolones, MLS, trimethoprim) reported for this clone in the literature[14–16,20,79,80]. As most resistance-associated MGEs in the complete reference genome DAR4145 are mosaic-like and flanked by repetitive elements[16], we used specific diagnostic genes present as complete single copies in the reference annotation of DAR4145[16] to define presence of the IRP (*msrA*) and Tn*4001* (*aacA-aphD*). Mykrobe predictor simultaneously called the *grlA* mutations S80F and S80Y for quinolone resistant phenotypes. However, in all cases one of the variants was covered at extremely low median k-mer depth (< 20) and we consequently assigned the variant with higher median k-mer depth at *grlA* (Supplementary Table 6).

ARIBA[81] with default settings and the core Virulence Factor database were used to detect the

complement of virulence factors in ST772. We corroborated and extended our results with detailed *in-silico* microarray typing, including the presence of the *egc* gene cluster or *S. aureus* specific virulence factors such as the enterotoxin homologue ORF CM14. Differences in detection of relevant virulence factors between the *in silico* typing and ARIBA included, amongst others, *lukS/F-PVL* (337 vs. 336), *sea* carried on the φ-IND772 prophage (336 vs. 326), *sec* (333 vs 328) and *sak* (1 vs. 2). Since *in silico* microarray typing was based on assembled genomes and may therefore be prone to assembly errors, we used results from the read-based typing with ARIBA to assess statistical significance of virulence factors present in basal strains and ST772-A (Supplementary Figure 7).

Pan-genome analysis was conducted using Prokka annotated assemblies in Roary[82], with minimum protein BLAST identity at 95% and minimum percentage for a gene to be considered core at 99% (Supplementary Figure 13). A nucleotide BLAST comparison between the extrachromosomal plasmid 11809-03 of USA300, the integrated resistance plasmid in the ST772 reference genome DAR4145 and the integrated plasmid region in strain 11819-07 of ST80 was plotted with geneD3 (https://github.com/esteinig/geneD3/), showing segments > 1kb (Supplementary File 1). A gene synteny comparison between major SCC*mec* types was plotted with genoPlotR[83] (Supplementary Figure 8)

*Growth curves*

*S. aureus* strains were grown overnight in 5 mL tryptic soy broth (TSB, Fluka) with shaking (180 *rpm*) at 37 °C. Overnight cultures were diluted 1:1000 in fresh TSB and 200 µL was added to a 96 – well plate (Costar) in triplicate. Growth was measured 37 °C, with shaking (300 *rpm*) using a FLUOROstar fluorimeter (BMG Labtech) using an absorbance wavelength of 600 nm. Growth curves represent the mean of triplicate results.

*Cell culture conditions*

The monocyte-macrophage THP-1 cell line was maintained in suspension in 30 mL Roswell Park Memorial Medium Institute (RPMI-1640) medium, supplemented with 10% heat-inactivated fetal bovine serum (FBS), 1 µM L-glutamine, 200 units/mL penicillin, and 0.1 mg/mL streptomycin at 37 °C in a humidified incubator with 5% $CO_2$. Cells were harvested by centrifugation at 700 x *g* for 10 min at room temperature and re-suspended to a final density of 1–1.2 x $10^6$ cells/mL in tissue-grade phosphate buffered saline, typically yielding >95 % viable cells as determined by easyCyte flow cytometry (Millipore).

Human erythrocytes were harvested from 10 mL of human blood following treatment in sodium heparin tubes (BD). Whole blood was centrifuged at 500 x *g* for 10 min at 4 °C. Supernatant (plasma) was aspirated and cells were washed twice in 0.9 % NaCl and centrifuged at 700 x *g* for 10 min. Cell pellet was gently re-suspended in 0.9 % NaCl and diluted to 1 % (v/v).

