Antimicrobial resistance determinants are associated with Staphylococcus aureus bacteraemia and adaptation to the healthcare environment: a bacterial genome-wide association study

Bernadette C Young, Chieh-Hsi Wu, Jane Charlesworth, Sarah Earle, James R Price, N Claire Gordon, Kevin Cole, Laura Dunn, Elian Liu, Sarah Oakley, Heather Godwin, Rowena Fung, Ruth Miller, Kyle Knox, Antonina Votintseva, T Phuong Quana, Robert Tilley, Matthew Scarborough, Derrick W Crook, Timothy E Peto, A Sarah Walker, Martin Llewelyn, Daniel J Wilson

Publication date
27-11-2021

Document Version
Published version

Published in
Microbial Genomics

Link to external publisher version
https://doi.org/10.1099/mgen.0.000700

MICROBIAL GENOMICS

RESEARCH ARTICLE

Young *et al.*, *Microbial Genomics* 2021;7:000700

DOI 10.1099/mgen.0.000700

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Antimicrobial resistance determinants are associated with *Staphylococcus aureus* bacteraemia and adaptation to the healthcare environment: a bacterial genome-wide association study

Bernadette C. Young[1,2,*], Chieh-Hsi Wu[1], Jane Charlesworth[1], Sarah Earle[3], James R. Price[4,5], N. Claire Gordon[1,2], Kevin Cole[4,5], Laura Dunn[2], Elian Liu[2], Sarah Oakley[2], Heather Godwin[1], Rowena Fung[1], Ruth Miller[1], Kyle Knox[6], Antonina Votintseva[1], T. Phuong Quan[1,7,8], Robert Tilley[9], Matthew Scarborough[2], Derrick W. Crook[1,2,7,8], Timothy E. Peto[1,2,7,8], A. Sarah Walker[1,7,8]†, Martin J. Llewelyn[4,5]† and Daniel J. Wilson[3]†

Abstract

Staphylococcus aureus is a major bacterial pathogen in humans, and a dominant cause of severe bloodstream infections. Globally, antimicrobial resistance (AMR) in *S. aureus* remains challenging. While human risk factors for infection have been defined, contradictory evidence exists for the role of bacterial genomic variation in *S. aureus* disease. To investigate the contribution of bacterial lineage and genomic variation to the development of bloodstream infection, we undertook a genome-wide association study comparing bacteria from 1017 individuals with bacteraemia to 984 adults with asymptomatic *S. aureus* nasal carriage. Within 984 carriage isolates, we also compared healthcare-associated (HA) carriage with community-associated (CA) carriage. All major global lineages were represented in both bacteraemia and carriage, with no evidence for different infection rates. However, kmers tagging trimethoprim resistance-conferring mutation F99Y in *dfrB* were significantly associated with bacteraemia-vs-carriage (P=$10^{-8.9}$-$10^{-9.3}$). Pooling variation within genes, bacteraemia-vs-carriage was associated with the presence of *mecA* (HMP=$10^{-5.3}$) as well as the presence of SCCmec (HMP=$10^{-4.4}$). Among *S. aureus* carriers, no lineages were associated with HA-vs-CA carriage. However, we found a novel signal of HA-vs-CA carriage in the foldase protein *prsA*, where kmers representing conserved sequence allele were associated with CA carriage (P=$10^{-7.1}$-$10^{-19.4}$), while in *gyrA*, a ciprofloxacin resistance-conferring mutation, L84S, was associated with HA carriage (P=$10^{-7.2}$). In an extensive study of *S. aureus* bacteraemia and nasal carriage in the UK, we found strong evidence that all *S. aureus* lineages are equally capable of causing bloodstream infection, and of being carried in the healthcare environment. Genomic variation in the foldase protein *prsA* is a novel genomic marker of healthcare origin in *S. aureus* but was not associated with bacteraemia. AMR determinants were associated with both bacteraemia and healthcare-associated carriage, suggesting that AMR increases the propensity not only to survive in healthcare environments, but also to cause invasive disease.

Received 27 May 2021; Accepted 30 September 2021; Published 23 November 2021

Author affiliations: [1]Nuffield Department of Medicine, Experimental Medicine Division, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DU, UK; [2]Microbiology and Infectious Diseases Department, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford OX3 9DU, UK; [3]Big Data Institute, Nuffield Department of Population Health, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Old Road Campus, Oxford, OX3 7LF, UK; [4]Department of Infectious Diseases and Microbiology, Royal Sussex County Hospital, Brighton BN2 5BE, UK; [5]Department of Global Health and Infection, Brighton and Sussex Medical School, University of Sussex, Falmer BN1 9PS, UK; [6]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK; [7]National Institute for Health Research, Oxford Biomedical Research Centre, Oxford, UK; [8]NIHR Health Protection Unit in Healthcare Associated Infections and Antimicrobial Resistance at University of Oxford in partnership with Public Health England, Oxford, UK; [9]Department of Microbiology, University Hospitals Plymouth NHS Trust, Derriford Hospital, Plymouth PL6 8DH, UK.
***Correspondence:** Bernadette C. Young, bernadette.young@ndm.ox.ac.uk

Keywords: Bacterial pathogens; Bacteraemia; microbial genomics; microbial epidemiology; nosocomial infection.

Abbreviations: Agr, accessory gene regulator; AMR, antimicrobial resistance; CA, community associated; CC, clonal complex; GWAS, genome wide association study; HA, healthcare associated; HMP, harmonic mean *p*-value; LD, inkage disequilibrium; MLST, Multi locus sequence type; MRSA, Methicillin-resistant *S. aureus*; OR, odds ratio; PBP, Penicillin binding proteins; PVL, Panton Valentine leucocidin; Rsp, repressor surface protein; SAB, *Staphylococcus aureus* bacteraemia; *S. aureus*, *Staphylococcus aureus*; WGS, whole genome sequencing.

Sequenced bacterial isolates deposited in Short Read Archive. 7 isolates included as part of another study are found under accession number PRJNA369475 and the remaining 1994 accession number PRJNA690682. A complete listing of sequences is contained in Table S5.

†These authors contributed equally to this work

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Five supplementary tables and eleven supplementary figures are available with the online version of this article.

DATA SUMMARY

New data – submitted to NCBI SRA accession number PRJNA690682.

All sequenced bacterial isolates are deposited in Short Read Archive. Nine isolates included as part of another study are found under accession number PRJNA369475 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA369475) and sequences of the remaining isolates under accession number PRJNA690682 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA690682). A complete listing of sequences is contained in supplementary material (Table S5, available in the online version of this article).

The authors confirm all supporting data, code and protocols have been provided within the article or through supplementary data files.

Impact Statement

Bacteria like *Staphylococcus aureus* are part of our normal flora, carried by one in three adults, and are also found in serious, life-threatening disease like bloodstream infection (bacteraemia). Many studies have examined whether any part of the genetic code of *S. aureus* might increase the chances of these serious infections, and the results of these have been conflicting. In particular lineages (i.e. related strains), the presence of some genes and genetic mutations have all been reported as possibly increasing the risk, but these reports conflict. This may partly be because the population structure of *S. aureus* makes it difficult to disentangle true associations from false.

Here we compare 2001 *S. aureus* from bacteraemia and carriage, using methods which account for those challenges. We find all lineages are equally likely to cause bacteraemia, but some genes and mutations encoding antimicrobial resistance (AMR) are found more often in bacteraemia, while distinct AMR mutation factors are associated with healthcare-associated carriage.

Our study settles persisting questions about the propensity of differing lineages to cause bloodstream infections. It further raises the suggestion that AMR may predispose to serious infection, a possibility which should increase the urgency of our global fight against the growing threat of AMR.

BACKGROUND

Staphylococcus aureus is a common coloniser of human mucosal surfaces and skin but also a major human pathogen [1, 2]. It is a leading cause of hospital and community acquired infection and one of the leading causes of bloodstream infection worldwide [1–4]. Over 12000 cases occur each year in England, and the rate continues to increase, despite improving control of methicillin-resistant *S. aureus* (MRSA) bacteraemia [5]. Mortality following *S. aureus* bacteraemia (SAB) has not declined over recent years, and remains generally over 20% at 30 days, even with appropriate antimicrobial therapy [6–8].

S. aureus possesses diverse and variable virulence mechanisms facilitating tissue invasion, inflammation and evasion of host immune factors. These include a thick peptidoglycan wall, polysaccharide capsule, toxins [9, 10], complement control proteins [11], and bound adhesins [12]. With the exception of specific toxinoses such as toxic shock syndrome [13–15], and the role of Panton-Valentine leucocidin (PVL) in skin or soft tissue infections [16] and pyomyositis [17], evidence linking bacterial genetic variability to clinical disease phenotype is inconclusive.

The majority of evidence accrues from case-control studies using candidate gene approaches or microarrays to examine gene presence or absence. Such studies have implicated several specific genes encoding putative virulence factors in invasive *S. aureus* disease. Secreted enterotoxins [18–21], haemolysins [18] and leucotoxins [22]; surface proteins which mediate tissue attachment, invasion and immune evasion [18, 22], the presence of the intercellular adhesin locus [18, 23] and variation in the accessory gene regulator (*agr*) system [22] have all been shown to co-occur with invasive *S. aureus*. However, the evidence for these associations is inconsistent, and for every study reporting an association, there is at least one large study that shows no evidence of effect [18, 24].

The fact that a few *S. aureus* lineages account for the majority of *S. aureus* infections suggest important inter-lineage differences in virulence but again evidence is conflicting for a relationship between CC and invasive disease [22, 24–29]. These discrepant results may reflect the variable sensitivity of probes employed, or the inconsistent methods used to control for effects of population structure. In particular, associations cannot be reliably inferred without considering linkage disequilibrium between candidate genes and potential virulence factors elsewhere in the genome.

Different genomic associations have been identified for community-associated (CA) or healthcare-associated (HA) SAB [30]. Loss of function mutations in *agr*, a central controller of *S. aureus* expression, have been frequently found in HA bloodstream infections [31]. Reduced cytotoxicity and low *agr* expression were also independent predictors of mortality in a study of nosocomial MRSA pneumonia [32]. There is a direct relationship between *mecA* expression and *agr* dysfunction: the altered penicillin binding protein (PBP) expressed by some MRSA can directly reduce *agr*-mediated toxin production [33]. This dysfunction may be a fitness 'cost', overcome by the relative advantage of antimicrobial resistance in healthcare settings, and the relatively lower host defences found in hospitalised patients. However, evidence shows MRSA clones traditionally conceived of as HA lineages – such as CC-22 - are equally capable of transmission in the community, including household settings [34], questioning the notion that MRSA requires the healthcare setting to gain a relative advantage.

There is growing evidence that the relationship between toxin production and virulence in *S. aureus* is not straightforward. While superantigen toxins and leukocidins have been linked to certain disease phenotypes, genomic changes associated with reduced bacterial toxicity may actually enhance bacterial survival in the bloodstream, evidenced by lower lymphocyte toxicity and greater fitness in human serum exhibited by bacteraemia isolates compared to those found in nasal carriage or soft tissue infection [35]. *agr* defective strains have been found in association with persistent bacteraemia [36, 37], and associated with higher mortality [38]. Naturally occurring loss of function mutations in the regulatory protein repressor of surface proteins (Rsp) have been documented arising within-host and in bloodstream infections [39]. These mutations were associated with attenuated mortality in a mouse model of disease, but preserved the ability to disseminate and form abscesses [40], and have been shown to alter survival in blood and antibiotic tolerance [41]. A comprehensive survey of within-host evolution of *S. aureus* infection demonstrated evidence of bacterial genomic adaption, with protein altering variation in regulatory genes, and the cell surface proteins under control of key regulators [42]. Similar signals of adaptation were found in the genetic changes associated with prolonged bacteraemia [43].

Thus, while conflicting evidence for the role of gene presence in bacteraemia-vs-carriage arises from case-control studies, there are observations supporting the hypothesis that subtle genetic variation – including that of a type traditionally thought to diminish virulence – could increase the likelihood of SAB [35–44]. Recent developments in bacterial genome wide association studies (GWAS) demonstrate that these powerful tools can help delineate the genomic basis of bacterial infection. A study of bacteraemia caused by the ST-239 lineage investigated associations between bacterial genetic variants, toxin production and severity of disease in a mouse model [45]. Conversely, GWAS of *S. aureus* lineage CC-45 did not identify genomic predictors of bacteraemia-vs-carriage [46]. Genomic variants were integrated with bacterial phenotyping and clinical data in 300 adults with bacteraemia involving the CC-22 and CC-30 lineages, finding that bacterial predictors of mortality varied by lineage [47]. Investigating whether a genomic basis for invasive disease exists more generally at a population level requires careful control for population structure, an otherwise potent confounder. Bacterial GWAS incorporating such controls has recently identified PVL as the key determinant of *S. aureus* pyomyositis in a paediatric population [17].

Here we present a bacterial GWAS of SAB across bacterial lineages, studying population-representative cases of bacteraemia and nasal carriage controls, integrating clinical data with 2001 bacterial sequences to investigate whether bacterial lineage or genomic variation is associated with bacteraemia-vs-carriage. Within *S. aureus* carriage, we further examine genomic features associated with HA-vs-CA carriage.

METHODS

Identification of cases and controls

Cases of *S. aureus* bacteraemia (SAB) were identified from three UK hospital trusts between 2008–2014: Oxford University Hospitals NHS trust (Oxford UK), Brighton and Sussex University Hospitals NHS trust (Brighton, UK) and University Hospitals Plymouth NHS trust (Plymouth, UK). These sites were part of the UK Clinical Infection Research Group (UKCIRG) which established prospective cohort study of SAB in 2008 [6], and the International *Staphylococcus Aureus* Collaboration (ISAC) which established a multinational prospective cohort study of SAB in 2006 [7]. Sequential individuals from these studies over 13 years of age with *S. aureus* on blood culture were included if there was an isolate available for sequencing, with associated clinical data, and the blood culture had not been deemed to be a contaminant on local clinical review. We identified 1203 cases in patients that were not contaminants, after excluding repeat episodes (775, 232 and 196 at each centre respectively). Bacterial isolates were found from 724, 207 and 163 episodes at each centre respectively, and a minimum clinical data set (see below) was available for 674, 187 and 160 cases. We successfully sequenced 1017 of these 1021 cases for inclusion. These included 417 cases from a previously sequenced collection investigating identifying antimicrobial resistance [48] (Fig. S1).

S. aureus isolates from nasal carriage in individuals without *S. aureus* infection were identified from two studies of *S. aureus* carriage in Oxfordshire, UK (Fig. S2).

The first was a study of *S. aureus* nasal carriage in adults in the community between July 2009 and April 2013 [49]. Of 1123 individuals enrolled, 360 individuals carrying *S. aureus* at recruitment and 211 swab-negative individuals were invited to supply nasal swabs at two-monthly intervals [49]. Where co-habiting individuals carried the same *spa*-type, only one individual was considered for inclusion to avoid over-sampling of strains with household transmission. The second was an investigation of nosocomial carriage and transmission at the John Radcliffe Hospital, Oxford between September 2009 and August 2011 [50]. Individuals admitted to three study wards had nasal swabs for *S. aureus* carriage performed on admission and at fortnightly intervals until discharge. In total, 1146 individuals were found to have *S. aureus* carriage on any swab during the study (enrolled from Intensive Treatment Unit (ITU; 729 individuals), Trauma unit (352) or one of two elderly care wards (65)). Carriers with *S. aureus* originally isolated on the first or second day of admission and no overnight stay in the preceding 12 months were classified as community-associated controls (269). Carriers with *S. aureus* originally isolated more than 2 days after admission and carriers who had been admitted for three or more nights in the preceding 28 days were classified as healthcare-associated controls (335). All carriers with *S. aureus* isolated from a clinical sample in the previous 12 months were excluded. In total 984 asymptomatic carriage controls were successfully sequenced (Fig. S2). For carriers

with multiple positive samples, the latest sample of the longest carried *spa*-type was selected.

Epidemiological data

For each episode of SAB, the minimum dataset for inclusion was patient gender and age at the time of infection, date of admission, the number of days between admission and first blood culture from which *S. aureus* was cultured, and the number of days since the most recent discharge from hospital. If available, the clinically-determined focus of infection and vital status 90 days after infection were recorded. Epidemiological data on episodes of SAB was collected as part of on-going service evaluation studies, as part of multi-centre collaborations with the UK Clinical Infection Research Group (UKCIRG) [6] and the International *Staphylococcus aureus* Collaboration (ISAC) [7]. Further data, including for carriage controls were obtained from the Infections in Oxfordshire Research Database (IORD) which links information about patient attendances with results from pathology services in an anonymised research database [51].

Cases were deemed healthcare-associated (HA) if the first blood culture positive for *S. aureus* was collected on the third day or later of a hospital admission (healthcare onset cases), or if the patient had an inpatient admission in the previous 90 days (community onset, healthcare-associated cases) [52]. Cases were deemed community-associated (CA) if the first blood culture positive for *S. aureus* was collected on the first or second calendar day of admission, and there was no inpatient admission in the previous 90 days.

Microbiological methods

S. aureus isolates from blood culture were characterised using standard operating procedures of clinical laboratories at all three centres. Isolates for inclusion were retrieved and frozen in 15% glycerol stock prior to DNA extraction.

Hospital carriage swabs were collected by research nurses using dry cotton-tipped swabs. Community carriage study participants self-collected swabs, returning these by mail as previously described [49]. All swabs were incubated in 5% saline enrichment broth (Oxoid LTD, Basingstoke, UK) overnight at 37 °C before subculture onto SaSelect chromogenic agar (Bio-Rad, Watford, UK). Plates were examined after 24 h incubation and potential *S. aureus* colonies confirmed by catalase, DNase and Prolex Staph Xtra Latex kit (Pro-Lab Diagnostics, Birkenhead, UK). Isolates were stored at −80 °C in 15% glycerol. Isolates were *spa*-typed as previously reported [49].

Whole genome sequencing

For each bacterial culture, a single colony was sub-cultured and DNA was extracted from the sub-cultured plate using a mechanical lysis step on FastPrep homogeniser (MPBiomedicals, Santa Ana, CA) followed by extraction with the Quickgene-mini80 device (Autogen Inc, Holliston, MA), and sequenced at the Wellcome Trust Centre for Human Genetics, Oxford. Six hundred isolates were sequenced on the Illumina HiSeq 2500 platform (San Diego, California, USA), with paired-end reads 150 base pairs long. A further 417 isolates were sequenced for an earlier study, 26 on with Illumina HiSeq 2500, and 391 on the HiSeq 2000 platform, with paired-end reads of 99 base pairs.

Variant calling

Following established methods [17], we used Velvet [53] v1.0.18 to assemble reads into contigs *de novo*. Velvet Optimiser v2.1.7 was used to choose the kmer lengths on a per sequence basis. The median kmer length for assembly was 123, however this was affected by sequencing read length, being significantly lower for assemblies based on 99 bp reads (median k=69) than those based on 150 bp reads (median k=125) ($P<10^{-5}$, Wilcoxon rank sum test).

We used BLAST [54] to find the relevant loci, and defined multilocus sequence type (MLST) using the online database at http://saureus.mlst.net/ Strains that shared six of seven MLST loci were considered to belong to the same CC. Antibiotic sensitivity was predicted by interrogating the assemblies for a panel of resistance determinants as previously described [48].

We used Stampy [55] v1.0.22 to map reads against a reference genome (MRSA252, Genbank accession number NC_002952) [56]. Repetitive regions, defined by BLAST [54] comparison of the reference genome against itself, were masked prior to variant calling. Bases were called at each position using previously described quality filters [39, 57, 58]. Missing calls were imputed using ClonalFrameML [59].