*Cytotoxicity assay*

To monitor *S. aureus* toxicity, *S. aureus* strains were grown overnight in TSB, diluted 1:1000 in 5 mL fresh TSB and grown for 18 h at 37 °C with shaking (180 *rpm*). Bacterial supernatants were prepared by centrifugation of 1 mL of bacterial culture at 20,000 x *g* for 10 min. For assessing toxicity to THP-1 cells, 20 µL of cells were incubated with 20 µL of bacterial supernatant and incubated for 12 min at 37 °C. Both neat and 30% diluted supernatant (in TSB) were used as certain *S. aureus* strains were considerably more toxic than others. Cell death was quantified using easyCyte flow cytometry using the Guava viability stain according to manufacturer's instructions. Experiments were done in triplicate. For assessing haemolysis, 150 µL of 1% (v/v) erythrocytes were incubated with 50 µl of either neat and 30% supernatant in a 96 well plate for 30 min at 37°C. Plates were centrifuged for 5 min at 300 x *g* and 75 µL of supernatant was transferred to a new plate and absorbance was measured at 404nm using a FLUOROstar fluorimeter (BMG Labtech). Normalised fluorescence was achieved using the equation $(A_t – A_0) / (A_m / A_0)$ where $A_t$ is the haemolysis absorbance value of a strain, $A_0$ is the minimum absorbance value (negative control of 0.9% NaCl) and Am is the maximum absorbance value (positive control of 1 % triton X-100).

*Lipase assay*

Bacterial supernatants used in the above cytotoxicity assays were also used to assess lipase activity, using the protocol published by Cadieux *et al*. [84] with modifications. Briefly, 8mM *para*-nitrophenyl butyrate (pNPB), the short chain substrate, or *para*-nitrophenyl palmitate (pNPP), the long chain substrate, (Sigma) was mixed with a buffer (50mM Tris-HCl (pH 8.0), 1mg/ml gum Arabic and 0.005% Triton-X100) in a 1:9 ratio to create assay mixes. A standard curve using these assay mixes and *para*-nitrophenyl (pNP) (Sigma) was created, and 200µl of each dilution was pipetted into one well of a 96-well plate (Costar). 180µl of the assay mixes was pipetted into the remaining wells of a 96-well plate, and 20µl of the harvested bacterial supernatant was mixed into the wells. The plate was placed in a FLUOstar Omega microplate reader (BMG Labtech) at 37°C, and a reading at 410nm was taken every 5 min.s for 1h. The absorbance readings were converted to µM pNP released/min. using

588 the standard curve.

589

590 *Biofilm formation*

591

592 Semi-quantitative measurements of biofilm formation on 96-well, round-bottom, polystyrene plates
593 (Costar) was determined based on the classical, crystal violet method of Ziebuhr et al.[85]. 18 h bacterial
594 cultures grown in TSB were diluted 1:40 into 100 µL TSB containing 0.5 % glucose. Perimeter wells
595 of the 96-well plate were filled with sterile $H_2O$ and plates were placed in a separate plastic container
596 inside a 37°C incubator and grown for 24 h under static conditions. Following 24 h growth, plates
597 were washed five times in PBS, dried and stained with 150 µL of 1% crystal violet for 30 min at room
598 temperature. Following five washes of PBS, wells were re-suspended in 200 µL of 7% acetic acid,
599 and optical density at 595 nm was measured using a FLUOROstar fluorimeter (BMG Labtech). To
600 control for day to day variability, a control strain (E-MRSA15) was included on each plate in
601 triplicate, and absorbance values were normalised against this. Experiments were done using six
602 technical repeats from 2 different experiments.