Reconstructing the phylogenetic tree

We constructed a maximum likelihood phylogeny of mapped genomes for visualization using RAxML [60] assuming a general time reversible (GTR) model, and fine-tuned the estimates of branch lengths using ClonalFrameML [59].

Kmer counting

We used a kmer-based approach to capture non-SNP variation [61]. Using the *de novo* assembled genomes, all unique 31 base haplotypes were counted using dsk [62]. If a kmer was found in the assembly it was counted (i.e. determined to be present for that genome), otherwise it was treated as absent. This produced a set of 23 860 793 variably present kmers, with the presence or absence of each determined per isolate. We identified a median of 2 760 000 kmers per isolate, including variably present kmers and kmers common to all genomes (IQR 2725000–2806000). The number of kmers found per isolate did not differ significantly with sequencing platform (P=0.4, Wilcoxon rank sum test). From a smaller set of 1610 isolates sequenced with 150 bp reads, we identified 22 284 204 variably present kmers.

Calculating heritability

We used the Genome-wide Efficient Mixed Model Association tool GEMMA [63] to fit a linear mixed model for association between a single phenotype (bacteraemia vs asymptomatic nasal carriage, [encoded as 1 and 0, retrospectively]). We

calculated the relatedness matrix from biallelic SNP and kmer presence for tests of each allele type. We used GEMMA to estimate the proportion of variance in phenotypes explained by genotypic diversity (i.e. heritability).

Genome wide association testing of SNPs and kmers

We performed association testing using an R package bacterialGWAS (https://github.com/jessiewu/bacterialGWAS), which implements a published method for locus testing in bacterial GWAS [64]. The association between each SNP and kmer with the phenotype was tested controlling for population structure and genetic background using the linear mixed model (LMM) implemented in GEMMA [63]. We included healthcare or community origin of case/control status as a fixed covariate in the model when testing for associations with the bacteraemia-vs-carriage phenotype. The parameters of the linear mixed model were estimated by maximum likelihood and a likelihood ratio test was performed against the null hypothesis (that each locus has no effect) using the software GEMMA [63], using a minor allele frequency of 0 to include all SNPs. GEMMA was modified to output the ML log-likelihood under the null and alternative hypothesis and $-\log_{10} P$ values were calculated using R scripts in the bacterialGWAS package.

Testing for lineage effects

We tested for associations between lineage and phenotype using principal components (PCs) in the R package bugwas (available at https://github.com/sgearle/bugwas), which implements a published method for lineage testing in bacterial GWAS [64]. PCs were computed based on biallelic SNPs using the R function prcomp. To test the null hypothesis of no background effect of each principal component, we used a Wald test [65] against a χ^2 distribution with one degree of freedom.

Kmer mapping and sequence alignment

We used Bowtie [66] to align all 31 bp kmers from short-read sequences to the reference genome MRSA252 [56]. For all 31 bp kmers significantly associated with case-controls status, the likely origin of the kmer was additionally determined by nucleotide sequence BLAST [54] of the kmers against a database of all *S. aureus* sequences in GenBank. We used BLAST [54] to identify the best match for coding sequences of interest in the *de novo* assembly, and used Jalview [67] to and visualize the assembled sequences.

Multiple testing correction

Multiple testing was accounted for by applying a Bonferroni correction [68]; the individual locus effect of a variant (PC, kmer or SNP) was considered significant if its p value was smaller than α/n_p, where we took $\alpha = 0.05$ to be the genome-wide false-positive rate and n_p to be the number of PCs, kmer phylopatterns or SNP phylopatterns. We defined each phylopattern to be a unique partition of individuals by the alleles at that kmer or SNP.

The Bonferroni correction represents a conservative approach to controlling for type 1 error. The harmonic mean p-value (HMP) has recently been developed as a method to combine alternative hypotheses against the null hypothesis, without sacrificing power, even when the tests are not independent [69]. The HMP was calculated for coding regions using the R package harmonicmeanp v3.0 (https://CRAN.R-project.org/package/harmonicmeanp). The HMP across a region was then adjusted for the proportion of kmers mapping to that region:

$$HMP_{adj} = HMP/\omega$$

(where ω=proportion of kmers mapping to that region, compared to the total number of kmers mapping to coding regions). The adjusted HMP was compared directly to the significance threshold α_L, being a nominal threshold α (0.05) adjusted for the number of unique p-values being tested [69].

RESULTS

Sequences from 2001 *S. aureus* isolates (1017 cases of bacteraemia and 984 asymptomatic nasal carriage controls) were analysed (Table 1). Cases were marginally more likely to be healthcare-associated (38 vs 33.5%, P=0.04, χ^2 test). Consistent with established risk factors for SAB [2], cases were significantly older (median age 68 years vs 59 years, $P<10^{-5,}$ Mann-Whitney U test) and more likely to be male (68.4 vs 51.9%, $P<10^{-5,}$ χ^2 test). Cases had a higher proportion of MRSA than controls (13.6 vs 5.5%, $P<10^{-5,}$ χ^2 test), including when comparing HA cases (63/386 (15.3%)) with HA controls (35/330 (10.6%), P=0.04 (χ^2 test)). Thus, even in

Table 1. Cases and controls included in study. HA includes both hospital-onset disease and community-onset, healthcare acquired disease

	***S. aureus* bacteraemia (cases)**	***S. aureus* nasal carriage (controls)**	
Number of sequences	1017	984	
Community-associated (CA)	631 (62.0%)	654 (66.5%)	
Healthcare-associated (HA)	386 (38.0%)	330 (33.5%)	P = 0.04 (χ2 test)
Age (median (IQR))	68 (53–79)	59 (38–76)	$P < 10^{-5}$ (Mann-Whitney U test)
Male sex (n (%))	659 (68.4%)	511 (51.9%)	$P < 10^{-5}$ (χ2 test)
MRSA (n (%))	138 (13.6%)	54 (5.5%)	P < 10-5 (χ2 test)

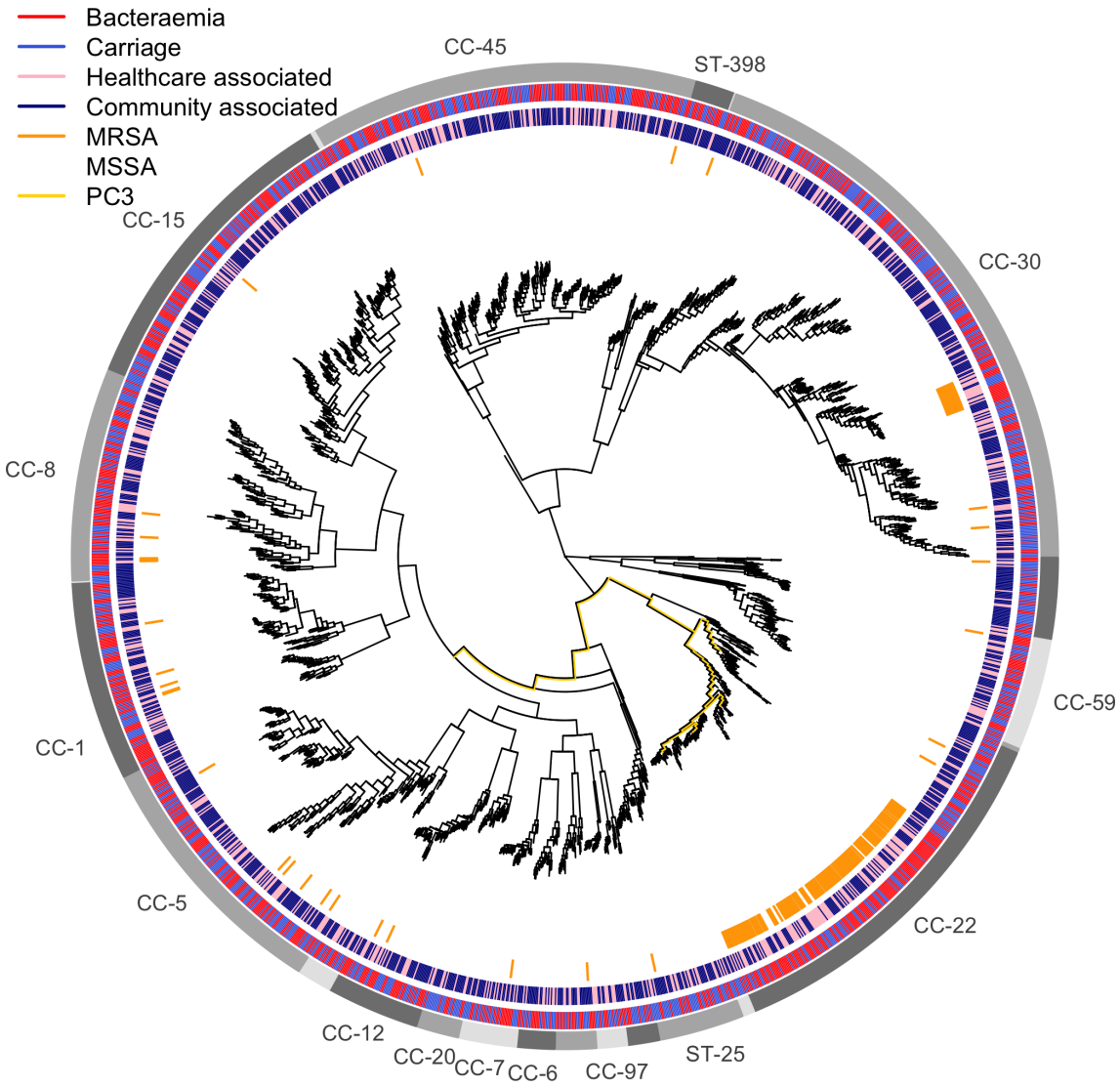

Fig. 1. Maximum likelihood phylogeny of 2001 isolates from bacteraemia and carriage. Branch lengths have been square-root transformed to better discriminate closely related lineages. The outer ring indicates clusters with a shared lineage; lineages with more than 20 isolates are named by the clonal complex (or ST if only a single ST was in the cluster). The second outermost ring indicates isolate source (blue carriage, red bacteraemia). The third outermost ring indicates whether each isolate was community (dark blue) or healthcare (pink) associated. The inner ring indicates isolates were MRSA (orange) or MSSA (white). The branches corresponding with the most significant principal component of variance (PC3, P=0.02, Wald test) with respect to bacteraemia-vs-carriage are highlighted in yellow.

individuals exposed to the healthcare environment, MRSA was found more often in bacteraemia than carriage. Reported focus of infection in cases of SAB showed soft tissue and vascular catheter infections to be the most commonly identified foci (Table S1). Mortality by 30 days was 26.5% (Table S2). These observations are consistent with previously reported UK cohort studies of SAB [6–8].

S. aureus lineages do not differ strongly in their propensity to cause bacteraemia

A phylogeny of 2001 cases and controls demonstrated that a broad diversity of *S. aureus* lineages among our cases (Fig. 1), with representatives from all clonal complexes (CC) as defined in the MLST scheme [70]. Two lineages dominated MRSA isolates – ST-22 (within CC-22) and ST-36 (within CC-30) – consistent with the epidemiology of MRSA in Oxfordshire and throughout the UK [34, 71–73]. HA cases and controls were distributed throughout the tree, and did not strongly cluster within the population.

Formal testing for lineage effects with *bugwas* [64] supported the absence of strong lineage effects. The third principal component (PC3), which identified the MRSA clade within the CC-22 lineage, was most strongly associated with

bacteraemia-vs-carriage (P=0.02, Wald test), but this was not statistically significant after adjusting for multiple testing (Figs 1 and S3). The overall sample heritability was predicted to be low (2.1%, 95% CI 0.0–5.3%). This comprehensive survey of SAB and carriage indicates that lineages of *S. aureus* do not differ substantially in their intrinsic propensity to cause bacteraemia.

Antimicrobial resistance determinants are associated with *S. aureus* bacteraemia

Testing all identified SNPs for association with case/control status in 2001 isolates did not identify any statistically significant associations between individual SNPs and bacteraemia at the genome wide level when controlling for population structure (Fig. S4). However, the SNP coming closest to a statistically significant association (P=$10^{-5.6}$, likelihood-ratio test (LRT)) was an A to T mutation at position 1497290 in the MRSA252 reference genome. This SNP encodes a phenylalanine to tyrosine substitution at codon position 99 in dihydrofolate reductase (*dfrB*); this F99Y mutation confers trimethoprim resistance [48]. It was relatively rare, being found in 41 cases and five controls, and correlated with resistance to other antibiotics: 36/46 (78%) of isolates with this variant were also methicillin resistant. This variant was found most commonly in isolates from CC-22 (63%) and ST-36 (22%).

To further investigate associations between genomic content or sequence variation and the bacteraemia-vs-carriage phenotype, we used a kmer approach [60] to detect variation in the accessory genome, and variants such as small insertions or deletions that are not well captured by mapping SNPs.

When testing kmers found in all 2001 isolates, we found 1214 kmers significantly associated with carriage. These kmers mapped to multiple sites across the genome, most of which were repeat regions, including 16S rDNA and transposon insertion sequences (Fig. S5a). However, the presence of these kmers was strongly affected by the Illumina sequencing read length, being found in isolates sequenced using 150 bp but not 99 bp reads (Fig. S5b). *De novo* assembly was repeated using a constrained kmer length in assembly (up to 79 bp) to control for the variation in read length, and when these new assemblies were used, 924/1214 (76%) of the previously identified 31 bp kmers were no longer significantly associated with the phenotype (Fig. S5c). We concluded that varying length of sequencing was a major source of confounding in kmer-based associated estimates.

To avoid false positive results, we therefore restricted the investigation of kmers associated with bacteraemia-vs-carriage to isolates sequenced with 150 bp reads (Table 2). This reduced set of cases showed similar epidemiological characteristics to the larger group (Table 1).

Kmers tagging antimicrobial resistance (AMR) conferring mutations were significantly associated with bloodstream infection. In total, we identified 22 284 204 kmers, occurring in 930 702 unique patterns across 1610 isolates with 150 bp reads. Twenty-three kmers, occurring in two phylopatterns, were significantly associated with SAB (Fig. 2). These kmers, mapping to a 52 bp region in *dfrB*, exhibited 11.2–11.6-fold increased odds of being found in a disease-causing, rather than carried, *S. aureus*. This association remained significant after controlling for population structure (P=$10^{-8.9}$-$10^{-9.3}$, LRT). When mapped, these kmers centred on MRSA252 position 1497290, where three known single nucleotide variants are capable of conferring trimethoprim resistance, including the F99Y variant identified by our SNP GWAS. Like the trimethoprim resistance conferring SNP, these kmers were found in low frequencies (35/626 (5.6%) cases and 5/984 controls (0.5%)).

No further individual kmers met the threshold for significance, but there were distinctive peaks enriched for small p-values in the Manhattan plot (Figs 2 and S6). We calculated HMP to perform aggregate kmer-based tests of association across coding regions of the genome. At the whole-genome level, the pooled evidence for association between coding sequence variation and bacteraemia was considerable (HMP=$10^{-3.3}$, LRT). The evidence for kmers associated with bacteraemia, when pooled, was significant in several loci (Fig. 2), including the *dfrB* locus, SAR1439 (HMP=$10^{-6.9}$, HMP_{adj}=$10^{-3.3}$, LRT)). The presence of kmers mapping across the SCCmec region was also significantly associated with bacteraemia (HMP=$10^{-4.4}$, HMP_{adj}=$10^{-2.2}$, LRT). The strongest evidence within SCCmec was for *mecA* (SAR0039, HMP=$10^{-5.3}$, HMP_{adj}=0.02, LRT), which encodes PBP2a, a transpeptidase which has low beta-lactam affinity and confers methicillin resistance.

Table 2. Cases and controls sequenced at 150 bp read-length and included in the kmer study. HA includes hospital-onset disease and community-onset, healthcare-acquired disease

	***S. aureus* bacteraemia (cases)**	***S. aureus* nasal carriage (controls)**	
Number of sequences	626	984	
Community-associated (CA)	410 (65.5%)	654 (66.5%)	
Healthcare-associated (HA)	216 (34.5%)	330 (33.5%)	P = 0.7 (χ2 test)
Age (median (IQR))	68 (53-79)	59 (38-76)	$P < 10^{-5}$ (Mann-Whitney U test)
Male sex (n (%))	399 (63.7%)	511 (51.9%)	$P < 10^{-5}$ (χ2 test)
MRSA (n (%))	78 (12.4%)	54 (5.5%)	$P < 10^{-5}$ (χ2 test)