603

604 *Statistical analysis*

605

606 All statistical analyses were carried out in R or python and considered significant at $p < 0.05$, except
607 for comparisons of proportions across the multiple virulence and resistance elements, which we
608 considered be statistically significant at $p < 0.01$. Veterinary samples ($n = 39$) were restricted to one
609 isolate (one patient, Staff_E1A) for statistical comparison of region of isolation, proportion of
610 resistance, virulence and MSSA between basal strains and ST772-A ($n = 302$, Main, Figures 3d,
611 Supplementary Figure 7). Differences in pairwise SNP distance and nucleotide diversity between all
612 regions were assessed using non-parametric Kruskal-Wallace test and post-hoc Dunn's test for
613 multiple comparisons with Bonferroni correction, as distributions were assumed to be not normally
614 distributed (Figure 2b, $n = 340$, Supplementary Figure 2). Phenotypic differences were assessed for
615 normality with Shapiro-Wilk tests. We consequently used either Welch's two-sided t-test or the non-
616 parametric two-sided Wilcoxon rank-sum test (Figure 4, Supplementary Table 8).

*Code availability*

Core analyses, including parameter settings, cluster resource configurations and versioned software distributions are reproducible through the *bengal-bay-0.1* workflow, which be found along with other scripts and data files at our GitHub repository (https://github.com/esteinig/ST772). The workflow implements Anaconda virtual environments, including software distributed in the Bioconda[86] channel and is executable through Snakemake[87]. Analyses were conducted on the Cheetah cluster at Menzies School of Health Research, Darwin.

*Data availability*

Short-read sequences have been deposited at ENA under accession numbers detailed in Supplementary Table 2. Additional isolates from India are available from the SRA under accession numbers SRR404118, SRR653209, SRR653212 and SRR747869-SRR747873. Outgroup strains used in the context phylogeny are available from ENA under accession numbers SRR592258 (MW2), ERR217298, ERR217349, ERR221806, ERR266712, ERR279022, ERR279023, ERR278908. ERR279026, ERR716976, ERR717011 (ST573). The ST772 reference genome DAR4145 is available at GenBank under accession number CP010526.1.

*Author contributions*

EJS, ST conducted the bioinformatics analysis; SD performed the dating analysis; SM, PS, PA performed *in silico* typing and provided support for bioinformatics analysis; DS provided support on the computing cluster; MY, ML, RM conducted phenotyping experiments; DAR, DW, AK, RG, ED, RE, SM, MI, MO, GC, AP, GB, AS, DC, AP, AM, HdL, HW, NK, HH, BS, FL, SP, SW, HA, LS, SH provided strains and relevant meta-data; EJS, ST, DAR, SM, MTGH wrote the manuscript; all authors contributed to critical review of the manuscript. ST directed the project with support from SB and JP.

*Acknowledgements*

We thank the library construction, sequencing, and core informatics teams at the Wellcome Trust Sanger Institute. We also extend our gratitude to Anand Manoharan for comments on the manuscript and strains from India. ST is supported by an Australian National Health and Medical Research Council Career Development Award (#1065736). DAR is supported by NIH grant GM080602. DC

652 and AS are supported by an Irish Health Research Board grant HRA-POR-2015-1051. MO is
653 supported by an NHMRC project grant (#1065908).