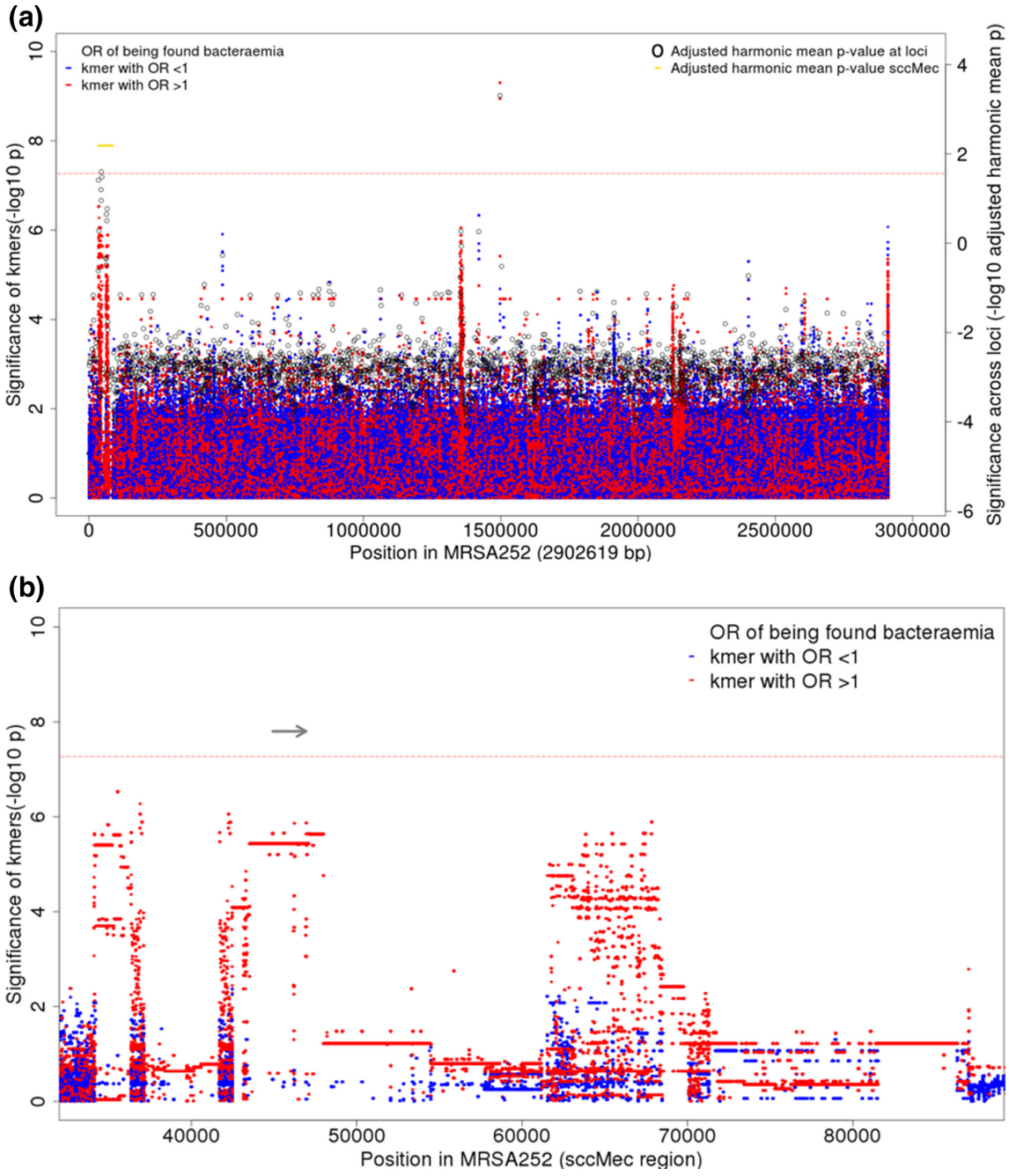

Fig. 2. Association of kmers with *S. aureus* bacteraemia-vs-carriage, controlling for population structure and HA or CA origin. (a) Manhattan plot showing significance of association (-log10 *P*-value, left axis) for individual kmers (red, kmers with OR>1 of being found in bacteraemia-vs-carriage; blue, kmers with OR<1 of being found in bacteraemia-vs-carriage). Unmapped kmers are plotted at the end of the genome. Pooled evidence (adjusted harmonic mean *P*-value, right axis) across each CDS is shown by black open circles. Pooled evidence across SCCmec is shown in gold. A threshold of significance is plotted in a red horizontal line. For kmers this a Bonferroni-corrected threshold of significance, adjusted for the number of individual kmer patterns ($10^{-7.2}$, left axis). For evidence across loci, the same family wide error rate of 0.05 ($10^{-1.6}$, right axis) was applied by adjusting the p-values for the numbers of variants tested. (b) Kmers mapping to the region of staphylococcal chromosome cassette sccMec are shown in greater detail. The coding sequence of *mecA* (SAR0039) is marked by a horizontal grey arrow.

When this region was examined in closer detail, high-risk kmers covered the entirety of the SAR0039 locus, encoding PBP2a, a PBP with low affinity for beta-lactams (Fig. 2b). There are no alternate low-risk kmers in this region, suggesting the presence of this gene, rather than variation within it, is associated with bacteraemia. Thus, genomic sequences associated with AMR to both trimethoprim and beta-lactams, but not other antimicrobial classes, were significantly associated with bacteraemia.

Genomic signals of healthcare-associated carriage include antimicrobial resistance factors, as well as variation in a virulence determinant, *prsA*

Regulatory gene changes have been associated with persistent bacteraemia and increased mortality, and it has been hypothesised these changes either convey or accompany relative survival advantages in healthcare environments [31, 32]. We compared the genomic factors associated with healthcare environments by conducting a GWAS for HA-vs-CA among carriers (330 HA, 654 CA). We focused only on carriage isolates because we had more extensive data about hospital admissions in this group, and epidemiological data to further demonstrate that isolates from CA-carriage truly reflected community origin. The MRSA CC-22 lineage showed the strongest association with HA-vs-CA carriage (P=0.06, Wald test, Fig. S7), but this lineage effect was even less significant than for bacteraemia-vs-carriage (P=0.02, Wald test, Fig. S3), suggesting that no lineages were strongly associated with healthcare acquisition of *S. aureus* carriage.

In total 124 SNPs were associated with HA-vs-CA carriage after controlling for population structure, and adjusting for the 77 597 SNP phylopatterns in the population (Fig. S8, Table S3). The most significant were non-coding (intergenic) variants in a region encoding tRNAgly at 2034022–2034039. However, basecalls at these sites were only made in 380/984 (38. 6%) with calls imputed in the remaining 604/984 (61.4%). Excluding isolates with imputed calls at these sites reduced the unadjusted OR for finding these SNPs in HA-vs-CA carriage from 2.1 to 0.87, suggesting that the observed association was a product of SNP imputation. The most significant coding variants included a G to A substitution at 2 417 648 in MRSA242 (P=$10^{-7.2}$, LRT), encoding a proline to leucine substitution at 707 in SAR2345, an AcrB/AcrD/AcrF family protein, which is a multidrug efflux system subunit. They also included a C to T substitution at 7255 in MRSA252 (P=$10^{-7.2}$), which encodes L84S in *gyrA* and confers quinolone resistance [48]. A variant at these positions was called in all 984 isolates, being found in 43/330 (13.0%) HA carriage isolates and 23/264 (3.5%) CA carriage isolates, with no calls imputed. A significant association was also seen with a band of 91 low frequency SNPs, with shared minor alleles co-inherited predominantly in the MRSA sub-clade of CC-22 (Fig. S9). The *dfrB* mutation seen associated with bacteraemia was not significantly associated with HA-vs-CA carriage (P=0.4).

All carriage isolates were sequenced with 150 bp reads, so all were included in a study of kmer associations. We found 188 31 bp kmers in 59 unique patterns were significantly associated with HA carriage (Fig. 3a, Table S4); 123/188 (65.4%) kmers comprising 52 phylopatterns mapped to a single gene – *prsA* – and their absence was associated with HA carriage (P=$10^{-7.2}$-$10^{-19.4}$, LRT) (Fig. 3b). A small peak of 11 significant kmers mapped to a hypothetical protein SAR0061, covering to a 41 bp region from 67 816 (P=$10^{-7.3}$, LRT). Three significant kmers mapped to a putative transcriptional regulator SAR2394, covering a 33 bp region at 2 465 937 (P=$10^{-7.2}$, LRT). Of the remaining significant kmers, 26/188 (13.8%) did not map to the reference genome, and 25/188 (13.3%) mapped to non-coding regions at 1.3 MB, 2.7 MB and 2.8 MB.

While CA carriage isolates showed conservation of the *prsA* sequence across the population, multiple patterns of variation arising in different lineages were seen in HA carriage isolates (Fig. S10), consisting of both SNPs and deletions (Fig. S11). When a conditional SNP GWAS was performed including the presence or absence of any of the kmers most strongly associated with HA-vs-CA carriage as a covariate in the model, no SNPs reached the threshold for significance, suggesting the signal of association accompanying these SNPs was better explained by *prsA* variation. Variation in *prsA* was common in the MRSA sub-clade of CC-22 but was not limited to this lineage. PrsA is a surface bound foldase protein, responsible for post-translational processing of virulence factors, including proteases and cell surface proteins [74, 75] which has also been shown to modulate susceptibility for both beta-lactam and glycopeptide antibiotics [76]. No peak of significant kmers occurred at the *prsA* locus in the bacteraemia-vs-carriage GWAS (Fig. 2a), suggesting that variation at this locus is specifically associated with carriage arising from the healthcare environment, but not with bacteraemia.

DISCUSSION

These findings reflect analysis of a large collection, representing the populations of *S. aureus* circulating in community and healthcare settings, including the whole genomic content of the population, with control of population structure. In doing so, we address the conflicting results from smaller case-control studies which have found evidence for [26, 29] and against [18, 24] differing invasiveness between *S. aureus* lineages. We conclude that *S. aureus* lineages do not differ in the frequency with which they caused bacteraemia, compared to their frequency in carriage. However, across lineages we found evidence that genetic variants underlying AMR were associated with increased odds of bacteraemia versus carriage. These included some determinants of methicillin resistance, contrasting with previous research indicating that methicillin resistance incurs a fitness cost for *S. aureus*, reducing its pathogenicity [77, 78].

The most obvious possible explanation for association between methicillin resistance and bacteraemia would be survival benefit in the presence of beta-lactam antibiotics. However more than half of our MRSA bacteraemia cases were community-associated, being first detected prior to or

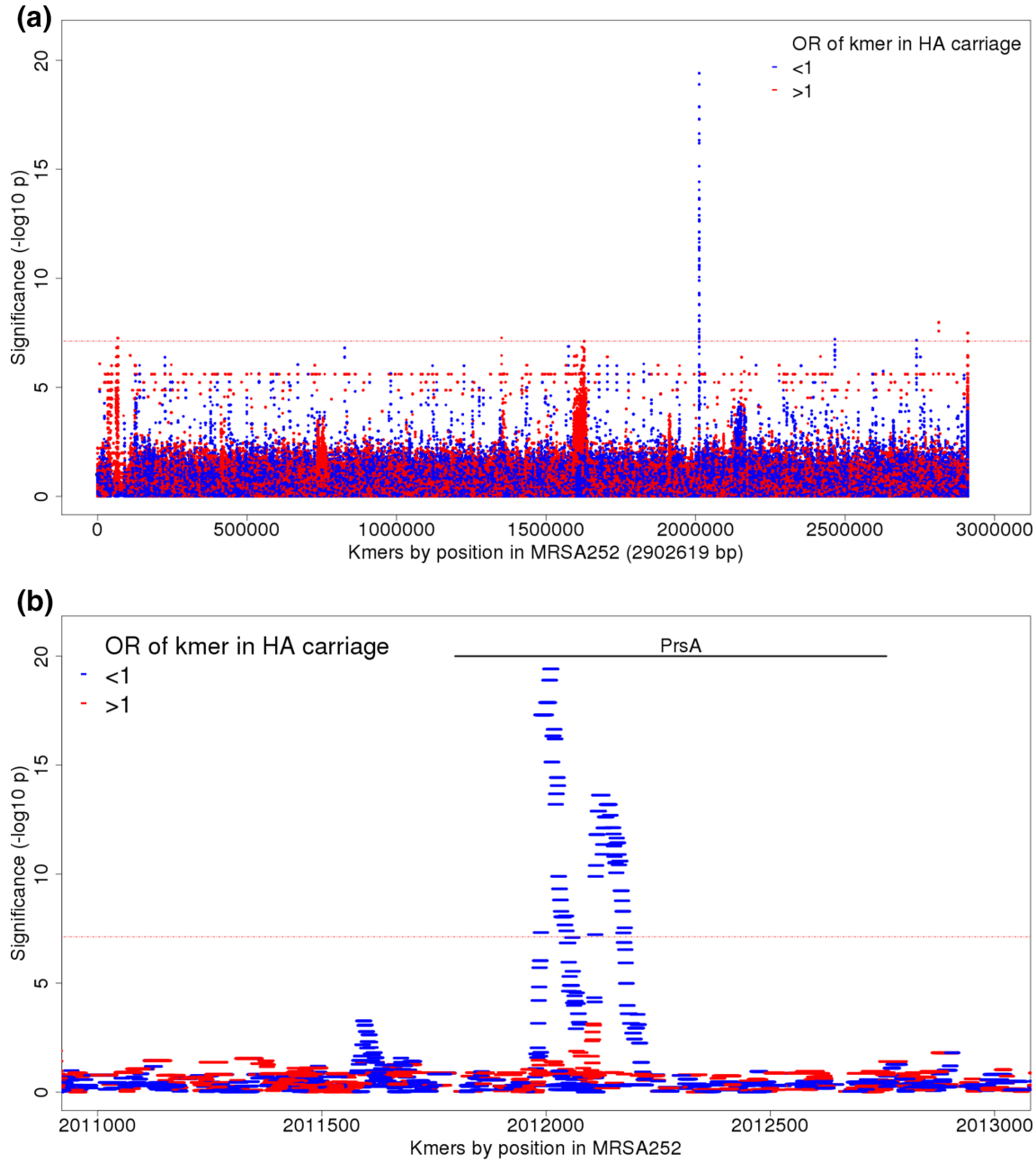

Fig. 3. Association of kmers with HA-vs-CA carriage, controlling for population structure. (a) Manhattan plot showing significance of association (-log10 *P*-value, left axis) for individual kmers (red, kmers found more often in HA carriage; blue, kmers found more often in CA carriage). The Bonferroni significance threshold, adjusting for the number of kmer phylopatterns, was $10^{-7.1}$ (red horizontal line). (b) The region of *prsA* is shown in greater detail. Kmers showing significant association with HA-vs-CA carriage cover the region 2011973–2012202 in the reference, which corresponds to bases 177–394 in *prsA*.

within 48 h of hospital admission. While we do not have data about pre-hospital antibiotic treatment, it is unlikely that the majority of patients with CA bacteraemia were on beta-lactam antibiotics at the time that bacteraemia developed. It is also possible that other between-group differences, such co-morbid illnesses, may also account for the different prevalence of MRSA observed between patients with bacteraemia and asymptomatic carriers, and we have not been able to measure patient factors confounding between MRSA carriage and invasive infection.

However, there is *in vitro* evidence that *mecA* – the primary determinant of methicillin resistance – modifies bacterial virulence independently of antibiotic selection pressure, through modulation of *agr*-mediated toxin expression [33]. Previous evidence has implicated the altered PBP encoded

by *mecA* in persistent and complicated bacteraemia, and our observations concords with those findings. Analysis of SAB over 21 years in the USA showed that within CC-8, the MRSA lineage USA300 was associated with higher rates of metastatic infection compared to the rest of the *spa*-t008 complex after adjusting for patient and clinical variables [79]. *S. aureus* strains which are phenotypically methicillin susceptible but contain the *mecA* gene have been reported to have a higher risk of persistent bacteraemia, even when treated with vancomycin [80]. AMR elements have been implicated as virulence determinants in another Gram-positive species; *Streptococcus pneumoniae*, where a PBP variant associated with penicillin tolerance (though not resistance) was associated with meningitis among isolates found in invasive pneumococcal disease [81].

Our finding of trimethoprim resistance associated with bacteraemia is not easily explained through direct antibiotic selection since trimethoprim containing antibiotics (e.g. co-trimoxazole) are not advised in therapeutic guidelines for treatment of skin and soft tissue infections in the United Kingdom, where CA-MRSA rates are low. It is plausible that patients may receive trimethoprim for urinary tract infections in the community, and this could reflect an impact on commensal flora, enabling invasion by trimethoprim-resistant *S. aureus* strains. It is also possible that mutations in *dfrB*, a metabolic gene important in bacterial DNA synthesis, affect bacterial persistence: trimethoprim resistant *S. aureus* have shown variable growth rate and survival under environmental stress according to the mechanism of resistance [82].

Overall, we found MRSA was more prevalent in bacteraemia than carriage when sampling the population representatively. In the UK at this time MRSA was almost exclusively healthcare associated. Consequently, one potential explanation for this finding is unmeasured healthcare exposure among CA bacteraemia cases for which healthcare exposure data was only available for the preceding 12 weeks. However even restricting to HA associated cases and controls MRSA was associated with bacteraemia. Furthermore, by controlling for population structure, we can be confident the association demonstrated between *mecA* and bacteraemia is not simply a reflection of healthcare adapted lineages, and that such lineage effects could have been detected using the GWAS methods employed here. In fact, while variation within *prsA* was strongly associated with HA carriage, variation in this gene was not associated with bacteraemia-vs-carriage. Likewise, a quinolone resistance mutation in *gyrA* was associated with HA carriage but not with bacteraemia. The association with quinolone resistance is consistent with previous reports of quinolone resistance associated with HA *S. aureus* infections in the absence of quinolone treatment [83]. However these contrasting observations suggest that distinct bacterial factors favour healthcare environment adaptation or transmission compared with those favouring bloodstream infection.

The surface bound foldase protein PrsA has been implicated as a secondary resistance factor for both beta-lactam and glycopeptide antibiotics [84]. *In vitro* assays have shown that *S. aureus* can survive despite disruption to *prsA*, but strains with *prsA* disruption are more susceptible to oxacillin [84]. PrsA reduces the membrane quantity of PBP2a without altered transcription of *mecA,* regulating expression of a methicillin-resistant phenotype independently of *mecA*. Additionally, PrsA plays an important role in the post-translational processing of virulence factors, including proteases and cell surface proteins [74, 75], and while the secretion of some virulence factors is decreased when *prsA* is deleted [74], PrsA-deficient bacteria have enhanced aggregation and adherence [75], changes which might favour survival or transmission in the healthcare environment.

Our study demonstrates some constraints and pitfalls for bacterial GWAS. Firstly, sequencing read length was a strong source of confounding in our data which, without adequate control, produced false positive results. This is an ongoing challenge for studies pooling existing sequencing data where it is not possible to randomize cases and controls across sequencing batches. We dealt with this using the conservative option of excluding 391 cases from kmer analysis, representing a substantial sacrifice of power. A further limitation was that population structure appeared to be incompletely controlled for low frequency variants [85], as exhibited by a set of 91 SNPs in strong linkage disequilibrium (LD) associated with HA-vs-CA carriage. Linear mixed models are able to control for lineages, and cryptic population structure through the use of a relatedness matrix, but they make the assumption that closely related isolates are unlikely to have large differences in phenotype. This assumption means that they can incompletely account for lineage effects arising from closely related strains which vary significantly in frequency between phenotypes [85]. In our study, the SNPs in LD were found in the predominantly HA-MRSA isolates within CC-22. This may represent either a true association with that lineage, or co-carriage with another genomic element in that lineage. In this case, *prsA* variation was common in CC-22, but not detected in the SNP based study (as the variation was a deletion rather than a nucleotide substitution), and the signal of association accompanying these SNPs was better explained by *prsA* variation. Overall, the kmer-based methods were more fruitful in our study, both in their ability to detect non-SNP based variation (such as deletions in *prsA*, and the presence of *mecA*), and retaining the ability to identify significant SNPs (including a *dfrB* variant).

In previous studies we have identified within-host adaptation of *S. aureus* genes associated with development of invasive disease from a colonising isolate – particularly involving the *agr* locus, and the cell wall proteins under regulatory control of *agr* and *rsp* [42]. Such variation was not associated with bacteraemia in this population-based study, perhaps because it provides only a short-term advantage to the bacteria, and in the longer term is detrimental, by adversely affecting transmission. In contrast, variation which confers AMR is likely to confer a bacterial survival advantage in carriage in the face of antibiotic selection pressure in the healthcare environment, as well as in the bloodstream, allowing these variants to survive in the population.

CONCLUSIONS

In a study of over 2000 isolates from *S. aureus* bacteraemia and nasal carriage in the UK, we found strong evidence that all *S. aureus* lineages are equally capable of causing bloodstream infection, and of being carried in the healthcare environment.

We found that genomic variation in *prsA* (encoding a foldase protein) was a novel genomic marker of healthcare adaptation in *S. aureus*. This predictor of healthcare-associated carriage was not associated with bacteraemia, while AMR determinants were associated with both bacteraemia and healthcare-associated carriage, raising the suggestion that in addition to enabling survival in healthcare environments, AMR functions as a virulence factor, promoting invasive disease.

Given studies demonstrating a direct effect of *mecA* on toxin expression [33] and reduced toxicity enhancing bloodstream survival [35], we hypothesise that lowered expression of toxic virulence factors seen in MRSA may be one method by which *S. aureus* gains a short-term survival advantage and causes bloodstream infection.

Funding information
This study was supported by the Oxford NIHR Biomedical Research Centre, a Mérieux Research Grant, the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at Oxford University in partnership with Public Health England (PHE) (grant HPRU-2012–10041), and the Health Innovation Challenge Fund (a parallel funding partnership between the Wellcome Trust (grant WT098615/Z/12/Z) and the Department of Health (grant HICF-T5-358)). Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. Financial support was provided by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z. D.J.W. is a Sir Henry Dale Fellow, jointly funded by the Wellcome Trust and the Royal Society (Grant 101237/Z/13/B). D.J.W. is supported by a Big Data Institute Robertson Fellowship. B.C.Y. is an NIHR Academic Clinical Lecturer, and this work was funded by a Research Training fellowship from Wellcome Trust (Grant 101611/Z/13/Z). T.E.P., D.W.C. and A.S.W. are NIHR Senior Investigators. The views expressed are those of the author(s) and not necessarily those of the NHS, PHE, the NIHR or the Department of Health.

Acknowledgements
We thank the International *Staphylococcus aureus* Consortium and the United Kingdom Clinical Infection Research Group for sharing data and bacterial strains from bloodstream infection.