654

655 *Competing financial interests*

656

657 There are no competing financial interests to declare.

658

659 *Materials and correspondence*

660

661 Steven Y.C. Tong

## References


1. Kumarasamy, K. K. *et al.* Emergence of a new antibiotic resistance mechanism in India, Pakistan, and the UK: a molecular, biological, and epidemiological study. *Lancet Infect. Dis.* **10,** 597–602 (2010).
2. Laxminarayan, R. *et al.* Antibiotic resistance—the need for global solutions. *Lancet Infect. Dis.* **13,** 1057–1098 (2013).
3. Chung The, H. *et al.* South Asia as a reservoir for the global spread of ciprofloxacin-resistant Shigella sonnei: a cross-sectional study. *PLOS Med.* **13,** e1002055 (2016).
4. Holden, M. T. G. *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).
5. He, M. *et al.* Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. *Nat Genet* **45,** 109–113 (2013).
6. Tong, S. Y. C., Davis, J. S., Eichenberger, E., Holland, T. L. & Fowler, V. G. J. *Staphylococcus aureus* infections: epidemiology, pathophysiology, clinical manifestations, and management. *Clin. Microbiol. Rev.* **28,** 603–661 (2015).
7. Suaya, J. A. *et al.* Incidence and cost of hospitalizations associated with *Staphylococcus aureus* skin and soft tissue infections in the United States from 2001 through 2009. *BMC Infect. Dis.* **14,** 296 (2014).
8. Goering, R. V. *et al.* Molecular epidemiology of methicillin-resistant and methicillin-susceptible *Staphylococcus aureus* isolates from global clinical trials. *J. Clin. Microbiol.* **46,** 2842–2847 (2008).
9. Afroz, S. *et al.* Genetic characterization of Staphylococcus aureus isolates carrying Panton-Valentine leukocidin genes in Bangladesh. *Jpn. J. Infect. Dis.* **61,** 393–396 (2008).
10. Chen, C.-J. & Huang, Y.-C. New epidemiology of *Staphylococcus aureus* infection in Asia. *Clin. Microbiol. Infect.* **20,** 605–623 (2014).
11. D'Souza, N., Rodrigues, C. & Mehta, A. Molecular characterization of methicillin-resistant Staphylococcus aureus with emergence of epidemic clones of sequence type (ST) 22 and ST 772 in Mumbai, India. *J. Clin. Microbiol.* **48,** 1806–1811 (2010).
12. Nadig, S. *et al.* Staphylococcus aureus eye infections in two Indian hospitals: emergence of ST772 as a major clone. *Clin. Ophthalmol.* **6,** 165–173 (2012).
13. Manoharan, A. *et al.* An outbreak of post-partum breast abscesses in Mumbai, India caused by ST22-MRSA-IV: genetic characteristics and epidemiological implications. *Epidemiol. Infect.* **140,** 1809–1812 (2012).
14. Blomfeldt, A. *et al.* Emerging multidrug-resistant Bengal Bay clone ST772-MRSA-V in

Norway: molecular epidemiology 2004–2014. *Eur. J. Clin. Microbiol. Infect. Dis.* **36,** 1911–1921 (2017).

15. Chakrakodi, B., Prabhakara, S., Nagaraj, S., Etienne, J. & Arakere, G. High Prevalence of ciprofloxacin resistance in community associated *Staphylococcus aureus* in a tertiary care Indian hospital. *Adv. Microbiol.* **4,** 133–141 (2014).
16. Steinig, E. J. *et al.* Single-molecule sequencing reveals the molecular basis of multidrug-resistance in ST772 methicillin-resistant *Staphylococcus aureus*. *BMC Genomics* **16,** 388 (2015).
17. Ellington, M. J., Ganner, M., Warner, M., Cookson, B. D. & Kearns, A. M. Polyclonal multiply antibiotic-resistant methicillin-resistant *Staphylococcus aureus* with Panton-Valentine leucocidin in England. *J Antimicrob Chemother* **65,** 46–50 (2010).
18. Pokhrel, R. H. *et al.* Detection of ST772 Panton-Valentine leukocidin-positive methicillin-resistant *Staphylococcus aureus* (Bengal Bay clone) and ST22 S.aureus isolates with a genetic variant of elastin binding protein in Nepal. *New Microbes New Infect.* **11,** 20–27 (2017).
19. Madzgalla, S. *et al.* Molecular characterization of Staphylococcus aureus isolates causing skin and soft tissue infections in patients from Malakand, Pakistan. *Eur. J. Clin. Microbiol. Infect. Dis.* **35,** 1541–1547 (2016).
20. Brennan, G. I. *et al.* Emergence of hospital- and community-associated Panton-Valentine leukocidin-positive methicillin-resistant Staphylococcus aureus genotype ST772-MRSA-V in Ireland and detailed investigation of an ST772-MRSA-V cluster in a neonatal intensive care unit. *J. Clin. Microbiol.* **50,** 841–847 (2012).
21. Zanger, P. *et al.* Import and spread of Panton-Valentine Leukocidin-positive Staphylococcus aureus through nasal carriage and skin infections in travelers returning from the tropics and subtropics. *Clin. Infect. Dis.* **54,** 483–492 (2012).
22. Planet, P. J. *et al.* Parallel epidemics of community-associated methicillin-resistant *Staphylococcus aureus* USA300 infection in North and South America. *J. Infect. Dis.* **212,** 1874–82 (2015).
23. Nimmo, G. R. USA300 abroad: global spread of a virulent strain of community-associated methicillin-resistant *Staphylococcus aureus*. *Clin. Microbiol. Infect.* **18,** 725–734 (2012).
24. Stegger, M. *et al.* Origin and evolution of European community-acquired methicillin-resistant *Staphylococcus aureus*. *MBio* **5,** e01044-14 (2014).
25. 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).