Author contributions
Study conceptualisation B.C.Y., C.H.W., J.C., S.E., D.C.W., T.E.A.P., A.S.W., M.J.L., D.J.W. Sample collection and epidemiological investigation by B.C.Y., J.R.P., J.C.G., S.O., H.G., R.F., A.V., R.M., K.K., T.P.Q., R.T., M.S., A.S.W., M.J.L., genomic sequencing performed by K.C., L.D., E.L., B.C.Y. Formal analysis of data by B.C.Y., C.H.W., J.C., S.E., A.S.W., M.J.L., D.J.W. Discussion of results by B.C.Y., C.H.W., D.W.C., T.E.A.P., A.S.W., M.J.L., D.J.W. Writing – original manuscript preparation B.C.Y., and D.J.W. Writing – review and editing B.C.Y., C.H.W., J.C., J.R.P., R.T., M.S., A.S.W., M.J.L., D.J.W.

Conflicts of interest
The authors declare that there are no conflicts of interest.

Ethical statement
The study of community *S. aureus* carriage was approved by Oxfordshire Research Ethics Committee B (08/H0605/102). Ethical approval for linkage to patient data without individual patient consent in Oxford was obtained from the South Central Ethics Committee (14/SC/1069) and the Confidentiality Group [ECC5-017(A)/2009]. Data about *S. aureus* bacteraemia in Oxfordshire, Brighton and Plymouth were collected for evaluations of clinical service provision. Under UK National Research Ethics Service guidance at the time this data collection constituted a service evaluation involving routinely available, non-identifying patient data and therefore not requiring research ethics committee review.

References

1. Lowy FD. *Staphylococcus aureus* infections N. *N Engl J Med* 1998;339:520–532.
2. Tong SYC, Davis JS, Eichenberger E, Holland TL, Fowler VG Jr. *Staphylococcus aureus* infections: Epidemiology, pathophysiology, clinical manifestations, and management. *Clin Microbiol Rev* 2015;28:603–661.
3. Anderson DJ, Moehring RW, Sloane R, Schmader KE, Weber DJ, *et al*. Bloodstream infections in community hospitals in the 21st century: A multicenter cohort study. *PLoS One* 2014;9:e91713.
4. Shorr AF, Tabak YP, Killian AD, Gupta V, Liu LZ, *et al*. Healthcare-associated bloodstream infection: A distinct entity? insights from a large U.S. *Critical Care Medicine* 2006;34:2588–2595.
5. Public Health England. Annual epidemiological commentary: Gram-negative bacteraemia, MRSA bacteraemia, MSSA bacteraemia and C. difficile infections, up to and including financial year April 2019 to March 2020. PHE. 2020. https://www.gov.uk/government/statistics/mrsa-mssa-and-e-coli-bacteraemia-and-c-difficile-infection-annual-epidemiological-commentary
6. Thwaites GE. United Kingdom Clinical Infection Research Group (UKCIRG). The management of *Staphylococcus aureus* bacteremia in the United Kingdom and Vietnam: A multi-centre evaluation. *PLoS One* 2010;5:e14170.
7. Kaasch AJ, Barlow G, Edgeworth JD, Fowler VGJ, Hellmich M, *et al*. *Staphylococcus aureus* bloodstream infection: A pooled analysis of five prospective, observational studies. *J Infect* 2014;68:242–251.
8. Nambiar K, Seifert H, Rieg S, Kern WV, Scarborough M, *et al*. Survival following *Staphylococcus aureus* bloodstream infection: A prospective multinational cohort study assessing the impact of place of care. *J Infect* 2018;pii: S0163-4453(18)30260-3.
9. Alonzo F, Torres VJ. The bicomponent pore-forming leucocidins of *Staphylococcus aureus*. *Microbiol Mol Biol Rev* 2014;78:199–230.
10. Otto M. *Staphylococcus aureus* toxins. *Curr Opin Microbiol* 2014;17:32–37.
11. Foster TJ. Immune evasion by staphylococci. *Nat Rev Microbiol* 2005;3:948–958.
12. Foster TJ, Geoghegan JA, Ganesh VK, Höök M. Adhesion, invasion and evasion: The many functions of the surface proteins of *Staphylococcus aureus*. *Nat Rev Microbiol* 2013;12:49–62.
13. Spaulding AR, Salgado-Pabón W, Kohler PL, Horswill AR, Leung DY, *et al*. Staphylococcal and streptococcal superantigen exotoxins. *Clin Microbiol Rev* 2013;26:422–447.
14. Schlievert PM, Shands KN, Dan BB, Schmid GP, Nishimura RD. Identification and characterization of an exotoxin from *Staphylococcus aureus* associated with toxic-shock syndrome. *J Infect Dis* 1981;143:509–516.
15. Lina G, Gillet Y, Vandenesch F, Jones ME, Floret D, *et al*. Toxin involvement in staphylococcal scalded skin syndrome. *Clin Infect Dis* 1997;25:1369–1373.
16. Shallcross LJ, Fragaszy E, Johnson AM, Hayward AC. The role of the Panton-Valentine leucocidin toxin in staphylococcal disease: A systematic review and meta-analysis. *Lancet Infect Dis* 2013;13:43–54.
17. Young BC, Earle SG, Soeng S, Sar P, Kumar V, *et al*. Panton-Valentine leucocidin is the key determinant of *Staphylococcus aureus* pyomyositis in a bacterial GWAS. *Elife* 2019;8:e42486.
18. Peacock SJ, Moore CE, Justice A, Kantzanou M, Story L, *et al*. Virulent combinations of adhesin and toxin genes in natural populations of *Staphylococcus aureus*. *Infect Immun* 2002;70:4987–4996.