26. Castillo-Ramírez, S. *et al.* Phylogeographic variation in recombination rates within a global clone of methicillin-resistant *Staphylococcus aureus*. *Genome Biol.* **13,** 126 (2012).

27. Harris, S. R. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327,** 469–474 (2010).

28. Prabhakara, S. *et al.* Genome sequencing unveils a novel sea enterotoxin-carrying PVL phage in *Staphylococcus aureus* ST772 from India. *PLoS One* **8,** e60013 (2013).

29. Balakuntla, J., Prabhakara, S. & Arakere, G. Novel rearrangements in the staphylococcal cassette chromosome mec type V elements of Indian ST772 and ST672 methicillin resistant *Staphylococcus aureus* strains. *PLoS One* **9,** e94293 (2014).

30. Schijffelen, M. J., Boel, C. H. E., van Strijp, J. A. G. & Fluit, A. C. Whole genome analysis of a livestock-associated methicillin-resistant *Staphylococcus aureus* ST398 isolate from a case of human endocarditis. *BMC Genomics* **11,** 376 (2010).

31. Lee, S. M. *et al.* Fitness cost of staphylococcal cassette chromosome mec in methicillin-resistant Staphylococcus aureus by way of continuous culture. *Antimicrob. Agents Chemother.* **51,** 1497–1499 (2007).

32. Ender, M., McCallum, N., Adhikari, R. & Berger-Bächi, B. Fitness cost of SCCmec and methicillin resistance levels in Staphylococcus aureus. *Antimicrob. Agents Chemother.* **48,** 2295–2297 (2004).

33. Collins, J. *et al.* Offsetting virulence and antibiotic resistance costs by MRSA. *ISME J* **4,** 577–584 (2010).

34. Kennedy, A. D. *et al.* Complete nucleotide sequence analysis of plasmids in strains of Staphylococcus aureus clone USA300 reveals a high level of identity among isolates with closely related core genome sequences. *J. Clin. Microbiol.* **48,** 4504–4511 (2010).

35. Cheung, A. I., Projan, S. J., Edelstein, R. E. & Fischetti, V. A. Cloning, expression, and nucleotide sequence of a *Staphylococcus aureus* gene (fbpA) encoding a fibrinogen-binding protein. *Infect. Immun.* **63,** 1914–1920 (1995).

36. White, M. J., Boyd, J. M., Horswill, A. R. & Nauseef, W. M. Phosphatidylinositol-specific phospholipase C contributes to survival of *Staphylococcus aureus* USA300 in human blood and neutrophils. *Infect. Immun.* **82,** 1559–1571 (2014).

37. Truong-Bolduc, Q. C., Villet, R. A., Estabrooks, Z. A. & Hooper, D. C. Native efflux pumps contribute resistance to antimicrobials of skin and the ability of Staphylococcus aureus to colonize skin. *J. Infect. Dis.* **209,** 1485–1493 (2014).

38. Prabhakara, S. *et al.* Draft genome sequence of Staphylococcus aureus 118 (ST772), a major disease clone from India. *J. Bacteriol.* **194,** 3727–3728 (2012).

39. Prabhakara, S. *et al.* Genome sequencing unveils a novel sea enterotoxin-carrying PVL

767 phage in *Staphylococcus aureus* ST772 from India. *PLoS One* **8,** e60013 (2013).
768 40. Paterson, G. K. *et al.* Capturing the cloud of diversity reveals complexity and heterogeneity
769 of MRSA carriage, infection and transmission. *Nat. Commun.* **6,** 6560 (2015).
770 41. Morrison, M. A., Hageman, J. C. & Klevens, R. M. Case definition for community-
771 associated methicillin-resistant *Staphylococcus aureus*. *J. Hosp. Infect.* **62,** 241 (2006).
772 42. Fridkin, S. K. *et al.* Methicillin-resistant *Staphylococcus aureus* disease in three
773 communities. *N. Engl. J. Med.* **352,** 1436–1444 (2005).
774 43. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
775 sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
776 44. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using
777 exact alignments. *Genome Biol.* **15,** 1–12 (2014).
778 45. Bankevich, A. SPAdes: a new genome assembly algorithm and its applications to single-cell
779 sequencing. *J Comput Biol* **19,** 455–477 (2012).
780 46. Song, L., Florea, L. & Langmead, B. Lighter: fast and memory-efficient sequencing error
781 correction without counting. *Genome Biol.* **15,** 509 (2014).
782 47. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve
783 genome assemblies. *Bioinformatics* **27,** 2957–2963 (2011).
784 48. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
785 *ArXiv* (2013).
786 49. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–
787 2079 (2009).
788 50. Deorowicz, S., Kokot, M., Grabowski, S. & Debudaj-Grabysz, A. KMC 2: fast and resource-
789 frugal k-mer counting. *Bioinformatics* **31,** 1569–1576 (2015).
790 51. Walker, B. J. *et al.* Pilon: An Integrated Tool for Comprehensive Microbial Variant
791 Detection and Genome Assembly Improvement. *PLoS One* **9,** e112963 (2014).
792 52. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068–2069
793 (2014).
794 53. Monecke, S. *et al.* A field guide to pandemic, epidemic and sporadic clones of methicillin-
795 resistant *Staphylococcus aureus*. *PLoS One* **6,** e17936 (2011).
796 54. Monecke, S. *et al.* Diversity of SCCmec Elements in *Staphylococcus aureus* as Observed in
797 South-Eastern Germany. *PLoS One* **11,** e0162654 (2016).
798 55. Bradley, P. *et al.* Rapid antibiotic-resistance predictions from genome sequence data for
799 *Staphylococcus aureus* and *Mycobacterium tuberculosis*. *Nat. Commun.* **6,** 10063 (2015).
800 56. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide
801 polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w(1118);

802 iso-2; iso-3. *Fly (Austin).* **6,** 80–92 (2012).
803 57. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing.
804 *ArXiv* (2012).
805 58. Stamatakis, A. RAxML Version 8: a tool for phylogenetic analysis and post-analysis of large
806 phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).
807 59. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree
808 display and annotation. *Bioinformatics* **23,** 127–128 (2007).
809 60. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial
810 whole genome sequences using Gubbins. *Nucleic Acids Res.* **43,** 15 (2015).
811 61. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R
812 language. *Bioinformatics* **20,** 289–290 (2004).
813 62. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things).
814 *Methods Ecol. Evol.* **3,** 217–223 (2012).
815 63. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for
816 visualization and annotation of phylogenetic trees with their covariates and other associated
817 data. *Methods Ecol. Evol.* **8,** 28–36 (2017).
818 64. Jombart, T., Balloux, F. & Dray, S. adephylo: new tools for investigating the phylogenetic
819 signal in biological traits. *Bioinformatics* **26,** 1907–1909 (2010).
820 65. To, T.-H., Jung, M., Lycett, S. & Gascuel, O. Fast dating using least-squares criteria and
821 algorithms. *Syst. Biol.* **65,** 82–97 (2016).
822 66. Duchêne, S., Geoghegan, J. L., Holmes, E. C. & Ho, S. Y. W. Estimating evolutionary rates
823 using time-structured data: a general comparison of phylogenetic methods. *Bioinformatics*
824 **32,** 3375–3379 (2016).
825 67. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood
826 phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59,** 307–321 (2010).
827 68. Korber, B. *et al.* Timing the ancestor of the HIV-1 pandemic strains. *Science (80-. ).* **288,**
828 1789–1796 (2000).
829 69. Ramsden, C., Holmes, E. C. & Charleston, M. A. Hantavirus evolution in relation to its
830 rodent and insectivore hosts: no evidence for codivergence. *Mol. Biol. Evol.* **26,** 143–153
831 (2009).
832 70. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure
833 of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2,** vew007
834 (2016).
835 71. Duchêne, S., Duchêne, D., Holmes, E. C. & Ho, S. Y. W. The performance of the date-
836 randomization test in phylogenetic analyses of time-structured virus data. *Mol. Biol. Evol.* **32,**