19. Calderwood MS, Desjardins CA, Sakoulas G, Nicol R, Dubois A, *et al*. Staphylococcal enterotoxin P predicts bacteremia in hospitalized patients colonized with methicillin-resistant *Staphylococcus aureus*. *J Infect Dis* 2014;209:571–577.
20. Bhatty M, Ray P, Singh R, Jain S, Sharma M. Presence of virulence determinants amongst *Staphylococcus aureus* isolates from nasal colonization, superficial & invasive infections. *Indian J Med Res* 2013;138:143–146.
21. Van Belkum A, Melles DC, Snijders SV, van Leeuwen WB, Wertheim HF, *et al*. Clonal distribution and differential occurrence of the enterotoxin gene cluster, egc, in carriage- versus bacteremia-associated isolates of *Staphylococcus aureus*. *J Clin Microbiol* 2006;44:1555–1557.
22. Rasmussen G, Monecke S, Ehricht R, Söderquist B. Prevalence of clonal complexes and virulence genes among commensal and invasive *Staphylococcus aureus* isolates in Sweden. *PLoS One* 2013;8:e77477.
23. Yu F, Yang L, Pan J, Chen C, Du J, *et al*. Prevalence of virulence genes among invasive and colonising *Staphylococcus aureus* isolates. *J Hosp Infect* 2011;77:89–91.
24. Lindsay JA, Moore CE, Day NP, Peacock SJ, Witney AA, *et al*. Microarrays reveal that each of the ten dominant lineages of *Staphylococcus aureus* has a unique combination of surface-associated and regulatory genes. *J Bacteriol* 2006;188:669–676.
25. Melles DC, Gorkink RF, Boelens HA, Snijders SV, Peeters JK, *et al*. Natural population dynamics and expansion of pathogenic clones of *Staphylococcus aureus*. *J Clin Invest* 2004;114:1732–1734.
26. Wertheim HF, van Leeuwen WB, Snijders S, Vos MC, Voss A, *et al*. Associations between *Staphylococcus aureus* Genotype, Infection, and In-hospital mortality: A nested case-control study. *J Infect Dis* 2005;192:1196–2000.
27. Feil EJ, Cooper JE, Grundmann H, Robinson DA, Enright MC, *et al*. How clonal is *Staphylococcus aureus*? *J Bacteriol* 2003;185:3307–3316.
28. Fowler VG Jr, Nelson CL, McIntyre LM, Kreiswirth BN, Monk A, *et al*. Potential associations between hematogenous complications and bacterial genotype in *staphylococcus aureus* infection. *J Infect Dis* 2007;196:738–747.
29. Messina JA, Thaden JT, Sharma-Kuinkel BK, Fowler VG. Impact of bacterial and human genetic variation on *staphylococcus aureus* infections. *PLoS Pathog* 2016;12:e1005330.
30. Lenz R, Leal JR, Church DL, Gregson DB, Ross T, *et al*. The distinct category of healthcare associated bloodstream infections. *BMC Infect Dis* 2012;12:85.
31. Painter KL, Krishna A, Wigneshweraraj S, Edwards AM. What role does the quorum-sensing accessory gene regulator system play during *Staphylococcus aureus* bacteremia? *Trends Microbiol* 2014;22:676–685.
32. Rose HR, Holzman RS, Altman DR, Smyth DS, Wasserman GA, *et al*. Cytotoxic virulence predicts mortality in nosocomial pneumonia due to methicillin-resistant *Staphylococcus aureus*. *J Infect Dis* 2015;211:1862–1874.
33. Rudkin JK, Edwards AM, Bowden MG, Brown EL, Pozzi C, *et al*. Methicillin resistance reduces the virulence of healthcare-associated methicillin-resistant *Staphylococcus aureus* by interfering with the agr quorum sensing system. *J Infect Dis* 2012;205:798–806.
34. Coll F, Harrison EM, Toleman MS, Reuter S, Raven KE, *et al*. Longitudinal genomic surveillance of MRSA in the UK reveals transmission patterns in hospitals and the community. *Sci Transl Med* 2017;9:eaak9745.
35. Laabei M, Uhlemann AC, Lowy FD, Austin ED, Yokoyama M, *et al*. Evolutionary trade-offs underlie the multi-faceted virulence of *Staphylococcus aureus*. *PLoS Biol* 2015;13:e1002229.
36. Traber KE, Lee E, Benson S, Corrigan R, Cantera M, *et al*. Agr function in clinical *Staphylococcus aureus* isolates. *Microbiology (Reading)* 2008;154:2265–2274.
37. Fowler VG, Sakoulas G, Mcintyre LM, Meka VG, Arbeit RD, *et al*. Persistent bacteremia due to methicillin- resistant *Staphylococcus aureus* infection is associated with agr dysfunction and low- level in vitro resistance to thrombin- induced platelet microbicidal protein. *J Infect Dis* 2004;190:1140.
38. Schweizer ML, Furuno JP, Sakoulas G, Johnson JK, Harris AD, *et al*. Increased mortality with accessory gene regulator (agr) dysfunction in *Staphylococcus aureus* among bacteremic patients. *Antimicrob Agents Chemother* 2011;55:1082–1087.
39. Young BC, Golubchik T, Batty EM, Fung R, Larner-Svensson H, *et al*. Evolutionary dynamics of *staphylococcus aureus* during progression from carriage to disease. *Proc Natl Acad Sci U S A* 2012;109:4550.
40. Das S, Lindemann C, Young BC, Muller J, Österreich B, *et al*. Natural mutations in a *Staphylococcus aureus* virulence regulator attenuate cytotoxicity but permit bacteremia and abscess formation. *Proc Natl Acad Sci U S A* 2016;113:E3101-10.
41. Krishna A, Holden MTG, Peacock SJ, Edwards AM, Wigneshweraraj S. Naturally occurring polymorphisms in the virulence regulator Rsp modulate *Staphylococcus aureus* survival in blood and antibiotic susceptibility. *Microbiology (Reading)* 2018;164:1189–1195.
42. Young BC, Wu CH, Gordon NC, Cole K, Price JR, *et al*. Severe infections emerge from commensal bacteria by adaptive evolution. *elife* 2017;6:pii: e30637.
43. Giulieri SG, Baines SL, Guerillot R, Seemann T, Gonçalves da Silva A, *et al*. Genomic exploration of sequential clinical isolates reveals a distinctive molecular signature of persistent *Staphylococcus aureus* bacteraemia. *Genome Med* 2018;10:65.
44. Smeltzer MS. *Staphylococcus aureus* pathogenesis: The importance of reduced cytotoxicity. *Trends Microbiol* 2016;24:681–682.
45. Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, *et al*. Predicting the virulence of MRSA from its genome sequence. *Genome Res* 2014;24:839–849.
46. Roe C, Stegger M, Lilje B, Johannesen TB, Ng KL, *et al*. Genomic analyses of *Staphylococcus aureus* clonal complex 45 isolates does not distinguish nasal carriage from bacteraemia. *Microb Genom* 2020;6:mgen000403.
47. Recker M, Laabei M, Toleman MS, Reuter S, Saunderson RB, *et al*. Clonal differences in *Staphylococcus aureus* bacteraemia-associated mortality. *Nat Microbiol* 2017;2:1381–1388.
48. Gordon NC, Price JR, Cole K, Everitt R, Morgan M, *et al*. Prediction of *Staphylococcus aureus* antimicrobial resistance by whole-genome sequencing. *J Clin Microbiol* 2014;52:1182–1191.
49. Miller RR, Walker AS, Godwin H, Fung R, Votintseva A, *et al*. Dynamics of acquisition and loss of carriage of *Staphylococcus aureus* strains in the community: the effect of clonal complex. *J Infect* 2014;68:426–439.
50. Votintseva AA, Fung R, Miller RR, Knox K, Godwin H, *et al*. Prevalence of *Staphylococcus aureus* protein A (spa) mutants in the community and hospitals in Oxfordshire. *BMC Microbiol* 2014;14:63.
51. Finney JM, Walker AS, Peto TE, Wyllie DH. An efficient record linkage scheme using graphical analysis for identifier error detection. *BMC Med Inform Decis Mak* 2011;11:7.
52. Cardoso T, Almeida M, Friedman ND, Aragão I, Costa-Pereira A, *et al*. Classification of healthcare-associated infection: a systematic review 10 years after the first proposal. *BMC Med* 2014;12:40.
53. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Res* 2008;18:821–829.
54. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.
55. Lunter G, Goodson M. Stampy: A statistical algorithm for sensitive and fast mapping of illumina sequence reads. *Genome Res* 2011;21:936–939.
56. Holden MT, Feil EJ, Lindsay JA, Peacock SJ, Day NP, *et al*. Complete genomes of two clinical *Staphylococcus aureus* strains: evidence for the rapid evolution of virulence and drug resistance. *Proc Natl Acad Sci U S A* 2004;101:9786–9791.
57. Didelot X, Eyre DW, Cule M, Ip CL, Ansari MA, *et al*. Microevolutionary analysis of *Clostridium difficile* genomes to investigate transmission. *Genome Biol* 2013;13:R118.

58. Golubchik T, Batty EM, Miller RR, Farr H, Young BC, *et al*. Within-host evolution of *Staphylococcus aureus* during asymptomatic carriage. *PloS One* 2013;8:e61319.
59. Didelot X, Wilson DJ. ClonalFrameML: Efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol* 2015;11:e1004041.
60. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30:1312–1313.
61. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, *et al*. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. *Proc Natl Acad Sci U S A* 2013;110:11923–11927.
62. Rizk G, Lavenier D, Chikhi R. DSK: k-mer counting with very low memory usage. *Bioinformatics* 2013;29:652–653.
63. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* 2012;44:821–824.
64. Earle SG, CH W, Charlesworth J, Stoesser N, Gordon NC, *et al*. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat Microbiol* 2016;1:16041.
65. Wald A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans Am Math Soc* 1943;54:426–482.
66. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods* 2012;9:357–359.
67. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 2009;25:1189–1191.
68. Dunn OJ. Estimation of the medians for dependent variables. *Ann Math Statist* 1959;30:192–197.
69. Wilson DJ. The harmonic mean *P*-value for combining dependent tests. *Proc Natl Acad Sci U S A* 2019;116:1195–1200.
70. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res* 2018;3:124.
71. Miller RM, Price JR, Batty EM, Didelot X, Wyllie D, *et al*. Healthcare-associated outbreak of meticillin-resistant *Staphylococcus aureus* bacteraemia: role of a cryptic variant of an epidemic clone. *J Hosp Infect* 2014;86:83–89.
72. Wyllie DH, Walker AS, Miller R, Moore C, Williamson SR, *et al*. Decline of meticillin-resistant *Staphylococcus aureus* in Oxfordshire hospitals is strain-specific and preceded infection-control intensification. *BMJ Open* 2011;1:e000160.
73. Knight GM, Budd EL, Whitney L, Thornley A, Al-Ghusein H, *et al*. Shift in dominant hospital-associated methicillin-resistant *Staphylococcus aureus* (HA-MRSA) clones over time. *J Antimicrob Chemother* 2012;67:2514–2522.
74. Wiemels RE, Cech SM, Meyer NM, Burke CA, Weiss A, *et al*. An intracellular peptidyl-prolyl cis/trans isomerase is required for folding and activity of the *Staphylococcus aureus* secreted virulence factor nuclease. *J Bacteriol* 2017;199:e00453-16.
75. Lin MH, Li CC, Shu JC, Chu HW, Liu CC, *et al*. Exoproteome profiling reveals the involvement of the foldase PrsA in the cell surface properties and pathogenesis of *Staphylococcus aureus*. *Proteomics* 2018;18:e1700195.
76. Jousselin A, Manzano C, Biette A, Reed P, Pinho MG. The Staphylococcus aureus cChaperone PrsA is a new auxiliary factor of oxacillin resistance affecting penicillin-binding prols a New Auxiliary Factor of Oxacillin Resistance Affecting Penicillin-Binding Protein 2A. *Antimicrob Agents Chemother* 2015;60:1656–1666.
77. Beceiro A, Tomás M, Bou G. Antimicrobial resistance and virulence: a successful or deleterious association in the bacterial world. *Clin Microbiol Rev* 2013;26:185–230.
78. Watkins RR, David MZ, Salata RA. Current concepts on the virulence mechanisms of meticillin-resistant *Staphylococcus aureus*. *J Med Microbiol* 2012;61:1179–1193.
79. Souli M, Ruffin F, Choi SH, Park LP, Gao S, *et al*. Changing characteristics of *Staphylococcus aureus* bacteremia: Results from a 21-year, prospective, longitudinal study. *Clin Infect Dis* 2019;69:1868–1877.
80. Jones D, Elshaboury RH, Munson E, Dilworth TJ. A Retrospective analysis of treatment and clinical outcomes among patients with methicillin-susceptible Sstaphylococcus aureus bloodstream isolates possessing detectable mecA by a commercial PCR assay compared to patients with methicillin-resistant Sstaphylococcus aureus bloodstream isolates. *Antimicrob Agents Chemother* 2017;62:e01396-17.
81. Li Y, Metcalf BJ, Chochua S, Li Z, Walker H, *et al*. Genome-wide association analyses of invasive pneumococcal isolates identify a missense bacterial mutation associated with meningitis. *Nat Commun* 2019;10:178.
82. Coelho C, de Lencastre H, Aires-de-Sousa M. Frequent occurrence of trimethoprim-sulfamethoxazole hetero-resistant *Staphylococcus aureus* isolates in different African countries. *Eur J Clin Microbiol Infect Dis* 2017;36:1243–1252.
83. Witte W, Grimm H. Occurrence of quinolone resistance in *Staphylococcus aureus* from nosocomial infection. *Epidemiol Infect* 1992;109:413–421.
84. Jousselin A, Renzoni A, Andrey DO, Monod A, Lew DP, *et al*. The posttranslocational chaperone lipoprotein PrsA is involved in both glycopeptide and oxacillin resistance in *Staphylococcus aureus*. *Antimicrob Agents Chemother* 2012;56:3629–3640.
85. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* 2012;44:243–246.