1895–1906 (2015).

72. Murray, G. G. R. *et al.* The effect of genetic structure on molecular dating and tests for temporal signal. *Methods Ecol. Evol.* **7,** 80–89 (2016).

73. Stucki, D. *et al. Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages. *Nat Genet* **48,** 1535–1543 (2016).

74. Neuditschko, M., Khatkar, M. S. & Raadsma, H. W. NetView: a high-definition network-visualization approach to detect fine-scale population structures from genome-wide patterns of variation. *PLoS One* **7,** e48375 (2012).

75. Steinig, E. J., Neuditschko, M., Khatkar, M. S., Raadsma, H. W. & Zenger, K. R. NetView P : a network visualization tool to unravel complex population structure using genome-wide SNPs. *Mol. Ecol. Resour.* **16,** 216–227 (2016).

76. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 7821–7826 (2002).

77. Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 1118–1123 (2008).

78. Pons, P. & Latapy, M. Computing communities in large networks using random walks. *J. Graph Algorithms Appl.* **10,** 191–218 (2006).

79. Ellington, M. J., Ganner, M., Warner, M., Cookson, B. D. & Kearns, A. M. Polyclonal multiply antibiotic-resistant methicillin-resistant *Staphylococcus aureus* with Panton-Valentine leucocidin in England. *J. Antimicrob. Chemother.* **65,** 46–50 (2010).

80. Shore, A. C. *et al.* Panton-Valentine leukocidin-positive Staphylococcus aureus in Ireland from 2002 to 2011: 21 clones, frequent importation of clones, temporal shifts of predominant methicillin-resistant S. aureus clones, and increasing multiresistance. *J. Clin. Microbiol.* **52,** 859–870 (2014).

81. Hunt, M. *et al.* ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb. Genomics* **3,** e000131 (2017).

82. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31,** 3691–3693 (2015).

83. Guy, L., Roat Kultima, J. & Andersson, S. G. E. genoPlotR: comparative gene and genome visualization in R. *Bioinforma.* **26,** 2334–2335 (2010).

84. Cadieux, B., Vijayakumaran, V., Bernards, M. A., McGavin, M. J. & Heinrichs, D. E. Role of lipase from community-associated methicillin-resistant *Staphylococcus aureus* strain USA300 in hydrolyzing triglycerides into growth-inhibitory free fatty acids. *J. Bacteriol.* **196,** 4044–4056 (2014).

85. Ziebuhr, W. *et al.* Detection of the intercellular adhesion gene cluster (ica) and phase

variation in *Staphylococcus epidermidis* blood culture strains and mucosal isolates. *Infect. Immun.* **65,** 890–896 (1997).

86. Dale, R. *et al.* Bioconda: A sustainable and comprehensive software distribution for the life sciences. *bioRxiv* (2017).

87. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28,** 2520–2522 (2012).